Validation and Model Selection: Three Similarity Measures Compared

Robert E. Marks

Economics, University of New South Wales, Sydney, Australia

6 Vincent Street, Balmain NSW 2041, Australia
E-mail: r.marks@unsw.edu.au
url: http://www.agsm.edu.au/bobm

There are two types of simulation models: Demonstration models, essentially existence proofs for phenomena of interest, and Descriptive models, that attempt to track dynamic historical phenomena. Both types require verification. Descriptive models require validation against historical data as well. More broadly, we can think of a process of choosing the "best" of several models. This paper examines three measures of the similarity of two sets of vectors, here time series. The best known but flawed is the Kullback-Leibler information-theoretic construct. A second measure is what I have called the State Similarity Measure. The third measure is a set-theoretic measure of similarity, the Generalized Hartley Metric. For illustration, we use data from a dynamic simulation model of historical brand rivalry.

Keywords: model validation, State Similarity Measure, Kullback-Leibler, generalized Hartley metric.

1. Introduction

Critics of simulation models and modelling argue that, since modellers can make any assumptions they wish, such models are little more than toys. But serious modellers see their models as tools in the scientific enterprise. There are two types of simulation models: *Demonstration models*, essentially existence proofs for phenomena of interest, and *Descriptive models*, that attempt to track dynamic historical phenomena. Most early simulation models were demonstrative (or qualitative), such as Schelling's (1971) segregation model. Although demonstrative simulation models are useful, not least at performing "what if" exercises of exploration of different models, policy analysis requires validated, descriptive simulation models. Both types require verification. Descriptive models require validation against historical data as well. But validation of any but very simple simulation models has been slow in appearing in the literature.

This paper is an attempt to provide a new tool in validating serious simulation models: a means of measuring the distance (or similarity) between pairs of sets of vectors, such as time-series data. It outlines a new technique, the State Similarity Measure, for tackling the fourth core issue of Fagiolo et al. (2007): validating agent-based models

using historical data. The SSM can measure the distance between two sets of vectors, here time-series vectors; in effect, it measures the row-wise distance between pairs of matrices.

What is model validation? Surely it's an attempt to assure the reader that the model is "good" at being able to generate the observed data. Looked at from an information-theoretic framework instead of a statistical perspective, the observed data contain information, and the models we develop (from our theoretical understanding of the underlying processes generating the observed data) can be thought of as attempts, in one sense, to express this data in as compact a form as possible via a model. Paraphrasing Burnham and Anderson (2002, pp. 437): such a model represents a hypothesis and is then a basis for making inferences about the process or system that generated the observed data. All simulation models are existence proofs (Marks, 2007): there exists at least one model — this one — that is sufficient to generate data "close" to the observed data. Necessity is harder to establish.

A given set of observed data contains only a finite, fixed amount of information. The ultimate goal of modelling is to derive a model (or set of models) that produces the identical set of output data.[1] If this were achieved — although it's realistically unattainable — then no information would be lost in going from the observed data to a model of the information in the data. Since models are only approximations of reality, the idealised goal of a complete and accurate model is unattainable, and often undesirable because of over-fitting. With several contending models, validation might be able to point the researcher to the "best" model, in the sense that it loses least information.

We take a pluralist, realist approach, in which we compare models by measuring the distance between each model's brand price output traces and the historical brand price traces of the real world, in order to choose the best model. In doing so, our method is closest to the "indirect calibration approach" of Fagiolo et al. (2007): we focus on a single market (micro), using empirical data to validate our models' simulated outputs, although we do not then indirectly calibrate since our purpose here is to introduce our new SMM measure for comparing sets of vectors, such as time series, not to outline a full validation technique.

2. Our Simulation Model

To illustrate and compare the measures discussed here, we analyse historical data of markets in which rivalry among brands of vacuum-packed, canned, ground coffee results in a dynamic rivalrous dance, with abrupt changes in weekly prices and sales volumes, as shown in Figure 1,[2] from Midgley et al. (1997), part of an on-going research program

[1] This is case (e) in Marks (2007, Figure 2): the model is complete and accurate.

[2] These historical data represent the weekly prices and sales of nine brands in a supermarket chain over 50 weeks. We focus on the three most strategic of these: Folgers (red), Maxwell House (purple), and Chock Full O Nuts (green).

(see Midgley et al., 2007).

FIGURE 1 HERE.

We model this market behaviour as the heterogeneous brands choosing next week's price as a function of the state of the market, which is defined to include each of this week's prices (and possibly other marketing actions), but might also include the prices (and actions) of past weeks, depending on the brands' depths of memory.

We model the price P_{bw} of brand b in week w as a function f_b of the state of the market M_{w-1} at week $w-1$, where M_{w-1} in turn might be a product of the set of weekly prices S_{w-j} of all brands over several weeks (depending on the depth of memory), as shown in the following equation (see Appendix 1 for an example of deriving states of the market from sets of prices with different depths of memory):

$$P_{bw} = f_b(M_{w-1}) = f_b(S_{w-1} \times S_{w-2} \times S_{w-3} \cdots)$$

In our research program (Midgley et al., 1997), we use the Genetic Algorithm to search for "better" (i.e. more profitable) brand-specific mappings, f_b, from market state to pricing action, where each brand's fitness function (the maximand) is its weekly profit. We use estimates of each brand's cost function and the market's (asymmetric) response to each brand's price in any week, given its rivals' prices.[3]

Since we do not use the historical time series to estimate the parameters of the model, we can use the historical data as a yardstick against which to measure the performance of our models: this multi-model selection should allow us to determine the best of our GA-determined models, in a process of model selection masquerading as model validation. That is, following Burnham and Anderson (2002), we expand validation — asking whether the model output is close to the historical data — to choose one of several models which is "best," where this means the model that captures as much of the information in the historical data set as possible.

Can we measure the degree of similarity of the historical data to the output from a model of the phenomenon? If so, then we can use the measure to compare simulation models and choose the "best" model: that which generates a set of time series as its output which is "closest" to the set of historical time series.

We discuss three possibilities for deriving such a measure of similarity: first, the Kullback-Leibler construct, which is closely related to Shannon's information-theoretic measure of entropy; second, a new measure of the author's, the State Similarity Measure; and, third, a set-theoretic measure derived by Klir from an early measure of Hartley's, the Generalized Hartley Measure.

[3] Our program models each brand as an independent profit maximiser, seeking brand-specific mapping functions from market state to next period's price for that brand, to maximise that brand's profit, subject to supermarket moderation and to brand exhaustion. The GA is used in a co-evolving interaction among the strategic brands, each brand separately seeking the best mapping function, where the brands' costs and market demand responses are idiosyncratic. Midgley et al. (1997) describe this in detail.

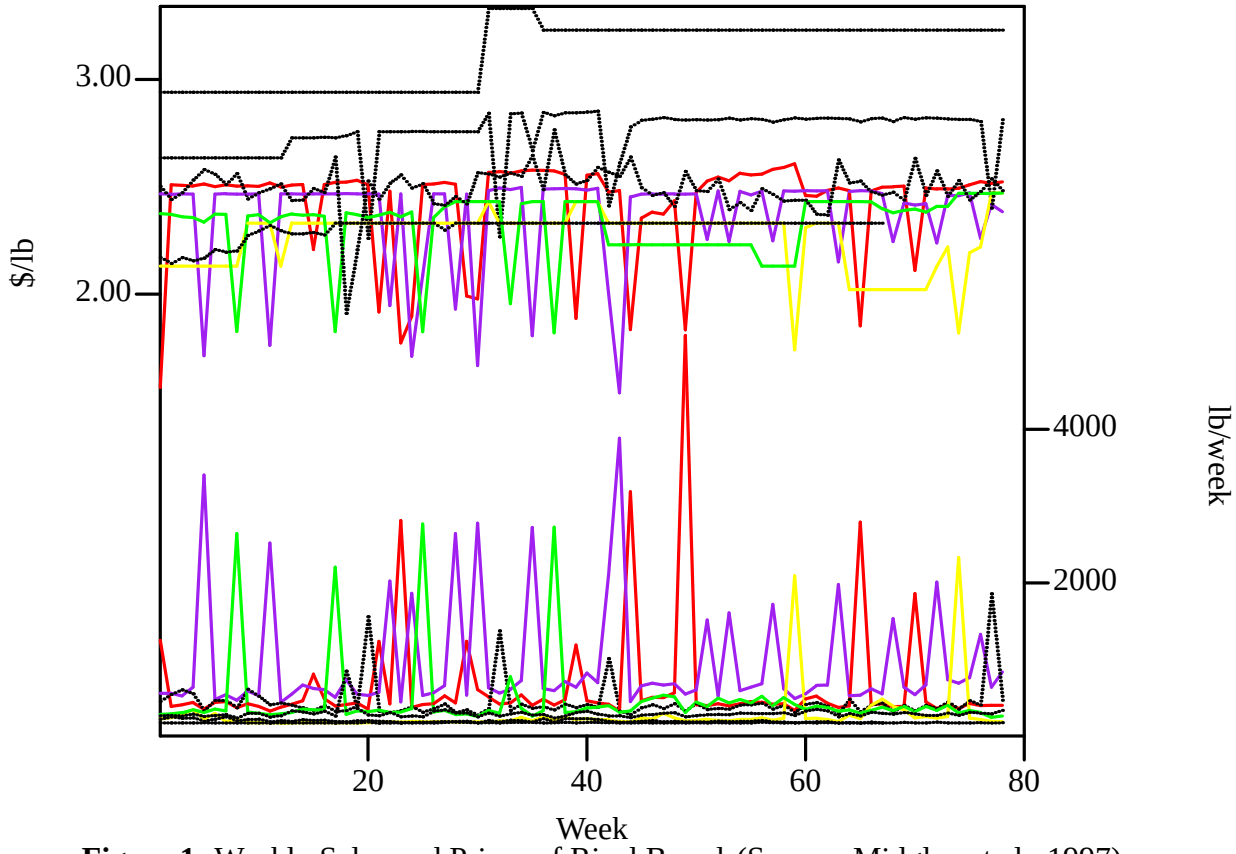

Figure 1: Weekly Sales and Prices of Rival Brands(Source: Midgley et al., 1997)

Before we discuss these three possible measures of similarity between historical data and the simulation models' outputs, we discuss how we can simplify the "rivalrous dance" into numerical measures, which can then be compared.

3. Defining States of the Historical and Simulated Markets

We focus on the brands' prices, although other marketing actions might also be used (Midgley et al., 1997). We face a curse of dimensionality with the historical data: every week, each brand can price anywhere[4] between about $1.50 per pound and $3.40 per pound, a choice of about 190 price points.

In modelling this, Marks and Midgley (1995) reduced the possible price points in the model to four: three high and one Low. To reduce the dimensionality still further, we here use a dichotomous partition of the historical prices: any brand's weekly price below that brand's midpoint price is designated "Low", and any price above that midpoint is "High." We define the state of the market in any week by the combination of the partitioned prices of the three strategic brands, Folgers, Maxwell House, and Chock Full O Nuts.[5] The number of possible states depends on the way in which the inputs and outputs are partitioned: the partitioning (coarsening) of the price space, and the depth of memory of the players (brands). We use dichotomous price partitioning in any week, so that with three brands, each pricing Low or High, there are $2^3 = 8$ possible states per week. Figure 2 shows the historical prices of the three strategic brands, after equivalent dichotomous partitioning, where each colour represents a distinct brand, and the time periods are weeks. These are the historical data against which we compare our models' outputs, using SSM and GHM.

FIGURE 2 HERE.

Figure 3 shows the non-partitioned output from one of our derived simulation models (Model 26a) with these three strategic brands.

FIGURE 3 HERE.

With two-week memory, there are $8^2 = 64$ possible states in any week, when each brand's response next week to actions this week and last can be considered; with three-week memory there are $64^2 = 512$ possible states, where third-order responses (a brand's response next week to others' actions this week, last week, and the week before) can be considered.

The distribution of the eight possible 1-week states in the historical chain store (H) from Figure 2, and in three models (11, 26a, 26b) of the models' outputs from Figure 3, using 50 weeks of data, are shown. in Table 1 below, "0" corresponding to a "High" price and "1" to a "Low" price. Modelling deeper memory for the brands results in

[4] Anywhere, that is, subject to moderation by the supermarket, to prevent more than one brand pricing Low in any week, and to prevent any brand pricing Low two weeks in succession. See Midgley et al. (1997).

[5] As seen in Figure 1, these three brands are the most dynamic in their pricing; they constitute an average market share of 77%.

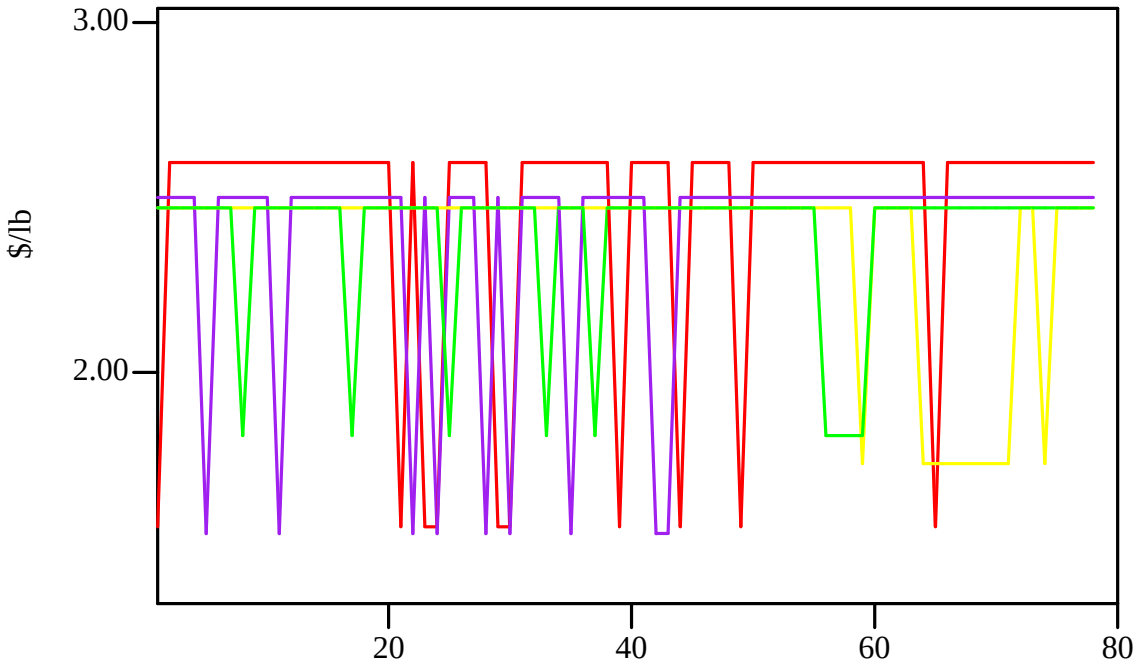

Figure 2: Partitioned Historical Weekly Prices of Four Brands

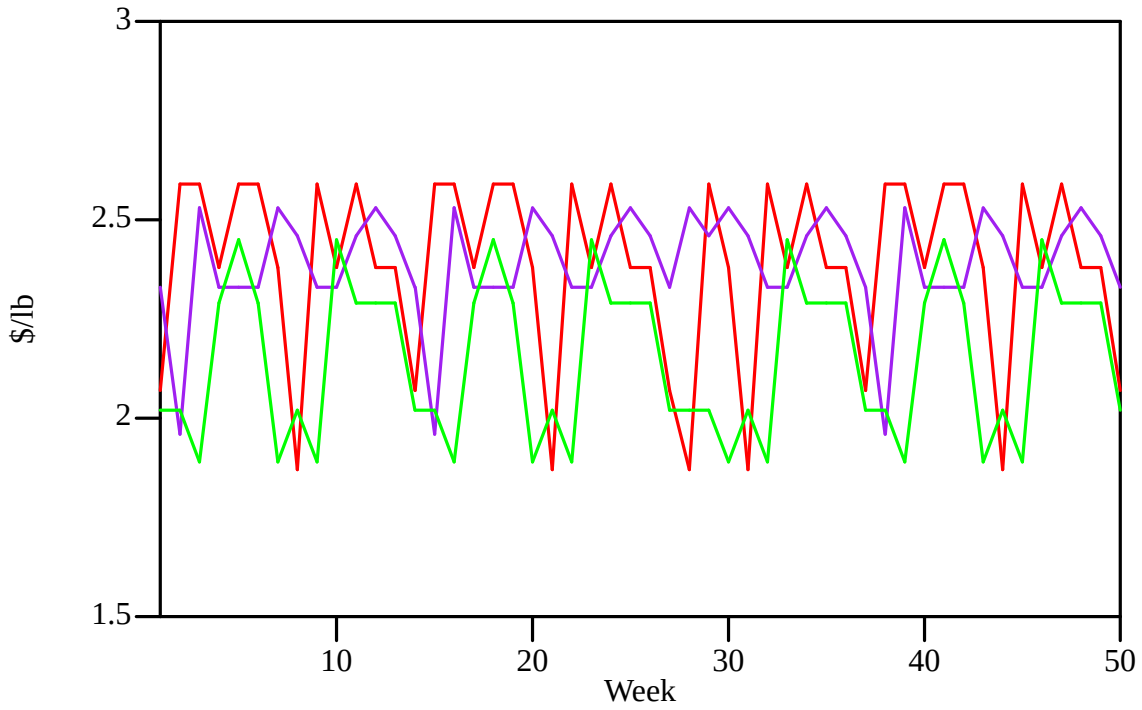

Figure 3: Prices from a Simulated Oligopoly (Marks et al., 1995)

Table 1
State frequencies from History (Chain 1) and three models.

State	History (Chain 1)	Model 11	Model 26a	Model 26b
000	32	0	30	20
001	2	18	11	10
010	6	15	3	7
011	1	0	0	0
100	7	16	5	12
101	0	0	0	0
110	2	0	1	1
111	0	1	0	0
Total	50	50	50	50

similar distributions, but the tables are 64 rows and 512 rows deep, with 2-week and 3-week memory, respectively.

How significant is the degree of partitioning? In Marks (2010), we used three-week memory and dichotomous price partitioning throughout, and for the historical data (from seven supermarket chains), considered three-brand and four-brand interaction. Comparing the outputs of three simulation models with a single set of historical time series of price (from Chain 1), we used three-brand interactions, with 1-, 2-, and 3-week memory.

4. Measures of Closeness or of Information Loss

A variety of proposals have been made to measure similarity, but most of them have been inspired by two measures: Shannon entropy and Hartley information. Shannon (1948) entropy (SE) is based on probability and can be defined as:

$$SE(p(x)|x \in X) = -\sum p(x)\log_2(p(x)),$$

where p is a probability distribution of random variable x. Function SE fulfills some useful properties such as additivity, branching, normalization and expansibility. Shannon entropy led to the Kullback-Leibler (1951) measure of information loss from historical to model, which has some attractions theoretically, but is not a true metric, as we shall see.

4.1. Kullback-Leibler information loss

The Kullback-Leibler (K-L) information loss provides a measure of the information lost when model g is used to approximate full reality f:

$$I(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x|\theta)}\right)dx$$

in the continuous version, where the models g are indexed by θ, or

$$I(f,g) = \sum_{i=1}^{k} p_i \times \log\left(\frac{p_i}{\pi_i}\right)$$

in the discrete case, with full-reality f distribution $0 < p_i < 1$, and model g distribution $0 < \pi_i < 1$, with $\sum p_i = \sum \pi_i = 1$. Here, there are k possible outcomes of the underlying process; the true probability of the ith outcome is given by p_i, while the $\pi_1, \ldots, \pi_k$ constitute the approximating model. Hence, f and g correspond to the p_i and π_i, respectively.

But the K-L information loss is not a true metric: it is not symmetric and does not satisfy the triangle inequality, since $I(f,g) \neq I(g,f)$: it is a semi-quasimetric. Moreover, both π_i and p_i must be positive[6] , while in our data, even for the coarse, dichotomous partition we are considering, one or both of these values is likely to be zero.[7]

Alternative measures are the author's State Similarity Measure (which uses rectilinear or Minkowski's L_1 or the cityblock distance), and Klir's Generalized Hartley Measure.

5. The State Similarity Measure (SSM)

The SSM counts the absolute difference in the frequency of each possible state in each of two sets of vectors (or time series), and sums these to obtain the SSM for the pair of sets of vectors.[8] In effect, SSM treats each time series set as a vector $\mathbf{p}$ in an n-dimensional, non-negative, real vector space with a fixed Cartesian coordinate system, where there are n possible states in the sets of vectors. The SSM between two sets $\mathbf{P}$ and $\mathbf{Q}$ of vectors (or time series) is calculated as the rectilinear or cityblock distance (Krause 1986) d_1 between their two constructed vectors $\mathbf{p}$ and $\mathbf{q}$, given by $d_1^{\mathbf{PQ}} = d_1(\mathbf{p},\mathbf{q}) =$

[6] The K-L measure is defined only if p_i = 0 whenever π_i = 0.

[7] As Akaike (1973) first showed, the negative of K-L information is Boltzmann's entropy. Hence minimizing the K-L distance is equivalent to maximizing the entropy; hence the term "maximum entropy principle." But, as Burnham & Anderson point out, maximizing entropy is subject to a constraint—the model of the information in the data. A good model contains the information in the historical data, leaving only "noise." It is the noise (or entropy or uncertainty) that is maximized under the concept of the entropy maximizing principle. Minimizing K-L information loss then results in an approximating model g that loses a minimum amount of information in the data f. The K-L information loss is averaged negative entropy, hence the expectation with respect to f.

Fagiolo et al. (2007, p. 211) note further that "K-L distance can be an arbitrarily bad choice from a decision-theoretic perspective ... if the set of models does not contain the true underlying model ... then we will not want to select a model based on K-L distance." This is because "K-L distance looks for where models make the most different predictions—even if these differences concern aspects of the data behaviour that are unimportant to us."

[8] I am grateful to Daskalova, who pointed out to me that the SSM is a version of what has been called cityblock distance, rectilinear distance, or taxicab geometry.

Table 2
SSMs calculated between the six pairs of sets.

	Pair	1-week memory	2-week memory	3-week memory
a	History (Chain 1), Model 11	70	88	92
b	History (Chain 1), Model 26a	18	36	54
c	History (Chain 1), Model 26b	28	48	68
d	Model 11, Model 26a	62	76	88
e	Model 11, Model 26b	42	60	80
f	Model 26a, Model 26b	22	42	60

$\sum_{i=1}^{n} |p_i - q_i|$, where p_i is the number of occurrences (or frequencies) of state i in vector set $\mathbf{P}$. That is, SSM is the sum of the absolute differences of the coordinates of the two sets of vectors as n-dimensional constructed vectors. (See the Appendix 1 for details of this procedure.)

We use three models from simulations undertaken in Marks et al. (1995). Each model has three interacting brands, and each brand agent independently chooses its price from its own set of four possible prices in order to maximise its weekly profit, in a process of co-evolution using the Genetic Algorithm. With 1-week memory, each agent's action is determined by the state of the market in the previous week, which means $4^3 = 64$ possible market states for each agent to respond to. The GA chooses the mapping from perceived state to action for each brand (with weekly profit as its "fitness").

Each model of the three brands' interactions 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 model.[9] Figure 3 (above) shows a fifty-week period of simulated interactions among three brand agents (Brands, 1, 2, and 5) in Model 26a, where each brand chooses from one of four possible prices per week.

The six pairs of SSMs between the partitioned prices of the three models‘ and the historical data (from Chain 1), using 50-week data series, are presented in Table 2 for 1-, 2-, and 3-week memory:[10]

Characteristics of the SSM measure (Marks, 2010): First, an SSM of zero means that the two sets of vectors are identical; larger SSMs imply less similarity. Second, the

[9] The three models differ in more than the frequencies of the eight states (Table 1): each model contains three distinct mappings from state to action, and, as deterministic finite automata (Marks, 1992), they are ergodic, with emergent periodicities. Model 26a has a period of 13 weeks, Model 26b of 6 weeks, and Model 11 of 8 weeks. It is not clear that the historical data exhibit ergodicity, absence of which will make simulation initial conditions significant (Fagiolo et al., 2007). Initial conditions might determine the periodicity of the simulation model.

[10] In Marks (2010) there was a bug in the code used to calculate the states of the three simulation models (11, 26a, and 26b), now corrected.

maximum D of an SSM measure occurs when the intersection between the states of the two sets of vectors is null, with $D = 2 \times S$, where S is the number of window states, which depends on the memory length, inter alia. Here, maximum D would be 100 for 1-week memory, $2 \times 49 = 98$ for 2-week memory, and $2 \times 48 = 96$ for 3-week memory, (given that there are 50 observations per set of time series). Third, we can, using Monte Carlo stochastic sampling (Marks, 2014), derive some statistics to argue that any pair of sets is not likely to include random series (see below).

As the partitioning becomes finer (with deeper memory of past actions), the SSMs increase as the two sets of vectors (or time series) become less similar. This should not surprise us. We also note that with these four sets of time series, the rankings do not change with the depth of memory: (from closer to more distant) (Chain 1, Model 26a), (Model 26a, Model 26b), (Chain 1, Model 26b), (Model 11, Model 26b), (Model 11, Model 26a), and (Chain 1, Model 11). Which of the three models is closest to the historical data of Chain 1? The SSM tells us that Model 26a is best, followed by Model 26b, with Model 11 bringing up the rear.

As defined here, the SSM is an absolute measure, where its maximum distance D is a function of the equal length of the pair of sets of vectors. The lower the SSM, the closer the two sets of vectors. It is possible to define a normalised measure, call it SSMN, where SSMN is between 0% and 100%:

$$SSMN \equiv 100 \times \left(1 - \frac{SSM}{D}\right),$$

where D is the maximum SSM distance apart of the two sets of vectors, equal to the length of each vector.[11] Hence SSMN = 100 implies identity between the two sets, and SSMN = 0 means maximum distance between the two, with null overlap.[12]

5.1. Monte Carlo simulations of the SSM

Table 3 presents the distances between historical Chain 1, and the three simulations, Model 11, Model 26a, and Model 26b from Marks et al. (1995), with 3-week memory. Model 11 is far from any of the other sets, and Model 26b is closest to Model 26a, but Model 26a is closer to the Chain 1 historical data (at 54/96) than it is to the closest other simulation, Model 26b (at 60/96). (Note: * in the Table indicates we cannot reject the null at the 5% level.)

Null Hypothesis: each of two sets of time series is random.

With this null hypothesis, we can set 1% and 5% one-sided confidence intervals to the SSM numbers. With three brands and S = 48, the maximum D is 96. 95% of pairs of sets of three random time series are at least 80 apart, and 99% of pairs of sets

[11] In our analysis, this length is a function of the number of observations (equal for historical and simulated data) and the depth of memory.

[12] I am grateful to a reviewer for suggesting this.

Table 3
SSMs between Historical Chain 1 and Three Models

	Chain 1	Model 11	Model 26a	Model 26b
Chain 1	0	92*	54	68
Model 11	92*	0	88*	80*
Model 26a	54	88*	0	60
Model 26b	68	80*	60	0

of three random time series are at least 76 apart.[13] This means that, in Table 3, we reject the null hypothesis of random data for the pairs (Chain 1, Model 26a), (Chain 1, Model 26b), and (Model 26a, Model 26b), since all SSMs here are less than 76, so the data are significantly non-random, and the null hypothesis is rejected. The other three pairs (all comparisons with Model 11), with SSMs above 80, are 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 (see Table 1).

We show these results in Figure 4, which plots the Cumulative Mass Function of the MC parameter bootstrap simulation against the six SSMs of the pairs.[14] 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 4 HERE.

The one-sided confidence interval at 1% corresponds to a SSM of 76, and at 5% 80. That means that any SSM above 80 could have resulted (with a 5% probability) from two sets of random vectors; or above 76 with a 1% probability. The higher the SSM, the greater the likelihood that the two sets are random, as is seen with the rising CMF. For SSMs to the left (below 76) we reject the random hypothesis. Thus we cannot reject the null hypothesis (random sets) for any pairs comparing Model 11; but reject the null (random) hypothesis for the other three pairs.

6. Classical Possibility Theory

Possibility theory offers a non-additive method of assigning a numerical value to the likelihood of a system assuming a specific state, one of a given set of states. The likelihood expressed is that of *possibility*; for this reason, the possibility assigned to a collection of possible events is the maximum (rather than the sum) of the individual possibilities (Ramer, 1989).

Hartley (1928) solved the problem of how to measure the amount of uncertainty

[13] 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.

[14] The CMF of a discrete random variable X is defined as $F_X = \Pr(X \leq x)$, where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x. Here, X is the SSM between two sets of random numbers.

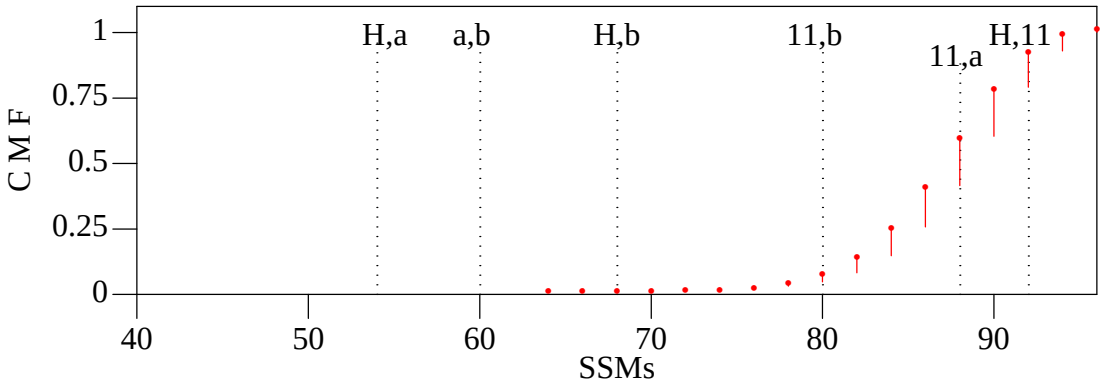

Figure 4: Chain 1 and Three Models; SSMs against Random CMF.

associated with a finite set E of possible alternatives: he proved that the only meaningful way to measure this dichotomous amount (when any alternative is either in or out: no gradations of certainty) is to use a functional of the form:

$$c \log_b \sum_{x \in X} |E|,$$

where set E contains all possible alternatives from the larger (finite) set X, and where $|E|$ denotes the cardinality of set E: b and c are positive constants, and it is required that $b \neq 1$. If b = 2 and c = 1 (or more generally, if $c \log_2 = 1$), then we obtain a unique functional, H, defined for any basic possibility function, r_E, by the formula:

$$H(r_E) = \log_2 |E|,$$

where the measurement unit of H is bits. This can also be expressed in terms of the basic possibility function r_E as

$$H(r_E) = \log_2 \sum_{x \in X} r_E(x).$$

H is called a *Hartley measure* of uncertainty, resulting from lack of specificity: the larger the set of possible alternatives, the less specific the identification of any desired alternative of the set E. Clear identification is obtained when only one of the considered alternatives is possible. Hence this type of uncertainty can be called *non-specific*.

This measure was first derived by Hartley (1928) for classical possibility theory, where any alternative element of set X is either possible (i.e. in set E) or not. The basic possibility function, r_E, is then

$$r_E(x) = \begin{cases} 0 & \text{when } x \in E, \\ 1 & \text{when } x \notin E. \end{cases}$$

and is derived explicitly in Klir (2006, pp. 28). To be meaningful, this functional must satisfy some essential axiomatic requirements.[15]

6.1. The Generalized Hartley Measure (GHM) for Graded Possibilities

Following Klir (2006), we relax the "either/or" characteristic of the earlier treatment and allow the basic possibility function[16] on the finite set X to take any value between zero and one: $r : X \rightarrow [0, 1]$ Note that

$$\max_{x \in X} \{r(x)\} = 1,$$

[15] See further discussion in Appendix 2.

[16] It is not correct to call function r a possibility *distribution* function, since it does not distribute any fixed value among element of the set X: $1 \leq \sum_{x \in X} r(x) \leq |X|$.

a property known as possibilistic normalization.

The Generalized Hartley Measure (GHM) for graded possibilities is usually denoted in the literature by U, and is called U-*uncertainty*. U-uncertainty can be expressed in various forms. A simple form is based on notation for graded possibilities: $X = \{x_1, x_2, \ldots, x_n\}$ and r_i denotes for each $i \in \mathbf{N}_n$ the *possibility* of the singleton event x_i. Possibilities can (although need not) be estimated by frequencies. Elements of X are appropriately rearranged so that the possibility profile:

$$\mathbf{r} = \ < r_1, r_2, \ldots, r_n >$$

is ordered in such a way that

$$1 = r_1 \geq r_2 \geq \ldots \geq r_n > 0,$$

where $r_{n+1} = 0$ by convention. Moreover, set $A_i = \{x_1, x_2, \ldots, x_i\}$ is defined for each $i \in \mathbf{N}_n$.

Using this simple notation, the U-uncertainty is expressed for each given possibility profile $\mathbf{r}$ by the formula

$$U(\mathbf{r}) = \sum_{i=1}^{n} (r_i - r_{i+1}) \log_2 |A_i|. \tag{1}$$

Since, clearly

$$\sum_{i=1}^{n} (r_i - r_{i+1}) = 1,$$

the U-uncertainty is a weighted average of the Hartley measure for sets A_i, $i \in \mathbf{N}_n$, where the weights are the associated differences $r_i - r_{i+1}$ in the given possibility profile. These differences are values of the basic possibility assignment function for sets A_i.

Since $|A_i| = i$ and $\log_2 |A_1| = \log_2 1 = 0$, the above equation can be rewritten in the simpler form

$$U(\mathbf{r}) = \sum_{i=2}^{n} (r_i - r_{i+1}) \log_2 i \tag{2}$$

or, alternatively, in the form

$$U(\mathbf{r}) = \sum_{i=2}^{n} r_i \log_2 \left(\frac{i}{i-1} \right). \tag{3}$$

U-uncertainty preserves the ordering of possibility profiles defined on the same set: $U(^1\mathbf{r}) \leq U(^2\mathbf{r})$ for any pair of probability profiles defined on the same set and such that $^1\mathbf{r} \leq^2 \mathbf{r}$. Moreover,

$$0 \leq U(\mathbf{r}) \leq \log_2 |X|$$

for any possibility profile $\mathbf{r}$ on X. The lower and upper bounds are obtained, respectively for the smallest (expressing no uncertainty) and the largest (expressing total ignorance) possibility profiles, $< 1, 0, \ldots 0 >$ and $< 1, 1, \ldots, 1 >$:

$$U(< 1, 0, \ldots 0 >) = 0$$

$$U(< 1, 1, \ldots, 1 >) = \log_2 |X|.$$

6.2. Ordering of Possibility Profiles

Possibility profiles of the same length can be partially ordered in the following way: given any two possibility profiles of length n (where $r_{n+1} = 0$ by construction):

$$^{j}\mathbf{r} =<^{j} r_1, ^{j} r_2, \ldots, ^{j} r_n >,$$

$$^{k}\mathbf{r} =<^{k} r_1, ^{k} r_2, \ldots, ^{k} r_n > .$$

Following Klir (2006), we define

$$^{j}\mathbf{r} \leq^{k} \mathbf{r} \iff {}^{j}r_i \leq^{k} r_i$$

for all $i \in \mathbf{N_n}$. For any $^{j}\mathbf{r}, ^{k}\mathbf{r} \in \mathbf{R_n}$, if $^{j}\mathbf{r} \leq {}^{k}\mathbf{r}$, then $^{k}\mathbf{r}$ represents greater uncertainty than does $^{j}\mathbf{r}$; that is, $^{j}\mathbf{r}$ contains more information than does $^{k}\mathbf{r}$.

Klir (2006, p. 160) notes something relevant to our purposes here: "Another important interpretation of possibility theory is based on the concept of *similarity*, in which the possibility $r(x)$ reflects the degree of similarity between x and an ideal prototype, x_P, for which the possibility degree is 1. That is, $r(x)$ is expressed by a suitable distance between x and x_P defined in terms of the relevant attributes of the elements involved. The closer x is to x_P according to *the chosen distance*, the more possible we consider x to be in this interpretation [our emphasis]."

6.3. Applying U-uncertainty to our data

From the frequencies of Table 1 (one-week memory), we can reorder[17] the possibilities (observed frequencies) of the three runs and the historical data, to get the four reordered, non-normalised[18] possibility profiles:

Using equation (2), the four Hartley measures are calculated:[19]

[17] It might be objected that this reordering loses information. But this overlooks the fact that the order of the states is arbitrary. It should not be forgotten that the definition of the states with more than one week's memory captures dynamic elements of interaction.

[18] Normalisation here means $r_1 = 1$, not $\sum r_i = 1$.

[19] For clarity, we have included the $(i = 1)$th element, $(r_1 - r_2) \log_2 1$, which is always zero, by construction, consistent with equation (2).

Table 4
The four possibility profiles, one-week memory.

History (Chain 1):	32	7	6	2	2	1	0	0
Model 11:	18	16	15	1	0	0	0	0
Model 26a:	30	11	5	3	1	0	0	0
Model 26b:	20	12	10	7	1	0	0	0

Table 5
GHMs calculated for three memory partitions.

Process	1-week memory	2-week memory	3-week memory
History (Chain 1)	0.383	0.495	0.782
Model 11	1.399	2.179	2.787
Model 26a	0.516	0.679	1.085
Model 26b	1.054	1.657	2.542

1. History (Chain 1:)

$$U(\mathbf{r}) = \frac{1}{32}(25\log_2 1 + 1\log_2 2 + 4\log_2 3 + 0\log_2 4 + 1\log_2 5 + 1\log_2 6)$$

$$= 0.383$$

2. Model 11:

$$U(\mathbf{r}) = \frac{1}{18}(2\log_2 1 + 1\log_2 2 + 14\log_2 3 + 1\log_2 4)$$

$$= 1.399$$

3. Model 26a:

$$U(\mathbf{r}) = \frac{1}{30}(19\log_2 1 + 6\log_2 2 + 2\log_2 3 + 2\log_2 4 + 1\log_2 5)$$

$$= 0.516$$

4. Model 26b:

$$U(\mathbf{r}) = \frac{1}{20}(8\log_2 1 + 2\log_2 2 + 3\log_2 3 + 6\log_2 4 + 1\log_2 5)$$

$$= 1.054$$

The GHMs for the three models and Chain 1 have been calculated for the three cases of 1-week, 2-week, and 3-week memory, as seen in Table 5.

These GHMs are true metrics (they satisfy the triangle inequality, unlike the K-L information loss), and so we can compare the differences of Table 6 between the four measures. We can readily see that Model 26a (0.516) is closest to the historical data of Chain 1 (0.374); next is Model 26b (0.516), with Model 11 (1.399) furthest from the

Table 6
GHM differences calculated for the six pairs of sets.

	Pair	1-week memory	2-week memory	3-week memory
a	History (Chain 1), Model 11	1.016	1.684	2.005
b	History (Chain 1), Model 26a	0.133	0.184	0.303
c	History (Chain 1), Model 26b	0.671	1.162	1.760
d	Model 11, Model 26a	0.883	1.500	1.702
e	Model 11, Model 26b	0.345	0.522	0.245
f	Model 26a, Model 26b	0.538	0.978	1.457

historical data. Moreover, we can see that Model 26a is closer to the historical Chain 1 data than it is to Model 26b.

Table 6 shows the six pairwise differences in GHM, derived from Table 5. It can be compared with the six pairwise SSMs of Table 2. For 1-week memory the maximum GHM, corresponding to 50 equi-likely states, is $\log_2 50 = 5.644$; for 2-week memory $\log_2 49 = 5.615$, and for 3-week memory $\log_2 48 = 5.585$. These numbers are the maximum pairwise difference between GHMs; the minimum difference is zero in all three depths of memory.[20]

7. Comparing the distances measured by SSM and GHM

From Table 2, for 1-week memory, the SSMs are ranked (closest to farthest): $\{$b, f, c, e, d, a$\}$; but, from Table 6, the GHM differences are ranked (smallest to largest): $\{$b, e, f, c, d, a$\}$. Model 26a is closest to History using either measure, and Model 11 is farthest, Note, however, from Table 2, that although the SSM rankings are the same for 1-, 2-, or 3-week memory, the GHM rankings are sensitive to the depth of memory. That is, the two methods do not always produce identical rankings, although the degree to which these two measures result in similar rankings of distances is noteworthy, given their quite different foundations.[21]

8. Conclusion

The two measures, SSM and GHM, are true metrics that allow us to measure the degree of similarity between two sets of vectors, here time series. The SSM between two sets of vectors is the absolute distance between two constructed vectors in non-negative, n-dimensional vector space, where n is the number of possible states that each set of vectors can exhibit. GHM is a measure of the possibility of any set $\mathbf{P}$ of vectors occurring as a vector $\mathbf{p}$ in n-dimensional space.

[20] We could also define a normalised GHM, as above.
[21] We postpone exploration of these differences to a later paper.

Since GHM is a metric, differences of sets of vectors' GHMs are meaningful. SSM is also a metric (symmetric and satisfying the triangle inequality). As such, both measures can be used to score the distance between any two sets of vectors, such as sets of time series, which previously was unavailable. The Kullback-Leibler measure, although based in information theory and Shannon's entropy measure, is not a true metric, despite its relationship to maximum likelihood methods.

The two measures, SSM and GHM, 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. The measures also allow us to measure the distance between any two sets of time series and so to estimate the parameters, or to help validate a model against history. The measures can thus be used to identify the model which is "closest" to history, as measured by comparing its output set of time series to the historical set of time series, that is, by identifying the model that has captured most information of the historical set of time series, given our definition of model states.

The SSM and the GHM have demonstrated closeness in measuring similarity of sets of time series, although the two measures' rankings of distances are not identical, as seen above. The SSM is intuitive: it uses the cityblock metric to count up the differences in the states between two constructed vectors. It can be described in six simple steps, as outlined in Appendix 1. The GHM is anything but intuitive, based on arcane possibility theory. It is developed and applied in four pages of mathematics above. Using Occam's Razor, the SSM, as a simpler, more transparent measure, is preferred.

The two measures, SSM and GHM, are not restricted to measuring the similarity of (or distance between) two sets of time series: they are more general, as we have reminded the reader, in that they can be applied to pairs of sets of (equal length) vectors. The data used here are illustrative only: the two measures can be applied to any simulated data and historical data, so long as the number of observations of the model output and the historical data are equal, with equal numbers of vectors, or observations. Even more generally, the two measures can be thought of alternative methods of measuring the row-wise distance between any two matrices of equal dimension.

Acknowledgments

I should like to thank Dan MacKinlay for his mention of the K-L information loss measure, Arthur Ramer for his mention of the Hartley or U-uncertainty metric and his suggestions, and Vessela Daskalova for her mention of the "cityblock" metric. I also thank the organisers of the Global Systems Dynamics & Policy Agent-Based Modeling Workshop at the Sorbonne, September 2011, for their generous support. The comments of the editor and an anonymous referee were very constructive.

Table 7
An example: three brands, 1-, 2-, and 3-week windows.

Week	Brand ($P'_{b,w}$)			1-Week	2-Week	3-Week
	Red	Purple	Green	S_w	M_{2w}	M_{3w}
18	0	0	0	0		
19	0	0	0	0	0	
20	0	0	0	0	0	0
21	1	0	0	4	32	256
22	0	1	0	2	20	160
23	1	0	0	4	34	276
24	1	1	0	6	52	418
25	0	0	1	1	14	116
26	0	0	0	0	1	14
27	0	0	0	0	0	1
28	0	1	0	2	16	128
29	1	0	0	4	34	272
30	1	1	0	6	52	418

Appendix 1: Calculating the SSM

1. First, construct the weekly states of the market: 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 $S_w = P'_{1,w} \times P'_{2,w} \cdots$.
   For a 3-brand time series, $S_w = 4 \times P'_{1,w} + 2 \times P'_{2,w} + P'_{3,w}$. Then construct the windowed states of the market (as in Table 7).
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 $M_w = S_w \times S_{w-1} \times S_{w-2} \cdots$
   For a 3-week window, $M_{3w} = 64 \times S_w + 8 \times S_{w-1} + S_{w-2}$. (The powers of 8 are because, with three brands, there are 8 possible states of the market S_w each week.) This means that for a 2-week memory there are 8^2 possible states, and for a 3-week memory, $8^3 = 512$ possibilities. Table 7 provides an example.
   Then construct the SSM:
4. Count the numbers of each state M_w observed for the set of time series over the given time period. Convey this by an $n \times 1$ vector $\mathbf{p}$, where $p_s \geq 0$ is the number of observations of window state s over the period.
   With T longitudinal observations, the maximum SSM distance apart of two sets of time series is $2 \times (T - w + 1)$, where w is number of weeks remembered. (This would happen when the two sets of states are disjoint.)
5. Subtract the number of observations in set P of time series from the number ob-

served in set Q, across all n possible states; $\mathbf{D^{PQ}} = \mathbf{p} - \mathbf{q}$.

6. Sum the absolute values of the differences across all possible states:

$$d_1^{PQ} = \sum |p_i - q_i|$$

This number $d_1^{PQ} = d_1^{QP}$ is the distance between two time series sets P and Q. This is the State Similarity Measure.

Appendix 2: Properties of the Hartley and Generalized Hartley Measures

8.1. Properties of the Hartley Measure.

Its uniqueness was proved on axiomatic grounds by Rény (1970), who showed that the only functional that satisfies the axioms of Branching, Monotonicity, and Normalisation is $H(n) = \log_2(n)$ (see Klir, 2006, pp. 29).

As Klir explains, the Additivity axiom ($H(n \times m) = H(n) + H(m)$) involves a set with $m \times n$ elements, which can be partitioned into n subsets, each with m elements. A characterization of an element from the full set requires the amount $H(m \times n)$ of information. But we can also proceed in two steps by taking advantage of the partition of the set. First, we characterize the subset to which the element belongs: the required information is $H(n)$. Then we characterize the elements within the subset: the required information is $H(m)$. Since these two amounts of information completely characterize an element of the full set, their sum should equal $H(m \times n)$, as required by the axiom.

The Monotonicity axiom ($H(n) \leq H(n+1)$) is obvious and necessary: when the number of possible alternatives increases, the amount of information needed to characterize any one of them cannot decrease.

8.2. Further Properties of the Hartley Measure

First, it is readily seen that the Hartley measure satisfies the inequalities:

$$0 \leq H(E) \leq \log_2 |X|$$

for any $E \in$ the power set $P(X)$: the lower bound is obtained when only one of the alternatives is possible; the upper bound is obtained when all alternatives are equally possible — complete ignorance.

If a given set of possible alternatives, E, is reduced by the outcome of an action to a smaller set $E' \subset E$, then the amount of information $I_{(A:E \to E')}$ generated by the action $A : E \to E'$ is measure by the difference $H(E) - H(E')$:

$$I_{A:E \to E'} = \log_2 \left(\frac{|E|}{|E'|} \right) = \log_2 |E| - \log_2 E'| = H(E) - H(E').$$

When there is only a single possibility (i.e. $|E'| = 1$), then we get $I_{A:E \to E'} = \log_2 |E| = H(E)$, so that $H(E)$ can also be viewed as the amount of information needed to characterize a single element of set E.

Hartley information (I) is based on set theory: given a finite set, X, of cardinality n, where each element represents a possible alternative to select, a sequence is defined as a set s selected alternatives, and therefore there can be n^s possible sequences. Hartley information is defined as:

$$I(n^s) = I(N) = \log_2(N).$$

8.3. Properties of the GHM

U-uncertainty is a special case of a family of functions that (as does Shannon's entropy) satisfy the five properties of Additivity, Subadditivity, Expansibility, Symmetry, and Continuity, together with a Branching property (Ramer 1989).

Klir (2006, p. 200) describes a definition of U-uncertainty without requiring an ordered possibility profile: Let $\mathbf{r} =< r(x)|x \in X >$ denote a possibility profile on X that may not be ordered and let

$$^{\alpha}r = \{x \in X | r(x) \geq \alpha\}$$

for each $\alpha \in [0, 1]$. Then

$$U(\mathbf{r}) = \int_0^1 \log_2 |^{\alpha}r| d\alpha \tag{4}$$

Klir (2006) provides proofs and elaborations of generalized possibility theory and Hartley functions. His Table 5.1 (p. 159) compares the mathematical properties of probability theory versus possibility theory for finite sets: Basis of each measure, Body of Evidence, Unique Representation, Normalisation, Additivity or Max/Min Rules, Total Ignorance, Conditionalities, Non-interactions, and Independence. Moreover, Ramer (1987) shows that the generalized Hartley function satisfies all the usual axioms (see Klir, 2006, pp. 197) of an information measure: Subadditivity, Additivity, Monotonicity, Continuity, Expansibility, Symmetry, Range, Branching/Consistency, Normalisation, and Coordinate Invariance.

References

[1] Akaike, H. (1973). "Information theory as an extension of the maximum likelihood principle," in B.N. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory.* Budapest: Akademiai Kiado, pp. 267–281.

[2] Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd. ed., New York: Springer.

[3] Fagiolo, G., 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(3): 195–226.

[4] Hartley, R.V.L. (1928). "Transmission of information." *The Bell System Technical Journal*, 7(3): 535–563.

[5] Klir, G.J. (2006). *Uncertainty and Information: Foundations of Generalized Information Theory*, New York: Wiley.

[6] Krause, E.F. (1986). *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*, New York: Dover. (First published by Addison-Wesley in 1975.)

[7] Kullback, J.L. and Leibler, R.A. (1951). "On information and sufficiency," *Annals of Mathematical Statistics*, 22: 79–86.

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

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

[10] Marks, R.E. (2010). "Comparing Two Sets of Time-Series: The State Similarity Measure," In *2010 Joint Statistical Meetings Proceedings – Statistics: A Key to Innovation in a Data-centric World*, Statistical Computing Section. Alexandria, VA: American Statistical Association, pp. 539–551.

[11] Marks, R.E. (2014). "Monte Carlo," in *The Palgrave Encyclopaedia of Strategic Management*, edited by David Teece and Mie Augier, London: Palgrave, forthcoming.

[12] Marks, R.E., Midgley, D.F., and Cooper, L.G. (1995). Adaptive behavior in an oligopoly, *Evolutionary Algorithms in Management Applications*, ed. by J. Biethahn and V. Nissen, (Berlin: Springer-Verlag), pp.225–239.

[13] Midgley, D.F., Marks, R.E., and Cooper, L.G. (1997). "Breeding competitive strategies," *Management Science*, 43(3): 257–275, March.

[14] Midgley, D.F., Marks, R.E., and 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(8): 884–893, August.

[15] Ramer, A. (1987). "Uniqueness of information measure in the theory of evidence," *Fuzzy Sets and Systems*, 24(2): 183–196.

[16] Ramer, A. (1989). "Conditional possibility measures," *International Journal of Cybernetics and Systems*, 20: 233–247. Reprinted in D. Dubois, H. Prade, and R. R. Yager, (eds.) *Readings in Fuzzy Sets for Intelligent Systems*, San Mateo, Calif.: Morgan Kaufmann Publishers, 1993, pp. 233–240.

[17] Rényi, A. (1970). *Probability Theory*, Amsterdam: North-Holland (Chapter 9, "Introduction to information theory," pp. 540–616).

[18] Schelling, T.C. (1971). "Dynamic models of segregation," *Journal of Mathematical Psychology*, 1: 143–186.

[19] Shannon, C.E. (1948). "A mathematical theory of communication," *Bell System Technical Journal*, 27: 379–423, 623–656, July, October.