BIGBOY1.2: Generating Realistic Synthetic Data for Disease Outbreak Modelling and Analytics

Raunak Narwal*[1] and Syed Abbas†[2]

[1]Department of Mathematics,IISER Mohali, India
[2]Department of Mathematics, IIT Mandi, India

July 23, 2025

Abstract

Modelling disease outbreak models remains challenging due to incomplete surveillance data, noise, and limited access to standardized datasets. We have created **BIGBOY1.2**, an open synthetic dataset generator that creates configurable epidemic time series and population-level trajectories suitable for benchmarking modelling, forecasting, and visualisation. The framework supports SEIR and SIR-like compartmental logic, custom seasonality, and noise injection to mimic real reporting artifacts. BIGBOY1.2 can produce datasets with diverse characteristics, making it suitable for comparing traditional epidemiological models (e.g., SIR, SEIR) with modern machine learning approaches (e.g., SVM, neural networks). .

Keywords: synthetic data; epidemiology; outbreak modelling; visual analytics; SEIR

1 Introduction

Infectious diseases have repeatedly challenged the global health system, economies, and societies. During the past few decades, we have witnessed outbreaks such as SARS (2003), H1N1 influenza (2009), Ebola (2014), and COVID-19 (2020), which have demonstrated how quickly pathogens can disrupt normal life and healthcare systems, causing unprecedented economic and social consequences. In such scenarios, epimedic modelling plays significant role in disease outbreak prediction, policymaking and timely intervention stratergies. But epidemiological modelling remains heavily dependent on the availability and quality of outbreak data. Missing dates, reporting delays and substandard datasets make it challenging to train and benchmark models. That has led to our increased interest in synthethic data generation, BIGBOY1.2 could generate reality mimicking datasets and visualtions which make them ideal for benchmarking models, stress testing algorithms, and conducting reproducible experiments.

1.1 Background

Accurate modeling and forecasting of disease outbreaks have been a crucial topic for public health planning and decision making. Classical epidemiological models, such as compartmental models (SIR, SEIR) have proven their effectiveness for understanding transmission dynamics. These can be used to estimate parameters like basic reproduction number R_0, beta effective and estimate interventions. However these models are idealistic and generally different from real world data. Real world data is noisy, incomplete and subject to irregular reporting due to many factors such as delays in case confirmation, underreporting and inconsistent testing

*ms23177@iisermohali.ac.in

†abbas@iitmandi.ac.in, internship supervisor

policies across different regions.
The COVID-19 pandemic further highlighted the need for high quality datasets for epidemic modelling. Most studies relied on the use of fragmented and incomplete datasets, which limited the reliability of forecasts and their ability to reproduce. Data inconsistencies such as negative incidence values (due to backlogs and correction) created major challengers for data-driven machine learning models, that require large, well structured datasets. As a result, forecasting methods on real world datsets is often inconclusive.
To limitations have compelled researcherrs to use synthethic data. Our synthethic data generator , BIGBOY1.2 allows for complete control over epidemic parameters, which includes population, layers, seasonality, interventions and stochastic variations. It provides an invaluable testbed for benchmarking forecasting models under controlled scenarios, enabling rigorous evaluation of algorithmic performance in condition where real world data would be insufficient or biased. But most existing synthethic dataset tools are either very simplistic and fail to mimic the complex nature of real world outbreaks or too specialized, designed for specific disease and narrow research goals.

1.2 Motivation for BIGBOY1.2

As discussed before, synthethic data generators fall short in key aspects of realism, flexibilty and usability. Important factors like seasonal variation, stochastic effects and reporting biases are ignored and are tightly coupled to specific diseases or parameters settings. As a consequence, researchers resort to creating ad-hoc datasets, which lack standardization, making it difficult to compare forecasting models across studies.
Moreover current tools rarely integrate visual analytics with the data generation pipeline. The ability to intuitively visualize compartmental dynamics, intervention impacts is very crucial for communicating findings effectively. Without built in visualization support, the user relies on external scripts and tools, increasing complexity to even perform a basic exploratory analyses.
We have proposed BIGBOY1.2 , a versatile and fully configurable synthethic dataset generator for disease outbreak modeling and analytics. BIGBOY1.2 allows users to simulate epidemics with customizable transmission parameters and intervention stratergies. It also generates visual plots such as time series plots, heatmaps, phase diagrams along with datasets. BIGBOY1.2 is lightweight and easy to use, unlike many heavy ML based dataset generators. By standardizing synthethic dataset creation, BIGBOY1.2 aims to improve reproducibility and enable fair benchmarking of disease outbreak modeling.

2 Related Work

Brief comparison with existing synthetic epidemic simulators, agent-based models, or benchmarking datasets.

3 Methods

3.1 Model Architecture

Describe compartments (S, E,I, R, etc.), transitions, sampling.

3.2 Parameterisation

Ranges, distributions, seasonality functions.

3.3 Data Outputs

CSV schema; metadata; reproducibility seed.

4 Experiments

4.1 Benchmark Models

E.g., SIR, SEIR, SVR, SVM-FDE hybrid.

4.2 Evaluation Metrics

RMSE, MAE, R^2, calibration, lead-time scores.

5 Results

Highlight key quantitative results; include tables and plots.

6 Visualisation Examples

Show streamgraph, stacked area, XY plot of beta vs. seasonality, etc.

7 Discussion

Strengths, limitations, intended uses, future work.

8 Conclusion

Short recap + what users can do with BIGBOY1.2.

Acknowledgments

Thank collaborators, funding (if any), and data providers.

Data and Code Availability

Links or DOIs to GitHub / Zenodo / institutional repos.