Universität Bielefeld

Faculty of Business Administration and Economics

Working Papers in Economics and Management

No. 04-2017
June 2017

# An empirical validation protocol for large-scale agent-based models

Sylvain Barde Sander van der Hoog

Bielefeld University
P.O. Box 10 01 31
33501 Bielefeld – Germany

ISSN 2196–2723

# An empirical validation protocol for large-scale agent-based models*

Sylvain Barde[a] Sander van der Hoog[b]

June 20, 2017

**Abstract**

Despite recent advances in bringing agent-based models (ABMs) to the data, the estimation or calibration of model parameters remains a challenge, especially when it comes to large-scale agent-based macroeconomic models. Most methods, such as the method of simulated moments (MSM), require in-the-loop simulation of new data, which may not be feasible for such computationally heavy simulation models.

The purpose of this paper is to provide a proof-of-concept of a generic empirical validation methodology for such large-scale simulation models. We introduce an alternative 'large-scale' empirical validation approach, and apply it to the Eurace@Unibi macroeconomic simulation model (Dawid et al., 2016). This model was selected because it displays strong emergent behaviour and is able to generate a wide variety of nonlinear economic dynamics, including endogenous business- and financial cycles. In addition, it is a computationally heavy simulation model, so it fits our targeted use-case.

The validation protocol consists of three stages. At the first stage we use Nearly-Orthogonal Latin Hypercube sampling (NOLH) in order to generate a set of 513 parameter combinations with good space-filling properties. At the second stage we use the recently developed Markov Information Criterion (MIC) to score the simulated data against empirical data. Finally, at the third stage we use stochastic kriging to construct a surrogate model of the MIC response surface, resulting in an interpolation of the response surface as a function of the parameters. The parameter combinations providing the best fit to the data are then identified as the local minima of the interpolated MIC response surface.

The Model Confidence Set (MCS) procedure of Hansen et al. (2011) is used to restrict the set of model calibrations to those models that cannot be rejected to have equal predictive ability, at a given confidence level. Validation of the surrogate model is carried out by re-running the second stage of the analysis on the so identified optima and cross-checking that the realised MIC scores equal the MIC scores predicted by the surrogate model.

The results we obtain so far look promising as a first proof-of-concept for the empirical validation methodology since we are able to validate the model using empirical data series for 30 OECD countries and the euro area. The internal validation procedure of the surrogate model also suggests that the combination of NOLH sampling, MIC measurement and stochastic kriging yields reliable predictions of the MIC scores for samples not included in the original NOLH sample set. In our opinion, this is a strong indication that the method we propose could provide a viable statistical machine learning technique for the empirical validation of (large-scale) ABMs.

**Keywords:** Statistical machine learning; surrogate modelling; empirical validation.

*We are grateful to Herbert Dawid, Jakob Grazzini, Matteo Richiardi, and Murat Yıldızoğlu for helpful discussions and suggestions. In addition, the paper has benefited from comments by conference participants at CEF 2014 held in Oslo, and CEF 2016 held in Bordeaux. SH acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 649186 - Project ISIGrowth ("Innovation-fuelled, Sustainable, Inclusive Growth").

[a]School of Economics, University of Kent. CT2 7NP United Kingdom. Email: s.barde@kent.ac.uk

[b]Department of Business Administration and Economics, Chair for Economic Theory and Computation Economics, Bielefeld University, Universitätsstrasse 25, 33615 Bielefeld, Germany. Email: svdhoog@wiwi.uni-bielefeld.de

# 1 Introduction

Despite recent advances in bringing agent-based models (ABMs) to the data, the estimation or calibration of model parameters remains a challenge, especially when it comes to large-scale agent-based macroeconomic models. Most methods, such as the method of simulated moments (MSM), require in-the-loop simulation of new data, which may not be feasible for models that are computationally heavy to simulate.[1] Nevertheless, ABMs are becoming an important tool for policy making and it is therefore a relevant issue to be able to compare ABMs to other policy-related models:

> "With regard to policy analysis with structural macroeconomic models, an important question is how agent-based models can be used to deliver answers to the type of questions policy makers typically ask of DSGE models. [...] A comparison of agent-based and DSGE models with regard to such questions should be tremendously useful for practical macroeconomic policy analysis." (Wieland et al., 2012, p.12).

Currently two main challenges exist for agent-based macroeconomists: bringing their models to the data and bringing them to the policy-makers. To address the Lucas Critique (Lucas, 1976) agent-based modellers should generate models that are both empirically validated and policy-relevant. For such empirically and policy-relevant ABMs, the replication of stylized facts does not appear to be a strong enough criterion for model selection since in principle multiple underlying causal structures could generate the same statistical dependencies and therefore match the same set of stylized facts equally well (Guerini and Moneta, 2016). In this view, model comparison and selection should occur at the level of the underlying causal structures, rather than at the level of the statistical dependencies that result from these structures. According to this approach, the objective for policy-relevant ABMs should be to minimize the distance between the causal mechanisms incorporated in the ABM and the causal mechanisms that underlie the real-world data generating process (RW DGP). At this stage, developing more rigorous methods to compare such causal mechanisms is one of the most important open problems in the agent-based modelling community. Resolving this issue will undoubtedly strengthen the reliability, trust and confidence that both academics and policy-makers put in the policy recommendations coming from such models.

An alternative approach is to remain agnostic about the underlying causal structures and instead try to match the conditional probability structures that are embedded in the data. In this view, the appropriate method for model comparison and selection is to minimize the distance between two distributions, namely the distribution of the data resulting from the model and the distribution of the empirical data. This is the approach we have adopted here.

The purpose of this paper is to provide a proof-of-concept of a generic empirical validation methodology for such large-scale simulation models. We introduce an alternative 'large-scale' empirical validation approach, and apply it to the Eurace@Unibi macroeconomic simulation model (Dawid et al., 2016). This model was selected because it displays strong emergent behaviour and is able to generate a wide variety of nonlinear economic dynamics, including endogenous business- and financial cycles. In addition, it is a computationally heavy simulation model, so it fits our targeted use-case.[2]

Our example application uses a large-scale agent-based macroeconomic model, but in principle the method is agnostic about the underlying DGP. For our method it is irrelevant how the model is implemented or how it is simulated, as long as the simulator produces sequential time series data. In addition, the model validation technique that we propose is applicable to any model structure (predictor device) that formally can be represented as a finite-state machine with closed suffix set (FSMX, Rissanen, 1986), or equivalently, by a finite-order Markov process. Most, if not all, computational models in economics can be represented using such formalisms.

In developing our method we rely on the literature on Design and Analysis of Simulation Experiments (DASE, Kleijnen, 2007), which emphasizes the importance of a good Design of Experiments (DoE). This

[1] A possibility for in-the-loop parameter adjustments would be to use computational steering methods, see Wagner et al. (2010), who develop an interactive data exploration framework. The parameter adjustments can take place either within-simulation (parameters are adjusted during a simulation run), or post-mortem-simulation (parameter adjustments occur after a simulation has finished).

[2] To illustrate the computational load, we required approximately 128,250 CPU hours to produce the simulation data needed for the parameter calibrations reported in this paper.

is particularly important when dealing with computationally heavy simulation models or data-intensive methodologies, such a laboratory or field experiments, where the experimenter is not able to generate unrestricted amounts of data. In such cases a sequence of carefully designed experiments is required to obtain a sufficient amount of data to cover the range of possible outcomes. In addition, we adopt a Response Surface Methodology (Box and Wilson, 1951), to ensure the validation protocol is able to handle computationally heavy simulation models. Broadly speaking, our proposed methodology consists of three stages, following Salle and Yıldızoğlu (2014):

1. Start with an experimental design and efficient sampling method, followed by data generation using the simulation model (computationally heavy step).

2. Training and scoring by considering the simulated data as a 'response surface' of the model. That is, as a mapping from a pre-determined set of parameter calibration points into a fitness landscape using the MIC score as fitness metric.

3. Surrogate modelling and validation by optimizing over the interpolated 'MIC response surface' to find new candidate sample points with possibly better performance.

At the first stage (experimental design, efficient sampling and data generation), we use the Nearly-Orthogonal Latin Hypercube sampling (NOLH) method of (Cioppa, 2002; Cioppa and Lucas, 2007) to generate an efficient experimental design matrix consisting of 513 parameter combinations for eight structural parameters of the Eurace@Unibi model. The resulting sample of the parameter space has two important properties which will be critical for the third stage. Firstly, the sample has good space-filling properties, ensuring good coverage of the parameter space. Secondly, the obtained parameter vectors are nearly orthogonal to each other, increasing the effectiveness of the surrogate model. Once the 513 sample points (parameter calibrations) have been generated by the NOLH sampling method these are used to generate corresponding sets of simulated time series using Monte Carlo simulations of the ABM with 1,000 replication runs per sample.

At the second stage (training and scoring) we use the recently developed Markov Information Criterion (MIC, developed by Barde, 2016a,b) to score the synthetic data sets against the empirical data using macroeconomic target variables for 30 OECD countries and the euro area. The MIC methodology is a generalization of the standard concept of an information criterion to any model that is reducible to a finite-order Markov process. Its main feature of interest is that different models can be scored against an empirical data set using merely the simulated time series data, remaining agnostic about the underlying causal mechanisms and without the need to construct a statistical structure. In order to learn the Markov transition probabilities of the underlying DGP, the Context Tree Weighting algorithm (CTW, developed by Willems et al., 1995) is applied to the simulated time series, yielding the conditional probabilities required to score each model against the empirical data. The CTW algorithm is proven to provide an optimal learning performance, in the sense that it achieves the theoretical lower bound on the learning error. As explained in Barde (2016a), this means that a bound correction procedure can be applied to the raw CTW score to correct the measurement error due to learning, thus enabling an accurate measurement of the informational distance between the model and the data.

By considering the model as an input/output response function, i.e. as a mapping from a space of inputs/parameter calibrations into a space of outputs/variables, we use the MIC score as a fitness measure for each parameter calibration. This concept is then used to construct a 'model response surface' that we call the 'MIC Response Surface' of the model consisting of the MIC scores for the 513 NOLH sample points. This surface is a fitness landscape over which we interpolate and minimize in the third stage.

At the third stage (surrogate modelling and validation) we use stochastic kriging (Krige, 1951; Kleijnen, 2017) to construct a surrogate model of the 'MIC Response Surface' generated at stage two. The surrogate model is an interpolation between the realized MIC scores yielding predicted MIC scores for model calibrations that have not yet been tried, possibly resulting in new sample points that are promising candidates for better model calibrations with higher fitness/lower MIC scores. We proceed by identifying the local minima across the interpolated MIC Response Surface, selecting those parameter combinations that provide the lowest predicted MIC score, i.e. providing the best fit to the empirical data by minimizing the relative Kullback-Leibler distance. Next, validation of the surrogate model is carried out by

generating supplementary training data for the newly identified best parametrisations using the original, computationally heavy, simulation model. After re-running the second stage of the analysis on these supplementary samples and cross-checking the realised MIC values against the predicted MIC scores of the surrogate model, we are able to confirm whether or not the surrogate model is able to detect local MIC minima of the MIC Response Surface. Such an internal validation procedure could be seen as an in-sample prediction test of the surrogate model.[3]

The results we obtain so far look promising as a first proof-of-concept for the methodology since we are able to validate the model using the empirical data sets for 30 OECD countries and the euro area. The internal validation procedure of the surrogate model also suggests that the combination of NOLH sampling, MIC measurement and stochastic kriging yields reliable predictions of the MIC scores for samples that are outside of the original NOLH sample set. In our opinion, this is a strong indication that the method we propose could provide a viable statistical machine learning technique for the empirical validation of (large-scale) ABMs.

This rest of this paper is organized as follows. In Section 2 we give an overview of related literature. Section 3 provides an overview of our methodology. Section 4 gives a brief overview of the model, the parameters selected for calibration, and the empirical data that we used. Section 5 discusses the results. Section 6 concludes with a discussion of current limitations of the protocol and suggestions for future extensions.

# 2 Literature

## 2.1 Key challenges for ABM validation

Ten years ago Windrum et al. (2007) identified several key issues concerning the empirical validation of ABMs in response to a perceived lack of discipline and robustness in the field of agent-based modelling in economics. The first issue is that the neoclassical community has consistently developed a core set of theoretical models and applied these to a wide range of issues. On the other hand, the agent-based community had not yet done so. A second issue is the lack of comparability between different ABMs: "Not only do the models have different theoretical content but they seek to explain strikingly different phenomena. Where they do seek to explain similar phenomena, little or no in-depth research has been undertaken to compare and evaluate their relative explanatory performance." (*ibid.*, p. 198). Finally, a third issue is the relationship between ABMs and the empirical data: "[E]mpirical validation involves examining the extent to which the output traces generated by a particular model approximate reality." (*ibid.*).

Over the course of the last ten years the agent-based modelling community has tackled the first of these important issues successfully, namely to apply the same ABMs to different policy questions and to adopt standard modelling tools. Examples run from agent-based macroeconomic models with deliberately simple structures that focus on a single market transmission mechanism (Arifovic et al., 2013; Assenza and Delli Gatti, 2013; De Grauwe and Macchiarelli, 2015; Riccetti et al., 2015) to models with a more holistic perspective that model a system of integrated markets (Mandel et al., 2010; Dosi et al., 2010; Dawid et al., 2016).[4] The field is at a stage where several large-scale, agent-based macroeconomic models have been constructed with the specific purpose of performing macroeconomic policy analyses (see e.g., Dosi et al., 2015; Dawid et al., 2017) and this now necessitates bringing these models to the data.

Progress has been slower, however, on the two issues of model comparison and empirical validation. A key debate, presented below, is to clarify what the goals of empirical validation of ABMs should be. The existing macro-ABMs are all slightly different in scope and in scale, mainly due to their application to different policy domains. This makes a comparison of their predictive accuracy an important aspect that needs to be investigated.[5]

[3] Note that the additional training data set produced after the kriging step is minor in relation to the initial data volume that needs to be generated for the original training data at stage 1. The supplementary data is needed in order to refine the empirical parameter validation at stage 3.

[4] Since these macro-ABMs do not assume a priori that there is simultaneous market clearing, but also do not assume that all markets are in disequilibrium all of the time, it would perhaps be more appropriate to characterize them as 'integrated market models', rather than as general (dis)equilibrium models.

[5] Indeed, even if two models share exactly the same model structure – for example, if they have the same market structure

The current paper seeks to address the second and third issues identified above by offering a methodology for model comparison and model selection against an empirical dataset. This method can be used to bridge the gap between model selection/comparison and model calibration/estimation since it works on the basis of a pre-existing set of parameter calibrations with no new parameter calibrations being generated iteratively (no in-the-loop simulation).

## 2.2 Validation concepts and methods for ABM

In order to provide some clarity within the wide array of different modelling choices, we classify current models in the Agent-based Macroeconomics literature into two main groups, either small or large-scale agent-based macroeconomic models (see also Richiardi, 2015 and Salle and Yıldızoğlu, 2014 for similar assessments). Due to the distinct approaches it seems quite difficult to make model comparisons or to do quality assessments, which would explain the lack of a general methodology for model comparison and empirical estimation. The problem gets aggravated due to the lack of objective model selection criteria and quantitative measures of fit to data sets.

A generic model classification scheme for model validation is proposed by Epstein and Axtell (1994), who categorize ABMs into four classes according to their level of empirical relevance:

- Level 0: The model is a caricature of reality, as established through the use of simple graphical devices (e.g., allowing visualization of agent motion).
- Level 1: The model is in qualitative agreement with empirical macro structures, as established by plotting, say, distributional properties of the agent population. (This can be associated to matching stylized facts.)
- Level 2: The model produces quantitative agreement with empirical macrostructures, as established through on-board statistical estimation routines.
- Level 3: The model exhibits quantitative agreement with empirical microstructures, as determined from cross-sectional and longitudinal analysis of the agent population.

The current literature on empirical validation of ABMs shows a progression from models at level-1, concerned with the qualitative matching of stylized facts, to models at level-2, concerned with quantitative estimation. The focus is currently shifting towards the development of more rigorous empirical validation techniques. However, all of the approaches that are currently proposed still use macro variables as the observables. The final step, moving towards models at level-3, would require observables at the micro level. A challenge for such a truly agent-based estimation methodology is data availability at the level of individual agents, which would require highly disaggregated data (see also Chen et al., 2014 and Grazzini and Richiardi, 2014).

Lux and Zwinkels (2017) survey the burgeoning literature on the empirical validation of agent-based models over the last decade. They discuss various methods for estimation and calibration of ABMs, covering reduced-form statistical models, Method of Simulated Moments (MSM), numerical Maximum Likelihood (ML) including Bayesian estimation, Markov Chain Monte Carlo (MCMC), Sequential Monte Carlo (SMC), Particle Filter Markov Chain Monte Carlo (PMCMC), and state-space methods. Since it is not our intention to provide a survey ourselves, below we just mention those contributions from the literature that are closest to our proposed methodology.

A particular issue of some importance for large-scale ABMs is in-the-loop data simulation versus post-mortem data analysis. In-the-loop data simulation is often used in estimation algorithms to search the parameter space by iteratively updating the parameter values and then simulating new data for the new parameter constellation. This works well when the fitness landscape is smooth and we can use a gradient search method, but may fail for rugged fitness landscapes. Another issue is that given a computational budget, a gradient search algorithm may be too costly for computationally heavy simulation models. For such cases, it may be better to start with a pre-specified, discrete set of parameter constellations followed by simulations for all points in this restricted space, and to perform a post-mortem analysis of the generated data sets.

and trading protocols – the application of this model to different policy questions may require to simulate it with a different population size or with different parameter constellations.

## 2.3 Surrogate modelling and meta-modelling

Meta-modelling or surrogate modelling could be a source of theoretical discipline for agent-based modellers since it forces the modeller to think about how to formulate a problem in terms of a structural, reduced-form statistical model (Bargigli et al., 2016). The benefit of such a surrogate modelling approach is that it makes it easier to compare two models. If we want to make a model comparison between two large, complex ABMs, we could first create a surrogate model of both of them, and then compare the structure of the two surrogates. The same holds if we want to make a comparison between a model and some empirical data set. We could first create a surrogate model of the synthetic data set and of the empirical data set and then compare the two surrogates. However, this may not hold if the modeller adopts a non-parametric statistical approach or uses machine learning techniques that are purely data-driven and agnostic about the underlying DGP.

A first surrogate modelling approach is to use a structural reduced-form statistical model as the metamodel (Gilli and Winker, 2001, 2003; Mandes and Winker, 2016; Bargigli et al., 2016; Guerini and Moneta, 2016). This method consists of estimating a statistical model on the data produced by the simulation model. If the original simulation model is a high-dimensional, non-linear stochastic model, a clear advantage of the reduced-form meta-model is that it can be used to circumvent the curse of dimensionality when used for practical, policy-relevant analyses. However, there are also several challenges, for example how to select the statistical structure, how many time-lags to use, and how many interaction terms should be included. An example would be to estimate a SVAR model on both the simulated data and the empirical data and then compare the two statistical structures (Guerini and Moneta, 2016).

A second method is to use a state-space representation as the metamodel (Salle and Yıldızoğlu, 2014). The advantage is that there is no need to specify any pre-defined statistical structure, so we can adopt a "let the data speak" methodology. A disadvantage of this approach is that the "let the data speak" methodology does not work in the age of big data, where statistical methods are over-determined by the data (there is too much of it), and a priori theoretical constraints become necessary to restrict the statistical methods being used.

A third approach is to use a statistical machine learning technique to directly extract a meta-model from the data generated by the model (Dosi et al., 2016). This method consists of applying a machine learning algorithm to the simulated time series data without first having to pre-define any particular statistical structure. However, since only the model's observable variables are used, this method remains at the surface. In particular, it does not take into account the mapping from parameters into observable variables in terms of a fitness landscape where the measure of fitness is the model's distance to the data, as a function of the parameter input.

The approach we take in this paper, which is similar in spirit to the third method, is to use indirect inference using statistical machine learning techniques. But we first construct intermediate metrics from the data of the original model, and only then extract a meta-model based on such metrics. The method consists of extracting the conditional probability structure of the state transition matrix for the underlying Markov process, using only the time series data from the original model, and then to measure the distance between the distribution of the simulated data and the distribution of an empirical data series.

The advantage of this approach is that it is purely data-driven and therefore agnostic about the underlying data generating process. No information is needed about the internal structure of the model or of the statistical structure of any surrogate model. This makes the method applicable to any process, even to those for which the data generator is inaccessible. The only requirement is that the data source is able to generate a sufficient amount of sequential time series data to train the algorithm.

A disadvantage of our method could be that without any information about the internal structure of the model, the method may be using the data inefficiently. Providing such additional information about the underlying statistical structure could then enhance the method's effectiveness. Another possible problem with this technique could be that meta-modelling using statistical machine learning could lead to computationally heavy estimation methods. However, a pragmatic trade-off exists between the time taken by a machine to perform extensive computations versus the time spent by an econometrician or other scientist to specify the appropriate statistical structure to estimate. Often, the required computational resources are cheap in comparison to the scientist's salary. Therefore this method may be said to sacrifice some efficiency for more generality, favoring a more generic method that is amenable to formalization in a machine language.

Resource costs that should be taken into account should not only involve the time it takes to perform the computations, but also storage space, maintenance costs of software and hardware, and data archiving costs. The scientific choice of what is the best modelling approach usually does not take such considerations into account, but in the era of Big Data Analytics these might have to be taken more seriously. Furthermore, such computational resource limits may even become prohibitive, depending on the size of the simulation model and the degree of accuracy required for the selected machine learning technique. For example, running certain algorithms on large-scale models such as climate models or large traffic simulations require high-end hardware and software that is usually not available at the level of individual academic institutions. A possible solution would then be to make use of High performance computing (HPC) centres which typically concentrate resources between institutions regionally.

# 3 Methodology

The proposed calibration and validation exercise relies on a combination of four existing methodological approaches which we detail below. These draw broadly on the recommendations of Barde (2016a,b) and Salle and Yıldızoğlu (2014) and provide the major advantage that they are all available as 'off-the-shelf' software, requiring only a coordination of their implementation.

## 3.1 Markov Information Criterion and Model Confidence Set

The Markov Information Criterion (MIC) is a recent model comparison methodology developed in Barde (2016a) that provides a measurement of the cross entropy between a model and an empirical data set for any model reducible to a Markov process of arbitrary order. In an analogous manner to a traditional information criterion (AIC, BIC, etc.), once the cross entropy is measured for each candidate model in the comparison set, taking differences across models provides a measurement of the relative Kullback and Leibler (1951) distance between a model and the data. Its key feature compared to other information criteria, however, is that it only requires an empirical data series and a simulated data series provided by the model, which makes it particularly appealing to agent-based models.

The intuition behind the MIC measurement is that the observed transitions in the simulated data from each model can be used to reconstruct the transition matrix for the corresponding Markov process underlying it. Once the transition matrix of each model is available, it can be used in combination with the observed transitions in the empirical data to provide a score for each model. In practice, this is done in two stages. In the first stage, the simulated data is processed using the the Context Tree Weighting (CTW) algorithm of Willems et al. (1995) in order to reconstruct the transition matrix, which is stored in a binary context tree. In a second stage, the Elias (1975) algorithm provides the MIC measurement by measuring the cross entropy of each observation of the empirical data, based on the conditional probabilities extracted from the context tree.

It is important to point out that the main design priority of the MIC is not the estimation of a model's parameters, but instead the provision of an accurate measurement of the distance between a model and the data, and the ability to statistically test any differences in distance across models to select between them. The first of these two properties is feasible due to the CTW algorithm that is proven to optimally reconstruct transition probabilities for all Markov processes of arbitrary order. As pointed out by Barde (2016a), the implication is that the bias incurred by having to use the frequencies observed in the simulated data to proxy the true underlying probabilities of the Markov process can also be measured and corrected, resulting in an unbiased measurement of the cross-entropy.

The second key property of the methodology is that given a set of at least two models, one can test the statistical significance of differences in the MIC scores across the models. This is possible because the cross-entropy is measured at the level of each individual observation, thus providing a vector of scores over the empirical data rather than a single scalar value. As an example, given two models and $N$ empirical observations, determining if the models are equivalent involves determining whether the mean of the vector of $N$ MIC score differences is statistically different from zero, following for example Diebold and Mariano (1995).

More generally, given a set $M$ of candidate models, the statistical identification of the best model from the $N \times M$ observation-level MIC scores can be carried out using the model confidence set (MCS)

procedure of Hansen et al. (2011). This method identifies the subset of models $\mathcal{M}_{1-\alpha}$ which cannot be distinguished in terms of their predictive ability at the $\alpha\%$ confidence level. As shown in Barde (2016a) and Barde (2016b), the MCS procedure can easily be integrated with the MIC measurements to provide a confidence interval in the space of models around the model that is identified as the 'best fit' according to the aggregate MIC scores.

The method of inference underlying the Model Confidence Set procedure is abduction (Peirce, 1997), also known as retroduction, or inference to the best explanation (Harman, 1965; Lipton, 2004; Halas, 2011). Abduction can be paraphrased as the elimination of implausible explanations from a set of possible explanations. The MCS consists of only those models that could not be eliminated as possible explanations of the empirical data.

One important caveat of the methodology, which will be discussed in more detail in Section 5, is that the current implementation of the MIC protocol (Barde, 2016a) is based on univariate time series. While theoretically there is no obstacle in mapping a multivariate model to its underlying Markov process using the CTW algorithm, the exponential increase in the dimension of the state space that results from integrating multiple variables requires a more efficient implementation of the algorithm, in order to maintain tractable run-times and memory requirements. As a result, in this paper the overall MIC scores for each calibration are obtained by summing over the univariate MIC scores for each individual target variable. This is clearly an extremely simplistic assumption, as it ignores any correlation between the variables. Nevertheless, this strategy is equivalent to that used in naive Bayes classifiers, where the features allowing the classification of an instance are treated as strictly independent, even though they may not be so in reality. Another argument to use the sum of the univariate MICs, rather than a multivariate variant, is that the MIC scores are in fact log-scores, and therefore summing over them is similar to taking the sum of log-likelihood scores. The univariate approach used here should therefore be seen as a first-order approximation, the accuracy of which will be tested by comparing the results to the results from a multivariate implementation of the MIC protocol in the near future.

### 3.2 Sampling design points: Nearly-Orthogonal Latin Hypercube Sampling

Because the MIC is not a parameter estimation methodology, but instead a criterion designed to support model comparison and model selection, it relies on the availability of pre-existing simulated data from a set of candidate models. Given the objective of this paper to evaluate the ability of the MIC to identify 'good' calibrations of the Eurace@Unibi model, this imposes a choice of sampling procedures over the parameter space of the model.

Following the recommendations of Salle and Yıldızoğlu (2014), we use the NOLH sampling method of Cioppa and Lucas (2007) in order to generate a nearly-orthogonal design matrix for the experimental design. These authors argue that an efficient experimental design requires a sampling from the parameter space that is orthogonal and possess good space-filling properties. Orthogonality of the parameter vectors facilitates the identification of the effect of univariate parameter variations on the output variable of interest, while good space-filling properties ensure there is sufficient coverage of the entire parameter space as well as a uniform density of sample points across the space.

The NOLH sampling design possesses both of these properties by construction. First of all, the sample forms a Latin hypercube in the parameter space, therefore every sample point takes a unique value in each dimension. Given a large enough sample size, this provides a high level of resolution over the chosen parameter interval. Secondly, while finding exactly orthogonal Latin hypercubes with good space filling properties can be a difficult problem to solve numerically, Cioppa (2002) shows that it is easier to construct Latin hypercubes with good space-filling characteristics for which the parameter vectors are nearly orthogonal.

Using the extension procedure outlined in Cioppa (2002), a $513 \times 8$ design matrix is constructed from the basic $129 \times 22$ matrix provided in appendix D of that paper. A two-way scatter plot of the sample is provided in Figure 1. The first benefit of using the NOLH sampling approach for the analysis carried out here is that it provides flexibility around the central calibration of the model parameters thus enabling the comparison of the same model calibration to 31 empirical data series.

The second reason is that, as shown in Barde (2016a), when combined with the MIC and MCS methodologies, the NOLH approach allows for the evaluation of the local sensitivity of the model fit

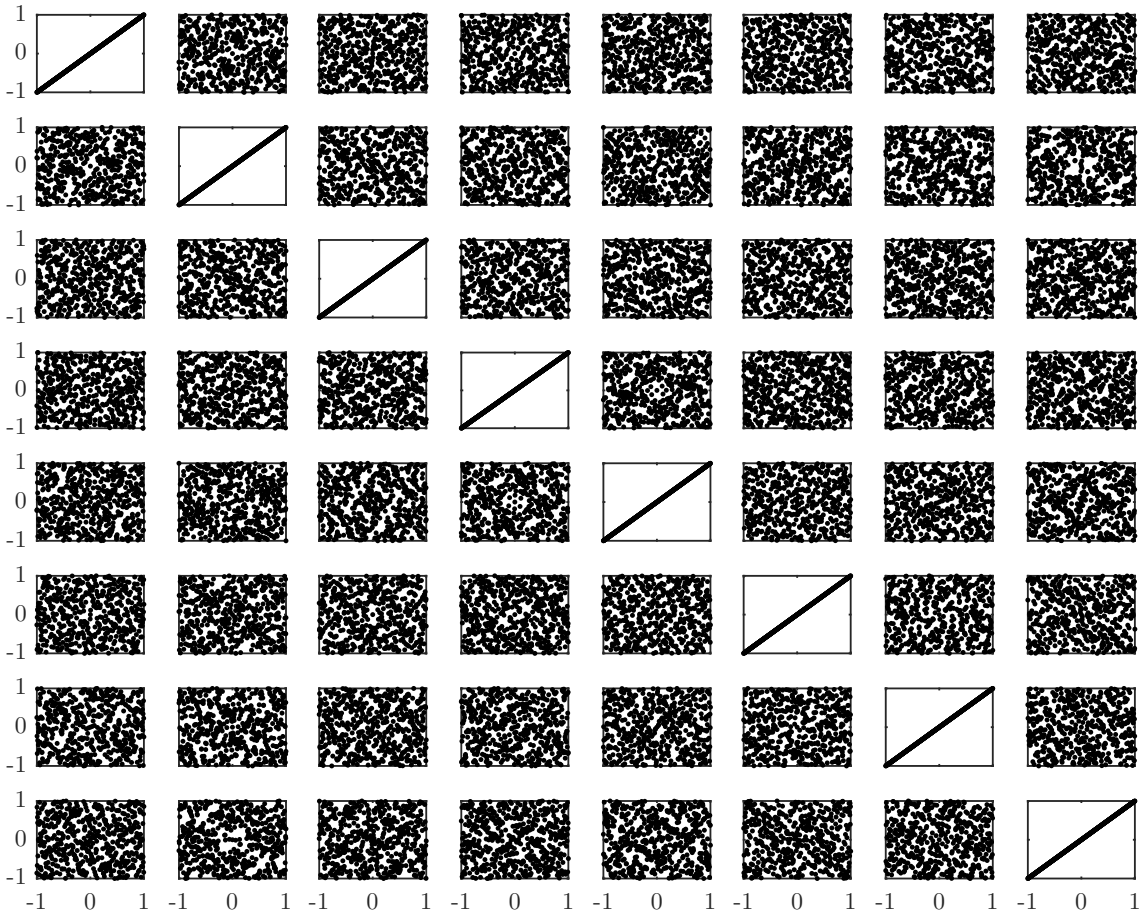



Figure 1: Scatter diagram with 513 samples from an 8-dimensional parameter space. Shown are the projections of $\mathbb{R}^8$ onto the 2-parameter subspaces in $\mathbb{R}^2$.

with respect to variations in the parameter values, essentially providing a confidence interval over the parameter space. This will be important in evaluating the ability of the MIC to discriminate amongst candidate calibrations, which is the main objective of this experiment.

### 3.3 Surrogate modelling: Stochastic Kriging

As previously stated, one of the aims of this paper is to evaluate the ability of the MIC to serve as the basis for the generation of a response surface in the parameter space, through the use of a surrogate model (also known as a meta-model). The motivation is that for large agent-based models the high dimensionality of the parameter space and the emergent behaviour of the model make it computationally prohibitive to specify an I/O mapping from the inputs (model parameters) to outputs (target variables). Instead, the literature suggests to use a surrogate model to approximate the responses of the full-scale ABM. In our case, the aim is to provide a surrogate model for the MIC values over the parameter space in order to identify good parameter calibrations.

Following the suggestion by Salle and Yıldızoğlu (2014) we select stochastic kriging as our surrogate modelling methodology. The main justification for this is theoretical, as kriging is known to provide the best linear unbiased prediction (BLUP). Furthermore, Salle and Yıldızoğlu (2014) argue that the combination of NOLH sampling and kriging is very efficient at providing good surrogate models, due to the near-orthogonality of the sample vectors and the BLUP property. This property is highly desirable in our case given the complexity of the Eurace@Unibi model and its relatively high dimensionality in both parameters and variables.

Table 1: Notation

| | | | |
|---|---|---|---|
| $M$ | Set of candidate models | $t$ | No. of observations in data series (1,000) |
| $\mathcal{M}$ | Set of selected models | $m$ | No. of models to compare (513) |
| $\|\mathcal{M}\|$ | No. of selected models | $n$ | No. of input parameters (8) |
| $\mathcal{K}$ | Set of models after kriging | $q$ | No. of target variables (3) |
| $\|\mathcal{K}\|$ | No. of selected kriging models | $N$ | No. of empirical data series (31) |
| $\Omega$ | Agent state space | $A$ | No. of agents |
| $\Omega^A$ | System state space | $\delta_i$ | partial state transition function |
| $\|\Omega\|$ | No. agent states | $\Delta$ | system state transition function |

A second motivation for this choice is more practical, as the ooDACE toolbox for Matlab already provides kriging as a direct surrogate modelling method. Given that both the MIC and MCS approaches developed by Barde (2016a) have also been implemented in Matlab, this allows for the construction of an integrated protocol using 'off-the-shelf' solutions. More specifically, the procedure used for building the surrogate model for each country is *stochastic kriging* (SK, see Kleijnen, 2017, Sect. 5), where the MIC values obtained for each sample point are treated as noisy measurements, as opposed to *ordinary kriging* (OK) where the observations are treated as deterministic signals. This is done to account for the fact that, as shown in Barde (2016a), the MIC measurement noisy, especially at relatively low levels of training. Using the terminology of Kleijnen (2007), the MIC measurement will already contain an element of intrinsic noise, which needs to be accounted for separately from the extrinsic noise process used by ordinary kriging.[6]

## 3.4 The validation protocol: a formal exposition

The validation protocol consists of a sequence of steps listed in Appendix 6. Below we go through these steps using a formal presentation. Table 1 provides an overview of notation. The first step is to define a simulation model as an Input/Output function, mapping model parameters into model variables.

**Definition 3.1.** (Input/Output Function) Let a simulation model be specified by $n$ parameters (model inputs) and by $q$ target variables (model outputs). Imposing bounds on the parameter ranges yields a domain $\mathcal{D} \subset \mathbb{R}^n$. Let an input signal $s \in \mathcal{D}$ be an $n$-vector, and the corresponding output response of the model denoted by $y \in \mathbb{R}^q$. The simulation model is defined by the Input/Output function

$$f : \mathcal{D} \to \mathbb{R}^q, \qquad s \mapsto y, \qquad y = f(s). \tag{1}$$

In Definition 3.2 we define the *Input/Output Correspondence* of a simulation model.

**Definition 3.2.** (Input/Output Correspondence) Let a simulation model be defined as an I/O function $f$ as in Def. 3.1. Further, let a set of candidate models $M$ be given by a collection of $m$ input signals, denoted by $S = \{s_1, ..., s_m\}$, where $S$ is an element of the sampling space $S \in \mathcal{S} \subset \mathbb{R}^{m \times n}$. The set of input signals $S$ is mapped to a set of output responses $Y = (y_1, ..., y_m)$, which is an element of the output response space $Y \in \mathcal{Y} \subset \mathbb{R}^{m \times q}$. The set $Y$ is obtained by applying the function $f$ element-wise to the elements of $S$, i.e. $\{y_1 = f(s_1), ..., y_m = f(s_m)\}$.

The *Input/Output Correspondence* (IOC) of the simulation model is a many-to-many correspondence,

[6] The implementation of stochastic kriging in the ooDACE toolbox in Matlab is called regression kriging. All settings of the ooDACE toolbox are set to their default values. In particular, the bounds on the 8 hyper parameters are set to [-2,2] (in $\log_{10}$ scale).

mapping a set of inputs $S \in \mathcal{S}$ to a set of output responses $Y \in \mathcal{Y}$:[7]

$$IOC : \mathcal{S} \to \mathcal{Y}, \quad S \mapsto Y, \tag{2}$$
$$\text{where} \quad s_i \mapsto y_i(s_i) \quad \text{for each} \quad i = 1, ..., m.$$

In the following, we will refer to the set of output responses $Y$ as the Output Response Data of the model, given the input signals $S$. In Definition 3.3 the *MIC Response Surface* (MICRoS) is defined, by applying the MIC-measurement function element-wise to the Output Response Data of the model.

**Definition 3.3.** (MIC Response Surface) Let $Y \in \mathcal{Y}$ be the Output Response Data of a model, as defined in Def.3.2. Suppose we have obtained, for each of the output responses, the MIC measurement wrt. some empirical data series. The (univariate) MIC Response Surface (MICRoS) of the model is defined as the mapping:

$$MIC : \mathcal{Y} \to \mathbb{R}^m, \qquad Y \mapsto MIC(Y), \tag{3}$$
$$\text{where} \quad y_{i,j}(s_i) \mapsto MIC(y_{i,j}(s_i)) \quad \text{for each} \quad i = 1, ..., m, \ j = 1, ..., q,$$
$$\text{and} \quad MIC(y_i(s_i)) := \sum_{j=1}^{q} MIC(y_{i,j}(s_i)) \quad \text{for each} \quad i = 1, ..., m. \tag{4}$$

The last line (4) indicates that we consider the univariate variant of the MIC measurement of a model, by taking the sum of MIC scores across the individual target variables $j = 1, ..., q$. Essentially, the MICRoS is an $m$-dimensional manifold embedded in an $(m+1)$-dimensional space. It consists of the realized MIC scores over the sample space $\mathcal{S}$. Figure 2 provides an example of a 2-dimensional response surface and its interpolation surface (Couckuyt et al., 2014). Note that the black dots correspond to the response surface proper, while the smooth surface is the interpolated response surface. For the experiment in this paper, the parameters are: $m = 513$, $n = 8$, $q = 3$, and the number of empirical data series (OECD countries) is $N = 31$. Since in our case the MIC Response Surface is an 8-dimensional manifold embedded in a 9-dimensional space, we cannot easily provide a visualization for it.

Note that the MIC measurements do not constitute an estimation method as such since they only provide us with a metric of the distance between a simulated data series and an empirical data series. Note however that the MIC Response Surface can give us some information about promising not-yet sampled calibration points that lie in between the points that we have actually sampled. To provide us with predicted MIC scores for such non-sampled points in the parameter space, we adopt a statistical surrogate modelling approach that provides us with an interpolation function over the MIC scores.[8] Specifically, this interpolation is carried out by applying stochastic kriging to the MIC Response Surface obtained in Def. 3.3. This final step is formally described in Def. 3.4.

**Definition 3.4.** (Interpolated MIC Response Surface) Let a MIC Response Surface be an $m$-dimensional manifold, as defined in Def.3.3. Applying stochastic kriging as the interpolation function over this manifold yields a continuous sub-manifold of $\mathbb{R}^m$. The result of this interpolation is called the *interpolated MIC Response Surface*, given by:

$$k : \mathbb{R}^m \to \mathbb{R}^m, \qquad MIC(Y) \mapsto k(MIC(Y)). \tag{5}$$

Here the function $k(.)$ represents the application of stochastic kriging to the MIC Response Surface. The entire validation protocol can now be summarized by the sequence of steps in Table 2 (see Appendix 6 for a pseudo algorithm).

[7] In the above definition, the collection of model output responses $y(\mathbf{s})$ is defined in terms of the 'target variables' $y_j$. Such target variables may refer either to variables that are directly observable from the model output (for example, the unemployment rate), or may refer to derived variables constructed after the simulation data has been obtained (target

Table 2: Sequence of steps for the protocol.

| | |
|---|---|
| 1. Parameter selection: | $\mathcal{D} \subset \mathbb{R}^n \rightarrow \mathcal{S} \subset \mathbb{R}^{m \times n}$ |
| 2. Efficient sampling: | $\mathcal{S} \subset \mathbb{R}^{m \times n} \rightarrow S \in \mathcal{S}$ |
| 3. Data generation: | $S \in \mathcal{S} \rightarrow Y \in \mathcal{Y} \subset \mathbb{R}^{m \times q}$ |
| 4. Scoring: | $Y \in \mathcal{Y} \subset \mathbb{R}^{m \times q} \rightarrow MIC(Y) \in \mathbb{R}^m$ |
| 5. Surrogate modelling: | $MIC(Y) \in \mathbb{R}^m \rightarrow k(MIC(Y)) \in \mathbb{R}^m$ |

Next, given the interpolated MIC Response Surface, we try to find all local minima of this surface using a constrained optimization algorithm. This is the final step of the protocol, and is done in order to identify promising new sample points that lie outside of the initial 513 NOLH sample points.[9] The reason we cannot simply take the global minimum of the interpolated surface is due to the intrinsic noisiness of the MIC measurement as a measure for the relative Kullback-Leibler distance, which could result in selecting a false local minimum as the global minimum.

The performance of the resulting kriging model as a surrogate model for the MIC Response Surface over the sample space $\mathcal{S}$ can be evaluated using the leave-one-out cross-validated prediction error (cvpe), which is obtained by successively treating each sample point as an out-of-sample test. Formally, let $\hat{k}_i(MIC(y))$ be the kriging predictor for the interpolated MIC Response Surface at sample point $s_i$, obtained by applying Def. 3.4 to the set of $m-1$ responses that excludes $y_i$. The cvpe is obtained by calculating the squared deviation between this leave-one-out predictor and the actual MIC measurement at $s_i$, as follows:

$$cvpe = \frac{1}{m} \sum_{i=1}^{m} \left( MIC\left(y_i\right) - \hat{k}_i\left(MIC\left(y\right)\right) \right)^2 \tag{6}$$

A lower value of the cvpe indicates that the kriging model is better at predicting MIC scores out-of-sample across the interpolated MIC response surface.

# 4 Application of the protocol

## 4.1 A brief overview of the Eurace@Unibi artificial macroeconomy

Below we give a brief and general overview of the Eurace@Unibi macroeconomic model. For the most up-to-date model description we refer to Dawid et al. (2016). An overview of the results from various policy applications is given by Dawid et al. (2017).

**Consumption goods producers' production output decision.** A typical Eurace@Unibi model economy contains 80 firms producing consumption goods. The output level of firm $i$ is determined according to a Leontief production function with complementarity between physical capital and human

variables could for instance be ratios of certain macroeconomic variables such as the debt-to-GDP ratio). Also, the target variable could refer to any statistics $\mathbf{m}(Y)$ that are constructed from the simulated data series (see Windrum et al., 2007).

[8] It is quite likely that the interpolation of the MIC Response Surface only works well in between the set of sample points already sampled, but not in regions outside of the sampled subspaces. This is another argument why we need to ensure a good initial coverage of the parameter space, and hence the combination of NOLH + SK seems to be a perfect match, as explained by Salle and Yıldızoğlu (2014).

[9] In practice, this is carried out using Matlab's build-in `fmincon` function. The minimization process converges fast and without errors, but is sensitive to the starting point, indicating the presence of local minima in the response surface. In order to ensure robustness, the optimization procedure is run with 1,000 different random starting points within the sample space. For each local minimum found, a count is kept of the number of initial conditions in its basin of attraction, and only those local minima are kept that attract at least 10 out of 1,000 initial conditions. In other words, the basin of attraction of the local minima should have a mass of at least 1%. This ensures only those local minima with a significant basin of attraction are selected.

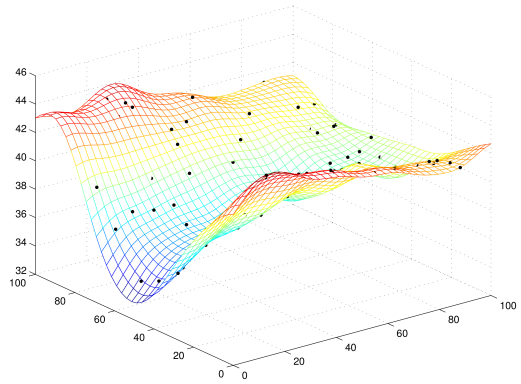



Figure 2: Illustration of a model's response surface. Black dots denote the output response of the model at selected sample points. In this example, the model's output is used directly to obtain a 2-dimensional interpolation surface. In our validation protocol, however, we need an intermediary step to compute the MIC scores, and then we interpolate over these to obtain the MIC Response Surface. Source: Lophaven et al. (2002, p.22).

capital, given by:

$$Q_{i,t} = \sum_{v=1}^{V} \left\{ \underbrace{\min\left[K_{i,t}^{v}, \max\left[0, L_{i,t} - \sum_{k=v+1}^{V} K_{i,t}^{k}\right]\right]}_{\text{effective no. machines used of vintage } v} \times \underbrace{\min\left[A^{v}, B_{i,t}\right]}_{\text{effective productivity}} \right\}, \tag{7}$$

where $v$ denotes the different vintages of physical capital $K_{i,t}^{v}$, with newer vintages being of higher quality and therefore possessing a higher productivity per unit of capital. $L_{i,t}$ denotes the workforce and $A^{v}$ denotes the productivity of capital of vintage $v$. Finally, $B_{i,t}$ denotes the average productivity of the firm's employees, determined by the average specific skill level of the workforce. Note that the sum outside of the outer brackets goes over the vintages $v$, so that we first determine the contribution of each vintage in isolation, and then sum over these to obtain the total production quantity $Q_{i,t}$. The individual contributions consist of two terms: the effective number of machines of vintage $v$ being used to produce, and the effective productivity of these machines, which depends on the average productivity of labour and capital due to the complementarity of these two input factors. In the first term of the overall product, i.e. in the effective number of machines used, $K_{i,t}^{v}$ is the number of units of physical capital of vintage $v$ that need to be operated by employees in a 1:1 ratio, demonstrating the complementarity in real terms. In productivity terms, the complementarity shows itself in the second term of the product, $\min\left[A^{v}, B_{i,t}\right]$.

**Households' consumption choice.** The default model economy is populated by 1600 households and the consumption decision of households is described by a discrete choice model. McFadden (1973, 1980) has shown that the conditional choice probabilities of a population of consumers can be derived as rational choice behavior of an individual consumer, by adopting a random utility framework. Let the utility of consumer $h$ from consuming good $i$ be given by the random utility function

$$u_h(p_{i,t}) = \bar{u}_h - \gamma^c \ln(p_{i,t}) + \epsilon_{h,i,t}, \tag{8}$$

where $\bar{u}_h$ is the base utility of the product (identical across firms) and $\epsilon_{h,i,t}$ captures the contribution of the (horizontal) product properties of product $i$ to the utility of consumer $h$ in period $t$. The term $-\gamma^c \ln(p_{i,t})$ represents the fact that consumers prefer cheaper products to more expensive ones, assuming they cannot discern any quality differences.

Assuming that in each period each consumer chooses the product with the highest utility and that $\epsilon_{h,i,t}$ is a random idiosyncratic term following an extreme value distribution, McFadden has shown that the conditional choice probability of consumer $h$ for product $i$ is given by

$$\mathbb{P}[\text{Consumer } h \text{ selects product } i] = \frac{\exp\left(-\gamma^c \ln(p_{i,t})\right)}{\sum_j \exp(-\gamma^c \ln\left(p_{j,t}\right))}. \tag{9}$$

Here the parameter $\gamma^c$ denotes the price sensitivity of consumers wrt. price differences between the goods to choose from. This parameter can also be interpreted as the intensity of price competition between the consumption goods producers in the model.

**Consumption goods producers' investment decision.** For the investment decision the firm needs to decide how much to invest, but also what vintage to purchase. For the latter choice, the firm considers the ratio between the effective productivity of a particular vintage, $\hat{A}_{i,t}^{eff}(v)$, and its price, $p_t^v$. The vintage choice then follows a similar specification as used for the consumer's choice described above, with the conditional choice probability by firm $i$ for vintage $v$ given by:

$$\mathbb{P}[\text{Firm } i \text{ selects vintage } v] = \frac{\exp\left[\gamma^v \ln\left(\frac{\hat{A}_{i,t}^{eff}(v)}{p_t^v}\right)\right]}{\sum_v \exp\left[\gamma^v \ln\left(\frac{\hat{A}_{i,t}^{eff}(v)}{p_t^v}\right)\right]}. \tag{10}$$

**Banks' interest rate setting on firm loans.** There are typically 20 banks in the economy that provide deposit accounts for households and firms, and maintain the payment settlement system. All money in the Eurace@Unibi artificial economy is stored in bank deposit accounts as electronic money, and all transfers are electronic as well, so there is no need for cash money in the usual sense. Banks provide credit to firms in order to finance their production or to make other payments (for instance, for debt servicing or dividend payouts). However, in the default model implementation, only the consumption goods producers can apply for loans. Household cannot get consumptive credits or mortgage loans to purchase real-estate. Also, the investment goods producer does not need any loans since it does not use any labour or capital inputs to produce, and the investment goods (machines) are produced on demand, so it also does not need any money to make advance payments.

The total volume of credit that can be created by the banks is restricted by banking regulations, possibly resulting in credit rationing for the firms. The floor level for the interest rate on commercial loans is given by the Central Bank's base rate $r^c$ (the policy rate). This is supplemented by a mark-up on the base rate that depends on the financial health of the firm, which is an increasing function of the firm's financial leverage (debt-to-equity ratio). The bank's own funding costs play only a minor role in the interest rate offered to borrowers and are added as a random idiosyncratic term $\epsilon_t^b$ for each new loan contract. The bank's offered interest rate is given by:

$$r_{i,t}^b = r^c \left(1 + \lambda^B \cdot PD_{k,t}^b + \epsilon_t^b\right), \quad \text{where} \quad \epsilon_t^b \sim U[0,1]. \tag{11}$$

Here $PD_{k,t}^b$ is the bank's assessment of the firm's Probability of Default on the loan, which is given by:

$$PD_{k,t}^b = \max\left\{3 \times 10^{-4}, 1 - \exp\left(-\nu(D_{i,t} + \mathcal{L}_{k,t}^b)/E_{i,t}\right)\right\}, \tag{12}$$

where $D_{i,t}$ and $E_{i,t}$ denote the current debt and equity of firm $i$, and $\mathcal{L}_{k,t}^b$ is the new loan indexed by $k$ that is to be added to the total debt $D_{i,t}$. Default values for the parameters are: $\nu = 0.10$ and $\lambda^B = 3$ (these parameters are not varied in this paper).

**Labour market, firms' base wage offer.** The labour market is modelled as a fully decentralized market with direct interaction between individual firms and unemployed job seekers. A firm makes a base wage offer $w_{i,t}^{base}$, driven by labour market tightness. The firm increases its base wage offer by a factor $\varphi$ if it is unable to fill all open vacancies:

$$w_{i,t+1}^{base} = (1 + \varphi) w_{i,t}^{base}. \tag{13}$$

Thus, the parameter $\varphi$ reflects the firm's willingness to pay for higher wages in case of labour market tightness.

## 4.2 Parameter selection and parameter calibration

In this subsection we provide more information on the design of experiments, which parameters were selected, how the parameter ranges were set, and what would be the influence of a variation of a parameter within the context of the model.

The model contains 33 parameters to model 5 markets and 1 household sector. A full list is given in the appendix in Table 13. From this list, eight parameters were selected to be used for the empirical validation experiment in this paper. These eight parameters are listed in Table 3. The parameter ranges were set based on domain knowledge and previous model explorations by the original authors of the model. Below we describe the influence of each of the eight selected parameters.

First, the income tax rate $\vartheta$ for the household sector is the tax rate on all forms of household income, including wages, unemployment benefits, dividend income from share holdings and interest income from bank deposits. Higher values of $\vartheta$ signify less disposable income and lower purchasing power for the household. Hence, this will typically result in lower demand, lower output and higher unemployment rates.

Second, on the consumption goods market, the parameter $\gamma^c$ is the logit parameter in the conditional choice probability of the households' consumption choice problem. In (9) this parameter reflects the sensitivity of consumers wrt. differences in prices between multiple firms selling the same consumption good in the Mall, and measures the intensity of price competition on the market. Higher values of $\gamma^c$ signify a more competitive market, typically resulting in a more unstable economy due to lower profit margins.

Third, on the investment goods market we select two parameters. The first is the parameter $\Delta$ controlling the slope of the technological frontier, which is reflected by a jump in the productivity of the best-practice technology after a successful innovation. Note that the occurrence of innovations is stochastic. Higher values of $\Delta$ signify a greater jump in technological progress, typically leading to higher productivity of the capital stock of the consumption goods producers. But note, however, that only those firms that actually invest in the new vintage will benefit from this increased productivity. Also, since the productivity of physical capital is complementary to the productivity of the labour force in the Leontief production function (7), the firm needs to hire workers with higher specific skill levels in addition, to take advantage of the increased physical capital productivity. The parameter value of $\Delta$ is also used to increase the price of the new vintage when it enters the market. Therefore a higher value of $\Delta$ will also mean that investments in the current best-practice technology becomes more expensive, which might lead some consumption goods producers to rather invest in older vintages first. Hence, the diffusion of new technologies could slow down for higher values of $\Delta$.

Fourth, the second parameter we select on the investment goods market is $\gamma^v$, which is the logit parameter in the conditional choice probability of the consumption goods firms, as given in (10). It controls the intensity of choice to select vintage $v$ over any of the other vintages available. Higher values of $\gamma^v$ imply that a firm is more sensitive to differences wrt. the ratios between the effective productivity and the price.

Fifth, on the credit market we select two parameters as well. The first is the parameter $T$ that reflects the length of the debt repayment period. Increasing the parameter $T$ will give firms more time to repay their debts, but it will also lead to more leverage. In tranquil times this will lead to higher levels of production and more investments, but in times of crisis it results in more financial instabilities.

Sixth, the second parameter on the credit market that we select is the Central Bank base rate $r^c$. In (11) higher levels of $r^c$ lead to higher interest rates on deposits and to higher interest rates on commercial loans. The overall effect is therefore ambiguous, as it may lead to higher income flows for households and firms alike on their deposit accounts through the interest channel, but it may also lead to higher interest payments for firms that need to service their debts.

Seventh, on the financial market we select the parameter $d$ that sets the dividend payout ratio for all (active and profitable) firms and banks. A higher dividend payout ratio $d$ has a positive effect on demand since the dividends are an income flow for the household sector, and thus functions as a purchasing power enhancing effect. But at the same time dividend payout are also an expenditure for the corporate sector, so it may crowd out investments.

Finally, the eight' parameter is $\varphi$, related to the labour market, which controls the factor by which firms adjust their base wage offer. In (13) a higher value of $\varphi$ signifies that firms will increase their base

Table 3: Selected list of parameters from the Eurace@Unibi model used in this paper.

| Parameter | Description | Default | Range |
|---|---|---|---|
| $\vartheta$ | Income tax rate | 0.05 | $[0.01, 0.10]$ |
| $\gamma^c$ | Intensity of consumption choice | 12 | $[0, 40]$ |
| $\Delta$ | Slope of technological frontier | 0.025 | $[0, 0.07]$ |
| $\gamma^v$ | Intensity of vintage choice | 30.0 | $[10, 50]$ |
| $T$ | Debt repayment period | 18 | $[6, 48]$ |
| $r^c$ | Central Bank policy rate | 0.05 | $[0.01, 0.10]$ |
| $d$ | Dividend payout ratio | 0.70 | $[0, 1]$ |
| $\varphi$ | Markup on base wage offer | 0.01 | $[0, 0.01]$ |

wage by a greater percentage in case they are unable to fill their open vacancies. Under conditions of labour market tightness this might lead to wage push inflation.

## 4.3 Simulation: Generation of the training data

Our use case has multiple parameter sets (typically $m = 513$ sets) and multiple Monte Carlo replication runs per set (typically $1,000$ runs per set). This yields an embarrassingly parallel computing problem, since all the individual simulation runs are independent from each other. Therefore they can be arbitrarily distributed across many compute nodes, and launched as a distributed computational problem on a computing cluster.[10]

The simulated data is divided into two sets, an in-sample data set consisting of 99% (990 series per NOLH sample) and an out-of-sample data set for the remaining 1% (10 series per NOLH sample), for each of the 513 NOLH samples. The 99% set forms the training data that is used by the CTW algorithm to build the set of 513 context trees corresponding to each calibration. The trees encode the reconstructed Markov transition matrices that can be used to extract the conditional probabilities required to score the calibrations on the empirical data. The out-of-sample 1% set is used for an internal validation exercise, in order to investigate whether the training data is sufficient to give the MIC discriminatory power over the 513 NOLH samples. Such an internal validation exercise could be seen as an out-of-sample prediction test. The exercise uses each trained context tree to score the $513 \times 10$ runs belonging to the out-of-sample data set in order to establish whether the trees are able to identify the specific calibration they were trained on. If this is the case, then we can conclude that the MIC is able to discriminate between the calibrations, thus validating the training stage.

## 4.4 Empirical data

The macroeconomic data are used for the empirical calibration exercise covers 30 OECD countries and the euro area. We use three target variables (the names in parenthesis refer to variable names in the Eurace@Unibi model): the harmonized monthly unemployment rate (unemployment rate), the monthly year-on-year growth rate of industrial production, considering only the manufacturing sector (output growth rate), and the monthly year-on-year growth rate of the CPI (inflation rate). The relatively large number of countries examined and the choice of a monthly data frequency are both motivated by the desire to have as large an empirical data set as possible, in order to facilitate the statistical analysis of the MIC scores using the MCS procedure.

A basic description of the data, including the countries used, number of observations and bounds for each series, is provided in Table 9 in appendix C. All data series are taken from the Stats OECD website,

[10] We have used a simple round-robin algorithm (like card shuffling) to allocate the runs to a fixed number of job lists. We used 4750 job lists consisting of 9 consecutive blocks of 12 runs each, yielding 108 runs in total per job list. Each job list takes almost exactly 2 hours of wall-time to complete on a compute node with 2 Westmere hexa-core processors (Xeon X5650, 2.66 GHz) with 48 GB of RAM. The total computational load for generating the training data is therefore $4750 \times 2$ hrs $= 9500$ hrs wall-time, and $114,000$ hrs CPU time or 13 CPU years, respectively.

and the start date corresponds to the point in time at which all three data series became available.[11] The harmonised unemployment rate data is directly available as a monthly rate, whereas the growth rates of manufacturing production and the CPI were downloaded as monthly index values and then transformed into monthly year-on-year growth rates by the authors. Similarly, the simulation data for aggregate firm output and the price index are produced as monthly values first, and then need to be transformed into monthly year-on-year growth rates before we start the validation protocol.

Because the MIC methodology treats both empirical and simulated data sets as if they are the output from a Markov process, all data series need to be discretized before the methodology can be applied. This requires choosing a set of bounds to define the support of each variable, as well as a choice of resolution $r$, in order to determine the number of discrete states of each variable ($2^r$). Discretizing the support of the target variables implies that information is inevitably discarded. However, as explained more fully in Barde (2016a), the MIC is not affected by the discretization procedure if the discretization error is pure white noise, i.e. distributed as an i.i.d. uniform random variable. As a result, one can check that a given choice of bounds and resolution are appropriate for the procedure by testing the error term for this.

The main difficulty encountered in the procedure is to set ranges on the target variables to account for the variability reported in Table 9. The resolution is set to $r = 7$ bits for all three variables while the unemployment rates are restricted to $[1\%, 25\%]$, manufacturing output growth rates are restricted to $[-30\%, 30\%]$ and the inflation rates are restricted to $[-2\%, 20\%]$, with any out-of-bounds observations taking the value of the corresponding bound. Tables 10, 11 and 12, also in appendix C, display the results of the discretization tests for these values of the upper and lower bounds and resolution. The Komogorov-Smirnov test is used to check that the discretization error is uniformly distributed, while two Ljung-Box tests are used to test for independence, one by testing for autocorrelation in the discretization error itself, the other by testing for cross-correlation of the error with the discretized variable.

These tables show first of all that the combination of bounds and resolutions are sufficient in nearly all cases to ensure that the discretization error is uniformly distributed. The few cases where this is not the case, such as the unemployment series for Greece, or the inflation series for Ireland and Mexico, seem to occur due to the relatively large number of out-of-bound observations. Autocorrelation of the discretization error does not seem to be a problem for unemployment, and to a lesser extent for manufacturing output growth, however quite a few countries fail the test for CPI inflation. The test that performs poorest for all three variables is the correlation with the discretized variable, suggesting that the discretization error still contains information with respect to the state variable. However, if the error is uniform and not autocorrelated, this is less of an issue as long as the correlation is low. Of the 21 failed cross-correlation tests over the 3 variables and 31 series, only six have an absolute Pearson correlation coefficient above 0.1, and the largest case observed is -0.134 for unemployment in Chile.

# 5 Results

## 5.1 Internal validation of the MIC

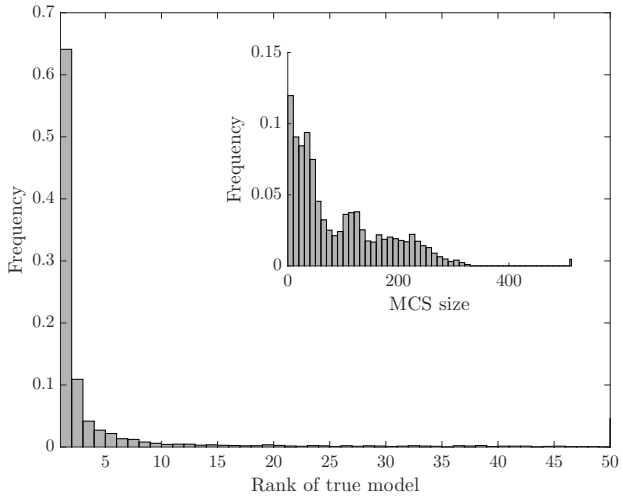
As explained in Section 4.3, after the first stage in which we compute MIC scores using 99% of the training data (for each NOLH sample separately), an internal validation test is carried out using the remaining 1% of the training data, in order to verify that the training data is sufficient to give the MIC discriminatory power over the 513 NOLH samples. This internal validation step is not required in order to obtain the MIC scores on the empirical data, but it is helpful as an intermediate confirmation of the robustness of the training data set.

The internal validation is carried out by treating the simulated runs in the 1% training data set as if they were empirical data. We use these to obtain the MIC score for each of the 513 trees that result from the CTW algorithm on the 99% training data set, for all 3 macroeconomic variables. The 1% training data set consists of $513 \times 10 \times 3$ series, corresponding to the set of 513 calibrations, the 10 runs in the sample and the 3 variables in the analysis. For each of these series, the internal validation test provides a MIC score for each of the 513 trees generated by using the 99% training data set.

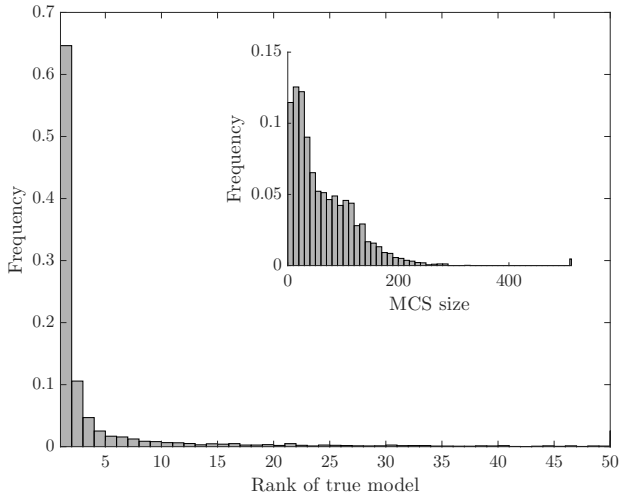
As this is a large set of test scores, the results are summarised using two main statistics. The first is the rank attributed to the 'true' calibration, i.e. the rank given where the training and testing calibrations

[11] `http://stats.oecd.org/`

(a) Unemployment rate




(b) Industrial output growth rate




(c) CPI inflation rate

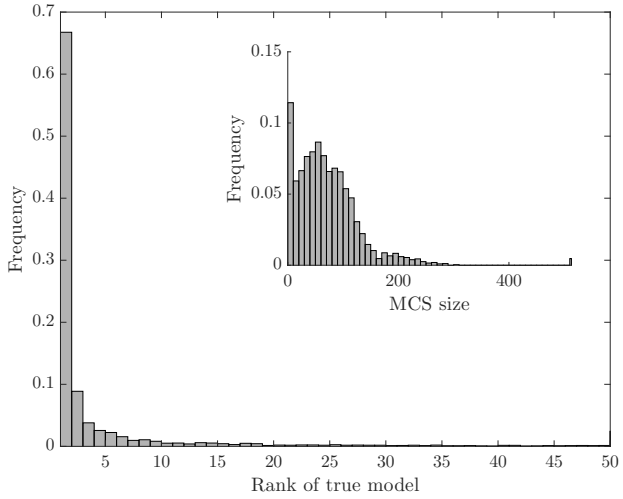



Figure 3: Distribution of true calibration rank
inset: distribution of MCS sizes

Table 4: Size of the MCS.

| Variable | Mean | S.dev | Median | $95^{\text{th}}$ pctl |
|---|---|---|---|---|
| Monthly unemployment rate | 90.78 | 84.63 | 57.00 | 248.00 |
| Monthly output growth rate | 64.55 | 61.14 | 46.00 | 167.00 |
| Monthly inflation rate | 71.26 | 59.88 | 61.00 | 178.00 |

are the same. This tests the ability of the MIC protocol to correctly identify the calibration on which it was trained. As an illustration, suppose that a series from calibration 25 is scored against all 513 trees. If the training is sufficient to enable the MIC to discriminate effectively between calibrations, the best score should come from the tree generated using training data from calibration 25. The second statistic is the size of the MCS at a 90% confidence level. As explained in Barde (2016a), due to the stochastic nature of the simulated data, there is always an element of noise in the MIC measurement which reduces the ability of the procedure to distinguish between two models that are very similar. The size of the model confidence set of 'best' calibrations provides a measurement of this noise, and of the resulting uncertainty in the model rankings.

The results for these two statistics are provided in Figure 3 and Tables 4 and 5. With regards to the rank of the true calibration, the mode at rank 1 for all three variables in Figure 3 shows that the MIC correctly identifies the true calibration in about 65% of cases. Furthermore, the same figure also reveals that for those cases where the true calibration is (incorrectly) ranked lower, the ranking remains high nevertheless: only in very few cases the true model is ranked outside of the top 10. Similarly, the average MCS size points towards a relatively accurate measurement. For industrial production the average MCS size shown in Table 5 is 64.55, suggesting that 88% of the initial 513 calibrations can be rejected. This falls to 82% for the unemployment rate, with an average MCS size of 90.78.

While this suggests that the MIC is reliable in identifying the true calibration and eliminating incorrect ones, the variability in MCS sizes, nevertheless suggests that the measurement can be quite noisy. Both the MCS size distributions in the inset figures of Figure 3 and the $95^{\text{th}}$ percentiles in Table 5 show that in a significant number of instances the MCS is much larger (in the 100-200 range). The MIC measurement using the 99% training data set is therefore quite noisy, which is expected to affect the empirical analysis. This will be discussed further in the following sections below.

Table 5: Rank of the true model.

| Variable | pct $\in$ MCS | Median Rank | Mean rank |
|---|---|---|---|
| Monthly unemployment rate | 0.97 | 1.00 | 7.73 |
| Monthly output growth rate | 0.96 | 1.00 | 6.01 |
| Monthly inflation rate | 0.95 | 1.00 | 6.07 |

## 5.2 Empirical scores on the target variables

As explained in Section 3.1, the aggregate MIC score for each of the 513 calibrations in the NOLH sample is simply the sum of MIC scores obtained from each of the 3 empirical macroeconomic variables, effectively assuming the variables are uncorrelated. An important consequence of this is that it is possible to run the MCS procedure for each variable individually, and assess the ability of the procedure to discriminate between calibrations at the level of a single variable. As will be shown later on, this intermediate step also allows us to establish that some variables are more noisy than others, in the sense that the MIC has less power for certain variables than for others.

Table 6 shows the size of the MCS per macroeconomic variable at the 90% significance level for all 31 empirical data series (countries and euro area). The main finding of this variable-level analysis is that the performance of the MIC varies significantly across variables: the MCS size for the unemployment

Table 6: MCS size per country and per target variable, at 90% significance.

| Country | Unemp. | Output | Infl. | Country | Unemp. | Output | Infl. |
|---|---|---|---|---|---|---|---|
| Austria | 1 | 66 | 71 | Japan | 5 | 97 | 210 |
| Belgium | 3 | 25 | 119 | Korea | 1 | 118 | 156 |
| Canada | 48 | 54 | 124 | Luxembourg | 3 | 26 | 155 |
| Chile | 12 | 42 | 113 | Mexico | 1 | 143 | 199 |
| Czech Republic | 15 | 154 | 152 | Netherlands | 15 | 110 | 67 |
| Denmark | 28 | 31 | 89 | Norway | 1 | 62 | 117 |
| Estonia | 64 | 131 | 93 | Poland | 1 | 81 | 167 |
| euro area | 3 | 101 | 100 | Portugal | 21 | 49 | 198 |
| Finland | 37 | 157 | 122 | Slovak Republic | 1 | 34 | 139 |
| France | 2 | 71 | 134 | Slovenia | 38 | 100 | 6 |
| Germany | 30 | 35 | 53 | Spain | 1 | 114 | 216 |
| Greece | 23 | 71 | 163 | Sweden | 90 | 58 | 74 |
| Hungary | 10 | 83 | 158 | Turkey | 3 | 67 | 47 |
| Iceland | 2 | 95 | 113 | United Kingdom | 38 | 188 | 161 |
| Ireland | 2 | 23 | 100 | United States | 23 | 13 | 188 |
| Italy | 3 | 109 | 136 | | | | |

Unemployment - Monthly harmonised unemployment rate
Output - Monthly year-on-year change in output
Inflation - Monthly year-on-year CPI inflation rate

variable is systematically much smaller than for the other two variables, with the exception of Slovenia. This suggests that very few of the unemployment dynamics generated by the model are close to the ones observed in reality, as measured by their MIC score, allowing the MCS procedure to eliminate a large majority of them. Conversely, the larger MCS sizes for the two remaining variables suggest that many model calibrations provide near-equivalent performance, and it is less easy to distinguish the best ones. It is important to note that the relatively smaller MCS sizes indicate that this is less of an issue for manufacturing growth compared to CPI inflation, even though both are affected to some extent.

## 5.3 Aggregate MCS and kriging sample

As stated in Section 3.1, summing over the observation-level vectors of MIC scores for the three individual macroeconomic variables results in an aggregate vector of scores for each calibration and each empirical data series. This set of aggregate scores is used in two ways. Firstly, as was the case for the individual variables in Table 6, an MCS analysis is run on the aggregate score in order to identify the subset of best-performing calibration points. The results of this analysis are shown in the first three columns of Table 7, which respectively identify the size of the MCS for each country, the ID of the best performing sample point, and the MIC score corresponding to this sample. A first interesting result is that while the variable-specific MCS sizes shown in Table 6 were relatively large for two out of three variables, the MCS sizes at the aggregate level are nevertheless reasonable in magnitude. A second result is that there is evidence of clustering of countries in terms of their best sample/calibration points. Calibration 227, for instance, offers the best performance for 10 out of 31 data series, including Germany and the UK. Similarly, calibration 266 is the best for 5 data series, including France, Italy and the euro area. Calibration 490 similarly covers 3 data series. Indeed, it therefore seems possible to classify 18 out of 31 empirical data series and associate them to three model calibrations.

As stated in Section 3.3, the set of 513 aggregate MIC scores is used to generate a surrogate model for each country by using stochastic kriging (SK, following Krige, 1951), which enables us to identify promising local MIC minima in the parameter space outside of the original NOLH sample. Optimising the per-country surrogate models using stochastic kriging yields a new set of 175 SK-sample points, denoted by the set $\mathcal{K}$. It is important to point out that as shown in column 5 of Table 7, this set only adds a few extra sample points per data series (on average 5.6), so only a small amount of additional training data is needed if one is only interested in a single country or region. This is in line with the

Table 7: MCS (based on aggregated MIC) and Kriging diagnostics.

| | NOLH MCS | | | Kriging | | | post-Kriging MCS | | |
|---|---|---|---|---|---|---|---|---|---|
| Country | $\lvert\mathcal{M}_{90}\rvert$ | Best | $\underline{\text{MIC}}$ | $\lvert\mathcal{K}\rvert$ | $E[\underline{\text{MIC}}]$ | cvpe | $\lvert\mathcal{M}_{90}\rvert$ | $\lvert\mathcal{K}_{90}\rvert$ | $\underline{\text{MIC}}$ |
| Austria | 15 | 461 | 10.95 | 4 | 10.85* | 7.78 | 15 | 0 | 11.29 |
| Belgium | 4 | 227 | 11.36 | 6 | 11.44 | 2.79 | 7 | 2 | 11.41 |
| Canada | 24 | 227 | 10.92 | 6 | 10.91* | 3.66 | 30 | 3 | 11.07 |
| Chile | 1 | 227 | 11.98 | 4 | 12.13 | 3.18 | 1 | 0 | 12.30 |
| Czech Republic | 46 | 89 | 12.48 | 5 | 12.49 | 4.24 | 45 | 1 | 12.49 |
| Denmark | 11 | 195 | 11.77 | 5 | 11.93 | 4.96 | 11 | 0 | 12.07 |
| Estonia | 18 | 227 | 12.32 | 5 | 12.62 | 3.00 | 18 | 0 | 12.65 |
| euro area | 34 | 266 | 9.13 | 6 | 8.78* | 2.47 | 35 | 3 | 9.26 |
| Finland | 38 | 227 | 11.30 | 7 | 11.78 | 2.16 | 42 | 3 | 11.34 |
| France | 61 | 266 | 9.79 | 10 | 9.18* | 2.29 | 63 | 2 | 9.82 |
| Germany | 12 | 227 | 11.83 | 5 | 12.05 | 2.94 | 13 | 1 | 11.87 |
| Greece | 3 | 227 | 11.72 | 5 | 11.98 | 2.31 | 10 | 1 | 11.88 |
| Hungary | 63 | 412 | 12.76 | 6 | 12.71* | 5.21 | 65 | 2 | 12.81 |
| Iceland | 26 | 67 | 14.18 | 3 | 14.13* | 7.74 | 26 | 0 | 14.72 |
| Ireland | 5 | 454 | 13.19 | 5 | 13.42 | 2.69 | 5 | 0 | 13.70 |
| Italy | 63 | 266 | 10.17 | 10 | 9.92* | 2.45 | 66 | 3 | 10.23 |
| Japan | 1 | 490 | 14.54 | 5 | 17.60 | 8.01 | 1 | 0 | 15.46 |
| Korea | 13 | 372 | 12.60 | 3 | 13.02 | 9.57 | 14 | 1 | 13.26 |
| Luxembourg | 4 | 490 | 13.75 | 4 | 15.64 | 6.45 | 4 | 0 | 15.51 |
| Mexico | 15 | 174 | 11.31 | 3 | 10.80* | 10.13 | 16 | 1 | 11.31* |
| Netherlands | 61 | 460 | 10.28 | 6 | 9.92* | 4.98 | 62 | 1 | 10.40 |
| Norway | 16 | 460 | 11.48 | 5 | 11.43* | 7.79 | 19 | 2 | 11.66 |
| Poland | 54 | 255 | 11.62 | 7 | 11.99 | 1.68 | 56 | 4 | 11.64 |
| Portugal | 40 | 227 | 11.94 | 5 | 12.14 | 3.22 | 41 | 2 | 12.07 |
| Slovak Republic | 3 | 454 | 13.10 | 4 | 13.45 | 2.33 | 4 | 1 | 13.08* |
| Slovenia | 48 | 227 | 12.06 | 4 | 12.17 | 5.11 | 53 | 3 | 12.25 |
| Spain | 4 | 266 | 10.89 | 6 | 11.33 | 1.52 | 5 | 1 | 10.88* |
| Sweden | 54 | 490 | 13.40 | 8 | 14.32 | 3.71 | 55 | 1 | 14.16 |
| Turkey | 23 | 266 | 14.03 | 4 | 14.03 | 6.07 | 23 | 2 | 14.15 |
| United Kingdom | 45 | 227 | 9.74 | 6 | 9.31* | 3.83 | 37 | 2 | 9.68* |
| United States | 26 | 256 | 9.43 | 7 | 9.21* | 5.71 | 26 | 1 | 9.42* |

$\lvert\mathcal{M}_{90}\rvert$ - Size of the model confidence set at the 90% level
Best. - ID of the best-performing sample point
$\underline{\text{MIC}}$ - MIC score of the best-performing sample point
$\lvert\mathcal{K}\rvert$ - Number of extra sample points from kriging model
$\lvert\mathcal{K}_{90}\rvert$ - Number of kriging sample points included in the model confidence set at the 90% level

objective presented in the introduction of minimising the amount of in-the-loop simulation data required for the validation of computationally heavy ABMs. The predicted MIC score of the best-performing kriging sample for each country is provided in column 6 of Table 7, with its cvpe as defined in (6) reported in column 7. Those series for which the kringing model predicts a lower MIC score than the best NOLH sample are identified with a star. Crucially, this shows that only in a few cases the kriging models are able to predict an improvement on the best NOLH sample.

A visualisation of the MCS and the kriging sample parameter values is provided in Figures 4 to 7 for Germany, the euro area, Mexico and the US, respectively.[12] These show two-dimensional scatter plots of the parameter values of the NOLH confidence set (in grey) and the kriging samples (in black). The best performing sample in each category is identified with a $\times$. The $\blacklozenge$-symbol identifies the kriging local minimum that dominates in the random 1,000 initial conditions used in the kriging model optimisation (see Section 3.3). For the case of Germany and Mexico, this dominant kriging minimum also provides the best performance (the $\times$ and $\blacklozenge$ coincide). But this is not always the case, as shown in the scatter plots of the euro area and the US. More importantly, these scatter plots reveal that the combination of NOLH sampling, MIC measurement, MCS analysis and kriging seems to be able to identify promising subspaces of the overall parameter space.

Re-running the MIC and MCS analysis on the new 175 kriging samples provides a corresponding set of realised MIC scores, which are shown in the last three columns of Table 7. These can be compared to the best NOLH samples and to the predicted MIC scores for the kriging samples, and again stars indicate data series for which the best realised kriging sample MIC score outperforms the best NOLH sample. This comparison confirms the finding mentioned above that the kriging samples offer relatively minor improvements over the original NOLH samples, and even for those that do, kriging only provides for small gains. However, one positive aspect for the overall calibration strategy developed in this paper is that the realised and predicted MIC scores are close for all the data series involved. This is confirmed in Figure 8, which presents the scatter plot of realised and predicted MIC scores for all 175 kriging samples, and establishes that they are very highly correlated. This immediately validates the 31 kriging surrogate models for the 31 data series, as they are able to reliably predict the MIC scores out-of-sample. As a result, the relative inability of kriging to improve on the initial NOLH samples does not seem to be due to the kriging procedure, but most likely stems from the lack in accuracy of the MIC as noisy measurement for the relative Kullback-Leibler distance between a model and the data.

[12] In the interest of space, the full set of scatter plots is not shown here. However, it is available as supplementary material.

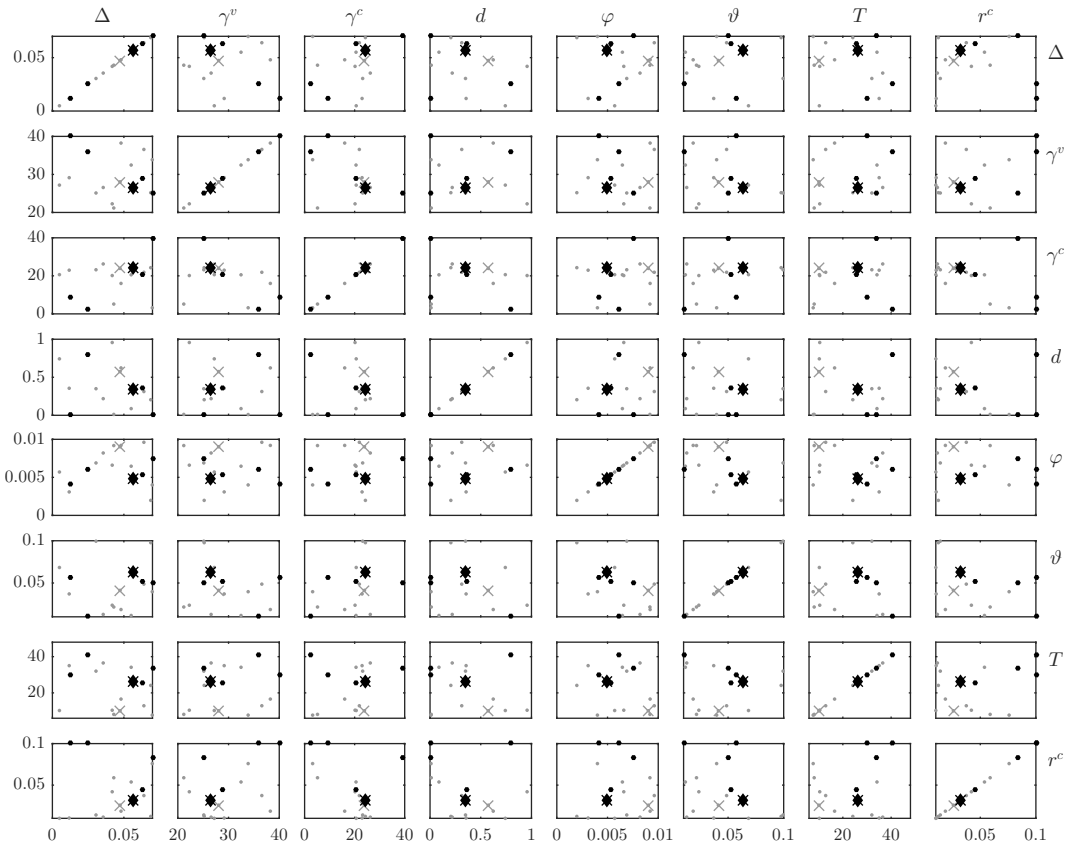



Figure 4: Germany, $|\mathcal{M}_{90}| = 12$, $|\mathcal{K}| = 5$, ♦ = 678

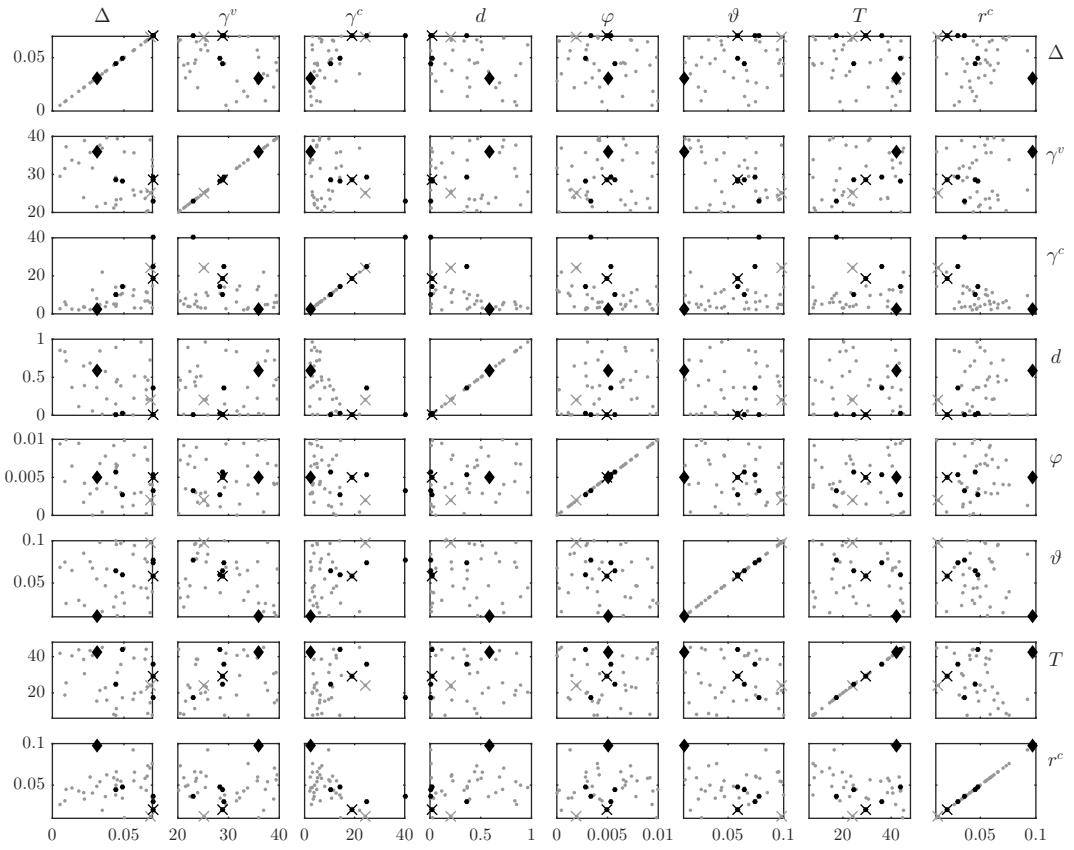



Figure 5: euro area, $|\mathcal{M}_{90}| = 34$, $|\mathcal{K}| = 6$, ♦ = 364

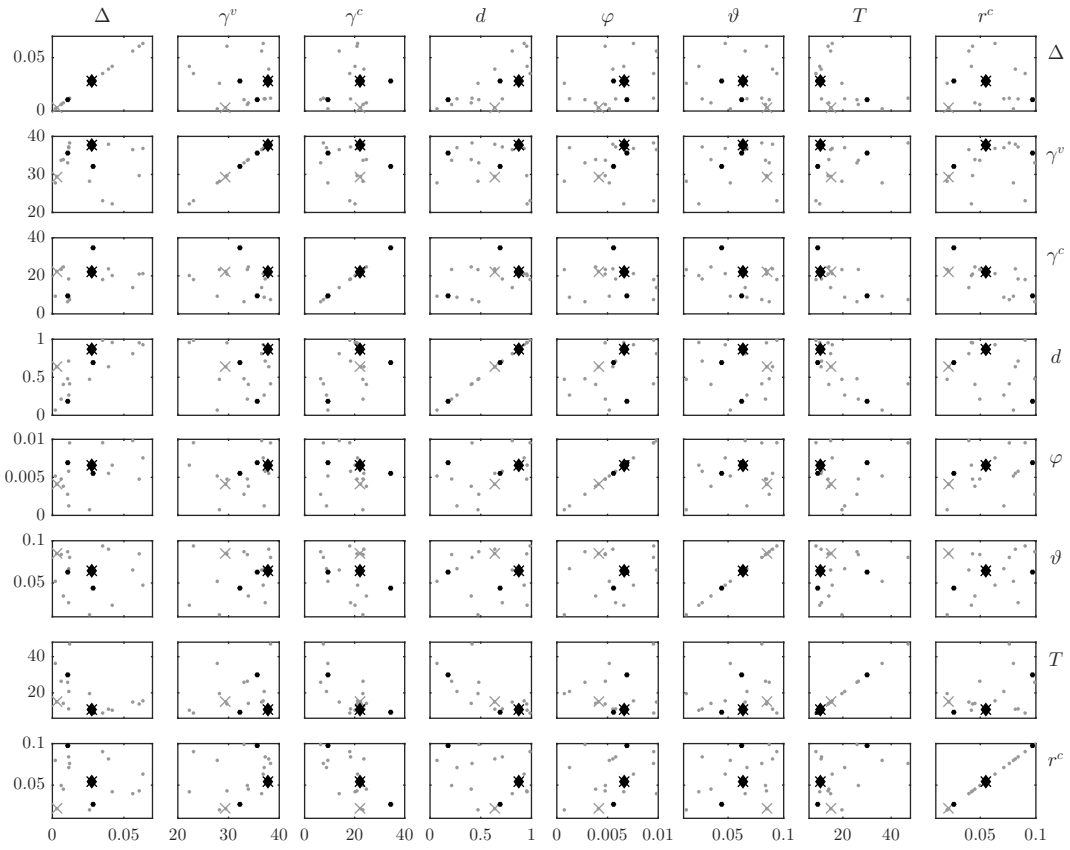



Figure 6: Mexico, $|\mathcal{M}_{90}| = 15$, $|\mathcal{K}| = 3$, ♦ = 759

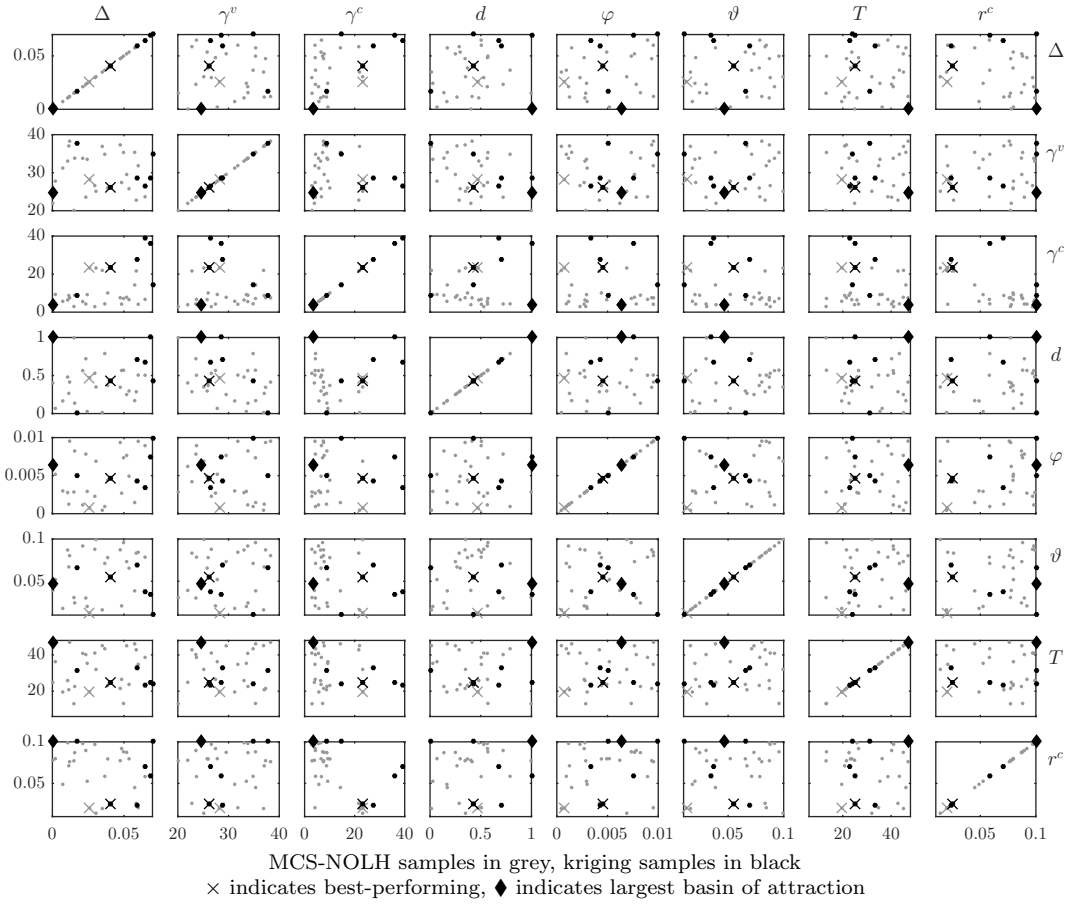



Figure 7: United States, $|\mathcal{M}_{90}| = 26$, $|\mathcal{K}| = 7$, ♦ = 696

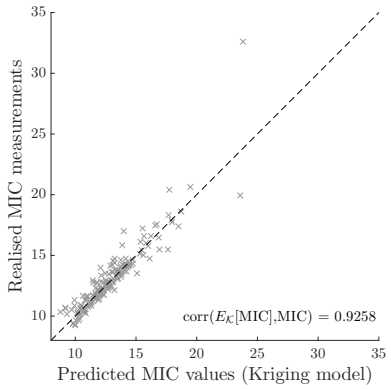



Figure 8: Predicted vs. realised MIC scores for kriging sample $\mathcal{K}$

Table 8: Parameter calibrations 227 and 266 and associated countries.

| Parameter: | $\Delta$ | $\gamma^v$ | $\gamma^c$ | $d$ | $\varphi$ | $\vartheta$ | $T$ | $r^c$ | Country list |
|---|---|---|---|---|---|---|---|---|---|
| Set 227 | 0.047 | 28.05 | 23.91 | 0.57 | 0.0057 | 0.041 | 10.10 | 0.026 | BE, Chile, CA, DE, ES FI, GR, PT, SL, UK |
| Set 266 | 0.069 | 25.20 | 24.30 | 0.21 | 0.0020 | 0.098 | 24.13 | 0.012 | euro area, FR, IT, SP, TR |

### 5.4 Economic interpretation of the results

In this section we discuss the model calibrations in the context of the Eurace@Unibi model and try to provide an economic interpretation of the results. We try to answer several questions. Do the optimal parameter calibrations make sense from an economic point of view? If we consider the results per country, can we classify the countries into groups with similar parameter constellations?

Column 3 of Table 7 identifies the best out of 513 NOLH samples that can be associated to each country. Two samples that appear quite often in the list are calibrations 227 and 266. Table 8 shows the parameter vector corresponding to these two model calibrations and the list of countries that can be associated to them, resp. Together, these two calibrations account for 15 out of 31 empirical data sets that we considered. For the euro area the best calibration is 266, while calibration 227 is associated to several other European countries.

The parameters with the most striking differences in value between the two sets appear to be $d$, $\vartheta$ and $T$. Parameter $d$ is the dividend payout ratio, which for calibration 227 is $d = 57\%$ and for calibration 266 it is lower, $d = 21\%$. Parameter $\vartheta$ is the income tax rate, which for calibration 227 is 4%, while for calibration 266 it is 10%. Finally, the parameter $T$ is the length of the debt repayment period in months. For calibration 227 this is 10 months, while for calibration 266 it is 24 months.

Summarizing, calibration 227 can be associated to high dividend ratios, low taxes, and a short debt repayment period, while calibration 266 can be associated to low dividend ratios, high taxes, and long debt repayment periods. If we would have to come up with an economic rational for why countries in the OECD can apparently be classified according to these two model calibrations, then we could hypothesize that calibration 227 represents economies in which the firms obtain short-term debt and have a higher dividend payout ratio, and these economies are characterized by the tendency to maximize shareholder value and short-termism in investment decisions. The second group of countries, that are associated to calibration 266, consists of economies in which firms obtain long-term debt and have low dividend ratios, which is an indication that firms are more prone to retain earnings and therefore this would correspond to economies with more long-term planing horizons and more patient investment decisions.

## 6 Discussion

We have shown that, in principle, it is feasible to empirically validate a large-scale, agent-based, macroeconomic simulation model, using a newly designed validation protocol. Even though the data generation step to produce the required amount of training data is computationally heavy, the actual application of the protocol after that data is in place is relatively fast (the steps that use the CWT algorithm and the MIC measurements). To avoid in-the-loop simulations with a computationally heavy simulation model we split the data generation stage from the training and testing stages. The main problem with in-the-loop simulation is not that it's impossible, it is rather that it would require a hyper-algorithm with computational steering to adjust the model's parameters in between successive iterations of the validation protocol. It is quite likely that such computational steering algorithms would have to be model-specific, since the parameter adjustments depend on how the model responds to inputs. This would require that the computational steering algorithm constructs an input-output response surface of the model, and an approximation of the local gradient and possibly also of the Hessian, followed by a gradient search on the response surface. Indeed, such hill-climbing algorithms with information about the higher-order derivatives is one of the methods adopted by current state-of-the-art machine learning techniques such as deep neural networks. Although at the moment this method does not appear computationally feasible for computationally heavy simulation models, one of us has provided some suggestions along these lines

(van der Hoog, 2016). In this paper we have circumvented the need for any in-the-loop simulations by sampling the parameter space beforehand, ensuring that the samples have sufficient coverage across the parameter space, and by using a nearly-orthogonal experimental design.

As stated in the introduction, the main purpose of this paper is to provide a proof-of-concept for a validation protocol of large-scale agent-based simulation models. An important question therefore is the assessment of the proof-of-concept and the performance of the validation protocol in general. The results outlined above suggest that the proof-of-concept with an application to the Eurace@Unibi model has been successful at identifying promising subspaces of the parameter space. The relatively tight identification across countries of the best-performing values for several parameters, such as the intensities of consumption- and vintage choices $\gamma^c$ and $\gamma^v$, the wage rate adjustment $\varphi$, and the income tax rate $\vartheta$ suggests that the protocol is effective at pinning down those parameters to which the model is most sensitive. Similarly, the tight correlation shown in Figure 8 between the MIC scores predicted by the kriging models and the realised MIC scores (post-kriging) suggest that the kriging procedure provides for an effective interpolation of the MIC response surface.

The results obtained nevertheless display some problems as well, which we should address. Many countries show relatively large MCS sizes for the initial NOLH sample, such as Hungary (63), Italy (63), France (61), the Netherlands (61). This suggests the presence of noise in the measurement which limits the ability of the validation protocol to discriminate between similar calibrations. Similarly, some parameters $(\Delta, d, T)$, are not identified as good as the ones mentioned above, which also points to limitations in the ability to identify good calibrations. These issues raise several concerns regarding the protocol as currently implemented at this proof-of-concept stage.

A first minor concern relates to the robustness of the NOLH/kriging combination advocated by Salle and Yıldızoğlu (2014) and Mandes and Winker (2016). Both seem to perform well here, as shown in Figure 8. However it is also important to verify their efficiency relative to existing alternatives. As for the sampling procedure, a major constraint of the NOLH sampling scheme is the fact that the number of samples is fixed ex-ante by the design matrix, and expanding the sampling space ex-post is not straightforward. An alternative would be to use quasi-Monte Carlo methods and Sobol sequences, which does allow additional samples to be created more easily. Although the challenge then becomes to maintain orthogonality in the resulting sample. Nevertheless, alternative sampling methods should be tested in order to ensure the robustness of the NOLH approach.

A similar case can be made for the choice of kriging as the surrogate modelling method to interpolate the MIC response surface. As pointed out by Kleijnen (2017), several alternatives to kriging already exist in the metamodelling literature, such as polynomial regression. In addition, machine learning and classifiers could also be used to generate surrogate models from training data (Mandel and Sani, 2016).

The main concern with the validation protocol, however, relates to the accuracy of the MIC measurements. As pointed out in Barde (2016a), although the MIC is an unbiased measurement of the cross-entropy between model and data, the Markov variation inherent in the training data implies that it will always contain some element of noise. For univariate processes this noise is shown to only affect the accuracy for extremely similar models, but this will not be the case in general for highly non-linear and multivariate processes such as ABMs. The 'off the shelf' approach of treating each macroeconomic variable as a univariate process implicitly assumes independence between the variable. However, because the underlying process is multivariate, the measurements obtained will not take into account the correlation between the variables and will therefore be inaccurate as a result. While the positive results obtained in this paper do support the overall design of the validation protocol, the large MCS sizes and the relative inability to pin down some of the parameters nonetheless point towards the need to expand the protocol in the near future with a dedicated multivariate implementation of the MIC.

# Annex A: Orthogonality and space-filling criteria for the NOLH sampling procedure

In Definition 6.1 we define the design matrix $\mathbf{X}$, that will later on be used to represent an efficient sampling of the parameter space.

**Definition 6.1.** Let a simulation model be specified by $n$ parameters as inputs and $q$ target variables as outputs. Imposing restrictions on the parameter ranges yields the restricted subspace $\mathcal{D} \subset \mathbb{R}^n$. The design matrix $\mathbf{X} \in \mathbb{R}^m \times \mathcal{D}$ is an $m \times n$ dimensional matrix with $m$ rows denoting the sample points and $n$ columns denoting the parameters.

The objective of an efficient sampling algorithm is to determine 'good' orthogonality and space-filling properties. The problem of finding an *sufficiently efficient* experimental design can be specified as the minimization of an objective function subject to two constraints (see Cioppa, 2002):

$$\text{Minimize} \quad \mathbf{f}(Mm, ML_2) \tag{14}$$

$$\rho_{max} \leq 0.03 \tag{15}$$

$$cond_2(\mathbf{X}^\top \mathbf{X}) \leq 1.13 \tag{16}$$

Here the objective function $\mathbf{f}$ is the rank sum of the two space-filling metrics $Mn$ and $ML_2$, and the constraints are formulated in terms of the maximum pairwise correlation $\rho_{max}$ and the condition number $cond_2$. The solution to this problem guarantees that the design matrix $\mathbf{X}$has columns (parameter vectors) that are nearly-orthogonal, and that it has good space-filling characteristics. Below we discuss these four metrics in more detail.

First, for the orthogonality property of the design matrix the criteria are that: (i) the pairwise correlation between any two columns of the design matrix $\mathbf{X}$, given by $\rho_{max} = \max\{|\rho_{ij}|\}$, should be as close to 0 as possible; (ii) the condition number of $\mathbf{X}^\top \mathbf{X}$, which is the ratio between the largest and smallest eigenvalues, should be as close to 1 as possible: $cond_2(\mathbf{X}^\top \mathbf{X}) = \frac{\Psi_1}{\Psi_2}$ where $\Psi_1 = max_i\{\lambda_i\}$ and $\Psi_2 = min_i\{\lambda_i\}$. For the $513 \times 8$ design matrix used in this paper, the largest pairwise correlation between any two parameter vectors is $\rho_{max} = 0.00202$.

Second, for the space-filling property the criteria are that: (i) the Modified $L_2$ discrepancy of all points across the entire experimental region should be as small as possible. This means that design points are close to each other and the design has better space-filling characteristics. The Modified $L_2$ metric $ML_2$ is given by:

$$ML_2 = \left(\frac{4}{3}\right)^k - \frac{2^{1-k}}{n} \sum_{d=1}^{n} \prod_{i=1}^{k} (3 - x_{di}^2) + \frac{1}{n^2} \sum_{d=1}^{n} \sum_{j=1}^{n} \prod_{i=1}^{k} [2 - \max(x_{di}, x_{ji})] \tag{17}$$

(ii) the Euclidean maxi-min distance between all $n(n-1)/2$ pairs of design points is maximized. Let the vector of pairwise distances be defined by $\mathbf{d} = (d_1, d_2, ..., d_{n(n-1)/2})$ with the elements ranked in increasing order, $d_1 \leq d_2 \leq ... \leq d_{n(n-1)/2}$. The smallest distance is $Mn \equiv d_1$ and the goal is to maximize this smallest distance, in order to guarantee that the design points are spread out as much as possible. That is, no two points should be closer to each other than the maximized smallest distance $d_1$.

## Annex B: The validation protocol in steps

In this appendix we present the validation protocol as a sequence of steps. This could be considered a pseudocode for the re-implemention of the algorithm.

1. Set bounds on the model parameters.[13]
2. Generate an efficient sample, for example by using NOLH sampling.
3. Simulate training data for each sample.
4. Apply the CTW algorithm on training data series to construct a context tree for each sample.
5. Apply the MIC algorithm to generate a MIC score for each sample, using the context tree corresponding to that sample and an empirical data series.[14]
6. Apply MCS analysis on the aggregate MIC scores to identify the subset of best-performing calibrations.[15]
7. Internal validation step 1: Test if the MIC has discriminatory power by performing an out-of-sample prediction test using 99% for training and 1% for testing.
8. Construct a surrogate model of the MIC Response Surface by a kriging procedure, and interpolate over the response surface.
9. Find local minima of the interpolated MIC Response Surface, to find new sample points that are promising candidates for better calibrations.
10. Simulate new training data for the new kriging sample points.
11. Apply the MIC algorithm again to generate actual MIC scores for the new samples.
12. Internal validation step 2: Test if the kriging method was able to correctly predict the MIC scores for the new samples.
13. Select the best samples that minimize the realized MIC scores from the original NOLH set and the kriging set, and construct the Model Confidence Set corresponding to each empirical data series, taking into account that the MIC is a noisy measurement of the cross entropy measure.

[13] The bounds on the parameters serve to specify a multi-dimensional parameter space. One sample from the parameter space is a model calibration.

[14] Multiple empirical data series can be used in this step, to make use of multiple target variables, or data for multiple countries. In our experiment, we used 3 target variables and 30 countries plus the euro area, increasing the overall computational load by a factor of 124 (3 variables and their sum makes 4 targets, multiplied by 31 makes 124 in total).

[15] The aggregate MIC score per country is computed using the individual MIC scores obtained per target variable.

# Annex C: Descriptive statistics and quantisation diagnostics

Table 9: Descriptive statistics of empirical data

| | | | Unemployment rate | | | | Industrial output growth | | | | Inflation rate | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Country | Start | N | Min | Max | 2.5 pctl | 97.5 pctl | Min | Max | 2.5 pctl | 97.5 pctl | Min | Max | 2.5 pctl | 97.5 pctl |
| Austria | Jan 1994 | 258 | 3.60 | 6.00 | 3.70 | 5.80 | −19.07 | 17.25 | −13.10 | 12.95 | −0.32 | 3.69 | 0.49 | 3.25 |
| Belgium | Jan 1984 | 378 | 6.30 | 10.90 | 6.40 | 10.80 | −17.12 | 18.27 | −10.77 | 13.09 | 0.67 | 6.58 | 0.96 | 6.00 |
| Canada | Jan 1963 | 630 | 2.90 | 13.00 | 3.60 | 11.80 | −17.35 | 19.50 | −12.08 | 13.39 | −0.36 | 12.15 | 0.61 | 11.03 |
| Chile | Dec 1999 | 187 | 5.48 | 10.47 | 5.76 | 10.35 | −18.57 | 29.65 | −9.07 | 12.28 | −2.80 | 6.99 | −1.32 | 5.73 |
| Czech Republic | Jan 1997 | 222 | 4.20 | 9.20 | 4.20 | 9.10 | −27.22 | 54.06 | −19.60 | 17.69 | −0.16 | 14.91 | 0.12 | 14.08 |
| Denmark | Jan 1984 | 378 | 3.10 | 9.90 | 3.50 | 9.30 | −22.91 | 28.16 | −15.88 | 16.03 | 0.59 | 7.01 | 0.95 | 6.28 |
| Estonia | Jan 1999 | 198 | 4.00 | 18.90 | 4.00 | 17.60 | −34.33 | 39.19 | −29.44 | 35.05 | −1.11 | 7.57 | −0.56 | 6.75 |
| euro area | Jan 1998 | 210 | 7.20 | 12.10 | 7.30 | 12.00 | −22.25 | 10.20 | −18.57 | 8.74 | 0.63 | 2.56 | 0.67 | 2.41 |
| Finland | Jan 1989 | 318 | 2.90 | 17.80 | 3.00 | 17.10 | −23.70 | 18.52 | −21.07 | 13.96 | −0.84 | 8.20 | −0.47 | 7.56 |
| France | Jan 1984 | 378 | 7.10 | 12.50 | 7.50 | 12.50 | −22.50 | 8.74 | −14.22 | 6.98 | 0.30 | 9.28 | 0.45 | 6.57 |
| Germany | Jan 1992 | 282 | 4.70 | 11.20 | 4.80 | 10.90 | −28.44 | 18.67 | −17.81 | 13.50 | 0.25 | 6.50 | 0.58 | 5.95 |
| Greece | Apr 1999 | 195 | 7.30 | 27.90 | 7.60 | 27.70 | −17.17 | 27.89 | −14.41 | 24.01 | −3.64 | 4.50 | −2.34 | 4.07 |
| Hungary | Jan 1997 | 222 | 5.50 | 11.40 | 5.50 | 11.30 | −28.21 | 24.88 | −23.37 | 22.30 | 0.57 | 20.31 | 0.86 | 18.10 |
| Iceland | Jan 2004 | 138 | 2.50 | 7.50 | 2.50 | 7.50 | −14.68 | 58.21 | −14.31 | 51.37 | 0.68 | 17.44 | 0.92 | 15.10 |
| Ireland | Jan 1984 | 378 | 3.70 | 17.10 | 3.80 | 16.80 | −17.36 | 42.61 | −9.65 | 29.92 | −6.27 | 10.03 | −4.21 | 7.09 |
| Italy | Jan 1991 | 294 | 5.80 | 13.00 | 6.20 | 12.60 | −26.53 | 11.64 | −19.90 | 8.36 | 0.39 | 6.21 | 0.70 | 5.97 |
| Japan | Jan 1957 | 702 | 1.00 | 5.50 | 1.10 | 5.30 | −34.70 | 31.78 | −13.74 | 24.39 | −1.53 | 22.46 | −1.06 | 13.01 |
| Korea | Jan 1991 | 294 | 1.90 | 8.20 | 2.00 | 7.50 | −21.97 | 36.72 | −12.63 | 28.24 | −0.70 | 9.30 | −0.35 | 8.83 |
| Luxembourg | Jan 1984 | 378 | 1.40 | 6.10 | 1.60 | 6.00 | −29.30 | 33.66 | −18.34 | 18.88 | −1.45 | 8.17 | 0.74 | 4.62 |
| Mexico | Jan 1988 | 330 | 2.10 | 7.60 | 2.30 | 6.20 | −12.96 | 13.70 | −8.62 | 11.32 | 2.22 | 181.60 | 2.46 | 97.24 |
| Netherlands | Jan 1984 | 378 | 3.10 | 9.50 | 3.20 | 9.10 | −13.62 | 15.62 | −8.55 | 9.10 | 0.32 | 4.21 | 0.58 | 3.79 |
| Norway | Jan 1990 | 306 | 2.40 | 6.80 | 2.40 | 6.70 | −10.68 | 9.14 | −7.13 | 7.69 | −0.58 | 4.03 | 0.08 | 3.80 |
| Poland | Jan 1998 | 210 | 6.80 | 20.40 | 6.90 | 20.20 | −13.51 | 24.82 | −8.17 | 19.41 | 0.25 | 15.70 | 0.38 | 14.49 |
| Portugal | Jan 1989 | 318 | 4.70 | 17.50 | 4.90 | 16.70 | −19.70 | 16.10 | −11.01 | 8.62 | −0.47 | 16.09 | 0.01 | 15.44 |
| Slovak Republic | Jan 1999 | 198 | 8.70 | 19.70 | 9.00 | 19.50 | −27.28 | 29.89 | −23.75 | 25.44 | −1.05 | 15.77 | −0.77 | 15.19 |
| Slovenia | Jan 2001 | 174 | 4.20 | 10.60 | 4.30 | 10.30 | −25.62 | 13.71 | −23.37 | 11.39 | −1.06 | 8.91 | −0.40 | 8.20 |
| Spain | Apr 1987 | 339 | 7.90 | 26.30 | 8.20 | 26.10 | −29.86 | 18.85 | −16.34 | 10.99 | −0.33 | 7.53 | −0.08 | 7.04 |
| Sweden | Jan 1984 | 378 | 1.30 | 10.50 | 1.50 | 9.90 | −23.34 | 18.10 | −19.29 | 14.57 | −1.81 | 12.80 | −1.13 | 10.02 |
| Turkey | Jan 2006 | 114 | 7.80 | 13.30 | 8.00 | 13.00 | −26.36 | 25.87 | −25.15 | 20.90 | 3.65 | 10.77 | 3.66 | 9.87 |
| United Kingdom | Jan 1984 | 378 | 4.60 | 11.30 | 4.70 | 11.20 | −12.97 | 9.14 | −9.54 | 7.09 | −0.10 | 10.85 | 0.20 | 9.36 |
| United States | Jan 1959 | 678 | 3.40 | 10.80 | 3.50 | 9.80 | −17.98 | 22.07 | −11.36 | 12.50 | 0.61 | 13.59 | 0.96 | 11.37 |

Table 10: Quantization statistics of monthly harmonised unemployment rate, 7-bit resolution.

| | $\notin$ | KS test | | LB autocorr. test | | LB crosscorr. test | |
|---|---|---|---|---|---|---|---|
| Country | $[1, 25]$ | Stat | p-value | Stat | p-value | Stat | p-value |
| Austria | 0 | 0.050 | 0.891 | 15.858 | 0.893 | 1.097 | 1.000 |
| Belgium | 0 | 0.032 | 0.990 | 31.136 | 0.150 | 0.952 | 1.000 |
| Canada | 0 | 0.044 | 0.553 | 30.357 | 0.173 | 30.479 | 0.169 |
| Chile | 0 | 0.075 | 0.655 | 25.655 | 0.371 | 34.328$^{*}$ | 0.079 |
| Czech Republic | 0 | 0.054 | 0.894 | 15.048 | 0.919 | 0.593 | 1.000 |
| Denmark | 0 | 0.032 | 0.990 | 19.777 | 0.709 | 0.313 | 1.000 |
| Estonia | 0 | 0.045 | 0.985 | 23.482 | 0.491 | 5.489 | 1.000 |
| euro area | 0 | 0.071 | 0.642 | 14.275 | 0.940 | 52.476$^{***}$ | 0.001 |
| Finland | 0 | 0.082 | 0.228 | 20.967 | 0.641 | 24.754 | 0.419 |
| France | 0 | 0.053 | 0.654 | 25.272 | 0.391 | 35.044$^{*}$ | 0.068 |
| Germany | 0 | 0.057 | 0.742 | 32.261 | 0.121 | 2.108 | 1.000 |
| Greece | 36 | 0.149$^{**}$ | 0.024 | 33.953$^{*}$ | 0.086 | 34.863$^{*}$ | 0.070 |
| Hungary | 0 | 0.036 | 0.998 | 26.805 | 0.314 | 1.956 | 1.000 |
| Iceland | 0 | 0.072 | 0.849 | 15.133 | 0.917 | 31.759 | 0.133 |
| Ireland | 0 | 0.034 | 0.977 | 30.475 | 0.170 | 0.992 | 1.000 |
| Italy | 0 | 0.034 | 0.995 | 19.564 | 0.721 | 2.135 | 1.000 |
| Japan | 2 | 0.021 | 0.997 | 19.105 | 0.746 | 3.405 | 1.000 |
| Korea | 0 | 0.037 | 0.985 | 31.348 | 0.144 | 0.304 | 1.000 |
| Luxembourg | 0 | 0.071 | 0.280 | 18.377 | 0.784 | 107.742$^{***}$ | 0.000 |
| Mexico | 0 | 0.030 | 0.998 | 33.729$^{*}$ | 0.090 | 5.567 | 1.000 |
| Netherlands | 0 | 0.037 | 0.954 | 25.715 | 0.368 | 1.039 | 1.000 |
| Norway | 0 | 0.059 | 0.652 | 23.688 | 0.480 | 51.408$^{***}$ | 0.001 |
| Poland | 0 | 0.048 | 0.968 | 27.090 | 0.300 | 3.222 | 1.000 |
| Portugal | 0 | 0.050 | 0.806 | 24.995 | 0.406 | 21.817 | 0.590 |
| Slovak Republic | 0 | 0.045 | 0.985 | 20.750 | 0.653 | 1.997 | 1.000 |
| Slovenia | 0 | 0.063 | 0.867 | 32.892 | 0.106 | 76.016$^{***}$ | 0.000 |
| Spain | 21 | 0.062 | 0.521 | 33.046 | 0.103 | 36.485$^{**}$ | 0.049 |
| Sweden | 0 | 0.066 | 0.369 | 24.366 | 0.441 | 46.361$^{***}$ | 0.004 |
| Turkey | 0 | 0.070 | 0.933 | 26.409 | 0.333 | 24.943 | 0.409 |
| United Kingdom | 0 | 0.053 | 0.654 | 24.242 | 0.448 | 28.535 | 0.238 |
| United States | 0 | 0.040 | 0.647 | 25.021 | 0.405 | 1.929 | 1.000 |

Test 1 - Kolmogorov-Smirnov test on discretization error.
$H_0$: Discretization error is uniformly distributed over $[0, 1]$.
Test 2 - Ljung-Box test on 25 lags of the discretization error
$H_0$: Discretization error is independently distributed (no autocorrelation).
Test 3 - Ljung-Box test of the discretization error against 25 lags of the discretization series
$H_0$: Discretization error is not correlated with discretized series.
'$*$' indicates significance at the 10% level, '$**$' at the 5% level and '$***$' at the 1% level.

Table 11: Quantization statistics of monthly year-on-year industrial output growth rate, 7-bit resolution.

| | $\notin$ | KS test | | LB autocorr. test | | LB crosscorr. test | |
|---|---|---|---|---|---|---|---|
| Country | $[-2, 20]$ | Stat | p-value | Stat | p-value | Stat | p-value |
| Austria | 0 | 0.039 | 0.989 | 22.062 | 0.576 | 2.822 | 1.000 |
| Belgium | 0 | 0.056 | 0.592 | 18.993 | 0.752 | 34.733* | 0.072 |
| Canada | 0 | 0.030 | 0.933 | 19.944 | 0.700 | 0.109 | 1.000 |
| Chile | 2 | 0.053 | 0.946 | 13.051 | 0.965 | 21.919 | 0.584 |
| Czech Republic | 0 | 0.072 | 0.596 | 22.941 | 0.523 | 38.372** | 0.032 |
| Denmark | 0 | 0.042 | 0.881 | 15.299 | 0.912 | 11.780 | 0.982 |
| Estonia | 0 | 0.040 | 0.996 | 27.118 | 0.299 | 35.045* | 0.068 |
| euro area | 0 | 0.048 | 0.968 | 24.401 | 0.439 | 7.660 | 0.999 |
| Finland | 0 | 0.041 | 0.949 | 60.501*** | 0.000 | 1.149 | 1.000 |
| France | 0 | 0.040 | 0.922 | 16.948 | 0.851 | 5.113 | 1.000 |
| Germany | 0 | 0.035 | 0.994 | 16.855 | 0.855 | 9.268 | 0.997 |
| Greece | 7 | 0.056 | 0.908 | 19.857 | 0.705 | 3.019 | 1.000 |
| Hungary | 1 | 0.036 | 0.998 | 45.794*** | 0.005 | 13.150 | 0.964 |
| Iceland | 0 | 0.065 | 0.922 | 24.840 | 0.415 | 2.045 | 1.000 |
| Ireland | 14 | 0.093* | 0.074 | 90.870*** | 0.000 | 145.022*** | 0.000 |
| Italy | 0 | 0.037 | 0.985 | 18.675 | 0.769 | 17.572 | 0.823 |
| Japan | 8 | 0.023 | 0.993 | 24.730 | 0.421 | 1.616 | 1.000 |
| Korea | 0 | 0.048 | 0.885 | 25.052 | 0.403 | 4.191 | 1.000 |
| Luxembourg | 0 | 0.029 | 0.997 | 16.734 | 0.860 | 6.292 | 1.000 |
| Mexico | 67 | 0.121** | 0.014 | 25.781 | 0.364 | 7.127 | 1.000 |
| Netherlands | 0 | 0.034 | 0.977 | 17.361 | 0.833 | 17.928 | 0.806 |
| Norway | 0 | 0.062 | 0.584 | 14.902 | 0.924 | 4.159 | 1.000 |
| Poland | 0 | 0.048 | 0.968 | 18.606 | 0.773 | 2.342 | 1.000 |
| Portugal | 0 | 0.038 | 0.975 | 28.016 | 0.259 | 4.778 | 1.000 |
| Slovak Republic | 0 | 0.051 | 0.958 | 26.252 | 0.340 | 33.861* | 0.087 |
| Slovenia | 0 | 0.075 | 0.700 | 43.501*** | 0.009 | 54.695*** | 0.000 |
| Spain | 0 | 0.035 | 0.982 | 27.349 | 0.288 | 19.131 | 0.745 |
| Sweden | 0 | 0.037 | 0.954 | 25.841 | 0.361 | 31.813 | 0.132 |
| Turkey | 0 | 0.096 | 0.642 | 14.546 | 0.933 | 1.398 | 1.000 |
| United Kingdom | 0 | 0.040 | 0.922 | 13.492 | 0.957 | 1.976 | 1.000 |
| United States | 0 | 0.025 | 0.982 | 34.707* | 0.073 | 28.158 | 0.253 |

Test 1 - Kolmogorov-Smirnov test on discretization error.
$H_0$: Discretization error is uniformly distributed over $[0, 1]$.
Test 2 - Ljung-Box test on 25 lags of the discretization error
$H_0$: Discretization error is independently distributed (no autocorrelation).
Test 3 - Ljung-Box test of the discretization error against 25 lags of the discretization series
$H_0$: Discretization error is not correlated with discretized series.
'$*$' indicates significance at the 10% level, '$**$' at the 5% level and '$***$' at the 1% level.

Table 12: Quantization statistics of monthly year-on-year CPI inflation rate, 7-bit resolution.

| | $\notin$ | KS test | | LB autocorr. test | | LB crosscorr. test | |
|---|---|---|---|---|---|---|---|
| Country | $[-2, 20]$ | Stat | p-value | Stat | p-value | Stat | p-value |
| Austria | 0 | 0.039 | 0.989 | 22.088 | 0.631 | 3.040 | 1.000 |
| Belgium | 0 | 0.056 | 0.592 | 19.152 | 0.790 | 36.047* | 0.071 |
| Canada | 0 | 0.030 | 0.933 | 19.981 | 0.748 | 0.109 | 1.000 |
| Chile | 2 | 0.053 | 0.946 | 13.051 | 0.976 | 22.524 | 0.605 |
| Czech Republic | 0 | 0.072 | 0.596 | 23.410 | 0.554 | 39.922** | 0.030 |
| Denmark | 0 | 0.042 | 0.881 | 15.836 | 0.920 | 12.218 | 0.985 |
| Estonia | 0 | 0.040 | 0.996 | 27.217 | 0.345 | 36.469* | 0.065 |
| euro area | 0 | 0.048 | 0.968 | 24.463 | 0.493 | 7.770 | 1.000 |
| Finland | 0 | 0.041 | 0.949 | 68.839*** | 0.000 | 1.249 | 1.000 |
| France | 0 | 0.040 | 0.922 | 16.960 | 0.883 | 5.189 | 1.000 |
| Germany | 0 | 0.035 | 0.994 | 18.018 | 0.842 | 9.450 | 0.998 |
| Greece | 7 | 0.056 | 0.908 | 20.093 | 0.742 | 3.209 | 1.000 |
| Hungary | 1 | 0.036 | 0.998 | 45.985*** | 0.006 | 13.404 | 0.971 |
| Iceland | 0 | 0.065 | 0.922 | 25.497 | 0.435 | 2.452 | 1.000 |
| Ireland | 14 | 0.093* | 0.074 | 95.827*** | 0.000 | 148.423*** | 0.000 |
| Italy | 0 | 0.037 | 0.985 | 19.113 | 0.792 | 18.273 | 0.831 |
| Japan | 8 | 0.023 | 0.993 | 24.748 | 0.477 | 1.751 | 1.000 |
| Korea | 0 | 0.048 | 0.885 | 33.304 | 0.124 | 4.199 | 1.000 |
| Luxembourg | 0 | 0.029 | 0.997 | 21.682 | 0.654 | 6.568 | 1.000 |
| Mexico | 67 | 0.121** | 0.014 | 25.975 | 0.409 | 7.286 | 1.000 |
| Netherlands | 0 | 0.034 | 0.977 | 17.560 | 0.860 | 19.527 | 0.771 |
| Norway | 0 | 0.062 | 0.584 | 15.747 | 0.922 | 4.161 | 1.000 |
| Poland | 0 | 0.048 | 0.968 | 18.620 | 0.815 | 2.363 | 1.000 |
| Portugal | 0 | 0.038 | 0.975 | 28.277 | 0.295 | 4.896 | 1.000 |
| Slovak Republic | 0 | 0.051 | 0.958 | 26.300 | 0.392 | 34.680* | 0.094 |
| Slovenia | 0 | 0.075 | 0.700 | 44.166** | 0.010 | 55.977*** | 0.000 |
| Spain | 0 | 0.035 | 0.982 | 27.955 | 0.310 | 20.129 | 0.740 |
| Sweden | 0 | 0.037 | 0.954 | 25.876 | 0.414 | 32.972 | 0.132 |
| Turkey | 0 | 0.096 | 0.642 | 14.789 | 0.946 | 1.421 | 1.000 |
| United Kingdom | 0 | 0.040 | 0.922 | 13.714 | 0.966 | 2.007 | 1.000 |
| United States | 0 | 0.025 | 0.982 | 38.525** | 0.041 | 30.495 | 0.206 |

Test 1 - Kolmogorov-Smirnov test on discretization error.
$H_0$: Discretization error is uniformly distributed over $[0, 1]$.
Test 2 - Ljung-Box test on 25 lags of the discretization error
$H_0$: Discretization error is independently distributed (no autocorrelation).
Test 3 - Ljung-Box test of the discretization error against 25 lags of the discretization series
$H_0$: Discretization error is not correlated with discretized series.
'$*$' indicates significance at the 10% level, '$**$' at the 5% level and '$***$' at the 1% level.

## Annex D: Eurace@Unibi parameter list

Table 13: Full list of parameters in the Eurace@Unibi model. Parameters with ranges indicated in column 4 have been used in this paper for the empirical validation exercise.

| Symbol | Description | Default | Range |
|---|---|---|---|
| **Household sector** | | | |
| $\vartheta$ | Income tax rate | 0.05 | $[0.01, 0.10]$ |
| $u$ | Unemployment benefit percentage | 0.70 | |
| $\kappa$ | Marginal propensity to save | 0.1 | |
| $\Phi$ | Target wealth/income ratio | 16.67 | |
| $T^h$ | Mean individual income periods | 6 | |
| **Consumption goods** | | | |
| $\gamma^c$ | Logit parameter for consumption choice | 12 | $[0, 40]$ |
| $\chi$ | Service level for the expected demand | 0.8 | |
| $\rho$ | Discount rate of forecast profit flows | 0.02 | |
| **Investment goods** | | | |
| $p_0^v$ | Initial capital price | 20 | |
| $\mathbb{P}$[Innovation] | Probability of successful innovation | 0.025 | |
| $\Delta$ | Slope of technological frontier | 0.025 | $[0, 0.07]$ |
| $\lambda$ | Bargaining power of the capital goods producer | 0.5 | |
| $\gamma^v$ | Logit parameter for vintage choice | 30.0 | $[20, 40]$ |
| $\delta$ | Capital depreciation rate | 0.01 | |
| **Credit market** | | | |
| $T$ | Debt repayment period | 18 | $[6, 48]$ |
| $\varphi^{debt}$ | Debt rescaling factor | 0.30 | |
| $r^c$ | Central Bank policy rate | 0.05 | $[0.01, 0.10]$ |
| $\underline{e} = \overline{e}$ | Markdown and markup on $r^c$ for interest rates[16] | 0.10 | |
| $\lambda^B$ | Weight of default probability in interest rate rule | 3 | |
| $\alpha$ | Max. risk-based leverage ratio | 10 | |
| $\alpha^{-1}$ | Min. Capital Adequacy Requirement (CAR) | 0.10 | |
| $\beta$ | Min. Reserve Ratio Requirement (RRR) | 0.10 | |
| **Financial market** | | | |
| $d$ | Dividend payout ratio | 0.70 | $[0, 1]$ |
| $\overline{m}$ | Threshold to pay full dividends (firms) | 0.5 | |
| $\lambda^{ix}$ | Parameter price adjustment rule | 1.0 | |
| $\underline{c} = \overline{c}$ | Limit on down/upward price changes of risky asset | 0.10 | |
| **Labour market** | | | |
| $\varphi^{base}$ | Adjustment rate of base wage offer | 0.01 | $[0, 0.01]$ |
| $\psi$ | Adjustment rate of reservation wage | 0.01 | |
| $\eta^{month}$ | Applications per month | 5 | |
| $\eta^{day}$ | Applications per day | 3 | |
| $[\underline{\varrho}, \overline{\varrho}]$ | Uniform distr. for random dismissals | $[0, 0.10]$ | |
| $\gamma^{gen}$ | Logit parameter applicant selection (general skills) | 0.5 | |
| $C^{comm}$ | Fixed commuting costs | 1.0 | |

# Acknowledgements

**FLAME:** The simulations for this paper were performed using the simulation framework FLAME (Flexible Large-scale Agent Modelling Environment, Coakley et al., 2012.). This work contains information using the FLAME Xparser and Libmboard library, which are made available under the Lesser General Public License (LGPL v3), and can be downloaded from:

`https://github.com/FLAME-HPC/`.

**R Project:** This paper has made use of software provided by the R Project (R Development Core Team, 2008).

**Source code of the Eurace@Unibi Model:** The results in this paper make use of the source code of the Eurace@Unibi Model and of data analysis scripts written in R, specifically developed for the purpose of post-processing data from agent-based models (Gemkow and van der Hoog, 2012). The exact version of the model that was used for this paper is called: Financial Fragility Network Model (version 1.0). The source can be downloaded from:

`https://pub.uni-bielefeld.de/data/2908396`

**CHEOPS:** We gratefully acknowledge the Universität zu Köln, for providing us with computational resources on the CHEOPS high-performance computing cluster, that allowed us to generate the large amounts of training data that were needed to perform the analyses reported in this paper.

**Matlab ooDACE toolbox:** This paper has made use of the Matlab ooDACE toolbox provided by Couckuyt et al. (2014, 2012). This software package can be downloaded from:

`http://www.sumo.intec.ugent.be/ooDACE`.

# Supplementary Material

## Availability of source code and data

The data publication Barde and van der Hoog (2017) contains source code (C code and Matlab scripts), as well as the training data generated from the Eurace@Unibi model that is required to reproduce our analysis.

## Additional plots

In Section 5 we have shown the scatter plots only for 4 empirical data series. The scatter plots for all 31 data series are included in the data publication as well.

# References


Arifovic, J., Bullard, J., Kostyshyna, O., 2013. Social learning and monetary policy rules. Economic Journal 123, 38–76.

Assenza, T., Delli Gatti, D., 2013. E pluribus unum: Macroeconomic modelling for multi-agent economies. Journal of Economic Dynamics and Control 37, 1659 – 1682.

Barde, S., 2016a. A practical, accurate, information criterion for Nth order Markov processes. Computational Economics, 1–44.

Barde, S., 2016b. Direct comparison of agent-based models of herding in financial markets. Journal of Economic Dynamics and Control 73, 329–353.

Barde, S., van der Hoog, S., April 2017. Data for the paper: An empirical validation protocol for large-scale agent-based models. Data Publication Bielefeld University, doi:10.4119/unibi/2908396.

Bargigli, L., Riccetti, L., Russo, A., Gallegati, M., 2016. Network calibration and metamodeling of a financial accelerator agent based model. Working Papers in Economics No. 01/2016, DISEI - Università degli Studi di Firenze.

Box, G. E. P., Wilson, K., 1951. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society 13 (1), 1–45.

Chen, C.-H., Chang, C.-L., Du, Y.-R., 2014. Agent-based economic models and econometrics. Knowledge Engineering Review 27 (2), 187–219.

Cioppa, T. M., 2002. Efficient nearly orthogonal and space-filling experimental designs for high-dimensional complex models. Doctoral Dissertation.

Cioppa, T. M., Lucas, T. W., 2007. Efficient nearly orthogonal and space-filling latin hypercubes. Technometrics 49, 45–55.

Coakley, S., Chin, L.-S., Holcomb, M., Greenough, C., Worth, D., 2012. Flexible Large-scale Agent Modelling Environment (FLAME). University of Sheffield and Rutherford Appleton Laboratories, STFC, License: Lesser GPL v3. `https://github.com/FLAME-HPC/`.

Couckuyt, I., Dhaene, T., Demeester, P., 2014. ooDACE Toolbox: A Flexible Object-Oriented Kriging Implementation. Journal of Machine Learning Research 15, 3183–3186.

Couckuyt, I., Forrester, A., Gorissen, D., Turck, F. D., Dhaene, T., 2012. Blind Kriging: Implementation and performance analysis. Advances in Engineering Software 49, 1–13.

Dawid, H., Gemkow, S., Harting, P., van der Hoog, S., Neugart, M., 2017. Agent-Based Macroeconomic Modeling and Policy Analysis: The Eurace@Unibi Model. In: Chen, S.-H., M., K. (Eds.), Handbook of Computational Economics and Finance. Oxford University Press.

Dawid, H., Harting, P., van der Hoog, S., Neugart, M., 2016. A Heterogeneous Agent Macroeconomic Model for Policy Evaluation: Improving Transparency and Reproducibility. Bielefeld Working Papers in Economics and Managment No. 06-2016.

De Grauwe, P., Macchiarelli, C., 2015. Animal spirits and credit cycles. Journal of Economic Dynamics and Control 59, 95–117.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 134–144.

Dosi, G., Fagiolo, G., M., N., Roventini, A., Treibich, T., 2015. Fiscal and monetary policies in complex evolving economies. Journal of Economic Dynamics and Control 52, 166–189.

Dosi, G., Fagiolo, G., Roventini, A., 2010. Schumpeter meeting Keynes: A policy-friendly model of endogenous growth and business cycles. Journal of Economic Dynamics and Control 34, 1748–1767.

Dosi, G., Pereira, M., Virgillito, M., 2016. On the robustness of the fat-tailed distribution of firm growth rates: a global sensitivity analysis. Journal of Economic Interaction & Coordination z, xx–yy.

Elias, P., 1975. Universal codeword sets and representations of the integers. IEEE Transactions on Information Theory IT-21, 194–203.

Epstein, J. M., Axtell, R., 1994. Agent-based models: Understanding our creations. SFI Bulletin.

Gemkow, S., van der Hoog, S., 2012. R Data Analysis Scripts. Bielefeld University, Bielefeld, Unpublished. License: GPL v3.

Gilli, M., Winker, P., 2001. Indirect Estimation of the Parameters of Agent Based Models of Financial Markets. FAME Research Paper Series rp38, International Center for Financial Asset Management and Engineering.

Gilli, M., Winker, P., 2003. A global optimization heuristic for estimating agent based models. Computational Statistics & Data Analysis 42 (3), 299–312.

Grazzini, J., Richiardi, M., 2014. Estimation of ergodic agent-based models by simulated minimum distance. Economics Papers 2014-W07, Economics Group, Nuffield College, University of Oxford.

Guerini, M., Moneta, A., Dec. 2016. A Method for Agent-Based Models Validation. LEM Papers Series 2016/16, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy.

Halas, M., 2011. Abductive reasoning as the logic of agent-based modelling. Proceedings of the 25th European Conference on Modelling and Simulation. European Council for Modelling and Simulation.

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The model confidence set. Econometrica 79, 453–497.

Harman, G., 1965. The inference to the best explanation. Philosophical Review, 88–95.

van der Hoog, S., 2016. Deep Learning in agent-based models: A prospectus. Working Papers in Economics and Management, No. 02-2016. Faculty of Business Administration and Economics, Bielefeld University.

Kleijnen, J. P., 2017. Regression and kriging metamodels with their experimental designs in simulation: A review. European Journal of Operational Research 256 (1), 1 – 16.

Kleijnen, J. P. C., 2007. Design and Analysis of Simulation Experiments, 1st Edition. Springer.

Krige, D., 1951. A statistical approach to some mine valuations and allied problems at the witwatersrand. Master's thesis, University of Witwatersrand.

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Lipton, P., 2004. Inference to the Best Explanation, 2nd Edition. Routledge, Abingdon.

Lophaven, S. N., Nielsen, H. B., Søndergaard, J., 2002. DACE, A Matlab Kriging Toolbox. Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby.

Lucas, R., 1976. Econometric policy evaluation: A critique. Carnegie-Rochester Conference Series on Public Policy 1 (1), 19–46.

Lux, T., Zwinkels, R. C. J., March 2 2017. Empirical validation of agent-based models. Tech. rep., Christian Albrechts Universität zu Kiel.

Mandel, A., Jaeger, C., Fuerst, S., Lass, W., Lincke, D., Meissner, F., Pablo-Marti, F., Wolf, S., 2010. Agent-based dynamics in disaggregated growth models. CES Working Paper.

Mandel, A., Sani, A., 2016. Learning Time-Varying Forecast Combinations. Documents de travail du Centre d'Economie de la Sorbonne 16036, Universit Panthon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne.

Mandes, A., Winker, P., 2016. Complexity and model comparison in agent based modeling of financial markets. Journal of Economic Interaction & Coordination x (y), z–z, doi:10.1007/s11403-016-0173-0.

McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105 – 142.

McFadden, D., 1980. Econometric models for probabilistic choice among products. Journal of Business 53 (3), S13–29.

Peirce, C. S., 1997. Pragmatism as a Principle and Method of Right Thinking. The 1903 Harvard Lectures on Pragmatism. SUNY Press, Albany, New York.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

Riccetti, L., Russo, A., Gallegati, M., 2015. An agent based decentralized matching macroeconomic model. Journal of Economic Interaction and Coordination 10 (2), 305–332.

Richiardi, M., 2015. The future of agent-based modelling. Economics Papers 2015-W06, Economics Group, Nuffield College, University of Oxford.

Rissanen, J., 1986. Complexity of strings in the class of markov sources. IEEE Transactions on Information Theory IT-32, 526–532.

Salle, I., Yıldızoğlu, M., 2014. Efficient sampling and meta-modeling for computational economic models. Computational Economics 44 (4), 507–536.

Wagner, C., Flatken, M., Meinel, M., Gerndt, A., Hagen, H., 2010. FSSteering: A Distributed Framework for Computational Steering in a Script-based CFD Simulation Environment. Working paper.

Wieland, V., Cwik, T., Mller, G. J., Schmidt, S., Wolters, M., 2012. A new comparative approach to macroeconomic modeling and policy analysis. Journal of Economic Behavior & Organization 83 (3), 523–541.

Willems, F. M. J., Shtarkov, Y. M., Tjalkens, T. J., 1995. The context-tree weighting method: Basic properties. IEEE Transactions on Information Theory IT-41, 653–664.

Windrum, P., Fagiolo, G., Moneta, A., 2007. Empirical validation of agent-based models: Alternatives and prospects. Journal of Artificial Societies and Social Simulation 10 (2), 8.