



UNDERSTANDING SWEDEN'S FHM COVID-19 DATA UPDATES.

Presented By : Olayemi Morrison

March 13, 2025

OVERVIEW

- Introduction
- Objectives
- Why Nowcasting?
- Data Description
- Methodology
- Implementation
- Results
- Conclusion

INTRODUCTION

The COVID-19 pandemic highlighted the need for timely and accurate reporting of case numbers to inform public health decisions. However, delays in case reporting can obscure the true state of an outbreak, necessitating statistical methods to estimate the actual number of cases in real time. This process, known as nowcasting, corrects for reporting delays by leveraging past patterns of data revisions.

In this seminar, I will present my ongoing thesis work on nowcasting COVID-19 cases in Sweden, using Bayesian hierarchical models to dynamically estimate the true number of cases. My approach extends existing nowcasting frameworks by incorporating demographic (age, gender) and regional breakdowns, allowing for more granular insights. The models leverage past reporting patterns and lagged data streams to predict missing case counts, improving real-time situational awareness.

This presentation will cover the methodology used, including data preprocessing, model formulation, and evaluation techniques. I will also discuss challenges encountered—such as handling large datasets and structuring them within an SQLite database—as well as next steps in refining the model to improve accuracy and computational efficiency.

OBJECTIVES

The primary objective of this study is to assess the quality of Sweden's COVID-19 case reporting over time and develop a Bayesian nowcasting framework to correct for reporting delays. Specifically, the study aims to:

First:

Evaluate reporting accuracy trends throughout the pandemic by analyzing how data quality has changed over time. This includes assessing whether early pandemic data (e.g., 2020) were less reliable than later periods (e.g., 2022) and examining the impact of reporting policy changes (e.g., transition from daily to weekly reports).

OBJECTIVES

The primary objective of this study is to assess the quality of Sweden's COVID-19 case reporting over time and develop a Bayesian nowcasting framework to correct for reporting delays. Specifically, the study aims to:

Second:

Determine the reliability of real-time data for nowcasting by identifying periods where reporting delays were minimal and assessing when data quality deteriorated, making nowcasting less reliable. This includes detecting systematic patterns in underreporting, such as delays associated with weekends, holidays, or specific timeframes.

OBJECTIVES

The primary objective of this study is to assess the quality of Sweden's COVID-19 case reporting over time and develop a Bayesian nowcasting framework to correct for reporting delays. Specifically, the study aims to:

Third:

Analyze regional and demographic variations in reporting quality by identifying which regions demonstrated consistently accurate or poor reporting practices and evaluating disparities in data completeness across demographic groups (age, gender, etc.).

WHY NOWCASTING?

Importance of Adjusting for Delays in Case Reporting

COVID-19 case counts are often underreported in real-time due to delays in test results, administrative processing, and data collection. These delays lead to a distorted view of the actual infection trends, which can mislead policymakers and public health officials. Nowcasting is essential because:

1

Delays can vary based on the day of the week, holidays, or changes in reporting policies (e.g., shifting from daily to weekly reports).

2

Without adjustments, real-time data underestimates true infections, potentially leading to late or inadequate public health interventions.

3

Nowcasting corrects these distortions by using statistical models to estimate the actual number of cases that occurred on a given date, even before all reports are finalized.

EXAMPLE

- **Scenario:**

If 500 cases are reported today, but historical data suggests that reporting is only 70% complete on the first day, nowcasting helps estimate the true number of cases, which could be closer to 700 cases rather than 500.

EXAMPLE

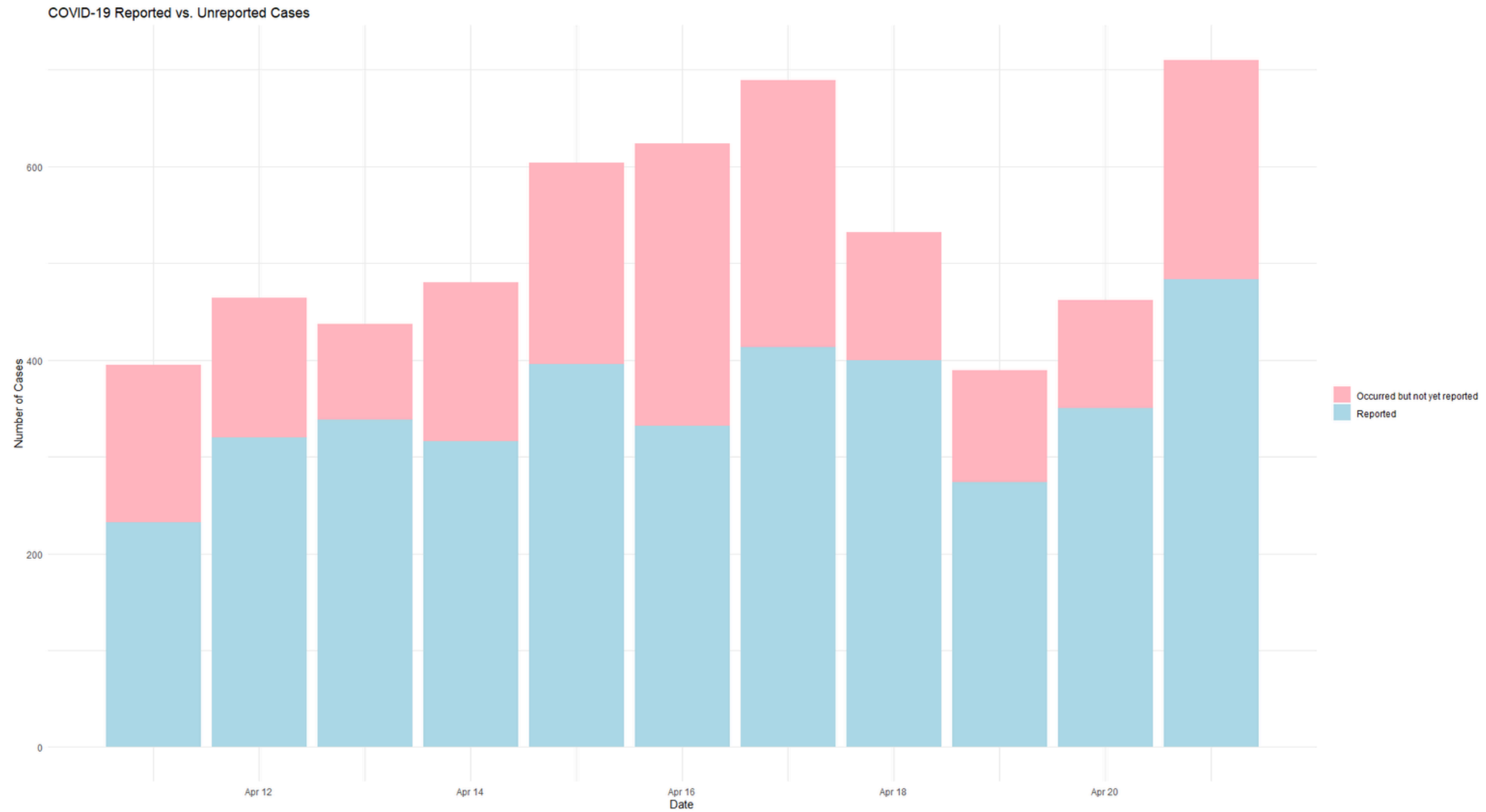
This visualization highlights the reporting delay in COVID-19 cases in Stockholm from April 11, 2020, to April 21, 2020, showing reported vs. unreported cases. The blue bars represent cases reported by the Swedish Public Health Agency as of each reporting date, while the pink bars indicate additional cases that had occurred but were not yet reported at the time.

| event_date | rep_date | reported_so_far | cases_occurred | cases_unreported |
|------------|-----------|-----------------|----------------|------------------|
| 4/11/2020 | 4/12/2020 | 232 | 395 | 163 |
| 4/12/2020 | 4/13/2020 | 320 | 464 | 144 |
| 4/13/2020 | 4/14/2020 | 338 | 437 | 99 |
| 4/14/2020 | 4/15/2020 | 316 | 480 | 164 |
| 4/15/2020 | 4/16/2020 | 396 | 604 | 208 |
| 4/16/2020 | 4/17/2020 | 332 | 624 | 292 |
| 4/17/2020 | 4/18/2020 | 413 | 689 | 276 |
| 4/18/2020 | 4/19/2020 | 400 | 532 | 132 |
| 4/19/2020 | 4/20/2020 | 274 | 389 | 115 |
| 4/20/2020 | 4/21/2020 | 350 | 462 | 112 |
| 4/21/2020 | 4/22/2020 | 483 | 710 | 227 |

EXAMPLE

10

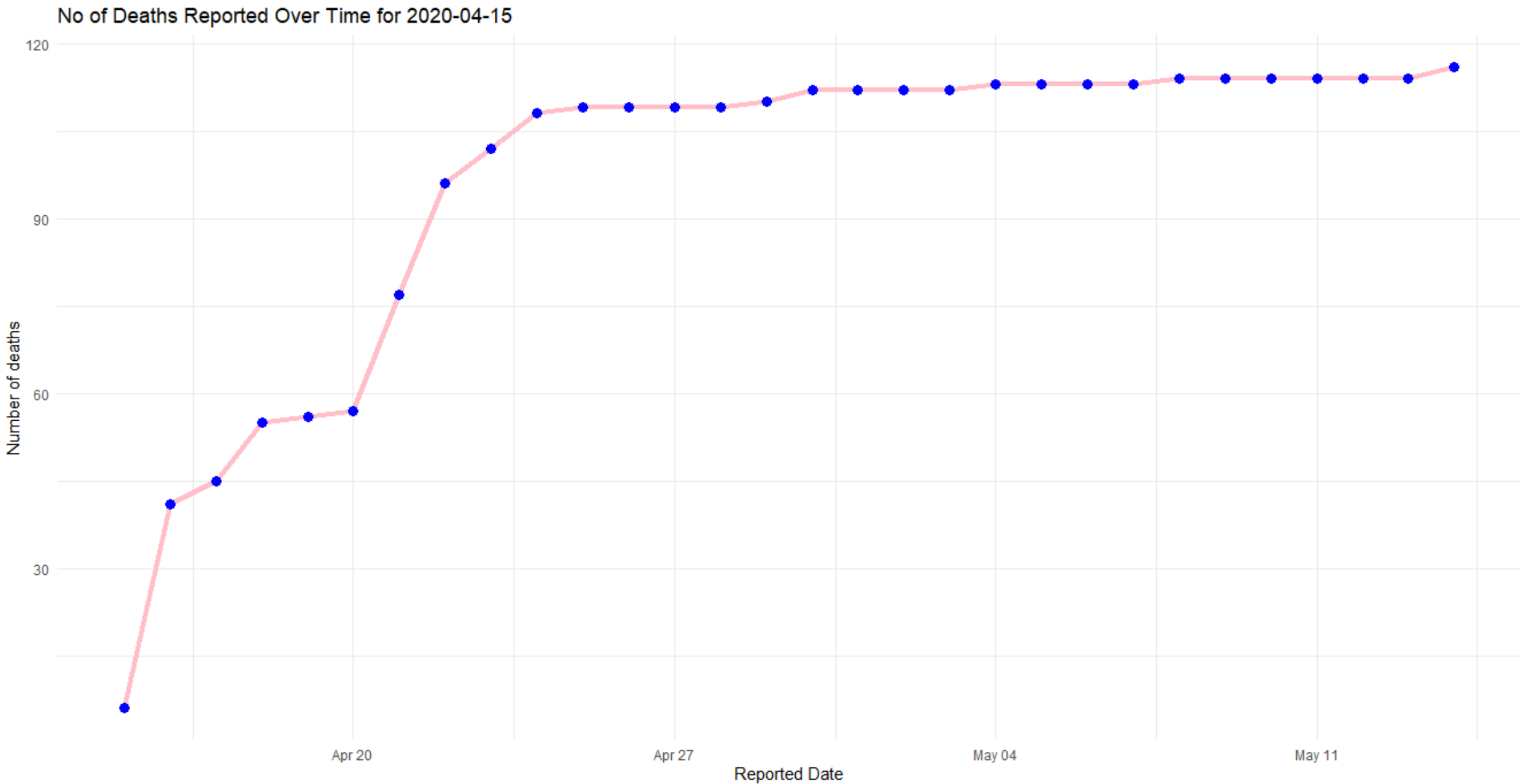
This visualization highlights the reporting delay in COVID-19 cases in Stockholm from April 11, 2020, to April 21, 2020, showing reported vs. unreported cases. The blue bars represent cases reported by the Swedish Public Health Agency as of each reporting date, while the pink bars indicate additional cases that had occurred but were not yet reported at the time.



EXAMPLE

This plot visualizes how the number of reported deaths evolved over time for cases corresponding to 2020-04-15, highlighting fluctuations in delayed reporting.

| <u>death_date</u> | <u>N</u> | <u>rep_date</u> |
|-------------------|----------|-----------------|
| 4/15/2020 | 6 | 4/15/2020 |
| 4/15/2020 | 41 | 4/16/2020 |
| 4/15/2020 | 45 | 4/17/2020 |
| 4/15/2020 | 55 | 4/18/2020 |
| 4/15/2020 | 56 | 4/19/2020 |
| 4/15/2020 | 57 | 4/20/2020 |
| 4/15/2020 | 77 | 4/21/2020 |
| 4/15/2020 | 96 | 4/22/2020 |
| 4/15/2020 | 110 | 4/29/2020 |
| 4/15/2020 | 113 | 5/7/2020 |
| 4/15/2020 | 114 | 5/8/2020 |
| 4/15/2020 | 114 | 5/9/2020 |
| 4/15/2020 | 114 | 5/10/2020 |
| 4/15/2020 | 114 | 5/11/2020 |
| 4/15/2020 | 114 | 5/12/2020 |
| 4/15/2020 | 114 | 5/13/2020 |
| 4/15/2020 | 116 | 5/14/2020 |



DATA DESCRIPTION

11

● The dataset used in this study originates from the Public Health Agency of Sweden (Folkhälsomyndigheten, FHM), which is responsible for monitoring and reporting COVID-19 statistics across Sweden. The dataset provides a comprehensive view of the pandemic's progression, covering the period from January 2020 to December 2024. It includes records of new hospital admissions due to COVID-19, differentiating between general hospitalizations and ICU treatment, offering insights into case severity.

Mortality data is also featured, with daily and weekly death counts at national and regional levels, alongside cumulative totals for long-term trends. Epidemiological metrics include weekly incidence rates per 100,000 inhabitants and 14-day incidence rates, which reflect broader infection trends. Additionally, cumulative case counts for infections, ICU admissions, and deaths are provided.

● Raw Data:

| Variable | Description |
|---|---|
| år (Year) | The calendar year of the reported data. |
| veckonummer (Week Number) | The week of the year (1–52). |
| Antal_fall_vecka (Weekly Cases) | The number of new COVID-19 cases reported in a given week. |
| Antal_fall_100000inv_vecka (Weekly Cases per 100,000 Inhabitants) | Weekly cases normalized by population size. |
| Antal_fall_100000inv_14dagar (14-Day Incidence Rate per 100,000) | The cumulative 14-day incidence rate, a common metric for epidemiological trends. |
| Kum_antal_fall (Cumulative Cases) | The cumulative total of reported cases up to and including the given week. |
| Antal_nyintensivvårdade_vecka (Weekly Intensive Care Admissions) | The number of new admissions to intensive care units. |
| Kum_antal_intensivvårdade (Cumulative ICU Admissions) | The cumulative total of ICU admissions. |
| Antal_avlidna_vecka (Weekly Deaths) | The number of COVID-19-related deaths reported during the week. |
| Kum_antal_avlidna (Cumulative Deaths) | The cumulative total of COVID-19-related deaths. |

DATA DESCRIPTION

● Key Datasets and Variables:

All files have been converted to .csv format for uniformity. Each dataset captures a different aspect of case reporting dynamics, contributing to a comprehensive understanding of how delays and underreporting affect real-time surveillance. The processed folder contains the following:

| File Name | Description | Key Variables |
|-------------------|---|--|
| acov19DAG.csv | Daily COVID-19 cases. | Region, Day, Cases per day |
| bcov19Kom.csv | Weekly data for each municipality. | Municipality, Indicator, Year and week, Cases by municipality and week |
| ccov19Reg.csv | Weekly confirmed cases per region. | Region, Indicator, Year and week, Confirmed cases |
| ccov19kon.csv | Weekly cases by gender and region. | Region, Indicator, Gender, Year and week, Cases by gender and region |
| dcov19ald.csv | Weekly cases by age group. | Indicator, Age group, Year and week, Cases by age group |
| xcov19ivavDAG.csv | Daily deaths and ICU admissions. | Indicator, Day, Intensive care and deceased per day |
| ycov19ivavald.csv | ICU and deaths by age group (weekly). | Indicator, Age group, Year and week, ICU cases and deaths |
| ycov19ivavkon.csv | ICU and deaths by gender (weekly). | Indicator, Gender, Year and week, ICU cases and deaths |
| ecov19sabo.csv | Cases among individuals 65+ receiving social services (weekly). | Region, Category, Year and week, Cases among 65+ |

DATA DESCRIPTION

● Handling Large Data:

Given the high volume of case reports, efficient data storage and processing are crucial for effective nowcasting. The following steps were taken to manage the dataset:

- Storage in SQLite:
 - Enables fast querying of large datasets.
 - Facilitates structured data retrieval without excessive memory overhead in R.

```
# Example usage
# Set your folder path here
folder_path <- "C:/Users/HP/Documents/FHM_project/data/20230406/"

csv_file_names <- function(folder_path) {
  # List all CSV files in the folder
  csv_files <- list.files(folder_path, pattern = "\\*.csv$", full.names = TRUE)

  # Remove files with "PCR" in the filename
  csv_files <- csv_files[!grepl("PCR", csv_files)]

  # Remove files with "test" in the filename
  csv_files <- csv_files[!grepl("test", csv_files)]

  # Check if any CSV files exist
  if (length(csv_files) == 0) {
    cat("No CSV files found in the specified folder.\n")
    return(character(0))
  }

  # Convert to vector and remove ".csv" extension
  csv_files <- tools::file_path_sans_ext(basename(csv_files))

  # Print list of all CSV files
  cat("List of CSV files:\n")
  print(csv_files)
}
```

```
sqlcon <- dbConnect(RSQLite::SQLite(), "fhmdata.sqlite")

create_tables_from_df <- function(db_connection, column_dataframe) {
  for (table_name in colnames(column_dataframe)) {
    # Extract column names (remove NAs)
    col_names <- na.omit(column_dataframe[[table_name]])

    # Define column definitions (all columns as TEXT, adjust as needed)
    col_definitions <- paste0(col_names, " TEXT", collapse = ", ")

    # Construct the SQL statement
    sql_query <- paste0("CREATE TABLE IF NOT EXISTS ", table_name, " (", col_definitions, ")")

    # Print SQL query
    # Execute the SQL command
    dbExecute(db_connection, sql_query)

    cat("Created table:", table_name, "\n")
  }
}
```

METHODOLOGY

Flexible Bayesian Nowcasting

This method aims to infer the total number of events on a given day t based on the N_t information available at a later day T . Since case reporting is delayed, the reported $T > t$ count on a given day underestimates the true number of infections. The observed and unobserved case counts can be formulated as:

$$N_t = \sum_{d=0}^{T-t} n_{t,d} + \sum_{d=T-t+1}^D n_{t,d}$$

where:

- $n_{t,d}$ is the number of events occurring on day t but reported with a delay of d days.
- The first sum represents observed cases, while the second sum accounts for unreported cases.

METHODOLOGY

Bayesian Model Components

Epidemic Curve (Latent Process Model)

The epidemic curve (epi curve) is a graph showing the number of new COVID-19 cases over time. It helps us see the rise, peak, and decline of an outbreak. It helps us predict missing or delayed data by looking at trends.

We model the expected number of new cases using a log-Gaussian process:

$$\log(\lambda_t) \sim N(\log(\lambda_{t-1}), \sigma^2)$$

where:

- $\lambda_t = E[N_t]$ is the expected number of cases at time t .
- This follows a Random Walk assumption, ensuring smooth epidemic progression.

To incorporate additional lagged predictors:

$$\log(\lambda_t) \sim N(\log(\lambda_{t-1}) + \beta_0 + \beta_1 \log(m_{t-l}), \sigma^2)$$

where:

- m_{t-l} is an auxiliary data stream lagged by l days (e.g., ICU admissions).
- β_0 and β_1 are regression coefficients capturing relationships between cases and auxiliary indicators.
- If $\beta_1 = 0$, we revert to the simpler random walk model.

METHODOLOGY

Bayesian Model Components

Delay Distribution (Reporting Model)

The delay distribution describes how long it takes for COVID-19 cases to be reported after they actually happened. The delay process is modeled as a hazard function:

$$P_{t,d} = P(\text{delay} = d | \text{infection at } t)$$

Following Günther et al. (2020), we assume a discrete-time hazard function:

$$h_{t,d} = L_{t,d} + W_{t,d}$$

where:

- $h_{t,d}$ is the reporting hazard at delay d .
- $L_{t,d}$ captures structured delay effects (e.g., weekday effects).
- $W_{t,d}$ represents residual variability.

Cases reported with delay d follow a negative binomial distribution:

$$n_{t,d} | \lambda_t, P_{t,d} \sim \text{NB}(\lambda_t P_{t,d}, \phi)$$

where ϕ is an overdispersion parameter accounting for variability in reporting delays.

METHODOLOGY

Implementation & Inference

The model is implemented using a hierarchical Bayesian framework, incorporating:

- Stan (MCMC sampling) for posterior inference.
- R (RStan, EpiNow2) for model estimation.

The Bayesian framework allows us to:

- Estimate missing case counts in real time.
- Adjust for uncertainty in delay patterns.
- Incorporate external predictors (ICU admissions, testing rates).

PERFORMANCE EVALUATION PLAN

● Root Mean Squared Error (RMSE)

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

● Continuous Ranked Probability Score (CRPS)

Measures the accuracy of the predicted probability distribution.

● Prediction Interval Coverage

Evaluating whether observed case counts fall within 75%, 90%, and 95% credible intervals.

● Log Scores:

Measuring how well a probability distribution assigns likelihood to the true observed value. A lower log score means the model assigned a high probability to the correct value, indicating a better forecast.




PERFORMANCE EVALUATION

- date → The reporting date.
- med_r → The median of the nowcasted case distribution.
- q5_r / q95_r → The 5th and 95th percentiles (prediction interval).
- n_true_retro → The actual number of cases reported retrospectively.
- err_r_7 → The absolute error between nowcasted and true values.
- log_r_7 → Log score for the nowcast (lower is better).
- crps_r_7 → Continuous Ranked Probability Score (CRPS), measuring forecast accuracy.
- rmse_r_7 → Measures how far the median prediction is from the true value.

| date | med_r | q5_r | q95_r | n_true_retro | err_r_7 | log_r_7 | crps_r_7 | rmse_r_7 |
|------------|-------|------|-------|--------------|---------|---------|----------|----------|
| 7/27/2020 | 10 | 3 | 24 | 6 | 4.7143 | 2.9304 | 3.1682 | 4.7143 |
| 8/13/2020 | 1 | 0 | 5 | 6 | 2.8571 | 2.3516 | 1.9692 | 2.8571 |
| 11/24/2020 | 63 | 25 | 152 | 69 | 6 | 3.8333 | 5.2580 | 6 |



NEXT STEPS AND TIMELINE

- Finalizing Model Variants (Case-based and ICU-based lagged predictors).
 - Running Full Model Evaluation (Scoring rules and interval coverage).
 - Developing a Real-Time Shiny Dashboard for Visualization.
 - Final Thesis Writing Plan (Structuring results and discussion).
- 
- 
- 

The background features three vertical stripes on the left: a wide reddish-pink stripe, a narrower teal stripe, and a tan stripe. On the right side, there are two rectangular areas filled with a grid of small, light red dots. The text "THANK YOU" is centered in a large, bold, black sans-serif font.

THANK YOU