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 28, 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)[1], Ebola[2], and COVID-19[3], which have shown how quickly pathogens can disrupt normal life and healthcare systems, causing unprecedented economic and social consequences. In such scenarios, epimedic modeling plays significant role in disease outbreak prediction, policymaking and timely intervention stratergies[4]. 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)[5], 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[6].
The COVID-19 pandemic further highlighted the need for high quality datasets for epidemic modelling[7]. 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[8].

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[9].
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[5], extended with dynamic parameters, seasonal influences, vaccination and multi wave outbreak structures[10]. It is a multi layered and age structured SEIR model which includes a noise and reporting module to simulate real world data irregularities[11]. 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 mimic 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[8].
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.

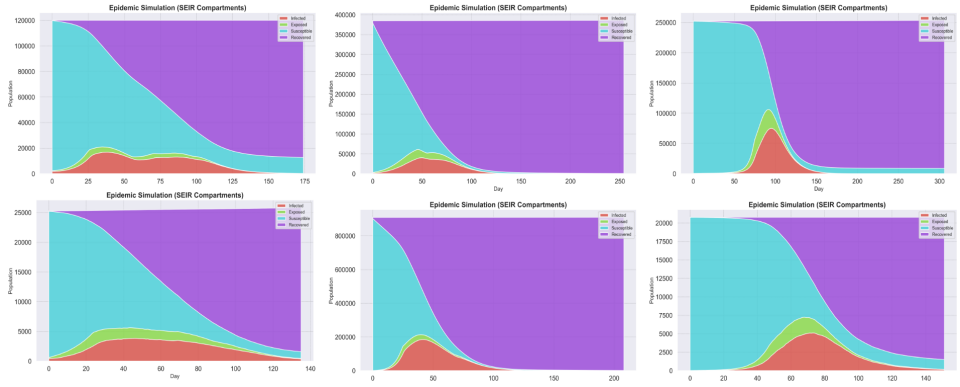

Figure 1: Batch mode (Python BIGBOY1.2.py batch 6 –plots all)

2.2 Mathematical Foundations of BIGBOY1.2

As discussed BIGBOY1.2 builds upon the foundational SEIR model, which categorizes the population into four compartments [5]. 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 [12, 13].

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

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 [14]:

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

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) [12].
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 [11].
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: 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

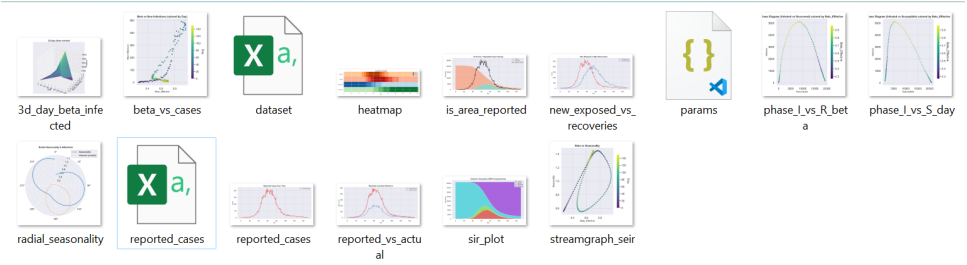

Figure 3: Directory snapshot of save CSVs, JSON and PNGs

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 Visual Plots and Demonstration

Visual Plots are an integral part of BIGBOY1.2, it presents epidemic simulation data into interpretable and high dimensional plots. These visual plots could be used as analytical tools that could validate moedl outputs and also reveal deeper insights like epidemic progression, interventions and control

4.1 Overview of the Plotting System

Upon simulation, BIGBOY1.2 allows the users to generate a diverse set of set plots by passing –plots all or –plots sir. It has a modular plotting engine that allows both minimal and advanced visualization and the output is aved as high resolution PNG files. All of the plots are further save in timestamped directories with accompanying metadata, ensuring reporducibility.

4.2 SEIR Compartment Plot

A stacked area plot showing the progression of Susceptible (S), Exposed (S), Infected (I) adn Recovered (R) populations over time, classical SEIR-style visualization is foundational for understanding the macro level behavior of the epidemic. Th purpose of SEIR-stacked chart is revealing key phases such as exponential growth, peak infections and herd immnunity thresholds.

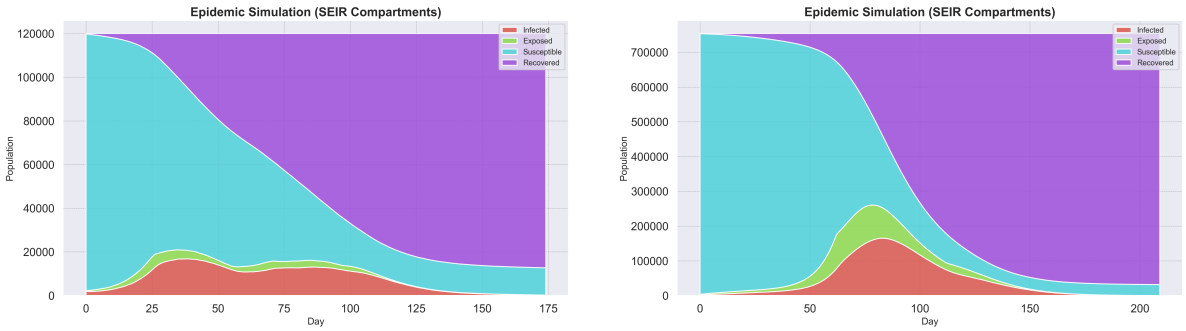

Figure 4: SEIR stacked plots from a random batch

4.3 Reported Cases Timeline

It displays the reported cases across days, this contracts the real world observed data with latent epidemic dynamics. It simulated the public health reporting pattern .

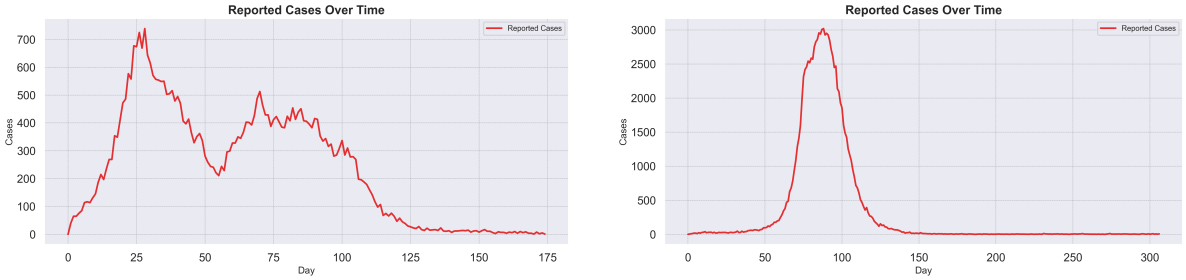

Figure 5: Reported cases from a random batch

4.4 3D plot of day $\times\beta$ $\times$infection

This 3D plot displays how contagious a disease is, as contingency and number of infections are not linear, 3D visualization shows how small changes in β can lead to explosive outbreaks under conditions. Lag effects can also be revealed visualizing this 3D plot, even if β rises sharply, infections may spike few days later, this helps in understadning incubation periods.
If an intervention like mask adherence is applied (or lockdown) , it causes β to drop and the flattening infection counts could be seen in the Z-axis. This shows casual evidence of policy effectiveness in time. In multiple wave simulations, the 3D plot clearly shows how subsequent waves differ in timing, strength and transmissibility. These could be compared early vs later variants of the disease visually.

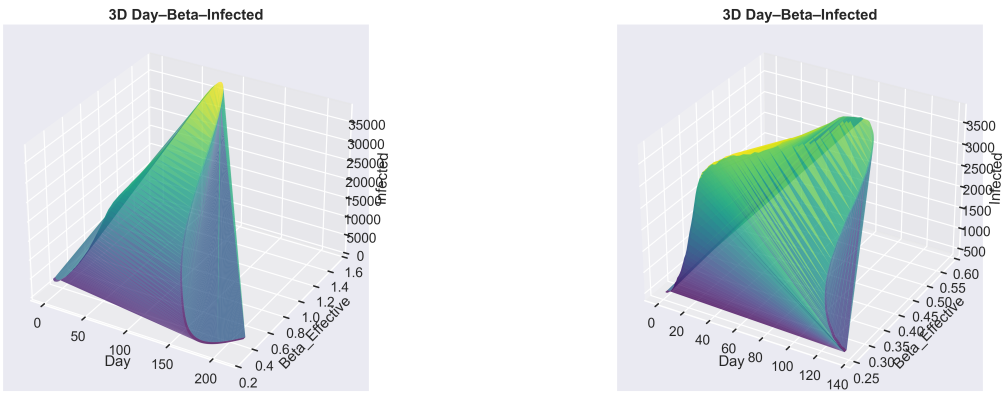

Figure 6: 3D plots from a random batch

4.5 β vs. New Infections (Colored by Day)

This scatterplot is crucial for understadning how chnages in β correlate with spikes or drops in new infections. Higher β generally increases the infections, but during later stanges, even higher β may not cause spikes due to immunity build up. These effects are shown by Day, the scatter dots are colored by day.

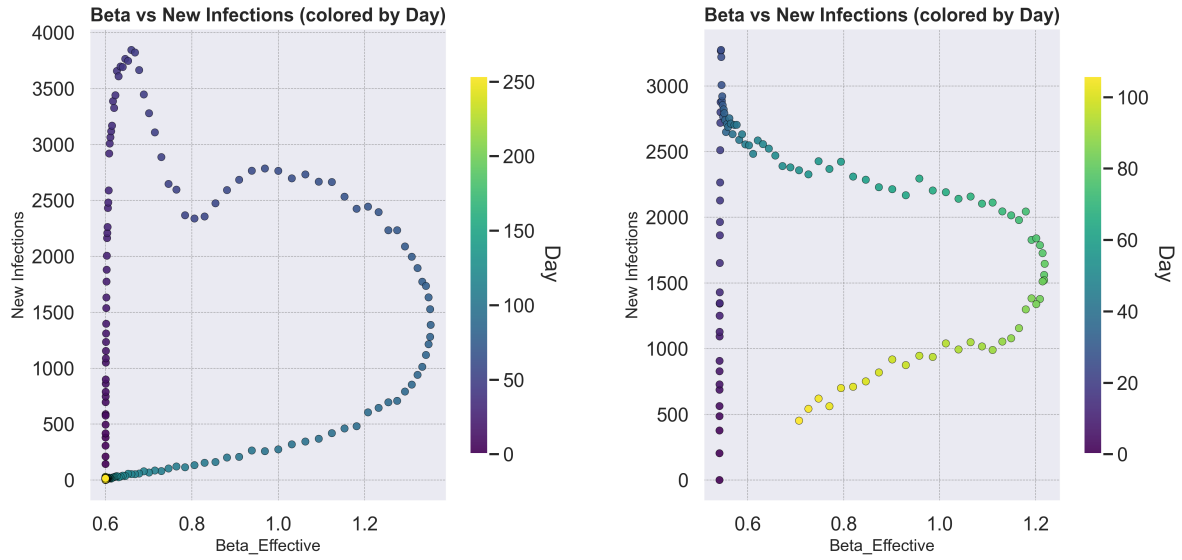

Figure 7: Beta vs. New Infections plot from a random batch

4.6 New Exposed vs New Recoveries

This plot helps in epidemic growth detection. If New exposed > New Recoveries, the infection is spreading faster than its being cleared and if new exposed < new recoveries, it indicates decline of epidemic.

This plots often shows a crossover point where the two curves intersect each other. The first corssover signals the wave onset and the second crossover suggests wave resolution or success in interventions.

Reduced transmission is demonstrated by sharp dips in new exposed after a ploicy change for example lockdown and vaccination. If new recoveries continue to rise , it means that the healthcare system is still managing the previous cases, but future burden is dropping.

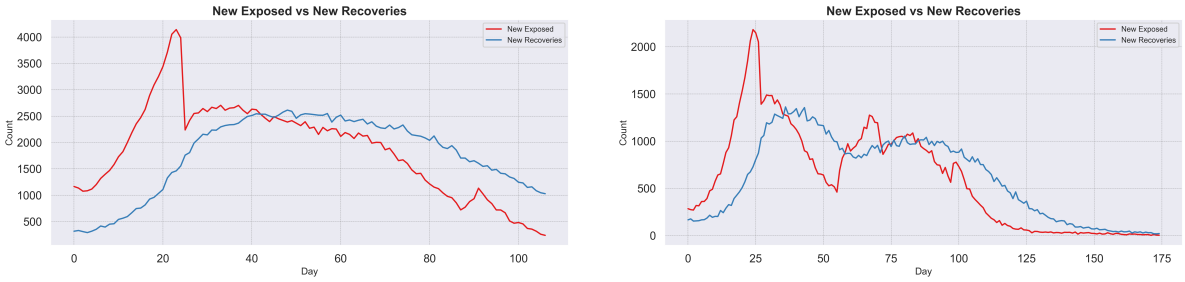

Figure 8: New Exposed vs New Recoveries plot from a random batch

4.7 Phase Diagrams

Two types of phase diagrams are displayed, the I vs S diagram and the I vs R diagram, both colored by β effective, in the I vs S phase plot, each point represents the system's state ona given day, as the epidemic goes on, the system traces a trajectory through its space forming a loop or an arc, highliting the rise and fall of infections relating to the shrinking susceptible population. The I vs R variant similarly tracks how infections transition into recoveris and it is particularly uselful for visualizing the cumulative impact of the epidemic. Unlike simpler plots, phase diagrams capture the entire system's evolution in a compact , geometric path , making them crucial for both theoretical exploration and practical modeling.

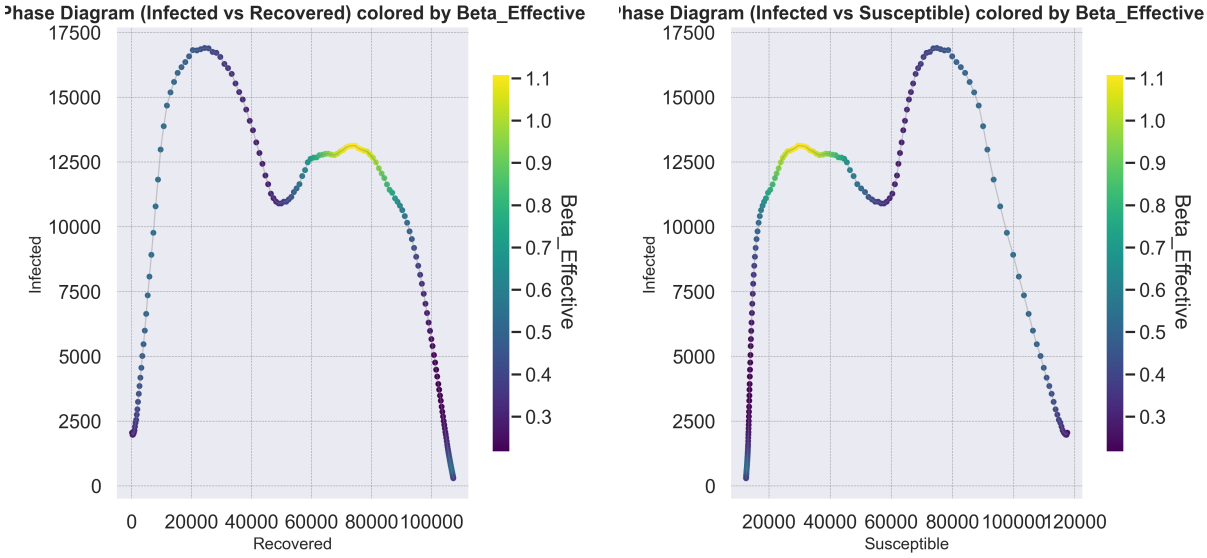

Figure 9: Phase Diagrams from a random batch

4.8 Othe plots

The **radial seasonality** plot uses a polar coordinate system to display cycles of seasons alongside infection levels, which makes it ideal for visualizing periodic disease where transmission rate is directly related to seasonality. The radial axis shows both seasonality strength and infection counts which allows user to correlate peaks in transmissiblity with actual outbreaks.
The **Reported vs Actual Infections** is a dual ine graph that compares reported cases with new infections which helps users to visualize under reporting, testing delays and observational noise. It is key for assessing the gap between observed date and true disease data , especialy in low testing regions and during surges.

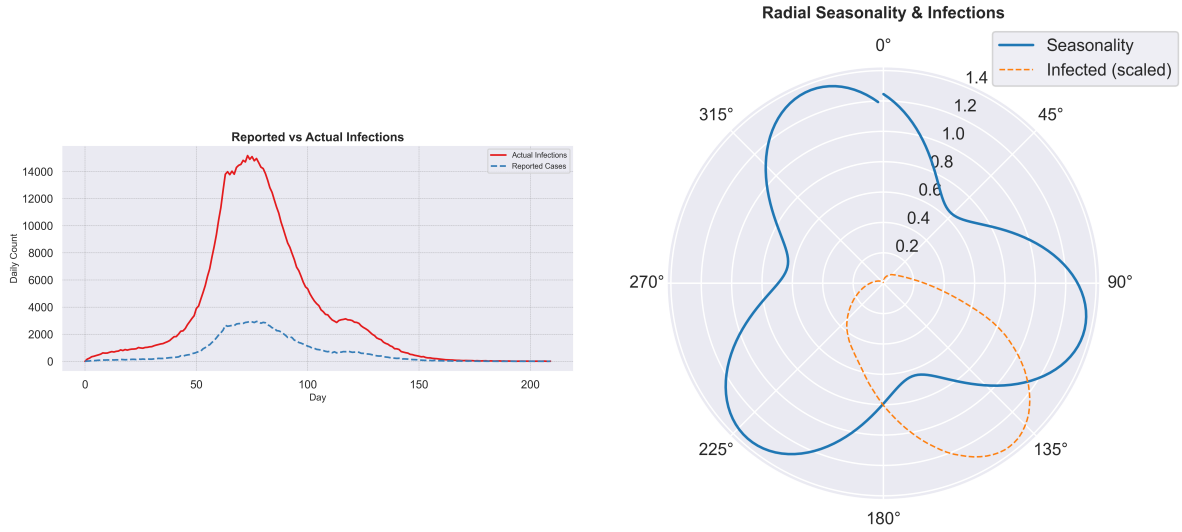

Figure 10

4.9 Comparison with real world date

To evaluate the realism and validity of the synthethic data generated by our model, we have compared it with the actual daily reported cases of COVID-19 in India (for the second wave).The data were taken from the official WHO website [15], first converted to BIGBOY format (it had only one column of reported cases) using a Python script (available on GitHub), and the comparison was run using another Python script. We used scaling to match the range and resolution of synthethic outputs. The overall epidemic curve shape, peak structure and rise-fall dynamics have remained remarkably consistent across both datasets, despite manual feeding of parameters to BIGBOY1.2. We have done metric based comparisons between the two data below.

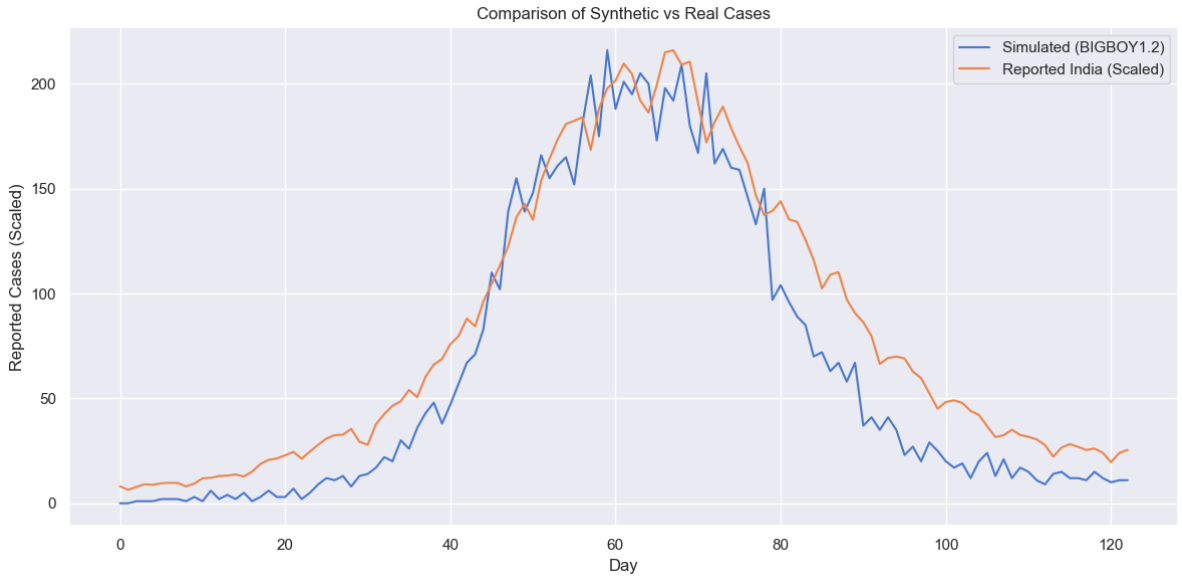

Figure 11: COVID19 2nd wave compared to BIGBOY1.2

Metric	Real Data	BIGBOY1.2 Simulated
Basic Reproduction Number (R_0)	1.29	1.16
Epidemic Duration (Days)	53 to 71	51 to 70
Peak Infection Days (Top 3)	67, 68, 59	59, 68, 67

Table 1: Comparison of COVID-19 vs BIGBOY1.2

4.10 Sources of Deviation

The shape of the waves depict promising overlap, there are still scopes of improvement, these deviations are not unexcepted and can be attributed to several factor like **real world data complexity** , publicly available data suffers from factors like inconsistent testing , region specific anomalies. Extracting granular data would improve graph alignment more. Another factor is **parameter calibration**, BIGBOY1.2 uses manually selected parameters, which is good for understanding curves and real world factors but when it comes to mimicking real world curves, automated hyperparameter tuning (like Bayesian optimization) would match real epidemics more precisely. These changes would be incorporated in the next version of BIGBOY.
These minor discrepancies between the simulated and observed curves do not stem out from model inaccuracies, rather it is the inherent stochasticity and undetermined variables that play in the real world outbreaks. In fact, real world epidemics are so chaotic that even the same virus would behave differently if replayed in the same conditions.

5 Future Work (BIGBOY1.3)

BIGBOY1.2 provides a great platform for generating synthethic disease outbreak data but its development is not completed yet. We have planned to include several powerful extensions for future versions, the roadmap includes the following improvements :

5.1 Enhanced Visualizations

We aim to include an animate mode as –animate X (X being different animated visuals) to make outbreak visualizations more dynamic and interactive. This mode would include layered epidemic progression, animated transmission waves and geospatial spread mapping. We would also be integrating agent based and grid based simulations to complement the compartmental SEIR model. This will allow us to see a individual level perspective, mobility and stochasticity, it would help in capturing phenomena like superspreading and localized interventions.

5.2 Disease mode

A new interface would be implemented in the CLI, which would allow the use to select from pre-configured templates for diseases like COVID-19, measles, influenza and more. Each of these templates would include pre loaded parameters, allowing faster and disease specific scenario generation and modelling. The mode could be accessed as –disease X (X being the disease template). We would also expand on the intervention parameters and healthcare system constraints would also be implemented. A better multiwave structure would also be configured.

5.3 Automated Hyperparameter Tuning

BIGBOY1.3 will feature automated hyperparameter tuning using techniques such as Bayesian Optimization and grid search [16], which will calibrate parameters directly from real datasets, which could be used in scenarios where limited amount of data is available for a particular disease and BIGBOY1.3 could be used to generated unlimited disease like data using AHT.

5.4 Country mode

BIGBOY1.3 will also feature a country mode where user could select from all the countries and a set of calibrated parameters would be applied to them. These parameters would include population, crowding, literacy (corresponding to mask adherence and vaccinations), country

specific behaviors, population pyraminds (dividing population into categories) and seasonality profies.
Let us understand capabilites of BIGBOY1.3 using an example

```
–disease EBOLA –country INDIA –state HARYANA –animate GRID
```

This would simulate disease outbreak of EBOLA in Haryana, India. Though EBOLA has never hit Haryana, but the model contains both the profiles for EBOLA and Haryana and will flawlessly simulate and generate datasets for Ebola Outbreak in Haryana.

5.5 Ramen1

We are working on several SVR, SIR, SEIR and FDE [17, 18] based models and their hybrids to make SOUP S-1 (SVR-SIR) , CRUM1 (SVR-SEIR), Broth-N (Naive baseline based SVR) and SVR-FDE models, further this model soup would be merged into a high fidelity ensemble system named Ramen1 , which would combine best of all approaches using weighted voting and time series.

6 Open Source and Contributions

BIGBOY1.2 is open source and could be used for unrestricted academic and non-commercial use. The complete codebase along with sample datasets from BIGBOY1, BIGBOY1.1 and BIGBOY1.2 are publicly available on GitHub.
A short demo video demonstrating how to run BIGBOY1.2 on your local machine is available on Youtube.
We encourage students and researchers to use BIGBOY1.2 to understand and simulate epidemic curves and we are open to contributors who would like to be a part of BIGBOY1.3 and further improvements. For contribution purposes, mail here.
Ramen1 is the bigger project which required generation of a synthethic dataset, that's how BIGBOY1 came into being. Ramen1 used many sub models to predict disease outbreak, it switches between various model depending on the phase of the outbreak (ensemble model). It is still work in progress, if anyone wants to contribute in Ramen1 , send a mail here.

6.1 Acknowledgements:

The author (Raunak Narwal) would like to thank **Prof. Syed Abbas** for this internship opportunity which led to this project.
The author would also like to thank **Rishu Narwal**, Phd candidate at IIT Delhi, for helping out with research literature, structure of the research paper and providing a computational support.
Special thanks to **Chirag Verma**, third year undergraduate studetn at IISER Mohali, for his theoretical insights and discussions in the area of epidemic outbreak modeling.

References

[1] World Health Organization. Summary of probable sars cases with onset of illness from 1 november 2002 to 31 july 2003. https://www.who.int/publications/m/item/summary-of-probable-sars-cases, 2004. Accessed: 2025-07-28.

[2] World Health Organization. Ebola virus disease. https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease, 2021. Accessed: 2025-07-28.

[3] World Health Organization. Coronavirus disease (covid-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019, 2020. Accessed: 2025-07-28.

[4] Matt J. Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.

[5] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[6] Ingrid Holmdahl and Caroline Buckee. Wrong but useful—what covid-19 epidemiologic models can and cannot tell us. *New England Journal of Medicine*, 383(4):303–305, 2020.

[7] Matteo Chinazzi, Jessica T. Davis, Marco Ajelli, Carlo Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, Cecile Viboud, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368(6489):395–400, 2020.

[8] Srinivasan Venkatramanan, Bryan Lewis, Jinho Chen, David Higdon, Anil Vullikanti, and Madhav Marathe. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*, 30:100383, 2020.

[9] Evan L. Ray, Nutcha Wattanachit, Jarad Niemi, Anindya H. Kanji, Kathryn House, Estee Y. Cramer, et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *medRxiv*, 2020. Preprint. Available at: https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v2.

[10] Amina Bakhta, Mahdi Hacene, and Benallal Rezak. Extending seir model to predict covid-19 dynamics with vaccine rollout and multiple waves. *Applied Mathematical Modelling*, 101:384–401, 2022.

[11] Kristina M. Gostic, Lauren McGough, Eben B. Baskerville, et al. Practical considerations for measuring the effective reproductive number, r_t. *PLoS Computational Biology*, 16(12):e1008409, 2020.

[12] Kate M Bubar, Katia Reinholt, Stephen M Kissler, et al. Model-informed covid-19 vaccine prioritization strategies by age and serostatus. *Science*, 371(6532):916–921, 2021.

[13] Jonathan Neipel, Michael Meyer-Hermann, et al. Power-law population heterogeneity governs epidemic waves. *PLoS One*, 15(12):e0239678, 2020.

[14] Kiesha Prem, Alex R Cook, and Mark Jit. Projecting contact matrices in 152 countries using contact surveys and demographic data. *PLOS Computational Biology*, 13(9):e1005697, 2017.

[15] World Health Organization. Who coronavirus (covid-19) dashboard: India, 2021. Accessed: 2025-07-26.

[16] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, volume 25, pages 2951–2959, 2012.

[17] Yihong Duan et al. Application of support vector regression for covid-19 pandemic prediction. *Journal of Physics: Conference Series*, 1644(1):012160, 2020.

[18] Ming Li and Gang Chen. Fractional-order seir model and its application to covid-19 in italy. *Nonlinear Dynamics*, 101:865–876, 2020.

"Thanks"