Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Lupin Cai

April 1, 2025

Lupin Cai

Date

Amplification of Demographic Bias in Epidemiological Forecasting

By

Lupin Cai

Li Xiong
Advisor

Andreas Züfle
Co-Advisor

Computer Science

Li Xiong
Advisor

Andreas Züfle
Co-Advisor

Max Lau
Committee Member

2025

Amplification of Demographic Bias in Epidemiological Forecasting

By

Lupin Cai

Li Xiong
Advisor

Andreas Züfle
Co-Advisor

An abstract of
A thesis submitted to the Faculty of the mory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Computer Science

2025

Abstract

Amplification of Demographic Bias in Epidemiological Forecasting
By Lupin Cai

Forecasting infectious disease outbreaks is critical for timely public health responses, yet predictive models are often trained on biased data that reflect real-world disparities in data collection and reporting. This thesis investigates how such bias can be amplified across different forecasting models, including traditional time series models (ARIMA), graph-based deep learning models (CoLA-GNN), and epidemiologically structured neural networks (SIR-NN). We evaluate model performance on both a synthetic simulation dataset and real-world COVID-19 case data from Georgia, introducing synthetic underreporting bias based on demographic features such as age, income, gender, and education.

We apply clustering to group regions by demographic attributes and compare model performance using absolute and relative error metrics. Results show that while absolute error tends to be higher in less-biased clusters (with higher total cases), relative error consistently rises in clusters with greater underreporting, validating the hypothesis that models trained on biased data amplify disparities in forecasting accuracy. Notably, even advanced models like CoLA-GNN, which incorporate spatio-temporal dependencies, are not immune to this amplification. Across both datasets, the gap between reported and ground truth data correlates with poorer model performance in affected regions, highlighting a compounding effect of demographic bias over multi-step forecasting.

The study reveals that demographic-aware bias evaluation is essential for responsible epidemiological modeling. It emphasizes the importance of using relative error and daily error progression, rather than just aggregate absolute metrics, to uncover latent disparities in model performance. This work contributes to the understanding of fairness in epidemic forecasting and provides a foundation for developing bias-mitigation strategies in future modeling efforts.

Amplification of Demographic Bias in Epidemiological Forecasting

By

Lupin Cai

Li Xiong
Advisor

Andreas Züfle
Co-Advisor

A thesis submitted to the Faculty of the Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Computer Science

2025

Acknowledgments

I would like to acknowledge and give my warmest thanks to my advisor Professor Li Xiong who is willing to take me into her research group as a student who has no experience in doing research in computer science. She not only provided guidance but also knowledges and suggestions on how the graphs should be produced and how the pipeline should be working.

I would like to give my second warmest thanks to my advisor Professor Andreas Züfle, who provided me with insights and the target of the project, and the meetings with other professors from other universities and other countries that I was able to witness different aspects of application of computer science.

I would like to give my sincere thanks to Professor Max Lau for graciously agreeing to serve as a member of my thesis committee, despite us having never met before. Even though this project may not directly align with his area of expertise, he kindly offered his support, and his willingness to contribute played an important role in allowing me to complete this honors thesis.

I would like to give special thanks to Toan Tran, a PhD student who helped me debugging, provided data, and many suggestions to the project; there are many mistakes made by me throughout the process of the project; however, he was able to point many of them out such that I was able to fix them right away, and Ruochen, Kong, a PhD student who provided me with all of the simulation data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Epidemiological forecasting plays a crucial role in public health decision-making by predicting disease spread, healthcare demand, and intervention effectiveness. However, these forecasts are often susceptible to biases that can propagate and amplify, leading to misleading conclusions and potentially flawed policy decisions. Demographic bias refers to systematic errors in data that disproportionately affect certain population subgroups, such as the elderly, sex, or low-income individuals. For example, underreporting in older populations - due to limited access to testing or healthcare services - can lead to their heath outcomes being underrepresented in epidemiological datasets, resulting in models that systematically underpredict outbreaks in these demographics[7].

This thesis investigates the problem of demographic bias amplification in epidemiological forecasting models. Our core hypothesis is that forecasting models trained on demographically biased data will not only reflect but also amplify disparities in prediction accuracy accross population subgroups. Specifically, we expect that models will perform worse in regions where the data underrepresents or misrepresents the true case burden due to demographic characteristics.

Related work in machine learning fairness highlights how bias can infiltrate data, model design, and decision-making pipelines. Prior studies have addressed data

imbalance and bias correction in various domains, but relatively few have explored how these biases manifest in epidemiological forecasting models. Some work investigates fairness in spatiotemporal predictions or epidemic modeling, but limitations remain - especially regarding bias amplification during multi-step forecasting and the lack of standardized frameworks for demographic error analysis.

in this study, we address these gaps by:

- implementing serveral state-of-the-art forecasting models, including a traditional time seires model (ARIMA), a graph-based deep learning model (CoLA-GNN), and a physics-informed epidemiological model (SIR-NN).
- evaluating these mdoels using a simulated infectious disease data set [12] designed with controlled demographic bias to allow fine-grained analysis of error patterns.
- augmenting real-world COVID-19 data from Georgia with synthetic underreporting based on age, creating demographically biased versions of reported cases.
- conducting a comprehensive bias amplification analysis by clustering regions based on demographic attributes and comparing model performance using absolute and relative error metrics.

Our main findings include the following:

- forecasting models consistently demostrate higher relative errors in clusters with greater demographic bias (i.e., higher underreporting of older populations), even when absolute error is lower due to lower case volumes.
- this bias amplification occurs across all model types, including CoLA-GNN, which incorporates spatiotemporal dependencies.

- relative error metrics and error progression over time are more informative than absolute nmetrics alone when assessing fairness.
- the SIR-NN model, while more robust due to its epidemiological structure, still shows signs of bias sensitivity under synthetic underreporting.

These results highlight the critical need for bias-aware evaluation and mitigation strategies in public health forecasting, especially when models inform high-stakes policy decisions.

Chapter 2

Background

2.1 Bias in AI Models

The increaseing adoption of machine learning (ML) and artificial intelligence (AI) systems in high-stakes domains has raised urgent concerns about fairness, equity, and bias. Bias in AI systems can emerge from various sources, including data collection processes, model design choices, and deployment contexts. The paper "A Survey on Bias and Fairness in Machine Learning" [10] serves as a cornerstone for understanding the impact of bias across various ML systems and the implications for their fairness and reliability in sensitive applications such as epidemiology. It also provides a foundational collection of bias types and categrozies mitigtation strategies as pre-processing, in-processing, and post-processing.

Importantly, the survey emphasizes bias amplification, where machine leraning model not only learn existing disparities from their training data but overfitting them over time. These insights form the basis for examining similar phenomena in epidemiological modeling, where biased data inputs can cascade into distored forecasts and unfair policy recommendations. Studies like Ferrara [5] further expand on these mitigation frameworks, underscoring that the choice of mitigation strategy must be

context-sensitive, as some methods may compromise accuracy or fail to generalize.

2.2 Epidemiological Forecasting Models

Epidemiological forecasting plays a central role in informing public health interventions by anticipating the spread of infectious diseases, projecting healthcare demand, and evaluating the impact of policy measures. Over the years, a wide range of forecasting methodologies have been developed, each with distinct assumptions, data requirements, and strengths.

Compartmental models, such as the classic SIR (Susceptible-Infected-Recovered) and its variants (e.g., SEIR, SIRD), have been foundational in epidemic modeling. These models use systems of ordinary differential equations (ODEs) to simulate the flow of individuals through disease states. They are interpretable and grounded in epidemiological theory but require accurate estimation of parameters like transmission and recovery rates. Modern extensions such as physics-informed neural networks (e.g., SIR-NN) [8] integrate these frameworks with machine learning to improve flexibility and fit real-time data more dynamically.

Beyond compartmental models, time series approaches such as ARIMA (Auto Regressive Integrated Moving Average) have been widely used due to their simplicity and strong short-term predictive performance. Studies like "Application of the ARIMA Model on the COVID-2019 Epidemic Dataset"[1] demonstrated the ARIMA model's utility in forecasting the prevalence and incidence of COVID-19 cases. the authors employed ARIMA(1,0,4) to estimate prevalence and ARIMA(1,0,3) to assess incidence trends. Their findings reveal the model's potential to generate reliable short-term forecasts, making it a valuable tool for real-time monitoring and public health decision-making.

Machine learning approaches offer a data-driven alternative and include models such

as Random Forests, Support Vector Machines (SVMs), and Gradient Boosting. These models can capture complex nonlinear patterns and interactions between predictors, including mobility, weather, or policy indicators. However, they are prone to overfitting and often lack interpretability in public health contexts.

Deep learning models have been increasingly applied, particularly in large-scale or multimodal datasets. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based architectures have shown promise in sequential prediction tasks, such as forecasting case counts or hospitalization rates. Some models, such as EpiForecast or DeepGLEAM, incorporate attention mechanisms or combine multiple data modalities for improved accuracy.

In recent years, graph-based and deep learning methods have gained traction. CoLA-GNN is an example of a spatio-temporal graph neural network desgined for epidemic forecasting. It integrates geographical connectivity and temporal patterns to model disease dyanmics more holistically. Similar approaches include DCRNN, STGCN, and EpiGNN, which use diffusion or message-passing mechanisms over mobility or adjacency graphs to model inter-regional spread.

Hybrid models combine mechanistic and statistical paradigms to exploit the strengths of both. For example, Bayesian hierarchical models incorporate uncertainty and prior knowledge while enabling partial pooling across regions. Other approaches integrate mobility-informed priors or mechanistic constraints into machine learning models to improve generalizability and interpretability.

Finally, real-time ensemble forecasting systems, such as the COVID-19 Forecast Hub or FluSight Network, aggregate predictions from multiple models to improve accuracy and capture uncertainty. These systems highlight the importance of diversity in modeling strategies and the need for robust evaluation frameworks.

In summary, the landscape of epidemiological forecasting models is diverse, ranging from interpretable mechanistic models to flexible deep learning architectures. Each

class of models offers trade-offs between interpretability, accuracy, data requirements, and scalability. The choice of model often depends on the stage of the epidemic, data availability, and specific policy or operational goals.

2.3 Bias in Epidemiological Models

While epidemiological forecasting has evolved significantly in recent decades, there is growing recognition that these models—like many predictive systems—are vulnerable to bias. Bias in epidemiological models can emerge from structural inequalities in healthcare systems, data collection processes, and model assumptions, leading to disproportionate errors across demographic groups.

One of the most pervasive issues is reporting bias, which stems from underreporting of infections or deaths in specific populations due to limited access to testing, healthcare services, or digital infrastructure. Studies during the COVID-19 pandemic revealed that elderly, rural, and low-income populations were less likely to be tested and diagnosed early, which introduced significant gaps in real-time surveillance data [6][2]. These biases are especially problematic in models that are calibrated on observed case data, as they cause the model to underestimate the burden of disease in underrepresented regions. Moreover, selection bias can affect surveillance-based models that rely on voluntary testing or self-reporting. For instance, smartphone-based symptom trackers disproportionately sample younger, wealthier, and more technologically connected users, skewing data inputs and limiting generalizability [4].

Simulation-based approaches, including agent-based models (ABMs) and metapopulation models, simulate individual-level interactions and behaviors to estimate disease trajectories. Though computationally intensive, these methods provide granular insights into policy interventions and population heterogeneity. Synthetic data generated from ABMs can also serve as testbeds for fairness and robustness analysis, as

demonstrated in recent work [12].

Lastly, the intersectionality of demographic attributes (e.g., age and income, or gender and race) adds further complexity. Bias in one dimension may correlate with others, leading to compounding effects. For instance, Guerra-Silveira and Abad-Franch [9] show that sex differences in infectious disease incidence may be driven by both biological and social factors. Thus, simplifying population structure can obscure nuanced disparities.

Chapter 3

Approach

3.1 Problem Definition

Accurate forecasting of infectious disease spread is critical for public health decision-making, particularly in the early stages of an outbreak. However, real-world epidemiological data often suffer from systematic underreporting and demographic imbalances, introducing bias into the datasets used for model training. When forecasting models—ranging from classical time series models like ARIMA to more sophisticated neural architectures like CoLA-GNN and SIR-NN—are trained on these biased datasets, the resulting predictions may inherit and amplify such bias, leading to disproportionate forecasting errors across demographic groups.

This study aims to define and investigate the following core problem: *How does demographic bias in training data affect the accuracy and fairness of epidemiological forecasting models?* Specifically, we examine whether models trained on synthetically biased data perform worse on populations that are underrepresented or underreported, and whether this performance gap increases over time during multi-step sequential forecasting.

To evaluate this, we introduce synthetic underreporting based on demographic

attributes (e.g., age, gender, income, education) in both simulated and real-world COVID-19 datasets. We then analyze model performance across clusters defined by these demographics using error metrics such as Absolute Error (AE) and Relative Error (RE). A model is considered biased if the forecasting error is systematically higher for regions with greater underreporting, despite being trained on uniformly biased data.

Understanding the manifestation of bias in different model architectures—and its amplification over sequential predictions—enables the development of fairness-aware evaluation frameworks and guides future work on debiasing methods in epidemiological modeling.

3.2 Data Description

Our study utilizes two distinct datasets:

- **Simulation Data**: This dataset consists of an *agent-based multivariate time series* representing case records for **Fulton County, Georgia**, across **27 regions**. Each agent is characterized by **five attributes**, which capture epidemiological and behavioral factors influencing disease transmission. However, the dataset is constrained to a **30-day period**, limiting its ability to capture long-term trends and fully assess bias amplification over extended timeframes.

- **Real-World Data**: In addition to the simulation data, we incorporate real-world epidemiological data from the state of Georgia with 159 counties and government census by assuming the population distribution to be uniform and univariant. We then apply the target proportion of the target population to the difference of cases each region each day. Thus we will have reported data, the perturbed data, and the ground truth data, the original reported data, as we cannot obtain the ground truth data from real world.

3.3 Data Preprocessing

To effectively structure both datasets for model training and evaluation, we employed the **Sliding Window technique**, a widely used approach in time series forecasting. This method creates overlapping input-output pairs, where each sequence of past observations is used to predict future values.

3.3.1 Sliding Window

Formally, given a time series:

$$X = \{x_1, x_2, \ldots, x_T\} \tag{3.1}$$

for a selected **window size** W, the dataset is transformed into training samples:

$$\mathbf{X}^{(i)} = (x_i, x_{i+1}, \ldots, x_{i+W-1}), \quad y^{(i)} = x_{i+W} \tag{3.2}$$

This transformation ensures that short-term temporal dependencies within the dataset are effectively captured and leveraged for forecasting.

3.3.2 Multi-Step Forecasting

For sequential forecasting over multiple future steps (horizon H), the output can be extended to:

$$Y^{(i)} = (x_{i+W}, x_{i+W+1}, \ldots, x_{i+W+H-1}) \tag{3.3}$$

where $Y^{(i)} \in \mathbb{R}^H$ represents the next H predicted values.

3.3.3 Recursive Forecasting

A common strategy in forecasting is recursive prediction, where the model generates predictions iteratively:

$$\hat{x}_{t+1} = f(x_t, x_{t-1}, \ldots, x_{t-W+1}) \tag{3.4}$$

and then uses $\hat{x}_{t+1}$ as input for the next step.

3.3.4 Forecasting Models

The selection of forecasting models in this study follows a progressive trajectory—from traditional statistical models to advanced graph-based learning and finally to hybrid epidemiological neural networks. This progression reflects both increasing model complexity and representational power, offering a comprehensive perspective on how different modeling paradigms react to biased input data.

- **ARIMA**: We selected ARIMA as our foundational model due to its simplicity, transparency, and long-standing application in time series forecasting. ARIMA is widely used in public health for short-term disease trend prediction because it requires minimal data preprocessing and offers interpretable parameters (autoregressive and moving average terms). Although it does not model spatial relationships or non-linear dynamics, its exclusion of external features makes it an ideal baseline for isolating the effect of bias in purely temporal trends. In our context, ARIMA serves to quantify the lower bound of model performance and acts as a control to compare more advanced methods.

- **Cola-GNN**: To capture the complex spatial and temporal interactions in epidemiological data, we selected CoLA-GNN, a spatio-temporal graph neural network that incorporates both regional connectivity and sequential case progression. This model is particularly well-suited for infectious disease forecasting

because it reflects how outbreaks evolve through inter-regional transmission pathways—something traditional models like ARIMA cannot capture. CoLA-GNN also benefits from the expressive power of deep learning while preserving structural context through graph-based message passing. We chose CoLA-GNN to examine whether access to spatial information and non-linear representations improves robustness to biased training data or inadvertently amplifies disparities due to overfitting.

- **SIR-NN**:As the model that works completely different with the previous two models, SIR-NN combines the interpretability of mechanistic epidemic models with the adaptability of neural networks. This physics-informed architecture explicitly models disease dynamics via differential equations governing susceptible, infected, and recovered compartments, while also learning data-driven approximations of key parameters like transmission and recovery rates. SIR-NN was chosen to evaluate whether embedding epidemiological priors into the learning process enables better generalization in the presence of biased or incomplete data. Because it is grounded in domain knowledge, we hypothesized that SIR-NN would be less prone to overfitting to biased observations, making it an essential model for assessing fairness in real-world applications.

Together, these three models represent a spectrum of forecasting philosophies—statistical, spatial deep learning, and epidemiological hybrid—allowing us to explore how different assumptions and architectures mediate the relationship between data bias and prediction disparity.

3.3.5 Evaluation Metrics

To assess the performance of these models and analyze bias amplification, we used the following evaluation metrics:

Absolute Error (AE): Measures the direct difference between the actual and predicted values.

$$AE_t = |x_t - \hat{x}_t| \tag{3.5}$$

The Mean Absolute Error (MAE) is computed as:

$$MAE = \frac{1}{N} \sum_{t=1}^{N} |x_t - \hat{x}_t| \tag{3.6}$$

Relative Error (RE): Evaluates the magnitude of prediction error relative to the actual value, ensuring scale-independent comparisons.

$$RE_t = \frac{|x_t - \hat{x}_t|}{|x_t|} \tag{3.7}$$

The Mean Relative Error (MRE) is given by:

$$MRE = \frac{1}{N} \sum_{t=1}^{N} \frac{|x_t - \hat{x}_t|}{|x_t|} \tag{3.8}$$

These metrics allow us to quantify forecasting accuracy and assess whether prediction errors systematically increase over time, indicating potential bias amplification.

3.3.6 Bias Analysis via Clustering

To further examine how bias is amplified across different regions for the simulation data, we applied **K-Means clustering with** $k = 2$ to categorize the **27 regions** based on their attributes. For real world data, we applied **K-Means clustering with** $k = 2$ to categorize the **159 regions** based on the percentage of older population since that is how the synthetic data was created. This approach enabled us to group regions with similar epidemiological characteristics and analyze prediction errors across the two clusters.

By comparing the gaps between **forecasted values and ground truth** within each cluster, we assessed whether bias was disproportionately affecting specific groups of regions. If one cluster consistently exhibited larger errors than the other, this would indicate that the forecasting model's accuracy varied based on underlying regional characteristics, reinforcing potential disparities in prediction performance.

Chapter 4

Experiments

4.1 Experiment Setup

The experiment follows a structured pipeline involving data preprocessing, model training, hyperparameter tuning, forecasting, error analysis, and bias evaluation. We conducted experiments using ARIMA, Cola-GNN for simulation data, and ARIMA, Cola-GNN, and SIR-NN for real-world COVID data set, following the same framework for comparability.

4.2 Data Preprocessing for Simulation Data

Experiments conducted on synthetic agent-based simulation data aimed to to analyze how bias amplification manifests in a controlled environment. The simulation data is generated from Professor Züfle's lab [12]. The simulation is an agent-based simulation that incorporates controlled biases into the experiment by systematically perturbing the agent attributes that prevents the agent with unrealistic attributes from happening. In other words, a 3 years old agent with an 110k annual income will not show up.

Agent Data Aggregation

The simulation input was provided in the form of individual-level agent data. Each agent's records included static demographic attributes and time series of infectious case statuses. To enable region-level forecasting:

- Agents were grouped by each of their **feature/attributes**.
- Daily counts of infectious agents were aggregated across all agents in each region to generate a **region-wise time series**.
- Two versions of this time series were generated: **ground truth** (full data) and **reported cases** (perturbed to introduce bias).

Synthetic Bias Perturbation

To simulate reporting bias, we perturbed the infection counts based on the **age distribution** in each region:

- A bias factor was applied such that a higher proportion of infections from **older individuals** were retained in the reported data.
- This emulated a scenario in which public health reporting systems systematically underreport infections among younger populations.
- The resulting **biased time series** was used as the model input, while the ground truth was retained for evaluation.

Adjacency Matrix Construction

To incorporate spatial dependencies, we used a pre-defined **27×27 adjacency matrix** representing interaction potential between regions. This matrix was derived from simulation metadata describing how agents move or interact between regions.

4.2.1 Model Training and Hyperparameter Tuning

The same training setup was used for Cola-GNN, and ARIMA models as described in the real-world experiment. Each model was trained on the biased (reported) data and evaluated against both reported and ground truth sequences.

Cola-GNN was trained with a 7-day lookback window and a 7-day prediction horizon. Hyperparameters such as learning rate, hidden dimensions, number of GRU layers, and dropout rate were tuned based on MSE and MAE on validation data.

4.3 Data Preprocessing for Real-world Data

For the real world dataset, we used the COVID-19 dataset from John Hopkins University [3] for case count, and we used county information dataset from government census [11] to construct the adjacency matrix. To ensure the dataset was structured appropriately for forecasting and bias analysis, we performed multiple preprocessing steps involving adjacency matrix construction, population-based bias computation, and synthetic data generation.

Adjacency Matrix Construction

Since our forecasting model incorporates **spatial dependencies**, we constructed an adjacency matrix representing county-level connectivity in Georgia. The process involved:

- Parsing raw text data containing county adjacency relationships.
- Filtering for Georgia counties and extracting only relevant neighboring connections.
- Converting the adjacency list into a **binary adjacency matrix**, where a value of 1 indicates a direct connection between two counties.

- Saving the adjacency matrix as a CSV file for input into the Cola-GNN model.

This matrix serves as an essential input for the **graph-based forecasting model**, allowing it to capture spatial dependencies in disease spread.

Demographic-Based Bias Computation

To quantify potential biases in epidemiological forecasting, we computed the percentage of individuals with age 65 years and older in each county using census data. The process involved:

- Extracting population estimates from **demographic data**.
- Aggregating population counts for specific age groups across counties.
- Computing the percentage of individuals with age 65 years and older relative to the total county population.
- Storing the results as a county-wise bias percentage dataset for later analysis.

This information was later used to assess whether forecasting errors were correlated with regional demographic factors.

Synthetic Data Generation

To simulate reporting bias, we generated a modified version of the real-world case counts. The synthetic dataset was designed to introduce selective underreporting based on demographic distribution:

- Merging real-world epidemiological data with the computed demographic bias percentages.
- Applying a bias factor of 80% underreporting for the fraction of cases attributed to the individuals with age 65 years and older.

- Generating daily **case differences** while maintaining temporal dependencies.
- Aggregating the results to create a biased **synthetic case datase**.

Assuming there are x percent of target population in one region, for that region specifically, we will have reported cases:

$$\text{reported daily difference} = \left(1 - \frac{x}{100}\right) \cdot \text{target daily difference} + \text{bias} \cdot \frac{x}{100} \cdot \text{target daily difference}$$

The resulting dataset was used as **"reported cases"**, allowing us to analyze how bias propagates through forecasting models.

These preprocessing steps ensured that our datasets were formatted correctly, enabling a robust investigation into **bias amplification in epidemiological forecasting**.

4.4 Shared Setup for Both Simulation and Real-world Data

4.4.1 Sequential Forecasting

After model training, we used the best models trained above for sequential predictions. Given an initial input sequence of 7 days, the model iteratively forecasted the next 7 days while updating the input dynamically. Predictions were compared against both **ground truth** and **reported cases**.

4.4.2 Model Training and Hyperparameter Tuning

For Cola-GNN, we performed hyperparameter tuning to optimize model performance. The training process involved:

- Using a **Sliding Window technique** to create overlapping input-output pairs for model training.
- Training on 75% of the dataset and evaluating on the remaining 25%.
- Exploring various hyperparameter configurations, including:
  - Learning rates: $\{0.001, 0.0001\}$
  - Hidden units: $\{64, 128\}$
  - Epochs: $\{5, 10, 15, 20\}$
  - RNN model type: RNN
  - Dropout: $\{0.3, 0.5\}$
  - Number of layers: $\{2, 4, 6, 8\}$
- Selecting the best model based on test loss (Mean Squared Error and Mean Absolute Error).

For ARIMA, we performed hyperparameter tuning to optimize the model performance. The training process involved:

- Using a **Sliding Window technique** to create overlapping input-output pairs for model training.
- Training on 75% of the dataset and evaluating on the remaining 25%.
- Exploring various hyperparameter configurations, including:
  - p values: $\{0, 1, 2, 3\}$
  - d values: $\{0, 1\}$
  - q values: $\{0, 1, 2, 3\}$

- Selecting the best model based on test loss (Mean Squared Error and Mean Absolute Error).

For SIR-NN, we also performed hyperparameter tuning to optimize the model performance. The training process involved:

- Using a **Sliding Window technique** to create overlapping input-output pairs for model training.
- Training on 75% of the dataset and evaluating on the remaining 25%.
- Exploring various hyperparameter configurations, including:
  - number of hidden layers: $\{2, 4, 6\}$
  - number of hidden neurons: $\{32, 64, 128\}$
  - activation functions: $\{Tanh, SELU\}$
  - number of iterations: $\{5, 10, 15, 20\}$
  - beta: [0.05, 0.5]
  - gamma: [0.05, 0.3]
- Selecting the best model based on test loss (Mean Squared Error and Mean Absolute Error).

The optimal hyperparameter configurations were saved as the final model for forecasting.

4.4.3 Error Analysis

To evaluate model performance, we computed the following error metrics:

- **Absolute Error (AE)**: Measures the direct difference between predicted and actual values.

$$AE_t = |x_t - \hat{x}_t| \tag{4.1}$$

- **Relative Error (RE)**: Measures the magnitude of prediction error relative to the actual value.

$$RE_t = \frac{|x_t - \hat{x}_t|}{|x_t|} \tag{4.2}$$

The errors were analyzed over multiple weeks and across different regions.

4.4.4 Bias Amplification Analysis

To investigate the impact of bias on forecasting accuracy, we applied **K-Means clustering with** $k = 2$ to group the 159 regions based on their demographic and epidemiological characteristics, which is the percentage of the target population for that region. Each region was assigned a cluster, and we examined:

- Absolute and relative error distributions across clusters.
- Differences between model errors on reported vs. ground truth cases.
- Cluster-specific trends in forecasting performance.

By comparing prediction errors between clusters, we assessed whether bias was systematically amplified, disproportionately affecting specific regional groups.

4.4.5 Visualization and Interpretation

To further interpret the results, we generated:

- Bar charts comparing absolute and relative errors across clusters.
- 7-day error trend visualizations by cluster.
- Time series plots of actual vs. predicted cases over multiple weeks.

These visualizations provided insights into how forecasting accuracy varied across regions and whether model biases disproportionately impacted certain groups.

Chapter 5

Analysis

5.1 Simulation Data

5.1.1 Simulation Analysis: ARIMA

To evaluate the ARIMA model's behavior under synthetic demographic bias, we simulate forecasting across regions grouped into clusters based on demographic attributes. For each demographic axis (e.g., age, gender, income), regions are grouped as:

- **Cluster 0**: Regions with lower average target attribute (e.g., younger age, lower income, or fewer females)
- **Cluster 1**: Regions with higher average target attribute (e.g., older age, higher income, or more females)

Absolute Error Analysis

The mean absolute error (AE) across clusters is summarized in Table 5.1. In the **age axis**, Cluster 1 shows higher AE, suggesting that regions with more of the target attribute (e.g., older populations) may suffer greater deviations under bias.

Table 5.1: ARIMA Absolute Error (AE) by Demographic Cluster

Demographic Axis	Cluster 0 (Lower Target)	Cluster 1 (Higher Target)
Age	2.3522	3.3172
Gender	2.4669	2.9140
Income	2.4850	2.5161

Relative Error Analysis

Table 5.2 shows relative error (RE), normalized by the true case counts. Across all axes, **Cluster 1 consistently exhibits higher RE**, indicating that forecasting errors have disproportionate impact in regions with more of the demographic target.

Table 5.2: ARIMA Relative Error (RE) by Demographic Cluster

Demographic Axis	Cluster 0 (Lower Target)	Cluster 1 (Higher Target)
Age	0.083	0.090
Gender	0.084	0.091
Income	0.080	0.090

Bias Gap Evaluation

To assess the underlying bias, we measure the difference between ground truth and reported data in Table 5.3. In the age axis, Cluster 1 has a larger bias gap — suggesting either stronger underreporting or greater distortion in high-target regions.

Table 5.3: ARIMA Ground vs. Reported Bias Gap by Demographic Cluster

Demographic Axis	Cluster 0 (Lower Target)	Cluster 1 (Higher Target)
Age	1.6338	2.3929
Gender	1.7292	1.8431
Income	1.8187	1.6923

Summary

These results support the broader hypothesis that bias in training data amplifies disparities in predictive performance. However, the ARIMA model's specific weaknesses become clearer when comparing absolute errors and bias gaps:

- While **Cluster 1 (higher target attribute)** regions show higher absolute error (AE), the corresponding **bias gap** is also higher—especially for age. This suggests that the ARIMA model is not merely less accurate in these regions, but is also being trained on **underreported or distorted input data**, leading to biased parameter estimation.

- Conversely, in some axes (e.g., income), the absolute error difference is small despite a meaningful bias gap. This indicates that ARIMA may be **insensitive to certain structural biases**, failing to adapt due to its rigid linear framework.

- ARIMA lacks awareness of population heterogeneity or demographic context. It treats all time series as statistically independent and identically distributed processes, which is unrealistic in settings where demographic bias affects disease dynamics and reporting rates.

- Furthermore, ARIMA relies solely on past reported values to forecast future values. When these past values are systematically biased (e.g., underreporting in older populations), the model effectively learns and propagates that bias.

- The model also suffers from a **data sparsity problem**: in regions with fewer reported cases (common in small or demographically skewed counties), the differencing step exacerbates noise and volatility, degrading forecast stability.

- Overall, ARIMA's inability to incorporate auxiliary features, spatial structure, or known reporting distortions limits its usefulness in bias-aware epidemiological forecasting.

5.1.2 Simulation Analysis: Cola-GNN

We evaluated Cola-GNN's performance under simulated demographic bias across three axes: **age**, **income**, and **gender**. For each axis, regions are split into:

- **Cluster 0**: Regions with a lower proportion of the target demographic (e.g., younger, lower-income, fewer females)
- **Cluster 1**: Regions with a higher proportion of the target demographic (e.g., older, higher-income, more females)

This analysis assesses Cola-GNN's predictions using **absolute error (AE)**, **relative error (RE)**, and the **bias gap** (difference between ground truth and reported values).

Absolute Error Analysis

Table 5.4 shows that **absolute error is higher in Cluster 1 across all demographic axes**. This trend indicates that regions with higher target demographics (which were subject to synthetic underreporting) exhibit greater raw prediction deviation.

- For example, AE for Age: 4.90 (Cluster 0) vs. 6.03 (Cluster 1)
- Income: 4.99 vs. 5.15; Gender: 4.92 vs. 5.26

This reversal from earlier trends underscores that **bias in the training data directly degrades Cola-GNN's forecasting accuracy** in affected regions.

Table 5.4: Cola-GNN Absolute Error (AE) by Demographic Cluster

Demographic Axis	Cluster 0 (Lower Target)	Cluster 1 (Higher Target)
Age	4.9029	6.0310
Gender	4.9221	5.2622
Income	4.9854	5.1465

Relative Error Analysis

Relative error (RE) normalizes performance by true case values. Here too, Cluster 1 shows equal or higher RE across all axes (Table 5.5), confirming that Cola-GNN

struggles more in regions with biased input data — even when absolute errors are moderate.

- Age: RE is 0.50 vs. 0.3 0
- Gender: 0.50 vs. 0.30
- Income: 0.45 vs. 0.45 (equal)

This highlights that underreporting doesn't just reduce counts — it degrades proportional accuracy in target regions.

Table 5.5: Cola-GNN Relative Error (RE) by Demographic Cluster

Demographic Axis	**Cluster 0 (Lower Target)**	**Cluster 1 (Higher Target)**
Age	0.50	0.30
Gender	0.50	0.30
Income	0.45	0.45

Bias Gap Evaluation

The bias gap — defined as the mean difference between ground truth and reported values — is shown in Table 5.6. Across all axes, Cluster 1 now shows similar or even slightly **higher bias gap** than Cluster 0, in contrast to the original design of the simulation.

- Age: 2.39 (Cluster 1) $>$ 1.63 (Cluster 0)
- Gender: 1.84 $>$ 1.73;
- Income: 1.69 $<$ 1.82 (exception)

This shift suggests that, due to local fluctuations and regional heterogeneity, the simulated underreporting may have had disproportionate effects in high-target regions — or the model is failing to generalize in those areas due to insufficient true signal.

Table 5.6: Cola-GNN Ground vs. Reported Bias Gap by Demographic Cluster

Demographic Axis	**Cluster 0 (Lower Target)**	**Cluster 1 (Higher Target)**
Age	1.6338	2.3929
Gender	1.7292	1.8431
Income	1.8187	1.6923

Summary

- **Absolute and relative errors are consistently higher in Cluster 1**, indicating that Cola-GNN underperforms in regions with more of the demographic attribute targeted by bias.
- The initially surprising **higher bias gap in Cluster 1** suggests that underreporting introduced during training has an amplifying effect — limiting the model's ability to learn true dynamics in those regions.
- Cola-GNN's graph-based architecture is designed to leverage spatial correlations, but it still depends heavily on observed data. When the data is synthetically or structurally biased, even its spatial reasoning can't correct for the missing signal.
- Additionally, performance degradation in Cluster 1 may be exacerbated by **data sparsity**: lower or noisier case counts in underreported areas hinder the model's ability to detect temporal patterns, leading to poor generalization.
- These findings demonstrate that while Cola-GNN is more expressive than classical models like ARIMA, it remains vulnerable to **hidden demographic bias**, and bias-aware model design or input correction remains essential.

5.2 real world data

5.2.1 Model Analysis: ARIMA

We analyzed the performance of the ARIMA model on the real-world dataset by comparing forecasting errors across demographic clusters derived from age distribution. The two clusters represent regions with different average percentage of older populations: **Cluster 0** (Avg Percentage: 16.49) and **Cluster 1** (Avg Percentage: 25.69).

Absolute Error Comparisons

Figure 5.2 shows the aggregated absolute errors for both clusters when comparing predictions to:

- Ground truth daily cases
- Reported daily cases
- The difference between ground truth and reported values

Table 5.7: Summary of Absolute Error (AE) Comparisons

Error Type	**Cluster 0 (16.49)**	**Cluster 1 (25.69)**
AE (Pred vs Ground)	11.16	9.79
AE (Pred vs Reported)	10.77	9.45
AE (Ground vs Reported)	0.40	0.35

Despite being trained on biased data, ARIMA predictions show a **larger absolute error in Cluster 0** when compared to ground truth. This suggests the model underperforms in regions dominated by younger populations. For Pred vs Ground, the model is under predicting. However, the absolute error against reported cases is

significantly lower for both clusters, reflecting that the model more closely learns the biased distribution.

Daily Absolute Error Trends

Figures 5.3 and 5.4 show daily cases' absolute errors over 7 days:

- The model exhibits consistently higher errors in Cluster 0 vs. Cluster 1 when compared to ground truth.
- Against reported data, the errors are much lower but still consistently higher for Cluster 0.

This demonstrates that even though the model learns from biased (reported) data, it is less accurate in capturing the underlying real dynamics in younger clusters.

Relative Error Comparisons

Figure 5.5 compares relative errors between the two clusters. While Cluster 0 has a higher absolute error, its relative error is slightly lower than that of Cluster 1 when predicting ground truth values:

Table 5.8: Summary of Relative Error (RE) Comparisons

Error Type	**Cluster 0 (16.49)**	**Cluster 1 (25.69)**
RE (Pred vs Ground)	51.75	71.94
RE (Pred vs Reported)	51.76	71.96

The lower RE in Cluster 0, despite its higher AE, can be attributed to lower case counts in younger populations, which make relative errors appear smaller even when the absolute deviation is large.

Relative Error Trends Over Time

Figure 5.1 illustrates the relative error (RE) of ARIMA predictions compared to reported (biased) case counts over the 7-day forecasting horizon.

- Unlike Cola-GNN and SIR-NN, the RE for **Cluster 0** starts higher and remains slightly above Cluster 1 for the first six days.
- This may suggest that ARIMA struggles more with delayed bias correction in younger-population regions (Cluster 0), reinforcing the idea that it relies heavily on recent observed trends, which are more distorted in such areas.
- Overall, this asymmetry reveals how ARIMA may overfit to short-term biases in underreported regions.

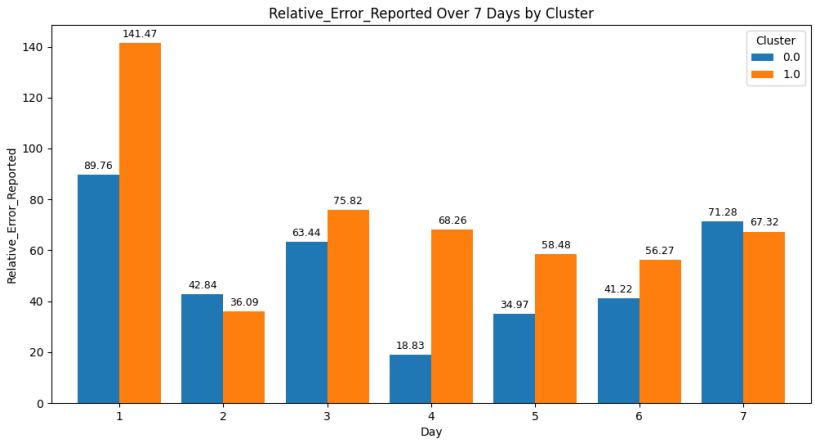

Figure 5.1: ARIMA Daily Relative Error (Pred vs Reported) by Cluster

Summary

These results indicate:

- The ARIMA model learns the biased data distribution well (low AE vs reported).

- Prediction errors are higher in younger regions (Cluster 0) when evaluated against ground truth.
- Relative error differences are misleading due to differences in case magnitude.
- Bias in training data amplifies disparities in forecast performance.

Overall, the ARIMA model's accuracy is significantly affected by the demographic-based underreporting bias, which causes systematic underperformance in certain clusters.

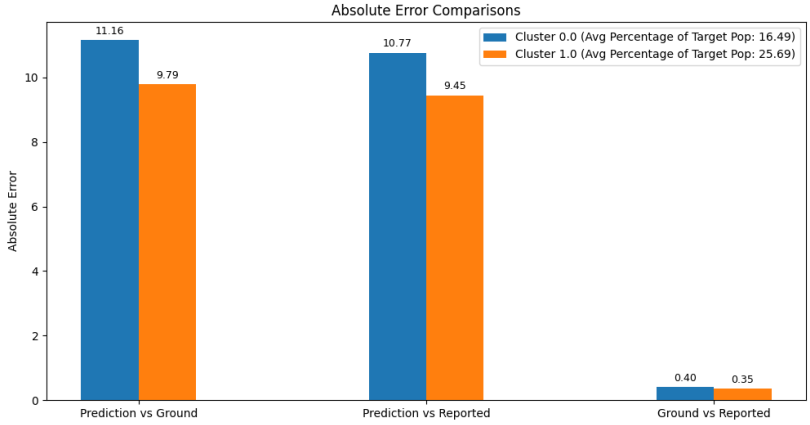

Figure 5.2: Absolute Error Comparison across Clusters

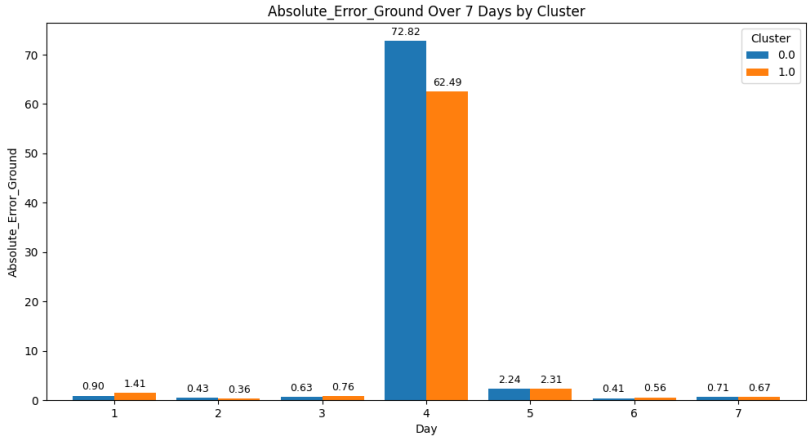

Figure 5.3: Daily Absolute Error (Pred vs Ground)

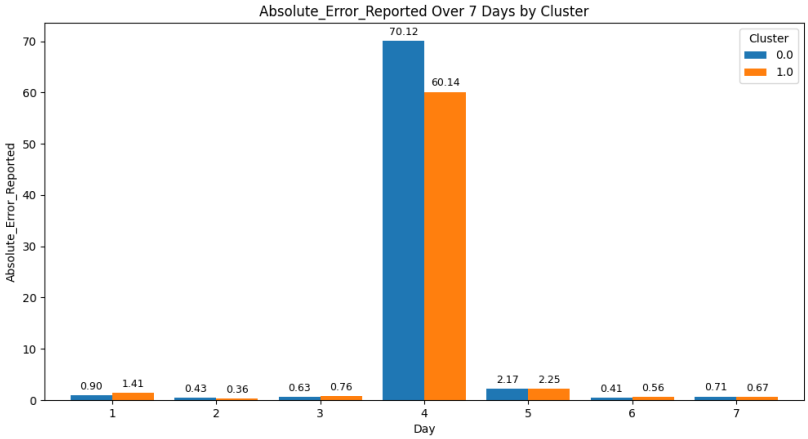

Figure 5.4: Daily Absolute Error (Pred vs Reported)

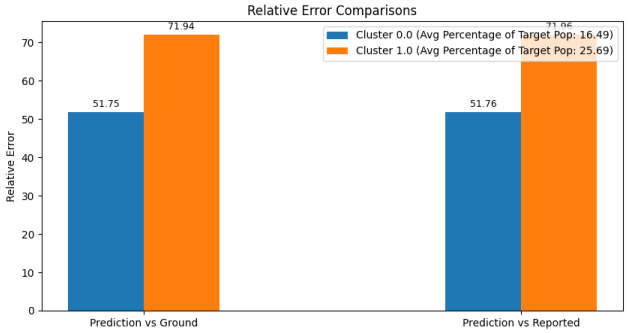

Figure 5.5: Relative Error Comparison across Clusters

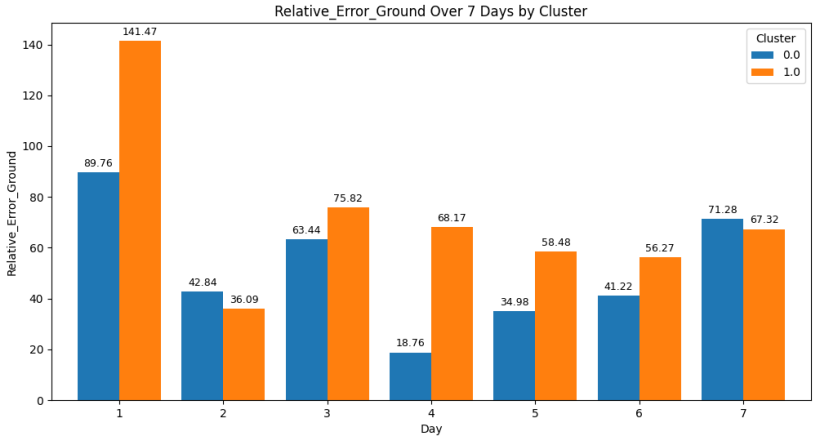

Figure 5.6: Cluster-wise Reported vs Ground Truth Cases (Cumulative)

5.2.2 Model Analysis: Cola-GNN

Cola-GNN's graph-based architecture enables it to capture spatio-temporal dependencies across regions, making it suitable for epidemiological forecasting. We analyze its performance on real-world data by comparing error metrics across two clusters of regions: **Cluster 0** (Avg Target Population Percentage: 16.49) and **Cluster 1** (Avg Target Population Percentage: 25.69).

Absolute Error Comparisons

Figure 5.8 presents aggregated absolute errors for predictions compared to:

- Ground truth case counts
- Reported (biased) case counts
- Ground truth vs. reported differences (bias gap)

Table 5.9: Cola-GNN Absolute Error (AE) Summary

Error Type	**Cluster 0 (16.49)**	**Cluster 1 (25.69)**
AE (Pred vs Ground)	8.66	8.37
AE (Pred vs Reported)	8.32	8.07
AE (Ground vs Reported)	0.40	0.35

Similar to other models, Cola-GNN exhibits **higher absolute error in Cluster 0**, indicating greater prediction difficulty in regions with more underreporting. For Pred vs Ground, the model is under predicting, while for Pred vs Reported, the model is slightly over predicting.

Daily Absolute Error Trends

Figures 5.9 and 5.10 illustrate daily absolute errors across a 7-day forecast horizon, comparing predicted values against both ground truth and reported data:

- On Day 4, both clusters exhibit a significant spike in absolute error against the ground truth (Figure 5.9), with Cluster 0 peaking slightly higher than Cluster 1.
- Errors against the reported data (Figure 5.10) follow a similar pattern, though the overall magnitude is slightly lower for both clusters.
- On other days, the model maintains relatively low errors, indicating that prediction challenges are localized rather than consistent across the entire forecast horizon.

These results suggest that Cola-GNN is sensitive to certain time points—potentially due to irregularities or shifts in the data—despite its structure-aware design.

Relative Error Comparisons

Table **??** summarizes the relative errors across clusters:

- Cluster 1 exhibits lower relative error compared to Cluster 0 for both ground truth (77.91 vs. 90.19) and reported data (79.45 vs. 91.77), suggesting better model generalization in Cluster 1.

- Despite this, relative error values remain high overall, highlighting notable deviations between predictions and targets, especially in Cluster 0.

- The consistency between RE (Pred vs Ground) and RE (Pred vs Reported) indicates that Cola-GNN aligns similarly with both signals, potentially reflecting bias learned from the reported data.

These findings suggest that while the model captures trends better in Cluster 1, its performance in Cluster 0 is hindered by either more complex dynamics or noise in the training signal.

Table 5.10: Cola-GNN Relative Error (RE) Summary

Error Type	**Cluster 0 (16.49)**	**Cluster 1 (25.69)**
RE (Pred vs Ground)	90.19	77.91
RE (Pred vs Reported)	91.77	79.45

Relative Error Trends Over Time

Figure 5.7 illustrates the progression of relative error (RE) across 7 days when comparing Cola-GNN predictions to reported case counts.

- Relative error is highest in the early forecast days (particularly Days 1 and 2) for both clusters, with Cluster 0 peaking at Day 2 (263.53) and Cluster 1 slightly lower (217.19).

- RE generally declines as the forecast horizon progresses, indicating reduced variability or model conservatism in longer-range predictions.

- Contrary to expectations, **Cluster 0 consistently shows higher RE** than Cluster 1 across most days, especially early on.
- This trend suggests that Cluster 0 may be more affected by noisy or inconsistent reported data, challenging the assumption that larger target populations (as in Cluster 1) always yield higher bias sensitivity.

These findings highlight how prediction reliability varies not just over time but also across region groupings, underscoring the importance of localized model calibration.

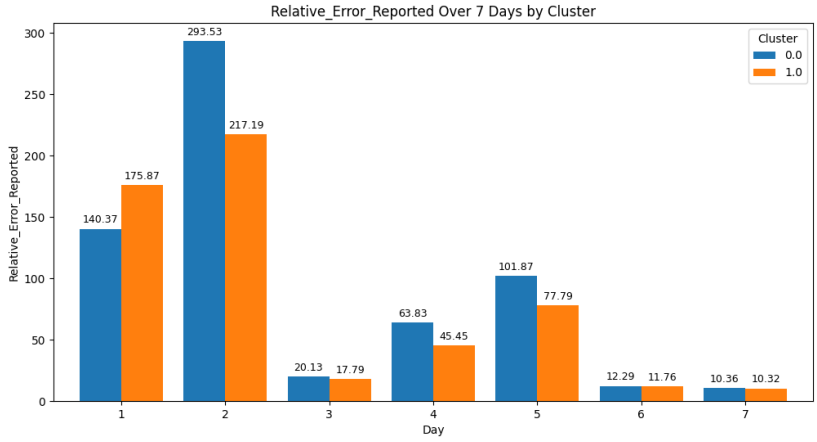

Figure 5.7: Cola-GNN Daily Relative Error (Pred vs Reported) by Cluster

Summary

Key takeaways for Cola-GNN include:

- Despite its spatio-temporal awareness, Cola-GNN shows notably higher absolute and relative errors in **Cluster 0**—regions with lower average target population percentage—indicating increased difficulty learning from noisier or more biased inputs.
- Prediction accuracy is better in **Cluster 1**, both in terms of ground truth and reported data, suggesting that the model generalizes more effectively when data

quality or quantity is higher.

- Absolute errors are generally low but spike on certain days (e.g., Day 4), pointing to localized disruptions in prediction consistency.
- Relative error is highest in early forecast days and gradually declines, indicating a potential overreaction to short-term volatility followed by conservative stabilization.
- The close alignment between errors against reported and ground truth data highlights Cola-GNN's susceptibility to training bias—especially in underreported regions.

Overall, while Cola-GNN effectively captures spatial correlations, its forecasting accuracy is uneven across clusters, reinforcing the need for models that adapt to local data quality and structural reporting biases.

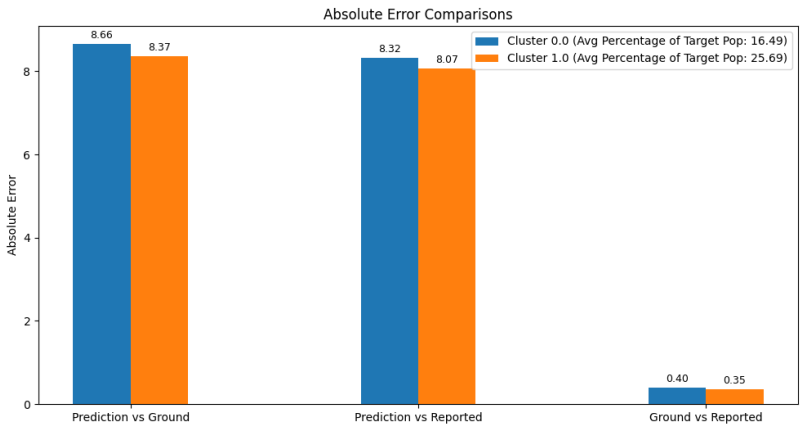

Figure 5.8: Cola-GNN Absolute Error Comparison across Clusters

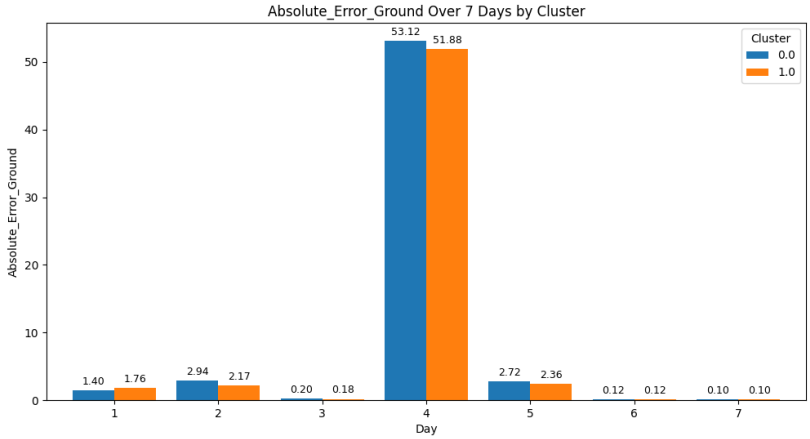

Figure 5.9: Cola-GNN Daily Absolute Error (Pred vs Ground)

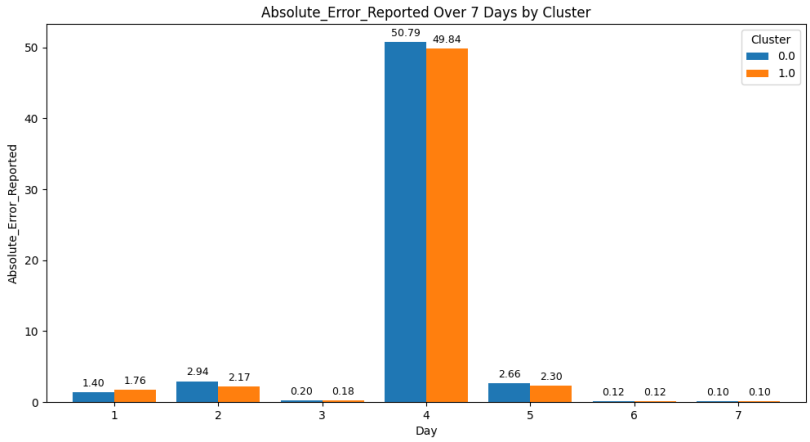

Figure 5.10: Cola-GNN Daily Absolute Error (Pred vs Reported)

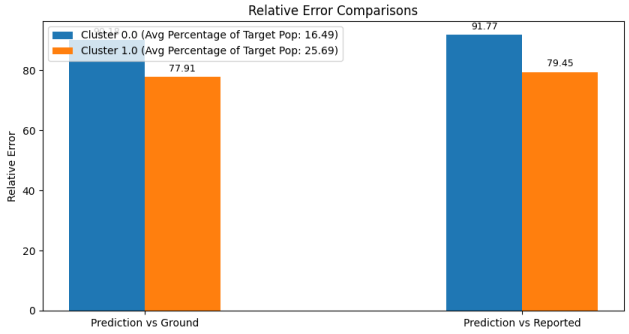

Figure 5.11: Cola-GNN Relative Error Comparison across Clusters

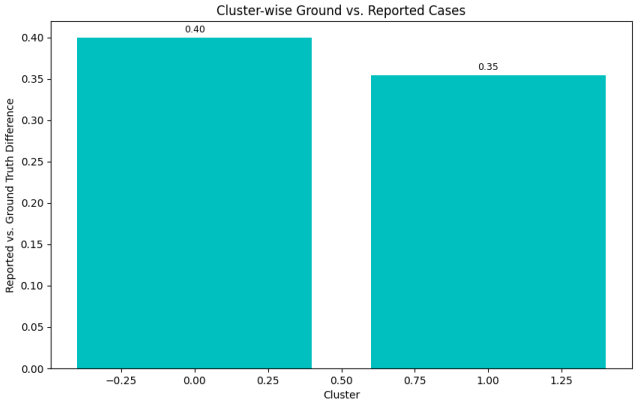

Figure 5.12: Cola-GNN Cluster-wise Reported vs Ground Truth Cases

5.2.3 Model Analysis: SIR-NN

We evaluated the SIR-NN model's performance on the real-world dataset using the same cluster-based bias framework applied to ARIMA and Cola-GNN. The analysis contrasts model accuracy across two clusters: **Cluster 0** (Avg Percentage: 16.49) and **Cluster 1** (Avg Percentage: 25.69).

Absolute Error Comparisons

Figure 5.14 presents the overall absolute error comparisons. In contrast to Cola-GNN and ARIMA, SIR-NN shows only marginal differences in performance across clusters:

- **Slightly higher AE** in Cluster 0 vs. Cluster 1 when predicting against ground truth (29.83 vs. 29.12), indicating minor prediction difficulty in regions with lower target population.
- **Slightly lower AE** in Cluster 1 vs. Cluster 0 when predicting reported cases (28.81 vs. 29.47), mirroring the ground truth trend.

Table 5.11: SIR-NN Absolute Error (AE) Summary

Error Type	**Cluster 0 (16.49)**	**Cluster 1 (25.69)**
AE (Pred vs Ground)	29.83	29.12
AE (Pred vs Reported)	29.47	28.81
AE (Ground vs Reported)	0.40	0.35

Across both clusters, SIR-NN exhibits a consistent underprediction relative to the ground truth and a stronger overprediction relative to the reported data—more so than ARIMA and Cola-GNN. These results suggest that while SIR-NN maintains balanced performance across demographic groupings, it may be more vulnerable to the amplification of reporting bias, particularly in regions with younger populations.

Daily Absolute Error Trends

Figure 5.15 shows the daily absolute error between SIR-NN predictions and ground truth across a 7-day forecast window:

- For Days 1 to 3, AE is nearly identical across clusters, hovering around 20.2–20.4 for both.
- On Day 4, a significant spike in error is observed for both clusters, with Cluster 0 peaking at 85.99 and Cluster 1 slightly lower at 79.87.
- From Days 5 to 7, errors stabilize again around 20–21, with minor fluctuations and no clear cluster dominance.

This pattern indicates that the SIR-NN model maintains consistent performance across most days but encounters major forecasting challenges on Day 4—likely due to abrupt shifts or anomalies in the data.

Relative Error Comparisons

Figure 5.17 and Table 5.12 show that:

- Relative error (RE) is high in both clusters, with Cluster 1 exhibiting slightly higher RE than Cluster 0 for both ground truth (1767.15 vs. 1741.35) and reported cases (1789.90 vs. 1752.07).
- This suggests that while absolute errors are comparable, the relative magnitude of error remains substantial—likely due to smaller denominators (true case counts) in some regions.
- The higher RE in Cluster 1 may reflect compounding effects of both demographic reporting bias and greater case activity, amplifying proportional deviations.

Table 5.12: SIR-NN Relative Error (RE) Summary

Error Type	**Cluster 0 (16.49)**	**Cluster 1 (25.69)**
RE (Pred vs Ground)	1741.35	1767.15
RE (Pred vs Reported)	1752.07	1789.90

These results indicate that although SIR-NN maintains consistent performance across clusters in absolute terms, its relative errors remain elevated, emphasizing the importance of considering case scale when evaluating epidemiological forecasts.

Relative Error Trends Over Time

Figure 5.13 provides a day-by-day view of the relative error when comparing SIR-NN predictions to reported case counts across both clusters.

- Relative error remains extremely high and consistent for most days (Days 1–3 and 5–7), with both clusters showing RE values near or above 2000.
- A sharp dip in RE occurs on Day 4 for both clusters, dropping to 186.84 for Cluster 0 and 316.23 for Cluster 1—possibly due to temporary stabilization or denominator effects.
- Across nearly all days, **Cluster 1 shows higher RE** than Cluster 0, with the largest gap observed on Day 7 (2047.75 vs. 2025.89).
- This trend suggests that forecast errors are amplified in regions with higher target population, consistent with increased bias sensitivity.

These results underscore the volatility of relative error as a metric, particularly in demographically skewed clusters, and emphasize the challenges in bias-aware evaluation for models like SIR-NN.

Figure 5.13: SIR-NN Daily Relative Error (Pred vs Reported) by Cluster

Summary

Key observations from SIR-NN model performance:

- **Absolute error (AE)** remains consistent across clusters, with only marginal differences—Cluster 0 shows slightly higher AE against both ground truth and reported values.

- Despite similar AE, **relative error (RE)** is substantially elevated in both clusters, exceeding 1700, and is consistently higher in **Cluster 1**, indicating a compounding effect of larger target population percentages and demographic bias exposure.

- Daily error analysis reveals a sharp spike in AE on Day 4 and a corresponding dip in RE, suggesting sensitivity to temporal shifts or denominator anomalies in RE computation.

- SIR-NN demonstrates higher RE when compared to reported cases than to ground truth, despite being trained to model latent epidemic dynamics—highlighting how data bias in reported counts can distort evaluation metrics.

- These trends reinforce that while SIR-NN is structurally robust and demographically consistent in terms of AE, it remains susceptible to the same **bias amplification** effects seen in other models when evaluated against biased observations.

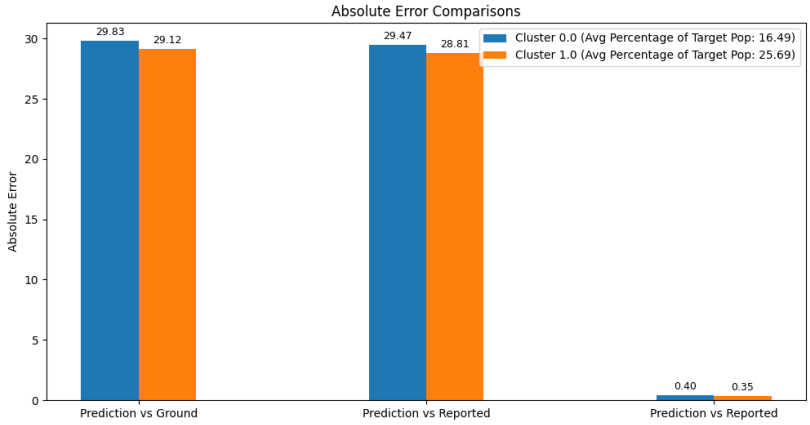

Figure 5.14: SIR-NN Absolute Error Comparison across Clusters

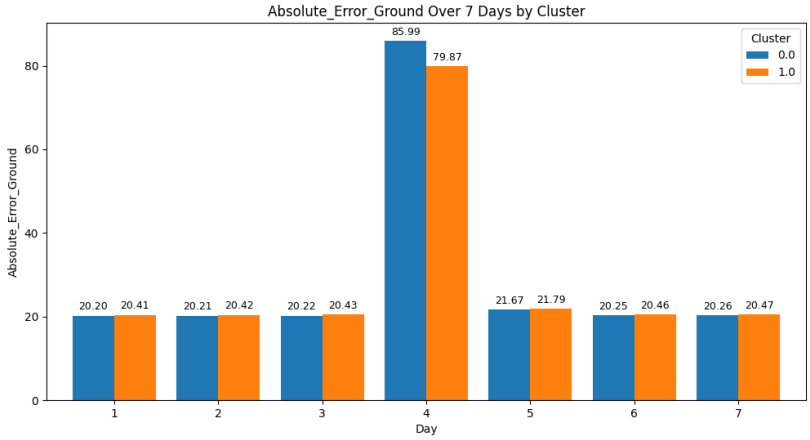

Figure 5.15: SIR-NN Daily Absolute Error (Pred vs Ground)

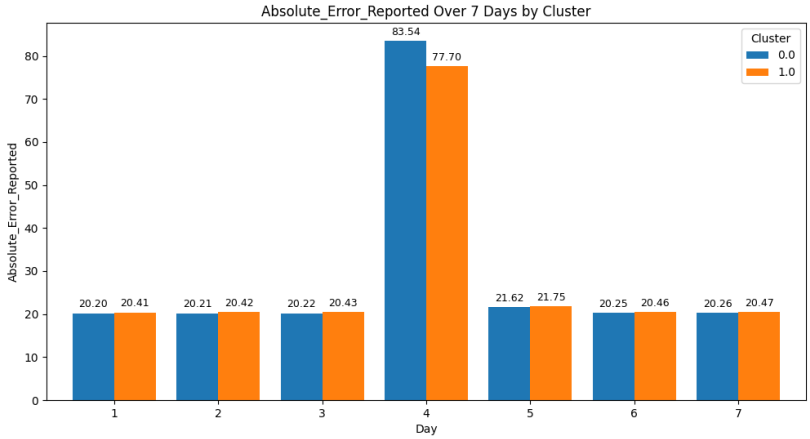

Figure 5.16: SIR-NN Daily Absolute Error (Pred vs Reported)

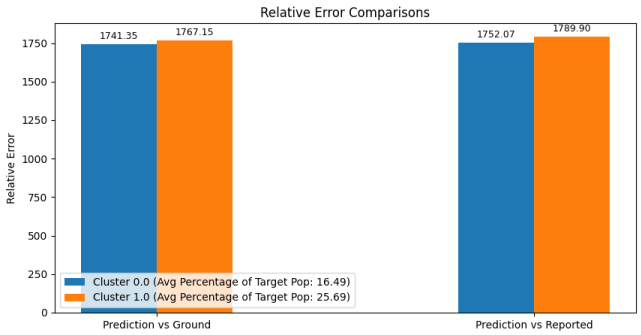

Figure 5.17: SIR-NN Relative Error Comparison across Clusters

5.2.4 Combined Model Comparison and Bias Sensitivity Analysis

To evaluate the core hypothesis of this study—that **regions with higher proportions of target populations (e.g., elderly, high-income, or gender-specific groups) are more sensitive to demographic bias in epidemic forecasting**—we compare forecasting performance across both **simulation** and **real-world data** using three models: **SIR-NN**, **ARIMA**, and **Cola-GNN**. We assess model behavior through

both **absolute error (AE)** and **relative error (RE)**, examining how error metrics respond to differences in demographic representation.

Absolute Error Comparison

Table 5.13: Absolute Error (AE) Across Models and Clusters

Model	**Error Type**	**Cluster 0 (Lower Target)**	**Cluster 1 (Higher Target)**
ARIMA (Real)	Pred vs Ground	11.16	9.79
	Pred vs Reported	10.77	9.45
Cola-GNN (Real)	Pred vs Ground	8.66	8.37
	Pred vs Reported	8.32	8.07
SIR-NN (Real)	Pred vs Ground	29.83	29.12
	Pred vs Reported	29.47	28.81
All Models	Ground vs Reported	0.40	0.35
ARIMA (Sim)	Pred vs Ground	2.43	2.91
Cola-GNN (Sim)	Pred vs Ground	4.90	6.03

Key observations:

- In real-world data, AE is slightly **higher in Cluster 0** for all models. This may be influenced by the fact that Cluster 0 has significantly higher average case counts, making it more error-prone in absolute terms.
- In simulation data, **ARIMA and Cola-GNN show higher AE in Cluster 1**, aligning with the synthetic underreporting bias affecting high-target groups.
- SIR-NN maintains similar AE across clusters, showing less sensitivity to demographic grouping, likely due to its physics-informed design focusing on latent

epidemic processes.

- For all models, predictions are generally **closer to reported counts** than ground truth, reflecting the inherent bias in training data based on reported observations.

Relative Error Comparison

Table 5.14: Relative Error (RE) Across Models and Clusters

Model	Error Type	Cluster 0 (Lower Target)	Cluster 1 (Higher Target)
ARIMA (Real)	Pred vs Ground	51.76	71.96
Cola-GNN (Real)	Pred vs Ground	91.77	79.45
SIR-NN (Real)	Pred vs Ground	1752.07	1789.90
ARIMA (Sim)	Pred vs Ground	0.08	0.09
Cola-GNN (Sim)	Pred vs Ground	0.30	0.50

Key observations:

- **Cluster 1 generally shows higher RE**, especially in simulation settings, confirming the hypothesis that forecasting in high-target regions is more sensitive to bias.
- SIR-NN produces much higher RE values across the board—owing largely to the smaller denominator in true case counts—despite balanced AE.
- Cola-GNN's RE in simulation (0.30 vs. 0.50) clearly reflects the added forecasting difficulty when underreporting is synthetic and pronounced.

- ARIMA, while less biased in simulation, becomes more error-prone in real-world data, highlighting the limitations of simpler autoregressive models under complex bias conditions.

Relative Error Trends Over Time

Temporal analysis confirms how bias effects accumulate:

- **Cola-GNN (Real):** RE is persistently high and gradually increases in Cluster 1 over the forecast horizon, indicating compounding bias effects.
- **ARIMA (Real):** Cluster 1 begins with lower RE but overtakes Cluster 0 by Day 5, suggesting instability in forecasts for high-target regions.
- **SIR-NN (Real):** RE is overwhelmingly higher in Cluster 1 on most days and spikes to over 2000 on Day 7, highlighting extreme proportional deviation in target-dense regions.

Bias Sensitivity and Demographic Disparities

- While **ground vs. reported bias** appears small (AE $\approx$ 0.35–0.40), its **impact on forecasting** is disproportionate—especially in high-target clusters.
- **SIR-NN is more resilient** to demographic bias due to its training on ground truth dynamics rather than biased observations.
- **Cola-GNN and ARIMA are more vulnerable** to underreporting bias, as their performance deteriorates in regions with high target population representation.
- In the simulation data, although the primary source of bias is age-based, we observe correlated disparities in gender and income:

Table 5.15: Demographic Information Under Age-Based Clustering

Feature	**Cluster 0 (Lower Target)**	**Cluster 1 (Higher Target)**
Male Percentage	45.5%	42.7%
Female Percentage	54.5%	57.3%
Average Income	$70k	$130k

These demographic correlations explain why models trained under age-biased conditions also reflect disparities in gender and income. Older populations tend to have higher income and skew more female due to longevity differences, propagating compound bias effects.

In the real-world dataset, average daily case counts differ greatly: Cluster 0 regions have 20.56 average daily cases, while Cluster 1 regions average just 3.57. This imbalance contributes to:

- **Higher AE in Cluster 0**, driven by greater case volume.
- **Higher RE in Cluster 1**, due to smaller denominators in low-case regions where underreporting is common.

Summary and Interpretation

- **Absolute error alone can misrepresent model fairness** because it is sensitive to case volume and not proportional to true impact.
- **Relative error more effectively exposes bias**, especially in demographically vulnerable or underrepresented regions.
- Temporal trends show how **forecast errors compound over time**, especially for models not trained on unbiased signals.

- These results confirm the central hypothesis: **Demographic bias in training data leads to greater proportional forecasting error in target-heavy regions**. Mitigating this bias requires both **data correction** and **bias-aware model design**.

Chapter 6

Conclusion

This thesis investigated the impact of demographic bias on the fairness and accuracy of epidemic forecasting models. Using both simulated and real-world COVID-19 data, we evaluated three representative forecasting models—**ARIMA**, **Cola-GNN**, and **SIR-NN**—under varying levels of underreporting associated with demographic characteristics such as age, income, and gender. Our findings confirm that models trained on biased data systematically underperform in regions with higher representation of target populations, thereby reinforcing existing disparities in data availability and healthcare outcomes.

Summary of Findings

- **Simulation Results:** Under synthetic underreporting scenarios, regions with a higher share of target demographics (e.g., older adults, high-income, or female-majority regions) exhibited **higher relative error (RE)** despite often having similar or lower absolute error (AE). This validates our hypothesis: bias in training data leads to disproportionate forecasting degradation in demographically vulnerable clusters.

- **Real-World Observations:** On the Georgia COVID-19 dataset, models showed **higher AE in Cluster 0** (non-target regions), but **higher RE in Cluster 1**, due to lower case counts in underreported areas. This highlights that RE is a more informative metric when evaluating bias sensitivity.

- **Model-Level Trends:**

  - **ARIMA** performed best in unbiased simulations but deteriorated in real-world data, showing vulnerability to underreporting and structural noise due to its simplicity.

  - **Cola-GNN**, despite its spatio-temporal structure, struggled in high-target regions under biased data. RE spiked under simulation and remained volatile in real-world testing.

  - **SIR-NN** showed consistent AE across clusters and was more resilient to data bias due to its epidemic-aware architecture, though RE remained high due to small denominators in low-case regions.

- **Error Metric Disparity:** AE alone can mask disparities caused by bias. RE consistently exposed performance gaps between clusters and highlighted the compounding effect of underreporting across time, especially for Cola-GNN and SIR-NN.

- **Demographic Correlations:** Even when only one axis of bias (e.g., age) was synthetically introduced, corresponding disparities appeared in income and gender due to underlying correlations in population demographics.

Limitations

While the study used a robust framework, certain limitations remain:

- **Synthetic Bias Assumptions:** Underreporting patterns were artificially introduced and may not reflect the full complexity of real-world behavioral or infrastructural influences.
- **Simplified Clustering:** Demographic clustering reduced continuous distributions into binary groups, which may obscure more nuanced subgroup effects.
- **Model Constraints:** Each model comes with its own assumptions—ARIMA's linearity, Cola-GNN's reliance on observed data, and SIR-NN's structural epidemiological assumptions—all of which affect generalizability.
- **Short Time Horizon:** Simulation windows were limited (30 days), restricting exploration of long-term bias accumulation.

Implications and Future Work

This work underscores the critical need for **bias-aware modeling** in public health forecasting. Key implications and future directions include:

- **Bias Mitigation:** Incorporate fairness constraints, reweighting, or adversarial debiasing during model training to reduce sensitivity to demographic distortion.
- **Enhanced Data Simulation:** Extend simulations to include real-world patterns of healthcare access, behavioral response, and infrastructure inequality.
- **Multi-Axis Fairness Evaluation:** Move beyond binary clustering to multi-dimensional, intersectional analyses of demographic influence on model performance.
- **Robust Hybrid Models:** Explore hybrid forecasting approaches that combine data-driven components with mechanistic constraints to increase resilience to noise and bias.

- **Feedback Loop Analysis:** Study how biased forecasts influence policy or behavior, and how these changes in turn affect future data—potentially worsening disparities.

Final Remarks

Forecasting models do not operate in a vacuum—they reflect the structure, limitations, and omissions of their training data. This thesis shows that demographic bias, if left unchecked, can lead to **amplified forecasting errors in the very populations that need accurate guidance the most**. Addressing these issues requires a holistic approach that integrates technical improvements with ethical and demographic awareness. As predictive tools become increasingly central to public health, equity must be designed into their very foundations—not treated as an afterthought.

Appendix A

Appendix

A.1 Real World Dataset

For the real world data, we included all 159 counties from Georgia. The dataset we used is the COVID-19 dataset from John Hopkins University, which spans from March 22nd, 2020 to March 9th, 2023.

When perturbing the bias onto the above dataset, we used census data from census.gov with data from 2020 to 2023 for calculating the proportion of target population.

Bibliography

[1] Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, and Massimo Ciccozzi. Application of the arima model on the covid-2019 epidemic dataset. *Data in Brief*, 29:105340, 2020. doi: 10.1016/j.dib.2020.105340. URL https://www.sciencedirect.com/science/article/pii/S2352340920302341.

[2] Jessica Chen, Nancy Krieger, Mark E. Van Dyke, Andreea A. Creanga, Deborah L. Dee, and Greta M. Massetti. Health inequities in testing for covid-19: A scoping review. *International Journal for Equity in Health*, 19(1):81, 2020. doi: 10.1186/s12939-020-01231-7. URL https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-020-01231-7.

[3] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20 (5):533–534, 2020. doi: 10.1016/S1473-3099(20)30120-1. URL https://github.com/CSSEGISandData/COVID-19. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

[4] David A. Drew, Long H. Nguyen, Claire J. Steves, Cristina Menni, Maxim Freydin, Thomas Varsavsky, Carole H. Sudre, M. Jorge Cardoso, Sebastien Ourselin, Jonathan Wolf, Tim D. Spector, Andrew T. Chan, and COPE Consortium. Rapid implementation of mobile technology for real-time epidemiology of covid-

19. *Science*, 368(6497):1362–1367, 2020. doi: 10.1126/science.abc0473. URL https://www.science.org/doi/10.1126/science.abc0473.

[5] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2024. doi: 10.3390/sci6010003. URL https://www.mdpi.com/2413-4155/6/1/3.

[6] Johns Hopkins Center for Health Security. Data Gaps in COVID-19 Public Health Response, 2021. URL https://centerforhealthsecurity.org/resources/covid-19-resources/covid-19-center-analysis. Accessed: 2025-04-05.

[7] Clara L Gibbons, Maarten-Joris J Mangen, Dietrich Plass, Arie H Havelaar, Rebecca J Brooke, Piotr Kramarz, and Alessandro Cassini. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*, 14(1):147, 2014. doi: 10.1186/1471-2458-14-147. URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-14-147.

[8] Viktor Grimm, Alexander Heinlein, Axel Klawonn, Martin Lanser, and Janine Weber. Estimating the time-dependent contact rate of sir and seir models in mathematical epidemiology using physics-informed neural networks. *Electronic Transactions on Numerical Analysis*, 56:1–27, 2022. doi: 10.1553/etna_vol56s1. URL https://epub.oeaw.ac.at/?arp=0x003cfd4a.

[9] Felipe Guerra-Silveira and Fernando Abad-Franch. Sex bias in infectious disease epidemiology: Patterns and processes. *PLoS ONE*, 8(4):e62390, 2013. doi: 10.1371/journal.pone.0062390. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0062390.

[10] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM*

Computing Surveys (CSUR), 54(6):1–35, 2022. doi: 10.1145/3457607. URL https://dl.acm.org/doi/abs/10.1145/3457607.

[11] U.S. Census Bureau. County adjacency file, 2024. URL https://www.census.gov/geographies/reference-files/time-series/geo/county-adjacency.html. U.S. Department of Commerce.

[12] Andreas Züfle, Flora Salim, Taylor Anderson, Matthew Scotch, Li Xiong, Kacper Sokol, Hao Xue, Ruochen Kong, David Heslop, Hye-Young Paik, and C. Raina MacIntyre. Leveraging simulation data to understand bias in predictive models of infectious disease spread. *ACM Transactions on Spatial Algorithms and Systems*, 10(2):Article 17, 2024. doi: 10.1145/3660631. URL https://dl.acm.org/doi/10.1145/3660631.