Department of Economics and Statistics
"COGNETTI DE MARTIIS"

Working Paper Series

52/13

ANALYSIS OF AGENT-BASED MODELS

JAKOB GRAZZINI, MATTEO RICHIARDI and LISA SELLA

Department of Economics and Statistics "Cognetti de Martiis"
Campus Luigi Einaudi, Lungo Dora Siena 100A, 10153 Torino (Italy)
www.unito.it/de

The Department of Economics and Statistics "Cognetti de Martiis" publishes research papers authored by members and guests of the Department and of its research centers. ISSN: 2039-4004

Analysis of Agent-based Models

Jakob Grazzini Matteo Richiardi Lisa Sella

November 15, 2013

1 Introduction

This paper deals with the problem of analyzing the behavior of an agent-based (AB) model. The problem is similar to the one faced by any modeling methodology: the researcher sets up the rules of the game, but does not know in advance the implications of those rules. Actually, it is in this uncertainty about the model outcomes, and about the relationship between the model outputs and the model inputs, that rests the value of having a model. However, the techniques to gain understanding about the model behavior differ substantially across modeling methodologies, and they remain quite under-explored in the AB modeling literature. In nuts, in a simulation model only *inductive* knowledge about its behavior can be gained, by repeatedly running the model under different samples from the parameter space.

The analysis of this inductive evidence has to confront with the *a priori* unknown stochastic properties of the model. The easiest case is when, for any values of the parameters, the model is stationary and ergodic: in these circumstances it is generally possible, with a reasonable number of experiments, to characterize both the equilibrium properties of the model and the adjustment dynamics to the equilibria. On the other hand, non-stationarity renders the analytical concepts of equilibrium and adjustment dynamics inapplicable, while non-ergodicity might hinder the same possibility of fully describing the behavior of the model. A preliminary analysis to discriminate between these cases is therefore necessary, and it can only be done by statistical testing. In the chapter we provide examples of the tests that can be used to detect both stationarity and ergodicity.

These properties in turn affect the types of subsequent analyses that can be performed, and the interpretation of the results. The techniques that are used to describe the relationships between the different variables of the model are referred to in general terms as *sensitivity analysis* (SA). Although a complete survey of these techniques is outside the scope of this chapter, we briefly describe them and offer an example of how they can be applied to a specific AB model.

The chapter is structured as follows: section 2 presents a simple characterization of AB models as recursive systems, which allows to discuss the differences with other modeling strategies and provides an analytical framework for discussing the problem of interpreting the model outcomes. Section 3 describes how it is possible to characterize the model in terms of its (statistical) equilibria. The following section 4 surveys the main techniques to perform such experiments, that fall into the broad discipline of SA. Finally, section 5 offers our conclusions.

2 AB models as recursive systems

A rather common misunderstanding about simulations is that they are not as sound as mathematical models. Computer simulations are, according to a popular view, characterized by an intermediate level of abstraction: they are more abstract than verbal descriptions but less abstract than "pure" mathematics. This is nonsense. Simulations *do* consist of a well-defined (although not concise) set of functions, which relate inputs to outputs. These functions describe a fully recursive system and unambiguously define the macro dynamics of the system. In this

respect, AB models are no different from any other model: they are logical theorems saying that, given the environment and the rules described by the model, outputs necessarily follow from inputs. As any other model, they provide sufficiency theorems: the environment and the rules are sufficient conditions to obtain the results, given the inputs. The resort to computer simulations is only an efficient way —given some conditions— to obtain the results.

Let us start from the following general characterization of dynamic micro models. At each time t an agent i, $i \in 1 \dots n$, is fully described by some state variables $\boldsymbol{x}_{i,t} \in \Re^k$.[1] Let the evolution of her state variables be specified by the difference equation:

$$\boldsymbol{x}_{i,t+1} = \boldsymbol{f}_i(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{-i,t}, \boldsymbol{\theta}_i, \boldsymbol{\xi}_{i,t}). \tag{1}$$

where $\boldsymbol{\xi}_{i,t}$ are stochastic terms. The behavioral rules may be individual-specific both in the functional form $\boldsymbol{f}_i(.)$ and in the parameters θ_i, and may also depend on the state $\boldsymbol{x}_{-i}$ of all agents other than i.[2] The set of structural equations (1), defined at the individual level, specifies the *data generating process* (DGP) of the model.

At any point in time, the system is in a state $\boldsymbol{X}_t = (\boldsymbol{x}_{i,t})$ which is the matrix of all individual states. By replacing eq. (1) in the definition above, we obtain

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}(\boldsymbol{X}_t, \boldsymbol{\theta}, \boldsymbol{\xi}_t), \tag{2}$$

the *transition equation* of the system.

Often, we are interested in some aggregate (observable) statistics of our economy. A vector of aggregate variables $\boldsymbol{Y}_t$ is defined as a function over the state of the system, that is as a projection from $\boldsymbol{X}$ to $\boldsymbol{Y}$:

$$\boldsymbol{Y}_t = \boldsymbol{G}(\boldsymbol{X}_t, \boldsymbol{\kappa}_t). \tag{3}$$

where $\boldsymbol{\kappa}_t$ are extra random terms that accounts for measurement errors and other shocks to the observables, if any. This is the *measurement equation*, which together with the transition equation forms the *state space representation* of the system.

The question is whether it is possible to solve equation (3) for each t, regardless of the specification adopted for $f_i(.)$, and the answer is that a solution can always be found, which traces the stochastic evolution of $\boldsymbol{Y}_t$ back to the initial state of the system and the values of the parameters. Expliciting this relationship is complicated because of the random terms $\boldsymbol{\xi}$ and $\boldsymbol{\kappa}$ that enter at every stage. As the behavioral rules f_i and the transformation function $\boldsymbol{G}$ need not to be linear, these random terms cannot be netted out by taking expectations. Therefore, the only way to analyze the mapping of $(\boldsymbol{X}_0, \boldsymbol{\theta})$ into $\boldsymbol{Y}_t$ is by means of Monte Carlo analysis, by simulating the model for different initial states and values of the parameters, and repeating each simulation experiment many times to obtain a distribution of $\boldsymbol{Y}_t$.

However, because digital computers are deterministic machines, it is possible to further pin down the formalization above.[3] In a digital computer random terms are not truly random: they are generated by an algorithm which produces sequences of numbers that resemble the properties of random numbers. Accordingly, these numbers are referred to as *pseudo-random*, and the algorithm is called *random number generator*. Each sequence is identified by a seed (the so-called *random seed*) which guarantees reproducibility, and which is either user-defined or taken from some environmental variable (as the computer clock). Therefore, the random terms

[1] Categorical variables can be indexed with integer values (e.g. 0 for unemployed, 1 for employed).

[2] Here and in the following we use "behavioral rules" and similar terms in a loose sense that encompasses the actual intentional behaviors of individuals as well as other factors such as technology *etc.*

[3] *Analog computers* exist in which continuously variable physical quantities, such as electrical potential, fluid pressure, or mechanical motion, are used to represent (analogously) the quantities in the problem to be solved. Answers to the problem are obtained by measuring the variables in the analog model. Analog computers are not deterministic, as physical quantities cannot be measured with absolute precision. Though digital computing has taken the lead, analog computers have been widely used in simulating the operation of aircraft, nuclear power plants, and industrial chemical processes.

$\boldsymbol{\xi}$ and $\boldsymbol{\kappa}$ are a deterministic function of one or more random seeds $\boldsymbol{s}$, and equations (2)-(3) reduce to

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}(\boldsymbol{X}_t, \boldsymbol{\theta}, \boldsymbol{s}) \tag{2'}$$
$$\boldsymbol{Y}_t = \boldsymbol{G}(\boldsymbol{X}_t, \boldsymbol{s}). \tag{3'}$$

Now, define $\boldsymbol{Z}_0 = \{\boldsymbol{X}_0, \boldsymbol{s}\}$ as the *initial conditions* of the system, which include the initial state plus the random seeds. By iteratively substituting $\boldsymbol{X}_{t+1}$ with $\boldsymbol{X}_t$ using (2′), we get

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{F}(\boldsymbol{F}(\cdots \boldsymbol{F}(\boldsymbol{Z}_0, \boldsymbol{\theta}) \cdots)) \\ &= \boldsymbol{F}^t(\boldsymbol{Z}_0, \boldsymbol{\theta}) \qquad (2'') \\ \boldsymbol{Y}_t &= \boldsymbol{G}\big(\boldsymbol{F}^t(\boldsymbol{Z}_0, \boldsymbol{\theta})\big). \qquad (3'') \end{aligned}$$

The law of motion (3″) uniquely relates the value of $\boldsymbol{Y}$ at any time t to the initial conditions of the system and to the values of the parameters, and is known as the *input-output transformation* (IOT) function. The word "function" is appropriate here, as any particular input given to the computer model will lead to only one output (different inputs might lead to the same output, though).

From a practical point of view, knowledge of the IOT function can be obtained by Monte Carlo analysis, by simulating the model for different initial states, parameter values, and random seeds.

Given this framework, it is easy to discuss the alleged differences in terms of "mathematical soundness" between analytical models and computer simulations. In analytical models, the behavioral rules (1) typically have a simple structure, with either limited or global interaction, and heterogeneity is kept to a minimum. Functional forms are often linear (or linearized). Aggregation is performed on selected variables by taking expectations over the stochastic elements, which are conveniently specified. On the contrary, an AB model poses little restrictions on the specification of equation (1), but this freedom comes at two prices: (i) the modeler has to exert self-discipline in order to stick with the KISS (*keep it simple, stupid*) principle and connect with the existing literature, and (ii) the equation for the macro dynamics (3″) can easily grow enormous, hindering any attempt at symbolic manipulation. Nevertheless, the functions (3″) are completely specified. It is thus possible to explore their local behavior by analyzing the artificial time series produced by the simulation.

To proceed in this analysis, a first step is to choose which variables $\boldsymbol{Y}$ to focus on, and over which time period. The goal being to understand the behavior of the system, we have to characterize its *regularities*.

3 Absorbing and transient equilibria

3.1 Definitions

As we have seen in chapter 2, one important difference between analytical models and AB models lies in the definition of equilibrium. In analytical models, equilibrium is defined as a consistency condition in the behavioral equations: agents (whether representative or not) must act consistently with their expectations, and the actions of all the agents must be mutually consistent. This is the methodological prescription of *rational expectations*, and logically operates at an individual level before the state space representation of eqs. (2) and (3). The system is therefore always in equilibrium, even during a phase of adjustment after a shock has hit. AB models, on the other hand, are characterized by *adaptive expectations*, according to which consistency might or might not arise, depending on the evolutionary forces that shape the system. An equilibrium can therefore be defined only at the aggregate level and only in statistical terms, after the macro outcomes have been observed.

Definition If, in a given time window $[\underline{T}, \bar{T}]$, an aggregate outcome of the AB model Y_t is stationary, the model is said to have a *statistical equilibrium* $Y^*(\boldsymbol{Z}_0, \boldsymbol{\theta})$, with respect to Y and for given initial conditions $\boldsymbol{Z}_0$ and values of the parameters $\boldsymbol{\theta}$.

Stationarity of Y_t implies that each observation carries information about some constant properties of the DGP. By stationarity, here and in the rest of the chapter, we mean weak stationarity. A stochastic process Y_t is (weakly) stationary if $E(Y_t) = \mu$, that is, its expected value is independent of t, and if $\text{Cov}(Y_t, Y_{t-h})$ exists, is finite and depends only on h and not on t. This is different from strict stationarity, which requires that Y_t has the same distribution for every t, and the joint distribution of $(Y_t, Y_{t+1}, Y_{t+2}, ..., Y_{t+h})$ depends only on h and not on t. Note that strict stationarity does not necessarily imply weak stationarity, as finite variance is not assumed in the definition of strong stationarity. An example of a stationary process is white noise $y_t \sim WN(0, \sigma^2)$, with[4]

$$Cov(y_t, y_{t+h}) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0 \end{cases}$$

Note that white noise is stationary but may not be strict stationary.[5] Examples of non-stationary series are the returns in a stock market, where there is clustered volatility (the variance changes over time); trend stationary series that can be transformed to stationary series by subtracting a function of time, and difference stationary series that can be transformed into stationary series by first differentiation.

We distinguish between two types of statistical equilibrium: absorbing and transient.

Definition A statistical equilibrium is said to be *absorbing* if $Y^*(\boldsymbol{Z}_0, \boldsymbol{\theta})$ is stationary in $[\underline{T}, \bar{T} + \tau], \tau \to \infty$.

Definition A statistical equilibrium is said to be *transient* if $Y^*(\boldsymbol{Z}_0, \boldsymbol{\theta})$ is no longer stationary in $[\underline{T}, \bar{T} + \tau], \tau > 0$.

Absorbing equilibria are stable regularities of the system: once the system is in an absorbing equilibrium, it cannot move out. Because the system —conditional on the random seed $\boldsymbol{s}$— is deterministic, for any given value of the initial conditions (including random seeds) and the parameters, there can be at most one absorbing statistical equilibrium Y_a^*. However, there might be many transient statistical equilibria Y_j^*: for instance, a model can oscillate between two (or more) transient statistical equilibria.

3.2 Unique and multiple equilibria

The condition under which the model, irrespective of the initial conditions, will always converge to the same statistical equilibria $Y^*(\boldsymbol{Z}_0, \boldsymbol{\theta}) = Y^*(\boldsymbol{\theta})$ is *ergodicity* of the time series Y_t. Intuitively, this means that changing the random seed, or the initialization, of the simulation does not affect the results, in statistical terms.

Ergodicity is a property that concerns the memory of a process. An ergodic process is characterized by weak memory, so that as the length of the time series we observe increases, our understanding of the process increases as well. In a non-ergodic process, by converse, persistence is so high that little information is provided by analyzing a sub-sample of the process, no matter how long this time series is.

[4] We use the lowercase y to denote examples of generic stochastic processes, while we keep the uppercase Y to refer to the aggregate outcome of an AB model, or in the general discussion.

[5] A Gaussian white noise, where y_t is identically independent distributed (i.i.d.) Gaussian $y_t \sim N(0, \sigma^2)$, is strict stationary.

Ergodicity is sometimes defined as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Cov(Y_t, Y_{t-k}) = 0 \tag{4}$$

which means that events far away from each other can be considered as almost independent. This implies that if some event *can* happen, by waiting for long enough it *will* happen, regardless of what has happened in the past or what will happen in the future.

If Y_t is ergodic, the observation of a unique time series provides information that can be used to infer about the model IOT function (3″): the process is not persistent and in the long run different time series (produced by the same IOT function) will have the same properties. If the number of observations of one single time series increases, the information we have about the IOT function increases as well.

Note that stationarity and ergodicity are different concepts, and one does not imply the other. A typical example of a stationary non-ergodic process is a constant series. Consider a process that consists in the draw of a number y_1 from a given probability distribution, and remains constant thereafter: $y_t = y_1$ for every t. The process is strictly stationary (yet degenerated since y_t is extracted from a distribution with mean y_1 and variance 0), and non-ergodic. Any observation of a given realization of the process provides information only on that particular process and not on the IOT function. An example of a non-stationary but ergodic process, that will be discussed below, is $y(t) = y(t-1) + u(t)$, $u(t) \sim N(0,1)$.

An implication is that if Y_t is ergodic and the model displays one absorbing statistical equilibrium, this equilibrium is unique and depends only on the parameters: $Y_a^*(\boldsymbol{Z}_0, \boldsymbol{\theta}) = Y_a^*(\boldsymbol{\theta})$. Therefore, if the expected value and the variance exist and are finite, the simulated mean $m(Y_a^*, \boldsymbol{\theta})$ converges, both over time and over teh replications r, to the theoretical limiting expected value of the underlying IOT conditional on the parameters used for the simulation, $\mu(Y_a^*, \boldsymbol{\theta})$:[6] If the equilibrium is absorbing, convergence takes place both over time and over the replications r:

$$\lim_{t\to\infty} m_t(Y_a^*, \boldsymbol{\theta}) = \lim_{r\to\infty} m_r(Y^*, \boldsymbol{\theta}) = \mu(Y^*, \boldsymbol{\theta}) \tag{5}$$

The properties of the time series generated by the model are constant both in time and across replications: therefore, they can be inferred from the sample moments.[7]

Moreover, the transient equilibria, if any, are also invariant across different instantiations of the model. Across different replications, $\{Y_{j,r}^*\}\ \forall j$ is a sequence of independent and identically distributed random variables. By the law of large numbers, as the number of replications increases any simulated moment $m_r(Y_j^*)$ converges to the theoretical limiting moment of the underlying DGP conditional on the parameters used for the simulation, $\mu(Y_j^*, \boldsymbol{\theta})$ (if it exists):

$$\lim_{r\to\infty} m_r(Y_j^*) = \mu(Y_j^*, \boldsymbol{\theta}),\ \forall j. \tag{6}$$

and the simulated moments of Y_j^* provide reliable information about the theoretical limiting moments, for $t \in [\underline{T}_i, \bar{T}_i]$.[8]

On the contrary, if Y_t is stationary but not ergodic, different (absorbing and/or transient) equilibria are obtained, for the same values of the parameters, depending on the initial conditions. This can be regarded as a case of multiple statistical equilibria.[9] In order to gain

[6] If Y_t is strictly stationary, any simulated moment —not only the mean— converges to its theoretical counterpart, if it exists.

[7] In other words, the simulated mean of Y, computed either over time or over replications, is a consistent estimator of the theoretical limiting expected value, for $t > \bar{T}$.

[8] The simulated moments are consistent (over the replications) estimators of the theoretical limiting moments, for $t \in [\underline{T}_i, \bar{T}_i]$.

[9] Note the two different flavors of multiple equilibria, over time (different transient equilibria) and over replications (e.g. different absorbing equilibria, depending on the initial conditions, if the model is not ergodic).

understanding about the behavior of the model we need to analyze many artificial time series produced with identical values of the parameters but with different initial conditions, and describe the distribution of statistical equilibria obtained.[10]

Note how flexible and powerful this descriptive apparatus is. For instance, a model can show an absorbing statistical equilibrium for, say, GDP. This means that after an initial adjustment period up to $\underline{T}$, the GDP series becomes stationary, with constant mean and variance. If the system receives a transitory shock, it moves away from the statistical equilibrium. However, once the shock has passed, if the model is ergodic it comes again to the previous steady state, after an adjustment phase. If we rerun the model and shock it 100 times, it would always come down to the same equilibrium: its properties are then suited for estimation. Moreover, it might happen that during the adjustment process some other transformation of the state of the system, for instance the speed of adjustment to the equilibrium level of GDP, becomes stationary. This new regularity breaks down when GDP reaches its steady state: it is therefore a transient statistical equilibrium. Again, if the model is ergodic the properties of the transient equilibrium are invariant across simulation runs and can be used for estimation.

It is possible that a model displays no absorbing equilibrium for a given variable of interest. To continue with our example, think of the evolution of GDP, with business cycle fluctuations of different amplitude and duration, and intermittent crises. This is an interesting case for many AB modelers, who essentially see the world as a disequilibrium process. Even in such situations however, it might be possible to find statistical regularities with respect to some other variable, for instance the distribution of the lenght and depth of recessions. If this is the case, we are back to equilibrium analysis, and if the model is ergodic we might be able to estimate it. Moreover, these regularities might be only transient, and vanish as the simulated time goes by (think for example of the effects of fiscal policies on GDP growth, which are very different depending on whether the economy is close to full employment or not). Again, if they are stable enough across different replications of the model, we might use them for estimation. If, on the other hand, the model exhibits no regularities whatsoever, no matter how the data are transformed, then the model is not suited for estimation, and one might argue that it is also of limited explanatory (not to say predictive) help: "everything can happen" is hardly a good theory. So, when AB researchers speak of disequilibrium or out-of-equilibrium analysis, what they have in mind, we believe, is really transient statistical equilibrium analysis of some sort, and this is in principle amenable to estimation.

To recap, understanding whether a simulated time series produced by the model is stationary and whether it is ergodic is crucial for characterizing the model behavior. The prescription therefore —following Hendry (1980)— can only be "test, test, test".

A number of stationary tests are available and can be performed on the simulated time series. In the Appendix, we describe as an example a non-parametric test for the stationarity of given moments of a simulated time series. Non-parametric tests are in general more suited for AB models, as they do not impose structure on the IOT function of the model, which at this stage of the analysis is still largely unknown. Moreover, the limited power of many non-parametric tests can in principle be overcome by increasing at will the length of the artificial time series, something that cannot obviously be done with real data.[11]

On the contrary, ergodicity is in general not testable in the real data, as we typically have only one historical time series available. This of course does not mean that the issue must be ignored in empirical applications: if the real world ("Nature") is non-ergodic, we cannot use the observed data to make inference about the *real world* IOT function, nor about the structure of the underlying DGP. Indeed, it is difficult to claim that Nature is in facts ergodic and that present events are not affected by (at least some) event of the past. All the researcher is left

[10] In many applications Y_t may be *subergodic*, that is, ergodic on subsets of the parameters space. A subergodic model is conditionally (to a known set of initial conditions) ergodic.

[11] Computing time can of course be an issue, in practice.

with in this case are statements about the true DGP that are *conditional* on the realization of these past events.[12]

In an AB model, the function to be described is the *model* IOT function; fortunately, the "veil of ignorance" about this function is much lighter than with real data, as the model DGP –the structural equations (1)– is known, while the real world DGP is unknown. In other words, the researcher is the God of her artificial world, although a non-omniscient God: she sets up the rules of the game, but does not know in advance what the results will be. However, she can re-play her artificial worlds at will, thus generating new time series that can provide more information on the behavior of the system.

This difference has an important implication: the ergodic properties of a simulation model are in principle testable, as we can produce as many time series as we wish, as long as we wish. And they should be tested, as we cannot content with conditional statements on initial conditions in understanding our system behavior: real initial conditions being in some sense legitimized by history, while the initial conditions chosen by the experimenter being often more arbitrary.

In the Appendix, we suggest an application of the same non-parametric test used for stationarity for testing ergodicity of the artificial time series generated by an AB model.

3.3 Implications of the stationarity and ergodicity analyses

To summarize, if the model is ergodic — with respect to an outcome Y and for given values of the structural parameters $\boldsymbol{\theta}$— each simulated time series Y_t can be used to characterize the IOT function, at the given values of the parameters, once "enough" time is passed to wash away the memory of the initial conditions. If, in addition, the model is also stationary in a given time interval, the time series can be used to characterize the (absorbing or transient) equilibria of the system.

On the other hand, if the model is non ergodic, each time series Y_t is only informative of one possible outcome, given the values of the parameters. Then, multiple runs of the model should be used and variation *across* runs exploited in order to characterize, in distributional terms, the properties of the system at the sampled values of the parameters.

A natural question then arises whether it is more convenient to always treat the model as non-ergodic, and examine the outcomes of multiple runs –i.e. many "short" time series– rather than only one "long" time series. The answer is that it is often important to characterize the equilibrium of the system, that is its stationary behavior, possibly after a prolonged phase of adjustment: analyzing "long" time series allow to test for stationarity and identify the equilibrium.

A second remark concerns the fact that the stationarity / ergodicity analysis below is only valid locally, i.e. for specific values of the parameters: the model DGP can be stationary or ergodic for some values of the parameters, and non-stationary or non-ergodic for some other values.[13]. Hence, in principle the analysis should be repeated for every sampled point of the IOT function, assuming by a continuity argument that the results also hold in between different points in the parameter space. When the form of the model DGP induces to expect some discontinuity in the behavior of the system for specific values of the parameters, these values should be included in the experimental design and hence duly explored. More in general, the choice of the points in the parameter space to be sampled, together with the overall design of the experiments[14] that are performed in order to gain understanding about the IOT function, is the focus of sensitivity analysis.

[12] Whether this is satisfactory or not depends on the objectives of the analysis.

[13] See the examples in the Appendix

[14] (Box et al., 1978; Kleijnen and van Groenendaal, 1992)

4 Sensitivity analyis of model output

The statistical techniques to analyze the behavior of the IOT function are called sensitivity analysis (SA). SA represents not only the final step in analyzing the model behavior, but can also be regarded as an essential step in the model building process itself, since it provides the analytical tools which allow to simplify the model structure by identifying its nonrelevant parts.

More in details, SA can be defined as "the study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input" (Saltelli et al., 2004). Such definition reflects the modeler's imperfect knowledge of the system, *i.e.* imperfect knowledge of the IOT function. By means of SA the relative importance of the parameters in influencing the model output can be assessed. This also allows to identify possible interactions among the input factors and hence critical regions in the input factor space, with respect to the conditions of most sensitivity of the model output to some specific factors.

4.1 Settings for SA

There exist three main settings for SA, namely factor screening, local SA, and global SA (Saltelli, 2000).

1. *Factor screening* aims at designing experiments to identify the most influential factors in models characterized by a large number of inputs. Often, only a few input factors have a significant effect on the model output. Screening experiments can be used to rank the input factors in order of importance. The experiments are generally one-at-time (OAT) designs, which evaluate the main effect of changes in single factors (Daniel, 1973), as well as factorial experiments, which evaluate both the main effects and the impact of factor interactions.[15]

2. *Local SA* focuses on the impact of small variations in the input factors around a chosen nominal value (base point). It generally assumes linear input-output relationships and involves the evaluation of the partial derivatives of the output functions with respect to the input factors. The experiments are generally OAT designs.

3. *Global SA* involves the estimation of the factor probability density functions, investigates the role of their scale and shape, and allows for the simultaneous variation of all factors over the whole factor space. The sensitivity is measured over the entire range of each input parameter. Global SA is particularly relevant for AB models as linear OAT sensitivities are ill-suited for nonlinear models characterized by high factor interaction and input uncertainty of various order of magnitude (Cukier et al., 1973).

4.2 Strategies for SA

Different SA strategies may be applied, depending on the setting. Moreover, given the manifold purposes of SA, a preliminary characterization of its objectives is essential. In particular, of fundamental importance is the adoption of the most suitable measure of sensitivity depending on the desired definition of factor importance. In fact, each importance measure generally produces its own factor ranking. Most measures rely on variance decomposition formulas of the model output with respect to the input factors, since the variance is generally regarded as a proxy for uncertainty.[16] In choosing the appropriate sensitivity measure, a model-free

[15] In particular, full factorial designs and fractional factorial designs are commonly adopted. A full factorial design is applied when the factors assume discrete values and considers all possible combinations of values across all factors, allowing to assess both the main effects and the impact of factor interactions. A fractional factorial design consists of a carefully chosen subset of the experimental runs of the corresponding full factorial design.

[16] Other measures can also be used, e.g. entropy (Saltelli et al., 2000)

approach should be followed, *i.e.* choosing a sensitivity measure which is independent of the model characteristics, such as linearity, monotonicity, additivity.

(Saltelli et al., 2008) describe four basic strategies, together with some associated sensitivity measures, namely factor prioritization, factor fixing, factor mapping, and metamodeling.

1. *Factor prioritization* identifies as the most influential the factor X_i causing on average, keeping its distribution fixed, the greatest reduction in the variance of the output Y. The associated sensitivity measure is the first-order sensitivity index S_i of X_i on Y, *i.e.* the average partial variance of Y conditional on the distribution of X_i. In formulas,

$$S_i = \left(V_{X_i}\left(E_{X_{-i}}\left(Y|X_i\right)\right)\right)/V(Y) \tag{7}$$

   where X_{-i} indicates all factors but X_i. The numerator represents the variance, over all possible values of X_i, of the conditional expectation of Y taken over all factors but X_i. The denominator is the unconditional variance of Y.

2. *Factor fixing* aims at simplifying the model by fixing the factors which do not appreciably affect the output in their range of variation. This has to be evaluated taking into account both the first-order effect S_i, which describes the direct effect of X_i on Y, and the higher-order effects, which describe the impact of the interactions between X_i and the other input factors. The sum of all-order effects due to X_i is called the total effect S_{Ti} and represents a suitable sensitivity measure in this setting. Considering the case of a three-factor model $Y = f(X)$, where $X = (X_1, X_2, X_3)$, the first-order effect of X_1 on Y is labeled S_1; the second-order effects of X_1 on Y are S_{12} and S_{13}, respectively representing the effect of the interactions between the couples of factors (X_1, X_2) and (X_1, X_3); finally, the third-order effect S_{123} measures the impact of the interaction among all terms. The total effect of X_1 on Y is given by $S_{T1} = S_1 + S_{12} + S_{13} + S_{123}$.

3. *Factor mapping* concerns the analysis of critical regions in the output distribution, such as threshold areas. It aims at identifying the factors producing realizations of Y into the critical range, rather than those driving the variance of the model output. A useful mapping method is the so called Monte Carlo filtering (Rose et al., 1991), which provides Monte Carlo realizations of Y corresponding to different sampled points in the input factor space. Next, it filters the realizations into two subsets depending on whether they belong or not to the critical region. Then, statistical hypothesis testing is performed to check whether the two subsets represent samples from the same distribution. An input factor is identified as important if the distribution functions of the generated samples prove to be statistically different (Saltelli et al., 2004).

4. *Metamodeling*, or model approximation, aims at identifying an approximation of the IOT function, i.e. a simple relationship between the input factors and the model output that fits the original model well enough. This simplification is due to regularity assumptions that allow to infer the value of the output at untried points in the input space, based on information from nearby sampled points. Hence, a surrogate model is identified, which contains the subset of the input factors accounting for most of the output variability. Clearly, this approach generally misses relevant high-order interaction terms and fails in the case of heavily discontinuous mapping.

This list of SA strategies is not exhaustive and other strategies can be defined, depending on both the specific objective of the SA and further considerations about the model under investigation, e.g. its computational burden, the number of input factors and their theoretical interactions, other features such as linearity, monotonicity, additivity.[17]

[17] Also, many software products for SA exist; (Chan et al., 2000) offer a brief review of some of them.

4.3 SA and AB modeling: some applications

The literature provides just a few examples of SA applied to AB modeling. (Kleijnen et al., 2003) assess the soundness of design of experiments techniques when carefully applied on a small subset of input factors. As an example, (Happe, 2005) and (Happe et al., 2006) propose a combined design of experiment and metamodeling setting applied to AgriPoliS, a spatial and dynamic simulation model developed to investigate the impact of agricultural policies on regional structural change. At first, a full factorial design is used to investigate both the first- and second-order effects of some selected factors on a target response variable. The stochastic nature of the model is faced by running a number of Monte Carlo experiments for each experiment. Then, the simulation output is analyzed by both graphical methods and metamodeling. In particular, an additive polynomial metamodel is estimated to assess the statistical significance of the main effects and the two-factor interactions. A stepwise Ordinary Least Squares procedure is applied to isolate and exclude those factors characterized by low significance levels. Similarly, (Lynam, 2002) adopts a fractional factorial design to investigate the mean effects of a selection of factors in a multi-agent model.

Another SA application to AB modeling is described in (Deffuant et al., 2002). The authors propose an AB model of innovation diffusion to investigate the effects of incentives for the conversion to organic farming in a French department. They explore the huge parameter space and evaluate factor importance by a decision tree approach (Breiman et al., 1984) on a composite error, which estimates the deviance between the real and the simulated data on the number of adopters and their proximity to the initial organic farmers. In particular, their SA algorithm selects the factors and the values corresponding to the smallest deviance by defining a learning set. It identifies the regions in the parameter space which are compatible with the real data. Hence, this represents an example of SA in the factor mapping setting.

4.4 A simple example: SA on a Bass diffusion model with local interaction

To gain further practical intuition in SA, we illustrate a very simple SA exercise based on a modified AB version of the Bass diffusion model. The classical Bass model (Bass, 1969) describes the process of adoption of new products by the interaction between actual and potential users. In particular, the probability of adoption for any potential user at time t depends on an innovation coefficient p, an imitation coefficient q, and the cumulative distribution function of adoptions $F(t)$, *i.e.*

$$Pr_t(adoption) = (f(t))/(1 - F(t)) = p + qF(t) \tag{8}$$

where $f(t)$ is the probability density function of adoptions, and $p + q < 1$. The innovation coefficient p measures an external (advertising) effect, the imitation coefficient q represents the internal (word-of-mouth) effect, and $F(t)$ can be interpreted as a global interaction term. In fact, this is an aggregated model describing diffusion dynamics in an homogeneous population in a fully-connected network, where each individual is aware of and influenced by the adoption behavior of the whole population. The model has an analytical solution,

$$f(t) = \frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p}e^{-(p+q)t}} \tag{9}$$

and captures the typical S-shaped adoption curve of many products. By reviewing the empirical work applying the Bass model to new products introduction, (Mahajan, Muller, Bass, 1995) find the average value of p and q to be 0.03 and 0.38, respectively, with p often less than 0.01 and q in the range $[0.3, 0.5]$.

In the original Bass model, every potential customer is linked to all others by the function $F(t)$. Alternative formulations of the network structure yield to different diffusion dynamics (Fibich and Gibori, 2010). In particular, we present the case of a small-world network characterized by an average number n of bidirectional connections per agent. Thus, the probability of

adoption for the i-th potential user does no more depend on the global interaction term $F(t)$, but on a local interaction term $A_{i,t}$, defined as the share of individuals connected to agent i who have adopted, *i.e.*

$$Pr_{i,t}(adoption) = p + qA_{i,t} \tag{10}$$

The analysis of this model is particularly simple because its stochastic properties are immediate to check: the model is ergodic, with a deterministic absorbing equilibrium (everybody eventually adopts) which is achieved in finite time, given $p > 0$, irrespective of q and n. So, our interest lies in characterizing how the three input parameters (p, q, n) affect the adjustment process to the equilibrium, i.e. the adoption dynamics.

The following SA exercise focuses on the effects of the parameters onto two output statistics Y: the cumulated curve of adoptions and the time of adoption of the 50th percentile of the population.

Figure 1 shows the sensitivity of the cumulated adoption curve at the variation of one parameter at time around a reference parameter configuration, *i.e.* $p = 0.03, q = 0.4, n = 5$ on a population of 1,000 agents. To get rid of random effects in the generation of the network, average results over 50 runs for every triple of parameters are reported. In particular, the top panel shows how adoption drastically speeds up when the effect of the external influence increases, for values of p sampled in the range $[0, 1]$ by increasing steps of 0.05; the middle panel illustrates how the adoption dynamics become more and more similar when the internal influence is higher, for values of q sampled in the range $[0, 1]$ by increasing steps of 0.05; finally, the bottom panel shows that the cumulated adoption curves are quite indistinguishable for an average number of connections per agent greater than 6, for integer values of n sampled in the range $[1, 30]$. Hence, n seems to have the weakest effect on the adoption dynamics of the population (for large n), while p somehow the strongest (for small p).

Similar results are obtained when analyzing the OAT effect of the parameters on the average time of adoption of the 50th percentile over 50 runs. In fact, the bottom panel of figure 2 shows a flat distribution for values of n greater than 6. Moreover, the impact on the adoption speed of high values of q is quite similar, while the 50th percentile adopts in no more than 2 periods for values of p greater than 0.2.

However, the results of an OAT analysis are local, i.e. they are generally strongly influenced by the chosen reference point, and give no information about the eventual impact of the interactions among inputs. In order to overcome this limitation, a global analysis is performed by evaluating a metamodel $Y = g(\boldsymbol{X})$ on artificial data generated by allowing all parameters to change. The metamodel imposes a relationship between the inputs $\boldsymbol{X}$ and the output Y with an arbitrary functional form g, which crucially includes interaction terms (Kleijnen and Sargent, 2000). As an example, we perform a multivariate analysis on 1,000 parameter configurations, obtained by random sampling the inputs from uniform distributions. In particular, p is sampled in the range [0,0.2], q in [0,0.8] and the integer n in [1,30].

The preferred specification is an OLS regression of the average time of adoption of the 50th percentile on a third order polynomial of the innovation coefficient, the imitation coefficient and the average number of connections per agent, plus the second-order interaction terms between p and q and between p and n (the remaining second-order and third-order interaction terms, qn and pqn, turn out to be non significant at the 90% confidence level). Given that pq and pn are strongly significant (see table 1), the OAT analysis confirms to have just local implications.

Moreover, this metamodeling exercise allows us to quantify and compare the impact of variations in the parameter values. Starting from our reference point ($p = 0.03$, $q = 0.4$, $n = 5$), a 20% increase in the value of p lowers the average adoption time of the 50th percentile of about 11%; the same increase in n lowers the adoption time of about 2%, while a 20% increase in q causes a 8.7% variation in the output.

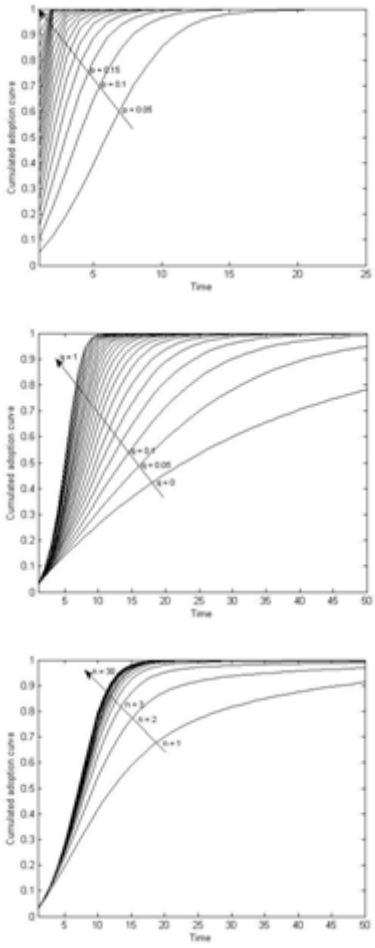


Figure 1: OAT analysis: average cumulated adoption curves over 50 runs, obtained by varying p in the range $[0 : .5 : 1]$ (top panel), q in $[0 : .5 : 1]$ (central panel), and n in $[1 : 1 : 30]$ (bottom panel). Reference parameter configuration: $(pqn) = (0.030.45)$, 1000 agents.

Furthermore, the exercise confirms a weak impact of variations in n when n is high, e.g. a 20% increase when $n = 20$ yields a 2.4% decrease in the average adoption time of the 50th percentile.

5 Conclusions

The discussion above should warn against the use of AB models as an "easy" way of model building that simply allows to bypass the difficulties of deriving analytical results. Indeed, given the higher complexity of AB models (which precludes the derivation of analytical solutions), one has to expect a lot of work to understand their behavior. To sum up, four stages are involved in the analysis of an AB model:

1. definition of the output variable(s) of interest, $\boldsymbol{Y}$;
2. design of an appropriate experimental design, with the definition of the points in the parameter space to be explored;
3. analysis of the stationarity / ergodicity properties of the system at the chosen points;
4. sensitivity analysis of the output variables $\boldsymbol{Y}$ with respect to other variables of the model $\boldsymbol{X}$ and of the structural parameters $\boldsymbol{\theta}$.

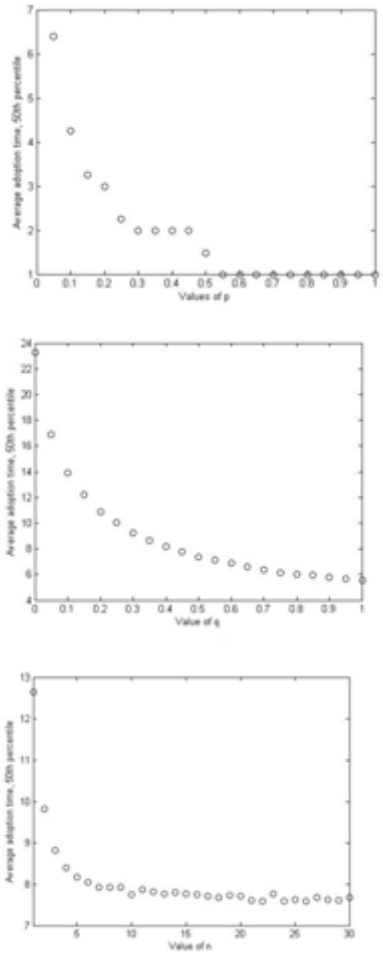


Figure 2: OAT analysis: average time of adoption of the 50th percentile over 50 runs, obtained by varying p in the range $[0 : .5 : 1]$ (top panel), q in $[0 : .5 : 1]$ (central panel), and n in $[1 : 1 : 30]$ (bottom panel). Reference parameter configuration: $(pqn) = (0.030.45)$, 1000 agents.

These steps should not be necessarily undertaken in the order specified above, as there may be feedbacks so that loops might become necessary: for instance, SA could be used to simplify the model structure (the model DGP), which in turn might affect the choice of the output variables Y and the design of the experiments. Similarly, finding that the system is non-ergodic might imply the need to reconsider the design of the experiments, with a higher attention to the effects of the initial conditions.

Unfortunately, such a complete analysis of the model behavior is very rarely done in the literature. In particular, stationarity is often simply checked by visual inspection, ergodicity generally disregarded, and sensitivity analysis at best implemented with a local OAT approach around a baseline configuration. True, global SA strategies with extensive testing for stationarity and ergodicity at every sampled point are very difficult to implement in large and complex models, computationally burdensome and characterized by many output variables. On the other hand, OAT designs around one central configuration (or a limited number of combinations) of the parameters are generally easier to understand, and reduces the need to test for ergodicity and stationarity, given the appropriate continuity assumptions: the tests can be performed at pre-specified intervals of the variable that is allowed to change, assuming that the results also hold for the other sampled values in between.[18]

[18]This can also be done with multi-dimensional designs; however, the identification of an adequate neighborhood of the tested points in which the continuity assumption is supposed to hold becomes more complicated.

time_50	Coeff.	Std.Err.	
p	-417.18	11.25	***
p^2	2824.05	125.46	***
p^3	7264.88	410.95	***
q	-32.31	2.64	***
q^2	28.86	7.52	***
q^3	-14.66	6.19	
n	-0.41	0.08	***
n^2	-0.02	11.25	***
n^3	0.0003	0.0001	
pq	107.44	4.57	***
pn	0.92	0.12	***
$cons$	30.86	0.51	***

*** Significant at the .01%

Table 1: Metamodeling: OLS regression on 1,000 different parameter configurations, obtained by random sampling from uniform distributions in the range p [0,0.2], q [0,0.8], n [1,30]. In order to get rid of random effects, the time of adoption of the 50th percentile is averaged over 50 runs. Adjusted R-squared = 0.84.

These difficulties notwithstanding, the importance of proper formal analysis of AB models should not be downplayed, if the methodology has to gain full respectability among the scientific community. Jointly considered, the techniques reviewed here retain a fundamental role in building and analyzing simulation models; they represent a compelling procedure in model developing, providing tools that map the input factor space into the prediction space and back, as well as techniques to evaluate alternative model structures and the relative importance of each input factor.

A A stationarity test for AB models

The test which we propose to check stationarity is called Runs Test (or Wald-Wolfowitz test). The Runs Test was developed by Wald and Wolfowitz (1940) to test the hypothesis that two samples come from the same population. In particular we employ an extension of the Runs Test to check the fit of a given interpolating function (Gibbons, 1985). Suppose that we have a time series and a function of time that is used to describe its trend. If the trend function fits the time series well, the observations should be randomly distributed above and below the function, regardless of the distribution of errors. The Runs Test tests whether the null hypothesis of randomness can be rejected or not. Given the estimated function, a 1 is assigned to any observation above the fitted line, and a 0 to any observation below the fitted line. Supposing that the unknown probability distribution is continuous, there is a 0 probability that a point lies exactly on the fitted line (if, by accident, it does happen, the point has to be disregarded). The process is then described by a sequence of ones and zeros that represents the sequence of observations above and below the fitted line. The statistics we use to test the null hypothesis is the number of runs, where a run is defined as "a succession of one or more identical symbols which are followed and preceded by a different symbol or no symbol at all" (Gibbons, 1985). For example in the sequence 1,0,0,1,1,1,0 there are 4 runs: {1},{0,0},{1,1,1} and {0}. The number of runs, too many or too few runs, may reflect the existence of non-randomness in the sequence. Following the notation of Wald and Wolfowitz (1940), we define the U-statistic as the number of runs, m as the number of points above the fitted function and n as the points below the fitted function. Under the null hypothesis of randomness around the trend, the mean and variance of the U-statistic are

$$E(U) = \frac{2mn}{m+n} + 1 \tag{11}$$

$$Var(U) = \frac{4mn(2mn - m - n)}{(m+n)^2(m+n-1)}. \tag{12}$$

The asymptotic distribution of U, as m and n tend to infinity (as the observations tend to infinity) is a normal distribution.[19]

To sum up, the Runs test tests the hypothesis that a set of observations is randomly distributed around a given fitting function; it tests whether the function provides a good representation of the observed data. The idea is to use the test described above to check the stationarity of a time series produced by the AB model. The first step is to divide the time series into w windows (sub-time series). Then we compute the moment of order m for each window:

$$\mu_m = \frac{1}{T}\sum_{t=1}^{T} Y_t^m \tag{13}$$

If the moment is constant, then the "window moments" are well explained by the moment of the same order computed over the whole time series ("overall moment"). Here is where the Runs Test is used: if the sample moments are randomly distributed around the "overall moment", it is concluded that the hypothesis of stationarity for the tested moment cannot be rejected. A strictly stationary process will have all stationary moments, while a stationary process of order m in this framework means that the first m non-centered moments are constant.

To run the test we have to choose the length of the artificial time series to be analyzed, together with the length of the windows. Under the null hypothesis, longer windows imply a better estimation of the subsample moments, but at the same time they imply fewer windows (given the length of the time series) and a worse approximation of the distribution of runs toward the normal distribution. The trade off can be solved by using long series and long

[19] The derivation of the finite sample properties and of the asymptotic distribution of U can be found in (Wald and Wolfowitz, 1940) and in (Gibbons, 1985).

windows, a solution which is often feasible in AB models (the only drawback being increased computational time), while it is generally not at hand with real data.

To describe the properties of the test, we check the stationarity of the first moment (mean) of an autoregressive function of the first order:

$$y_t = \theta y_{t-1} + \epsilon_t \tag{14}$$

with $\theta = 0$ (strictly stationary, figure 3 (a)), $\theta = 0.99$ (stationary, figure 3 (b)), and $\theta = 1$(non-stationary, figure 3 (c)), and ϵ_t a random error with uniform distribution $U(-1, 1)$.[20]

We show experiments with different window length s (1,10,50,100,500,1000,5000,10000) using a time series of 100,000 observations/periods. The performance of the test is evaluated, for every window length, over 100 Monte Carlo replications of the stochastic process. By changing the length of the windows we change the number of sub-samples (since the length of the time series is fixed).

Figure 3 shows the sensitivity of the test to different sample length.

The null hypothesis is that the first moment is constant, and in turn that the sub-time series moments are randomly distributed around the overall first moment. We set the tolerated probability of a type I error equal to 0.05: hence, we expect to reject the null hypothesis when the null is true in 5% of the cases; this happens with both $\theta = 0$ and $\theta = 0.99$. It is interesting to note that the length of the windows has no influence in the case of a strictly stationary process. In particular, since every observation has the same distribution, the stationarity can be detected even when the window length is equal to one. However, if $\theta = 0.99$, longer windows are needed to detect the stationarity property in order to allow the sub-time series to converge toward the overall mean; in other words more observations are needed to obtain a good estimation of the sub-sample moments. Non-stationarity is also simple to detect; the test has full power (it can always reject the null when the null is false) for all the window lengths but the ones that reduce the number of windows under the threshold of good approximation of the normal distribution (the test can detect non-stationarity as long as the number of samples is higher than 50).

As an additional experiment, we analyze a time series produced by an AR(1) process as in eq. 14 with $\theta = 0$, but with an error term that is distributed as $U(-1, 1)$ in the first part of the time series and as $U(-8, 8)$ in the second part. Figure 4 shows the distribution of the sub-sample moments around the overall moments. The test (correctly) does not reject stationarity for the the first moment, while it refuses the null hypothesis for the second moment.

The experiment shows the flexibility and the limits of the test. If the length of the time series and the number of windows are properly set, the test is highly reliable, with a power approaching 1. In case non-stationarity is found, standard methods may be used to transform the series in stationary ones (for example detrending or differentiating the series); the non parametric test can then be used on the transformed series.

B An ergodicity test for AB models

The test described below is a test of ergodicity of the moment of order m, where we test its invariance between different replications of the same DGP. To this aim, the Runs test is used again, but this time in the original version presented in Wald and Wolfowitz (1940) to test whether two samples come from the same population. Using the notation of Wald and Wolfowitz, suppose that there are two samples $\{a_t\}$ and $\{b_t\}$, and suppose that they come from the continuous distributions $f_a(a)$ and $f_b(b)$. Let Z be the set formed by the union of a_t and b_t and arrange the set Z in ascending order of magnitude. Eventually, create the set V, that is a sequence defined as follows: $v_i = 0$ if $z_i \in \{a_t\}$ and $v_i = 1$ if $z_i \in \{b_t\}$. We define a run as

[20]The experiment with $\theta = 0.99$ is shown to "test" the test in an extreme case, where the null and the alternative hypothesis are nearly indistinguishable from each other.

in the previous section, and use the number of runs, the U-statistic, to test our null hypothesis $f_a(\cdot) = f_b(\cdot)$. In the event that the null is true, the distribution of U is independent of f_a (and f_b). A difference between $f_a(\cdot)$ and $f_b(\cdot)$ will tend to decrease U. If we define m as the number of elements coming from the sample $\{a_t\}$ (number of zeros in V) and n as the number of elements in Z coming from the sample $\{b_t\}$ (number of ones in V), $m+n$ is by definition the total number of observations. Under the null, the mean and the variance of the U-statistics are (11) and (12), if m and n are large, the asymptotic distribution of U is Normal with asymptotic mean and variance.[21] Given the actual number of runs, U, computed over the two samples, we reject the null hypothesis if U is too low (we test U against its null distribution with the one-tailed test).

The aim here is to use this test as an ergodicity test, supposing that the process has already passed a stationarity test. Intuitively, if the process is ergodic the "horizontal" distribution of moments *within* one (long enough) time series should be the same as the "vertical" distribution of moments *between* different time series, created by different replications of the model.[22] To test the ergodicity of a given moment one long time series is therefore created and divided into sub-samples. As in the previous section, thinks of 100,000 periods for the entire time series, divided into 100 sub-samples of 1,000 periods each. The first sample of moments used for the Runs test (say $\{a_t\}$) is formed by the moments of the 100 sub-samples of this (long) time series. For the second sample of moments (say $\{b_t\}$) we create 100 new time series (by running each time the simulation model with a different random seed, or with different initial conditions) of 1,000 observations each and compute the moment of interest in each of them. Given the two samples of moments we can then apply the Runs test as described above (merge the two samples, arrange the observations in ascending order and compute the runs over the sequence of ones and zeros). Under the null hypothesis, the two samples of moments $\{a_t\}$ and $\{b_t\}$ have the same distribution.

The moments in the two samples have to be computed over time series of the same length (in our example 1,000 periods), because, under the null hypothesis, the variance of the moments depends on the number of observations over which they are computed. If we used longer time series to build the second sample we would produce sample of moments with lower variance, and the Runs test would consider the two samples as coming from different distributions.

As regards the implementation of the test, particular care should be taken when the time series under analysis converges during the simulation toward an asymptotic mean. Suppose that we have a time series that converges to a long run mean in a given number of periods and then stays around that mean for ever. In this case, the stationarity test would correctly deem the process as stationary (with a sufficient long simulation), but even if the process is ergodic, the ergodicity test will result in a refusal of the hypothesis with the process being classified as non-ergodic, since in the time series used to create the second sample of moments the memory of the initial conditions matters more than in the sub-samples coming from the long time series used to create the first sample of moments.

As an example, consider a simple AR(1) process $y_t = 0.99y_{t-1} + u_t$, where $u_t \sim N(1,1)$. The process is ergodic; it starts from zero and converges toward the asymptotic mean $E(y_t) = 100$.[23]

Figure 5 illustrates the problem: the top and middle panels show the long time series used to create the first sample of moments and (one of) the short time series used to create the second sample of moments, respectively; the bottom panel shows the moments computed from then 100 sub-samples of the long time series (dots) and the moments computed from the 100 short time series (squares). The different effect of the initial conditions in the two series is clearly

[21] As in the stationarity test we use the exact mean and variance to implement the test.

[22] Often a replication of a simulation model –an instance of the model producing a stream of artificial data– is called a "run". Here, to avoid confusion with the definition of run used by the Runs test –a sequence of equal values (0 or 1) of an opportunely defined indicator– we use "replication" instead.

[23] In general, for a process AR(1) $y_t = \theta_0 + \theta_1 y_{t-1} + u_t$ where θ_0 represent constant coefficient and $u_t$has zero mean and a given variance, the asymptotic mean is $E(y_t) = \frac{\theta_0}{1-\theta_1}$.

visibile, and creates a convergence problem. The ergodicity test will find significant differences between the two samples, and detect non-ergodicity.

If the series converge toward a long run mean, we have to use its stationary part only for the ergodicity test (we know the series is stationary, at least from some point onward, from the stationarity analysis). Often, the stationary part of the series can be identified by simple visual inspection. For example, to build the second sample we can create a set of time series with 2,000 observations and compute the moments using the last 1,000 observations.

The ergodicity test tells us, for example, whether the first moments of a series can be used as estimates of the true moment of the IOT. The test have to be replicated for every moment under consideration.

To check the performances of the test we define the following process:

$$y_t = \theta y_{t-1} + u_t \tag{15}$$

where $u_t \sim N(l, 1)$. If l is a random variable extracted at the beginning of the process, the process mimics a situation in which the starting conditions have an everlasting effect on the process, and it is not ergodic, as different replications of the process entail different extractions of l. If by converse l is fixed once and for all, the process is ergodic.

In order to assess the performances of the test, we run 5 experiments of 100 replications each of the test, in three different settings ($\theta = 0, \theta = .99, \theta = 1$), both in the case of an ergodic process ($l = \tilde{l} = 0$) and in the case of a non-ergodic process ($l \sim U(-5, 5)$).[24]

Figure 6 shows the result. When the process is ergodic (top panel), with $\theta = 0$, $\theta = 0.99$ the test suggests non-ergodicity in about 5% of the cases (this is the chosen level for the type I error). However, if the process is non-stationary, $\theta = 1$, the test (erroneously) always rejects ergodicity: this is due to the fact that the test cannot distinguish between non-ergodicity and non-stationarity.

The bottom panel of figure 6 shows the results for of a non-ergodic process, where an initial random draw determines the asymptotic mean of the process. The test can detect non-ergodicity with power 1.

To further clarify how the test works, figure 7 shows the two samples of moments used for the test for an ergodic (top panel) and a non-ergodic (bottom panel) process. Simple visual inspection confirms that the two samples come from the same distribution in the case of an ergodic process, but not in the case of a non-ergodic process (the dots come from the first sample, while the squares come from the second sample).

Of course, a process may be ergodic in the first moment but non-ergodic in the second moment or in other order moments.

To analyze the performance of the test with respect to ergodicity in second moments we use the same framework as before (eq. 15) with $u_t \sim N(0, l)$, extracted at the beginning of the process. We consider the case of $l \sim U(1, 5)$ for a non-ergodic process (the variance of the error changes across different replications), and the case of $l = \tilde{l} = 1$ for an ergodic process.

To test the second moment we simply have to build the first sample using the second moment of the 100 sub-samples of the long time series, and the second sample using second moment of the 100 short time series. The test is exactly as above but for the fact that we are comparing second moments.

Figure 8 is the analog of figure 6, and shows the performance of the test on the second order moment of an ergodic (top panel) and a non ergodic (bottom panel) process. Only the case of $\theta = 0$ is considered, corresponding to strictly stationarity, as with $\theta \to 1$ (e.g. $\theta = 0.99$) the

[24] If $\theta = 0.99$, $l = \tilde{l} \neq 0$ and the starting point is $y_0 = 0$ the process generates time series as in figure 5: we may therefore reject ergodicity even if the process is ergodic. To solve this convergence problem, as already discussed, we compute the moments of the second sample (the moments of the 100 replications of the model) by creating time series of 2,000 periods and computing the moments only in the last 1,000 observations.

variance of an AR(1) process tends to infinity [25]: the test always detects non-ergodicity when the process is non-ergodic (full power), and rejects ergodicity in 5% of the cases ((the chosen level for the type I error) when the process is ergodic:

$$var(y_t) = \frac{\sigma_u^2}{1-\theta^2} \tag{16}$$

Finally, note that when the process is non-ergodic in the second moment but ergodic in the first moment, as in the bottom panel of figure 8, the ergodicity test on the first order moment (mean) gives between 20% and 30% of non-ergodicity results. This is because the different variance of the error implies a different variance in the first moments, so despite the fact that the different processes have the same mean, the test detects that "something is wrong".

For completeness, we report the analog of figure 7 for the second moments (figure 9).

[25] more observations are needed in this case to regain full power

References

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, 15(5):215–27.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for experimenters. An introduction to design, data analysis and model building.* Wiley, New York,.

Breiman, L., Freidman, J. H., Olshen, R. A., and Stone, J. H. (1984). *Classification and regression trees.* Chapman & Hall, New York,.

Chan, K., Scott, E. M., and Andres, T. (2000). In Saltelli, A., Chan, K., and Scott, E. M., editors, *Sensitivity analysis*, chapter Software for sensitivity analysis - a brief review. John Wiley & Sons Ltd., Chichester,.

Cukier, R., Fortuin, C. M., Schuler, K. E., Petschek, A. G., and Schaibly, J. H. (1973). Study of the sensitivity of coupled reaction systems to uncertainties in the rate coefficients. i theory. *Journal of Chemical Physics*, 59:3873–78.

Daniel, C. (1973). One-at-time plans. *Journal of the American Statistical Association*, 68:353–60.

Deffuant, G., Huet, S., Bousset, J. P., Henriot, J., Amon, G., and Weisbuch, G. (2002). Agent-based simulation of organic conversion in allier dpartment. In Janssen, M. A., editor, *Complexity and ecosystem management.* Edward Elgar, Cheltenham,.

Fibich, G. and Gibori, R. (2010). Aggregate diffusion dynamics in agent-based models with a spatial structure. *Operations Research, submitted.*

Gibbons, J. D. (1985). *Nonparametric Statistical Inference.* Marcel Dekker Inc., New York, second edition.

Happe, K. (2005). Agent-based modeling and sensitivity analysis by experimental design and metamodeling: an application to modeling regional structural change. In *Proceedings of the XI International Congress of the European Association of Agricultural Economists.*

Happe, K., Kellermann, K., and Balmann, A. (2006). Agent-based analysis of agricultural policies: an illustration of the agricultural policy simulator agripolis, its adaptation, and behavior. *Ecology and Society*, 11(1).

Hendry, D. F. (1980). Econometrics-alchemy or science? *Economica*, 47(188):387–406.

Kleijnen, J. P. C., Sanchez, S. M., Lucas, T. W., and Cioppa, T. M. (2003). A user's guide to the brave new world of designing simulation experiments. CentER Discussion paper, Tilburg University.

Kleijnen, J. P. C. and Sargent, R. G. (2000). A methodology for the fitting and validation of metamodels in simulation. *European Journal of Operational Research*, 1:14–29.

Kleijnen, J. P. C. and van Groenendaal, W. (1992). *Simulation a statistical perspective.* John Wiley & Sons Ltd., New York,.

Lynam, T. (2002). Scientific measurement and villagers' knowledge: an integrative multi-agent model from the semi-arid areas of zimbabwe. In Janssen, M. A., editor, *Complexity and ecosystem management*, pages 188–217. Edward Elgar, New York,.

Rose, K. A., Smith, A. P., Gardner, R. H., Brenkert, A. L., and Bartell, S. M. (1991). Parameter sensitivities, monte carlo filtering, and model forecasting under uncertainty. *Journal of Forecasting*, 10:117–33.

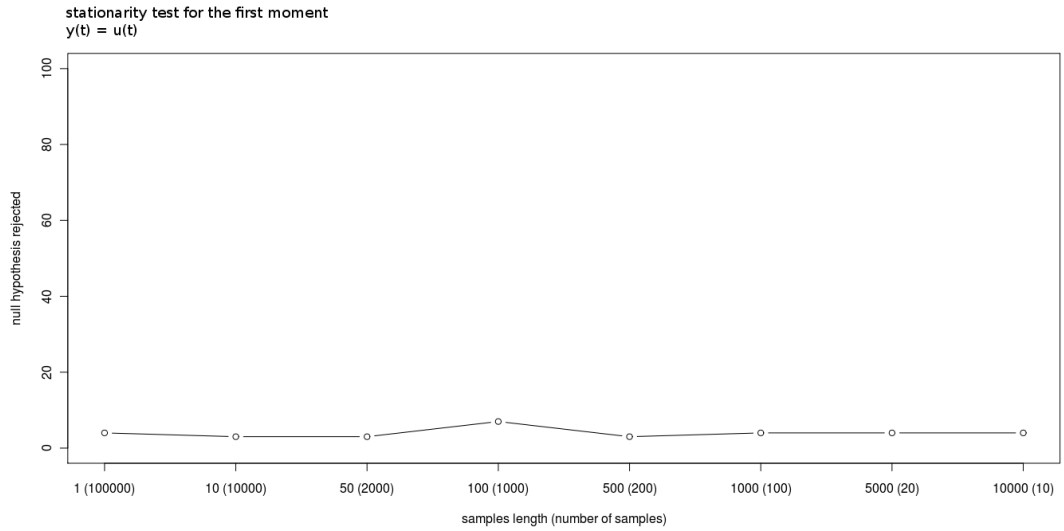
Saltelli, A. (2000). Fortune and future of sensitivity analysis. In Saltelli, A., Chan, K., and Scott, E. M., editors, *Sensitivity analysis*. John Wiley & Sons Ltd., Chichester, West Sussex England.

Saltelli, A., Chan, K., and Scott, E. M. (2000). *Sensitivity analysis*. Jonh Wiley & Sons Ltd., Chichester, West Sussex England.

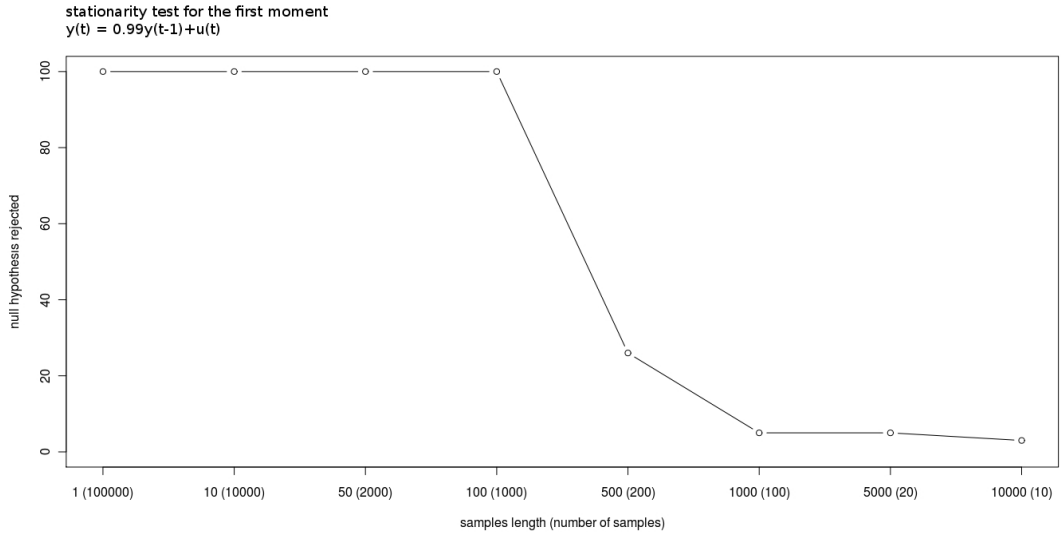
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis. The primer*. John Wiley & Sons, Ltd., Chichester, West Sussex England.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons, Ltd.

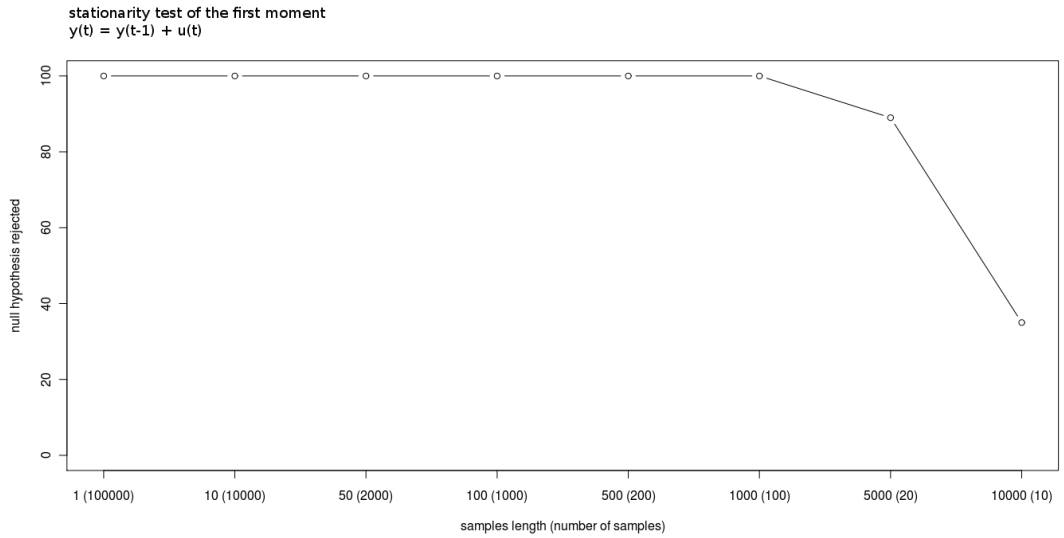
Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162.



(a) Strictly stationary process ($\theta = 0$)



(b) Stationary process ($\theta = .99$)



(c) Non-stationary process ($\theta = 1$)

Figure 3: Rejection rate for the null hypothesis of stationarity (%).

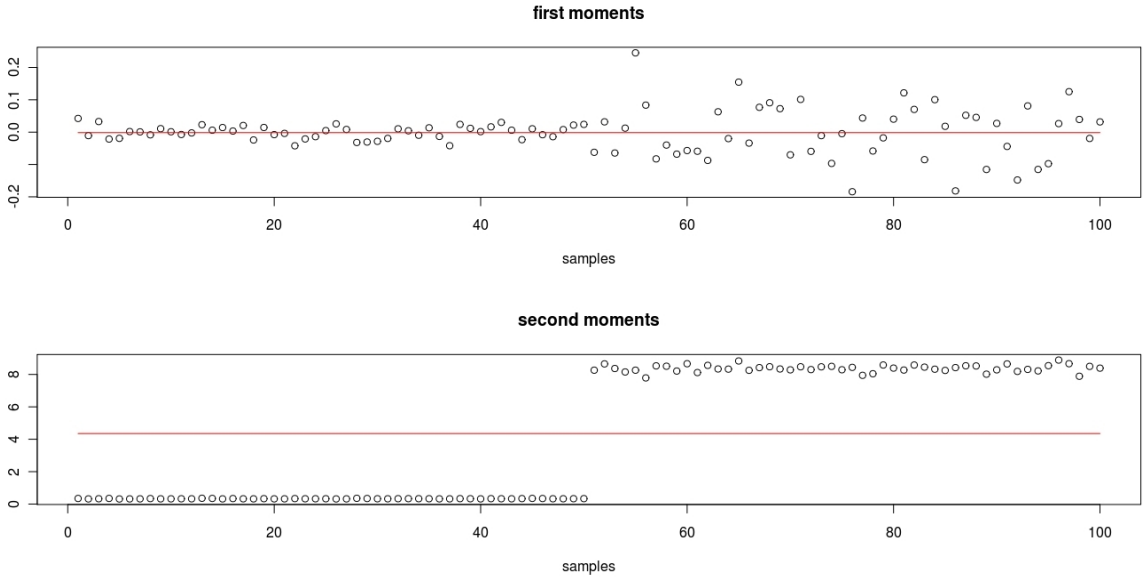


Figure 4: The dots are the sub-sample moments, the line is the overall moment. The first moments are randomly distributed around the overall mean (above). The second moments are not randomly distributed around the overall moments (below).

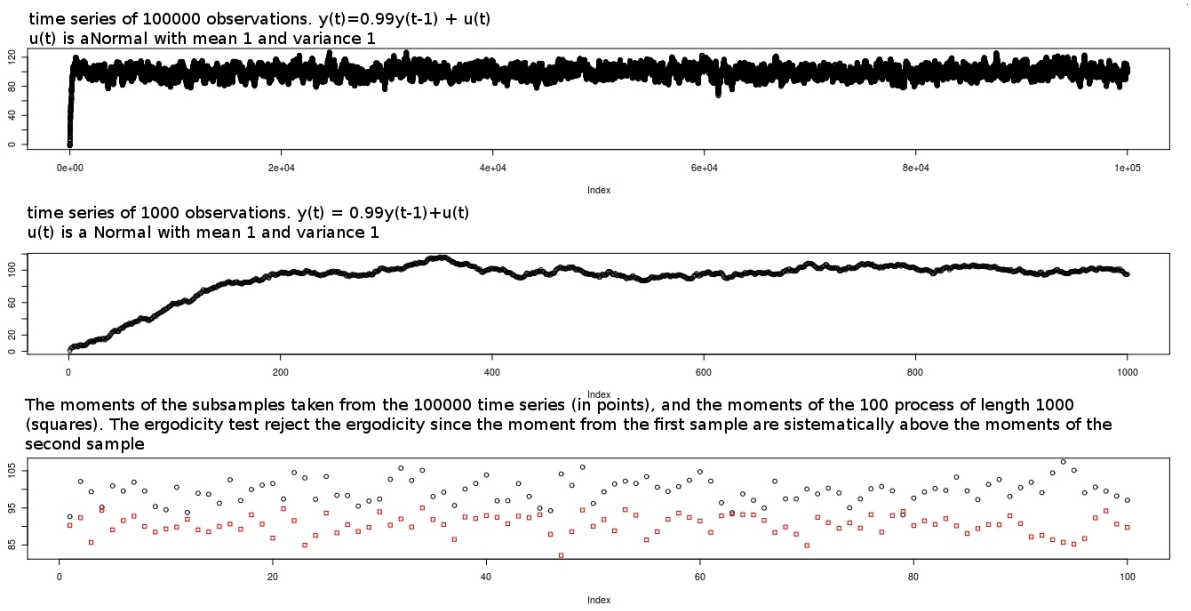


Figure 5: The long process (above), a short process (middle) and the moments computed from the sub-samples of the long process (points) and the moments computed from the short processes (squares).

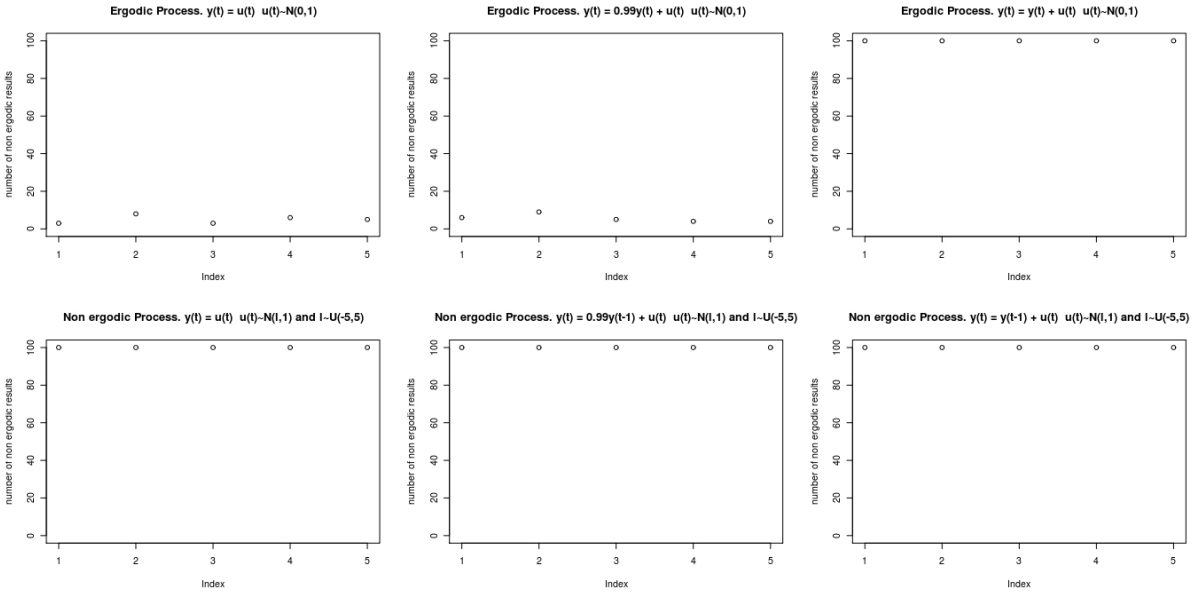


Figure 6: The performance of the ergodicity test. In the top panel the process is ergodic. In the bottom panel the process is non ergodic. One experiment is made by testing 100 times the same process using different random seeds. The experiment is made 5 times for each setting. The graphics displays the number of times the test reject the null hypothesis of ergodicity.

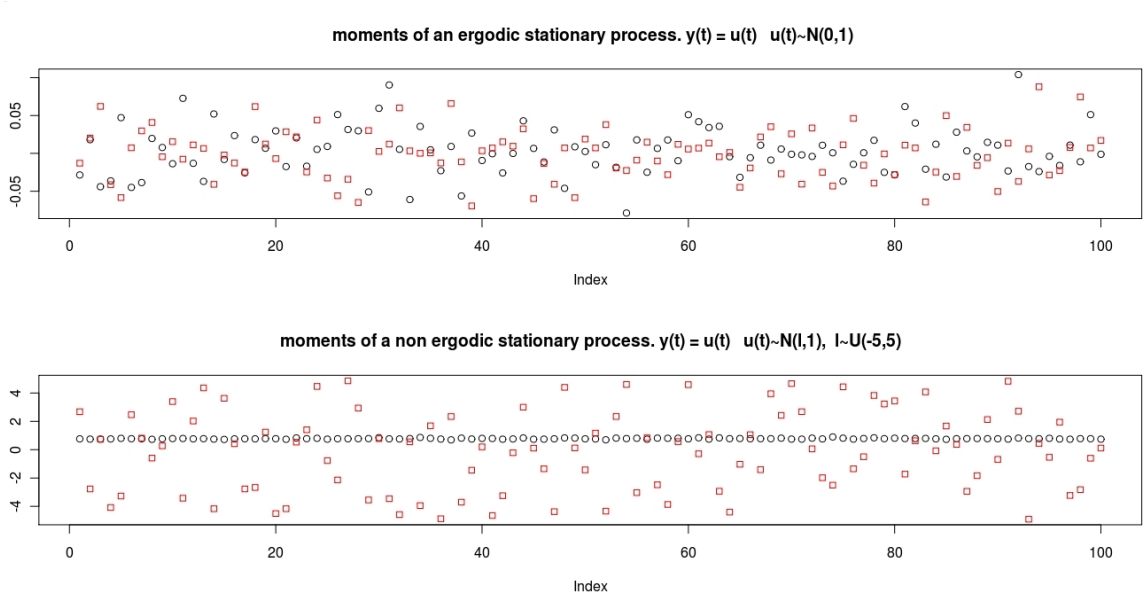


Figure 7: The test checks whether there is a significant difference between two samples of moments: one coming from sub-samples of a (long) time series produced by a single replication of the simulation model (dots) and the other coming from (short) time series produced by multiple replications of the simulation model (squares). Top panel: an ergodic process. Bottom panel: a non-ergodic process.

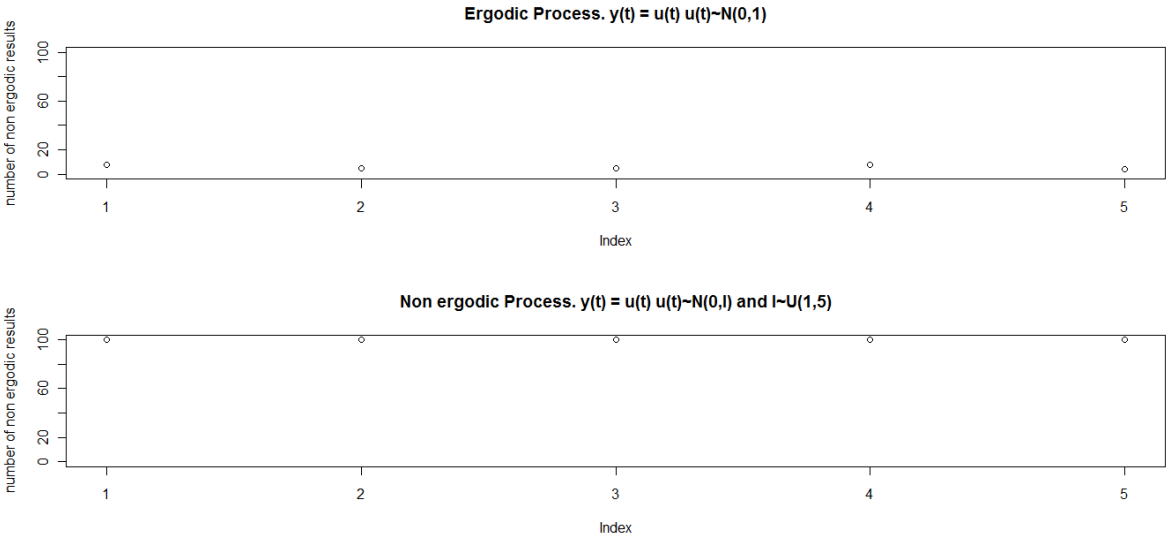


Figure 8: The performance of the ergodicity test on second moments. In the top panel the process is ergodic. In the bottom panel the process is non ergodic. One experiment is made by creating 100 instances of the same process with different random seeds. The experiments has been done 5 times.

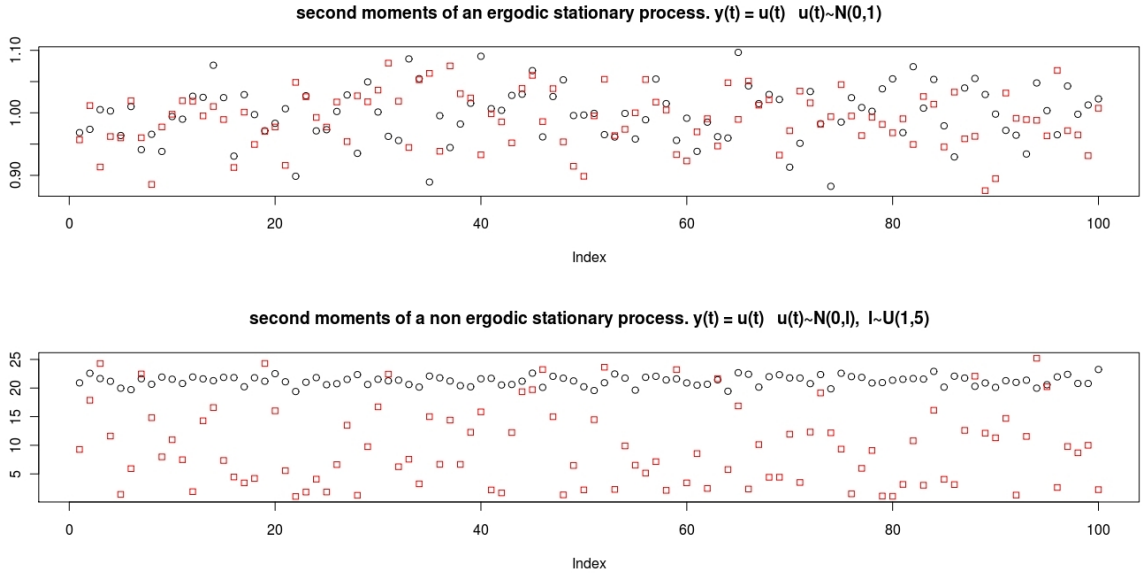


Figure 9: The test checks whether there is a significant difference between two samples of moments: one coming from sub-samples of a (long) time series produced by a single replication of the simulation model (dots) and the other coming from (short) time series produced by multiple replications of the simulation model (squares). Top panel: an ergodic process. Bottom panel: a non-ergodic process.