INSTITUTE
OF ECONOMICS

Scuola Superiore
Sant'Anna

LEM | Laboratory of Economics and Management

Institute of Economics
Scuola Superiore Sant'Anna

Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy
ph. +39 050 88.33.43
institute.economics@sssup.it

LEM
WORKING PAPER SERIES

From ABM back to real data: time series visualization and model selection in the K+S agent-based model

Giovanni Dosi [a]
Marcelo C. Pereira [b,a]
Gabriel Petrini [b]
Andrea Roventini [a]
Maria Enrica Virgillito [a]

[a] Institute of Economics, Scuola Superiore Sant'Anna, Pisa, Italy
[b] University of Campinas, Brazil

2025/17 **April 2025**
ISSN(ONLINE): 2284-0400
DOI: 10.57838/sssa/7aav-nk67

From ABM back to real data: time series visualization and model selection in the K+S agent-based model*

Giovanni Dosi[1], Marcelo C. Pereira[2, 1], Gabriel Petrini[†2, 1], Andrea Roventini[1], and Maria Enrica Virgillito[1]

[1]Institute of Economics, Scuola Superiore Sant'Anna, Pisa, Italy
[2]University of Campinas, Brazil

April 17, 2025

Abstract

Agent-Based Models (ABMs) provide powerful tools for economic analysis, capturing micro-to-macro interactions and emergent properties. However, integration with empirical data has been a persistent challenge. To address it, we propose a protocol for integration between empirical data and ABM, building a new multidimensional similarity index that aggregates different similarity measures into a composite score, specifically designed to quantify alignment between simulated and real-world data. This metric enables a complete model ranking procedure, facilitating a streamlined model selection. The protocol is designed to be model-agnostic and flexible, allowing its application to a wide range of models beyond ABMs, including aggregate dynamical systems and any type of computational model. As an example, we apply our methodology to different configurations and model versions of the Schumpeter meeting Keynes (K+S) ABM family (Dosi, Fagiolo, and Roventini, 2010) using US data (from 1948Q1 to 2019Q1). Next, we propose a policy-informed application, attributing different weights to variables associated with policy-making decisions and technological change. The exercise is done in order to showcase the capacity of the procedure to target specific policy variables of interest, allowing for the design of empirically informed scenario analyses and projections on real-world dynamics.

*This work has been written for the forthcoming *Resilient Complex Adaptive Systems: Modeling, Simulation, Visualization and Engineering Approaches*, edited by Claudia Szabo. The authors wish to thank participants at the "Agent-based models for political economy" workshop, Pisa (2025), for their helpful comments and suggestions. Gabriel Petrini gratefully acknowledges the financial support in the form of a research grant from the Brazilian National Council for scientific and Technological Development (CNPq) under Grant 140721/2020-7 and from the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) under Grant 88887.716528/2022-00.

[†]Corresponding author. Email address: gpetrinidasilveira@gmail.com

Keywords: Agent-Based Models, Model selection, Validation, Similarity measurement
JEL classification: C63, C52, C18

1 Introduction

Understanding and evaluating how much a model is similar/dissimilar *vis-à-vis* the empirical evidence is certainly a challenge for model development. The latter is particularly relevant for the complex system approach to socio-economic phenomena (Dosi, 2023), capable of integrating micro-meso-macro dynamics and analyzing the evolution of aggregates as a result of interactions and structures occurring at lower levels of aggregation. Within the complex system approach, Agent-Based Models (ABMs) stand out as a growingly important technique to study evolving economies. One of the aims of building an agent-based model is to understand mechanisms and processes that guide the real world dynamics and they often represent thought experiments or metaphorical models (Bouchaud, 2023), that is simulated complex maps that can be used to build scenario analyses and test for the relevance of parameters/interactions, providing diagnosis. The incipient May (1972)'s model of population ecology, where high ecological diversity leads to unstable rather than stable outcomes, serves as a classic example of a metaphorical model. Another common usage of ABMs is comprehending the mechanisms of a social phenomenon, as Epstein (2007) referred to as Generative Social Science.

Such metaphorical or diagnostic models are not primarily designed to replicate any specific empirical dataset but to investigate how the proposed laws of motion yield coherent qualitative results. With this virtual laboratory, the modeler can generate scenarios and data that are typically unobservable or unavailable for empirical studies. ABMs applied to economic phenomena are medium-/large-scale simulated models that cover micro, industry, and macro aggregate variables, providing a powerful tool for scenario analysis and policy experiments (Dosi, Fagiolo, and Roventini, 2010; Dawid, Gemkow, Harting, and Neugart, 2012; Dawid, Gemkow, Harting, van der Hoog, et al., 2018; Cincotti et al., 2010; Delli Gatti et al., 2010, to name a few). One of the substantial advancements of ABMs in economic modeling is the ability to create simulated environments for diagnostic exercises, focusing on qualitative empirical patterns. Overall, agent-based models are designed to generate stylized facts resulting from the interactions of numerous disaggregated units connected by various forms of structure and interactions, allowing the emergence of aggregated properties, which are tested against robust qualitative patterns.

While ABMs have proven helpful in capturing emergent properties and complex interactions, their integration with empirical data remains an open issue. The relationship between empirical data and models in economics generally goes in the realm of four different domains, which are estimation, calibration, validation, and selection. While the first two problems mainly pertain to

parameter definitions, validation, and selection pertain instead to model mechanisms, which depend on the inner structure of interactions and the structure of variables.

The literature has addressed such issues in the last years, distinguishing between input and output validation and highlighting the importance of external validity (Fagiolo, Moneta, et al., 2007; Lamperti, 2018; Marks, 2007; Windrum et al., 2007; Fagiolo, Guerini, et al., 2019a). External validity is a dimension that agent-based models particularly require, because of their attempt to be empirically disciplined models (Dosi and Roventini, 2019). Recently, there has been growing interest in applying ABMs to forecasting exercises (Poledna et al., 2023) and to inform specific policy decisions (Delli Gatti and Reissl, 2022), stressing the shift towards the quantitative dimension.[1] Overall, this indicates a rich research venue for data-friendly ABMs.

However, this involves closely examining and utilizing data. Still today there is a lack of a comprehensive approach able to confront with the inner complexity of such models. Reconciling simulated models with real-world data is non-trivial, especially in socio-economic systems impacted by interacting processes and environmental changes like crises, catching-ups, economic development, and stagnation. Developers of socio-economic models encounter numerous challenges, starting with determining the appropriate data selection methods, including aggregation levels, historical periods, and types of series to be analyzed. In addition, there is still a lack of a protocol able to synthetically rank models in their performance ability to be close to the data. Metrics of distance between simulated and real time series, such as simulated minimum distance, root mean squared errors, or correlation coefficients, fail to capture the complete range of information and structure encoded in the data. Indeed, unidimensional distance measures, such as Euclidean distance, may fail to capture the global model dynamics and typical attributes of time series produced by ABMs, such as persistent deviations from average behavior, path dependency, and structural breaks. In addition, information regarding the signal domain of the series is not usually embraced by unidimensional distance measures.

Signal processing and the frequency domain of economic variables are crucial information that should be embraced when comparing data with a time-series structure. Economic time series are, in fact, characterized by some distinctive features. Indeed, a common characteristic of this data type is a smoothly decreasing spectral shape, which indicates the significance of low-frequency phenomena, such as long-term fluctuations (Granger, 1966). Long waves, for example, have been associated with low-frequency events in the literature (Kondratieff and Stolper, 1935), particularly in the context of technological change (Clark et al., 1981; Silverberg, 2007) and structural change (Dosi, Pereira, et al., 2022). In this specific context, for example, large-scale adoption of new technological paradigms can also be driven by short-term euphoria (Perez, 2009). Neglecting the frequency domain would overlook the distinctions associated with short versus medium versus long

[1] It is important to note that this quantitative dimension should not be seen as an alternative test of model validity but rather as a complementary step to the diagnostic capabilities of the models.

run cycles. Indeed, a similarity measure of time series should be able to embrace the time spectra, including time intervals, lag structures, and misalignment. There is still a lack of a single indicator able to account for different dimensions, therefore the quest for a multidimensional approach.

Responding to the increasing demand for data-friendly ABMs and the challenges of integrating data back into the models, we propose a step to complement existing techniques in the literature. This chapter introduces an alternative approach to similarity measurement, focusing on model selection through a composite similarity index aggregating multiple dimensions of comparison. The primary purpose of this approach is to provide a systematic method for model selection, which can assist other methods in integrating data back into models.

The contribution of our approach lies in constructing a composite similarity index that aggregates multiple statistical and dynamic measures into a single ranking system. Instead of relying on a single indicator, we combine several metrics that capture different aspects of similarity, ensuring a more detailed assessment. The ranking system is designed to iteratively refine model selection by penalizing those models that achieve the highest aggregate dissimilarity score. This iterative selection process ensures a continuous refinement of model selection, enabling a structured analysis that can assist in integrating data back into models.

We apply the protocol to different versions of the Schumpeter meeting Keynes Agent-Based Models (Dosi, Fagiolo, and Roventini, 2010; Dosi, Fagiolo, Napoletano, and Roventini, 2013; Dosi, Fagiolo, Napoletano, Roventini, and Treibich, 2015; Dosi, Pereira, et al., 2017; Dosi, Pereira, et al., 2020; Dosi, Pereira, et al., 2022). We use US data from 1948Q1 to 2019Q1 for our empirical dataset throughout the analysis. Our protocol is designed to be easily adaptable to different empirical cases and sets of simulated models, with the flexibility to apply to models beyond ABMs.

This protocol consists of two main phases. During the first phase, we systematically analyze the empirical-simulated pair for each variable across different K+S model versions. Specifically, we conduct a battery of tests to dynamically compare time series dissimilarities, taking into account different alignments in the time spectra. This comparison examines statistical moments, different frequencies, and paths of the signal as a whole. We obtain a variable-wise multidimensional index by applying this procedure to every variable across all models under scrutiny. In the next phase, we use the information from the multidimensional index to perform an iterative model selection. In this way, we advance the analysis of time series similarity and model selection.

We finally present an extension to this protocol by assigning different weights (*ex-ante*) to fine-tune the model selection for specific purposes. We apply the experiment to variables typically associated with policymaking decisions and with technological change-related variables. The analysis reveals that the finance-augmented K+S model demonstrates robust performance compared to recent developments, particularly when examining different frequencies. Likewise, the labor-augmented model, adjusted for competitive institutional settings, yields satisfactory results across most configurations.

The remainder of this chapter is structured as follows. Section 2 briefly introduces the key aspects of the K+S models. The information about the structure of the models grounds the rationale behind pairing decisions between the simulated and empirical data. Section 3 describes the data selection used to build the empirical dataset for comparison with the simulated models. Section 4 outlines the first phase of the protocol, which consists of creating a comprehensive evaluation of model similarity to empirical data. Next, Section 5 presents the model selection routine to identify the least dissimilar model. This section also presents a purpose-guided model selection to evaluate models' performance with a subset of target variables. Finally, Section 6 concludes the chapter and offers insights into other potential uses of this protocol and possible improvements.

2 K+S family model

This section briefly presents the evaluated models and their underlying structural assumptions. These assumptions will guide the data selection described in Section 3. Table 1 lists the different model versions. The K+S model (Dosi, Fagiolo, and Roventini, 2006; Dosi, Fagiolo, and Roventini, 2008; Dosi, Fagiolo, and Roventini, 2010) is a macroeconomic agent-based model where Schumpeterian and Keynesian mechanisms influence business cycle fluctuations and long-run economic growth. The original version encompasses two industries, each populated by heterogeneous firms. The households, banking, and government sectors are stylized in a parsimonious way and refined in other versions. In this artificial economy, all agents make decisions using bounded-rational rules and simple heuristics. This model was designed on purpose to capture a rich ensemble of micro-macroeconomic regularities (stylized facts). As a consequence, this model (and its later extensions) was not intended to precisely replicate the empirical time series of any specific economy or time period. Therefore, some degree of dissimilarity with empirical data is expected.

Main reference	Contribution	Label
Dosi, Fagiolo, and Roventini (2010)	First version of the K+S model	`K+S-original`
Dosi, Fagiolo, Napoletano, and Roventini (2013)	Detailed financial sector	`K+S-finance`
Dosi, Pereira, et al. (2018b) and Dosi, Pereira, et al. (2018a)	Decentralized labor market and institutional growth regimes	[`K+S-Ford`, `K+S-Comp`]
Dosi, Pereira, et al. (2022)	Disruptive technological change and multiple sectors	`K+S-multi`

Table 1: Collection of model versions ($\mathcal{V}$) considered

Before comparing the models with empirical data, it is important to understand their general structure. The models feature two distinct sectors. The first sector consists of capital goods firms that produce heterogeneous, durable machine tools using only labor. These firms operate on a made-to-order basis, producing capital goods in response to the investment plans of the other industry, underscoring the interdependence between them. The labor productivity growth rate is

endogenous and evolves according to the level of R&D expenditures, solely comprised of labor. Thus, this sector is the primary locus of endogenous innovation embedded in capital goods of different vintages. Consumption-good firms employ a stock of machine tools whose technologies evolve over time. Consumption goods are demanded by households and can be stored if unsold. Both industries have access to bank loans (up to a limit) and operate under conditions of imperfect information. The capital goods sector, in particular, faces Schumpeterian competition. Firms in both sectors set prices based on a mark-up rule over production costs.

The model includes a single bank, representing a stylized financial sector. At each time step, the bank collects deposits from the private sector and offers interest-bearing loans to finance firms' production and investment activities. The bank does not employ any workers for its operations, and only firms have access to credit. If a firm is credit-constrained, it cannot carry out its production or investment decisions as planned, requiring further revisions. The central bank sets the policy rate, completing the financial circuit of the model. The prime rate anchors deposit and loan rates, which are determined using mark-down and mark-up rules.

The household sector is also stylized and populated by homogeneous workers facing a centralized labor market, receiving a homogeneous wage if employed. The government sets both the tax rate and unemployment benefits. In this setup, all firms pay the same wage, and unemployed households receive uniform unemployment benefits. The government is responsible for setting and collecting taxes, issuing bonds to finance the public deficit, and maintaining a sustainable public debt trajectory over the long term. Notably, the government does not provide goods or services to the private sector. Like the banking sector, the government does not employ any workers.

Despite its stylized representation, the original K+S model successfully captures a wide range of micro- and macroeconomic regularities. Table 2, adapted from Dosi, Pereira, et al. (2022), lists the stylized facts (SFs) replicated by this model. At the macroeconomic level, the `K+S-original` model generates endogenous, self-sustained growth with persistent fluctuations. It also produces fat-tailed growth rate distributions, cross-correlations among macroeconomic variables, and cyclicality in GDP components, among other features. At the firm level, the model exhibits lumpy investment patterns and persistent productivity heterogeneity across firms. The table also highlights the SFs associated with major extensions of the original model. Most of these extensions build on the core principles outlined above, but are tailored to address specific research questions.

The first major modification to the model was introduced by Dosi, Fagiolo, Napoletano, and Roventini (2013), who added a more complex banking sector. This extension aimed to explore the relationship between finance and economic growth. Subsequently, Dosi, Fagiolo, Napoletano, Roventini, and Treibich (2015) introduced bank heterogeneity. In this version, banks decide how much credit to extend to firms and prioritize borrowers based on a pecking-order list determined by the firms' liquidity ratios. The amount of credit granted also depends on the bank's net worth, subject to Basel-type regulatory capital adequacy constraints. These refinements provide a more

Microeconomic stylized facts	Macroeconomic stylized facts
Skewed firm size distribution[a]	Endogenous self-sustained growth with persistent fluctuations and crises[a]
Fat-tailed firm growth rates distribution[a]	Fat-tailed GDP growth rate distribution[a]
Heterogeneous productivity across firms[a]	Endogenous volatility of GDP, consumption and investment[a]
Persistent productivity differentials[a]	Cross-correlation of macro variables[a]
Lumpy investment rates of firms[a]	Pro-cyclical aggregate R&D investment and net entry of firms in the market[a]
Firm bankruptcies are counter-cyclical[b]	Cross-correlations of credit-related variables[b]
Firm bad-debt distribution fits a power-law[b]	Cross-correlation between firm debt and loan losses[b]
	Banking crises duration is right skewed[b]
	Fiscal costs of banking crises to GDP distribution is fat-tailed[b]
Heterogeneous skills distribution[c]	Persistent and counter-cyclical unemployment[c]
Fat-tailed unemployment time distribution[c]	Endogenous volatility of productivity, unemployment, vacancy, separation and hiring rates[c]
Fat-tailed wage growth rates distribution[c]	Unemployment and inequality correlation[c]
Cross-sectional Engel Law[d]	Pro-cyclical workers skills accumulation[c]
Heterogeneous propensity to save and consume[d]	Beveridge curve[c]
	Okun curve[c]
	Wage curve[c]
	Matching function[c]
	Engel Law[d]
	Non-satiation in luxury goods[d]
Technology-level stylized facts[d]	**Sectoral-level stylized facts[d]**
Stepwise increase in technological frontier	Product life-cycle
Lower rate of radical versus incremental innovation	Exponential age distribution
Fast diffusion of dominant techniques	Sectoral wage and productivity differentials

[a] Since Dosi, Fagiolo, and Roventini (2006) and Dosi, Fagiolo, and Roventini (2010).
[b] Since Dosi, Fagiolo, Napoletano, and Roventini (2013) and Dosi, Fagiolo, Napoletano, Roventini, and Treibich (2015).
[c] Since Dosi, Pereira, et al. (2017), Dosi, Pereira, et al. (2018a), and Dosi, Pereira, et al. (2020).
[d] Since Dosi, Pereira, et al. (2022).

Table 2: Stylized Facts Replicated by the K+S family models

detailed representation of banks' balance sheets and lending channels. This enables the model to reproduce systemic financial phenomena, particularly those associated with banking crises and their macroeconomic consequences (Reinhart and Rogoff, 2009). Notably, the model demonstrates that banking crises can engender fiscal burdens for the government (Laeven and Valencia, 2008). The incorporation of diverse transmission channels enriches the model's descriptive capacity and expands the scope for policy experiments, enabling revisiting the implications of both fiscal and monetary rules (Amendola and Pereira, 2024)

Later, the centralized labor market was replaced with a decentralized one, inaugurating the labor-augmented versions (Dosi and Roventini, 2019; Dosi, Pereira, et al., 2017; Dosi, Pereira, et al., 2020; Dosi, Pereira, et al., 2018b; Dosi, Pereira, et al., 2018a). In these versions, the economy is

populated by workers with heterogeneous skills. This means labor market conditions now influence productivity growth. When the decentralized labor market opens, firms and workers engage in a search-and-matching process. In addition to imperfect information, this market has other forms of imperfect competition. For example, larger firms receive more job applications than smaller ones. Unlike in earlier versions, the assumption of a single nominal wage no longer holds. The evolution of the labor market will be contingent upon the institutional arrangements within the economy.

Dosi, Pereira, et al. (2017) examine a setting that resembles a Fordist labor market in the spirit of Bouchaud (2023)'s metaphorical model. In this regime, wage determination is insensitive to the tightness of the labor market, while the productivity gains are passed on to wages. On the behavioral side, firms and workers maintain longer-lasting contractual relationships. This includes conditional firing schemes and less frequent job switching. While unemployed, households receive government unemployment benefits provided by a welfare state, which also guarantees a minimum wage. The authors also evaluate a "competitive regime" that reflects post-1980s changes. Unlike the Fordist regime, wages respond more rapidly and markedly to changes in unemployment levels. Firms have greater flexibility to fire workers in response to production plans, facilitating reductions in excess workforce. Both minimum wage and its indexation to productivity are no longer guaranteed. Workers also exhibit increased job-seeking behavior, even while employed. Regardless of the institutional setting, the labor-augment versions can replicate an additional collection of stylized facts, also listed in Table 2. These include wage, Beveridge, and Okun curves, unemployment, and vacancy rate volatility. This chapter applies the proposed protocol to both regimes, labeling the first as `K+S-Ford` and the second as `K+S-Comp`.[2]

While earlier model versions featured only two sectors, each with a single industry, Dosi, Pereira, et al. (2022) introduced a multi-sector version (`K+S-multi`). The key contribution is the introduction of disruptive technological change and the inclusion of multiple consumption goods sectors. One sector represents basic non-durable consumption goods, as in previous versions. The second simulates a luxury goods sector, produced through multiple stages. This extension, combined with a decentralized labor market, enables the exploration of relationships between income distribution, consumption patterns, and the impact of technological paradigms on job creation and displacement. These advancements enable the reproduction of a new batch of stylized facts, such as Engel's law, non-satiation in luxury goods, heterogeneous propensity to save, paradigm changes, and product life-cycle.

One of the advantages of this family of macroeconomic ABMs is its flexibility in generating different metaphorical economic systems. While each model has been empirically validated by comparing its outputs with real-world regularities, the extent to which they resemble/mirror empirical

[2] In the original papers (Dosi, Pereira, et al., 2018b; Dosi, Pereira, et al., 2017), the authors examine the effect of an institutional change during the simulations. Here, we compare the two regimes, ignoring the institutional transition phase.

data remains an open question. This chapter aims to address this question tentatively. We simulate each model version discussed above using their benchmark configurations for 600 time steps without changing the source code. All simulations were conducted using the Laboratory for Simulation Development (Valente and Pereira, 2023). The results were integrated into R using the LSDinterface library (Pereira, 2022). The first 350 periods are excluded from the analysis to ensure that the effects of institutional changes, which occur during simulations in some model versions, do not skew the results. The following section outlines the process of selecting empirical data, guided by the theoretical framework of the models presented here.

3 Data selection

The first step of the analysis consists of the collection of data, which will be later contrasted with the simulated models presented in Section 2. The analysis is restricted to the variables common to the most stylized model version presented in Dosi, Fagiolo, and Roventini (2010), due to the different levels of disaggregation among the model versions. As those models are calibrated for quarterly data, we downsample the empirical time series with higher frequency (*e.g.* monthly) by aggregating the inter-quarter data using the mean, a method chosen for its simplicity and ability to preserve the overall trend of the series. When available, we make use of seasonally adjusted series. For each variable considered, we estimate the trend (`TREND`) and the cycle (`CYCLE`) for both empirical and simulated series using the Christiano and Fitzgerald (2003) filter to the series in levels. This enables us to evaluate whether the models under consideration are similar to empirical data in other frequency domains alongside series without filtering (`Unfiltered`).

A critical aspect of the analysis is developing a mapping procedure to associate each simulated series with its corresponding empirical counterpart. The mapping strategy needs to be robust to what we refer to as *ex-ante dissimilarity*. This concept refers to the necessary degree of divergence between simulated and empirical data that ensures the model does not artificially reproduce the observed series due to structural simplifications or missing real-world complexities.[3] We illustrate the issue of *ex-ante* dissimilarity with an example. Suppose a model produces a simulated GDP series identical to the observed one. Given that the model lacks crucial real-world elements such as the external sector or more granularity, this perfect match suggests that the model might be similar through artifacts rather than correctly reproducing the economic dynamics of the series. In this case, a slight dissimilarity might indicate a more realistic representation of the core economic mechanisms, given how stylized the models are. Still, determining the tolerance level of dissimilarity is not straightforward.

One way to reduce the occurrence of spurious matches is by selecting time series that have

[3] Although this concept is not documented in the literature, it illustrates the challenges of comparing simulated models with real data, which is not restricted to the K+S family or ABMs.

a similar structure to the simulated ones, that is, selecting the empirical series whose generative mechanisms are included in the model. Therefore, for this exercise, the theoretical aspects of the model and its underlying assumptions serve as the selection criteria — even though it implies using shorter series — so the empirical data possess a structure most akin to the model. This increases the likelihood that models closely resembling the data are the best candidates for accurately representing a higher similarity. However, it is important to note that controlling for *ex-ante* dissimilarity is particularly challenging for GDP series, as simply aggregating demand subcomponents consistent with the model does not fully capture the indirect effects necessary to replicate GDP dynamics coherently. Since this is out of the scope of this chapter, we chose to select this variable in real terms (GDPC1) without further modification.

Table 3 shows the empirical-simulated mapping for the remaining evaluated variables.[4] We use US data from the FRED database and, where possible, we select time series spanning from 1948Q01 to 2018Q04. The decision to restrict the dataset before 2019 removes any potential effects of the COVID-19 crisis — a phenomenon not intended to be captured by the models under scrutiny.

As Section 2 describes, capital-good firms produce durable machine tools, which are subsequently supplied to firms producing consumer goods. The aggregate production of these firms is mapped to empirical data on durable industrial goods (IPG333S), reflecting their role in producing long-lasting capital equipment. The real series is obtained using the implicit deflator of industrial equipments (Y033RD3Q086SBEA). Additionally, we are repurposing this series to compare with the producer price inflation rate. Furthermore, firms in this sector also invest in R&D (Y006RC1Q027SBEA), allocating a portion of their workforce to innovative activities, which must be considered. Specifically for the R&D series, we use the empirical R&D series as it is, deflating it by the intellectual products implicit deflator (Y001RD3Q086SBEA), which lacks an equivalent in the model. This particular implicit deflator is not utilized further.

Firms producing consumer goods combine capital and labor to produce goods consumed by households. In the model, the consumption decisions exhibit characteristics akin to non-durable personal consumption expenditures (PCND), as purchasing such goods in one period does not affect decisions in the following periods. In obtaining the real series, we used the implicit price deflator for non-durable PCE goods (DNDGRD3Q086SBEA) instead of the implicit GDP deflator and repurposed this implicit deflator for the basic goods inflation rate. Because the production of consumer goods requires industrial equipments, it can be mapped to the production of manufactured non-durable (and non-energy) goods (IPB51210S). Likewise, we are using the observed capacity utilization rate for the non-durable goods sector (CAPUTLGMFNS) to contrast it with the simulated series. As for aggregate investment decisions, we focus on industrial equipment (A680RC1Q027SBEA), the closest sub-component to capital goods in the simulated models. This approach, mirroring that used for the capital goods sector, employs the implicit price deflator for

[4] We delay the explanation of its last column to Section 4.

industrial equipment at the finest level of disaggregation available. Additionally, the pairing basic goods/non-durable goods is used to compare labor productivity in the consumption goods sector by contrasting the simulated series with labor productivity in the manufacturing sector (OPHMFG). Although this decision has certain limitations, it is primarily driven by data availability.

For government expenditures, federal non-defense government consumption (A542RC1Q027SBEA) is used for comparison with the corresponding model variable, as public investment and defense expenditures are excluded from the models. The real government consumption series is derived by applying the federal non-defense implicit deflator (A825RD3Q086SBEA); however, since no equivalent exists in the models, it is not used further. Additionally, total federal debt as a share of GDP (GFDEGDQ188S) is directly contrasted with the model counterpart.

Labor-market variables are examined by comparing model-generated data with the general unemployment rate. For real wages, hourly compensation for all workers in the non-farm business sector (COMPNFB) is used and deflated using the GDP implicit deflator (GDPDEF). Although the model presents specific wages and unemployment by sector, we employ aggregate variables because the weight of the labor market in the capital-goods sector is very tiny. All data selected and corresponding variables in the model are listed in Table 3.

Private inventories are not considered due to the high level of aggregation of empirical series. Similarly, variables related to the financial sector are omitted primarily due to data availability constraints, which limit establishing a credit/debt relationship in the analysis. Future stages of this comparative exercise may incorporate additional data sources to address these limitations.

4 Multidimensional similarity index

4.1 Motivation and data pre-processing

This section presents a systematic approach to capturing various aspects of similarity between a set of simulated models and their empirical counterparts. Our point of departure is that an analysis limited to only one aspect of similarity — as the Euclidean distance, or statistical moments — fails to properly capture all information embedded into signals. For this reason, we propose a multidimensional composite index of similarity. The key principle guiding the development of this protocol is to create a versatile and scalable procedure that can be applied to different pairs of objects across the time domain, regardless of whether they originate from agent-based models. This multidimensional approach is necessary because of the inherent complexity of time series data and the absence of a single measurement capable of capturing all relevant aspects of similarity at once.

In order to motivate the importance of building a multidimensional similarity index, we start with a counter-example, showing the case of a unidimensional similarity metric and the weakness of such an approach. Figure 1 presents the trajectories of two time series. Suppose that the solid gray one corresponds to the simulated series produced by a computational model, while the solid

FRED Code	Empirical series	Simulated variables	Transformation
GDPC1	Real GDP	`GDP`	Growth rate
PCND	Nominal personal non-durable goods expenditure	`ScBas`	Growth rate
DNDGRD3Q086SBEA	Non-durable goods implicit price index	`CPIbas`	Growth rate
A542RC1Q027SBEA	Nominal federal non-defense government expenditure	`Gnom`	Growth rate
A680RC1Q027SBEA	Private investment on industrial equipment	`Inom`	Growth rate
Y033RD3Q086SBEA	Private equipment implicit price index	`PPI`	Growth rate
Y006RC1Q027SBEA	R&D	`RDnom`	Growth rate
Y001RD3Q086SBEA	Intellectual property products implicit price index	-	-
PRS32006093	Labor productivity	`A`	Growth rate
UNRATE	General Unemployment Rate	`U`	Divided by 100
IPG333S	Industrial production of machinery	`Q1`	Growth rate
IPB51210S	Non-durable and non-energy industrial production	`QcBas`	Growth rate
CAPUTLGMFNS	Capacity utilization rate in the non-durable manufacturing sector	`QcUbas`	Divided by 100
GFDEGDQ188S	Total public debt as percent of GDP	`DebGDP`	Divided by 100

Table 3: Empirical time series and corresponding equivalents in the models. Real series are estimated using the deflators indicated in the text.

black line is the empirical counterpart that this series was designed to represent. One immediate way to measure how similar these two series are is to compute the point-wise Euclidean distance between them, indicated by the dashed lines. A naive interpretation of this metric might suggest that the model lacks closeness to empirical data because the two series are substantially distant. This conclusion changes when we reveal that the data-generating processes of the two series are $\sin(t)$ and $2\cos(t)$, implying that although the data-generating processes are different, the two series might overlap by means of a time shift transformation. This simple example highlights the consequences of disregarding time shifts when measuring similarity.

Time shift is not the only drawback of the use of a Euclidean distance. Because computational models can generate as many data points as desired, series lengths may differ. Additionally, models may produce series with mismatches in scale and frequency. Labeling a model as similar/dissimilar without properly addressing these issues might lead to wrongly ranking model distance *vis-à-vis* the data structure.

Series with similar temporal structures may be deemed dissimilar solely due to differences in scale. Typically, non-stationary series may initially share a similar scale but quickly diverge. While taking the first difference is a standard method to transform a non-stationary series into a stationary one, it does not address the scaling issue. To mitigate this issue, we calculate growth rates when comparing two non-stationary series, thereby minimizing the problem of mismatched scales. The

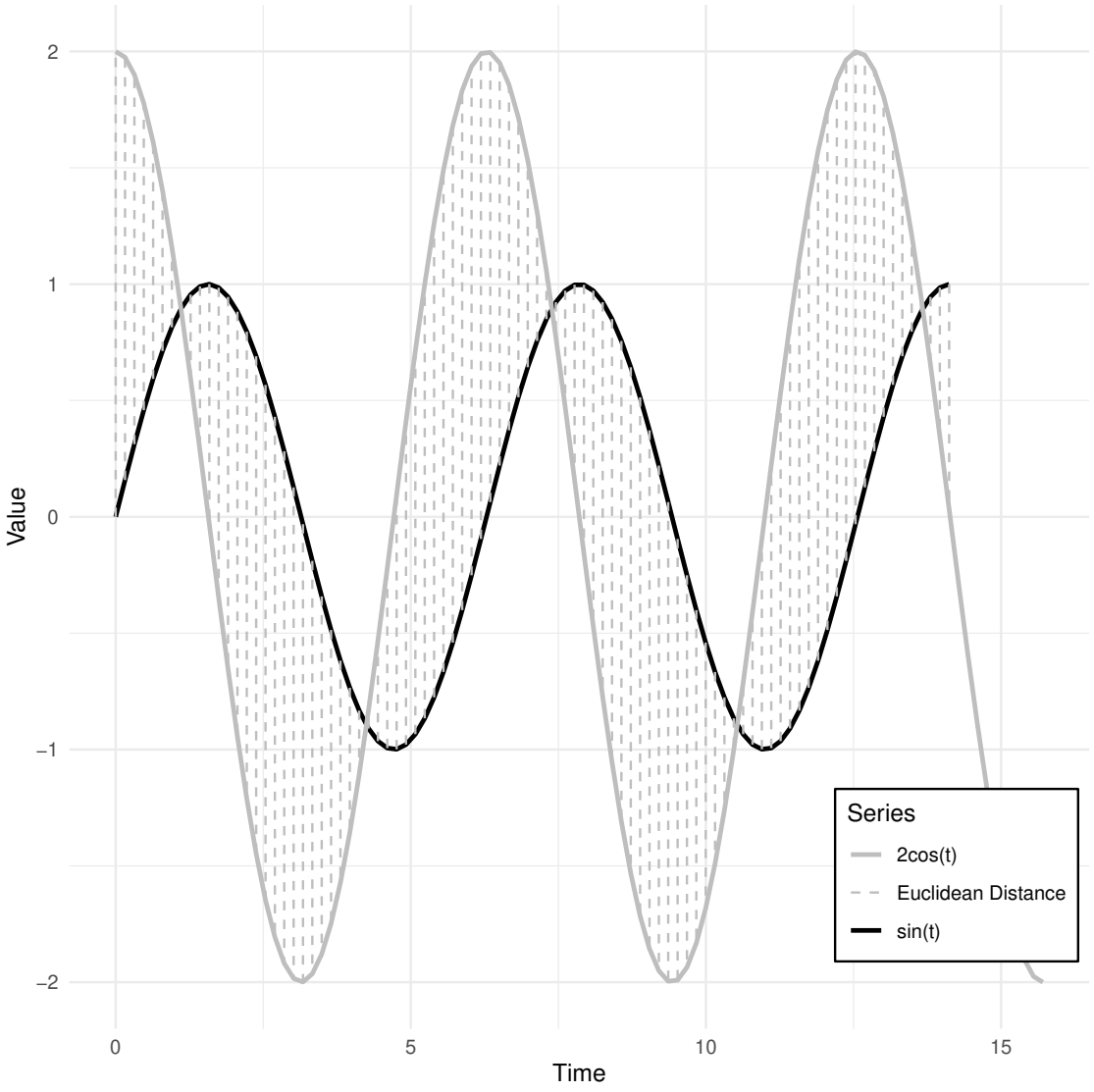


Figure 1: Example of measuring similarity of two generic and misaligned time series

last column of Table 3 denotes whether the series is being considered in terms of growth rates. Whenever further normalization is necessary, we utilize the z-score recommended by Paparrizos and Gravano (2015)[5]. This normalization is applied to the cyclical components of the series while computing growth rates to control scale mismatches in trends if the original series is non-stationary.

Another source of incomparability is length mismatches. Although some similarity measurements can handle differing lengths, others cannot. A common approach is to interpolate the shorter series to match the length of the longer one or to pad zeros to the shorter series. However, this requires another layer of decisions, such as the interpolation method, which can introduce artifacts into the signals. Instead, we truncate the longer series by removing excess time steps to ensure that

[5]See Keogh and Kasetty (2003) for a discussion about the necessity of normalization before measuring the distances.

both series have the same length. This procedure is not harmful in the case of ergodic series. The next decision concerns the placement of truncation. We opt for a centered downsampling, ensuring that an equal number of observations are dropped from the beginning and end of the series.

To provide an intuitive understanding of our protocol, Figure 2 visually illustrates the concept of multidimensional similarity. We rely on multiple indices to measure different aspects of similarity between time series, as indicated by the inputs on the left side of the figure. Each index will be thoroughly explained in the following subsections, and we will revisit the motivation example when appropriate. Each measurement is computed for all models for a single variable. This process is repeated for each variable across different model versions represented by the inputs on the right side. This results in a univariate multidimensional similarity index, which will later be used to develop a ranking system for model selection. For the sake of clarity, we defer the details of the computation of our multidimensional index to section 4.7.

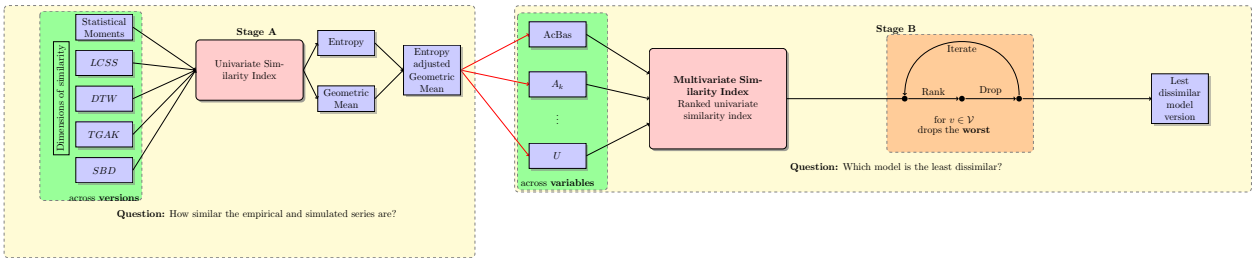


Figure 2: Flowchart representation of the protocol

We now introduce a general notation to ensure broad applicability across different model classes. Let $\mathbf{x}$ represent the empirical time series vector (reference) with a length of m. The models discussed in Section 2, denoted as $\mathcal{V}$, will be considered. A generic model version is indexed by v, enabling us to represent the simulated (query) time series vector for model v as $\mathbf{y}_v$, with a time-length of n data points. It is important to note that the reference/query pair may not have the same length, with $m \ll n$ being a typical case. Some similarity measures require all possible trajectory alignments between two series, denoted as $\mathcal{A}(m, n)$, while $\pi \in \mathcal{A}(m, n)$ represents one possible trajectory that belongs to this space. Specific alignments have a special property referred to as $\pi^\star$, while some measurements impose restrictions on the alignment space, denoted as ω. Using $\mathcal{M}$ to represent a generic index, we define the normalized distance as $\mathcal{D}_\mathcal{M}$, and the corresponding similarity as $\mathcal{S}_\mathcal{M} = 1 - \mathcal{D}_\mathcal{M}$ when the respective distance measurement is a metric and limited to closed range.

In the following subsections, we describe the individual measurements that compose the multidimensional similarity index. The index shows how closely different K+S model versions resemble their empirical counterparts. All similarity and distance measurements are applied to trend, cyclical, and unfiltered components of the series. When presenting each measurement, we use the capacity utilization rate as a variable of reference, as this series does not require any further transformation.

4.2 Statistical moments

The first measurement, likely one of the most intuitive, evaluates the difference between selected summary statistics of the simulated and empirical series. This index aims to measure dissimilarity in terms of the probability distribution of the series. We consider the mean, standard deviation, skewness, kurtosis, and the number of structural breaks. This selection is guided by the significance of these moments in capturing key patterns in time series data, particularly for economic series.

Each moment provides distinct insights into the distributional characteristics of the data. The mean μ conveys both the central tendency and the scale of the distribution. The standard deviation (SD) measures volatility, a fundamental aspect of business cycles (Stock and Watson, 1999; Napoletano et al., 2006, and references therein). Skewness ($Skew$) and kurtosis ($Kurt$) further describe the shape of the distribution. Skewness measures asymmetry and is relevant for analyzing the magnitude and duration of fluctuations in economic time series (McKay and Reis, 2008; Reinhart and Rogoff, 2009; Dosi, 2007). Kurtosis evaluates tailedness and the presence of fat tails, which are empirically observed features of output growth rates (Fagiolo, Napoletano, et al., 2008; Bottazzi and Secchi, 2006; Bottazzi, Li, et al., 2019; Laeven and Valencia, 2008), for instance. Finally, the inclusion of structural breaks (SB) accounts for sudden changes in both simulated and empirical series using the Bai and Perron (2003) method.

Let $\mu_{\mathbf{x}}$ denote the mean of empirical series and $\mu_{\mathbf{y}_v}$, denote the mean of the simulated series v. The relative deviation of the mean for model v, denoted as ($\text{rel}_{\text{mean},v}$) is computed as follows:

$$\text{rel}_{\text{mean},v} = \left| \frac{\mu_{\mathbf{y}_v} - \mu_{\mathbf{x}}}{\mu_{\mathbf{x}}} \right| \cdot \gamma \qquad \forall v \in \mathcal{V}$$

Normalization using the empirical mean mitigates scale mismatches. This procedure is applied consistently to standard deviation, skewness, and kurtosis, each normalized by its empirical counterpart. Specifically, for skewness and kurtosis, we introduce a penalty of $\gamma = 2$ when their sign differs from the empirical series and set $\gamma = 1$ otherwise. The computation of the distance in terms of the number of structural breaks is slightly different, as shown below:

$$\text{rel}_{\text{SB},v} = \left| \frac{SB_{\mathbf{y}_v} - SB_{\mathbf{x}}}{\max(1, SB_{\mathbf{x}})} \right| \cdot \gamma \qquad \forall v \in \mathcal{V}$$

The modification in the denominator ensures that the metric remains well-defined even when the empirical series has no structural breaks.

After computing the relative deviations for the selected moments, we take their geometric mean to form a composite index. This measure indicates how much a given simulated series deviates from the empirical counterpart. A value closer to zero suggests higher similarity. Note that this index is non-metric, ranging from $[0, \infty)$ meaning there are no strict boundaries. Thus, we define only the

distance measure $\mathcal{D}_{\text{statistical}}$ without an absolute similarity scale.

Table 4 presents the computed statistical moments for the capacity utilization rate and their relative deviations from the empirical series (indicated in bold). This allows for a direct assessment of how different K+S model versions replicate key distributional properties. Examining each moment individually, we observe that the original version (`K+S-original`) performs best in terms of mean and skewness, aligning more closely with the empirical values, while the multi-sector model (`K+S-multi`) exhibits the highest deviation in both moments. Besides the good relative performance in terms of the mean, the original version does not perform well in terms of standard deviation, displaced by the finance-augmented (`K+S-finance`). Regarding structural breaks, most models display similar occurrences, except for `K+S-Comp`, which presents a lower frequency. In terms of skewness, the multi-sector version (`K+S-multi`) stands out for having an opposite sign compared to the empirical series, whereas the Fordist version (`K+S-Ford`) shows a similar discrepancy in kurtosis. Aggregating these aspects — using the geometric mean —, the statistical moments similarity index identifies the original version as the closest to the empirical data. This suggests that, among the given alternatives, it better captures the distributional attributes of the observed capacity utilization series.

	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	Empirical
Mean	0.9887	0.7099	0.9839	0.3972	0.7028	**0.8061**
Mean diff	0.2265	0.1193	0.2205	0.5072	0.1282	**0.0000**
SD	0.0002	0.0130	0.0016	0.0107	0.0040	**0.0416**
SD diff	0.9953	0.6879	0.9610	0.7414	0.9047	**0.0000**
Skewness	-0.1934	-0.0516	-0.8938	0.6245	-0.4619	**-0.4109**
Pen. Skewness diff	0.5292	0.8743	1.1753	5.0396	0.1241	**0.0000**
Kurtosis	-0.3428	-0.8460	0.2072	-0.3846	-0.4024	**-0.5809**
Pen. Kurtosis diff	0.4099	0.4565	2.7134	0.3379	0.3072	**0.0000**
Breaks	2.0000	5.0000	3.0000	4.0000	4.0000	**4.0000**
Breaks diff	0.5000	0.2500	0.2500	0.0000	0.0000	**0.0000**
Statistical Moments Similarity Index	0.5125	0.4517	0.8908	0.8421	0.2586	**0.0000**

Table 4: Statistical moments similarity index contrasting simulated and empirical capacity utilization series across different K+S model versions

4.3 Longest Common Subsequence (LCSS)

Time series data are characterized by their dynamic nature, often representing evolving entities that do not necessarily progress at the same pace or share the same length. As a consequence, an appropriate metric is needed to measure the similarity of the trajectories between the two series. Wagner and Fischer (1974) introduced a method later called edit-based distance, which measures the cost of transforming one sequence into another through a series of insertions, deletions, or substitutions. This approach has been widely employed in diverse fields, including writing recognition and biolog-

ical pattern identification (Gruber et al., 2010; Shyu and Tsai, 2009, for example). Based on this method, Vlachos et al. (2002) proposed the Longest Common Subsequence (LCSS), a modification of the edit-based framework explicitly tailored for time series analysis.

The Longest Common Subsequence (LCSS) Distance is computed by first constructing a distance matrix that classifies pairs of points from time series $\mathbf{x}$ and $\mathbf{y}_v$ as either matching or non-matching. A pair (x_i, y_j) is considered a match if its Euclidean distance is smaller than or equal to a given threshold ϵ; in this case, their distance is set to 0; otherwise, it is set to 1. Using this matrix, a dynamic programming (DP) algorithm iteratively builds a cost matrix $C(i, j)$, where each entry represents the length of the longest common subsequence found up to that point. The cost matrix is updated using the recurrence relation:

$$C(i,j) = \begin{cases} C(i-1, j-1) + 1, & \text{if } |x_i - y_j| \leq \epsilon \quad \text{(match)} \\ \max(C(i-1,j), C(i,j-1)), & \text{otherwise (gap)} \end{cases}$$

If x_i and y_j match, the LCSS count increases based on previous matches; otherwise, the algorithm propagates the maximum value from the adjacent entries, effectively allowing gaps without imposing an explicit penalty. This flexibility enables LCSS to handle time shifts and noise by permitting unmatched regions while focusing on the longest sequence of similar values. Another notable feature of this index is the ability to accommodate cases where the length of the series diverges ($m \neq n$). The final entry of the cost matrix, $C(m, n)$, contains the length of the LCSS distance measurement between the two series:

$$LCSS = C(m,n) \tag{1}$$

By definition, a larger value of $LCSS$ indicates a greater degree of similarity, implying that the two sequences share a longer contiguous subsequence within the given tolerance ϵ. To ensure comparability across time series of different lengths, the similarity index is normalized by the length of the shorter series:

$$\mathcal{S}_{LCSS} = \frac{LCSS(\mathbf{x}, \mathbf{y}_v, \epsilon)}{\min(m,n)}$$

which allows us to compute the *normalized* LCSS distance measurement as:

$$\mathcal{D}_{LCSS} = 1 - \mathcal{S}_{LCSS}$$

Figure 3 revisits the prototypical example and presents the length of LCSS under different ϵ values. The occurrence in the matching condition is indicated with gray lines. This example illustrates the relevance of the choice of ϵ in determining the sensitivity of the LCSS measure: a larger ϵ increases the likelihood of matching between observations, thereby producing a longer common subsequence. To mitigate the discretion in selecting ϵ, an alternative approach is to

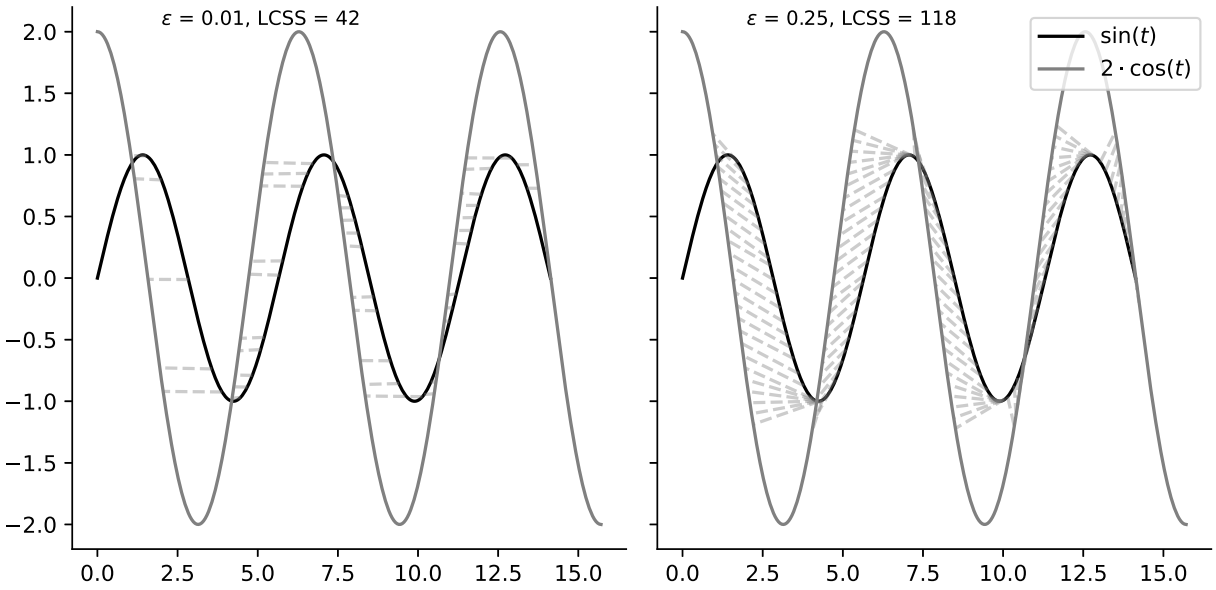


Figure 3: Time series similarity using LCSS for the prototypical example under different ϵ values

compute an overall distance measure by averaging across multiple distance computations up to a predefined maximum threshold, $\epsilon_{max} = 0.005$.[6] The interpretation of ϵ is meaningful only when applied to time series that are consistently scaled. For example, directly comparing the GDP levels across different series may yield few common subsequences due to differences in scale. To address this, it is often necessary to preprocess data. This is why we transformed variables into growth rates where relevant, as indicated in the last column of Table 3.

Table 5 presents the length of the LCSS for the simulated and observed capacity utilization rate according to Equation 1 under different values of ϵ. As expected, relaxing the restriction on ϵ leads to increased matches. A closer inspection of the index reveals that the original (`K+S-original`) and finance-augmented (`K+S-finance`) versions outperform the others, regardless of the threshold value selected. Notably, the other model versions fail to have any common subsequence with the observed data, suggesting the need for finer calibration. Still, this should not be interpreted as a justification for discarding such models outright; instead, it highlights the importance of evaluating similarity across multiple dimensions.

4.4 Dynamic Time Warping (DTW)

When using a pointwise method to compare two sequences, there is a risk of falsely identifying them as dissimilar when they are simply out of phase. This issue, as illustrated in the opening example of this section, can lead to the exclusion of relevant candidates. The Dynamic Time Warping (DTW) algorithm addresses this problem by providing a distance measure robust to time shifts,

[6] In the context of growth rates, the value of $\epsilon_{max} = 0.005$ is not overly stringent.

	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
0.001	0	6	0	0	2
0.005	0	7	0	0	4
0.01	0	8	0	0	5
0.015	0	13	0	0	6
0.025	0	29	0	0	8
0.05	0	60	0	0	33
0.075	0	171	0	0	64
0.1	0	272	0	0	249

Table 5: LCSS index contrasting simulated and empirical capacity utilization series across different K+S model versions

making it well-suited for comparing non-aligned time series (Berndt and Clifford, 1994; Lei and Sun, 2007).[7] Due to this property, DTW has been widely applied across various fields, including speech recognition (Sakoe and Chiba, 1978; Itakura, 1975), gene expression analysis in bioinformatics (Aach and Church, 2001; Bar-Joseph et al., 2002), handwriting comparison (Rath and Manmatha, 2003), and, to a lesser extent, economics (Franses and Wiemann, 2020; G.-J. Wang et al., 2012; Raihan, 2017). As an alternative to conventional L_p norms, such as Euclidean distance (L_2 norm), DTW is widely regarded as a state-of-the-art approach for sequential pattern matching (Cuturi, 2011; Lei and Sun, 2007).

The DTW algorithm aims to compute the optimal alignment between two sequences by minimizing the distance between corresponding data points, even when they are not temporally aligned (as Euclidean distance requires). Instead, DTW allows for temporal shifts by aligning points along the time axis (Keogh and Ratanamahatana, 2005). The alignment cost is computed iteratively by recursively updating the distortion function d, which measures discrepancies between elements of the two sequences. The distortion cost is updated for $i = 2, 3, \ldots, m$ and $j = 2, 3, \ldots, n$, while the direction of alignment follows a predefined step function. Following Cuturi (2011), we adopt the symmetrical step pattern with no slop restrictions, initially proposed by Sakoe and Chiba (1978), which remains a standard choice in the literature (Giorgino, 2009). The cost of aligning two sequences using DTW is formally defined as follows:

$$D(i,j) = d(x_i, y_{v,j}) + \min \begin{cases} D(i-1,j) & (\rightarrow) \\ D(i,j-1) & (\uparrow) \\ D(i-1,j-1) & (\nearrow) \end{cases} \tag{2}$$

Equation 2 represents the recurrence relation for computing the alignment cost along the step

[7] We refer to Kruskal and Liberman (1999) for a more detailed discussion about DTW and Sardá-Espinosa (2019) for its implementation.

pattern. The alignment proceeds monotonically by either moving right (skipping an element in the observed sequence), upward (skipping an element in the simulated sequence), or diagonally (aligning both). Consistent with Sardá-Espinosa (2019), we use the L_1 (Manhattan distance) as the discrepancy function for computing the distortion costs.

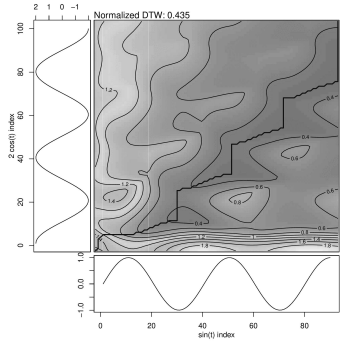
After computing the distortion costs, the algorithm constructs an $m \times n$ Local Cost Matrix (LCM), which accumulates alignment costs along the alignment space $\mathcal{A}(m, n)$.[8] Starting from the first element of each sequence $(x_1, y_{v,1})$ and proceeding to the last $(x_m, y_{v,n})$, the LCM enables the identification of multiple possible warping paths (π) that traverse the alignment space. One limitation of DTW is that it requires the first and last elements of the query/reference pair to be aligned as a boundary condition. In other words, it needs to start at x_1, y_1 and end at x_m, y_n. Some modifications can relax this restriction by allowing the algorithm to choose the starting/ending point with the lowest alignment costs. However, these modifications can lead to matches with little significance. For instance, the algorithm might suggest a warping path in which only a small portion $m' \ll m$ of the sequences are aligned. In other words, the resulting alignment could show that the series are alike, but only for a short time window. Thus, the decision was made to maintain the first/last alignment constraint. These paths define how indices (*i.e.*. time steps in the case of time series) are remapped between two sequences, effectively capturing distortions in the time axis. DTW aims to determine the optimal warping path $\pi^\star$ that minimizes the cumulative distortion over time. This is expressed as:

$$\mathcal{D}_{DTW} = DTW(\mathbf{x}, \mathbf{y}_v) = \min_{\pi \in \mathcal{A}(m,n)} D_{\mathbf{x},\mathbf{y}}(\pi)$$

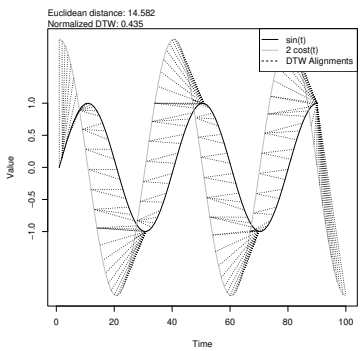
At this stage, it is important to emphasize that our goal is not to align the sequences directly but rather to measure their similarity. Hence, our primary focus is on computing distortion costs rather than optimizing the alignment.

Before delving into how DTW is applied in our comparative exercise, we revisit the motivating example presented at the beginning of this section. This example highlights two common issues in time series similarity measurement: misalignment (time shifts) and scale mismatches. Figure 4 exemplifies how the DTW algorithm works. In all subfigures, the horizontal axis represents the time indices of the $\sin(t)$ series, while the vertical axis corresponds to the time indices of $2 \cdot \cos(t)$. The cumulative local cost matrix for our prototypical example, computed using Equation 2, is in the main plot of Figure 4a. The solid line represents the warping path, which traverses regions of minimal distortion cost within the alignment space. This warping path provides an optimal alignment of the two series, as shown in Figure 4b. As demonstrated in this figure, the Euclidean distance substantially overestimates dissimilarity compared to DTW, underscoring DTW's robustness to time shifts.

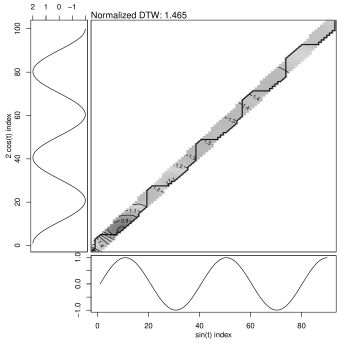
[8] One drawback of this algorithm is the computational cost of obtaining the matrix, which is of the order of $\mathcal{O}(m \cdot n)$. These aspects also motivate the departures from it, which will be discussed in the sub-section 4.5.



(a) Cumulative cost matrix and sequence alignment (no constraints)



(b) Time series alignment using DTW



(c) Cumulative cost matrix and sequence alignment (with global constraints)

Figure 4: Computing similarity using DTW: Revisiting the prototypical example

DTW ensures that the warping path follows the trajectory with the least temporal distortion. Still, this alignment may not always be economically meaningful. To address this issue, we impose global window constraints (ω), which restrict the warping path to align points within ω time periods, regardless of the distortion costs involved.[9] Paths falling outside this range are excluded from the analysis, even when producing a costless alignment. By imposing these constraints, we ensure that the computed alignments remain within an economically relevant time frame.[10] Figure 4c illustrates the distortion cost between the two series, restricted to a defined window. The resulting distance measurement increases as the algorithm is no longer allowed to traverse paths with minimal temporal distortion.

Having established the operationalization of DTW, we now outline its application in our analysis. Our approach involves computing the normalized DTW distance to align various model versions with a common empirical reference.[11] A lower DTW distance indicates greater similarity, with a zero value representing perfect alignment.[12] Table 6 presents the results for the simulated and observed capacity utilization rate time series. The distance computation is performed under different window sizes. As expected, increasing the window size reduces alignment costs, as the algorithm can explore a broader range of warping paths. To summarize these results, we compute the average distance up to a maximum window size ($\omega_{\max} = 8$), which is further utilized for constructing the

[9] In signal processing, various windowing procedures have been described in the literature. Sakoe and Chiba (1978) introduced one of the most well-known methods. However, this technique applies when the time series dimensions are equal ($m = n$). To address cases where $m \neq n$, rather than employing downsampling, we have chosen to utilize a slanted band window function (Giorgino, 2009).

[10] Another alternative would be to adopt a lower-bound DTW proposed by Lemire (2009) and Keogh and Ratanamahatana (2005). However, this algorithm only applies in cases where $m = n$.

[11] Normalization is important to yield comparable results in cases where $m \neq n$.

[12] It is worth noting that this distance is non-metric — not defined in a closed range — so only $\mathcal{D}_{DTW}$ is defined. Hence, additional re-scaling procedures are necessary to enable comparison with other measurements. Further details on this topic will be discussed in subsection 4.7.

multidimensional similarity index.

	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
2	0.177	0.090	0.172	0.402	0.097
4	0.172	0.084	0.167	0.396	0.092
8	0.166	0.075	0.161	0.387	0.083
20	0.158	0.060	0.153	0.372	0.068
40	0.153	0.047	0.148	0.357	0.054
60	0.149	0.042	0.145	0.352	0.050

Table 6: DTW index contrasting simulated and empirical capacity utilization series across different K+S model versions

The similarity index for these specific sequence pairs reveals that the finance-augmented model version (`K+S-finance`) shows the lowest dissimilarity, indicating minimal temporal distortion relative to the empirical series. The multi-sector version (`K+S-multi`), on the other hand, has the highest time distortion. Notably, the original model version (`K+S-original`) ranks second in similarity. The two labor-augmented institutional settings produce relatively close results. As these two models share the same structure, this outcome is expected.

4.5 Triangular Global Alignment Kernel (TGAK)

A key limitation of Dynamic Time Warping (DTW) is that it only explores a subset of possible alignments between two sequences, constrained by a predefined step pattern. This restriction prevents DTW from considering the full spectrum of alignments in the alignment space $\mathcal{A}(m, n)$. Cuturi et al. (2007) introduced the Global Alignment Kernel (GAK), a kernel-based method that measures similarity by aggregating contributions from all possible alignments to address this restriction. Unlike DTW, which focuses on the minimum-cost alignment, GAK computes the exponentiated soft-minimum of all alignment costs, as defined in Equations 3.[13] This approach provides a more comprehensive assessment of dissimilarity, capturing both strong and weak alignments. As a result, GAK offers a richer similarity measure compared to DTW, particularly when sequences have significant divergence.

The GAK is formally defined as:

$$k_{\mathrm{GA}}(\mathbf{x}, \mathbf{y}_v) = \sum_{\pi \in \mathcal{A}(m,n)} \prod_{i=1}^{|\pi|} \kappa\left(\pi_x(i), \pi_{y,v}(i)\right) \tag{3}$$

[13]The computation of the kernel requires estimating a hyperparameter (σ), and we adopt the suggestion of Cuturi (2011) implemented by Sardá-Espinosa (2019), which relies on a subsample of both series. Unlike previous methods, this sampling introduces randomness into the distance estimation, which can be controlled by setting a random seed.

where $\kappa(\mathbf{x}, \mathbf{y}_v) = \exp \phi_\sigma(\mathbf{x}, \mathbf{y}_v)$ is a local kernel computed over the aligned points in $\mathbf{x}$ and $\mathbf{y}_v$, while ϕ_σ is a negative definite kernel given by:

$$\phi_\sigma(\mathbf{x}, \mathbf{y}_v) = \frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{y}_v\|^2 + \log\left(2 - \exp - \frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \tag{4}$$

Building on the GAK framework, Cuturi (2011) introduced the Triangular Global Alignment Kernel (TGAK), which incorporates a triangular local kernel $T(i, j)$ to constrain the alignment space — as the global constraint in DTW.[14] Unlike DTW, where a larger constraint window typically reduces the distortion cost, TGAK does not guarantee that relaxing the window constraint will lead to lower dissimilarity. A larger ω window may result in higher dissimilarity due to the inclusion of suboptimal alignments. The TGAK is defined as:

$$T(i, j) = \left(1 - \frac{|i - j|}{\omega}\right)_+ \tag{5a}$$

$$\text{TGAK}(\mathbf{x}, \mathbf{y}_v, \sigma, \omega) = \tau^{-1}\left(T \otimes \frac{1}{2}\kappa\right)(i, \mathbf{x}; j, \mathbf{y}_v) = \frac{T(i, j)\kappa(\mathbf{x}, \mathbf{y}_v)}{2 - T(i, j)\kappa(\mathbf{x}, \mathbf{y}_v)} \tag{5b}$$

where τ^{-1} is the inverse mapping function, and ω controls the width of the triangular kernel[15].

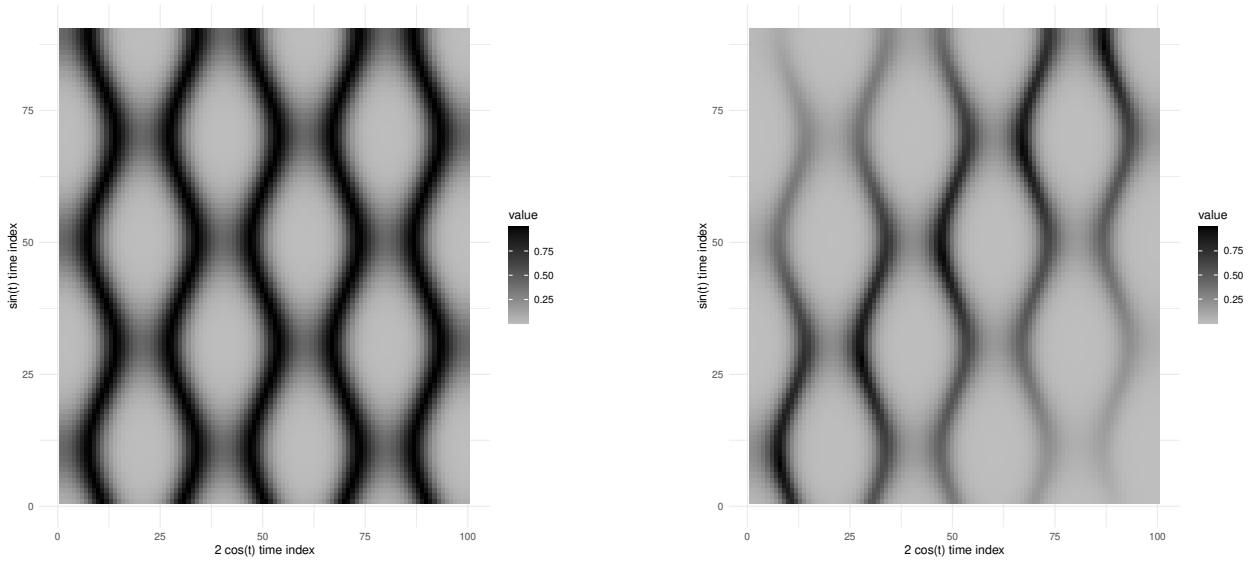
Figure 5 exemplifies the local similarity matrix (LSM) computed with the local kernel using Equation 4 for the prototypical $\sin(t)$ and $2\cos(t)$ example. Each entry of this matrix contains pairwise similarities between individual elements of the two time series being compared. The difference between the two sub-figures is the presence of a global constraint over the alignment space. In the case paired with the triangular kernel (Figure 5b), the similarity is reduced if the compared points are far apart from the time restriction modulated by ω. This ensures that only temporally close points contribute substantially to the overall similarity measurement.

The next step consists of accumulating the similarity scores for all alignments possible. By doing this, TGAK returns similarity scores ($\mathcal{S}_{TGAK} = TGAK$) ranging from zero (perfect similarity) to one (maximum dissimilarity), providing a proper metric distance $\mathcal{D}_{TGAK} = 1 - \mathcal{S}_{TGAK}$. Similar to the implementation of DTW, described in subsection 4.4, we explore different global constraints ω, averaging results up to $\omega_{\max} = 8$.

Table 7 presents the results of applying TGAK to compare empirical capacity utilization across different model versions. The results show that the similarity scores are numerically close across all

[14] According to Cuturi (2011), one technical advantage of using the triangular kernel is reduce the complexity from $\mathcal{O}(mn)$ to $\mathcal{O}(\min(m, n))$.

[15] A practical challenge with GAK is diagonal dominance, which occurs when the lengths of the sequences being compared differ considerably ($2 \cdot m \ll n$ or $m \gg 2 \cdot n$). In such cases, the kernel values $\kappa(\mathbf{x}, \mathbf{y}_v)$, may be biased toward self-alignments, distorting the similarity analysis (Cuturi et al., 2007). This is a probable scenario as we encounter cases where $m \ll n$. To mitigate this issue, we apply centered downsampling, truncating the longer sequence by removing an equal number of elements from both the beginning and the end. This ensures that the lengths of the sequences are comparable, reducing the risk of diagonal dominance.



(a) LSM without global constraint

(b) LSM paired with triangular kernel

Figure 5: Local Similarity Matrix (LSM) for the prototypical example

	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
2	0.023	0.020	0.023	0.023	0.021
4	0.026	0.022	0.025	0.026	0.023
8	0.027	0.024	0.027	0.028	0.024
20	0.029	0.025	0.029	0.030	0.026
40	0.030	0.026	0.030	0.030	0.027
60	0.030	0.026	0.030	0.031	0.027

Table 7: TGAK index contrasting simulated and empirical capacity utilization series across different K+S model versions

model versions, regardless of the window constraint ω. The finance-augmented model (`K+S-finance`) produces the most similar series, closely followed by the original model (`K+S-original`), while other models fall within a narrow range[16]. This suggests that TGAK alone may not isolate the most similar model. To refine the analysis, we use a min-max normalized distance index, which provides a finer-grained assessment to compare similarity. This highlights the need for a multidimensional similarity approach to distinguish between closely related models, discussed in sub-section 4.7.

4.6 Shape-Based Distance (SBD)

The previous indices do not explicitly account for the lag structure of the series, a commonly neglected factor in data mining literature (see X. Wang et al. (2013) for a review). The Shape-

[16]It is worth noting that these findings are consistent with the ones produced by DTW.

Based Distance (SBD) measurement, introduced by Paparrizos and Gravano (2015), addresses this gap and has been successfully applied in time series clustering (Fahiman et al., 2017). This index captures the temporal shape of two series by utilizing the normalized cross-correlation ($NCC(\mathbf{x}, \mathbf{y}_v)$, Equation 6). In signal processing, normalized cross-correlation — or sliding inner-product — is a measure of similarity between two signals as a function of the time lag slid over one of them. The result is scaled to a range of $[-1, 1]$, facilitating interpretable comparisons.[17] For two sequences $\mathbf{x}$ and $\mathbf{y}_v$ the discrete form normalized cross-correlation is given as:

$$\mathrm{NCC}(\mathbf{x}, \mathbf{y}_v, k) = \frac{CC(\mathbf{x}, \mathbf{y}_v, k)}{\sqrt{\sum_{n=1}^{n}(x[n] - \bar{x})^2 \cdot \sum_{n=1}^{n}(y[n+k]_v - \bar{y}_v)^2}} \tag{6}$$

$$CC(\mathbf{x}, \mathbf{y}_v, k) = \sum_{n=1}^{n}(x[n] - \bar{x})(y[m+k]_v - \bar{y}_v) \tag{7}$$

where $CC(\mathbf{x}, \mathbf{y}_v)$ is the cross-correlation between $\mathbf{x}$ and the shifted version of $\mathbf{y}_v$, and k represents the time index used to compute the shift.

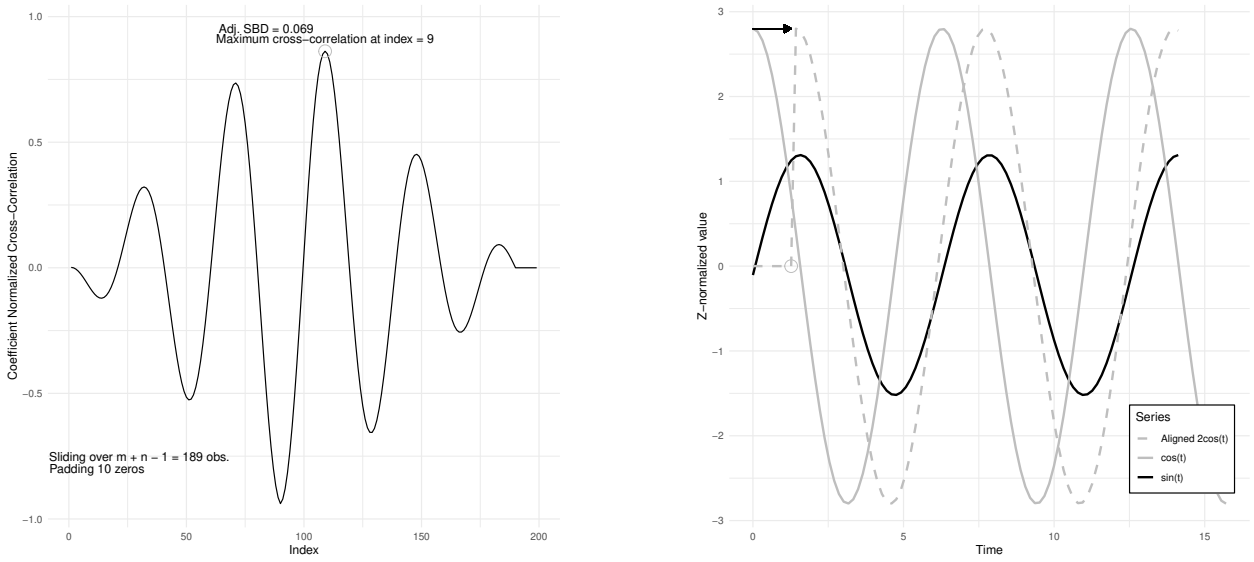
The cross-correlation is computed over $m + n - 1$ time steps. The algorithm identifies the lag i at which the two series have the highest cross-correlation, effectively capturing the similarity of out-of-phase signals (see Equation 8). Like Dynamic Time Warping (DTW) and the Triangular Global Alignment Kernel (TGAK), SBD is robust to temporal shifts, making it particularly suitable for comparing time series with phase differences.

$$SBD(\mathbf{x}, \mathbf{y}_v) = 1 - \max\{(NCC(\mathbf{x}, \mathbf{y}_v, k))\}_{k=1}^{k=m+n-1} \tag{8}$$

The SBD metric is constrained to the range $[0, 2]$, where zero indicates identical temporal shapes between the two series. For comparability with other measurements, we normalize the results by dividing them by two (details in section 4.7). As SBD provides a closed-form distance measurement, we can represent the results in terms of the distance $\mathcal{D}_{SBD} = SBD(\mathbf{x}, \mathbf{y}_v)$ and similarity $\mathcal{S}_{SBD} = 1 - \mathcal{D}_{SBD}$.

Before applying SBD, several practical issues must be addressed. First, the metric is sensitive to the scale of the input series. Following Paparrizos and Gravano (2015), we standardize each series by computing its z-score, ensuring scale invariance. Second, length mismatches between series can affect results as the SBD implementation utilizes the convolution theorem to reduce the computational complexity. In cases of length mismatch, zero-padding is applied during convolution to equalize series lengths. As this operation may introduce artifacts into the series, potentially distorting the signals, we adopt the centered downsampling approach, resizing both series to match the length of the shorter one.

[17] While the signal processing definition emphasizes time-domain analysis and signal detection, the statistical and econometric interpretation focuses on assessing the linear relationship and lead-lag dynamics between variables, often using a similar normalization to quantify correlation strength.



(a) Normalized Cross-Correlation

(b) Series alignment based on maximum NCC

Figure 6: Computing similarity using SBD: Revisiting the prototypical example (z-normalized)

Figure 6 illustrates the SBD alignment process for the prototypical example. Figure 6a shows the normalized cross-correlation (after z-normalization) between two series, while Panel 6b displays the same series after shifting one of them by the lag that maximizes cross-correlation. Visual inspection confirms the robustness of SBD to phase differences, as it correctly identifies the shift required to realign the sequences. The distance measure is computed at this optimal lag, reflecting the shape similarity of the series.

Table 8 presents the similarity analysis of simulated series compared to the empirical capacity utilization, using the $\mathcal{D}_{SBD}$ distance measure. Because SBD requires no additional parameterization, we apply it directly to the raw (unfiltered), trend, and cyclical components of the series. For the unfiltered series, the results align with earlier findings: labor-augmented versions perform relatively worse, while the original (`K+S-original`) model has the least dissimilarity. The finance-augmented (`K+S-finance`) and multi-sector (`K+S-multi`) models show comparable performance. This finding persists in the cyclical component, where the original model remains among the most similar while the other models have smaller relative differences. For the trend, the finance-augmented is considerably dissimilar compared to the other models, while the original version — indicated as the least dissimilar at other frequencies — no longer stands out.

4.7 Building the Multidimensional Index

The preceding sections introduced various measurements designed to capture distinct aspects of similarity between two series along different domains. Subsection 4.2, presents an index incorporating information on the probability distribution of the series. Subsection 4.3 utilized the Longest

	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
UNFILTERED	0.385	0.233	0.311	0.267	0.192
TREND	0.349	0.307	0.352	0.342	0.350
CYCLE	0.370	0.378	0.370	0.374	0.367

Table 8: SBD index contrasting simulated and empirical capacity utilization series across different K+S model versions

Common Subsequence (LCSS) to account for the similarity of the series' trajectories. Subsections 4.4 and 4.5 describe Dynamic Time Warping (DTW) and Triangular Global Alignment (TGAK), which are elastic measurements that handle temporal misalignments between sequences. Finally, subsection 4.6 discussed the Shape-Based Distance (SBD), which is robust to time shifts and provides insights into the lag structure of the series.

These measurements highlight a critical insight: no single similarity metric can fully capture the complexity embedded in the signals produced by the models. To address this limitation, we propose a multidimensional index that combines these diverse measurements, thereby achieving a more comprehensive approximation of similarity that embraces both the sequential and distributional domains of the data structure.

Certain adjustments are necessary to create our multidimensional index. First, we address the issue of differing ranges across measurements. Except for TGAK and SBD, the other criteria operate on non-closed ranges. Additionally, the LCSS measurement, as implemented, interprets similarity in the opposite direction compared to the others — higher values indicate greater similarity.[18] To ensure consistency, we transform all measurements to convey the same information.

As we have non-closed and non-metric indexes, we present the results in terms of dissimilarity. Thus, we use the normalized LCSS distance ($\mathcal{D}_{LCSS}$) so that higher values indicate greater dissimilarity. Next, we rescale all measurements using min-max normalization, where one represents the most dissimilar and zero represents the least dissimilar. This normalization ensures that all measurements are on a comparable scale. We then compress the information from all measurements into a single index using the geometric average.[19] This approach is chosen because it is less sensitive to extreme values than the arithmetic mean, making it more robust for combining dissimilarity metrics. Equation 9 presents the geometric mean computation.

$$\text{Geometric Average}_v = \sqrt[5]{\mathcal{D}_{Statistical,v} \cdot \mathcal{D}_{NLCSS,v} \cdot \mathcal{D}_{DTW,v} \cdot \mathcal{D}_{TGAK,v}, \mathcal{D}_{SBD,v}} \tag{9}$$

We introduce a penalizing factor based on Shannon (1948) entropy, to refine the index further. The entropy is computed to quantify the consistency of a model's performance across all measure-

[18] This is the case because implementing the LCSS in the TSdist package (Mori et al., 2014) returns the occurrence of contiguous subsequences that respect the matching condition, so higher values indicate greater similarity.

[19] We use the geometric average extension robust to the presence of zeroes, as proposed by Cruz and Kreft (2019).

ments. Specifically, we calculate the number of times it occupies each position in the ranking system for each model across all measurements. This is achieved by tabulating the positions of each model across the rankings and computing the entropy of the resulting distribution. A high entropy value indicates that a model performs inconsistently, appearing in multiple positions across the indices. Conversely, a low entropy value suggests that the model consistently occupies a similar position across all measurements.[20] To ensure comparability, we normalize the entropy by dividing it by the maximum possible entropy, which occurs when a model occupies a different position for each measurement. The normalized entropy, ranging from zero to one, is then used as a penalizing factor in the geometric average, which we refer to as the final multidimensional index. This formulation ensures that models with high entropy (inconsistent performance) are appropriately down-weighted in the final index. The penalized index is computed as follows:

$$\text{Entropy Adj. Geometric Average}_v = \text{Geometric Average} \cdot (1 + \text{Entropy}) \tag{10}$$

To showcase our approach, we present results for the capacity utilization rate as a representative variable in Table 9. The numbers in parentheses represent the rank across measurements, with one indicating the least dissimilar model and five indicating the most dissimilar.[21] We will use this example to discuss the multidimensional index in more detail and how to interpret it. According to the multidimensional index, the original model `K+S-original` outperforms the others, while the multi-sector version is the most dissimilar. This result is driven by the original model's good performance in replicating similar statistical moments and lag structures as observed in the empirical data (captured by SBD). Although the finance-augmented version performs well in terms of trajectory similarity (captured by LCSS) and alignment paths (captured by DTW and TGAK), the original model's consistent performance across all measurements explains its position as the relatively most similar model.

Our findings also provide valuable insights into domains where specific models require improvement. For instance, the `K+S-multi` model has poor performance in terms of statistical moments and shows notable dissimilarity regardless of temporal elasticity (as indicated by DTW and TGAK). In particular, the labor-augment and multi-sector versions perform poorly in capturing the trajectories of the series, with no common subsequences identified.

Figure 7 provides a comprehensive overview of the relative similarity across all variables, models, and filters using the multidimensional index[22]. The heatmap is divided into three subplots – Unfiltered, Trend, and Cycle – each corresponding to a different frequency component. The Y-axis

[20] It is worth noting that a low entropy does not necessarily mean that the model under consideration has a relatively good performance. In other words, a model with a consistently poor performance will present a low entropy.

[21] Due to space limitations, we present the results for the unfiltered series, although the same process is repeated for the filtered variables.

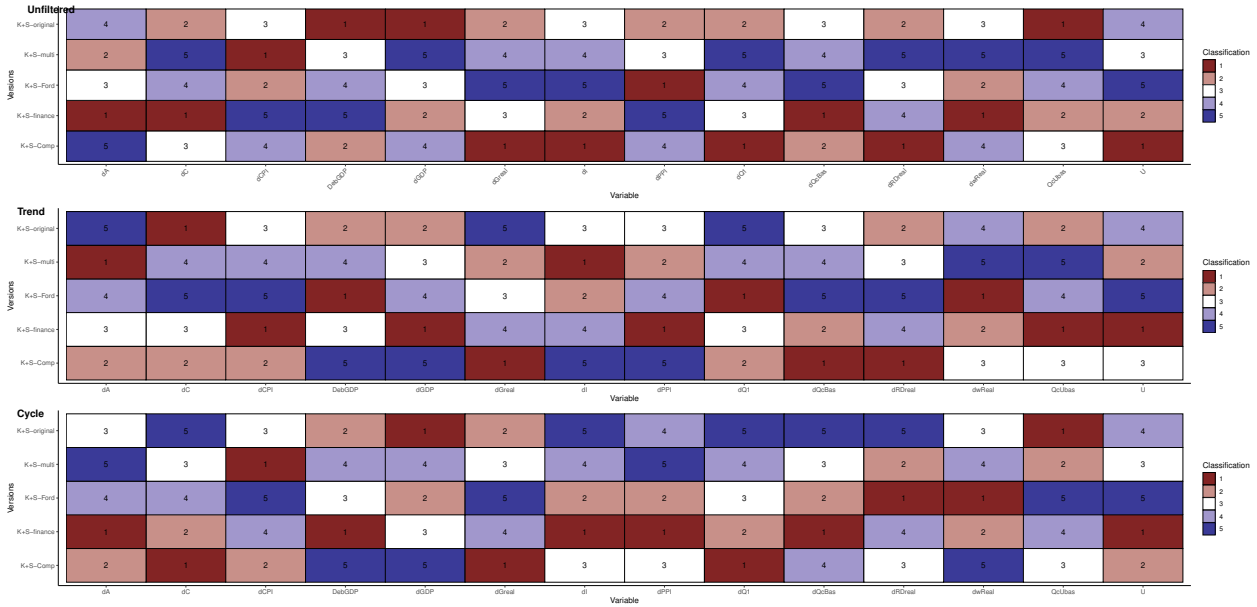
[22] Table A.1 in Appendix A offers a more detailed breakdown for each model, variable, and filter across all measures.

	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
Statistical	0.512 (3)	0.452 (2)	0.891 (5)	0.842 (4)	0.259 (1)
LCSS	1 (3)	0.155 (1)	1 (3)	1 (3)	0.375 (2)
DTW	0.171 (4)	0.083 (1)	0.167 (3)	0.395 (5)	0.091 (2)
TGAK	0.025 (4)	0.022 (1)	0.025 (3)	0.026 (5)	0.023 (2)
SBD	0.385 (5)	0.233 (2)	0.311 (4)	0.267 (3)	0.192 (1)
Geometric Average	6e-01 (3)	3e-03 (2)	7e-01 (4)	8e-01 (5)	1e-03 (1)
Entropy	7e-01 (4)	4e-01 (1)	6e-01 (3)	7e-01 (4)	4e-01 (1)
Entropy adj Geom. Avg.	1e+00 (3)	4e-03 (2)	1e+00 (4)	1e+00 (5)	1e-03 (1)

Table 9: Example of the computation of the multidimensional index for the unfiltered capacity utilization index across different K+S model versions

represents the model versions, while the X-axis denotes the variables listed in Table 3[23]. Each cell displays the ranking (one to five) within the same variable, where lower values indicate superior relative performance.

Figure 7: Relative performance of all models according to the multidimensional index across all variables and filters



In the Unfiltered dataset, model rankings appear more dispersed, suggesting that when all frequency components are considered, differences in model performance become more pronounced. This could be attributed to unfiltered data incorporating both short-term fluctuations and long-term trends, requiring models to balance these dynamics effectively. Conversely, in the Trend and Cycle decompositions, the rankings exhibit more structured patterns, indicating that specific

[23]Variables transformed into growth rates are prefixed with "d" to reflect this adjustment. For instance, dGDP represents the growth rate of real GDP.

models (notably `K+S-Comp` and `K+S-finance`) better capture long-term structural movements or cyclical dynamics. The systematic shifts in rankings across these panels underscore the importance of aligning model selection with the specific frequency components of interest in economic analysis.

Overall, these findings emphasize the sensitivity of model performance to the frequency domain and variable characteristics. For instance, while the `K+S-original` and `K+S-Comp` models perform well in the unfiltered dataset, their rankings deteriorate under different frequency filters. Specific economic indicators, such as GDP growth rate (dGDP) and general price inflation (dCPI), show relatively stable rankings across filtering methods, suggesting that model performance for these variables is more robust to data transformations. In contrast, variables like government consumption (dGreal) and consumer goods production (dQcUbas) growth rates show substantial ranking shifts depending on the filtering approach. This underscores the necessity of selecting a model tailored to the specific characteristics of the dataset and research objectives.

This section has demonstrated how to capture different aspects of similarity between two signals. Several key insights emerge from this approach. First, relying on a single similarity index may overlook critical information embedded in the signals, reinforcing the importance of considering multiple dimensions of similarity. Second, our method identifies specific aspects where a model requires refinement (indicated with higher ranking positions), enabling targeted improvements. Finally, the multidimensional index makes constructing a ranking system that sorts models based on their similarity to empirical data possible. In the next section, we leverage the ranking system derived from our proposed index to conduct model selection exercises.

5 Model selection

Following the workflow illustrated in Figure 2, we have computed a multidimensional index for each variable across all models being compared. This index quantifies the degree of similarity between each model and the empirical data, providing a comprehensive basis for comparison, discussed in the previous section. A striking result from this analysis is the absence of a single model version that consistently outperforms the others across all variables. Performance rankings vary notably depending on the specific variable and the frequency component analysed, highlighting the need for a nuanced approach to model selection.

A wide array of complementary approaches exists in model selection, each tailored to specific modeling contexts and objectives. A critical aspect of model selection is its coexistence with validation and calibration exercises, which ensure that the chosen model replicates key empirical patterns. One well-known set of techniques, rooted in statistical and econometric traditions, relies on information criteria and forecasting accuracy (Lamperti, 2018; Martinoli et al., 2024; Poledna et al., 2023). Another group prioritizes the ability of models to generate realistic qualitative outcomes, such as stylized facts, through history-friendly calibration (Windrum et al., 2007; Fagiolo, Guerini,

et al., 2019b) or pattern-oriented modeling (Grimm et al., 2005). Despite this diversity, the predominant focus has been on predictive performance or analyzing impulse response functions rather than measuring the direct similarity between models and empirical data.

This chapter introduces an alternative selection scheme based on our multidimensional index, departing from traditional practices. Starting from the ranking system described in the previous section, we remove models with a higher occurrence of the most dissimilar variables rather than selecting those with the highest frequency of similar cases. Algorithm 1 outlines the pseudocode for this iterative process. The first selection exercise does not attribute a distinct importance to any variables. Later, we imposed different weights to adopt a purpose-driven selection framework, similar to what Leombruni and Richiardi (2005) did.

Algorithm 1 Pseudo code of the iterative process

Require: dimension(MultidimensionalIndex) $\equiv$ Variables $\times$ $\mathcal{V}$ ▷ *Dataframe produce in phase 1*

Require: length(TieBreaker) $\equiv$ length($\mathcal{V}$) ▷ *TieBreaker is a vector of strings*

Require: length(Weights) $\equiv$ length(Variables) ▷ *Weights is a numeric vector*

function IterativeSelection(MultidimensionalIndex, TieBreaker, Weights)
  output Least dissimilar model configuration
  Versions = $\mathcal{V}$
  repeat
    RankedTable = matrix(Variables, Versions) ▷ *Reset at each iteration*
    for var **in** variables **do**
      RankedTable[var,Versions] ← rank(MultidimensionalIndex[var, Versions]) ▷ *Rank from 1 to lenght(Versions)*
    WorstLevel = max(RankedTable) ▷ *Gets the higher dissimilarity*
    for var **in** variables; ver **in** Versions **do**
      ▷ *Transforms the ranked table in a binary matrix* ◁
      if RankedTable[var, ver] $\equiv$ WorstLevel **then**
        RankedTable[var, ver] ← 1
      else
        RankedTable[var, ver] ← 0
    RankedTable ← RankedTable × Weights ▷ *Apply weights penalization to all columns*
    OccuranceWorst = matrix(Versions) ▷ *To collect number of occurances of worst cases*
    for ver **in** Versions **do**
      OccuranceWorst[ver] = ← $\sum$ RankedTable[:, ver]
    MostDissimilar = name(max(OccuranceWorst)) ▷ *Find the most dissimilar*
    if length(MostDissimilar) $>$ 1 **then**
      MostDissimilar ← sort(MostDissimilar, TieBreaker)
      ▷ *Apply the tie breaker criteria only if there is a tie* ◁
    MostDissimilar ← first(MostDissimilar) ▷ *Collects the most dissimilar*
    Versions ← drop(Versions, MostDissimilar) ▷ *Remove the most dissimilar from the list*
  until length(Versions) $\equiv$ 1
  return Versions ▷ *This is the least dissimilar version*

The first group of Table 10 provides the number of relative "worst" cases for each variable

using this algorithm. Although this analysis is applied to both cyclical and trend components, we illustrate our approach using the unfiltered series for clarity. In the first iteration, the `K+S-multi` version is removed, as it is the most dissimilar for the highest number of variables. The ranking may change after each removal because dissimilarity is measured in relative terms. We then sort the remaining candidates, assigning a value of one to the least dissimilar and four (the number of kept models, $\mathcal{V} - 1$) to the most distant one. This iterative process stops when only one model remains. Continuing this procedure, the next removed models are `K+S-Ford`, `K+S-finance`, and `K+S-Comp`. Based on this, we find that the `K+S-original` is the least dissimilar for the unfiltered series, while the `K+S-Comp` ranks as the second best across all frequencies. When analyzing other components of the frequency domain, the `K+S-finance` outperforms the other alternative versions.

Although the iteration ends here, this should not be seen as a final step. We suggest revisiting the performance of the selected candidates to refine the analysis. We refer to Figure 7, which provides a visual representation of the model performance across variables. A quick inspection of this heatmap reveals variables where the models could be improved. For example, the `K+S-original` shows relatively higher dissimilarity for productivity and unemployment series, indicating variables requiring further attention. Not only is the `K+S-original` the least dissimilar model for the unfiltered series, but it also has no occurrences of relatively worst variables before the iterative removal process begins. The same applies to the `K+S-finance` model for other frequencies. This indicates that the selection criteria successfully identify a good candidate as the least dissimilar model version.

Table 10: Number of relativelly worse variables across different K+S versions, filters, and purpose-driven experiments

	Unfiltered					Trend					Cycle				
Iterations	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
Baseline															
1	1	3	4	6	0	4	0	5	2	3	3	0	4	2	5
2	3	4	7	-	0	4	1	-	5	4	4	1	5	4	-
3	5	6	-	-	3	5	2	-	-	7	4	4	-	6	-
4	8	-	-	-	6	8	6	-	-	-	8	6	-	-	-
Policy-Oriented															
1	1	6	4	12	0	10	0	5	5	3	6	0	7	5	5
2	6	7	10	-	0	-	1	13	6	3	6	6	-	5	6
3	11	9	-	-	3	-	2	-	13	8	7	7	-	9	-
4	-	17	-	-	6	-	4	-	-	19	14	9	-	-	-
Research-Oriented															
1	4	6	7	12	0	10	0	8	2	9	3	0	4	8	14
2	6	10	13	-	0	-	4	13	3	9	4	4	5	16	-
3	8	15	-	-	6	-	8	-	7	14	12	4	13	-	-
4	14	-	-	-	15	-	13	-	16	-	17	12	-	-	-

5.1 Targeting variables: a policy-oriented exercise

It is worth noting that the selection procedure discussed so far is variable-agnostic. However, model design often targets specific research questions or policy-oriented debates. To address this, we propose an alternative approach that incorporates external criteria to evaluate the performance of any simulated model collection. At this stage, it is important to emphasize that the similarity of any two series does not depend on these criteria. Thus, we assign penalization weights during the iterative selection process rather than adjusting the multidimensional similarity index. Table 11 indicates the penalization weights associated with each variable in every experiment, in which a higher value forces the algorithm to demote the variable if the model has a relatively worse performance. Weights might be due to higher importance assigned to some specific variable, such as inflation in a phase of inflation spiral, unemployment in case of recessions, or productivity growth in case of prolonged stagnation.

Table 11: Penalization weights for purpose-guided model selection

	Policy-Oriented	Research-Oriented
dA	1	4
dC	2	1
dCPI	4	2
DebGDP	1	1
dGDP	4	4
dG	2	1
dI	2	2
dPPI	2	2
dQ1	2	2
dQcBas	2	1
dRDreal	1	4
dwReal	1	1
QcUbas	2	1
U	4	4

Once the priorities are set, we apply the iterative selection process with slight changes. Our goal is to identify the most similar model, so we use these priority weights as a penalization factor during the elimination step. As before, we first count the number of occurrences of the worst relative performance for each model under scrutiny. Unlike the previous exercise, a model is penalized by

these weights when it is more dissimilar with respect to the target variables. This approach allows us to guide the iterative procedure toward a desired purpose.

Policy-oriented considerations inspire the first experiment without aiming to replicate any specific policy decision. Through this experimental setup, we employ arbitrary weights to demonstrate the refined model selection mechanism[24]. For instance, suppose policymakers are interested in targeting specific macroeconomic variables, such as real GDP growth rate, inflation, and unemployment rate. Labor productivity and R&D variables are assigned lower priority, as they are indirectly affected by government actions in the short run. The iterative process then begins[25]. The second group of Table 10 displays the models removed at each iteration. Interestingly, the same model selection emerges in this experiment, with the `K+S-original` producing similar results for the unfiltered series, while the `K+S-finance` remains the least dissimilar for other frequencies. Except for the trend component, the `K+S-Comp` ranks as the second-best model in this targeted exercise.

As previously discussed, the models evaluated here were designed to reflect general regularities rather than a specific economy or time period. Therefore, it is more appropriate to contrast the model collections with their theoretical aims. For this reason, we also evaluate their relative performance in terms of similarity to variables associated with technological change, such as labor productivity, R&D expenditures, and their effects on real GDP and unemployment rates. The procedure remains the same as described earlier, with the only difference being the weights assigned to each variable, as shown in the second column of Table 11.

The third group of Table 10 presents the models removed at each iteration. Unlike previous exercises, the `K+S-original` is no longer the best candidate for the unfiltered series, displaced by the finance-augmented version, which also ranks as the best model for other frequencies. This result is primarily driven by the fact that the original version does not generate similar series for innovation-related variables and the unemployment rate. As a consequence, the finance-augmented version is more suitable for analyses involving these variables. The `K+S-Comp` maintains its relatively good performance, again ranking as the second-best model. Overall, this model performs well regarding labor productivity and unemployment rates, which is consistent with its theoretical design.

Considering all this, model performance and selection must be evaluated in light of the specific elements and objectives the model was designed to address. The `K+S-original` and `K+S-finance` models emerge as the least dissimilar in different contexts, with their performance varying substantially depending on the variables and frequency components analysed. These findings underscore the importance of aligning model selection with theoretical and practical objectives, offering a flexible framework for identifying the most suitable model for specific applications.

[24]The ranging of the values does not have an explicit reasoning, but rather imposes an ordinal hierarchy on the variables. The fine-tuned selection can be performed as long as there is an unambiguous hierarchy among variables.

[25]It is important to note that there are two models with the same number of worst cases on the second iteration of the cyclical component. This requires a tie-break rule. We remove the model with the higher mean value according to our proposed multidimensional index (see table A.1 in the appendix). While this tiebreaker criterion is not the only possible option, the number of tied cases is small, as shown in Table 10.

6 Concluding remarks

The primary contribution of this chapter lies in the development of a protocol that enables modelers to evaluate the similarity between their models and empirical data, with specific reference to the K+S macroeconomic ABM. We argue that the similarity of a model *vis-à-vis* the empirical evidence should be measured from a multidimensional perspective, considering various aspects when comparing two objects along the time domain. Building on this premise, we have created an index to capture these diverse dimensions, providing a detailed understanding of the relative performance of each model.

Our work is part of a larger effort to bridge the gap between simulated models and their real-world counterparts, focusing on a specific category of macroeconomic agent-based models. The motivation behind this exercise is the increasing demand for data-friendly models that can enhance forecasting capabilities for policymakers. While our application is rooted in economic literature, the protocol's utility is not limited to this field. With broader applications in mind, the design of the protocol aims to be as adaptable and scalable as possible, making it suitable for numerous research and policy challenges.

In addition to the protocol, we have proposed two potential uses. First, an iterative procedure identifies the most similar model. This procedure is further complemented by a refinement that considers an external criterion in the selection process, assigning weights to specific target variables. The analysis reveals that the finance-augmented K+S model demonstrates robust performance compared to recent developments, particularly when examining different frequencies. Likewise, the labor-augmented model, adjusted for competitive institutional settings, yields satisfactory results across most configurations. Second, we have performed a policy analysis exercise, targeting specific variables and showing that the protocol can be employed to assign specific weight and relevance to variables of interest. The iterative selection process can address several research questions and policy challenges by integrating domain-specific weights and criteria.

Another potential application, not explored due to space limitations, is the interactive calibration of models based on their similarity to empirical data. This approach could assist modelers in determining parameters that are not directly observable. For example, the modeler can compare different parametric settings of the same model against empirical data and choose the one that is more similar accordingly to the iterative model selection.

While the selection protocol and multidimensional index improve the understanding of the models' similarity with empirical data, they are not without limitations and offer opportunities for further refinement. At a higher level of abstraction, the lack of interaction among the variables under consideration is a notable limitation. Therefore, although the index covers multiple dimensions of similarity, it should not be interpreted as a global measure. Nonetheless, the protocol serves as a first effective step to identify a set of alternative configurations and/or models before resorting to more

computationally intensive methods. Finally, the applicability of the proposed method goes beyond ABM, enlarging the understanding of model performance under different theoretical assumptions.

References

Aach, John and George M. Church (June 1, 2001). "Aligning Gene Expression Time Series with Time Warping Algorithms". In: *Bioinformatics* 17.6, pp. 495–508. DOI: 10.1093/bioinformatics/17.6.495.

Amendola, Marco and Marcelo C. Pereira (2024). "Linear and State-Dependent Impulse Responses in Agent-Based Models: A New Methodology and an Economic Application". In: *Available at SSRN 4740360*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4740360 (visited on 06/04/2024).

Bai, Jushan and Pierre Perron (2003). "Computation and Analysis of Multiple Structural Change Models". In: *Journal of Applied Econometrics* 18.1, pp. 1–22. DOI: 10.1002/jae.659.

Bar-Joseph, Ziv et al. (Apr. 18, 2002). "A New Approach to Analyzing Gene Expression Time Series Data". In: *Proceedings of the Sixth Annual International Conference on Computational Biology*. RECOMB02: 6th Annual International Conference on Computational Molecular Biology. Washington DC USA: ACM, pp. 39–48. DOI: 10.1145/565196.565202.

Berndt, Donald J. and James Clifford (July 31, 1994). "Using Dynamic Time Warping to Find Patterns in Time Series". In: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. AAAIWS'94. Seattle, WA: AAAI Press, pp. 359–370.

Bottazzi, Giulio, Le Li, and Angelo Secchi (June 1, 2019). "Aggregate Fluctuations and the Distribution of Firm Growth Rates". In: *Industrial and Corporate Change* 28.3, pp. 635–656. DOI: 10.1093/icc/dtz016.

Bottazzi, Giulio and Angelo Secchi (2006). "Explaining the Distribution of Firm Growth Rates". In: *The RAND Journal of Economics* 37.2, pp. 235–256. DOI: 10.1111/j.1756-2171.2006.tb00014.x.

Bouchaud, Jean-Philippe (Dec. 2023). "From Statistical Physics to Social Sciences: The Pitfalls of Multi-Disciplinarity". In: *Journal of Physics: Complexity* 4.4, p. 041001. DOI: 10.1088/2632-072X/ad104a.

Christiano, Lawrence J. and Terry J. Fitzgerald (2003). "The Band Pass Filter". In: *International Economic Review* 44.2, pp. 435–465. JSTOR: 3663474. URL: https://www.jstor.org/stable/3663474 (visited on 05/07/2024).

Cincotti, Silvano, Marco Raberto, and Andrea Teglio (Dec. 1, 2010). "Credit Money and Macroeconomic Instability in the Agent-based Model and Simulator Eurace". In: *Economics* 4.1, p. 20100026. DOI: 10.5018/economics-ejournal.ja.2010-26.

Clark, John, Christopher Freeman, and Luc Soete (Aug. 1, 1981). "Long Waves, Inventions, and Innovations". In: *Futures* 13.4, pp. 308–322. DOI: 10.1016/0016-3287(81)90146-4.

Cruz, Roberto de la and Jan-Ulrich Kreft (Apr. 4, 2019). *Geometric Mean Extension for Data Sets with Zeros*. DOI: 10.48550/arXiv.1806.06403. arXiv: 1806.06403 [stat]. Pre-published.

Cuturi, Marco (2011). "Fast Global Alignment Kernels". In: Proceedings of the 28th International Conference on Machine Learning. Bellevue, WA, USA.

Cuturi, Marco et al. (Apr. 2007). "A Kernel for Time Series Based on Global Alignments". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pp. II-413-II–416. DOI: 10.1109/ICASSP.2007.366260. arXiv: cs/0610033.

Dawid, Herbert, Simon Gemkow, Philipp Harting, and Michael Neugart (July 1, 2012). "Labor Market Integration Policies and the Convergence of Regions: The Role of Skills and Technology Diffusion". In: *Journal of Evolutionary Economics* 22.3, pp. 543–562. DOI: 10.1007/s00191-011-0245-1.

Dawid, Herbert, Simon Gemkow, Philipp Harting, Sander van der Hoog, et al. (Feb. 22, 2018). "Agent-Based Macroeconomic Modeling and Policy Analysis: The Eurace@ Unibi Model". In: *The Oxford Handbook of Computational Economics and Finance*. Ed. by Shu-Heng Chen, Mak Kaboudan, and Ye-Rong Du. Oxford University Press, pp. 490–519. DOI: 10.1093/oxfordhb/9780199844371.013.19.

Delli Gatti, Domenic, Edoardo Gaffeo, and Mauro Gallegati (Dec. 2010). "Complex Agent-Based Macroeconomics: A Manifesto for a New Paradigm". In: *Journal of Economic Interaction and Coordination* 5.2, pp. 111–135. DOI: 10.1007/s11403-010-0064-8.

Delli Gatti, Domenico and Severin Reissl (2022). "Agent-Based Covid Economics (ABC): Assessing Non-Pharmaceutical Interventions and Macro-Stabilization Policies". In: *Industrial and Corporate Change* 31.2, pp. 410–447. URL: https://ideas.repec.org//a/oup/indcch/v31y2022i2p410-447..html (visited on 05/21/2024).

Dosi, Giovanni (2007). "Statistical Regularities in the Evolution of Industries: A Guide through Some Evidence and Challenges for the Theory". In: *Perspectives on Innovation*. Ed. by Franco Malerba and Stefano Brusoni. Cambridge: Cambridge University Press, pp. 153–186. DOI: 10.1017/CBO9780511618390.009.

— (2023). *The foundations of complex evolving economies: Part one: Innovation, organization, and industrial dynamics*. Oxford University Press.

Dosi, Giovanni, Giorgio Fagiolo, Mauro Napoletano, and Andrea Roventini (Aug. 1, 2013). "Income Distribution, Credit and Fiscal Policies in an Agent-Based Keynesian Model". In: *Journal of Economic Dynamics and Control*. Rethinking Economic Policies in a Landscape of Heterogeneous Agents 37.8, pp. 1598–1625. DOI: 10.1016/j.jedc.2012.11.008.

Dosi, Giovanni, Giorgio Fagiolo, Mauro Napoletano, Andrea Roventini, and Tania Treibich (Mar. 1, 2015). "Fiscal and Monetary Policies in Complex Evolving Economies". In: *Journal of Economic Dynamics and Control* 52, pp. 166–189. DOI: 10.1016/j.jedc.2014.11.014.

Dosi, Giovanni, Giorgio Fagiolo, and Andrea Roventini (May 15, 2006). "An Evolutionary Model of Endogenous Business Cycles". In: *Computational Economics* 27.1, pp. 3–34. DOI: 10.1007/s10614-005-9014-2.

— (Aug. 1, 2008). "The Microfoundations of Business Cycles: An Evolutionary, Multi-Agent Model". In: *Journal of Evolutionary Economics* 18.3, pp. 413–432. DOI: 10.1007/s00191-008-0094-8.

— (Sept. 1, 2010). "Schumpeter Meeting Keynes: A Policy-Friendly Model of Endogenous Growth and Business Cycles". In: *Journal of Economic Dynamics and Control*. Computational Perspectives in Economics and Finance: Methods,Dynamic Analysis and Policy Modeling 34.9, pp. 1748–1767. DOI: 10.1016/j.jedc.2010.06.018.

Dosi, Giovanni, Marcelo C. Pereira, et al. (Aug. 1, 2017). "When More Flexibility Yields More Fragility: The Microfoundations of Keynesian Aggregate Unemployment". In: *Journal of Economic Dynamics and Control*. International Conference "Large-scale Crises: 1929 vs. 2008" 81, pp. 162–186. DOI: 10.1016/j.jedc.2017.02.005.

— (Dec. 1, 2018a). "Causes and Consequences of Hysteresis: Aggregate Demand, Productivity, and Employment". In: *Industrial and Corporate Change* 27.6, pp. 1015–1044. DOI: 10.1093/icc/dty010.

— (Oct. 1, 2018b). "The Effects of Labour Market Reforms upon Unemployment and Income Inequalities: An Agent-Based Model". In: *Socio-Economic Review* 16.4, pp. 687–720. DOI: 10.1093/ser/mwx054.

— (May 1, 2020). "The Labour-Augmented K+S Model: A Laboratory for the Analysis of Institutional and Policy Regimes". In: *EconomiA* 21.2, pp. 160–184. DOI: 10.1016/j.econ.2019.03.002.

— (Dec. 1, 2022). "Technological Paradigms, Labour Creation and Destruction in a Multi-Sector Agent-Based Model". In: *Research Policy* 51.10, p. 104565. DOI: 10.1016/j.respol.2022.104565.

Dosi, Giovanni and Andrea Roventini (Mar. 2019). "More Is Different ... and Complex! The Case for Agent-Based Macroeconomics". In: *Journal of Evolutionary Economics* 29.1, pp. 1–37. DOI: 10.1007/s00191-019-00609-y.

Epstein, Joshua M (Jan. 2007). *Generative Social Science*. Princeton Studies in Complexity. Princeton, NJ: Princeton University Press.

Fagiolo, Giorgio, Mattia Guerini, et al. (2019a). "Validation of Agent-Based Models in Economics and Finance". In: *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*. Ed. by Claus Beisbart and Nicole J. Saam. Cham: Springer International Publishing, pp. 763–787. DOI: 10.1007/978-3-319-70766-2_31.

Fagiolo, Giorgio, Mattia Guerini, et al. (2019b). "Validation of Agent-Based Models in Economics and Finance". In: *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*. Ed. by Claus Beisbart and Nicole J. Saam. Simulation Foundations, Methods and Applications. Cham: Springer International Publishing, pp. 763–787. DOI: 10.1007/978-3-319-70766-2_31.

Fagiolo, Giorgio, Alessio Moneta, and Paul Windrum (Oct. 1, 2007). "A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems". In: *Computational Economics* 30.3, pp. 195–226. DOI: 10.1007/s10614-007-9104-4.

Fagiolo, Giorgio, Mauro Napoletano, and Andrea Roventini (2008). "Are Output Growth-Rate Distributions Fat-Tailed? Some Evidence from OECD Countries". In: *Journal of Applied Econometrics* 23, pp. 639–669.

Fahiman, Fateme et al. (July 2017). "Fuzzy C-Shape: A New Algorithm for Clustering Finite Time Series Waveforms". In: *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Naples, Italy: IEEE, pp. 1–8. DOI: 10.1109/FUZZ-IEEE.2017.8015525.

Franses, Philip Hans and Thomas Wiemann (June 1, 2020). "Intertemporal Similarity of Economic Time Series: An Application of Dynamic Time Warping". In: *Computational Economics* 56.1, pp. 59–75. DOI: 10.1007/s10614-020-09986-0.

Giorgino, Toni (Aug. 14, 2009). "Computing and Visualizing Dynamic Time Warping Alignments in R: The Dtw Package". In: *Journal of Statistical Software* 31, pp. 1–24. DOI: 10.18637/jss.v031.i07.

Granger, C. W. J. (1966). "The Typical Spectral Shape of an Economic Variable". In: *Econometrica* 34.1, pp. 150–161. DOI: 10.2307/1909859. JSTOR: 1909859.

Grimm, Volker et al. (Nov. 11, 2005). "Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology". In: *Science* 310.5750, pp. 987–991. DOI: 10.1126/science.1116681.

Gruber, C. et al. (2010). "Online Signature Verification with Support Vector Machines Based on LCSS Kernel Functions". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 40.4, pp. 1088–1100. DOI: 10.1109/TSMCB.2009.2034382.

Itakura, F. (Feb. 1975). "Minimum Prediction Residual Principle Applied to Speech Recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23.1, pp. 67–72. DOI: 10.1109/TASSP.1975.1162641.

Keogh, Eamonn and Shruti Kasetty (2003). "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration". In: *Data Mining and Knowledge Discovery* 7.4, pp. 349–371. DOI: 10.1023/a:1024988512476.

Keogh, Eamonn and Chotirat Ann Ratanamahatana (Mar. 1, 2005). "Exact Indexing of Dynamic Time Warping". In: *Knowledge and Information Systems* 7.3, pp. 358–386. DOI: 10.1007/s10115-004-0154-9.

Kondratieff, N. D. and W. F. Stolper (1935). "The Long Waves in Economic Life". In: *The Review of Economics and Statistics* 17.6, pp. 105–115. DOI: 10.2307/1928486. JSTOR: 1928486.

Kruskal, Joseph B. and Mark Liberman (1999). "The Symmetric Time Warping Algorithm: From Continuous to Discrete". In: *Time Warps, String Edits, and Macromolecules*. Ed. by David Sankoff and Joseph B. Kruskal. Reissue ed. David Hume Series. Stanford, CA: Center for the Study of Language and Information.

Laeven, L. and F. Valencia (2008). *Systemic Banking Crises: A New Database*. Working paper WP/08/224, International Monetary Fund.

Lamperti, Francesco (Apr. 1, 2018). "Empirical Validation of Simulated Models through the GSL-div: An Illustrative Application". In: *Journal of Economic Interaction and Coordination* 13.1, pp. 143–171. DOI: 10.1007/s11403-017-0206-3.

Lei, Hansheng and Bingyu Sun (Dec. 2007). "A Study on the Dynamic Time Warping in Kernel Machines". In: *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*. 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System SITIS. Shanghai, China: IEEE, pp. 839–845. DOI: 10.1109/SITIS.2007.112.

Lemire, Daniel (Sept. 1, 2009). "Faster Retrieval with a Two-Pass Dynamic-Time-Warping Lower Bound". In: *Pattern Recognition* 42.9, pp. 2169–2180. DOI: 10.1016/j.patcog.2008.11.030.

Leombruni, Roberto and Matteo Richiardi (Sept. 1, 2005). "Why Are Economists Sceptical about Agent-Based Simulations?" In: *Physica A: Statistical Mechanics and its Applications*. Market Dynamics and Quantitative Economics 355.1, pp. 103–109. DOI: 10.1016/j.physa.2005.02.072.

Marks, Robert Ernest (Oct. 1, 2007). "Validating Simulation Models: A General Framework and Four Applied Examples". In: *Computational Economics* 30.3, pp. 265–290. DOI: 10.1007/s10614-007-9101-7.

Martinoli, Mario, Alessio Moneta, and Gianluca Pallante (Dec. 1, 2024). "Calibration and Validation of Macroeconomic Simulation Models by Statistical Causal Search". In: *Journal of Economic Behavior & Organization* 228, p. 106786. DOI: 10.1016/j.jebo.2024.106786.

May, Robert M. (Aug. 1972). "Will a Large Complex System Be Stable?" In: *Nature* 238.5364, pp. 413–414. DOI: 10.1038/238413a0.

McKay, Alisdair and Ricardo Reis (May 2008). "The Brevity and Violence of Contractions and Expansions". In: *Journal of Monetary Economics* 55.4, pp. 738–751. DOI: 10.1016/j.jmoneco.2008.05.009.

Mori, Usue, Alexander Mendiburu, and Jose A. Lozano (2014). *TSdist: Distance Measures for Time Series Data*. Version 3.7.1. Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=TSdist (visited on 06/25/2024).

Napoletano, Mauro, Andrea Roventini, and S. Sapio (2006). "Are Business Cycles All Alike? A Bandpass Filter Analysis of the Italian and U.S. Cycles". In: *Rivista Italiana degli Economisti* 1, pp. 87–118.

Paparrizos, John and Luis Gravano (May 27, 2015). "K-Shape: Efficient and Accurate Clustering of Time Series". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD/PODS'15: International Conference on Management of Data. Melbourne Victoria Australia: ACM, pp. 1855–1870. DOI: 10.1145/2723372.2737793.

Pereira, Marcelo C. (2022). *LSDinterface: Interface Tools for LSD Simulation Results Files*. Version 1.2.1. Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=LSDinterface (visited on 06/24/2024).

Perez, Carlota (July 1, 2009). "The Double Bubble at the Turn of the Century: Technological Roots and Structural Implications". In: *Cambridge Journal of Economics* 33.4, pp. 779–805. DOI: 10.1093/cje/bep028.

Poledna, Sebastian et al. (Jan. 1, 2023). "Economic Forecasting with an Agent-Based Model". In: *European Economic Review* 151, p. 104306. DOI: 10.1016/j.euroecorev.2022.104306.

Raihan, Tasneem (2017). "Predicting US Recessions: A Dynamic Time Warping Exercise in Economics".

Rath, T.M. and R. Manmatha (2003). "Word Image Matching Using Dynamic Time Warping". In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* CVPR 2003: Computer Vision and Pattern Recognition Conference. Vol. 2. Madison, WI, USA: IEEE Comput. Soc, pp. II-521-II–527. DOI: 10.1109/CVPR.2003.1211511.

Reinhart, C. and K. Rogoff (2009). "The Aftermath of Financial Crises". In: *American Economic Review* 99, pp. 466–472.

Sakoe, H. and S. Chiba (Feb. 1978). "Dynamic Programming Algorithm Optimization for Spoken Word Recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1, pp. 43–49. DOI: 10.1109/TASSP.1978.1163055.

Sardá-Espinosa, Alexis (2019). "Time-Series Clustering in R Using the Dtwclust Package". In: *The R Journal* 11.1, pp. 22–43. URL: https://journal.r-project.org/archive/2019/RJ-2019-023/index.html (visited on 04/26/2024).

Shannon, Claude E (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3, pp. 379–423.

Shyu, Shyong Jian and Chun-Yuan Tsai (Jan. 1, 2009). "Finding the Longest Common Subsequence for Multiple Biological Sequences by Ant Colony Optimization". In: *Computers & Operations*

Research. Part Special Issue: Operations Research Approaches for Disaster Recovery Planning 36.1, pp. 73–91. DOI: 10.1016/j.cor.2007.07.006.

Silverberg, G. (July 26, 2007). "Long Waves: Conceptual, Empirical and Modelling Issues". In: *Elgar Companion to Neo-Schumpeterian Economics*. Ed. by Horst Hanusch and Andreas Pyka. Edward Elgar Publishing. DOI: 10.4337/9781847207012.00060.

Stock, J. H. and Mark Watson (Jan. 1, 1999). "Chapter 1 Business Cycle Fluctuations in Us Macroeconomic Time Series". In: *Handbook of Macroeconomics*. Vol. 1. Elsevier, pp. 3–64. DOI: 10.1016/S1574-0048(99)01004-6.

Valente, Marco and Marcelo C. Pereira (2023). *Laboratory for Simulation Develpment - LSD*. Version 8.1. URL: https://labsimdev.org (visited on 07/22/2022).

Vlachos, M., G. Kollios, and D. Gunopulos (2002). "Discovering Similar Multidimensional Trajectories". In: *Proceedings 18th International Conference on Data Engineering*. 18th International Conference on Data Engineering. San Jose, CA, USA: IEEE Comput. Soc, pp. 673–684. DOI: 10.1109/ICDE.2002.994784.

Wagner, R.A. and M.J. Fischer (1974). "The String-to-String Correction Problem". In: *Journal of the ACM (JACM)* 21.1, pp. 168–173. DOI: 10.1145/321796.321811.

Wang, Gang-Jin et al. (Aug. 15, 2012). "Similarity Measure and Topology Evolution of Foreign Exchange Markets Using Dynamic Time Warping Method: Evidence from Minimal Spanning Tree". In: *Physica A: Statistical Mechanics and its Applications* 391.16, pp. 4136–4146. DOI: 10.1016/j.physa.2012.03.036.

Wang, Xiaoyue et al. (Mar. 1, 2013). "Experimental Comparison of Representation Methods and Distance Measures for Time Series Data". In: *Data Mining and Knowledge Discovery* 26.2, pp. 275–309. DOI: 10.1007/s10618-012-0250-5.

Windrum, Paul, Giorgio Fagiolo, and Alessio Moneta (2007). "Empirical Validation of Agent-Based Models: Alternatives and Prospects". In: *Journal of Artificial Societies and Social Simulation* 10.2, p. 8. URL: https://www.jasss.org/10/2/8.html.

A Dissimilarity across all model versions, variables, and filter

Table A.1: Dissimilarity across different K+S versions, variables, and filters

Metrics	Unfiltered					Trend					Cycle				
	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
dA															
Statistical	1.574 (5)	1.163 (3)	1.326 (4)	**0.886** (1)	1.077 (2)	3.399 (5)	**1.917** (1)	3.108 (4)	2.167 (2)	2.834 (3)	1.021 (4)	**0.486** (1)	0.912 (3)	1.141 (5)	0.589 (2)
LCSS	0.025 (4)	**0.022** (1)	0.024 (3)	0.022 (2)	0.026 (5)	0.021 (4)	0.019 (2)	0.020 (3)	**0.012** (1)	0.022 (5)	**0.083** (1)	0.089 (3)	0.113 (5)	0.107 (4)	0.087 (2)
DTW	0.003 (3)	**0.003** (1)	0.004 (4)	0.006 (5)	0.003 (2)	**0.002** (1)	0.002 (3)	0.002 (2)	0.003 (5)	0.002 (4)	**0.357** (1)	0.449 (4)	0.430 (3)	0.415 (2)	0.451 (5)
TGAK	0.018 (3)	**0.016** (1)	0.019 (4)	0.025 (5)	0.017 (2)	0.013 (2)	0.013 (3)	0.014 (4)	**0.013** (1)	0.014 (5)	0.021 (3)	**0.014** (1)	0.025 (4)	0.020 (2)	0.026 (5)
SBD	0.367 (3)	0.370 (4)	**0.357** (1)	0.438 (5)	0.359 (2)	0.321 (2)	0.351 (3)	**0.317** (1)	0.434 (5)	0.364 (4)	**0.267** (1)	0.272 (3)	0.293 (4)	0.344 (5)	0.269 (2)
Geometric Average	0.284 (5)	**7.4e-04** (1)	0.051 (3)	0.030 (2)	0.137 (4)	0.029 (2)	0.048 (3)	0.054 (4)	**0.007** (1)	0.529 (5)	0.007 (2)	**0.002** (1)	0.686 (4)	0.741 (5)	0.222 (3)
Entropy	0.590 (2)	0.590 (2)	0.590 (2)	0.590 (2)	**0.311** (1)	0.828 (4)	**0.590** (1)	0.828 (4)	0.655 (2)	0.655 (2)	0.590 (2)	0.655 (3)	0.655 (3)	0.655 (3)	**0.418** (1)
Entropy adj Geom. Avg.	0.451 (5)	**0.001** (1)	0.081 (3)	0.048 (2)	0.179 (4)	0.053 (2)	0.077 (3)	0.099 (4)	**0.011** (1)	0.875 (5)	0.011 (2)	**0.003** (1)	1.136 (4)	1.227 (5)	0.315 (3)
dC															
Statistical	1.106 (3)	0.866 (2)	1.910 (5)	1.315 (4)	**0.619** (1)	1.033 (4)	0.736 (2)	1.952 (5)	0.795 (3)	**0.567** (1)	0.832 (3)	**0.475** (1)	0.528 (2)	0.833 (4)	0.869 (5)
LCSS	0.006 (3)	**0.006** (1)	0.009 (4)	0.015 (5)	0.006 (2)	0.005 (4)	0.005 (3)	0.008 (5)	0.005 (2)	**0.005** (1)	0.047 (2)	0.058 (4)	0.049 (3)	**0.046** (1)	0.062 (5)
DTW	0.004 (3)	0.003 (2)	0.004 (4)	0.008 (5)	**0.003** (1)	0.003 (3)	0.003 (2)	0.003 (4)	0.003 (5)	**0.002** (1)	**0.335** (1)	0.450 (5)	0.416 (2)	0.438 (3)	0.439 (4)
TGAK	0.008 (4)	**0.007** (1)	0.008 (3)	0.013 (5)	0.007 (2)	**0.004** (1)	0.004 (2)	0.005 (4)	0.005 (5)	0.004 (3)	0.011 (5)	0.008 (2)	0.009 (3)	0.011 (4)	**0.007** (1)
SBD	0.429 (4)	**0.409** (1)	0.413 (3)	0.410 (2)	0.441 (5)	**0.390** (1)	0.425 (4)	0.406 (2)	0.421 (3)	0.435 (5)	0.320 (4)	0.295 (2)	0.317 (3)	**0.284** (1)	0.334 (5)
Geometric Average	0.201 (3)	**6.1e-06** (1)	0.233 (4)	0.451 (5)	9.7e-04 (2)	0.003 (2)	0.160 (3)	0.700 (5)	0.284 (4)	**0.003** (1)	**0.048** (1)	0.057 (2)	0.369 (5)	0.155 (3)	0.364 (4)
Entropy	**0.418** (1)	**0.418** (1)	0.655 (4)	0.590 (3)	0.655 (4)	0.655 (3)	**0.590** (1)	0.655 (3)	0.655 (3)	**0.590** (1)	1.000 (5)	0.828 (4)	**0.418** (1)	0.655 (3)	0.590 (2)
Entropy adj Geom. Avg.	0.285 (3)	**8.7e-06** (1)	0.385 (4)	0.717 (5)	0.002 (2)	0.006 (2)	0.255 (3)	1.158 (5)	0.471 (4)	**0.005** (1)	**0.096** (1)	0.105 (2)	0.524 (4)	0.257 (3)	0.580 (5)
dCPI															
Statistical	1.163 (2)	1.311 (5)	1.278 (3)	**0.997** (1)	1.298 (4)	1.298 (4)	1.254 (3)	0.894 (2)	3.102 (5)	**0.618** (1)	**0.648** (1)	0.798 (2)	1.159 (3)	1.415 (4)	1.745 (5)
LCSS	0.033 (4)	0.078 (5)	0.014 (2)	**0.014** (1)	0.015 (3)	0.008 (4)	**0.006** (1)	0.008 (3)	0.018 (5)	0.007 (2)	0.050 (2)	0.060 (5)	0.056 (4)	**0.047** (1)	0.054 (3)
DTW	0.009 (3)	0.016 (5)	**0.005** (1)	0.011 (4)	0.006 (2)	0.003 (4)	**0.003** (1)	0.003 (3)	0.023 (5)	0.003 (2)	0.513 (2)	0.541 (4)	0.518 (3)	1.085 (5)	**0.482** (1)
TGAK	0.015 (4)	0.017 (5)	0.013 (2)	0.014 (3)	**0.013** (1)	0.009 (4)	0.008 (2)	0.009 (5)	**0.006** (1)	0.009 (3)	0.011 (3)	0.013 (5)	0.010 (2)	**0.010** (1)	0.011 (4)
SBD	0.385 (2)	0.417 (3)	0.451 (5)	**0.376** (1)	0.445 (4)	**0.408** (1)	0.422 (2)	0.453 (4)	0.439 (3)	0.458 (5)	0.327 (3)	0.309 (2)	0.364 (5)	**0.294** (1)	0.361 (4)
Geometric Average	0.300 (4)	0.883 (5)	0.006 (2)	**0.003** (1)	0.008 (3)	0.014 (3)	**0.008** (1)	0.179 (5)	0.169 (4)	0.014 (2)	0.023 (2)	0.311 (4)	0.318 (5)	**0.008** (1)	0.091 (3)
Entropy	0.655 (3)	**0.311** (1)	0.828 (4)	0.590 (2)	0.828 (4)	**0.311** (1)	0.655 (3)	0.828 (4)	0.590 (2)	0.828 (4)	0.655 (2)	0.655 (2)	0.828 (4)	**0.590** (1)	0.828 (4)
Entropy adj Geom. Avg.	0.497 (4)	1.157 (5)	0.011 (2)	**0.004** (1)	0.015 (3)	0.018 (2)	**0.012** (1)	0.327 (5)	0.268 (4)	0.025 (3)	0.038 (2)	0.514 (4)	0.581 (5)	**0.013** (1)	0.166 (3)
DebGDP															
Statistical	1.898 (2)	9.839 (5)	4.326 (4)	2.171 (3)	**0.469** (1)	34.886 (5)	1.365 (3)	**0.793** (1)	2.269 (4)	1.074 (2)	3.905 (5)	**0.681** (1)	1.775 (3)	2.469 (4)	0.876 (2)
LCSS	1.000 (2)	1.000 (2)	1.000 (2)	1.000 (2)	**0.016** (1)	0.040 (5)	0.021 (4)	**0.009** (1)	0.010 (2)	0.012 (3)	0.060 (3)	**0.056** (1)	0.060 (4)	0.070 (5)	0.058 (2)
DTW	0.540 (2)	146.100 (5)	0.555 (3)	3.798 (4)	**0.091** (1)	0.246 (5)	0.058 (4)	0.007 (2)	0.012 (3)	**0.006** (1)	0.848 (5)	**0.448** (1)	0.459 (2)	0.547 (4)	0.468 (3)
TGAK	0.022 (2)	0.024 (5)	0.022 (3)	0.023 (4)	**0.001** (1)	0.010 (5)	**0.005** (1)	0.008 (3)	0.005 (2)	0.009 (4)	0.011 (2)	0.012 (3)	0.016 (5)	0.014 (4)	**0.010** (1)
SBD	0.138 (3)	0.163 (4)	0.250 (5)	0.053 (2)	**0.037** (1)	0.412 (2)	0.424 (5)	**0.411** (1)	0.417 (4)	0.413 (3)	0.370 (4)	0.371 (5)	0.271 (2)	0.325 (3)	**0.265** (1)
Geometric Average	0.183 (2)	0.901 (5)	0.261 (4)	0.203 (3)	**0.0e+00** (1)	0.586 (5)	0.016 (3)	**2.4e-05** (1)	0.070 (4)	0.009 (2)	0.518 (4)	**0.001** (1)	0.175 (3)	0.555 (5)	0.002 (2)
Entropy	0.590 (3)	0.311 (2)	0.655 (4)	0.828 (5)	**-0.0e+00** (1)	**0.311** (1)	0.828 (4)	0.590 (2)	0.655 (3)	0.828 (4)	0.828 (4)	**0.590** (1)	0.828 (4)	**0.590** (1)	0.655 (3)
Entropy adj Geom. Avg.	0.291 (2)	1.181 (5)	0.432 (4)	0.371 (3)	**0.0e+00** (1)	0.769 (5)	0.029 (3)	**3.8e-05** (1)	0.116 (4)	0.016 (2)	0.946 (5)	**0.002** (1)	0.319 (3)	0.882 (4)	0.003 (2)
dGDP															
Statistical	1.690 (5)	1.195 (2)	1.532 (4)	1.449 (3)	**0.707** (1)	2.365 (5)	1.507 (3)	1.571 (4)	0.941 (2)	**0.918** (1)	1.088 (4)	1.439 (5)	1.059 (3)	0.542 (2)	**0.412** (1)
LCSS	0.009 (4)	**0.006** (1)	0.008 (3)	0.010 (5)	0.007 (2)	0.008 (5)	**0.005** (1)	0.006 (3)	0.006 (2)	0.007 (4)	0.050 (2)	**0.048** (1)	0.059 (4)	0.052 (3)	0.065 (5)
DTW	0.004 (2)	0.004 (3)	0.004 (4)	0.005 (5)	**0.004** (1)	0.003 (5)	**0.003** (1)	0.003 (3)	0.003 (2)	0.003 (4)	0.469 (4)	0.458 (3)	**0.376** (1)	0.445 (2)	0.564 (5)
TGAK	0.007 (3)	0.007 (2)	0.007 (4)	0.009 (5)	**0.007** (1)	**0.005** (1)	0.005 (2)	0.005 (4)	0.005 (5)	0.005 (3)	0.012 (4)	0.008 (2)	0.012 (5)	0.010 (3)	**0.008** (1)

Table A.1: Dissimilarity across different K+S versions, variables, and filters *(continued)*

	Unfiltered					Trend					Cycle				
Metrics	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
SBD	0.411 (2)	0.422 (3)	**0.398** (1)	0.430 (4)	0.434 (5)	**0.385** (1)	0.404 (4)	0.391 (3)	0.409 (5)	0.389 (2)	0.316 (2)	0.341 (4)	**0.273** (1)	0.357 (5)	0.316 (3)
Geometric Average	0.255 (4)	0.010 (2)	0.043 (3)	0.923 (5)	**0.001** (1)	0.467 (5)	**0.003** (1)	0.405 (4)	0.244 (3)	0.049 (2)	0.432 (5)	0.062 (3)	0.026 (2)	0.335 (4)	**0.022** (1)
Entropy	0.828 (5)	0.655 (4)	**0.590** (1)	**0.590** (1)	**0.590** (1)	**0.418** (1)	0.828 (4)	**0.418** (1)	**0.418** (1)	0.828 (4)	**0.418** (1)	1.000 (5)	0.828 (4)	0.655 (2)	0.655 (2)
Entropy adj Geom. Avg.	0.466 (4)	0.017 (2)	0.069 (3)	1.468 (5)	**0.002** (1)	0.662 (5)	**0.005** (1)	0.574 (4)	0.347 (3)	0.089 (2)	0.612 (5)	0.124 (3)	0.048 (2)	0.555 (4)	**0.036** (1)
dGreal															
Statistical	**2.013** (1)	5.395 (3)	8.931 (5)	7.605 (4)	2.980 (2)	0.801 (4)	0.570 (2)	0.724 (3)	**0.561** (1)	1.360 (5)	**0.460** (1)	1.038 (4)	1.158 (5)	0.639 (2)	0.859 (3)
LCSS	1.000 (3)	0.129 (2)	1.000 (3)	1.000 (3)	**0.121** (1)	**0.014** (1)	0.018 (4)	0.014 (2)	0.016 (3)	0.026 (5)	0.051 (2)	0.054 (4)	0.052 (3)	0.054 (5)	**0.050** (1)
DTW	0.633 (2)	0.846 (3)	9.664 (5)	3.480 (4)	**0.410** (1)	**0.029** (1)	0.044 (3)	0.031 (2)	0.046 (4)	0.083 (5)	0.726 (4)	0.764 (5)	0.661 (2)	**0.660** (1)	0.718 (3)
TGAK	0.018 (5)	**0.012** (1)	0.018 (4)	0.016 (3)	0.016 (2)	**0.003** (1)	0.003 (4)	0.003 (3)	0.003 (2)	0.004 (5)	0.011 (5)	**0.009** (1)	0.010 (4)	0.010 (3)	0.009 (2)
SBD	**0.390** (1)	0.432 (3)	0.410 (2)	0.432 (4)	0.434 (5)	**0.333** (1)	0.404 (4)	0.382 (3)	0.409 (5)	0.368 (2)	**0.268** (1)	0.340 (3)	0.358 (5)	0.357 (4)	0.330 (2)
Geometric Average	**0.001** (1)	0.008 (3)	0.844 (5)	0.690 (4)	0.005 (2)	**0.037** (1)	0.219 (4)	0.135 (3)	0.043 (2)	0.858 (5)	**0.008** (1)	0.248 (4)	0.329 (5)	0.066 (3)	0.032 (2)
Entropy	0.655 (4)	**0.590** (1)	**0.590** (1)	**0.590** (1)	0.655 (4)	**0.311** (1)	0.590 (4)	0.418 (3)	1.000 (5)	**0.311** (1)	0.828 (2)	0.828 (2)	0.828 (2)	1.000 (5)	**0.655** (1)
Entropy adj Geom. Avg.	**0.002** (1)	0.013 (3)	1.342 (5)	1.098 (4)	0.009 (2)	**0.048** (1)	0.348 (4)	0.192 (3)	0.087 (2)	1.124 (5)	**0.015** (1)	0.453 (4)	0.602 (5)	0.132 (3)	0.053 (2)
dI															
Statistical	**1.140** (1)	2.006 (4)	2.412 (5)	1.527 (2)	1.985 (3)	0.891 (4)	0.992 (5)	0.846 (3)	0.810 (2)	**0.638** (1)	0.613 (2)	**0.470** (1)	0.745 (3)	0.962 (5)	0.822 (4)
LCSS	**0.011** (1)	0.053 (5)	0.040 (4)	0.032 (3)	0.016 (2)	0.010 (5)	0.010 (4)	0.009 (2)	**0.009** (1)	0.009 (3)	0.048 (2)	0.048 (3)	**0.040** (1)	0.053 (5)	0.053 (4)
DTW	**0.019** (1)	0.074 (5)	0.061 (4)	0.050 (3)	0.026 (2)	0.012 (2)	0.013 (5)	**0.012** (1)	0.012 (4)	0.012 (3)	0.449 (4)	0.399 (2)	0.408 (3)	**0.396** (1)	0.475 (5)
TGAK	**0.008** (1)	0.015 (5)	0.015 (4)	0.014 (3)	0.010 (2)	0.004 (2)	**0.004** (1)	0.005 (5)	0.005 (3)	0.005 (4)	0.009 (2)	0.009 (4)	0.012 (5)	0.009 (3)	**0.007** (1)
SBD	0.430 (2)	**0.416** (1)	0.439 (4)	0.433 (3)	0.444 (5)	0.431 (5)	0.416 (4)	0.411 (3)	**0.394** (1)	0.408 (2)	**0.253** (1)	0.353 (3)	0.268 (2)	0.360 (4)	0.378 (5)
Geometric Average	**0.060** (1)	0.197 (2)	0.838 (5)	0.536 (4)	0.317 (3)	0.444 (5)	0.146 (4)	0.018 (2)	**0.007** (1)	0.031 (3)	0.077 (3)	**0.024** (1)	0.035 (2)	0.127 (4)	0.199 (5)
Entropy	**0.311** (1)	0.590 (4)	**0.311** (1)	**0.311** (1)	0.590 (4)	**0.655** (1)	**0.655** (1)	0.828 (3)	0.828 (3)	0.828 (3)	**0.590** (1)	0.828 (3)	0.828 (3)	0.828 (3)	0.655 (2)
Entropy adj Geom. Avg.	**0.078** (1)	0.314 (2)	1.099 (5)	0.702 (4)	0.504 (3)	0.735 (5)	0.242 (4)	0.032 (2)	**0.013** (1)	0.056 (3)	0.122 (3)	**0.044** (1)	0.064 (2)	0.232 (4)	0.329 (5)
dPPI															
Statistical	1.794 (5)	1.740 (4)	**1.024** (1)	1.038 (2)	1.056 (3)	1.770 (2)	1.915 (4)	1.780 (3)	3.902 (5)	**0.806** (1)	3.384 (4)	**1.773** (1)	2.806 (3)	4.057 (5)	2.481 (2)
LCSS	0.026 (4)	0.070 (5)	**0.006** (1)	0.012 (3)	0.007 (2)	0.008 (3)	**0.007** (1)	0.008 (4)	0.007 (2)	0.009 (5)	0.069 (4)	**0.052** (1)	0.052 (2)	0.069 (5)	0.052 (2)
DTW	0.013 (4)	0.020 (5)	**0.004** (1)	0.005 (3)	0.004 (2)	0.004 (3)	**0.004** (1)	0.004 (2)	0.009 (5)	0.004 (4)	**0.506** (1)	0.599 (2)	0.663 (4)	1.036 (5)	0.607 (3)
TGAK	0.014 (4)	0.016 (5)	**0.009** (1)	0.009 (3)	0.009 (2)	0.009 (4)	0.008 (2)	0.008 (3)	**0.008** (1)	0.009 (5)	0.010 (3)	0.011 (4)	**0.009** (1)	0.009 (2)	0.012 (5)
SBD	**0.436** (1)	0.440 (3)	0.439 (2)	0.441 (4)	0.451 (5)	0.453 (5)	0.445 (3)	0.446 (4)	**0.394** (1)	0.432 (2)	0.317 (3)	**0.239** (1)	0.338 (4)	0.257 (2)	0.345 (5)
Geometric Average	0.086 (4)	0.772 (5)	**0.028** (1)	0.070 (3)	0.043 (2)	0.364 (5)	**0.009** (1)	0.293 (4)	0.010 (2)	0.034 (3)	0.113 (3)	**7.1e-04** (1)	0.021 (2)	0.388 (5)	0.283 (4)
Entropy	0.590 (2)	0.590 (2)	**0.311** (1)	0.590 (2)	0.590 (2)	0.828 (3)	0.828 (3)	**0.655** (1)	**0.655** (1)	0.828 (3)	0.655 (3)	0.590 (2)	0.655 (3)	**0.418** (1)	0.655 (3)
Entropy adj Geom. Avg.	0.137 (4)	1.228 (5)	**0.037** (1)	0.111 (3)	0.068 (2)	0.666 (5)	**0.016** (1)	0.486 (4)	0.017 (2)	0.062 (3)	0.187 (3)	**0.001** (1)	0.034 (2)	0.551 (5)	0.469 (4)
dQ1															
Statistical	**1.646** (1)	2.483 (3)	3.603 (5)	2.427 (2)	2.852 (4)	0.611 (3)	**0.244** (1)	0.479 (2)	1.492 (5)	1.224 (4)	0.759 (3)	**0.503** (1)	0.970 (5)	0.756 (2)	0.935 (4)
LCSS	**0.020** (1)	1.000 (3)	1.000 (3)	1.000 (3)	0.041 (2)	0.009 (2)	0.011 (4)	**0.009** (1)	0.011 (5)	0.010 (3)	**0.054** (1)	0.073 (4)	0.062 (3)	0.057 (2)	0.076 (5)
DTW	**0.020** (1)	0.107 (3)	0.154 (5)	0.123 (4)	0.044 (2)	**0.006** (1)	0.011 (4)	0.007 (2)	0.014 (5)	0.007 (3)	**0.415** (1)	0.458 (4)	0.497 (5)	0.449 (3)	0.439 (2)
TGAK	**0.018** (1)	0.026 (3)	0.027 (5)	0.026 (4)	0.023 (2)	0.008 (3)	0.007 (2)	0.012 (5)	**0.007** (1)	0.008 (4)	**0.010** (1)	0.013 (4)	0.018 (5)	0.013 (3)	0.013 (2)
SBD	**0.416** (1)	0.419 (2)	0.420 (3)	0.427 (4)	0.443 (5)	0.416 (4)	0.414 (3)	**0.307** (1)	0.413 (2)	0.426 (5)	0.341 (3)	0.380 (5)	**0.199** (1)	0.298 (2)	0.367 (4)
Geometric Average	**0.0e+00** (1)	0.496 (3)	0.691 (5)	0.655 (4)	0.267 (2)	0.019 (2)	0.030 (3)	**0.002** (1)	0.318 (4)	0.445 (5)	**0.006** (1)	0.091 (2)	0.110 (3)	0.352 (4)	0.607 (5)
Entropy	**-0.0e+00** (1)	0.590 (3)	0.311 (2)	0.590 (3)	0.590 (3)	0.828 (4)	0.828 (4)	0.655 (2)	**0.590** (1)	0.655 (2)	**0.418** (1)	0.590 (3)	0.590 (3)	**0.418** (1)	0.655 (5)
Entropy adj Geom. Avg.	**0.0e+00** (1)	0.789 (3)	0.906 (4)	1.042 (5)	0.424 (2)	0.035 (2)	0.055 (3)	**0.003** (1)	0.505 (4)	0.737 (5)	**0.009** (1)	0.145 (2)	0.175 (3)	0.499 (4)	1.004 (5)
dQcBas															

Table A.1: Dissimilarity across different K+S versions, variables, and filters *(continued)*

Metrics	Unfiltered					Trend					Cycle				
	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
Statistical	3.444 (5)	2.490 (3)	3.033 (4)	**1.917** (1)	2.154 (2)	2.007 (5)	0.887 (2)	1.444 (4)	1.196 (3)	**0.801** (1)	0.840 (3)	1.108 (5)	**0.700** (1)	0.727 (2)	0.879 (4)
LCSS	0.018 (3)	**0.010** (1)	0.019 (4)	0.019 (5)	0.013 (2)	0.013 (5)	0.009 (2)	0.012 (4)	**0.008** (1)	0.009 (3)	0.074 (5)	0.062 (4)	0.060 (2)	**0.056** (1)	0.060 (2)
DTW	0.005 (3)	**0.004** (1)	0.005 (4)	0.008 (5)	0.005 (2)	**0.003** (1)	0.003 (3)	0.003 (4)	0.003 (5)	0.003 (2)	0.438 (3)	**0.371** (1)	0.398 (2)	0.540 (5)	0.443 (4)
TGAK	0.013 (3)	**0.011** (1)	0.014 (4)	0.017 (5)	0.012 (2)	0.007 (2)	**0.007** (1)	0.007 (4)	0.007 (3)	0.007 (5)	0.015 (4)	**0.009** (1)	0.010 (2)	0.018 (5)	0.013 (3)
SBD	**0.332** (1)	0.398 (4)	0.388 (2)	0.397 (3)	0.419 (5)	**0.314** (1)	0.406 (5)	0.348 (2)	0.393 (3)	0.394 (4)	**0.217** (1)	0.320 (3)	0.327 (4)	0.267 (2)	0.356 (5)
Geometric Average	0.050 (2)	**0.002** (1)	0.536 (5)	0.232 (3)	0.244 (4)	**0.007** (1)	0.012 (2)	0.424 (5)	0.064 (4)	0.019 (3)	0.081 (4)	**0.013** (1)	0.033 (2)	0.037 (3)	0.463 (5)
Entropy	0.590 (3)	0.590 (3)	**0.311** (1)	0.590 (3)	**0.311** (1)	0.655 (3)	0.828 (4)	**0.311** (1)	0.590 (2)	1.000 (5)	0.828 (3)	0.828 (3)	0.828 (3)	**0.655** (1)	**0.655** (1)
Entropy adj Geom. Avg.	0.080 (2)	**0.004** (1)	0.703 (5)	0.369 (4)	0.320 (3)	**0.011** (1)	0.021 (2)	0.556 (5)	0.102 (4)	0.039 (3)	0.149 (4)	**0.023** (1)	0.060 (2)	0.061 (3)	0.766 (5)
dRDreal															
Statistical	**0.416** (1)	0.934 (5)	0.919 (3)	0.932 (4)	0.749 (2)	0.859 (5)	0.785 (4)	0.518 (2)	**0.423** (1)	0.726 (3)	0.832 (3)	**0.411** (1)	0.800 (2)	0.842 (4)	1.163 (5)
LCSS	**0.008** (1)	0.050 (3)	0.057 (4)	0.058 (5)	0.012 (2)	**0.007** (1)	0.008 (4)	0.007 (2)	0.008 (5)	0.007 (3)	0.062 (5)	0.058 (4)	0.049 (2)	**0.041** (1)	0.056 (3)
DTW	**0.007** (1)	0.038 (4)	0.042 (5)	0.036 (3)	0.018 (2)	**0.005** (1)	0.007 (5)	0.005 (2)	0.007 (4)	0.006 (3)	0.465 (4)	0.475 (5)	**0.448** (1)	0.455 (2)	0.458 (3)
TGAK	**0.008** (1)	0.015 (4)	0.016 (5)	0.015 (3)	0.012 (2)	0.006 (5)	**0.005** (1)	0.006 (4)	0.005 (2)	0.005 (3)	**0.009** (1)	0.010 (2)	0.010 (5)	0.010 (4)	0.010 (3)
SBD	0.421 (3)	0.436 (4)	**0.406** (1)	0.447 (5)	0.420 (2)	0.418 (2)	0.429 (4)	0.420 (3)	0.442 (5)	**0.411** (1)	0.351 (2)	0.386 (5)	**0.327** (1)	0.351 (3)	0.356 (4)
Geometric Average	**0.042** (1)	0.870 (4)	0.594 (3)	0.950 (5)	0.312 (2)	**0.010** (1)	0.174 (4)	0.281 (5)	0.089 (3)	0.059 (2)	0.101 (3)	0.129 (4)	**0.016** (1)	0.068 (2)	0.612 (5)
Entropy	0.311 (2)	0.590 (3)	0.828 (5)	0.655 (4)	**-0.0e+00** (1)	0.655 (4)	0.590 (2)	0.590 (2)	0.828 (5)	**0.311** (1)	1.000 (5)	0.828 (3)	0.655 (2)	0.828 (3)	**0.590** (1)
Entropy adj Geom. Avg.	**0.055** (1)	1.383 (4)	1.085 (3)	1.573 (5)	0.312 (2)	**0.016** (1)	0.277 (4)	0.446 (5)	0.163 (3)	0.077 (2)	0.201 (3)	0.235 (4)	**0.027** (1)	0.124 (2)	0.974 (5)
dwReal															
Statistical	2.294 (3)	2.301 (4)	1.666 (2)	**1.362** (1)	2.643 (5)	1.604 (3)	2.170 (5)	1.239 (2)	**0.893** (1)	2.091 (4)	0.958 (3)	**0.754** (1)	0.893 (2)	1.001 (4)	1.727 (5)
LCSS	0.011 (3)	**0.010** (1)	0.012 (4)	0.018 (5)	0.011 (2)	0.006 (2)	0.008 (5)	0.006 (3)	**0.005** (1)	0.007 (4)	0.060 (5)	0.046 (2)	**0.042** (1)	0.058 (4)	0.051 (3)
DTW	0.004 (3)	**0.003** (1)	0.005 (4)	0.008 (5)	0.004 (2)	**0.002** (1)	0.002 (4)	0.002 (2)	0.003 (5)	0.002 (3)	0.475 (4)	0.524 (5)	**0.426** (1)	0.457 (3)	0.441 (2)
TGAK	0.008 (3)	**0.005** (1)	0.009 (4)	0.013 (5)	0.007 (2)	0.004 (2)	**0.004** (1)	0.004 (4)	0.004 (5)	0.004 (3)	0.010 (4)	0.010 (5)	**0.008** (1)	0.009 (3)	0.009 (2)
SBD	0.407 (4)	**0.390** (1)	0.395 (2)	0.442 (5)	0.402 (3)	0.401 (4)	0.377 (2)	**0.370** (1)	0.431 (5)	0.381 (3)	0.315 (2)	**0.263** (1)	0.333 (3)	0.369 (5)	0.368 (4)
Geometric Average	0.323 (4)	**0.090** (1)	0.259 (3)	0.704 (5)	0.236 (2)	0.061 (3)	0.045 (2)	**0.028** (1)	0.467 (5)	0.420 (4)	0.517 (4)	0.010 (2)	**5.9e-04** (1)	0.521 (5)	0.455 (3)
Entropy	**0.311** (1)	**0.311** (1)	0.418 (4)	**0.311** (1)	0.590 (5)	0.828 (3)	0.828 (3)	0.828 (3)	**0.418** (1)	**0.418** (1)	0.828 (4)	0.655 (2)	**0.590** (1)	0.655 (2)	0.828 (4)
Entropy adj Geom. Avg.	0.424 (4)	**0.118** (1)	0.367 (2)	0.923 (5)	0.376 (3)	0.111 (3)	0.081 (2)	**0.051** (1)	0.662 (5)	0.595 (4)	0.945 (5)	0.016 (2)	**9.4e-04** (1)	0.863 (4)	0.832 (3)
QcUbas															
Statistical	0.512 (3)	0.452 (2)	0.891 (5)	0.842 (4)	**0.259** (1)	0.695 (3)	**0.638** (1)	0.698 (4)	0.672 (2)	0.727 (5)	**0.279** (1)	0.369 (3)	0.283 (2)	0.482 (5)	0.381 (4)
LCSS	1.000 (3)	**0.155** (1)	1.000 (3)	1.000 (3)	0.375 (2)	0.163 (3)	**0.030** (1)	0.171 (4)	1.000 (5)	0.035 (2)	0.012 (5)	0.011 (2)	0.011 (4)	0.011 (3)	**0.010** (1)
DTW	0.171 (4)	**0.083** (1)	0.167 (3)	0.395 (5)	0.091 (2)	0.162 (4)	**0.079** (1)	0.157 (3)	0.377 (5)	0.086 (2)	0.019 (5)	0.016 (2)	0.019 (4)	0.016 (3)	**0.016** (1)
TGAK	0.025 (3)	**0.022** (1)	0.025 (4)	0.026 (5)	0.023 (2)	0.023 (4)	**0.018** (1)	0.023 (3)	0.025 (5)	0.019 (2)	0.014 (4)	0.014 (3)	0.014 (5)	**0.013** (1)	0.013 (2)
SBD	0.385 (5)	0.233 (2)	0.311 (4)	0.267 (3)	**0.192** (1)	0.349 (3)	**0.307** (1)	0.352 (5)	0.342 (2)	0.350 (4)	0.370 (3)	0.378 (5)	0.370 (2)	0.374 (4)	**0.367** (1)
Geometric Average	0.615 (3)	0.003 (2)	0.682 (4)	0.816 (5)	**9.2e-04** (1)	0.443 (3)	**0.0e+00** (1)	0.450 (4)	0.781 (5)	0.101 (2)	0.078 (3)	0.159 (4)	0.292 (5)	0.023 (2)	**4.8e-04** (1)
Entropy	0.655 (4)	**0.418** (1)	0.655 (4)	0.590 (3)	**0.418** (1)	0.418 (2)	**-0.0e+00** (1)	0.655 (5)	0.418 (2)	0.590 (4)	0.828 (4)	0.655 (2)	0.655 (2)	0.828 (4)	**0.590** (1)
Entropy adj Geom. Avg.	1.018 (3)	0.004 (2)	1.129 (4)	1.298 (5)	**0.001** (1)	0.628 (3)	**0.0e+00** (1)	0.746 (4)	1.108 (5)	0.161 (2)	0.143 (3)	0.264 (4)	0.484 (5)	0.043 (2)	**7.6e-04** (1)
U															
Statistical	**0.419** (1)	0.884 (2)	1.139 (3)	1.327 (4)	1.464 (5)	4.081 (5)	1.835 (3)	2.422 (4)	**0.491** (1)	0.624 (2)	4.376 (4)	3.905 (3)	2.980 (2)	5.649 (5)	**1.467** (1)
LCSS	0.138 (3)	**0.017** (1)	1.000 (4)	0.027 (2)	1.000 (4)	0.095 (3)	**0.014** (1)	1.000 (4)	0.019 (2)	1.000 (4)	0.009 (3)	**0.006** (1)	0.010 (4)	0.009 (2)	0.010 (5)
DTW	0.027 (2)	**0.017** (1)	0.042 (4)	0.033 (3)	0.048 (5)	0.027 (2)	**0.017** (1)	0.042 (4)	0.034 (3)	0.048 (5)	0.004 (2)	0.004 (5)	0.004 (3)	**0.004** (1)	0.004 (4)
TGAK	0.016 (2)	**0.013** (1)	0.017 (4)	0.016 (3)	0.017 (5)	0.016 (2)	**0.014** (1)	0.018 (4)	0.016 (3)	0.018 (5)	0.009 (4)	**0.008** (1)	0.009 (5)	0.009 (2)	0.009 (3)
SBD	0.332 (3)	**0.309** (1)	0.373 (5)	0.333 (4)	0.316 (2)	**0.279** (1)	0.314 (4)	0.337 (5)	0.305 (3)	0.298 (2)	**0.219** (1)	0.289 (2)	0.353 (3)	0.385 (5)	0.353 (4)

Table A.1: Dissimilarity across different K+S versions, variables, and filters *(continued)*

	Unfiltered					Trend					Cycle				
Metrics	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original	K+S-Comp	K+S-finance	K+S-Ford	K+S-multi	K+S-original
Geometric Average	**0.037** (1)	0.055 (2)	0.878 (5)	0.253 (3)	0.635 (4)	0.036 (3)	**0.003** (1)	0.837 (5)	0.010 (2)	0.415 (4)	0.135 (2)	**0.017** (1)	0.691 (5)	0.189 (3)	0.256 (4)
Entropy	0.655 (3)	**0.311** (1)	0.655 (3)	0.655 (3)	**0.311** (1)	0.828 (5)	0.590 (3)	**0.418** (1)	0.590 (3)	**0.418** (1)	0.828 (2)	0.828 (2)	0.828 (2)	**0.655** (1)	0.828 (2)
Entropy adj Geom. Avg.	**0.061** (1)	0.072 (2)	1.454 (5)	0.418 (3)	0.833 (4)	0.066 (3)	**0.005** (1)	1.188 (5)	0.016 (2)	0.588 (4)	0.248 (2)	**0.031** (1)	1.263 (5)	0.313 (3)	0.468 (4)

Note:

Number in parenthesis represents the relative position, in which (1) indicates the relative best; (5) the relative worse.