Using Synthetic Data to Limit Confounding in the Measurement of Performance of Partially Data-driven Models of Simulation

Working Paper

Juste Raimbault

Date

Abstract

1 Motivation

When evaluating data-driven models, or even more simple partially data-driven models involving simplified parametrization, an unavoidable issue is the lack of control on "underlying system parameters" (what is a ill-defined notion but should be seen in our sense as parameters governing system dynamics). Indeed, a statistics extracted from running the model on enough different datasets can become strongly biased by the presence of confounding in the underlying real data, as it is impossible to know if result is due to processes the model tries to translate or to a hidden structure common to all data.
Let illustrate the issue with a simple example.

2 Framework

3 Application

3.1 Level of Aggregation of Synthetic Data

3.2 Model for Generating Synthetic Data

4 Case Study