Tackling Model Selection and Validation: an Information Theoretic Criterion

Francesco Lamperti
Sant'Anna School of Advanced Studies
piazza Martiri della Libertà 33, 56127, Pisa, Italy
Email: f.lamperti@sssup.it

***Abstract*—Simulated economies suffer intrinsically from validation and comparison problems. The choice of a suitable indicator quantifying the distance between the model and the data is pivotal to model selection. However, how to validate and discriminate between models are still open problems calling for further investigation, especially in light of the increasing use of simulations in social sciences. In this paper I present a new information theoretic criterion to measure how close models' synthetic output replicates the properties of observable time series without the need to resort to any likelihood function or to impose stationarity requirements. This indicator is sufficiently general to be applied to any kind of model able to simulate or predict time series data, from simple univariate models such as Auto Regressive Moving Average (ARMA) and Markov processes to more complex objects including agent-based or dynamic stochastic general equilibrium models. More specifically, I use a simple function of the L-divergence computed at different block lengths in order to select the model that is better able to reproduce the distributions of time changes in the data. To evaluate the L-divergence, probabilities are estimated across frequencies including a correction for the systematic bias. Finally, using a known data generating process, I show how this indicator can be used to validate and discriminate between different univariate models providing a precise measure of the distance of each model from the data.**

I. Introduction

THE USE of simulations as a tool to investigate real phenomena has increased steadily in the last two decades, covering almost every field of the social sciences (see [1]). Acknowledging this trend, one fundamental issue has become to establish what a good simulation is. According to [2] the answer to this question must be: a good simulation is one that achieves its aim. But just what the aim or goal of a simulation might be is not obvious. Simulations might be used to explain a phenomenon, to predict its behaviour or to explore the internal structure of the phenomenon itself. Moreover, the aim of the simulation depends largely both on the modeller and the model. In [3] two kinds of models are recognized: *demonstration models*, essentially existence proofs for phenomena of interest, and *descriptive models*, that attempt to track dynamic historical phenomena. Most early simulation models are considered as demonstrative and a nice example could be the well known Schelling ([4])'s segregation model. Despite these models are extremely useful tools for explorative analysis and "as if" stories, policy analysis requires descriptive, validated models. The argument is simple: if you wanted to advise a policy maker on the basis of results from your model, you should be able, at least, to show that your model can replicate the behaviour of observed data. When this does not happen, it would be difficult for the policy maker to trust your advice or, at least, it should. Hence, for a descriptive model, offering good simulations means these simulations can be successfully validated against historical data.

How to validate a model is still an open issue for simulation studies[1] (see [9], [10] and more recently [11]). Finding appropriate tools to so is crucial both for the scientific debate and for policy analysis; the academia needs to develop theories whose implications fits with empirical evidence and policy makers needs information coming from reliable models. Establishing the fit of different models with empirical data is exactly what I am doing in this paper where I introduce, discuss and estimate a new information theoretic criterion.

Mason ([12]) distinguishes between *output validation* and *structural validation*. The latter asks how well the simulation model represents the (prior) conceptual model of the real-world system; the former asks how successfully the simulations' output exhibits the historical behaviours of the real-world target system. Output validation can be directly related to what Leombruni et al. ([13]) define as *empirical validity* of a model, i.e. validity of the empirically occurring true value relative to its indicator. Leombruni et. al introduce other four validity concepts that theory- and data-based simulation studies must consider: theory (the validity of the theory relative to the simuland), model (the validity of the model relative to the theory), program (the validity of the simulating program relative to the model), operational (the validity of the theoretical concept to its indicator or measurement). Any-time simulations exhibit lacks with respect to one or more of these validities, empirical validity is in turn affected and thereby reduced.
Following [14], it is useful to think of two parallel unfoldings: the evolution of the real economy (or market or whatever) and the evolution of the model of this real-world phenomenon. If the model is properly specified and calibrated, then its

[1] In what follows I refer to the Agent Based literature, where simulations are at the very core of the scientific inquiry process (see [5]); however, validation is crucial also for more standard approaches, especially in economics (see [6]), which would certainly benefit from reading and confronting with the literature I am referring to (see [7]). On the latter theme see also [8]

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

evolution should mirror the historical evolution of the real-world phenomenon: we could observe the evolution of the model or the real-world evolution and both should reveal similar behaviour of the variables of interest.

In this paper I focus on establishing whether and the extent to a simulation is able to reproduce and predict the behaviour of a phenomenon. This procedure is carried out by defining and computing an information theoretic criterion, based on a simple function of the L-divergence ([15]), which measures the distance between the actual, observed data and the synthetic series generated by different, competing models. This criterion, named *Generalized Subtracted L -divergence*, shortly *GSL-div*, allows to validate the output of a model by capturing its ability to reproduce the distributions of time-changes (that is, changes in the process' values from time to time) in the real-world observed process, without the need to resort to any likelihood function or to impose stationarity requirements. In turns, the procedure I am going to describe allows for direct model selection, identifying a precise measure which expresses empirical validity and selecting the model which exhibits the highest value with respect to this metric. The approach is in line with [3] and, as the State Similarity Measure (SSM) proposed therein, it tackles the fourth issue raised in [9]: validating agent based models using historical data. It is to be noticed that, in this paper, I use the *GSL-div* to compare univariate time series; however the approach can be extended to multivariate data structures.

II. Measuring the distance between time series: the GSL-div

As well explained in [3], using the glasses of information theory rather than statistics, the observed data contain information, and the (descriptive) models we develop (from our theoretical understanding of the underlying processes generating the observed data) can be seen of as attempts to reproduce the highest possible fraction of these information, in the most compact way. When several models referring to the same phenomenon are available, empirical validation should be able to point out the "best" model, that is the model whose output lose the least amount of information with respect to the real-world data.

The *GSL-div* measures the distance between the real and model's time series. The former is the unique realization of the unknown data generating process, the latter are taken to be M series generated by the same model, with the same (post-calibration) parameters' value. The use of an ensemble of replicated series provides a double advantage: it allows to correct for the systematic bias in the estimation of information theoretic quantities (see below) and it captures the behaviour of the model washing away the effects of particular realizations. The approach used to develop this criterion could be thought as the result of an extension of the work provided in [16].

Distance or divergence measures are widely used in a number of theoretical and applied statistical inference and data processing problems, including estimation, detection, compression and model selection ([17]). Most of them rely largely on the concept of Shannon's entropy ([18]), which expresses the amount of uncertainty associated with a random variable. Among these measures, one of the best known is the Kullback-Leibner divergenge (*KL-div*) between two distributions, $D(\mathbf{p}||\mathbf{q})$, or *relative entropy* ([19]). It is a measure of the inefficiency of assuming that the distribution is $\mathbf{q}$ when the true one is $\mathbf{p}$. The following discussion will be limited to discrete probability distributions, but results can be generalized to probability density functions.
Let X be a discrete random variable with support indicated by *X* and probability mass function $p(x)$, $x \in X$. If $q(x)$ is another probability mass function defined on the same support *X*, the *KL-div* is defined as

$$D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right), \tag{1}$$

where the logarithm is, usually, in base 2. Throughout the paper the following conventions will be used: $0 \log(0/0) = 0$ and, on the basis of continuity arguments, $0 \log(0/q) = 0$, independently of the logarithm's base. It is immediate to see that if there exist any symbol $x \in X$ such that $p(x) > 0$ and $q(x) = 0$ then, $D_{KL}(\mathbf{p}||\mathbf{q})$ is undefined. This means that distribution $\mathbf{p}$ has to be absolutely continuous with respect to $\mathbf{q}$ for the *KL-div* to be defined [20]. In addition, the $D_{KL}(\mathbf{p}||\mathbf{q})$ is non-negative, additive but not symmetric. In order to overcome these problems Lin defined a new symmetric measure, called L-divergenge, shortly *L-div*:

$$D_L(\mathbf{p}||\mathbf{q}) = D_{KL}(\mathbf{p}||\mathbf{m}) + D_{KL}(\mathbf{q}||\mathbf{m}), \tag{2}$$

where $\mathbf{m} = (\mathbf{p} + \mathbf{q})/2$ is the "mean" probability mass function. As the names suggest the $L-div$ is the basic building block I will use to construct the *GSL-div*.
It is immediate to see that $D_L(\mathbf{p}||\mathbf{q})$ vanishes only if $\mathbf{p} = \mathbf{q}$ and that the L-divergence is bounded above by 2. This is more evident when expressing the L-divergence in terms of the Shannon entropy, that is

$$D_L(\mathbf{p}||\mathbf{q}) = 2H \left(\frac{\mathbf{p} + \mathbf{q}}{2} \right) - H(\mathbf{p}) - H(\mathbf{q}), \tag{3}$$

i.e. the difference between twice the mean distribution and the sum of the entropies of $\mathbf{p}$ and $\mathbf{q}$. The generalization of the *L-div* is the Jensen-Shannon divergence (see [15]), defined as

$$Div_{JS}(\mathbf{p}, \mathbf{q}) = H(\pi_1 \mathbf{p} + \pi_2 \mathbf{q}) - \pi_1 H(\mathbf{p}) - \pi_2 H(\mathbf{q}), \tag{4}$$

where the weights π_1 and π_2 must satisfy $\pi_1, \pi_2 \geq 0$ and $\pi_1 + \pi_2 = 1$. It is straightforward that $D_L(\mathbf{p}||\mathbf{q}) = 2Div_{JS}(\mathbf{p}, \mathbf{q})$ for $\pi_1 = \pi_2 = 1/2$. It is to be noticed that the *KL-div*, and consequently the *L-div*, does not satisfy the triangle inequality, and hence cannot be considered a proper

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

metric.[2]
With reference to the use of these measures as quantities for model validation and selection Marks ([3]) outlines their inadequacy due to the previous problem. However, if models' data-distributions (say **q**) are always compared directly with the real data-distribution (say **p**), and not among themselves, model selection does not need a metric satisfying triangle inequality. Moreover, Endres et al. ([22]) found that the square root of the *L-div* is a metric and they called this new information metric the Jensen-Shannon distance.

In this paper I use the *L-div* as a measure that captures the distance between the distributions of time-changes in the real-world process and those generated by the synthetic output of simulated, competing models. Time-windows of different lengths are taken into consideration for the generation of the state space, which is represented by the set of values the series might take at each instant of time. The *L-div* is estimated for each length of the time-window and results are finally aggregated into a single information criterion, the Generalized Subtracted L-divergence (shortly *GSL-div*). It is worth noticing that this new measure is designed to capture similarities in the behaviour of the time series, and not in their levels. This reflects the opinion that is not relevant for a simulation to mirror the same values of the real data but to display the same behaviour in terms of trends, variabilities, trajectories and their shape. These elements are captured by the *GSL-div*. Furthermore, given two series sharing the same behaviour but different levels, it is sufficient to change initial conditions to notice they are in effect identical, and this would amount to add or subtract a drift to one of the two. Finally, levels depend largely on the unit of measure used by different models, while series' behaviour does not.

A. *The* GSL-divergence

Consider a random variable x taking values from the set $\mathbf{x} = (x_1, ..., x_k)$ with probabilities $\mathbf{p} = (p_1, ..., p_k)$. Assume we observe real-world or simulated time series both of length N; from $x(t)$, $t = 1, ..., N$ it is possible to build an histogram $\mathbf{n} = (n_1, ..., n_k)$, where n_i is the number of times the outcome was x_i. The frequency vector $\mathbf{f} = (f_1, ..., f_k) = (n_1/N, ..., n_k/N)$ is an estimator of the probability distribution $\mathbf{p}$.
Within this framework it is important to notice that I consider a discretization of the state space: time series are assumed to take only a finite set of values. How to conduct this procedure is crucial. In particular, for each time series $\{x(t)\}_{t=1}^{N}$, I take the original, real interval $[x_{min}; x_{max}]$ and I partition it in $b \in \mathbf{N}_0$ subintervals, each of equal length. These intervals are numbered increasingly from 1 to b, with 1 assigned to $[x_{min}; x_{min} + \frac{(x_{max} - x_{min})}{b})$. The time series is then symbolized straightforwardly: each observation is mapped into the number assigned to the interval it falls within. The parameter b controls for the precision of the symbolization: for $b = 1$ the symbolized series takes one and only one value (namely 1) while for $b \to \infty$ we are back to the (scaled) real-valued process. The symbolization is simple and works as follows: each $\{x(t)\}_{t=1}^{N}$ is mapped into the natural number corresponding to the partition interval where it falls.
For example, consider the following realization of the stochastic process $x(t)$ with $t = 3$: $\{0; 0.65; 1\}$. Choosing $b = 2$, the symbolized series will be $x^s(t) = \{1, 2, 2\}$, while choosing $b = 10$ the symbolized series becomes $x^s(t) = \{1, 7, 10\}$. It is immediate to see that increasing b the information loss about the behaviour of the stochastic process due to the symbolization becomes smaller and smaller. However, as it typically happens, increasing the precision of the symbolization has a cost: higher b translates also in higher size of the alphabet, that is the total number of words that could be created using symbols $\{1, ..., b\}$. The size of the alphabet corresponds to the cardinality of the state space and increasing it might require larger time series to conveniently estimate probability distributions. However, as will be shown, the *GSL-div* does not suffer from the use of low values of b. Additionally, it is important to notice that using high precision of the symbolization procedure is not a problem when a large amount of data are available, for example in high frequency models of financial markets (see, among others, these recent contributions [23] and [24][3]). A detailed discussion about the partitioning of the state space when dealing with information theoretic functional is provided in [25].

Once the time series are properly symbolized, they are subdivided in successive blocks of equal length l; this operation is recursive for $l = 1, ..., L$, where L is the maximum block's length (time-window) considered. Since there are N observations for each series, N/l blocks will be obtained for each value of l. L represents the maximum length of the windows which are used to compare the behaviour of the real data with the synthetic ones. It has to be chosen considering both (i) the nature of the phenomenon of interest and (ii) the size of the available real-world time series which can be used to validate the models. The first criterion, (i), reflects the time-horizon one considers when analysing a given phenomenon. For example, if the focus is centred on business cycles, data will be typically quarterly and the time-window around eight or twelve periods; conversely, in case one considers economic growth in the long run, data will be annual and the window considerably enlarged. The second criterion, (ii), puts a constraint on the comparability of real v.s. simulated data: when a real-world time series of length N is the only available source of information about the phenomenon under study, it makes a non-sense to compare it with a double-length simulated series. On the other hand it could be perfectly reasonable to take an ensemble of replicated series each of length N, both to wash away across-simulation variability and to solve the small sample problem ([26]). Using symbolized

[2] A metric is a distance function which must satisfy non-negativity, symmetry, coincidence and triangle inequality (see [21]).

[3] Here the methodology described in the paper is directly applicable to the time series of stock prices.

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

series, $l = 1, ..., L$ represents also the length of the words which compose the corresponding alphabet.

For each value of l, a subtracted version of the *L-div* is estimated from the data. It provides a measure of how close the behaviour the synthetic data replicates the real one when the series are studied along windows of length l. The *GSL-div* aggregates subtracted *L-div* values using weights increasing in l. In such a way it integrates the distances between the distributions of two time series for multiple (namely L) time-windows. Greater weights are assigned to values of the *L-div* considering longer windows.
Let $\{x(t)\}_{t=1}^{N}$ and $\{y(t)\}_{t=1}^{N}$ be two time series of total length N and indicate with $x^s(t)$ and $y^s(t)$ their symbolized version according to the procedure described above. It is important the precision level b chosen in the symbolization procedure to be the same for the two time series to be comparable.
The *GSL-div* takes the following functional form:

$$D_{GSL}(x(t)||y(t)) = \sum_{i=1}^{L} w_i [-2 \sum_{x \in X_i} m_i(x) \log_{a_i} m_i(x) + \sum_{x \in X_i} q_i(x) \log_{a_i} q_i(x)] \quad (5)$$

$$= \sum_{i=1}^{L} w_i \left(2H^{X_i}(m_i) - H^{X_i}(q_i)\right) \quad (6)$$

where the symbol $H^{X_i}(\cdot)$ indicates the Shannon entropy of a distribution over the state space X_i.
On the right hand side of the first line of (5) the big square brackets contain the subtracted L-divergence computed at different block lengths l. In particular I take the *L-div* ([15]) and I subtract the entropy for the time series $\{x(t)\}_{i=1}^{N}$. This can be justified in two ways. On the one hand it is due to the fact that $x(t)$ is always taken to be the real-world time series and it can be observed only once. This means it is not possible to replicate this series and create an ensemble, as it will be done for the time series produced by models. As a consequence, it cannot be corrected for the systematic bias stemming from the fact that its entropy is computed using an estimator (the frequency over the state space) and not the true probabilistic structure (see [27] and [26]). On the other hand, being the *GSL-div* always applied to real data against models' output, when one compares the distance of different simulated data with respect to the real counterpart the entropy of the latter will always be washed away.

The logarithm is always in base a_i with $i = 1, ..., L$, which corresponds to the cardinality of the alphabet available at length $l = i$.
Consider for example the following symbolized time series of length 8 obtained selecting $b = 4$:

$$\{x(t)\}_{t=1}^{8} = \{0, 2, 1, 2, 1, 3, 4, 0\}.$$

When $l = 1$ the time-window corresponds to one period, the series is sub-divided into 8 blocks and each of them is associated to one out of four symbols, namely $1, 2, 3$ or 4. When $l = 2$, $N/l = 8/2 = 4$ blocks are obtained and each is mapped to one of the following 2^4 symbols: $\{(11); (12); (13); ...; (43); (44)\}$. The mapping between blocks and corresponding symbols is straightforward: the series' first block of length 2 is $\{0, 2\}$ and it is associated to the symbol (02); the second block, $\{1, 2\}$, is associated to (12) and so on. The cardinality of the alphabet available when the selected time-window has length l corresponds to the number of different symbols the series' blocks might be associated to. Hence

$$a_i = 2^{X_i} = b^l, \ \forall l = 1, ..., L \quad (7)$$

where b is, as usual, the precision level used in the symbolization.
It is worth recalling that, in equation (5), $m_i(x)$ indicates the "mean"distribution of $p_i(x)$ and $q_i(x)$:

$$m_i(x) = \frac{p_i(x) + q_i(x)}{2}, \quad (8)$$

where $p_i(x)$ is the estimated probability (frequency) assigned by the real-world process to the symbol x, while q_i is the counterpart assigned by (one series of) the simulated process to the same symbol.

As introduced above, each of the subtracted L-divergences entering the *GSL-div* is assigned an increasing weight. This reflects the grater importance assigned to the ability of the simulated data to match the behaviour of the real process over a longer time-window and, additionally, it compensates for the increasing value of the logarithm basis a_i. In particular, weights are chosen to guarantee that their first differences are constant; that is, the weight assigned at a given length of the time window is equal to the one assigned at the previous length plus a constant term. As usual, the normalization condition must hold, $\sum_{i=1}^{L} w_i = 1$. The following weights are obtained[4]:

$$w_i = w_{i-1} + \frac{2}{L(L+1)} \qquad i = 1, ..., L \quad (9)$$

where $w_0 = 0$. As will be shown, the choice of the weights is robust to changes and even assuming equal weights across the length of the time-windows results are unaffected.

B. The systematic bias

When an information theoretic function is computed without knowing the exact probability of each symbol, a systematic error might arise. In particular this the case when the true probabilistic structure of a process has to be estimated from a finite sequence of observations (see [27], [26], [28], [29], [30]).
Even knowing the true distribution $\mathbf{p}$ of a time series $x(t)$ over a state space *X*, when one computes any of the *KL-div*, *JS-div*,

[4]proof omitted here both for sake of brevity and its simplicity.

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

L-div or *GSL-div* between **p** and **q** estimated from $\{x(t)\}_{t=1}^{N}$ with $N < \infty$, the result would be larger than zero. Obviously, the bias is also present when computing the distance between two frequency vectors that are estimated from two realizations of the same stochastic process.

The concept of systematic bias for the numerical values of information theoretic functional is well known in the literature and it follows directly from Jensen inequality (see [31]). In particular, the bias is identified with the expectation value $E[f(\mathbf{f})]$ being lower than $f(\mathbf{p})$ where **f** is an estimator of the true probability distribution **p**. Applying this result to the Shannon entropy one obtains

$$E[H(\mathbf{f})] \leq H(\mathbf{p}), \tag{10}$$

where the expectation is defined over the ensemble of finite-length i.i.d sequences generated by the probability distribution **p**.
Following [32] it can be shown that the expected value of the observed entropy is systematically biased downwards from the true entropy:

$$E[H(\mathbf{f})] = H(\mathbf{p}) - \frac{B-1}{2N} + O(N^{-2}), \tag{11}$$

where N is the length of each time series and B is the number of states $x \in X$ such that $f(x) > 0$. This result was originally obtained by Basharin ([27]) and Herzel ([26]) who found also that, up to the first order $O(N^{-1})$, the bias is independent of the actual distribution **p**. The term $O(N^{-2})$ contains unknown probabilities **p** and cannot be estimated in general (see [28], [29], [32]).

Dealing with a model it is always possible to generate an ensemble of time series; conversely, it becomes impossible with the unknown real data generating process, which produces an unique observable series for each phenomenon. This help justify the fact I subtract the entropy of the real series when I define the *GSL-div*.
Applying the previous correction of the systematic bias to the *GSL-div* one obtain the following expression

$$\begin{aligned} D_{GSL}(x(t)||y(t)) = &\sum_{i=1}^{L} w_i[-2\sum_{x\in X_i} m_i(x)\log_{a_i} m_i(x) \\ &+ \sum_{x\in X_i} q_i(x)\log_{a_i} q_i(x)] \\ &+ \sum_{i=1}^{L} w_i \left(\frac{B_i^m - 1}{N_i} - \frac{B_i^q - 1}{2N_i}\right) \end{aligned} \tag{12}$$

where the second line captures the correction terms for the systematic bias.

Finally it is important to recall that the *GSL-div* is bounded both from above and below. In particular it is possible to show that

$$0 \leq D_{GSL} \leq 2. \tag{13}$$

However, due to the subtraction with respect to the *L-div*, this is not the case in practice, and the lower bound for the *GSL-div* is the unknown entropy of the real-world time series, which is, apart from special cases, positive. However, this is not a problem for model selection and validation and the only thing which matters is to have an upper bound for the criterion, which can be used as a comparison term. To the purposes of model selection lower the *GSL-div* the better the ability of the model to reproduce the behaviour of the observed real data.

III. A SIMPLE EXAMPLE

In this section I show the performance and the precision of the *GSL-div* criterion in distinguishing between three ad hoc created time series. $x(t)$ is chosen to be the observed series while $x_A(t)$ and $x_B(t)$ are to be intended as the output of two models (A and B respectively) trying to simulate $x(t)$. These series are consciously chosen to have $x_A(t)$ much more close to the behaviour of $x(t)$ with respect to $x_B(t)$. Their plot is reported in figure 1.

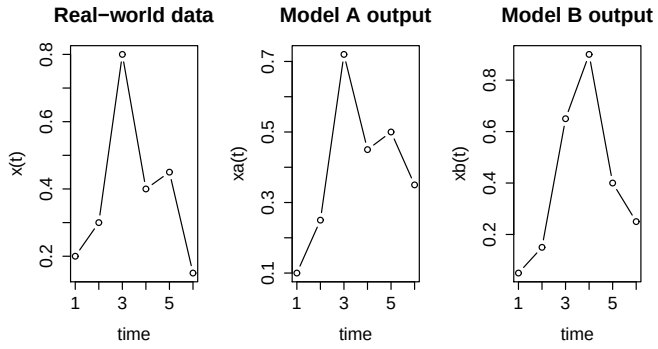

Fig. 1. Behaviour of three selected time series

I expect the *GSL-div* criterion to show a lower distance between the observed time series coming from the unknown data generating process and model A's output.
Before showing the results I present the symbolization process. The three series take values in the real interval $[0, 1]$ and a very small sample consisting of six observations is chosen:

$$x(t) = \{0.2; 0.3; 0.8; 0.4; 0.45; 0.15\},$$
$$x_A(t) = \{0.1; 0.25; 0.72; 0.45; 0.5; 0.35\},$$
$$x_B(t) = \{0.05; 0.15; 0.65; 0.9; 0.4; 0, 25\}.$$

The precision of the symbolization is set to $b = 3$; this choice leads to the following partition of the original state space: $[0; 0.33); [0.33; 0.66); [0.66; 1]$. Despite the choice of b is arbitrary results are robust to changes in the value of this parameter. The use of a low b can be justified here by the fact that the time series are very short; in addition, representing it the the precision of the symbolization process, the use of a

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

low value for b makes it more difficult to distinguish between the series. The ability of the *GSL-div* to recognize the most similar even when the symbolization is relatively imprecise would confirm the power of this new criterion.

According to the chosen parametrization, the three symbolized time series are:

$$x(t) = \{1, 1, 3, 2, 2, 1\},$$

$$x_A(t) = \{1, 1, 3, 2, 2, 2\},$$

$$x_B(t) = \{1, 1, 2, 3, 2, 1\}.$$

By inspection it is possible to notice that x_A is much more closer to x than x_B: while the former exhibits the same behaviour of the real data apart form the very last period, the latter displays twice the opposite one (it increases from $t = 3$ to $t = 4$ when $x(t)$ is decreasing and vice-versa in the following period).

Given the use of short time series, the maximum value of the time-window's length along which the three processes are compared cannot be set above $L = 3$; otherwise one and only one block would be available, the probability distribution over the alphabet would appear constant and its entropy pushed to zero. Hence, I am considering blocks and corresponding alphabets for $l = 1, 2, 3$. Respectively, six, three and two observations are obtained and used to estimate the frequencies. As it is obvious, these are very rough estimates of the probabilities the three process assign to symbols $x \in X_l$.
Notwithstanding this limitation the performance of the *GSL-div* in selecting model A and validating its output against real data is excellent. Table 1 (with progressive weights) and Table 2 (with uniform weights) provide evidence of this result.

TABLE I
GSL-DIV FOR $x(t)$ AND BOTH $x_A(t)$ AND $x_B(t)$ WITH PROGRESSIVE WEIGHTS

			Subtracted L-div	
	block length	weights	model A	model B
	1	0.17	0.948161	0.920620
	2	0.33	0.710310	0.710310
	3	0.50	0.420620	0.630930
GSL-div			**0.605900**	**0.706373**

TABLE II
GSL-DIV FOR $x(t)$ AND BOTH $x_A(t)$ AND $x_B(t)$ WITH UNIFORM WEIGHTS

			Subtracted L-div	
	block length	weights	model A	model B
	1	0.33	0.948161	0.920620
	2	0.33	0.710310	0.710310
	3	0.33	0.420620	0.630930
GSL-div			**0.693030**	**0.753953**

Two observations deserve attention. First, the subtracted L-divergence at blocks' length equal to one is lower for model B's than for model A's series. This is driven by the fact that $x_B(t)$ and $x(t)$ have been chosen to exhibit the same frequency distribution over the alphabet available for $l = 1$, $X_1 = \{1, 2, 3\}$, while $x_A(t)$ has not. This means that it becomes relatively more difficult to recognise $x_A(t)$ as the series most similar to $x(t)$. However, the distribution of time-changes is completely different between $x(t)$ and $x_B(t)$. The result is that when one move to $l = 2, 3$, corresponding to capture longer trends and trajectories, $x_A(t)$ equals and overcome $x_B(t)$'s performance in simulating the behaviour of $x(t)$. In addition, this justifies the choice of using progressive weights in the definition of the *GSL-div*: a model matching the distribution of changes for a longer time window should always be preferred and selected.
Secondly, the three time series have been selected ad hoc to show the performance of the *GSL-div*. Not having a proper model it is not possible to replicate simulations and correct for the systematic bias[5].
In the next section I move away from this example and I show the precision of the *GSL-div* in validating and selecting the most appropriate among 9 univariate stochastic models; the correction term for the systematic bias is added to the estimation of the criterion.

IV. SELECTING AND VALIDATING ARMA MODELS

A set of 9 Auto Regressive Moving Average (ARMA) models is analysed. The *GSL-div* is used to select the model which minimizes the distance with respect to the distribution of time changes in the real data. Real data are assumed to be a realization of a Gaussian $AR(1)$ process with autoregressive order-one parameter $\phi_1 = 0.1$. Figure 2 provides a plot of this process. It is obviously stationary, causal and invertible[6].

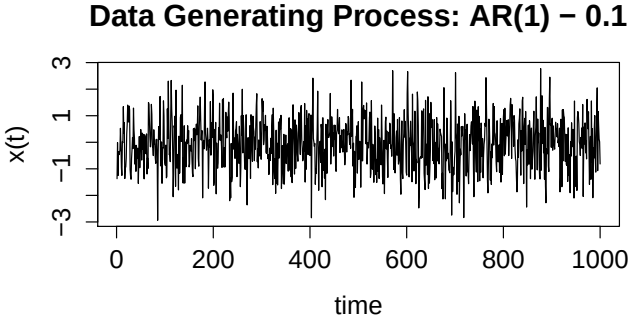

Fig. 2. The real-world time series

Table 3 summarizes the main features of the models which are considered for replicating the behaviour of the real data.

[5]The only meaningless solution would be assuming deterministic models producing always the same realization.
[6]see [33] for a definition of these properties

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

All of them are Gaussian $N(0,1)$ ARMAs and are used to produce an ensemble of $M = 1000$ Monte Carlo replications, each of length $N = 1000$. These series are symbolized using precision $b = 5$.

TABLE III
MAIN FEATURES OF THE NINE MODELS CONSIDERED

		parameters		properties	
	model	ϕ	θ	stationary	invertible
1	AR(1)	0.1	0	yes	yes
2	AR(1)	0.2	0	yes	yes
3	AR(1)	0.5	0	yes	yes
4	AR(1)	0.01	0	yes	yes
5	AR(1)	0.9	0	yes	yes
6	ARMA(1,1)	0.2	0.9	yes	yes
7	ARMA(1,1)	0.5	2	yes	no
8	AR(1)	1	0	no	yes
9	AR(1)	2	0	no	yes

The majority of the models considered are stationary and, even if not reported directly, they are also causal. In addition, most of them are invertible. This allows to conclude that six out of nine are unique, meaning that there is a one-to-one correspondence between the family of the finite dimensional distributions of the process and its finite parametric representation (see [33]). This applies also to the Data Generating Process (*DGP*).

The *GSL-div* is expected to recognize the model which is most similar to the *DGP*: model 1 exhibits exactly the same parametric representation of the data generating process from which $x(t)$ is taken. In addition one should ask the *GSL-div* to identify models producing series completely inconsistent with the real world data $x(t)$: model 9 is strongly non-stationary and exhibits an explosive behaviour. Therefore, within the class of models considered here, I expect the *GSL-div* to reach a minimum when model 1 and $x(t)$ are evaluated and a maximum when model 9 is compared to observed data.
Figures 3 and 4 provide a plot of a realization for model 1 and 9 respectively.

Tables 4 and 5 (in the next page) show the performance of the *GSL-div* in evaluating the distance between the distributions of the real-world time-series and different models, after correcting for the systematic bias. The maximum length of the time-window (or block-length) is chosen to be six. Numbers in bold indicate the estimated *GSL-div* while those in plain represent its partial values, that is subtracted L-divergences.

Expectations are perfectly confirmed: model 1 turns out be the closest to the real data while model 9 the most distant.

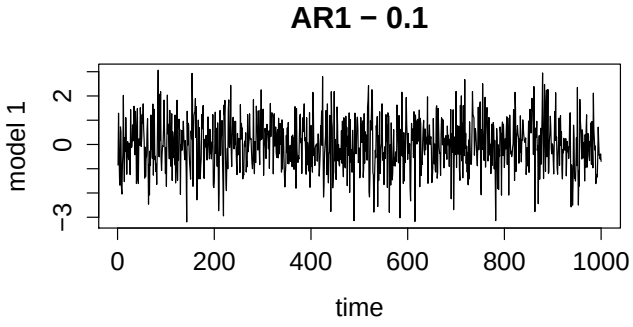

Fig. 3. A realization of model 1

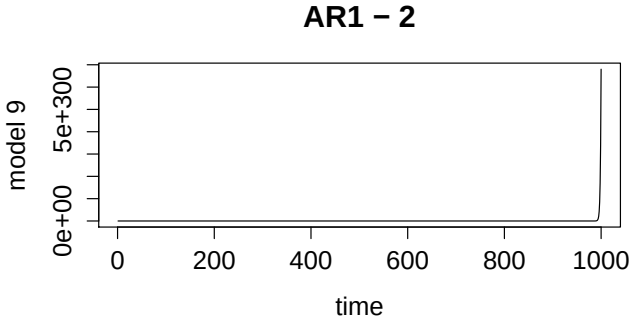

Fig. 4. A realization of model 9

In general the *GSL-div* is shown to distinguish clearly among models: non-stationary processes are the most distant from the real data and when a Moving Average component is added to the process the distance from the real data increases. This is true especially when the MA part is non-invertible (model 7). Moreover, among the same class of processes (AR(1)) the criterion is able to recognize those having a parametric representation which is closer to the *DGP*.
It is worth noticing that results are robust to the choice of the weights in the functional representation (12) of the *GSL-div*. Finally, the correction term for the systematic bias is, in absolute value, considerably low with respect to the estimated value of the *GSL-div* criterion, and it becomes even smaller the longer the time series. In particular, the correction never affects results and the ordering of models' distance from $x(t)$.

V. CONCLUSIONS

Validation of simulated models is still an open issue. One way of tackling this problem is via the identification of a measure of the distance between simulated and real-world data. This paper provides an information theoretic criterion, the *GSL-div*, which captures this distance without any requirement of stationarity nor the need to resort to any likelihood function. This constitutes a direct advantage with respect to other approaches aimed at characterizing times series and their behaviour.

TABLE IV
GSL-div FOR $x(t)$ AND NINE ARMA MODELS WITH PROGRESSIVE WEIGHTS

block length	*weights*	*AR1 - 0.1*	*AR1 - 0.2*	*AR1 - 0.01*
1	0.047619	0.035086	0.035110	0.035135
2	0.095238	0.069981	0.070112	0.070133
3	0.142857	0.104996	0.105102	0.105244
4	0.190476	0.140687	0.140874	0.140918
5	0.238095	0.173118	0.172980	0.173164
6	0.285714	0.194703	0.195057	0.195459
GSL-div		**0.718571**	**0.719235**	**0.720053**

block length	*weights*	*AR1 - 0.5*	*AR1 - 0.2 MA1 - 0.9*	*AR1 - 0.5 MA1 - 2*
1	0.047619	0.035154	0.035147	0.035203
2	0.095238	0.071077	0.071422	0.073134
3	0.142857	0.106422	0.106308	0.109892
4	0.190476	0.142626	0.144963	0.146516
5	0.238095	0.173115	0.174841	0.175495
6	0.285714	0.194625	0.195291	0.195260
GSL-div		**0.723019**	**0.727973**	**0.735501**

block length	*weights*	*AR1 - 0.9*	*AR1 - 1*	*AR1 - 2*
1	0.047619	0.035500	0.040350	0.072301
2	0.095238	0.077123	0.087890	0.109458
3	0.142857	0.114550	0.126863	0.143475
4	0.190476	0.151631	0.163332	0.174783
5	0.238095	0.178279	0.188724	0.196458
6	0.285714	0.196556	0.202831	0.207266
GSL-div		**0.753639**	**0.809990**	**0.903741**

TABLE V
GSL-div FOR $x(t)$ AND NINE ARMA MODELS WITH UNIFORM WEIGHTS

block length	*weights*	*AR1 - 0.1*	*AR1 - 0.2*	*AR1 - 0.01*
1	0.17	0.122801	0.122884	0.122972
2	0.17	0.122467	0.122696	0.122733
3	0.17	0.122495	0.122619	0.122784
4	0.17	0.123101	0.123265	0.123303
5	0.17	0.121183	0.121086	0.121215
6	0.17	0.113577	0.113784	0.114018
GSL-div		**0.725624**	**0.726333**	**0.727025**

block length	*weights*	*AR1 - 0.5*	*AR1 - 0.2 MA1 - 0.9*	*AR1 - 0.5 MA1 - 2*
1	0.17	0.123037	0.123015	0.123211
2	0.17	0.124385	0.124989	0.127985
3	0.17	0.124159	0.124026	0.128208
4	0.17	0.124798	0.126843	0.128202
5	0.17	0.121181	0.122389	0.122846
6	0.17	0.113531	0.113920	0.113902
GSL-div		**0.731091**	**0.735181**	**0.744353**

block length	*weights*	*AR1 - 0.9*	*AR1 - 1*	*AR1 - 2*
1	0.17	0.124250	0.141225	0.253055
2	0.17	0.134965	0.153807	0.191552
3	0.17	0.133642	0.148007	0.167387
4	0.17	0.132677	0.142915	0.152935
5	0.17	0.124795	0.132107	0.137521
6	0.17	0.114658	0.118318	0.120905
GSL-div		**0.764987**	**0.836380**	**1.023355**

My approach leaves two free parameters: the precision of the symbolization process, namely b, and the maximum length of the time-window used to identify blocks of the time series, namely l. Both can be increased when the size of real time series against which models are evaluated is large; however, I showed that using relatively low parameters' values ($b = 5$, $l = 6$) the *GSL-div* is extremely precise in selecting and ordering the models which are better able to reproduce the distributions of time-changes observed in the real data.

In this paper the *GSL-div* is applied to univariate models. Extensions to multivariate settings are possible. There, I explicitly account for the fact that a multivariate model which perfectly matches one real time series but poorly replicates the others should not, in general, be better than one which decently simulates the behaviour of all the considered series.

REFERENCES

[1] G. Ballot G., Weisbuch, *Applications of Simulations to Social Sciences*. Oxford: Hermes Science Publishing, 2000.

[2] R. Marks, "Validating simulation models: A general framework and four applied examples," *Computational Economics*, vol. 30, no. 3, pp. 265–290, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10614-007-9101-7

[3] ——, "Validation and model selection: Three similarity measures compared," *Complexity Economics*, vol. 2, no. 1, 2013.

[4] T. C. Schelling, "Dynamic models of segregation," *The Journal of Mathematical Sociology*, vol. 1, no. 2, pp. 143–186, 1971. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/0022250X.1971.9989794

[5] R. Axelrod, *Advancing the art of simulation in the social sciences,* in Handbook of Research on Nature Inspired Computing for Economy and Management, jean-philippe rennard, hersey ed. PA: Idea Group, 2006.

[6] A. Paccagnini, "Model validation in the dsge approach: A survey," 2009, working paper, mimeo.

[7] N. Poudyal and A. Spanos, "Confronting theory with data: Model validation and dsge modeling," 2013, working Paper, mimeo.

[8] G. Fagiolo and A. Roventini, "Macroeconomic Policy in DSGE and Agent-Based Models," *Revue de lâĂŹOFCE*, vol. 124, pp. 67–116, 2012.

[9] P. Windrum, G. Fagiolo, and A. Moneta, "Empirical validation of agent-based models: Alternatives and prospects," *Journal of Artificial Societies and Social Simulation*, vol. 10, no. 2, p. 8, 2007. [Online]. Available: http://jasss.soc.surrey.ac.uk/10/2/8.html

[10] G. Fagiolo, C. Birchenhall, and P. Windrum, "Empirical validation in agent-based models: Introduction to the special issue," *Computational Economics*, vol. 30, no. 3, pp. 189–194, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10614-007-9109-z

[11] C. Macan, "On validating multi-agent system applications," in *14th International Workshop on Multi Agent Based Simulation*, 2013.

[12] M. S.M., *Validation and verification of multi-agent systems, in Complexity and Ecosystem Management*, edited by m.a. janssen ed. Cheltenham: Edward Elgar, 2002.

[13] M. Richiardi, R. Leombruni, N. J. Saam, and M. Sonnessa, "A common protocol for agent-based social simulation," *Journal of Artificial Societies and Social Simulation*, vol. 9, no. 1, p. 15, 2006. [Online]. Available: http://jasss.soc.surrey.ac.uk/9/1/15.html

[14] R. R., *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. Oxford: Pergamon, 1985.

[15] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, pp. 145–151, 1991.

[16] G. Curato and F. Lillo, "Multiscale model selection for high-frequency financial data of a large tick stock by means of the jensenâĂŞshannon metric," *Entropy*, vol. 16, no. 1, pp. 567–581, 2014. [Online]. Available: http://www.mdpi.com/1099-4300/16/1/567

[17] M. Basseville, "Review: Divergence measures for statistical data processing-an annotated bibliography," *Signal Process.*, vol. 93, no. 4,

pp. 621–633, Apr. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2012.09.003

[18] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, 1948.

[19] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 49–86, 1951.

[20] S. Kullback, *Information Theory and Statistics*. New York: Dover Publications, 1968.

[21] W. Rudin, *Real and complex analysis*, ser. Mathematics series. McGraw-Hill, 1987. [Online]. Available: http://books.google.it/books?id=Z_fuAAAAMAAJ

[22] D. Endres and J. Schindelin, "A new metric for probability distributions," *Information Theory, IEEE Transactions on*, vol. 49, no. 7, pp. 1858–1860, July 2003.

[23] G. Curato and F. Lillo, "Modeling the coupled return-spread high frequency dynamics of large tick assets," *ArXiv e-prints*, Oct. 2013.

[24] G. Fagiolo, J. Leal, M. S. Napoletano, and A. Roventini, "Rock around the Clock: An Agent-Based Model of Low- and High-Frequency Trading," Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy, LEM Papers Series 2014/03, 2014. [Online]. Available: http://ideas.repec.org/p/ssa/lemwps/2014-03.html

[25] S. Panzeri and A. Treves, "Analytical estimates of limited sampling biases in different information measures," *Network: comput. in Neur. Sys.*, vol. 7, pp. 87–107, 1996.

[26] H. Herzel, A. Schmitt, and W. Ebeling, "Finite sample effects in sequence analysis," *Chaos, Solitons & Fractals*, vol. 4, no. 1, pp. 97 – 113, 1994, chaos and Order in Symbolic Sequences. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0960077994900205

[27] G. Basharin, "On a statistical estimate for the entropy of a sequence of independent random variables," *Theory Prob. App.*, vol. 4, pp. 333–338, 1959.

[28] P. Grassberger, "Finite sample corrections to entropy and dimension estimates," *Phys. Lett. A*, vol. 128, pp. 369–373, 1988.

[29] P. Grassberger and T. Schurmann, "Entropy estimation of symbol sequencences," *Chaos*, vol. 6, 1996.

[30] I. Samengo, "Estimating probabilities from experimental frequencies," *ArXiv: cond. math. stat. mech.*, 2002.

[31] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons: Hoboken, New Jersey, 2006.

[32] M. Roulston, "Estimating errors on measured entropy and mutual information," *Physica D*, vol. 125, pp. 285–294, 1999.

[33] J. Hamilton, *Time Series Analysis*. Princeton University Press: Princeton, New Jersey, 1994.