ONE HEALTH Lab
with Computational Epidemiology

가톨릭대학교
THE CATHOLIC UNIVERSITY OF KOREA

Dr Lee Jong-Wook Fellowship

Surveillance Analysis and Application 3

11 September 2025

Achangwa Chiara, M.Sc., Ph.D.
Department of Preventive Medicine,
The Catholic University of Korea, Seoul, Korea,
ciaraacha@gmail.com

1

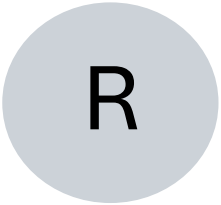
Lecture Content

Definition and scope

Why surveillance matters!

Components of Surveillance

Analysis of surveillance data



Practice in R

Key takeaways

2

Requirements

R

R studio

3

Definition and Scope

✓Public health surveillance

- Ongoing, systematic collection, analysis, interpretation, and dissemination of health data
- Used for action: prevention, control, and policy

✓Analysis

- Transform raw data into signals and insights (time, place, person)

✓Application

- Translate insights into actions: alerts, prevention and control measures, vaccination, and evaluation

4

Why surveillance?

Detect	Detect outbreaks early and guide rapid response
Monitor	Monitor disease trends, severity, and inequities
Measure	Measure program impact (vaccination, NPI, case management)
Allocate	Allocate resources efficiently and plan preparedness
Support	Support risk communication and public trust

5

Components of Surveillance

Planning and design	Objectives, case definitions, data flows, governance
Data Collection	Case reports, labs, sentinel sites, digital/event-based sources
Data analysis	Quality checks, descriptive epi, models, thresholds
Interpretation & dissemination	Bulletins, dashboards, situation reports
Application & evaluation	Control measures, after-action reviews, system evaluation

6

Analysis of Surveillance data

Core Descriptive Analysis

- Time: time-series analysis, interrupted time series analysis, epidemic curves,
- Place: maps, incidence choropleths, cluster detection
- Person: age/sex/ attack rates and CFR

Spatiotemporal Analysis

- Spatial, space–time analysis through Moran's I

Disease transmission analysis / modeling

- Incubation period and serial interval
- Basic reproduction number (R_0) and (R_t)
- SEIR (Susceptible – Exposed – Infected – Recovered)

Time Series Analysis for Surveillance

A Time Series is a collection of observations *y* made sequentially in time *t*, that is a series of values collected over a period of **time**

- Every year

Year	2010	2011	2012	2013
Prevalence of HBV	2.3	2.6	3.1	3.3

- Every quarter

Quarter	Quarter 1	Quarter 2	Quarter 3	Quarter 4
N° of Measles	3.6	5.2	13.2	4.3

- Every month

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
N° of IFV cases	120	123	133	125	130	132	132	132	133

- Every week

Week	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
N° of NTD cases	520	523	733	465	450	451	554	185	253	158

- Everyday

Days	1	2	3	4	5	6	7	8	9	10	11	12
N° of NTD cases	120	123	133	125	130	132	132	132	133	125	128	128

✓**A discrete time series** is a collection of data with distinct values or categories, like counts of events, categorical statuses, etc, recorded at specific times.

Week	Influenza A Cases	Vaccination Coverage (%)	School Closure
1	15	12	0
2	22	15	0

✓**A continuous time series** is a collection of data, a continuous range of values, like temperature, height etc, at regular time intervals.

Date	Anti S IgG Level	Vaccination Status	Age Group	Infection Status
2025-05-01	95.3	Vaccinated	30–50	Infection
2025-05-02	108.5	Not vaccinated	30–50	Infection

9

Importance of time series in surveillance

- ✓Understanding the past behavior of a dataset
- ✓Can forecast future trends/activities, hence can be used to plan future interventions
- ✓Evaluate current interventions
- ✓Facilitates comparison of periods/intervention

10

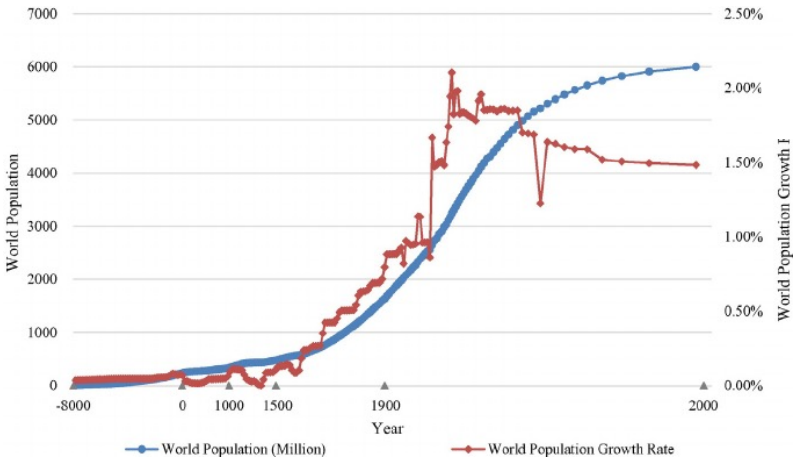
Patterns in Time Series for Surveillance.

1. Trends:

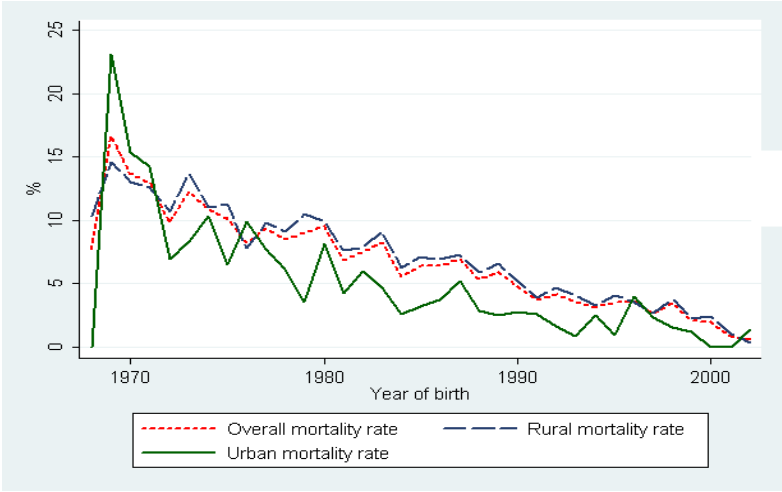
the upward, downward, or no pattern observed

Why measure trends?

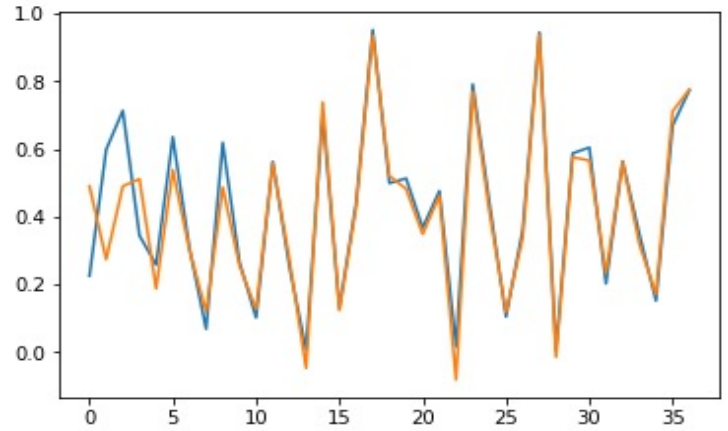
- Past behavior
- Estimation e.g, estimating infection peaks
- Other components e.g, comparative analysis among groups



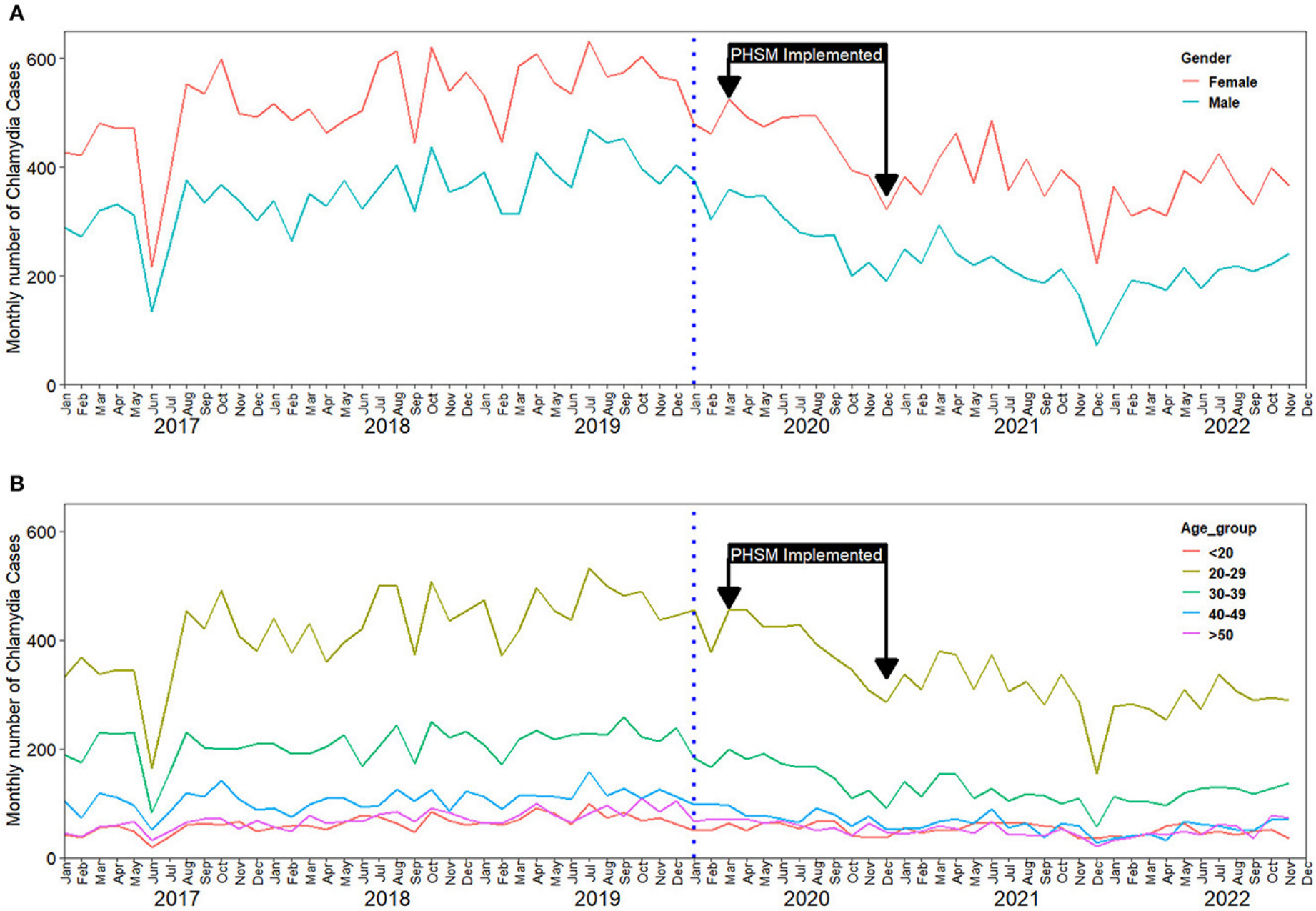
Upward pattern



Downward pattern



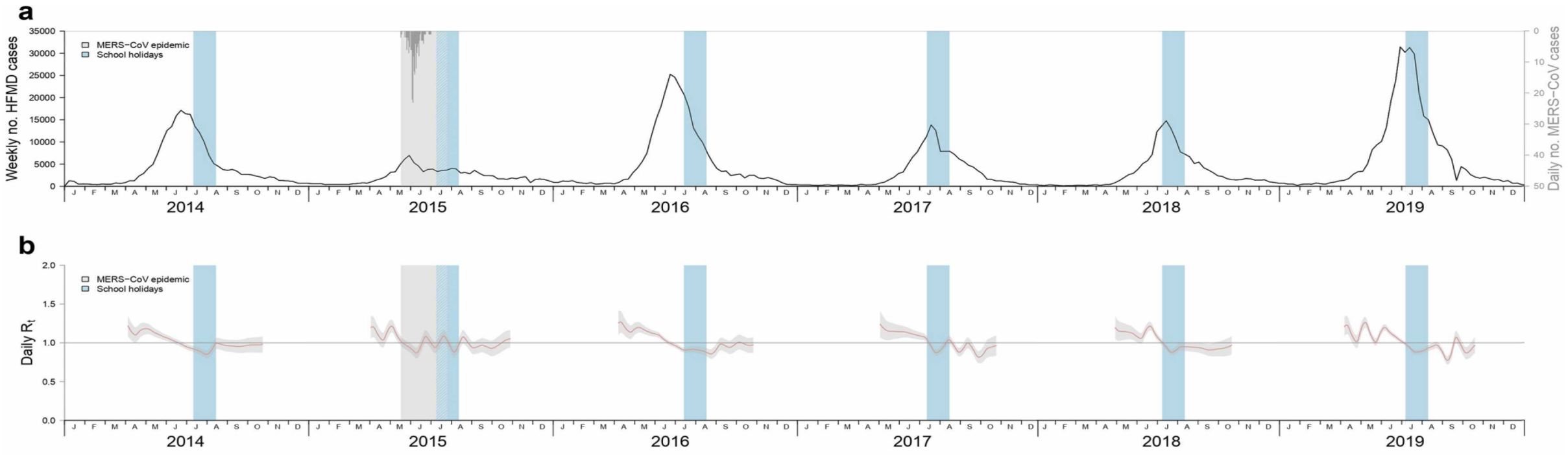
No pattern



Achangwa *et al.*, *Frontiers in Public Health, 2023.*

2. Periodicity: Repetition of behavior in a regular pattern

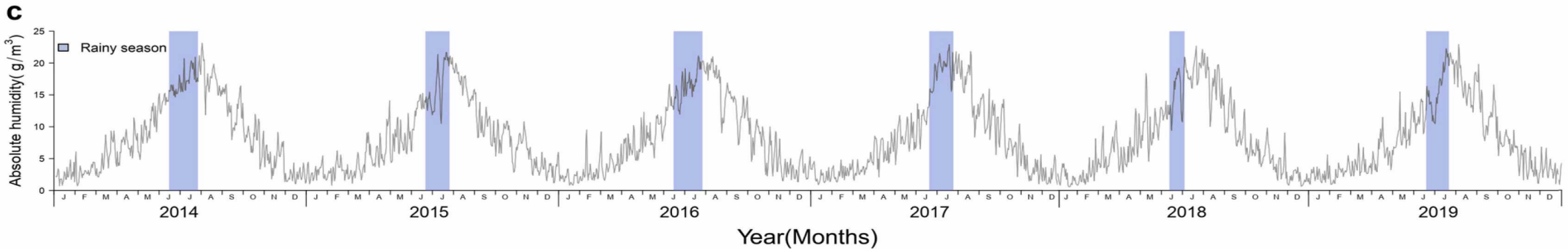
✓Repeated cycles of variation usually over a period of less than one year



Ryu et al., *Journal of Infection and Public Health*, 2023.

3. Seasonality: Periodic behavior with a known period (hourly, monthly, every 2 months...)

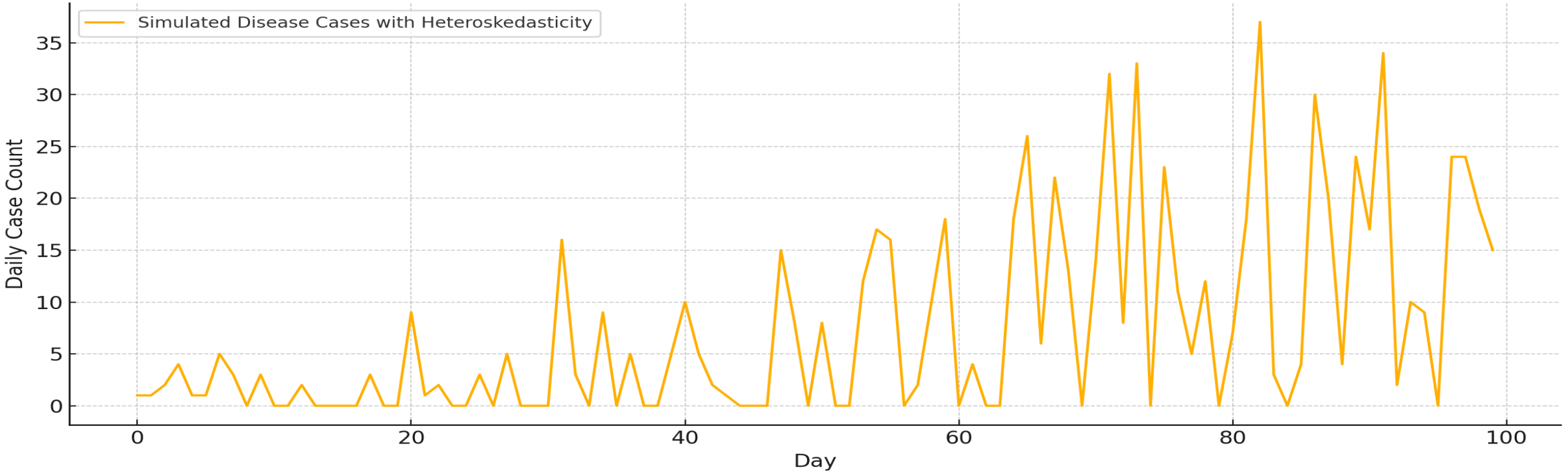
Concern	Seasonal Pattern	Explanation
Influenza (Flu) cases	Peaks in winter months (Nov–Feb)	Cold, dry air facilitates virus survival and indoor crowding increases spread.
Allergy-related hospital visits	Peaks in spring and fall	Due to increased pollen (spring) and mold (fall).
Malaria cases	Peaks in the rainy season in tropical regions	Rain creates stagnant water, ideal for mosquito breeding.
Vitamin D deficiency	Higher in winter	Reduced sunlight exposure in winter months.



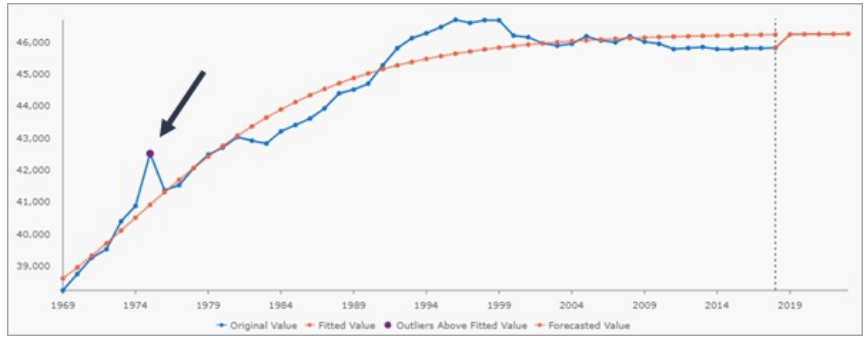
Ryu et al., *Journal of Infection and Public Health*, 2023.

4. Heteroskedasticity: changing variance

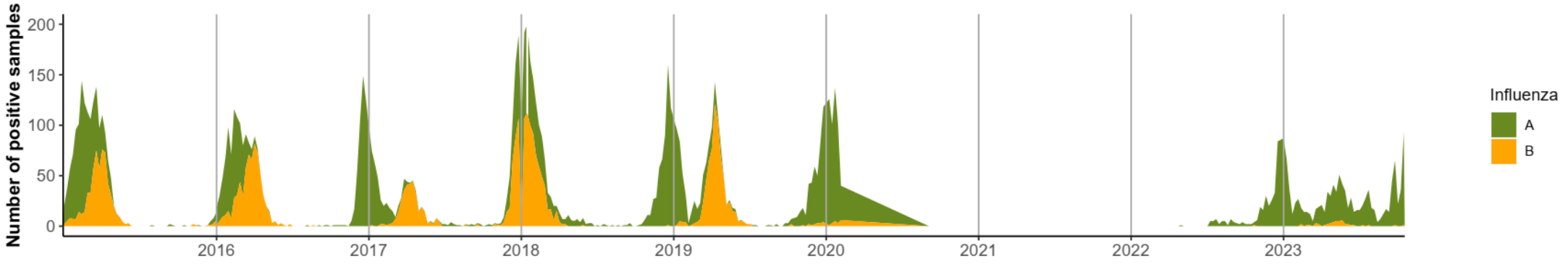
✓Changing the spread involving periods of low and high peaks across a period



5. Missing data, outliers, and breaks



- ✓ Data entry errors and data reporting delays
- ✓ Extreme values that deviate significantly from the trend or seasonality.
- ✓ Sudden changes in the underlying process (e.g., policy shifts, lockdowns).



Time series models

1-Classical models

Model	Description
AR (Autoregressive)	Predicts future values based on past values: $X_t=\phi_1 X_{t-1}+\epsilon_t$
MA (Moving Average)	Model where the current value of the series is expressed as a linear combination of past white noise error terms $X_t=\mu+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+\cdots+\theta_q\epsilon_{t-q}+\epsilon_t$
ARMA	Combines AR and MA: good for stationary series. $X_t=\phi_1 X_{t-1}+\phi_2 X_{t-2}+\cdots+\phi_p X_{t-p}+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+\cdots+\theta_q\epsilon_{t-q}+\epsilon_t$
ARIMA **Auto-Regressive Integrated Moving Average**	Includes differencing to handle non-stationarity.
SARIMA	ARIMA with seasonal components.
Seasonal Decomposition (STL)	Breaks a series into seasonal, trend, and residual components.

Time series models

2- Heteroskedasticity models

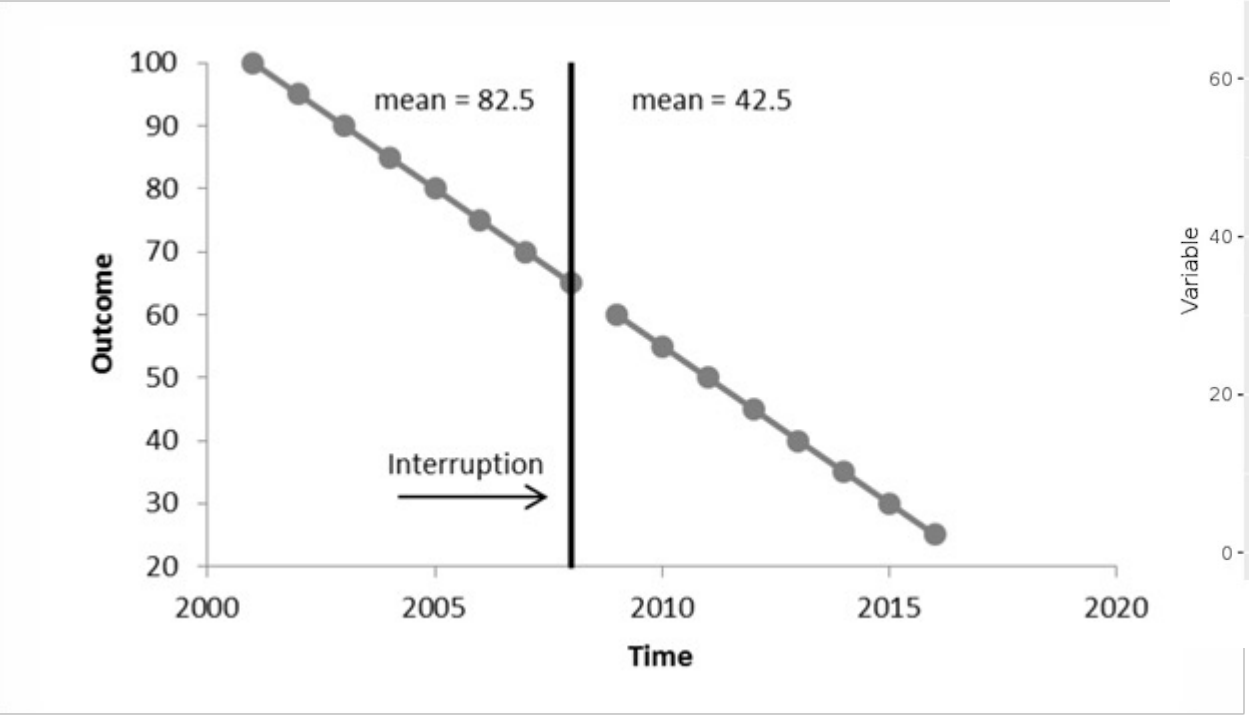
Model	Description
ARCH (Autoregressive Conditional Heteroskedasticity)	Models time-varying variance based on past squared errors.
GARCH (Generalized ARCH)	Extends ARCH by modeling variance as a function of both past errors and past variances.

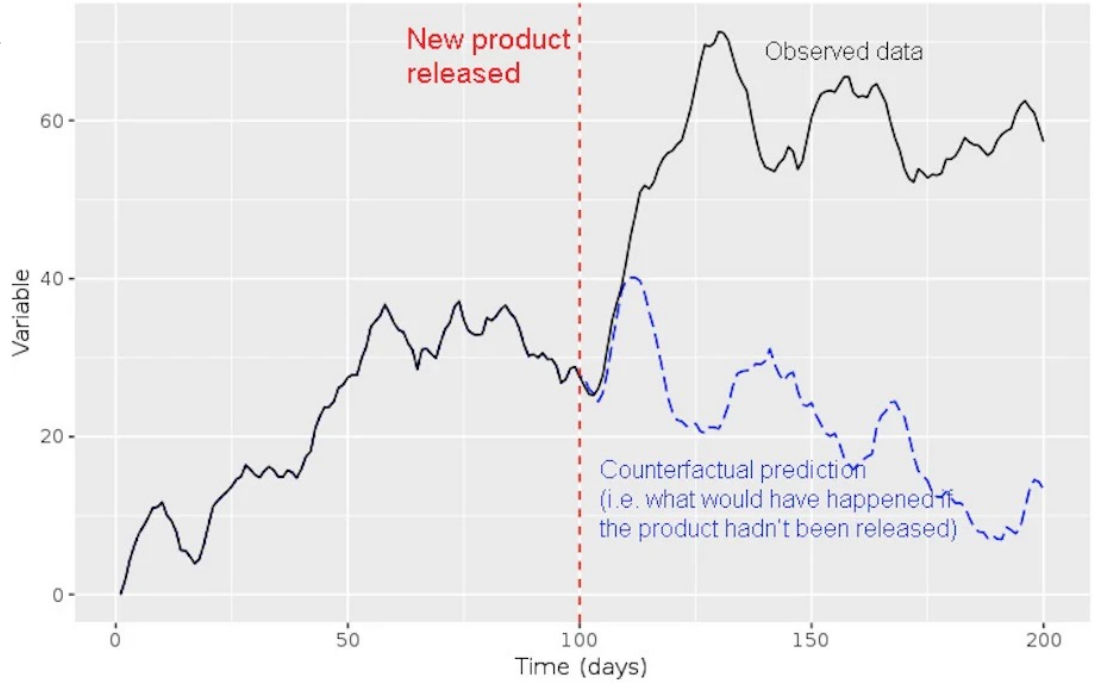
Interrupted Time Series (ITS) Analysis for Surveillance

✓Analysis of time series data (i.e., an outcome measured over time)

✓Comparison before and after an intervention or interruption

✓Particularly useful for assessing the impact of policy or some other healthcare initiative (pre- to post-comparison)





ITS Analysis

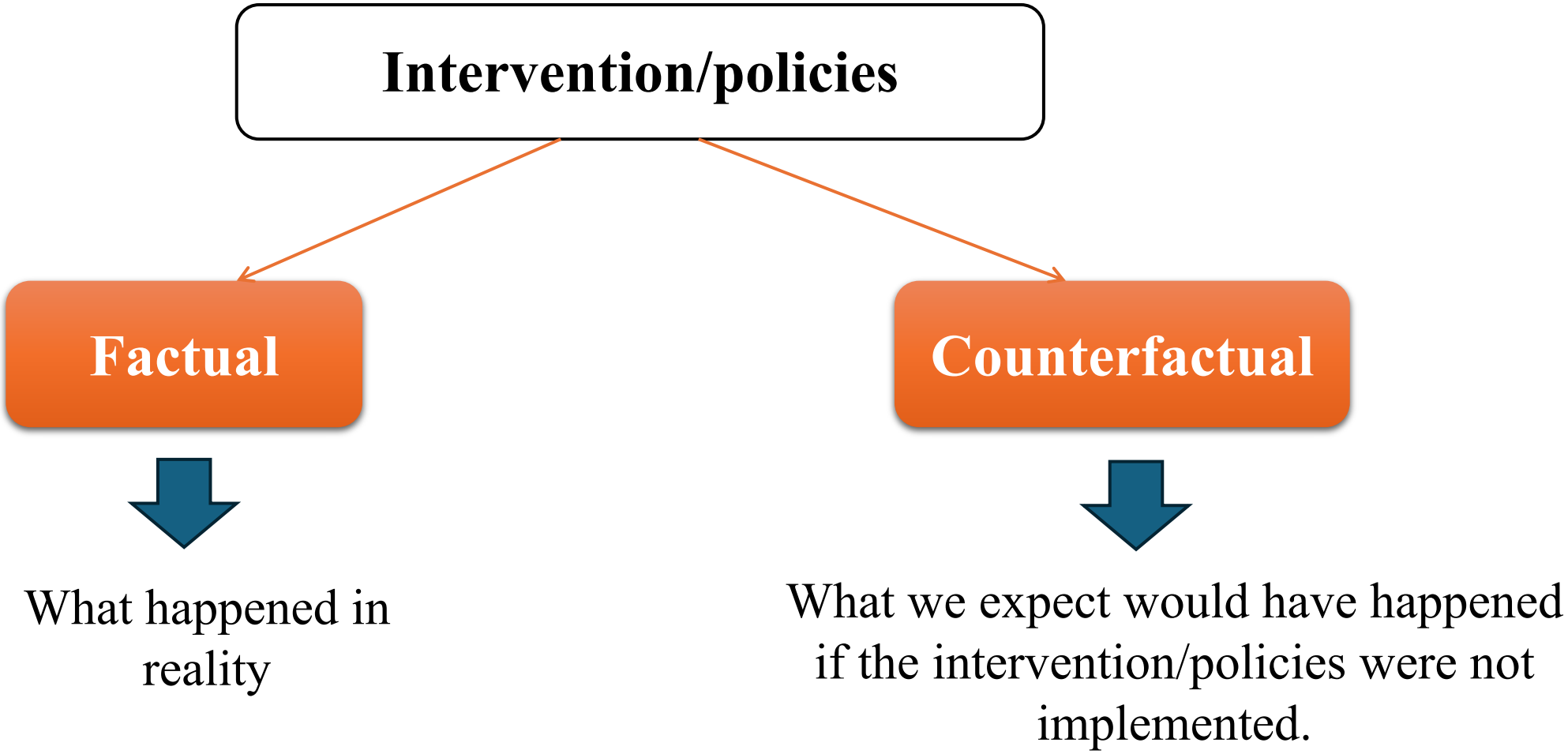
Examples:

1-Changes in the daily number of COVID-19 cases after the implementation of PHSM and vaccination

2- Changes in the number of live births after the implementation of the child encouragement policies

1. Was the implementation of PHSM and vaccination policies correlated with a decrease in the number of COVID-19 cases?
2. Was the implementation of child encouragement policies correlated with an increase in the number of live births?

ITS Analysis



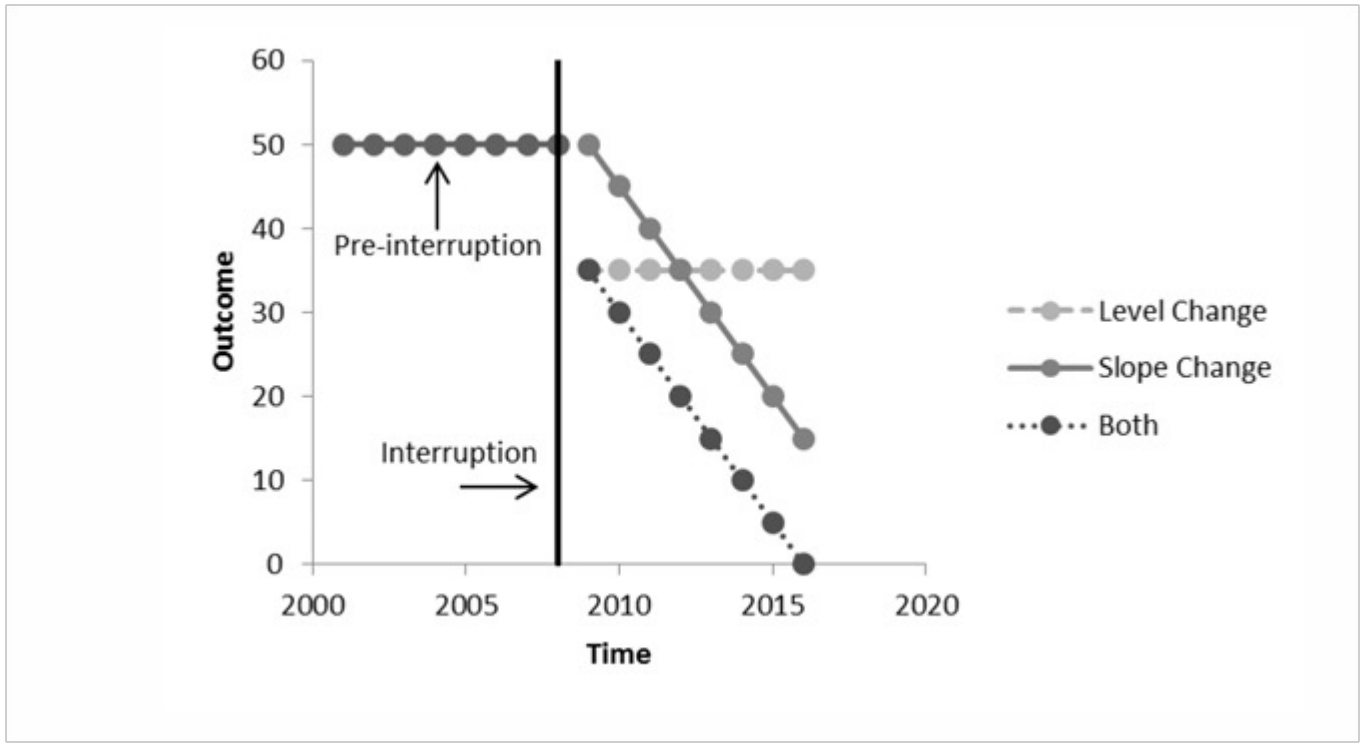
ITS Analysis

The change could take two forms, including;

✓A level/step change immediately after the intervention

✓A slope/trend change after the intervention

✓Both



Are there significant changes in level and/or slope following the intervention?

Types of ITS analysis used for surveillance

Uncontrolled ITS, one intervention

e.g Evaluate the effect of the nationwide COVID-19 lockdown (implemented in March 2020) on the incidence of influenza.

Uncontrolled ITS, two interventions

e.g To evaluate how two national interventions (1993-family planning policy, 2004 – childbirth promoting activities) affected the number of births in Korea between 1975 and 2022.

Controlled ITS, one intervention

e.g Evaluating the Impact of a Smoking Ban on Hospital Admissions for Asthma

ITS model equations

Factual

gls(quantity.x ~ Time + Intervention + Post.intervention.time, data = data, method="ML")

Counterfactual

gls(quantity.x ~ Time, data = data, method="ML")

Practice and Application

Data

Download the mock data sets from the link below:

http://onehealth.or.kr/surveillance.html

Datasets to be used

1 - Daily number of SARS-CoV-2 cases 2020 – 2022 - mockdata1

2 - Monthly reported cases of Chlamydia 2017 – 2022 - mockdata2

3 - Weekly number of influenza cases - mockdata3

4 - Yearly fertility rate in Korea 1975 – 2022 - mockdata4

26

South Korea reports 26 new coronavirus cases

Published : May 13, 2020 - 13:07:28

27

Plotting Time Series Plots in R

Scenario 1: Time series analysis using daily data

You are a worker at the Ministry of Health in your country. You have been asked to brief the Lee Jong-wook Public Health Fellowship 2025 cohort on the COVID-19 trends between 2020 and 2022. Using the daily COVID-19 case data (2020-2022) provided by the COVID-19 Surveillance System from KCDA;

1) Plot an overall time series of the daily number of reported cases

2) Briefly describe and interpret your results

Note: Use mockdata1 and R for your analysis.

✓Open R Studio (Installed from: https://posit.co/download/rstudio-desktop/)

✓Open a new R script file in R Studio

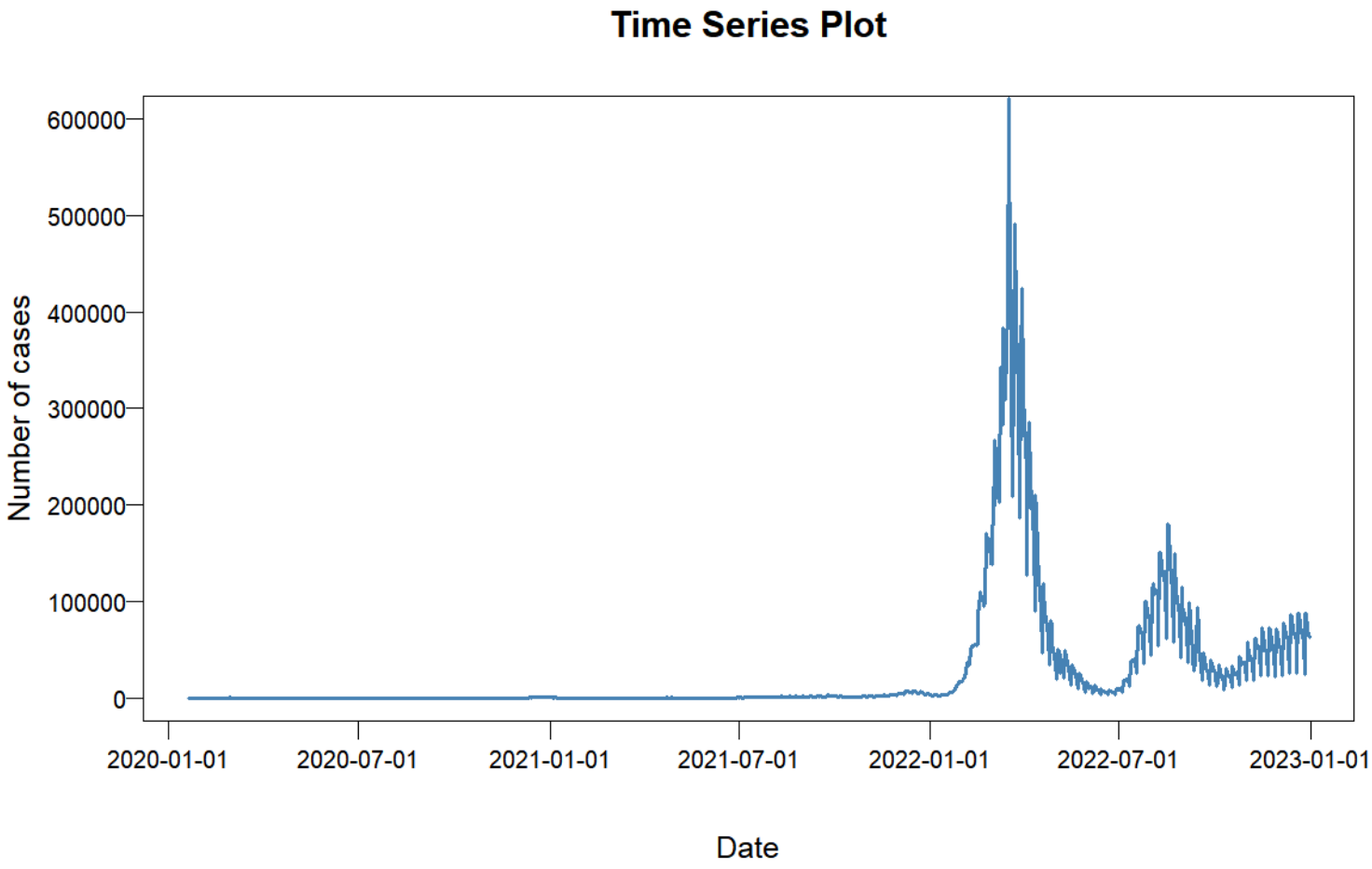
✓Load the following R libraries;

```
# install.packages("package") if not yet installed
library(readr)
library(dplyr)
library(psych)
library(forecast)
```

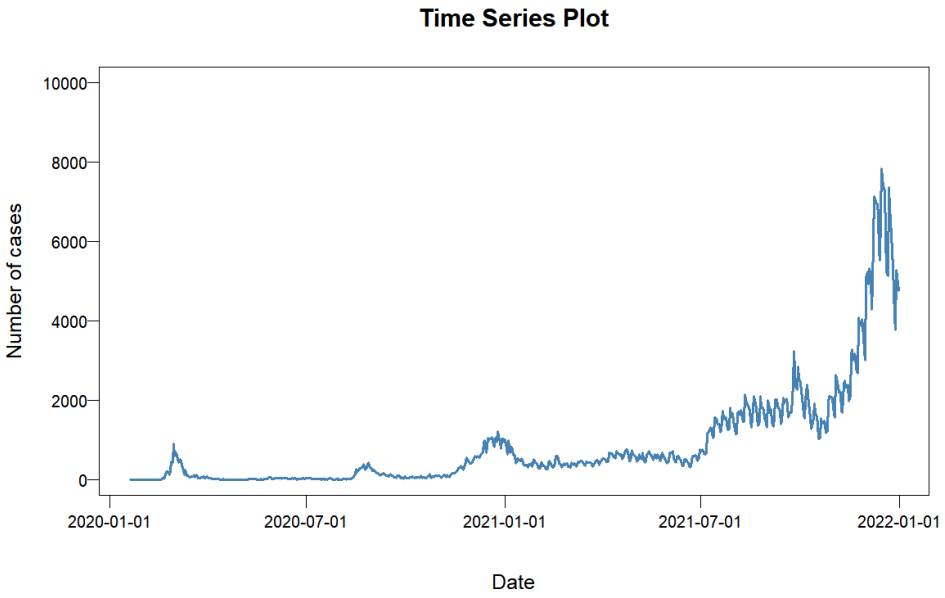
```
# open a new R script file in R studio
# load the following R packages
library(readr)
library(psych)
library(dplyr)
library(forecast)

# get and set your working directory
getwd()
setwd("C:/Users/ACHANGWA CHIARA/Desktop")

# import your data
df <- read_csv("C:/Users/ACHANGWA
CHIARA/Desktop/Dissertation_2023/CFR_Region/mockdata1.csv")

# Run summary statistics
stats_df <- psych::describe(df$Number_of_new_cases)
print(stats_df)

# Set Date column to Date format
df$Date <- as.Date(df$Date)
```

```
#setting the margin
par(mgp = c(3, 0.5, 0))
par(font.lab = 1)

# Plot time series using base R
plot(df$Date, df$Number_of_new_cases,
    type = "l",                  # line plot
    col = "steelblue",           # line color
    lwd = 2,                     # line width
    main = "Time Series Plot",   # title
    xlab = "Date",               # x-axis label
    ylab = "Number of cases",    # y-axis label
    ylim = c(0, 600000),         # y-axis limits
    xaxt = "n",                  # suppress x-axis for custom formatting
    yaxt = "n",                  # suppress y-axis for custom formatting
    cex.lab = 1.0)    # increase axis label font size

# Customize axes
axis(1, at = pretty(df$Date), labels = format(pretty(df$Date), "%Y-%m-
%d"), las = 1, cex.axis = 0.8)
axis(2, at = seq(0, 600000, by = 100000), las = 1, cex.axis = 0.8)

# Set file path and image size
png("timeseries_plot.png", width = 1200, height = 800, res = 150)

dev.off()
```

Output


Time Series Plot
600000
500000
400000
300000
200000
100000
0
Number of cases
2020-01-01
2020-07-01
2021-01-01
2021-07-01
2022-01-01
2022-07-01
2023-01-01
Date

```r
### subset Jan 2020 to Dec 2021

subset_data <- df[df$Date >= as.Date("2020-01-01") & df$Date <= as.Date("2021-12-31"), ]

stats_subset <- psych::describe(subset_data$Number_of_new_cases)
print(stats_subset)
#median(subset_data$Number_of_new_cases)
# View the first few rows of the subset
View(subset_data)
# Plot the subset
png("subset_timeseries_plot.png", width = 1200, height = 800, res = 150)

par(mgp = c(3, 0.5, 0))

par(font.lab = 1)
# Plot time series using base R
plot(subset_data$Date, subset_data$Number_of_new_cases,
     type = "l",                          # line plot
     col = "steelblue",                   # line color
     lwd = 2,                             # line width
     main = "Time Series Plot",           # title
     xlab = "Date",                       # x-axis label
     ylab = "Number of cases",            # y-axis label
     ylim = c(0, 10000),                  # y-axis limits
     xaxt = "n",                          # suppress x-axis for custom formatting
     yaxt = "n",                          # suppress y-axis for custom formatting
     cex.lab = 1.0)    # increase axis label font size

# Customize axes
axis(1, at = pretty(subset_data$Date), labels = format(pretty(subset_data$Date), "%Y-%m-%d"), las = 1, cex.axis = 0.8)
axis(2, at = seq(0, 10000, by = 2000), las = 1, cex.axis = 0.8)

dev.off()
```



Scenario 2: Time series analysis using monthly data

You are a worker at the Ministry of Health in your country. You have been asked to brief the Lee Jong-wook Public Health Fellowship on how the COVID-19 pandemic influenced the Chlamydia infection trends. Using the monthly surveillance data (2017-2022) provided by the Chlamydia surveillance system from KCDA;

1) Plot an overall time series of the reported cases

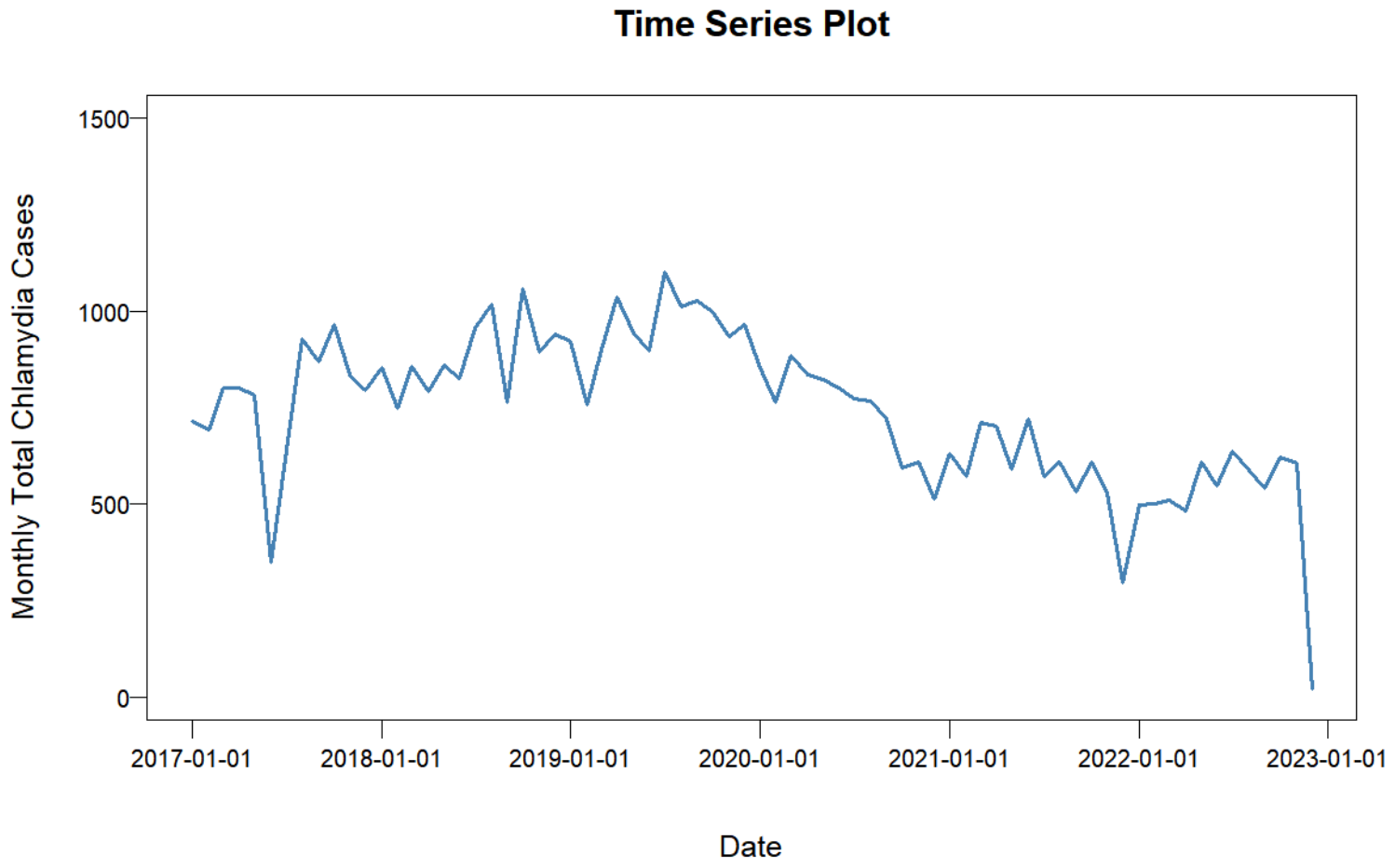
2) Briefly describe and interpret your results

Note: Use mockdata2 and R for your analysis.

34

```r
data <-  read_csv("lee fellowship/mockdata2.csv")
View(data)
data$Month <- match(data$Month, month.abb)

data$Time <- data$Year + data$Month / 12

data$Time <- as.Date(paste(data$Year, data$Month, "01", sep = "-"))

par(mgp = c(3, 0.5, 0))
par(font.lab = 1)

# Plot time series using base R
plot(data$Time, data$Total,
     type = "l",                          # line plot
     col = "steelblue",                   # line color
     lwd = 2,                             # line width
     main = "Time Series Plot",           # title
     xlab = "Date",                       # x-axis label
     ylab = "Monthly Total Chlamydia Cases",          # y-axis label
     ylim = c(0, 1500),                   # y-axis limits
     xaxt = "n",                          # suppress x-axis for custom formatting
     yaxt = "n",                          # suppress y-axis for custom formatting
     cex.lab = 1.0)    # increase axis label font size

# Customize axes
axis(1, at = pretty(data$Time), labels = format(pretty(data$Time), "%Y-%m-%d"), las = 1, cex.axis = 0.8)
axis(2, at = seq(0, 1500, by = 500), las = 1, cex.axis = 0.8)
png("Chlamydia_timeseries_plot.png", width = 1200, height = 800, res = 150)
dev.off()

```

Output


Time Series Plot
Monthly Total Chlamydia Cases
1500
1000
500
0
2017-01-01
2018-01-01
2019-01-01
2020-01-01
2021-01-01
2022-01-01
2023-01-01
Date

36

Plotting an interrupted time series in R

Data

1 - Daily number of SARS-CoV-2 cases 2020 – 2022 - mockdata1

2 - Monthly reported cases of Chlamydia 2017 – 2022 - mockdata2

3 - Weekly number of influenza cases - mockdata3

4 - Yearly fertility rate in Korea 1975 – 2022 - mockdata4

R code

-Download the corresponding syntax

Scenario 3: Impact of an Intervention on Influenza Cases

You are an epidemiologist at the National Center for Infectious Disease Control. You have been tasked with assessing whether a new influenza vaccination campaign introduced in January 2020 had an impact on the number of influenza-like illness (IFV) cases reported through the national surveillance system.

1) Plot an ITS to assess the impact of the intervention

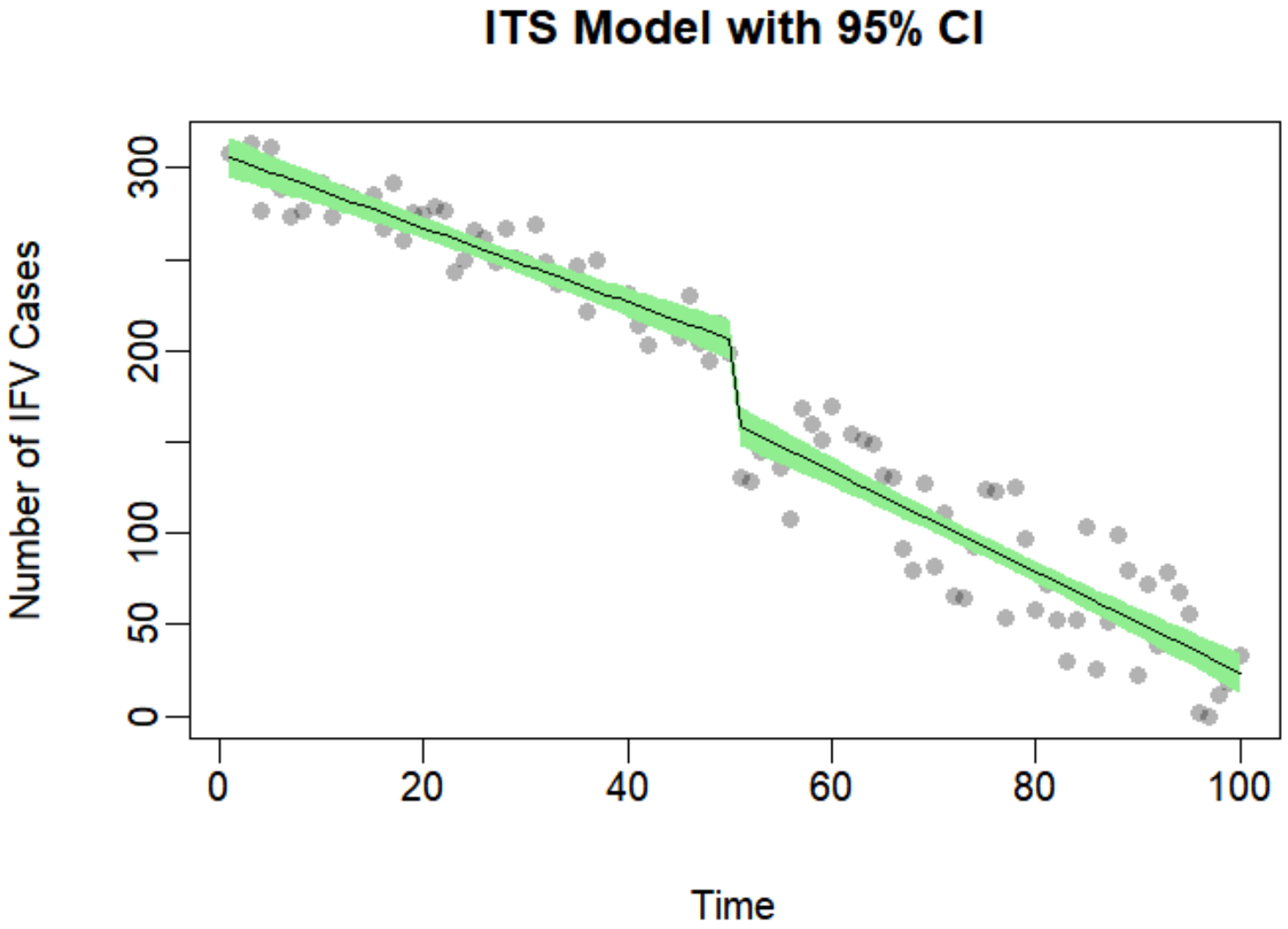
2) Briefly describe and interpret your results

Note: Use mockdata3 and R for your analysis.

The factual model

```r
library(tidyverse)
library(nlme)
library(AICcmodavg)
## ITS model

data <-  read_csv("lee fellowship/mockdata3.csv")
head(data)

#Model factual scenario
model.a = gls(Number_of_IFV_cases ~ Time + Intervention + Post_intervention_time, data = data,method="ML")

# Show a summary of the model
summary(model.a)

#  Add predicted values and standard errors
data <-data %>% mutate(
  model.a.predictions = predictSE.gls (model.a, data, se.fit=T)$fit,
  model.a.se = predictSE.gls (model.a, data, se.fit=T)$se
)
#plot time series
plot(data$Time, data$Number_of_IFV_cases,
     pch = 16, col = rgb(0, 0, 0, 0.3), # semi-transparent black
     xlab = "Time", ylab = "Number of IFV Cases",
     main = "ITS Model with 95% CI")

# Add the confidence ribbon using polygon
polygon(c(data$Time, rev(data$Time)),
        c(data$model.a.predictions - 1.96 * data$model.a.se,
          rev(data$model.a.predictions + 1.96 * data$model.a.se)),
        col = "lightgreen", border = NA)

# Add the model prediction line
lines(data$Time, data$model.a.predictions, col = "black", lty = 1)
```

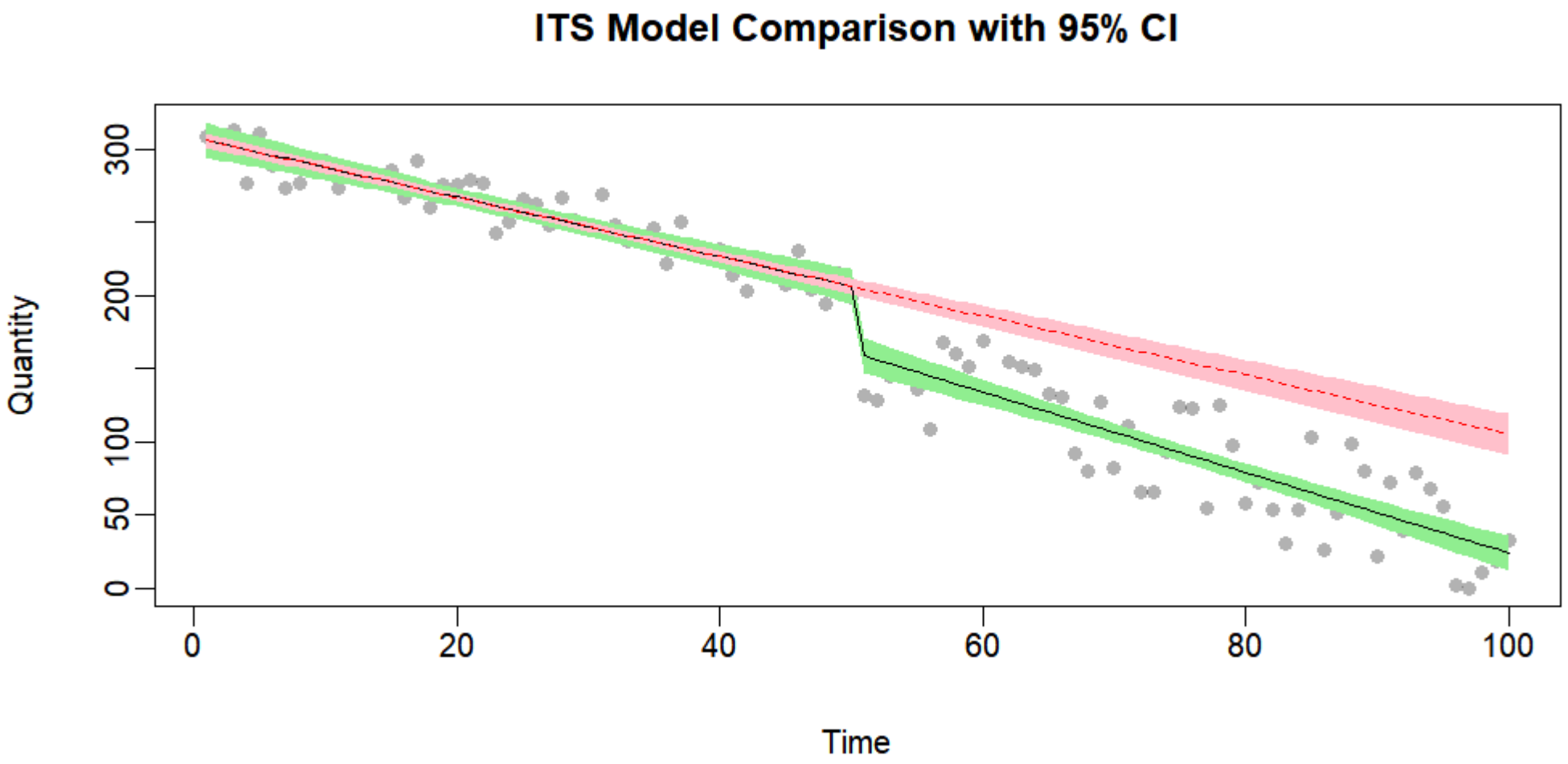
Output


ITS Model with 95% CI
Number of IFV Cases
0
50
100
200
300
0
20
40
60
80
100
Time

The counterfactual model

```r
##### Counterfactual model
model.b = gls(Number_of_IFV_cases ~ Time + Intervention + Post_intervention_time, data = data,method="ML", correlation= corARMA(p=2,q=2, form = ~ Time))
model.b
data<- data %>%
  mutate(
    model.b.predictions = predictSE.gls (model.b, data, se.fit=T)$fit,
    model.b.se = predictSE.gls (model.b, data, se.fit=T)$se
  )
df2<-filter(data,Time<51)
model.c = gls(Number_of_IFV_cases ~ Time, data = df2, correlation= corARMA(p=1, q=1, form = ~ Time),method="ML")
coefficients(model.c)
data<-data %>% mutate(
  model.c.predictions = predictSE.gls (model.c, newdata = data, se.fit=T)$fit,
  model.c.se = predictSE.gls (model.c, data, se.fit=T)$se
)

# Set up the plot with observed data points
plot(data$Time, data$Number_of_IFV_cases,
     pch = 16, col = rgb(0, 0, 0, 0.3), # semi-transparent black points
     xlab = "Time", ylab = "Quantity",
     main = "ITS Model Comparison with 95% CI",
     ylim = range(c(data$Number_of_IFV_cases,
                    data$model.b.predictions + 1.96 * data$model.b.se,
                    data$model.b.predictions - 1.96 * data$model.b.se,
                    data$model.c.predictions + 1.96 * data$model.c.se,
                    data$model.c.predictions - 1.96 * data$model.c.se)))

# Add the green ribbon for model B
polygon(c(data$Time, rev(data$Time)),
        c(data$model.b.predictions - 1.96 * data$model.b.se,
          rev(data$model.b.predictions + 1.96 * data$model.b.se)),
        col = "lightgreen", border = NA)

# Add the pink ribbon for model C
polygon(c(data$Time, rev(data$Time)),
        c(data$model.c.predictions - 1.96 * data$model.c.se,
          rev(data$model.c.predictions + 1.96 * data$model.c.se)),
        col = "pink", border = NA)

```

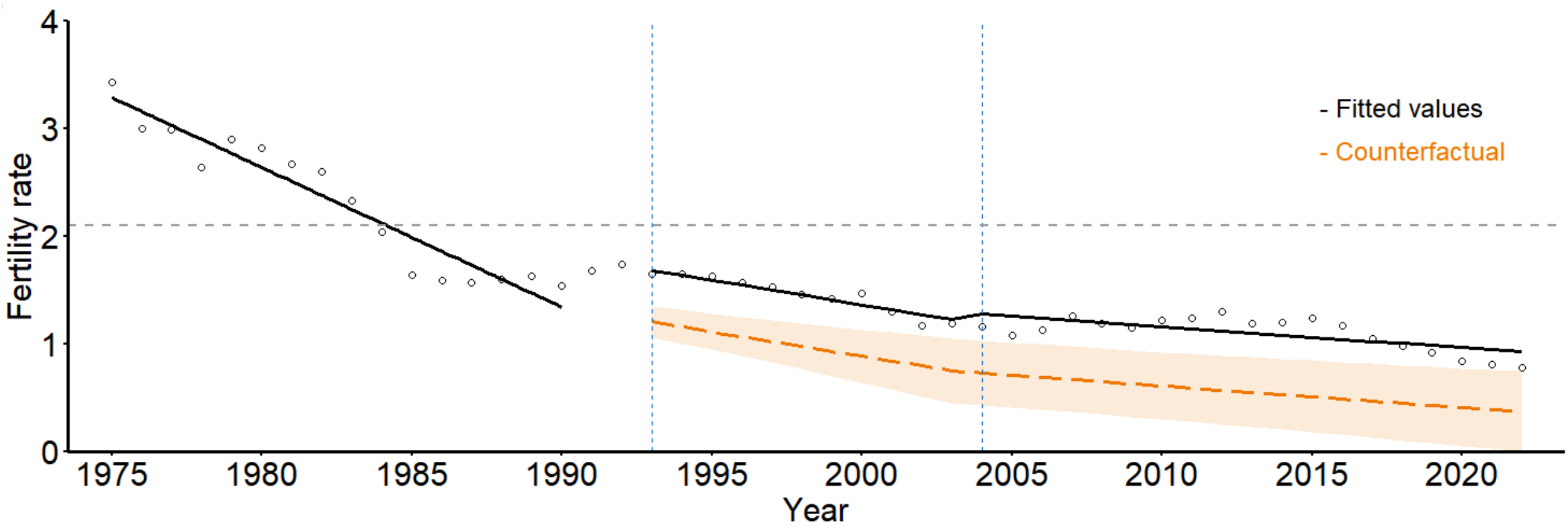
Output


ITS Model Comparison with 95% CI
Quantity
0
50
100
200
300
0
20
40
60
80
100
Time

Scenario 4

The total fertility rate (TFR) in South Korea has shown a continuous decline from 1975 to 2022. Between 1975 and 1992, the government actively promoted a family planning policy, which was discontinued in 1993. Later, in 2004, a birth encouragement policy was introduced to address declining fertility.

Using the provided mockdata4, create an Interrupted Time Series (ITS) plot in R to identify changes in the level and trend of TFR associated with these two policy interventions.



Achangwa et al., *JMIR Public Health and Surveillance, 2025.*

Key takeaways

✓Public health surveillance is action-oriented

- Involves systematic collection, **analysis, interpretation**, and dissemination of health data to guide prevention, control, and policy decisions/actions.

✓Core descriptive analysis is important for surveillance

- Examining data by time (trends, seasonality), place (maps, clusters), and person (age, sex, attack rates) provides the foundation for effective interventions and preparedness.

✓Time series analysis underpins surveillance

-It helps understand disease behavior, forecast future trends, evaluate the impact of interventions, and compare pre- and post-policy periods

ONE HEALTH Lab
with Computational Epidemiology

http://onehealth.or.kr/

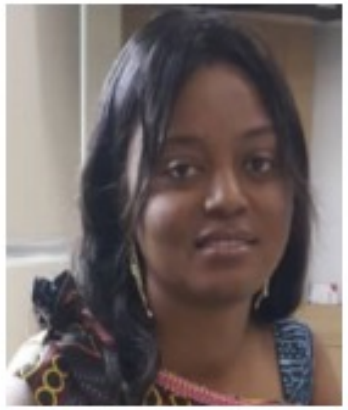
The Team

Sukhyun Ryu

ASSISTANT PROFESSOR

Chiara Achangwa

TEAM LEAD



Seonghui Cho

DATA ANALYST

Leah Yeongseon Ko

DATA ANALYST

Thank you!!