# Project B: Hybrid Forecasting and Sequential Intervention Assessment

**Author:**Emmanuel Adeleke, Olukayode Akinyode, Gbolahan Abioye & Nasif Ajilore

**Date:** December 2, 2025

**University:** Morgan State University

---

## Abstract

This project combines epidemiological forecasting with online change detection to assess the effectiveness of COVID-19 interventions (lockdowns, vaccination campaigns) in the United States. We apply SARIMA models, state-space Poisson/Negative Binomial models for weekly incidence forecasting, implement CUSUM and Bayesian Online Change-Point Detection (BOCPD) methods on forecast residuals to detect structural changes, and conduct counterfactual analysis to quantify policy impacts. Using counterfactual scenarios, we estimate avoided cases and simulate alternative policy timings (earlier vaccination vs. delayed lockdown). The analysis demonstrates that interventions significantly reduced COVID-19 cases and that timing of policy implementation is critical for epidemic control.

---

## Introduction

Understanding the effectiveness of public health interventions during the COVID-19 pandemic is crucial for future epidemic preparedness. This research addresses three key questions:

1. **Forecasting**: How accurately can we predict COVID-19 incidence using time series models?

2. **Change Detection**: Can we detect when interventions caused structural changes in disease transmission?

3. **Policy Impact**: What would have happened without interventions, and how does timing matter?

By combining forecasting models with sequential change detection and counterfactual analysis, we provide a comprehensive framework for evaluating intervention effectiveness. This approach benefits public health officials, policymakers, and researchers who need evidence-based assessments of epidemic control measures.

---

# Methodology

## Data Sources

- **Primary Dataset**: Johns Hopkins University (JHU) CSSE COVID-19 Data Repository

- USA Weekly COVID-19 Cases (January 2020 - March 2023)

- Source: `https://github.com/CSSEGISandData/COVID-19`

- Processing: Aggregated from daily cumulative cases to weekly incident cases

## Part 1: Forecasting Models

### 1.1 SARIMA Model

Seasonal Autoregressive Integrated Moving Average model:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

With seasonal component: SARIMA(1,1,1)×(1,1,1,52)

### 1.2 State-Space Poisson Model

$$y_t \sim \text{Poisson}(\lambda_t), \quad \log \lambda_t = \alpha + \beta t + u_t$$

where $u_t$ is a latent trend component.

### 1.3 State-Space Negative Binomial Model

$$y_t \sim \text{Negative Binomial}(\mu_t, r), \quad \log \mu_t = \alpha + \beta t + u_t$$

where $r$ is the dispersion parameter.

**Validation Metrics:**

- **RMSE**: $\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$

- **MAE**: $\frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$

- **MAPE**: $\frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$

## Part 2: Online Change Detection

### 2.1 CUSUM Test

Cumulative Sum test applied to SARIMA residuals:

$$S_t = \max(0, S_{t-1} + (x_t - k))$$

Signal when $S_t > h$, where $k = 0.5\sigma$ and $h = 5\sigma$.

### 2.2 Bayesian Online Change-Point Detection (BOCPD)

Recursive Bayesian update of run-length distribution for change-point detection, based on Adams & MacKay (2007).

## Part 3: Counterfactual Analysis

### 3.1 No-Intervention Scenario

Estimate pre-intervention growth rate from exponential model:

$$y_t = y_0 \exp(rt)$$

Project forward without policy interventions, capped at US population (335 million).

### 3.2 Avoided Cases Calculation

$$\Delta y_t = \tilde{y}_t^{\text{no policy}} - y_t^{\text{observed}}$$

Total avoided cases: $\sum_t \max(0, \Delta y_t)$

## Part 4: Policy Scenario Simulations

### 4.1 Earlier Vaccination Scenario

Simulate vaccination starting 2 months earlier (October 2020 instead of December 2020).

### 4.2 Delayed Lockdown Scenario

Simulate lockdown starting 2 months later (May 2020 instead of March 2020).

---

# Results

## Part 1: Forecasting Model Comparison

| Model | Value 1 | Value 2 | Percentage |
|---|---|---|---|
| SARIMA | 221,049 | 210,016 | 82.95% |
| Poisson State-Space | 696,691 | 693,658 | 276.79% |
| Negative Binomial State-Space | 998,235 | 995,260 | 395.44% |

| Model | RMSE | MAE | MAPE (%) |
|---|---|---|---|
| Negative Binomial State-Space | 998,235 | 995,260 | 395.44% |

**Key Findings:**

- **SARIMA** achieved the best performance with lowest RMSE (221,049), MAE (210,016), and MAPE (82.95%)

- State-space models showed higher error rates, likely due to overfitting or parameter estimation challenges

- All models captured the general trend but SARIMA best handled the seasonal patterns and extreme values

- All models captured the general trend but varied in handling extreme values

- Seasonal patterns were well-captured by SARIMA model

## Part 2: Change Point Detection

**CUSUM Results:**

- Number of change points detected: Multiple structural breaks identified

- Change points correspond to major intervention dates (lockdowns, vaccination rollout)

- CUSUM statistic exceeded threshold at key policy implementation dates

**BOCPD Results:**

- Number of change points detected: Multiple change points identified probabilistically

- Probabilistic approach identified periods of structural stability and change

- Run-length distribution shows clear reset points corresponding to interventions

**Visualization:** See `images/cusum_change_detection.png` and `images/bocpd_change_detection.png`

### Part 3: Counterfactual Analysis

**No-Intervention Scenario:**

- Total avoided cases: Significant (values capped at US population of 335 million)

- Average weekly avoided cases: Substantial reduction achieved

- Peak reduction: 100% (interventions prevented peak from reaching counterfactual levels)

**Key Insight:** Interventions prevented a substantial number of COVID-19 cases. Without policy measures, the epidemic would have grown exponentially, quickly reaching population limits.

**Visualization:** See `images/counterfactual_analysis.png`

### Part 4: Policy Scenario Simulations

| Observed (Actual) | 103,806,563 | Baseline | 5,648,740 |

| Earlier Vaccination | 40,180,381 | -61.3% | 1,977,059 |

| Delayed Lockdown | 44,353,721 | +57.3% | 1,977,059 |

| Scenario | Observed (Actual) | Baseline | Change from Baseline |
|---|---|---|---|
| **Total** | 103,806,563 | 5,648,740 | |
| Earlier Vaccination | 40,180,381 | 1,977,059 | -61.3% |
| Delayed Lockdown | 44,353,721 | 1,977,059 | +57.3% |

Scenario Total Cases Change vs Observed Peak Weekly Cases

---

| Scenario | Total Cases | Change vs Observed Peak | Weekly Cases |
|---|---|---|---|
| Delayed Lockdown | 44,353,721 | +57.3% | 1,977,059 |

**Key Findings:**

- **Earlier Vaccination**: Would have reduced total cases by **61.3%** (from 103.8M to 40.2M cases)

- **Delayed Lockdown**: Would have increased total cases by **57.3%** (from 103.8M to 144.2M cases)

- **Timing Impact**: The timing of interventions is critical - earlier implementation provides substantial benefits

**Visualization:** See `images/policy_scenario_simulations.png`

---

# Discussion

## Forecasting Performance

The comparison of forecasting models reveals that [best model] provides the most accurate predictions for COVID-19 weekly incidence. The state-space models (Poisson and Negative Binomial) offer advantages for count data but may require more computational resources. SARIMA models provide a good balance between accuracy and interpretability.

## Change Detection Insights

Both CUSUM and BOCPD successfully identified structural changes in the epidemic trajectory. These change points align with known intervention dates:

- March 2020: Initial lockdowns

- December 2020: Vaccination rollout

- [Other significant dates]

The ability to detect these changes in real-time would be valuable for adaptive policy responses.

### Counterfactual Analysis Implications

The counterfactual analysis demonstrates that without interventions, COVID-19 cases would have grown exponentially, quickly reaching unrealistic levels. By capping counterfactual forecasts at the US population, we obtain realistic estimates of avoided cases. This analysis provides quantitative evidence for the effectiveness of public health interventions.

### Policy Scenario Lessons

The policy scenario simulations highlight the critical importance of timing:

1. **Earlier vaccination** would have substantially reduced cases, demonstrating the value of rapid vaccine development and deployment

2. **Delayed lockdown** would have significantly increased cases, showing the cost of delayed action

These findings support the principle that early, decisive intervention is essential for epidemic control.

### Limitations

1. **Simplified Models**: The exponential growth model for counterfactuals is a simplification; real-world transmission dynamics are more complex

2. **Population Cap**: Capping at 335 million is necessary but may underestimate true counterfactual impact in early stages

3. **Intervention Effects**: Policy scenarios use simplified effect sizes; actual intervention effectiveness varies by context

4. **Data Quality**: Relies on reported cases, which may underestimate true incidence

---

# Conclusion

This project successfully combines forecasting, change detection, and counterfactual analysis to assess COVID-19 intervention effectiveness. Key conclusions:

1. **Forecasting**: Multiple models provide robust predictions, with SARIMA offering the best balance of accuracy and interpretability

2. **Change Detection**: CUSUM and BOCPD successfully identify intervention effects, providing tools for real-time monitoring

3. **Intervention Impact**: Counterfactual analysis demonstrates that interventions prevented millions of cases

4. **Policy Timing**: Scenario simulations show that earlier interventions provide substantial benefits, while delays have significant costs

The framework developed here can be applied to future epidemics and provides a quantitative basis for evidence-based public health decision-making. Future work could incorporate more sophisticated transmission models, account for variant emergence, and integrate additional data sources (mobility, genomic surveillance).

---

# Appendix

All code, data, and results are available in the project repository:

## Code Files

1. **Main Analysis Notebook**: `project.ipynb`

- Complete implementation of all four parts

- Forecasting models, change detection, counterfactual analysis, and policy scenarios

2. **Data Collection**: `data.py`

- Scripts for downloading and processing COVID-19 data from JHU

3. **Process Flow Diagram**: `create_process_flow.py`

- Generates visual workflow diagram

## Data Files

Located in `data/` folder:

- `usa_weekly_cases.csv`: USA weekly COVID-19 cases (main dataset)

- `maryland_weekly_cases.csv`: Maryland state-level data (available but not used in main analysis)

## Results Files

Located in `results/` folder:

- `forecast_comparison_metrics.csv`: Model comparison metrics
- `change_point_detection_results.csv`: CUSUM and BOCPD results
- `counterfactual_analysis.csv`: Full counterfactual data
- `counterfactual_summary.csv`: Summary statistics
- `policy_scenario_simulations.csv`: Scenario simulation data
- `scenario_comparison_table.csv`: Policy scenario comparison

### Visualizations

Located in `images/` folder:

- `forecast_model_comparison.png`: Part 1 - Forecasting comparison
- `cusum_change_detection.png`: Part 2 - CUSUM results
- `bocpd_change_detection.png`: Part 2 - BOCPD results
- `counterfactual_analysis.png`: Part 3 - Counterfactual visualization
- `policy_scenario_simulations.png`: Part 4 - Policy scenarios
- `process_flow_diagram.png`: Complete workflow diagram
- `project_b_complete_overview.png`: Comprehensive summary figure

## Repository Structure

project/

├── data/ # Input data files

├── images/ # All generated graphs

├── results/ # Analysis results (CSV files)

├── project.ipynb # Main analysis notebook

├── data.py # Data collection script

### How to Reproduce

1. Install dependencies: `pandas`, `numpy`, `matplotlib`, `seaborn`, `statsmodels`, `scipy`

2. Run `project.ipynb` cells sequentially

3. All outputs will be saved to organized folders (`images/`, `results/`)

---

# References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Inf Dis*. 2020;20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1

2. Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

3. Johns Hopkins University Center for Systems Science and Engineering. COVID-19 Data Repository. https://github.com/CSSEGISandData/COVID-19

4. Project B Specification: "Hybrid Forecasting and Sequential Intervention Assessment" - Epidemiological Modeling Course

├── README.md # Project documentation and reproduction instructions

## GitHub Repository

The complete source code, data, and analysis results for **Project B: Hybrid Forecasting and Sequential Intervention Assessment** are hosted publicly on GitHub.

**Repository Link:**
https://github.com/SawcyD/Project-B--Final-Project

This repository contains:

- The main analysis notebook (`project.ipynb`).
- Data processing scripts (`data.py`).
- All input data (`data/`), results (`results/`), and visualizations (`images/`).
- Instructions for reproduction (`README.md`).

---