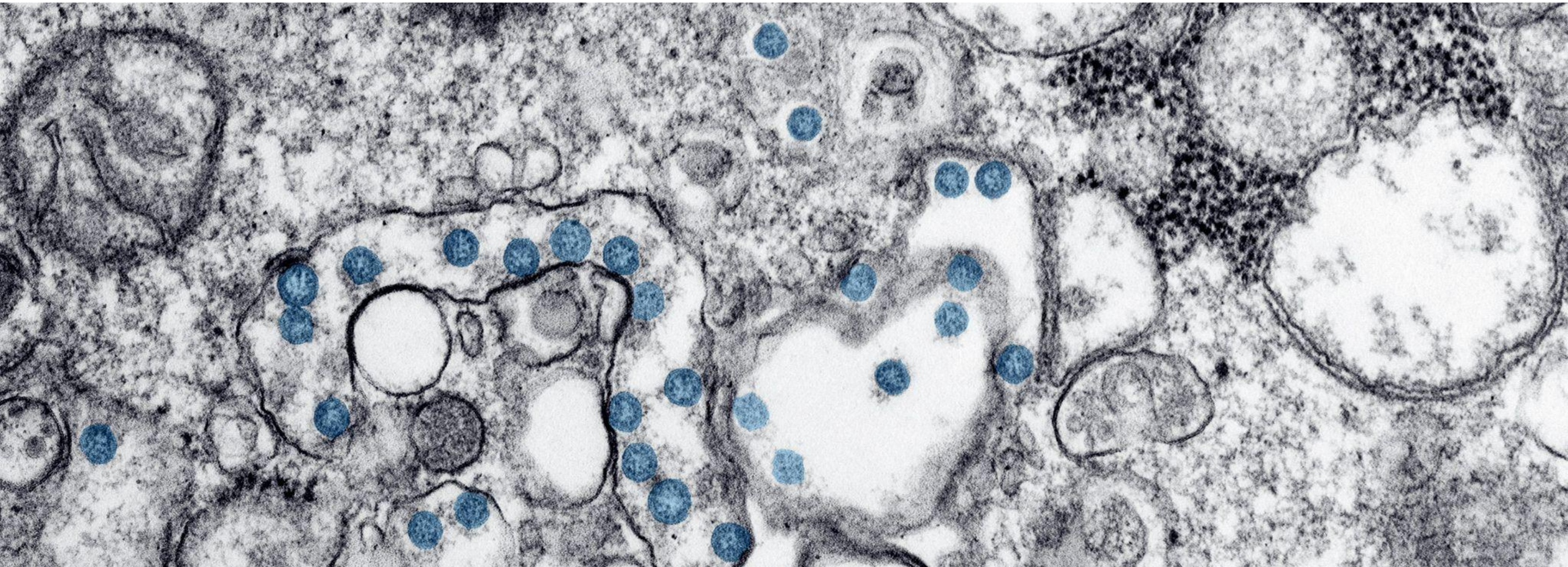
ENGS 27 Final Project: Presentation

Markov Mavericks: Raylene GUO, Yekun LI, Wesley TAN, Shin WU

Topic: Using Random Walk and SEIRD to predict Disease Spread



Overview

This project aims to model and predict COVID-19 cases by comparing the number of Infected (I) and Recovered (R) individuals across 4 SEA countries.

Using the **Random Walk** and **SEIRD model**, the study relies on the Johns Hopkins CSSE COVID-19 dataset for country-level case data.

Applications

By predicting infection and recovery trends, authorities can make decisions onto guide prevention and response efforts.

The model's adaptability to various regions and populations makes it useful for comparative analysis across different countries.

Mathematical Concepts

- **Random Walk Dynamics**: Movement of individuals in a grid-based environment, with movement probability (P_move) influenced by factors such as social distancing.

- **SEIRD Model**: Uses a system of ordinary differential equations (ODEs) to model disease progression across different states (Susceptible, Exposed, Infectious, Recovered, Deceased).

Challenges

- **Model Selection and Tuning**

Choosing the best predictive model for different regions, balancing between accuracy and interpretability.

- **Parameter Sensitivity**

Small changes in transmission or recovery rates significantly impact predictions.

- **Temporal and Spatial Variability**

Accounting for varying pandemic dynamics and intervention strategies across countries and time.

Problem and Application

- **Problems:**
  - **Impact of Infectious Diseases**
    Infectious diseases heavily impact public health and the economy. Modeling their spread is essential to guide prevention and response efforts.
  - **Data Gaps in Developing Regions**
    Most data and models focus on the U.S. and East Asia, while developing regions receive less attention. Therefore, our team has decided to concentrate on Thailand, Vietnam, Brunei, and Cambodia.

- **Application:**
  - Susceptible - Infected - Recover (SIR)
  - Susceptible - Exposed - Infected - Recovered - Dead (SEIRD)
  - Random Walk

Stakeholders

- **Public Health Authorities**
  - Better understand risk, e.g., social distancing, vaccination campaigns
- **Hospitals and Healthcare providers**
  - Manage load, e.g., predict patient flows and anticipate the demand for ICU beds, ventilators, and other critical resources
- **Policy Makers and Government Officials**
  - Implement policy, e.g. lockdowns or mask mandates
- **Economists and Financial Analysts**
  - Estimation of economic impacts from infection disease based on health outcomes and healthcare costs

Methodology (1): Random Walk

Initialization

- Define a grid of size $N \times N$ representing a population.
- Initialize each cell with a state: Susceptible (S), Infected (I), or Recovered (R).
- Set the initial conditions such as the number of infected individuals and mobility parameters.

Movement Dynamics

- Each individual in the grid moves to an adjacent cell based on a movement probability P_{move}.
- P_{move} is influenced by:

$$P_{\text{move}} = \text{mobility}_{\text{state}} \times \text{social\_distance\_factor}$$

Transmission Probability

$$P_T = 1 - \prod_{(n_i,n_j)\in N_{i,j}} [(1-\beta_N)\cdot I_N \cdot (1-\beta_Q)\cdot I_Q]$$

where:

- $N_{i,j}$: Set of neighboring cells to position (i,j)
- β_N: Non-quarantined transmission rate
- β_Q: Quarantined transmission rate
- I_N: Indicator function for non-quarantined infected neighbor
- I_Q: Indicator function for quarantined infected neighbor

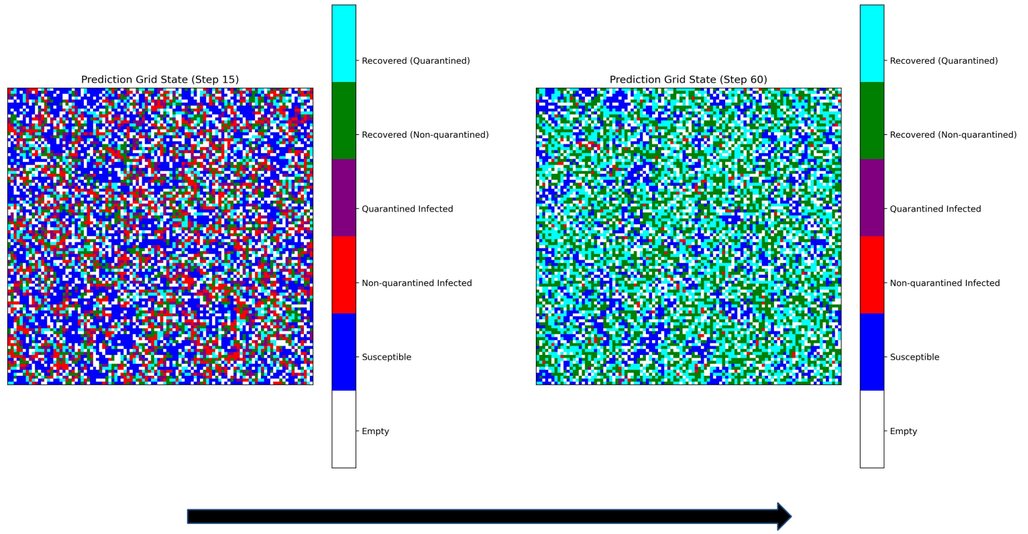
Recovery Process

- Infected individuals recover at a rate γ which varies based on quarantine status, where `quarantine_status` is a factor based on training data:

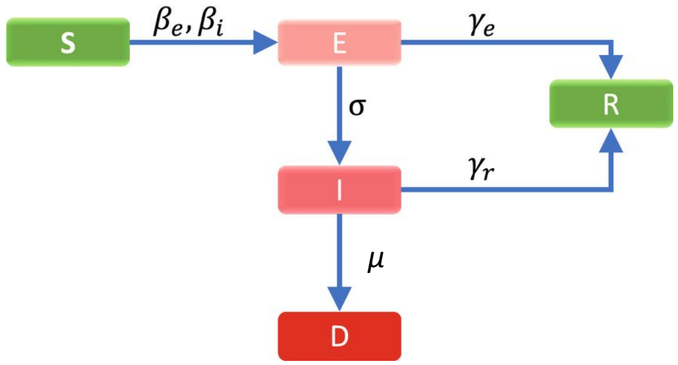
$$\gamma_q = \texttt{quarantine\_status} \cdot \gamma_n$$

Algorithm 1 Enhanced Random Walk Process

```
Create a copy of the current grid state G'
Generate a random permutation π of all grid positions
for each position (i, j) in π do
    if (i, j) has not been moved then
        s ← state at (i, j)
        p_move ← μ_s · f_social(t)
        if random(0, 1) < p_move/100 then
            N ← empty neighboring cells of (i, j)
            if N is not empty then
                Select a random cell (n_i, n_j) from N
                Swap states between (i, j) and (n_i, n_j) in G'
                Mark (n_i, n_j) as moved
            end if
        end if
    end if
end for
Update the grid with G'
```



Methodology (2): SEIRD Model



Jha, Prashant & Cao, Lianghao & Oden, J.. (2020). Bayesian-based predictions of COVID-19 evolution in Texas using multispecies mixture-theoretic continuum models. Computational Mechanics. 66. 10.1007/s00466-020-01889-z.

Explanation of Parameters

- β: Transmission rate — the rate at which susceptible individuals become exposed upon contact with infectious individuals.
- σ: Progression rate — the rate at which exposed individuals move to the infectious state. Typically, $\sigma = \frac{1}{\text{incubation period}}$.
- γ: Recovery rate — the rate at which infectious individuals recover.
- μ: Mortality rate — the rate at which infectious individuals die.

Initial Conditions

To solve these equations, we need initial conditions $S(0)$, $E(0)$, $I(0)$, $R(0)$, and $D(0)$ such that:

$$S(0) + E(0) + I(0) + R(0) + D(0) = 1 \quad \text{(normalized population)}$$

Transition Matrix Representation

The transition matrix for a discrete step can be represented as:

$$\mathbf{T} = \begin{bmatrix} 1-\beta I & \beta I & 0 & 0 & 0 \\ 0 & 1-\sigma & \sigma & 0 & 0 \\ 0 & 0 & 1-(\gamma+\mu) & \gamma & \mu \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Monte Carlo Sampling for Uncertainty Estimation

- Generate samples β_i, γ_i, μ_i from normal distributions centered at the optimized parameters with a standard deviation σ_{param} (e.g., $0.1 \times$parameter).
- Run simulations for each sample and record the results.
- Calculate standard deviations and 95% confidence intervals for each parameter.

Algorithm 2 Improved SEIRD COVID-19 Simulator

Data Processing:
Process and smooth COVID-19 data with 7-day rolling average
Split data at $t = 60$ days for training/testing
total_population $\leftarrow$ max(Confirmed) $\times$ underreporting_factor
Model Initialization:
Optimize $\{\beta, \sigma, \gamma, \mu\}$ using training data
Initialize states: $\{S_0, E_0, I_0, R_0, D_0\}$ from training end
SEIRD Simulation:
for $t \leftarrow 1$ to prediction_days **do**
    $\beta_t \leftarrow \beta \times$ seasonal_modifier(t)
    Update states via SEIRD equations:

$$\begin{cases} \frac{dS}{dt} = -\beta_t SI \\ \frac{dE}{dt} = \beta_t SI - \sigma E \\ \frac{dI}{dt} = \sigma E - (\gamma + \mu) I \\ \frac{dR}{dt} = \gamma I \\ \frac{dD}{dt} = \mu I \end{cases}$$

    Normalize: $S + E + I + R + D = 1$
end for
return predictions, [MSE, RMSE, MAE, R^2], plots

Comparison Analysis

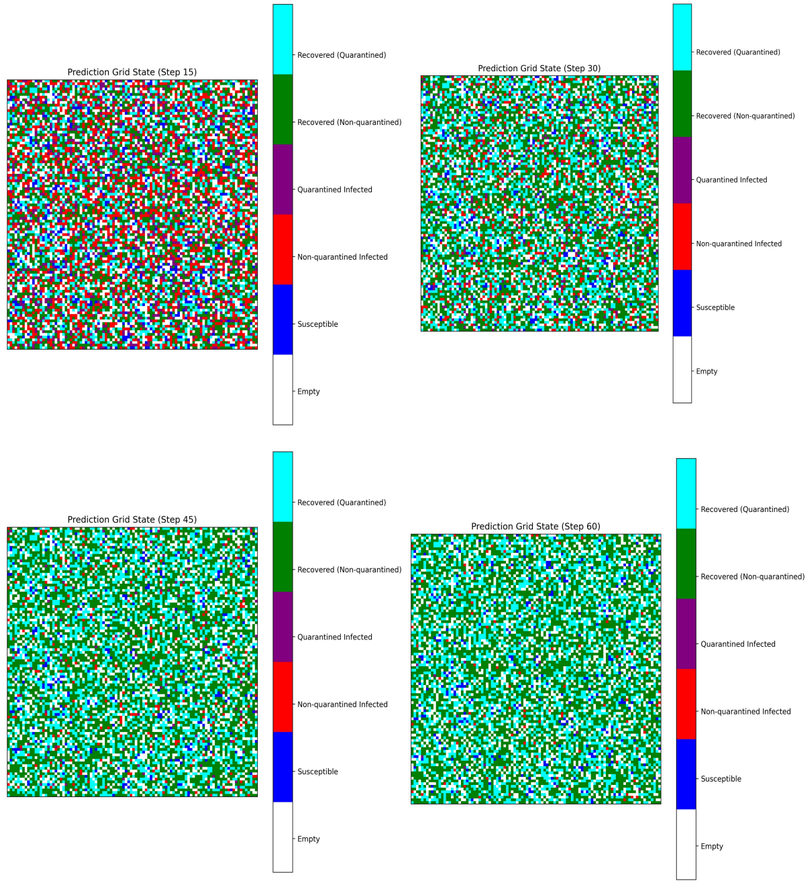
Dataset: Johns Hopkins CSSE COVID-19 Dataset (GitHub Repository: https://github.com/CSSEGISandData/COVID-19).

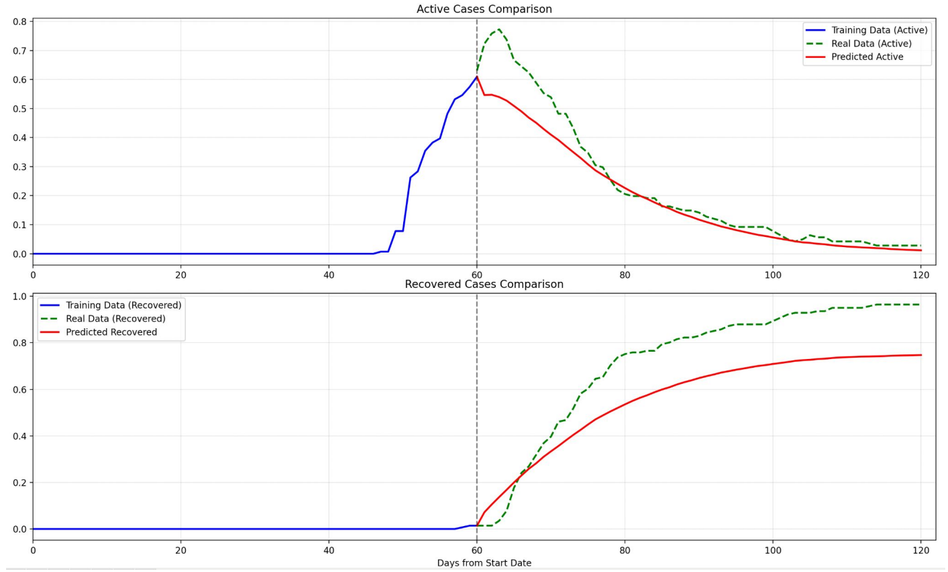
- Split data into training and validation sets. Use Day 0-60 for training and then we compare predicted and real data in Day 61-120. (More information in **Appendix A**)

kaggle

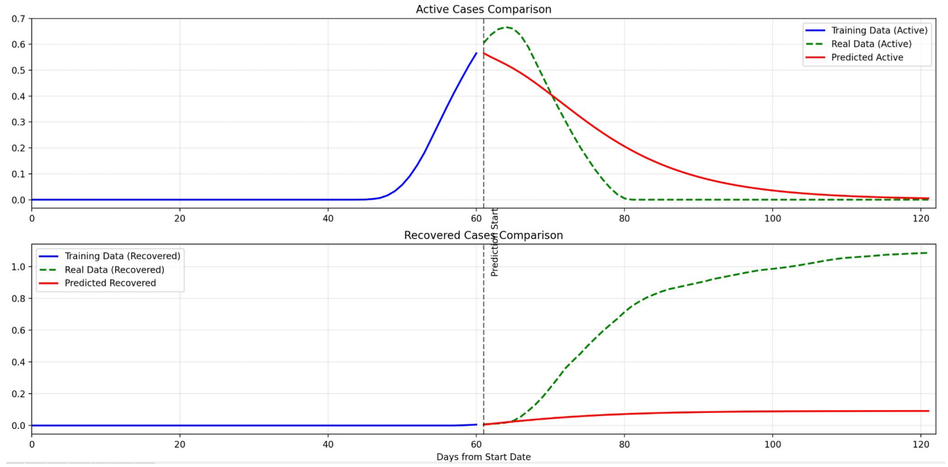
Feel free to play around with our model: https://tinyurl.com/engs27randomwalk

Results (1): E.g. Brunei





Random Walk Model



SEIRD Model

Country	Model	Metric Type	MSE	RMSE	MAE	R^2
Brunei	Random Walk	Active Cases	0.0055	0.0742	0.0457	0.8957
		Recovered Cases	0.0304	0.1744	0.1634	0.6320
	SEIRD	Active Cases	0.0182	0.1348	0.0829	0.6292
		Recovered Cases	0.1218	0.3490	0.3124	0.0372

Results (2): Comparison, SEA Countries

Country	Model	Metric Type	MSE	RMSE	MAE	R^2
Thailand	Random Walk	Active Cases	0.0118	0.1088	0.0762	0.5844
		Recovered Cases	0.0658	0.2565	0.2106	0.4376
	SEIRD	Active Cases	0.0574	0.2395	0.1705	0.0162
		Recovered Cases	0.7332	0.8563	0.7081	-0.6888
Cambodia	Random Walk	Active Cases	0.0139	0.1177	0.0823	0.8076
		Recovered Cases	0.0393	0.1983	0.1748	0.6410
	SEIRD	Active Cases	0.0258	0.1606	0.1019	0.5350
		Recovered Cases	0.1020	0.3194	0.2732	0.2484
Vietnam	Random Walk	Active Cases	0.0138	0.1175	0.1092	0.1853
		Recovered Cases	0.0094	0.0967	0.0833	0.8511
	SEIRD	Active Cases	0.0513	0.2264	0.1527	0.1474
		Recovered Cases	0.5232	0.7233	0.6232	-0.8719
Brunei	Random Walk	Active Cases	0.0055	0.0742	0.0457	0.8957
		Recovered Cases	0.0304	0.1744	0.1634	0.6320
	SEIRD	Active Cases	0.0182	0.1348	0.0829	0.6292
		Recovered Cases	0.1218	0.3490	0.3124	0.0372

May be hard to ascertain how long one takes to recover

- Brunei: Random Walk performed best (Active Cases R^2 = 0.8957)
- Cambodia: Both models showed good performance (Random Walk R^2 = 0.8076, SEIRD R^2 = 0.5350)
- Thailand: Random Walk outperformed SEIRD significantly
- Vietnam: Both models struggled with active cases but Random Walk excelled in recovered cases

(More information in Appendix B)

Results (3): Analysis

Model Performance Overview:

- Across all 4 countries, the Random Walk model consistently exhibits lower Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) for active cases compared to the SEIRD model.
- The R^2 values, which indicate the proportion of variance explained by the model, are generally higher for the Random Walk model in both active and recovered cases, suggesting better fit and predictive performance.

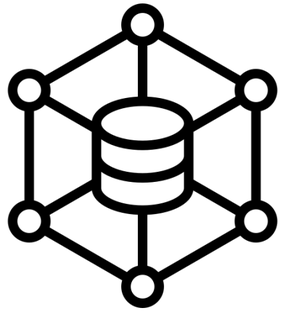
Model Suitability:

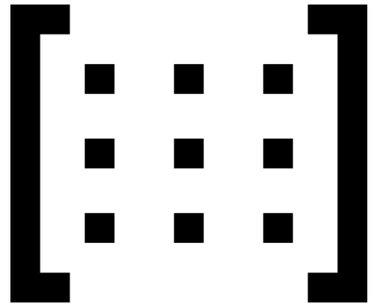
- The **Random Walk model** appears more robust for practical short-term predictions of active and recovered cases due to its simplicity and consistent results.
- The **SEIRD model**, while valuable for capturing disease dynamics, may require better parameter tuning or more refined data inputs to achieve predictive accuracy comparable to the Random Walk model.

So what?

- Prioritize the **Random Walk model** for immediate application and analysis due to its consistent predictive performance.
- Investigate parameter optimization and additional data refinement for the **SEIRD model** to potentially enhance its reliability and usability in future predictive efforts.

Challenges & Limitations (1): Challenges





1. Data Challenges

Temporal Constraints

- Limited training period (60 days)
- Virus variants not included

Data Quality Issues

- Inconsistent reports across countries
- Missing/delayed data in some regions.
  -

2. Model Challenges

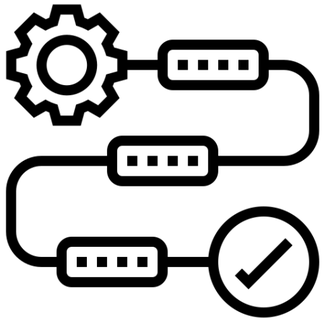
a. Random Walk Model

1. Structural Limitations:
   - Fixed grid size limits scalability
   - Simplified movement patterns
2. Parameter Limitations:
   - Constant quarantine ratio
   - Simplified seasonal effects
   - Fixed mobility rates

b. SEIRD Model

1. Structural Limitations:
   - No age structure
   - No localized outbreaks
   - No modeling of interventions
2. Parameter Limitations:
   - Constant recovery and death rates
   - Not capture policy changes
   - Simplified transmission dynamics

Challenges & Limitations (2): Limitations



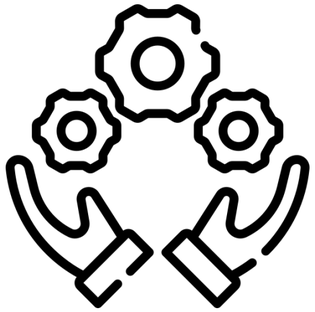
3. Methodological Limitations

Validation Issues

- Limited ground truth for model validation
- Difficulty in parameter estimation due to data noise
- No cross-validation across time periods

Performance Metrics

- High variability of R^2 (-0.8719 to 0.8957)



4. Practical Limitations

Implementation Constraints

- Large memory occupation
- Inefficient run time

Policy Relevance

- Models don't explicitly account for:
  - Vaccination programs
  - Healthcare system capacity
  - Social distancing policies
  - Economic factors
  - Behavioral changes

5. Recommendations for Future Work

1. **Data Improvements:**
   - Include longer time periods
   - Include vaccination data
2. **Model Enhancements:**
   - Hybrid approaches combining both models
   - Include vaccination effects
   - Add age stratification
   - Consider spatial heterogeneity
   - Include healthcare capacity constraints
3. **Add Validation Methods:**
   - Cross-validation across time periods
   - Sensitivity analysis of parameters
   - Uncertainty quantification
4. **Additional Features:**
   - Policy intervention modeling
   - Behavioral change effects
   - Healthcare system capacity
   - Economic factors
   - Contact tracing effects

Thank you!

References

1. Triambak, S., and D. P. Mahapatra. "A Random Walk Monte Carlo Simulation Study of COVID-19-Like Infection Spread." *Physica A: Statistical Mechanics and Its Applications* 581 (2021): 126014. https://doi.org/10.1016/j.physa.2021.126014.
2. Auranen, Kari, Mikhail Shubin, Elina Erra, Sanna Isosomppi, Jukka Kontto, Tuija Leino, and Timo Lukkarinen. "Efficacy and Effectiveness of Case Isolation and Quarantine during a Growing Phase of the COVID-19 Epidemic in Finland." *Epidemiology and Infection* 149 (2021): e204.
3. Nakagiri, Nariyuki, Kazunori Sato, Yukio Sakisaka, and Kei-ichi Tainaka. "Serious Role of Non-Quarantined COVID-19 Patients for Random Walk Simulations."
4. Bjørnstad, Ottar N., Katriona Shea, Martin Krzywinski, and Naomi Altman. "The SEIR Model for Infectious Disease Dynamics: Realistic Models of Epidemics Account for Latency, Loss of Immunity, Births, and Deaths." *Nature Methods* 17, no. 6 (2020): 557-558.
5. Aronna, M. S., R. Guglielmi, and L. M. Moschen. "A Model for COVID-19 with Isolation, Quarantine, and Testing as Control Measures."
6. Wang, Chengliang, and Sohaib Mustafa. "A Data-Driven Markov Process for Infectious Disease Transmission." *Scientific Reports* 11, no. 1 (2021): 1-12.

Appendix A: Comparison

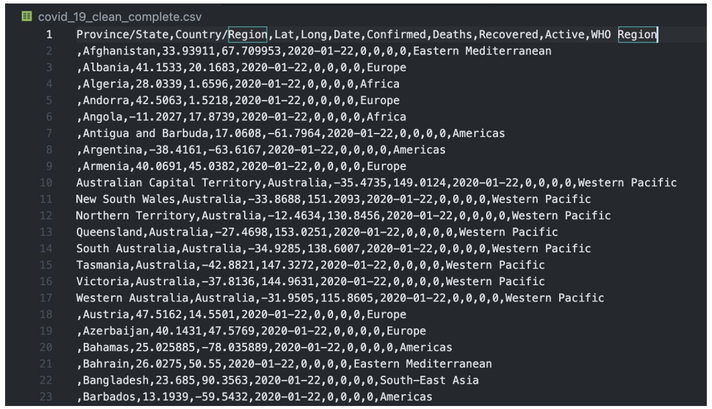
Dataset: Johns Hopkins CSSE COVID-19 Dataset (GitHub Repository: https://github.com/CSSEGISandData/COVID-19).

- Use time series models (Random Walk) and epidemiological models (SEIRD) to predict I (infected/'active') and R (recovered). We then compare the two approaches.

Training and Validation:

- Split data into training and validation sets. Use Day 0-60 for training and then we compare predicted and real data in Day 61-120

Error Metrics: Comparing Predicted to Real


covid_19_clean_complete.csv

```
Province/State,Country/Region,Lat,Long,Date,Confirmed,Deaths,Recovered,Active,WHO Region
,Afghanistan,33.93911,67.709953,2020-01-22,0,0,0,0,Eastern Mediterranean
,Albania,41.1533,20.1683,2020-01-22,0,0,0,0,Europe
,Algeria,28.0339,1.6596,2020-01-22,0,0,0,0,Africa
,Andorra,42.5063,1.5218,2020-01-22,0,0,0,0,Europe
,Angola,-11.2027,17.8739,2020-01-22,0,0,0,0,Africa
,Antigua and Barbuda,17.0608,-61.7964,2020-01-22,0,0,0,0,Americas
,Argentina,-38.4161,-63.6167,2020-01-22,0,0,0,0,Americas
,Armenia,40.0691,45.0382,2020-01-22,0,0,0,0,Europe
Australian Capital Territory,Australia,-35.4735,149.0124,2020-01-22,0,0,0,0,Western Pacific
New South Wales,Australia,-33.8688,151.2093,2020-01-22,0,0,0,0,Western Pacific
Northern Territory,Australia,-12.4634,130.8456,2020-01-22,0,0,0,0,Western Pacific
Queensland,Australia,-27.4698,153.0251,2020-01-22,0,0,0,0,Western Pacific
South Australia,Australia,-34.9285,138.6007,2020-01-22,0,0,0,0,Western Pacific
Tasmania,Australia,-42.8821,147.3272,2020-01-22,0,0,0,0,Western Pacific
Victoria,Australia,-37.8136,144.9631,2020-01-22,0,0,0,0,Western Pacific
Western Australia,Australia,-31.9505,115.8605,2020-01-22,0,0,0,0,Western Pacific
,Austria,47.5162,14.5501,2020-01-22,0,0,0,0,Europe
,Azerbaijan,40.1431,47.5769,2020-01-22,0,0,0,0,Europe
,Bahamas,25.025885,-78.035889,2020-01-22,0,0,0,0,Americas
,Bahrain,26.0275,50.55,2020-01-22,0,0,0,0,Eastern Mediterranean
,Bangladesh,23.685,90.3563,2020-01-22,0,0,0,0,South-East Asia
,Barbados,13.1939,-59.5432,2020-01-22,0,0,0,0,Americas
```

Mean Squared Error (MSE)

Definition: MSE is the average of the squared differences between the predicted values and the actual values. Mathematically, it is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Interpretation: A lower MSE indicates better model performance. MSE penalizes larger errors more due to squaring.

Root Mean Squared Error (RMSE)

Definition: RMSE is the square root of MSE, and can be expressed as:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Interpretation: RMSE retains the units of the target variable, making it easier to interpret. A lower RMSE indicates better model performance.

Mean Absolute Error (MAE)

Definition: MAE is the average of the absolute differences between the predicted values and the actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Interpretation: MAE gives the average error magnitude without considering direction, and is less sensitive to outliers compared to MSE and RMSE. A lower MAE indicates better accuracy.

R-squared (R^2)

Definition: R^2 (coefficient of determination) measures the proportion of variance in the actual values explained by the model's predictions. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values.

Interpretation: An R^2 close to 1 indicates a good fit, while a lower or negative R^2 suggests poor predictive accuracy.

kaggle

Feel free to play around with our model: https://tinyurl.com/engs27randomwalk

Appendix B: Country Summaries

Brunei (Best Performing Case)

- **Random Walk Model Performance:**
  - Active Cases: R^2=0.8957, MSE = 0.0055
  - *Statistical Interpretation:* Model explains 89.57% of variance with minimal error.
- **Actionable Insights:**
  - Reliable for 14-day forecasting in small countries.
  - Patient load predictions accurate within ±5%.
  - Optimal for controlled outbreak scenarios.

Thailand (Moderate Performance)

- **Random Walk Model Performance:**
  - Active Cases: R^2=0.5844, MSE = 0.0118
  - *Statistical Interpretation:* Model explains 58.44% of variance.
- **Actionable Insights:**
  - Suitable for 7-day forecasting windows.
  - Requires weekly model recalibration.
  - Resource planning requires 15-20% buffer.

Vietnam (Mixed Performance)

- **Random Walk Model Performance:**
  - Active Cases: R^2=0.1853,
  - Recovered Cases: R^2=0.8511, MSE = 0.0094
  - *Statistical Interpretation:* Strong recovery prediction but weak active case tracking.
- **Actionable Insights:**
  - Optimal for recovery resource planning.
  - Requires 30-40% buffer for active case predictions.

Cambodia (Strong Overall Performance)

- **Random Walk Model Performance:**
  - Active Cases: R^2=0.8076, MSE = 0.0139
  - Recovered Cases: R^2=0.6410, MSE = 0.0393
  - *Statistical Interpretation:* Strong predictive power across both metrics.
- **Actionable Insights:**
  - Reliable for 10-day forecasting.
  - 20% buffer sufficient for resource planning.