Validation Metrics: A Case for Pattern-Based Methods

Robert E. Marks

Economics, University of New South Wales, Sydney, Australia

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

Abstract: This chapter discusses the issue of choosing the best computer model for simulating a real-world phenomenon through the process of validating the model's output against the historical, real-world data. Four families of techniques are discussed that are used in the context of validation. One is based on the comparison of statistical summaries of the historical data and the model output. The second is used where the models and data are stochastic, and distributions of variables must be compared, and a metric used to measure their closeness. After exploring the desirable properties of such a measure, the paper compares the third and fourth methods (from information theory) of measuring closeness of patterns, using an example from strategic market competition. The techniques can, however, be used for validating computer models in any domain.*

Keywords: model validation, State Similarity Measure, Area Validation Metric, Generalized Hartley metric.

1. Introduction

Validation of a computer model broadly means determining whether the model is behaving as expected, given the modeller's knowledge of the real-world phenomenon being modelled; validating can aid in the choice of the best model, as discussed below. This chapter uses the example of agent-based models. Agent-based computer simulations (or multi-agent systems) are a special case of computer simulations which model autonomous or semi-autonomous rule-based agents dynamically interacting out of equilibrium, for the purpose of observing the emergence of patterns of behaviour at the micro (agent) level or at a higher, macro (group) level which might not otherwise be predicted.[1] For agent-based models, validation poses special issues since the emergent behaviour of such models might be previously unobserved or unexpected. This chapter explains tech-

* To appear in Beisbart, Claus, and Nicole J. Saam (eds.), *Computer Simulation Validation. Fundamental Concepts, Methodological Frameworks, Philosophical Perspectives* Cham: Springer International Publishing.

[1] For an overview of types of computer simulation modelling, see Gilbert & Troitzsch (2005).

niques of validation for such models, in particular the choice of a validation metric.[2] But the metrics to be discussed below are applicable in principle to validation of computer models against observed data — time series or cross-sectional — of applications in many fields in engineering, science, computer science, or the social sciences. Indeed, any phenomenon in which one set of multivariate variables is compared against another, time-series or cross section.[3]

Chapter 30 by Fagiolo et al. in this volume presents a clear overview of validation of agent-based simulation models.[4] They remark that there are many kinds of validation or validity: e.g. output validation, structural validation, theoretical validity, model validity, and operational validity. The simulation model is an attempt to include the relevant variables in a mechanism to reflect the behaviour and hopefully to explain the phenomenon being examined. The phenomenon exhibits a certain (historical) behaviour; the simulation model can generate simulated behaviour. How closely the simulation model's behaviour reflects the observed behaviour is one measure of how well the simulation model reflects the phenomenon being modelled (output validation). Another is to identify the causal structures underlying the real-world phenomenon, as revealed in the historical data, and to compare them with the causal structures of the simulation model or models. This chapter focuses on output validation, asking how well do the model data track existing real-world data, possibly micro (at the agent level), possibly macro (at the aggregate level).[5]

In this chapter, four broad families of measures that can be used in this respect will be explained: what might be called *empirical likelihood measures*, so-called *stochastic area measures*, so-called *information-theoretic measures*, and *pattern-based* or *strategic state measures*. There are trade-offs associated with these families of measures, and several metrics, so far, have been devised for each.[6]

In the fourth family, we describe in detail two metrics (the State Similarity Measure, SSM, and the Generalized Hartley Measure, GHM) which are applicable to validation of models the output of which is multivariate patterns, unlike other methods which assume univariate variables. The two measures can be thought of alternate methods of measuring the row-wise distance between any two matrices of equal dimension, $\mathbf{X}$ and $\mathbf{Y}$.

[2] We distinguish between broader *measure* and narrower *metric* — a metric is a measure, but a measure is not necessarily a metric — as discussed in Section 2.1.

[3] See Marks (2007), Midgley et al. (2007), Oberkampf & Roy (2010), and Liu et al. (2010) for further general discussions of validation.

[4] As Guerini & Moneta (2017) observe, the appearance of many measures to validate agent-based simulation models is an indication of "the vitality of the agent-based community."

[5] This chapter, in effect, focuses on techniques of output validation (Fagiolo et al.'s sections 4.2, 5.1 and 5.2), going into greater detail about three of the six measures they discuss.

[6] This chapter puts the work of Marks (2013) into a wider context.

2. Validation metrics

As an example from the social sciences, consider the interactions over time among several brands, where each brand's market decisions (prices, promotions, etc.) in any period affect the other brands' volumes sold and profits, and the other brands respond with their own market decisions in the following period. (See Section 5 below.) This "rivalrous dance," as I have called it, generates a complex dynamical pattern of prices, profits, volumes sold, etc. The problem is not specific to simulation models and phenomena in the social sciences: researchers in the biological sciences face the same issue and have made some seminal advances in our understanding of the issues (Mankin et al. 1977).

2.1. Four types of measurement scales

The variables compared in Ferson et al. (2008) and Roy & Oberkampf (2011), like almost all variables in scientific and engineering validation, share one property: they are *interval scales*. That is, they measure ordered magnitudes, defined so that the intervals of pairs of variables can be compared, or measured. (They could be *ratio scales*, such as Kelvin for temperature, with absolute zero and where ratios are meaningful,[7] but this is less common.)

Almost all validation methods in finance, science, and engineering are applicable to interval-scaled variables, but not to *order-scaled variables*, in which the variables might be increasing (decreasing) in one (or the other) direction but where distances in these directions are meaningless because order is their highest characteristic. And such validation methods cannot be applied when the variables are *nominal scales* only: when their order is arbitrary, and their highest characteristic is unique identity, with arbitrary, separate names or numbers.

The main focus of this chapter is on methods of validation which can deal with nominal-scaled variables, or *patterns*, such as those that are seen in the historical phenomena (and the computer programs written to simulate them) described in Section 5 below.

Two metrics in particular — the State Similarity Measure and the Generalized Hartley Measure — have been developed to deal with nominal-scaled data. These can be thought of as generalizations of the interval-scale-based metrics of Ferson et al. (2008). They also overcome an issue that does not arise with interval-scaled variables: the disappearance of any state in one but not the other of the two sets of matrices $\mathbf{X}$ and $\mathbf{Y}$: interval-scaled measures do not exhibit gaps in which one state appears in $\mathbf{X}$ but not in $\mathbf{Y}$, or vice versa.

[7] A temperature of 100K is twice as hot as 50K, but 100°C is not twice as hot as 50°C: K is a ratio scale, but °C is only an interval scale ("by how much?"); "hotter" and "colder" is only an ordered scale.

2.2. *The desirable properties of a validation metric.*

Ferson et al. (2008, p. 2415) state that "a validation metric is a formal measure of the mismatch between predictions [of the model] and data that have not previously been used to develop the model." And that the closer the match between the model output and the historical observations, the smaller the measure. Specifically, they argue that a desirable measure should exhibit six properties:

1. it should be objective (and quantitative) so that the same predictions and the same data will result in the same assessment, no matter who conducts it.
2. if there is a comparison between deterministic values without stochasticity, then the metric should generalize this in a reasonable way.
3. the metric should reflect all differences in the two distributions (of the predictions and of the history), not just the lower moments of these distributions (mean, standard deviation); it should not be too sensitive to outliers.
4. for ease of understanding, the unit of the metric should be the same as the unit of the variables, if possible.
5. the modulus of the measure should be unbounded above.
6. the measure should be a true *metric*: that is, it should be non-negative and symmetric, should satisfy the delta inequality:

$$d(x, y) + d(y, z) \geq d(x, z)$$

and should satisfy the identity of indiscernables:[8]

$$d(x, y) = 0 \iff x = y.$$

Property 6 defines a metric. Properties 1, 3, and 6 are, I believe, crucial to any validation measure. Property 4 is desirable for interval-scaled variables and Property 2 is desirable for validation of stochastic models. Property 5 is not necessary, and is inapplicable where the variables are order- or nominal-scaled.

The issue of measuring the distance between the dynamics of the output produced by a simulation model and the historical counterpart raises the question of how to define a metric to measure this distance. For simple phenomena (and simple models) the output might be simple too. Measuring the distance between two time series, say, is simple. When the phenomenon is dynamic and multivariate, with more than one interrelated time-series output, however, the issue of defining and measuring the distance between the pair of sets of outputs is not simple.

If, moreover, the variables of the data and the model predictions are not interval-scaled, but only nominal-scaled, then the units of the measure will not in general be those of the data and predicted variables (Property 4). And it is not clear whether Properties

[8] Lacking only symmetry, it is a quasi-metric; lacking only the identity of indiscernables, it is a semi-metric; lacking only the triangle inequality, it is a pseudo-metric.

2 and 3 will be satisfied. First, in the application of oligopolistic pricing below, the historical data and the model predictions are deterministic. (The computer simulation is a deterministic model, mapping from market state, determined by curtailed historical data, to the next period's marketing actions — here, prices). It is not clear how to generalize this to a stochastic model, except perhaps by Monte Carlo simulations (Marks 2016). Second, the metrics we propose below are no more sensitive to (less frequent) data than they are to more frequent data: the tails are not too influential.

3. Four families of validation measures

We can distinguish between four families of measures that are important in the context of validation. First, empirical likelihood measures; second, what might be called stochastic area measures; third, information-theoretic measures; and, fourth, strategic state measures that compare patterns of data.

3.1. Empirical likelihood measures

These measures include maximum likelihood, the generalized method of moments, the method of simulated moments, and indirect inference (see Chen et al. 2012); to a greater or lesser extent, these demand knowledge of the true probabilistic dynamics of the models' output, or require the use of assumptions about these dynamics.[9] But likelihood measures rely on summary statistics and do not explicitly compare the similarity of distributions or patterns between the data and the simulated data generated by the models.

In general, such measures satisfy Ferson et al.'s Properties 1, 4, 5, and 6, but, in only generating summary statistics, these measures ignore the information contained in the patterns, especially relevant in strategic, dynamic models; they do not satisfy Property 3. Moreover, such methods are usually seen not as validation methods, but as methods of calibration and estimation (see Ch. 30 by Fagiolo et al. in this volume, Section 4.1).

3.2. Stochastic area measures

These have been derived by Ferson et al. (2008) and Roy & Oberkampf (2011) and others. Specifically, these papers address models and observations with stochastic characteristics, and univariate response quantities. That is, the model output Y and the observed data X are single random variables. Unfortunately, the generalization to multivariate responses is not straightforward.

Following Ferson et al. (2008), there are a variety of ways to compare univariate random variables, expressed as probability density functions (p.d.f.s) or cumulative distribution functions:

[9] Guerini & Moneta (2017) present a new method of validation, based on comparing structures of vector autoregressive models estimated from both model and historical data.

First, the random variables are "equal" or "surely equal" if their p.d.f.s are identical.

Second (more weakly), the random variables are "equal in mean" if the expectations of the absolute values of the differences between X and Y are zero.

Third (more weakly), if not quite equal in means, the *mean metric* provides a measure of their discrepancy

$$dE(X,Y) = E(|X - Y|) \neq |E(X) - E(Y)|,$$

where E is the expectation operator. This can be generalized to higher-order moments of the distributions, where equality in the higher-order moments implies equality in all lower-order moments.

Fourth (more weakly), if the shapes of the distributions of the two variables are identical, then the random variables are "equal in distribution".

Fifth (more weakly), if the distributions are not quite equal in shape, there are many proposed measures, including the Kolmogorov-Smirnov distance:

$$dS(X,Y) = \sup_z |\Pr(X \leq z) - \Pr(Y \leq z)|,$$

which is the vertical distance between the cumulative distributions functions of the two random variables, where z takes on all values in the common range of historical observations X and model output Y. Other measures, such as the Kullback-Leibler divergence, are discussed below.

The variables compared in Ferson et al. (2008) and Roy & Oberkampf (2011), like almost all variables in finance, scientific, and engineering validation, share one property: they are interval scales. The Area Validation Metric (AVM) introduced by Ferson et al. (2008) can only be applied when the two variables are interval or ratio scales. The AVM measures the area between the cumulative distribution functions of the two random variables, that of the model predictions and of the historical data. The metric is not defined for ordered scales, in which the variables might be increasing (decreasing) in one (or the other) direction but where distances in these directions are meaningless because order is their highest characteristic.

And such interval-scale measures cannot be applied when the variables are nominal only: when their order is arbitrary, and their highest characteristic is unique identity, with arbitrary, separate names or numbers.

Indeed, the Smirnov distance is not applicable, even with an arbitrary ranking of the ordering of the states. But the the applications introduced below generate nominal-scale output, not interval-scaled.

Ferson et al.'s AVM, in measuring the divergence of the p.d.f. of the model output from the historical data, does take satisfy Property 3, but is limited in that it requires interval-scaled single variables of both output and observed data.

In what follows, we focus on methods that explicitly compare patterns in the data, both observed and simulated, and do not in general require interval-scaled variables. They are from the following two families.

3.3. Pattern-based measures I: information-theoretic measures

Information-theoretic measures are derived from Shannon's measure of entropy (Shannon 1948), and include the Kullback-Leibler construct (Kullback-Leibler 1951), and more recent measures that attempt to overcome shortcomings of Kullback-Leibler, such as the *GSL-div* (Lamperti 2018a. 2018b).

Such measures satisfy Properties 1, 3, and 5, but, as we discuss in Section 5 below, they do not in general satisfy Property 6, although that has not eliminated their use in model validation. I argue here that there are true metrics which should be considered instead.

3.4. Pattern-based measures II: strategic state measures

Strategic state measures include Marks' State Similarity Measure (Marks 2013) and Klir's 2006 Generalized Hartley Measure, from early set-theoretic work of Hartley's (Hartley 1928). These two measures satisfy Ferson et al.'s Properties 1, 2, 3, and 6, but not Property 4 (units of measurement), or Property 5 (the measures are bounded above); I argue that these two properties are not crucial for a validation metric.

4. Measures of closeness or of information loss

Turn now to the third family of measures. The broad idea behind evaluating a distance between the model output and the real-world data in order to choose the model "closest" to the real-world data is as follows. If the real data are information full, then models of the underlying process capture only some of the information. Choosing the model that loses least information compared to historical data is the criterion for choosing the "best" model.

Information is often measured using Shannon entropy (1948) (SE).[10] It 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 the probability distribution of random variable x. The function SE exhibits 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; it has some attractions theoretically, but is not a true metric, as we shall see.

[10] Another measure used for information is Hartley information (see Section 7). Both are special cases of Rényi entropy (Rényi 1960). Both derive from work done at the Bell Labs.

4.1. Kullback-Leibler information loss

The Kullback-Leibler (K-L) divergence or information loss (also known as relative entropy) 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 Property 6, since $I(f,g) \neq I(g,f)$.[11] Moreover, π_i must be positive for every i,[12] while in data, even for a coarse, dichotomous partition, this value is likely to be zero for some states, for either set of data (model predictions or real data).[13] As mentioned above, this is a stumbling block for the AVM technique of Ferson et al. (2008), although AVM is suitable for validation of models with univariate random variables for output and observations.

4.2. The Generalized Subtracted L-divergence (GSL-div)

To overcome shortcomings of the Kullback-Leibler divergence, the symmetric L divergence (Lin 1991) was developed. From this the *GSL-div* (Lamperti 2018) has been derived to measure the degree of similarity between real and simulated dynamics by

[11] It is a semi-quasimetric.

[12] The K-L measure is defined only if $p_i = 0$ whenever $\pi_i = 0$.

[13] 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 (2002) 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."

comparing the patterns of the time series. Lamperti discusses the procedure to obtain the *GSL-div*, and then presents results to discriminate among four different classes of stochastic processes. He also compares the *GSL-div* with alternative measures of fit (using several summary statistics) commonly used for calibrating ABMs, and concludes that *GSL-div* provides much more satisfactory performance at this (see Ch. 30 by Fagiolo et al. in this volume, Table 1). But neither K-L nor Lin's *L-div* (and hence *GSL-div*) satisfy Property 6, and, hence, are not proper metrics, despite the interesting properties of *GSL-div* (Lamperti 2018).[14]

Let us now turn to the fourth family of measures, the strategic state measures, which include the author's State Similarity Measure (Section 6) (which uses rectilinear or Minkowski's L_1 or the cityblock distance), and Klir's Generalized Hartley Measure (Section 7). Both are true metrics. Before we present the measures, we describe the models for our example.

5. The example: models and data

Return to our example of the interactions over time among several brands. We use three models from simulations described in Marks et al. (1995). Each model has three interacting brands, and each brand agent independently chooses its weekly price from its own set of four possible prices in order to maximize its weekly profit, in a process of co-evolution using the Genetic Algorithm (GA). 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. See results for 2- and 3-week memory below. The GA chooses the mapping from perceived state to action for each brand (with each brand's weekly profit as its "evolutionary fitness"). This means that the models are not derived from historical patterns of oligopolistic behaviour, and so can be used to predict these patterns.

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.[15] Figures 1 and 3 of Midgley et al. (1997) and also of Marks (2013) show, respectively, the observed historical weekly prices and volumes sold of several brands of coffee competing in a

[14] Although, as Lamperti (2018) points out, so long as the simulated data are always compared with the historical data, and not with simulated data from other models, *GSL-div* might still allow model choice.

[15] 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 A has a period of 13 weeks, Model B of 6 weeks, and Model C 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.

Table 1
State frequencies from History and three models.

State	History	Model A	Model B	Model C
000	32	30	20	0
001	2	11	10	18
010	6	3	7	15
011	1	0	0	0
100	7	5	12	16
101	0	0	0	0
110	2	1	1	0
111	0	0	0	1
Total	50	50	50	50

U.S. supermarket chain, and a fifty-week period of simulated interactions among three brand agents in Model A, where each brand chooses from one of four possible prices per week.

In order to reduce the number of degrees of freedom, we coarsen the partitioning of the data, using a dichotomous partition into High and Low prices for both the real data and the simulated data.

The distribution of the eight possible 1-week states in the historical chain store (H) with three brands or players and in three models (A, B, and C) [16] of the models' outputs, using 50 weeks of data, are shown in Table 1, with "0" corresponding to a player's "High" price and "1" to a player's "Low" price.[17] Modelling deeper memory for the brands results in similar distributions, but the tables are 64 rows and 512 rows deep, with 2-week and 3-week memory, respectively, corresponding to 64 and 512 states.

The important thing to note here is that these are models of *strategic* interaction: it is not sufficient to examine a single brand's time series of actions, since these have affected — and in turn have been affected by — its rivals' actions over time. This is essentially a multivariate validation problem.

6. The State Similarity Measure (SSM)

Introduced in Marks (2010), 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. 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

[16] In Midgley et al. (1997) and Marks (2013), Model A is called Model 26a, Model B is called Model 26b, and Model C is called Model 11.

[17] Figures 2 and 3 of Marks (2013) plot these behaviours. State 000 corresponds to all three players choosing High prices; State 001 corresponds to Players 1 and 2 choosing High prices and Player 3 choosing a Low price, etc.

Table 2
SSMs calculated between the six pairs of sets.

	Pair	1-week memory	2-week memory	3-week memory
b	History, Model A	18	36	54
f	Model A, Model B	22	42	60
c	History, Model B	28	48	68
e	Model C, Model B	42	60	80
d	Model C, Model A	62	76	88
a	History, Model C	70	88	92

SSM between two sets matrix $\mathbf{P}$ and matrix $\mathbf{Q}$ of vectors (or time series) is calculated as the rectilinear Minkowski's L_1 or cityblock distance (Krause 1986) d_1 between their two constructed vectors $\mathbf{p}$ and $\mathbf{q}$, given by

$$d_1^{\mathbf{PQ}} = d_1(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i|, \tag{1}$$

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 Marks (2013), Appendix 1 for details of this procedure.)

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.

The 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. In our example, 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). It is possible to define a normalised measure.

6.1. Results for the models

The six pairs of SSMs between the partitioned prices of the three models and the observed historical data, using 50-week data series, are presented in Table 2 for 1-, 2-, and 3-week memory. Table 3 presents the distances between History, and the three simulations, Model C, Model A, and Model B from Marks et al. (1995), with 3-week memory. Model C is far from any of the other sets, and Model B is closest to Model A, but Model A is closer to the History historical data (at 54/96) than it is to the closest other simulation, Model B (at 60/96).

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

Table 3
SSMs between Observed History and Three Models

	History	Model A	Model B	Model C
History	0	54	68	92*
Model A	54	0	60	88*
Model B	68	60	0	80*
Model C	92*	88*	80*	0

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) (History, Model A), (Model A, Model B), (History, Model B), (Model C, Model B), (Model C, Model A), and (History, Model C). Which of the three models is closest to the historical data of History? The SSM tells us that Model A is best, followed by Model B, with Model C bringing up the rear.

6.2. Monte Carlo simulations of the SSM

We can, using Monte Carlo stochastic sampling (Marks, 2016), derive some statistics to test whether any pair of sets is likely to include random series (see below).

As *Null Hypothesis* we choose: 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. (Note: * in Table 3 indicates we cannot reject the null at the 5% level.) 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 of three random time series are at least 76 apart.[18] This means that, in Table 3, we reject the null hypothesis of random data for the pairs (History, Model A), (History, Model B), and (Model A, Model B), 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 C), 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). Figure 4 of Marks (2013) plots the the Cumulative Mass Function (CMF) of the MC parameter bootstrap simulation against the six SSMs of the pairs.

7. 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

[18] 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.

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 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:

$$c \log_b |E|,$$

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.[19]

7.1. 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[20] on the finite set X to take any value between zero and one: $r : X \to [0, 1]$. Note that

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

[19] See further discussion in Marks (2013), Appendix 2.

[20] It is not correct to call the function r a possibility *distribution* function, since it does not distribute any fixed value among the elements 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 i = 1, ... 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, the set $A_i = \{x_1, x_2, \ldots, x_i\}$ is defined for each $i \in \{1, \ldots, n\}$.

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

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

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]."

7.2. Applying U-uncertainty to our data

From the frequencies of Table 1 (one-week memory), we can reorder[21] the possibilities (observed frequencies) of the three runs and the historical data, to get the four reordered, non-normalised[22] possibility profiles:

Using equation (2), the four Hartley measures are calculated:[23]

[21] 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.

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

[23] 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:	32	7	6	2	2	1	0	0
Model A:	30	11	5	3	1	0	0	0
Model B:	20	12	10	7	1	0	0	0
Model C:	18	16	15	1	0	0	0	0

Table 5
GHMs calculated for three memory partitions.

Process	1-week memory	2-week memory	3-week memory
History	0.383	0.495	0.782
Model A	0.516	0.679	1.085
Model B	1.054	1.657	2.542
Model C	1.399	2.179	2.787

1. History:

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

$$= 0.383$$

2. Model A:

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

$$= 0.516$$

3. Model B:

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

$$= 1.054$$

4. Model C:

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

$$= 1.399$$

The GHMs for the three models and History 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 Property 6, 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 A (0.516) is closest to the historical data of History (0.383); next is Model B (0.516), with Model C (1.399) furthest from the Historical

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

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

data. Moreover, we can see that Model A is closer to the Historical data than it is to Model B.

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.[24]

8. 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 A is closest to History using either measure, and Model C 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 (see Table 6). 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.[25]

9. Conclusions

Is a particular computer model the best model of a particular real-world phenomenon? “Best” can have several meanings, but here we mean whether the behaviour (“output”) of the simulation model is closest to the observed behaviour of the phenomenon. Measuring the closeness of the simulated behaviour and the observed (historical) behaviour might be simple (for example, for univariate, interval-scaled, deterministic variables) or not (for example, for multivariate output of nominal-scaled variables).

[24] We could also define a normalised GHM.
[25] Exploration of these differences awaits further research.

Measuring this closeness is necessary to validate any model, and can be used to choose the best model of set of contenders.

We have examined the appropriateness of measures from four families of techniques, as characterised by the kinds of output observed and generated. Using Ferson et al.'s "desirable properties" of validation metrics, and focussing on the kind of phenomenon (oligopolistic, strategic interactions among sellers) which exhibits multivariate, nominal-scaled behaviour, we have argued that two contenders — SSM and GHM — are appropriate.

These two strategic measures, SSM and GHM, are true metrics that allow us to measure the degree of similarity between two sets of vectors (or matrices $\mathbf{X}$ and $\mathbf{Y}$), here multivariate 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.

Since GHM is a metric, differences of sets of vectors' GHMs are meaningful. SSM is also a metric (satisfying Property 6). 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 SSM and GHM strategic state measures 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 tally the differences in the states between two constructed vectors. It can be described in six simple steps, as outlined in Marks (2013), Appendix 1. The GHM is anything but intuitive, based on arcane possibility theory.

Using Occam's Razor, the SSM, as a simpler, more transparent measure, is preferred.

The two strategic state 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 pairs of 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. The efforts of the editors of this volume and anonymous referees were very constructive, and

have greatly improved this chapter's presentation.

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] Chen S.-H., Chang C.-L., and Du Y.-R. (2012), "Agent-based economic models and econometrics." *The Knowledge Engineering Review* 27(2): 187–219.

[4] 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.

[5] Fagiolo G., Guerini M., Lamperti F., Moneta A., and Roventini A. (2017). "Validation of agent-based models in economics and finance," LEM Papers Series 2017/23, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy.

[6] Ferson S., Oberkampf W.L., and Ginzburg, L. (2008). "Model validation and predictive capability for the thermal challenge problem," *Computer Methods in Applied Mechanics and Engineering* 197: 2408–2430.

[7] Gilbert N. and Troitzsch K.G. (2005). *Simulation for the Social Scientist,* Open University Press, 2nd ed.

[8] Guerini M. and Moneta A. (2017), "A method for agent-based models validation," *Journal of Economic Dynamics & Control* 82: 125–141.

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

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

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

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

[13] Lamperti F. (2018a), "An information theoretic criterion for empirical validation of simulation models," *Econometrics and Statistics* 5: 83–106.

[14] Lamperti F. (2018b), "Empirical validation of simulated models through the GSL-div: an illustrative application", *Journal of Economic Interaction and Coordination* 13: 143-171.

[15] Lin J. (1991), "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory* 37(1): 145–151.

[16] Liu Y., Chen W., Arendt P., and Huang H.-Z. (2010), "Towards a better understanding of model validation metrics," 13th AIAA/ISSMO Multidisciplinary Analysis Optimization Conference, Multidisciplinary Analysis Optimization Conferences.

[17] Mankin, J.B., O'Neill, R.V., Shugart, H.H., and Rust, B.W. (1977), "The importance of validation in ecosystem analysis," in G.S. Innis, ed. *New Directions in the Analysis of Ecological Systems, Part 1, Simulation Council Proceedings Series*, Simulation Councils, La Jolla, California, 5: 63–71. Reprinted in H.H. Shugart and R.V. O'Neill, eds. *Systems ecology,* Dowden, Hutchinson and Ross, Stroudsburg, Pennsylvania, 1979, pp. 309–317.

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

[19] Marks, R.E. (2007). "Validating simulation models: a general framework and four applied examples,"

Computational Economics, 30(3): 265–290, October. http://www.agsm.edu.au/bobm/papers/s1.pdf

[20] 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.

[21] Marks, R.E. (2013). "Validation and model selection: Three similarity measures compared." *Complexity Economics,* 2(1): 41–61. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.401.6982&rep=rep1&type=pdf

[22] Marks, R.E. (2016). "Monte Carlo," in *The Palgrave Encyclopedia of Strategic Management*, edited by D. Teece and M. Augier, London: Palgrave.

[23] 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.

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

[25] 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.

[26] Oberkampf, W.L. and Roy, C.J. (2010), "Chapter 12: Model accuracy assessment," in *Verification and Validation in Scientific Computing,* Cambridge University Press, Cambridge, pp. 469–554.

[27] 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.

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

[29] Roy C.J. and Oberkampf W.L. (2011). "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing," *Computer Methods in Applied Mechanics and Engineering.* 200: 2131–2144.

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