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 24, 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 Methods

BIGBOY1.2 is a stochastic epidemic dataset and visual plot generator which builds upon BIGBOY1. With further refinements , it presents better and more realistic datasets.

2.1 Framework

BIGBOY1.2 is designed to simulate realistic infectious disease outbreaks with a high degree of configurability and realism. At its core, it is build on the SEIR (Susceptible, Exposed, Infectious, Recovered) model, extended with dynamic parameters, seasonal influences, vaccination and multi wave outbreak structures. It is a multi layered and age structured SEIR model which includes a noise and reporting module to simulate real world data irregularities. Unlike traditional simulations, BIGBOY1.2 produces data that closely resembles real world epidemic curves, and also retains full control over the underlying "ground truth" parameters. This allows researchers to test forecasting methods under controlled conditions.
the framework is modular in design, it consists of four key layers: Parameter Initialization, where user can manually define epidemiological and behavioral parameters; Simulation Engine, integrates the SEIR based equations and accounts for time varying transmission dynamics, interventions and stochastic effects; Noise and Reporting Layers, injects realistic data artifacts like under reporting, reporting delays and random fluctuations to mimicc real world surveillance

data and the last layer is Output and Visualizations, which exports the datasets in csv formats, parameters in JSON format and generates a range of visual graphs from simple time series plots to advanced 3D plots.
BIGBOY1.2 supports three operational modes through CLI : random mode (parameters taken from predefined ranges), interactive mode (user-driven configuration), and batch mode (generates many simulations at once). The user could also use various commands from the CLI like –plots all , –population X. More about this is discussed here in the demo video.

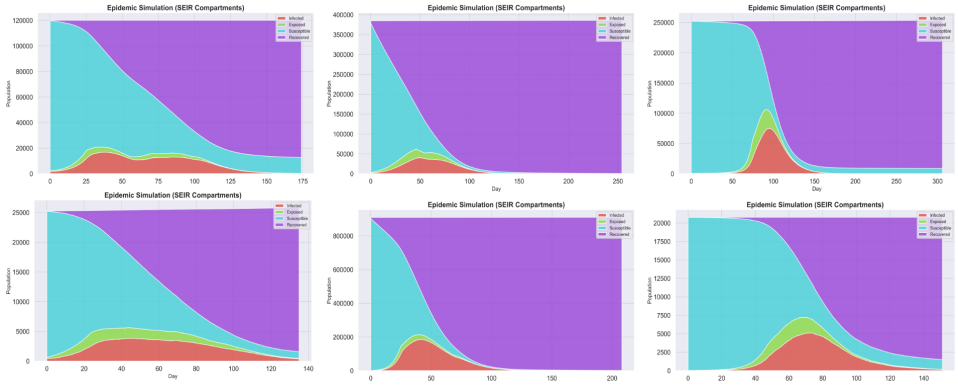

Figure 1: Batch mode (Python BIGBOY1.2.py batch 6 –plots all)

2.2 Mathematical Foundations of BIGBOY1.2

As discussed BIGBOY1.2 builds upon the foundational SEIR model, which categorizes the population into four compartments. At any point in time, each individual belongs to one of these compartments and transition between them is governed by a set of differential equations, this has been done to incorporate real world effects such as vaccination, behavioral factors and seasonality.

$$\begin{aligned} \frac{dS}{dt} &= -\beta(t) \cdot \frac{SI}{N} - \nu S \\ \frac{dE}{dt} &= \beta(t) \cdot \frac{SI}{N} - \sigma E \\ \frac{dI}{dt} &= \sigma E - \gamma I \\ \frac{dR}{dt} &= \gamma I + \nu S \end{aligned} \tag{1}$$

Here is a description of the model's parameters:

- $S(t)$, $E(t)$, $I(t)$, and $R(t)$ represent the number of susceptible, exposed, infectious, and recovered individuals at time t, respectively.
- N is the total population size (assumed constant).
- $\beta(t)$ is the time-varying transmission rate, which is crucial for defining how fast susceptible individuals become exposed.
- σ is the rate at which exposed individuals become infectious (the inverse of the incubation period).
- γ is the recovery rate.
- ν is the vaccination rate, which transfers susceptible individuals directly into the recovered class.

This extended form of SEIR formulation allows BIGBOY1.2 to simulate epidemic dynamics with the effects of public health interventions such as mass vaccination.
Time dependent transmission rate $\beta(t)$ is a powerful and novel feature of BIGBOY1.2. The framework does not assume a constant rate of disease transmission, rather it models $\beta(t)$ as a function of multiple interacting factors, each of which represents a real world influence on transmission dynamics.

$$\beta(t) = \beta_0 \cdot (1 - \theta_m m(t)) \cdot (1 + \theta_c c(t)) \cdot \left[1 + \alpha \sin\left(\frac{2\pi t}{T_s}\right)\right] \cdot \Phi(t) \tag{2}$$

Where

- β_0 is the baseline transmission rate, in the absence of external modifiers.
- $m(t)$ is the mask adherence score at time t, normalized between 0 and 1.Higher the mask adherence score, higher the compliance with mask wearing. The weight θ_m controls how strongly this factor suppresses transmission.
- $c(t)$ is the crowdedness score, also on a normalized scale. In this, the average density of human interaction is captured, with θ_c amplifying its effect on transmission.
- The sinusoidal term $\alpha \sin\left(\frac{2\pi t}{T_s}\right)$ models seasonality, it represents periodic increases or decreases in transmission due to environmental and behavioral cycles). T_s is the seasonal cycle period (typically 365 days).
- Finally, $\Phi(t)$ is a multi-wave adjustment factor, this allows the simulation to include multiple waves (due to new variants or changes in social behavior). It is defined as:

$$\Phi(t) = 1 + \sum_{j=1}^{W} (\phi_j - 1) \cdot \sigma_j(t) \tag{3}$$

Here each ϕ_j represents the peak multiplier of the j-th wave, and $\sigma_j(t)$ is a logistic ramp function that smoothly increases and decreases during the wave period. This component enables multiple waves having sharp rises and slow declines in transmission, a feature often seen in real epidemic data.
BIGBOY1.2 supports simulations with heterogeneous population structures, segmented by age and contact environments. This structure is implemented using an age and layer straified SEIR model. In confiugrations like this, the population is divided into L contact layers, such as household, workplace, school or community. A age groups, such as children, adults and the elders. For each combination the simulation tracks : $S_{l,a}$, $E_{l,a}$, $I_{l,a}$, and $R_{l,a}$.
Where, l = 1,2,...,L denotes the contact layer and a = 1,2,...,A denotes the age group.
The force of infection $\lambda_{l,a}(t)$, or the probability per unit time that a susceptible individual in group (l, a) becomes exposed, is calculated using summated contributions from all other groups based on this structured contact matrix:

$$\lambda_{l,a}(t) = \sum_{l'=1}^{L} \sum_{a'=1}^{A} \beta_{l,a,l',a'}(t) \cdot \frac{I_{l',a'}(t)}{N_{l',a'}} \tag{4}$$

This implies that the exposure risk for a school going kid within the community layer depends on how many infectious individuals exist in other age groups and settings, modulated by the contact matrix $C_{l,l'}$. This method brings realism into the simulation, and modeling of targeted interventions could also be enabled (like school closure or age prioritized vaccination).
The base SEIR model along with above extensions we have done, the BIGBOY1.2 provides a mechanistic ground truth view of an outbreak but real world surveillance data is noisy and subject to various distortions as well. To mimic this effect, BIGBOY1.2 introduces a post processing layer that applies multiple forms fo noise and uncertainity to the generated data.

Travel Noise in real epidemics, the geographical boundaries of a population is not sealed. People travel in and out of region for work, migration and emergencies. The local outbreak curves is affected by this movement, often introducing sudden spikes or dips. We have simulated this behavior through a travel noise generator, which adds or subtracts random infectious cases from the SEIR-generated curve. At each timestep t, the infectious compartment $I(t)$ is changed by:

$$I'(t) = I(t) + \Delta_{\text{travel}}(t)$$

Where $\Delta_{\text{travel}}(t) \sim \mathcal{N}(\mu, \sigma^2)$, a Gaussian-distributed noise term with mean μ and standard deviation σ. These parameters can also be fixed by the user. This gives noisy, jagged , heavy tailed curves that retains the overall trend of the outbreak and also includes short term fluctuations mimicking travel between cities.

Random Dropper , another realism challenge in epidemiology is underreporting of cases, all infections are not captured. This may be due to various reasons, maybe because a computer simulation is not really a real life oubtreak scenario. So, to reflect this we have included a random dropper, that hides a certain fraction of cases from the output.

$$\text{Reported}_I(t) \sim \text{Binomial}(I'(t), p_r)$$

Where:

- $I'(t)$ is the noisy infectious count after travel adjustment.
- $p_r \in [0, 1]$ is the reporting probability.

This same method can be applied to independent exposed, recovered, depending on the use case. We can define the output of BIGBOY1.2 as a function :

$$\mathcal{D}_{\text{BIGBOY1.2}} = \mathcal{R}\left(\mathcal{N}\left(\mathcal{S}_{\text{SEIR}}\left(\boldsymbol{\Theta}, \beta(t), \Phi(t), \nu, \mathbf{C}, \mathbf{M}, \mathbf{A}\right)\right)\right)$$

Where:

- $\mathcal{D}_{\text{BIGBOY1.2}}$: Final reported dataset
- $\mathcal{S}_{\text{SEIR}}$: SEIR simulator that generates compartment curves over time.
- $\boldsymbol{\Theta}$: Core epidemiological parameters $\{\beta_0, \gamma, \sigma, N\}$.
- $\beta(t)$: Time varying transmission function.
- $\Phi(t)$: Multi wave logistic ramp (captures new waves).
- ν: Vaccination rate.
- $\mathbf{C}$: Contact matrix.
- $\mathbf{M}, \mathbf{A}$: Layer $\mathbf{M}$ and age-group $\mathbf{A}$ structures.

Then:

- $\mathcal{N}(\cdot)$: Noise layer, which applies:
  - Travel noise: $\Delta_{\text{travel}}(t) \sim \mathcal{N}(\mu, \sigma^2)$.
  - Reporting delay.
  - Weekend or weekday bias.
- $\mathcal{R}(\cdot)$: Reporting layer, which applies:
  - Random dropper: $\text{Binomial}(I'(t), p_r)$.

– Reporting frequency control (daily, weekly, etc.).

The final synthetic dataset $\mathcal{D}_{\text{BIGBOY1.2}}$ is created by first running a SEIR simulation $\mathcal{S}_{\text{SEIR}}$. Then, stochastic noise $\mathcal{N}$ and distortions are injected. Finally, a reporting filter $\mathcal{R}$ simulates real-world underreporting.

3 Simulation Pipeline

BIGBOY1.2 is made as modular simulation engine , it is structured into well defined functional blocks. Each module processes data through a deterministic or stochastic transformation, allowing control and reproducibility. The system is configured using parameters.json and put together using python driver script.

Configuration Parsing and Preprocessing : At runtine, the simulation parses a structured paramter file that containing :

```
{
    "population": 20785,
    "days": 152,
    "initial_infected": 32,
    "mask_score": 10,
    "crowdedness_score": 7,
    "quarantine_enabled": "y",
    "seasonality_enabled": "y",
    "interventions_enabled": "n",
    "reporting_prob_min": 0.52,
    "reporting_prob_max": 0.72,
    "multi_wave": "n",
    "random_seed": 845114,
    "vaccination_enabled": "n",
    "daily_vaccination_rate": 0.016,
    "incubation_period": 5,
    "waves": [
        {
            "day": 60,
            "beta": 2.5,
            "seed": 100
        }
    ],
    "testing_rate": "medium",
    "mask_decay_rate": 0.0156,
    "travel_enabled": "n",
    "travel_max": 0,
    "mode": "random",
    "layers": 2,
    "age_groups": 3
}
```

Above is a sample params.json taken from a generated batch. The parser validates all input types, auto generates required times eries and prepares input buffers for the simulation.

Compartmental Simulation Layer: This module numerically integrates a layered, age structered SEIR system. It implements forward Euler integration over discrete time stamps, contains $L \times A$ compartment states in 4D tensors

$$S[L, A], \quad E[L, A], \quad I[L, A], \quad R[L, A]$$

Transmission rate β_t is calculated per time stamp by combining time dependent behavioral scores (mask , crowd), seasonal effects (sinusiodal) and wave ramp function. Cumulative states are kept in the memory and this engine supports toggling between homogenenous and stratified contact modes.
Multiwave Modulation, this submodule applies wave shpaed multipliers on beta effective.
Noise Injection Module, wraps the raw SEIR outputs and introduces realistic distortions such as tavel noise, delay, radnom modulators and zero clipping.
Output and Export Handlers are responsible for making time series CSVs for reported date , it includes 2 CSV files, one containing just the reported cases and the other containing: Day, Susceptible, Exposed, Infected, Recovered, New Exposed, New Infections, New Recoveries, Reported Cases, β_t, Seasonality, R_t. Output manager is also responsible for diagnostic logs (parameter hash, seed) and optional visualizations.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Day	Susceptible	Exposed	Infected	Recovered	New_Exposed	New_Infections	New_Recoveries	Reported_Cases	Beta_Effective	Seasonality	Rt
2	0	20746	7	30	2	7	0	2	0	0.270495086	0.558201195	0.270079
3	1	20744	8	29	4	2	1	2	0	0.273334695	0.564046169	0.272822
4	2	20738	13	29	5	6	1	1	1	0.27711915	0.571838218	0.276573
5	3	20732	15	31	7	6	4	2	1	0.281831893	0.581542541	0.281195
6	4	20730	15	32	8	2	2	1	1	0.28745229	0.593115797	0.286719
7	5	20720	21	33	11	10	4	3	2	0.293955733	0.606506297	0.293178
8	6	20714	24	34	13	6	3	2	2	0.301313768	0.621654235	0.300371
9	7	20709	22	39	15	5	7	2	2	0.309494239	0.638491957	0.308437
10	8	20702	22	42	19	7	7	4	1	0.318461452	0.656944262	0.317297
11	9	20694	24	40	27	8	6	8	3	0.32817636	0.676928737	0.326866

Figure 2: Dataset Snapshot from a random batch

Figure 3: Directory snapshot of save CSVs, JSON and PNGs

Reproducibility and Logging: reproduceibility is done by random seeds, this allows any experiment to be fully replicated or benchmarked until the user hasn't deleted the params.json file.

4 Results

Highlight key quantitative results; include tables and plots.

5 Visualisation Examples

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

6 Discussion

Strengths, limitations, intended uses, future work.

7 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.