

DOI: 10.34680/BENEFICIUM.2023.4(49).33-37

УДК 330.33:519.65

JEL C63, C67, E17



ORIGINAL PAPER

## USING INTERPOLATION FOR GENERATING INPUT DATA FOR THE GROSS DOMESTIC PRODUCT MONTE CARLO SIMULATION

A.M. Botchkarev, Financial University under the Government of the Russian Federation, Moscow, Russia

**Abstract.** Input modelling is a complex task within the Monte Carlo simulation, especially when the systems and processes under investigation reveal the non-linear behavior of several interdependent variables. Commonly used approaches for Monte Carlo simulation input modelling include selecting probability distributions and fitting them to existing data; resampling random variates from historical data; and using real-world data as an input model, which in the age of big data becomes more feasible. Each of the approaches comes with its own set of drawbacks. This note aims to describe a new method of input modelling for GDP Monte Carlo simulation based on interpolation of the GDP historical records. Also, this method has been implemented as a publicly available online tool using the Microsoft Azure Machine Learning Studio. A similar approach can be applied to other macroeconomic indicators, e.g., consumer price index (inflation) or current employment statistics. This note is intended for economists, data scientists, and operations research analysts interested in the GDP Monte Carlo simulation. It can also be used by academics, researchers, and practitioners in a broad subject area for generating input data for Monte Carlo simulation. Specifically, it can be of interest for Ph.D. candidates (VAC specialty 5.2.6) performing development of theory and methods of decision-making in economic and social systems, and application of artificial intelligence and big data methods in management.

**Keywords:** Gross Domestic Product, input modelling; interpolation, machine learning, Monte Carlo method, simulation

**Acknowledgement.** The views, opinions and conclusions expressed in this document are those of the author and do not necessarily represent the views of the author's current or former employers.

**For citation:** Botchkarev A.M. Using Interpolation for Generating Input Data for the Gross Domestic Product Monte Carlo Simulation // BENEFICIUM. 2023. Vol. 4(49). Pp. 33-37. DOI: 10.34680/BENEFICIUM.2023.4(49).33-37

ОРИГИНАЛЬНАЯ СТАТЬЯ

## ИСПОЛЬЗОВАНИЕ ИНТЕРПОЛЯЦИИ ДЛЯ ГЕНЕРАЦИИ ВХОДНЫХ ДАННЫХ ПРИ МОДЕЛИРОВАНИИ ВАЛОВОГО ВНУТРЕННЕГО ПРОДУКТА МЕТОДОМ МОНТЕ-КАРЛО

A.M. Бочкарев, Финансовый университет при Правительстве Российской Федерации, Москва, Россия

**Аннотация.** Моделирование входных данных является сложной задачей при использовании метода Монте-Карло, особенно когда исследуемые системы и процессы характеризуются нелинейным поведением нескольких взаимозависимых переменных. Обычно используемые подходы к моделированию входных данных методом Монте-Карло включают: выбор распределений вероятностей и их подгонку к существующим данным; повторную выборку случайных величин из исторических данных; использование реальных данных в качестве входной модели, что в эпоху больших данных становится более осуществимым. Каждый из подходов имеет свой набор недостатков. Цель этой статьи состоит в том, чтобы описать новый метод моделирования входных данных при моделировании ВВП по методу Монте-Карло, основанный на интерполяции исторических данных о ВВП. Показано, что этот метод может быть реализован с помощью общедоступного онлайн-инструмента с использованием Студии машинного обучения (Machine Learning Studio) Microsoft Azure. Аналогичный подход может быть применен для исследования других макроэкономических показателей, например, индекса потребительских цен (инфляция) или текущей статистики занятости. Эта статья предназначена для экономистов, специалистов по данным и аналитиков по исследованию операций, заинтересованных в моделировании ВВП методом Монте-Карло. Она также может использоваться исследователями и практиками в широкой предметной области для создания входных данных при моделировании методом Монте-Карло. В частности, результаты могут представлять интерес для аспирантов (специальность ВАК 5.2.6), занимающихся разработкой теории и методов принятия решений в экономических и социальных системах, а также применением методов искусственного интеллекта и больших данных в управлении.

**Ключевые слова:** валовой внутренний продукт, моделирование входных данных, интерполяция, машинное обучение, метод Монте-Карло, моделирование

**Благодарность.** Взгляды, мнения и выводы, изложенные в данной работе, принадлежат автору и не обязательно совпадают с мнением его нынешних или бывших работодателей.

**Для цитирования:** Botchkarev A.M. Using Interpolation for Generating Input Data for the Gross Domestic Product Monte Carlo Simulation // BENEFICIUM. 2023. Vol. 4(49). Pp. 33-37. (На англ.). DOI: 10.34680/BENEFICIUM.2023.4(49).33-37

## Introduction

Macroeconomic indicators are considered vital signs in the diagnostics of nations' health. Significant efforts have been invested in preparing data and calculating macroeconomic indicators such as Gross Domestic Product, Consumer Price Index (Inflation), Current Employment Statistics (CES), Interest Rate, etc. The business communities and governments keenly expect the publication of the newly calculated set of measures and use them as evidence in making far-reaching economic decisions. Of all fundamental macroeconomic indicators, Gross Domestic Product (GDP), probably, gets the most attention [1-4]. Over the years, scholars have formed a significant body of knowledge on a broad range of GDP-related research topics, e.g., predicting risks to economic growth measured through GDP [5], the relationship between economic growth and resource use (materials and energy) and greenhouse gas emissions [6], the impact of oil prices on GDP growth [7], the potential impact of COVID-19 on GDP [8], the long-run relationship between stock prices and GDP [9], etc.

Monte Carlo (MC) simulation is a well-known method of scientific analysis that includes a wide range of stochastic techniques used to quantitatively evaluate the behavior of complex systems or processes. This method has many applications in economics and finance, in general (e.g., [10-14]) and in GDP-related research, in particular (e.g., [15-17]). Conceptually, the structure of most MC experiments can be considered to include three components:

- a) generating input data to model uncertainty;
- b) randomly sampling through multiple repeated runs (simulations) of systems' or processes' models (simulation logic);
- c) quantitatively evaluating the characteristics of the model outputs (e.g., [11, 18, 19]).

This note is focused on modelling input data for MC simulation. A recent review of the input modelling for MC simulation identifies several most used approaches:

- a) selecting probability distributions and fitting them to existing data;
- b) resampling random variates from historical data;
- c) using real-world data as an input model, which in the age of big data becomes more feasible [20].

Each of the approaches comes with its own set of challenges, e.g., errors in selecting probability distributions and their parameters or insufficient volume of historical data. Misspecification of the input models may lead to errors in evaluating system outputs. Similar challenges are inherent to the MC simulation related to GDP research.

The purpose of this note is to describe a new method of input modelling for GDP MC simulation based on interpolation of the GDP historical records. Also, this method has been implemented as a tool using Microsoft Azure Machine Learning Studio, which is available in open access.

Several methodologies were used to achieve the research objective: identification of related peer-reviewed papers, critical literature review, critical thinking, and inductive reasoning.

This note is intended for economists, data scientists, and operations research analysts interested in the GDP Monte Carlo simulation. It can also be used by academics, researchers, and practitioners in a broad subject area for generating input data for Monte Carlo simulation.

For decades, academics have acknowledged the importance of input modelling (sometimes called input uncertainty or input model risk) in developing simulation models. It has gained increased attention in recent years. Recent literature reviews testify to this point (e.g., [18, 20]). Most of the reviewed papers are concerned with selecting probability distributions and adjusting their parameters to fit the stochastic behavior of real systems. The lack of knowledge about true systems and processes challenges this approach. Also, to simplify analytical descriptions and reduce the computational burden, selected probability distributions are often limited to the most common ones, such as normal, exponential or Weibull. Additionally, when a simulation scenario involves several variables, users of this approach tend to consider random variables as independent and identically distributed. Such simplifications contribute to input uncertainty and require serious validation. For many scenarios, the above simplifications are not possible, and input models must encompass multivariate, time-dependent, or interdependent (e.g., correlated) variables, leading to more complex input models with joint probability distributions (e.g., [21, 22]).

The volume of academic literature on using historical data as input for MC simulation is relatively low. For example, Shuanglong, et al. [23] used MC simulation based on historical data to forecast the charging load of electrical vehicles or Jeon and Hong [24] forecasted traffic speed on roads using historical big data. It can be noted that in both cited papers, historical data has been augmented using different pre-processing methods: in the former case, a filtering algorithm was used to find and remove data outliers, and in the latter case, transactional data were combined with data on the growth of the number of electric vehicles. An obvious advantage of the MC simulation using historical data is that there is

no need to build probability distributions and make related assumptions. Although, historical data may not be available in the volumes required for simulation.

The method proposed in this note is based on interpolation – a statistical technique that is widely used in many different fields, such as economics, finance, mathematics, computer science, etc. A recent review by Lepot, Aubin & Clemens [25] provides the state-of-the-art in this area. The implementation of the proposed method is performed in the Microsoft Azure Machine Learning (ML) Studio. The literature review confirmed that ML methods have been effectively used for interpolation of spatio-temporal observations in various research disciplines, e.g., to improve the seismic exploration by using support vector regression (SVR) for reconstructing seismic data from under-sampled or missing traces [26]; to enhance the database of wind components for climatological research by using neural networks and support vector machine [27]; to provide accurate monthly air temperature for biodiversity and ecosystem research by using a variety of methods, e.g., neural networks, support vector machine, k-nearest neighbours, random forest, etc. [28]; in material science, to develop enhanced materials by predicting their physical properties by using Gaussian process regression [29]; to support criminal justice decision makers to predict crime using kernel density estimation with Twitter-driven predictive analytics [30].

### Results and Discussion

The proposed method of preparing input data involves an MC simulation scenario, which requires GDP data and GDP components data. The US official GDP data are published quarterly by the Bureau of Economic Analysis (BEA) of the US Ministry of Commerce ([www.bea.gov](http://www.bea.gov)). Using the expenditure method, GDP is calculated as a sum of its major components: consumption, investment, government spending and net exports. The analysis of the major components as variables should consider their relationships (e.g., correlation, interdependences), which makes the development of the joint probability distribution a very complex and error-prone task (e.g., [31, 32]).

As it was mentioned, the use of historical data has been acknowledged in academic literature as a convenient alternative approach to input modelling for MC simulation. However, in many scenarios, the availability of historical data becomes a limiting factor in using this approach or totally excludes its application. MC simulation of the GDP-related processes is one of these scenarios. Being published quarterly, the data on GDP and its major components yield only a couple of hundred data points for a period of over 60 years (e.g., 262 points for a period from 1947Q1 to 2012Q2). The low volume of GDP data prohibits the direct use of this approach.

To resolve the problem of lack of data, the proposed method suggests the interpolation of the

historical data time series to produce additional points. Interpolation is a statistical method of constructing new data points within the range of a discrete set of known data points obtained by observation, sampling, or experimentation [33]. There are many mathematical techniques to realize interpolation, e.g., nearest-neighbor interpolation, polynomial interpolation (including linear, cubic, and spline), methods based on distance weighting, regression methods, machine learning methods, etc. [25]. There are no preferences which specific technique to select for preparing input data. From the proposed method's perspective, all techniques perform a common function: producing one or more additional points between the existing points. For example, for linear interpolation, additional points will be located on a straight line that passes through each pair of existing adjacent points. With a spline interpolation, a straight line is replaced by a curved pattern.

One interpolation parameter must be established to use the method – the number of additional points generated between each pair of the existing points. This parameter should be set depending on the size of the interpolated data set needed for the MC simulation experiment. Methods of evaluating the number of iterations and output accuracy have been established in the literature, e.g. [34, 35]. For the GDP-related simulations, we suggest setting this parameter to 99. This choice can be explained on an intuitive level. The existing GDP data are published quarterly. Adding 99 points between each pair of the existing points would allow viewing the resulting data set as an approximately daily representation of the GDP values and its major components. In this case, the initial data set of 262 points (in the example above) will yield a total number of 26.201 data points available for simulation. This volume of data will be sufficient for certain experiments.

A tool has been developed to implement the method proposed in Section 3. The tool has been realized as an experiment in MS Azure Machine Learning Studio and is publicly available online as part of the Azure Artificial Intelligence Gallery [36].

The overall structure of the experiment includes two data input modules, three modules with R script and two data output modules.

The first input data module contains a sample data set taken from the U.S. Bureau of Economic Analysis [37] publication on Gross Domestic Product (GDP) and its major components. The data set includes the following columns: index, GDP, Consumption (personal consumption expenditures), Investment (gross private domestic investment), Government (government consumption expenditures and gross investment), Exports (exports of goods and services) and Imports (imports of goods and services). Time series represent quarterly US GDP and its major components from 1947Q1 to 2012Q2. A total of 262 data points for each variable. The second input data module is used for preparing

placeholders for the interpolated data points. This data set has seven (7) columns (as the main data set) and 99 rows.

The data from the input modules are transferred to the Execute R Script module, which performs data set preparation for interpolation by inserting additional placeholder rows after each row of the main data set. Pre-processed data set is routed in parallel toward two modules. The first of them performs linear interpolation using the `na.approx()` function of the R zoo package, and the second module performs cubic spline interpolation using the `na.spline()` function of the R zoo package [38]. The results of the experiment are stored in two modules and available for download as .csv data sets. Each output data set (one with linear and one with cubic spline interpolation) has seven (7) columns (according to the main input data set) with 26.201 data points in each column.

### Conclusion

The proposed method of using interpolation of the GDP historical records as a technique to generate input data for the GDP-related MC simulation contributes to scientific knowledge by facilitating research experiments and allowing to avoid difficulties inherent to the approach of building joint probability distributions.

The practical contribution of the note is that the realization of the proposed method is publicly available online in the MS Azure Machine Learning Studio as part of the Azure Artificial Intelligence Gallery. Researchers and practitioners can either use this tool as is and download the stored data set or adjust the settings as needed for their own experiments.

Being based on historical data, the proposed method rests on the understanding that the current or future processes will be statistically similar to the past ones. The method has been illustrated by the US GDP numbers. Any generalizations of the results to other countries should be made with due diligence. Although it has been demonstrated that the proposed method increased the volume of the available data from several hundred to tens of thousands of data points, the new volume may be not sufficient for some simulation experiments and validation of the required number of MC iterations need to be performed for each experiment.

This paper has focused on the GDP-related area. Future efforts can apply a similar approach to other macroeconomic indicators, e.g., Consumer Price Index (Inflation), Current Employment Statistics (CES), etc.

### References

- [1] Bean C.R. Independent review of UK economic statistics. Publisher: HM Treasury, 2016. 254 p.
- [2] Feldstein M. Underestimating the Real Growth of GDP, Personal Income, and Productivity // *Journal of Economic Perspectives*. 2017. Vol. 31(2). Pp. 145-164.
- [3] Jorgenson D.W. Production and Welfare: Progress in Economic Measurement // *Journal of Economic Literature*. 2018. Vol. 56(3). Pp. 867-919. DOI: 10.1257/jel.20171358
- [4] Rehman F., Cooper R.J. Effective GDP: A Cross-Country Comparison // *The World Economy*. 2022. Vol. 46(3). Pp. 619-652. DOI: 10.1111/twec.13324
- [5] De Meo E., Tizzanini G. GDP-Network CoVaR: a Tool for Assessing Growth-at-Risk // *Economic Notes*. 2021. Vol. 50(2). DOI: 10.1111/ecno.12181
- [6] Wiedenhofer D., Virág D., Kalt G., Plank B., Streeck J., Pichler M., Mayer A., Krausmann F., Brockway P., Schaffartzik A. A systematic Review of the Evidence on Decoupling of GDP, Resource use and GHG Emissions, Part I: Bibliometric and conceptual Mapping // *Environmental Research Letters*. 2020. Vol. 15(6). Pp. 063002. DOI: 10.1088/1748-9326/ab8429
- [7] Charfeddine L., Klein T., Walther T. Reviewing the Oil Price–GDP growth Relationship: A replication Study // *Energy Economics*. 2020. Vol. 88. DOI: 10.1016/j.eneco.2020.104786
- [8] Maliszewska M., Mattoo A.A., Van der Mensbrugge D. The potential Impact of COVID-19 on GDP and Trade: A Preliminary Assessment // *Policy Research Working Paper*. 2020. DOI: 10.1596/1813-9450-9211
- [9] Österholm P. The Long-run Relationship Between Stock Prices and GDP in Sweden // *Economic Notes*. 2016. Vol. 45(2). Pp. 283-297. DOI: 10.1111/ecno.12057
- [10] Arnold U., Yildiz Ö. Economic risk analysis of decentralized renewable energy infrastructures—A Monte Carlo Simulation approach // *Renewable Energy*, Elsevier. 2015. Vol. 77(C). Pp. 227-239. DOI: 10.1016/j.renene.2014.11.059
- [11] Batkovskiy A.M., Konovalova A.V., Semenova E.G., Trofimets V.J., Fomina A.V. Study of economic systems using the simulation-based statistical modeling method // *Mediterranean Journal of Social Sciences*. 2015. Vol. 6(4). Pp. 369-380. DOI: 10.5901/mjss.2015.v6n4s4p369
- [12] Brandimarte P. Handbook in Monte Carlo simulation: applications in financial engineering, risk management, and economics. John Wiley & Sons Limited, 2014. 688 p.
- [13] Creal D.D survey of sequential Monte Carlo methods for economics and finance // *Econometric Reviews*. 2012. Vol. 31(3). Pp. 245-296. DOI: 10.1080/07474938.2011.607333
- [14] McLeish D.L. Monte Carlo simulation and finance. John Wiley & Sons Limited, 2011. 387 p.
- [15] Antonios A. An empirical analysis of simulated model of economic growth for United Kingdom // *Journal of Statistical and Econometric Methods*. 2019. Vol. 8(3). Pp. 1-19.
- [16] Gharehgozli O.A Monte Carlo Analysis of Robustness of the Synthetic Control Method and Dynamic Panel Estimation: A Comparative Case Study of a Policy Intervention // *Journal of Statistical and Econometric Methods*. 2020. Vol. 9(1). Pp. 67-87.
- [17] Lv J.P., Li Y.P., Sun J. Monte Carlo simulation based interval chance-constrained programming for regional ecosystem management – A case study of Zhuhai, China // *Ecological indicators*. 2018. Vol. 85. Pp. 214-228. DOI: 10.1016/j.ecolind.2017.10.027
- [18] Corlu C.G., Akcay A., Xie W. Stochastic Simulation under Input Uncertainty: A Review // *Operations Research Perspectives*. 2020. Vol. 7. Pp. 100162. DOI: 10.1016/j.orp.2020.100162
- [19] Botchkarev A. Assessing Excel VBA suitability for Monte Carlo simulation // *Spreadsheets in Education (EJSiE)*. 2015. Vol. 8(2). Pp. 4629. DOI: <https://sie.scholasticahq.com/article/4629>
- [20] Cheng R. History of input modeling // *Winter*

- Simulation Conference (WSC). 2017. Pp. 181-201. IEEE. DOI: 10.1109/WSC.2017.8247789
- [21] Cheng R.C., Currie C. S. Input modelling for multimodal data // *Journal of the Operational Research Society*. 2020. Vol. 71(6). Pp. 1038-1052. DOI: 10.1080/01605682.2019.1609887
- [22] Xie W., Nelson B.L., Barton R.R. Multivariate input uncertainty in output analysis for stochastic simulation // *ACM Transactions on Modeling and Computer Simulation (TOMACS)*. 2016. Vol. 27(1). Pp. 1-22. DOI: 10.1145/2990190
- [23] Shuanglong S., Zhe Y., Shuaihua L., Da M., Yuheng X., Huan X., Chunyan W. Charging demand forecasting method based on historical data // *IOP Conference Series: Earth and Environmental Science*. 2019. Vol. 295(3). Pp. 032002. DOI: 10.1088/1755-1315/295/3/032002
- [24] Jeon S., Hong B. Monte Carlo simulation-based traffic speed forecasting using historical big data // *Future generation computer systems*. 2016. Vol. 65. Pp. 182-195. DOI: 10.1016/j.future.2015.11.022
- [25] Lepot M., Aubin J.B., Clemens F.H. Interpolation in time series: An introductory overview of existing methods, their performance criteria and uncertainty assessment // *Water*. 2017. Vol. 9(10). Pp. 796. DOI: 10.3390/w9100796
- [26] Jia Y., Ma J. What can machine learning do for seismic data processing? An interpolation application // *Geophysics*. 2017. Vol. 82(3). Pp. V163-V177. DOI: 10.1190/GEO2016-0300.1
- [27] Reinhardt K., Samimi C. Comparison of different wind data interpolation methods for a region with complex terrain in Central Asia // *Climate Dynamics*. 2018. Vol. 51(9-10). Pp. 3635-3652. DOI: 10.1007/s00382-018-4101-y
- [28] Appelhans T., Mwangomo E., Hardy D.R., Hemp A., Nauss T. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania // *Spatial Statistics*. 2015. Vol. 14(A). Pp. 91-113. DOI: 10.1016/j.spasta.2015.05.008
- [29] Bélisle E., Huang Z., Le Digabel S., Gheribi A.E. Evaluation of machine learning interpolation techniques for prediction of physical properties // *Computational Materials Science*. 2015. Vol. 98. Pp. 170-177. DOI: 10.1016/j.commatsci.2014.10.032
- [30] Gerber M. S. Predicting crime using Twitter and kernel density estimation // *Decision Support Systems*. 2014. Vol. 61(1). Pp. 115-125. DOI: 10.1016/j.dss.2014.02.003
- [31] Crowley P.M., Hallett A.H. Correlations between macroeconomic cycles in the US and UK: What can a frequency domain analysis tell us? // *Italian Economic Journal*. 2016. Vol. 2(1). Pp. 5-29. DOI: 10.1007/s40797-015-0023-6
- [32] Qin D., Cagas M.A., Quising P., He X.H. How much does investment drive economic growth in China? // *Journal of Policy Modeling*. 2006. Vol. 28(7). Pp. 751-774. DOI: 10.1016/j.jpolmod.2006.02.004
- [33] Interpolation (2021). Wikipedia. URL: <https://en.wikipedia.org/wiki/interpolation> (accessed on 03.02.2021).
- [34] Heijungs R. On the number of Monte Carlo runs in comparative probabilistic LCA // *The International Journal of Life Cycle Assessment*. 2020. Vol. 25(2). Pp. 394-402. DOI: 10.1007/s11367-019-01698-4
- [35] Oberle W. Monte Carlo simulations: number of iterations and accuracy. Army Research Lab Aberdeen Proving Ground Md Weapons and Materials Research Directorate, 2015. 26 p.
- [36] Botchkarev A. Data Analysis and Interpolation Using R (2018). Experiment in Microsoft Azure Machine Learning Studio. Azure AI Gallery. URL: <https://gallery.azure.ai/Experiment/Data-Analysis-and-Interpolation-Using-R> (accessed on 05.02.2021).
- [37] U.S. Bureau of Economic Analysis. GDP and Other Major NIPA Series, 1929–2012: II. Part of Survey of Current Business. 2012. Vol. 92(8). Pp. 183-212.
- [38] Zeileis A., Grothendieck G. zoo: S3 Infrastructure for Regular and Irregular Time Series // *Journal of Statistical Software*. 2005. Vol. 14(6). Pp. 1–27. <https://doi.org/10.18637/jss.v014.i06>

#### About the Author / Информация об авторе

**Алексей Михайлович Бочкарев** – канд. техн. наук; старший научный сотрудник, Финансовый университет при Правительстве Российской Федерации, Москва, Россия / **Alexei M. Botchkarev** – Cand. Sci. (Engineering); Senior Research Associate, Financial University under the Government of the Russian Federation, Moscow, Russia  
 E-mail: AMBotchkarev@fa.ru  
 SPIN РИНЦ 7103-4875  
 ORCID 0000-0002-0689-8830  
 ResearcherID G-1173-2011  
 Scopus Author ID 24823757400

Дата поступления статьи: 28 сентября 2023  
 Принято решение о публикации: 25 ноября 2023

Received: September 28, 2023  
 Accepted: November 25, 2023