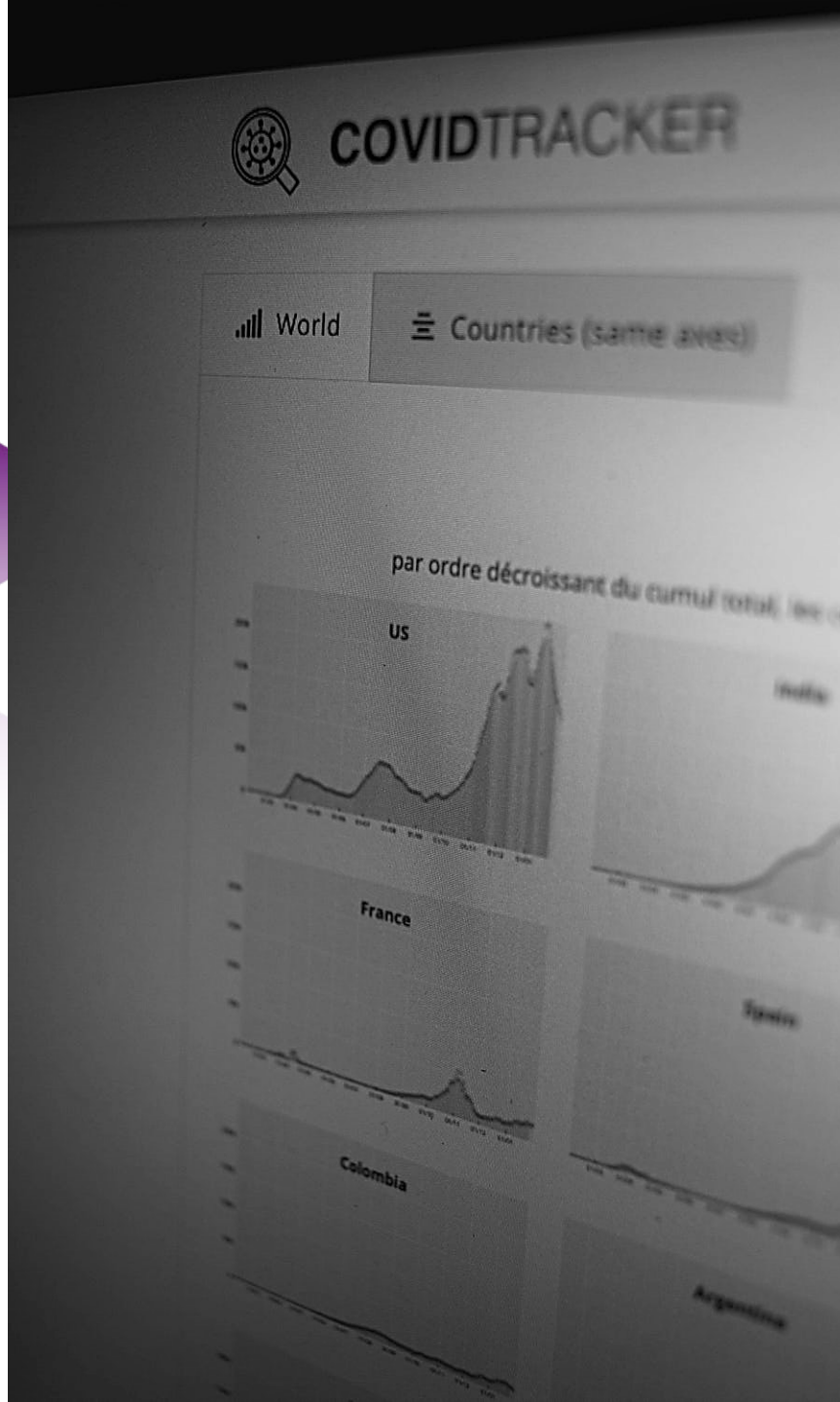
UK Data Service

An Introduction to Time Series Analysis and Forecasting

Nadia Kennar, UKDS



UKRI Economic and Social Research Council

Content

- What is time series data?
- What is time series analysis (TSA) ?
- Types of TSA
- Components of TSA
- Fitting Time Series Models
- Forecasting Techniques (ARIMA)
- Available Software

What is Time Series Data?

- “ A collection of observations obtained though repeated measurements of time”

x	y
2001	425
2002	500
2003	495
2004	700

Intervals (Sampling Frequency/Periodicity)

- Year
- Month
- Days
- Quarterly
- Hourly

Time Series Data Continued

- How is time-series data different to just having a time-field in your data set?
- Can longitudinal data sets be considered time-series?

Scenario: Web Application

Imagine you maintain a web application. You have been asked to analyse when a new user logs in

a) When a new user logs in, you may just update a “last_login” timestep for that user in a single row

b) Or, you treat each login as a separate event

Which option would you choose?

- Option A
- Option B

Option A

User	Company	Last_Login
A	X	01/09/2020 13.09.00
B	Y	01/07/2019 13.09.00
C	X	01/09/2020 13.09.00

Option B

User	Company	Login
A	X	01/09/2019 13.09.00
A	X	02/09/2019 14:01:17
A	X	03/09/2019 13.09.00
C	X	04/09/2019 14.10.12
B	Y	17/10/2021 09.00.00
B	Y	17/11/2021 10.01.01

To summarise

- Almost all data is recorded as a new entry
- The data typically arrives in time order
- Time-intervals can be regular (metrics) or irregular (events)

Different Aims of Time Series Analysis

1. Access the impact of a **single event** *(descriptive)*
2. To study **casual patterns** i.e. the effects of variables rather than events (*explanatory)*
3. To **forecast** future values of a time series using either previous values of one series (or values from others) (*prediction)*

Ashby 2020: *Initial evidence on the relationship between the coronavirus pandemic and crime in the United States*

- Aim: To understand crime patterns during the pandemic. Used police-recorded open crime data to understand how the frequency of certain crime types changed from the start of the pandemic

- Method: used SARIMA models of the frequency of crime types in 16 US cities between 01.01.2016 and 20.01.2020. Forecasts were created from these models to compare the actual calls to the expected

- Data and Code can be found here: https://osf.io/ef4dw/

Reka 2022: ‘Understanding Changing Demand Police during the Coronavirus Pandemic

- Method: Used a mixed-method approach of robust time series analysis and qualitative interviews with force call centre staff to explore changes in police demand during the pandemic. More specifically examining non-demand
- Findings: Reduction in overall calls than would be expected in the absence of the pandemic.

Report: https://www.n8prp.org.uk/wp-content/uploads/sites/315/2022/01/Reka-Solymosi-Police-Demand-Covid.pdf

Case Study

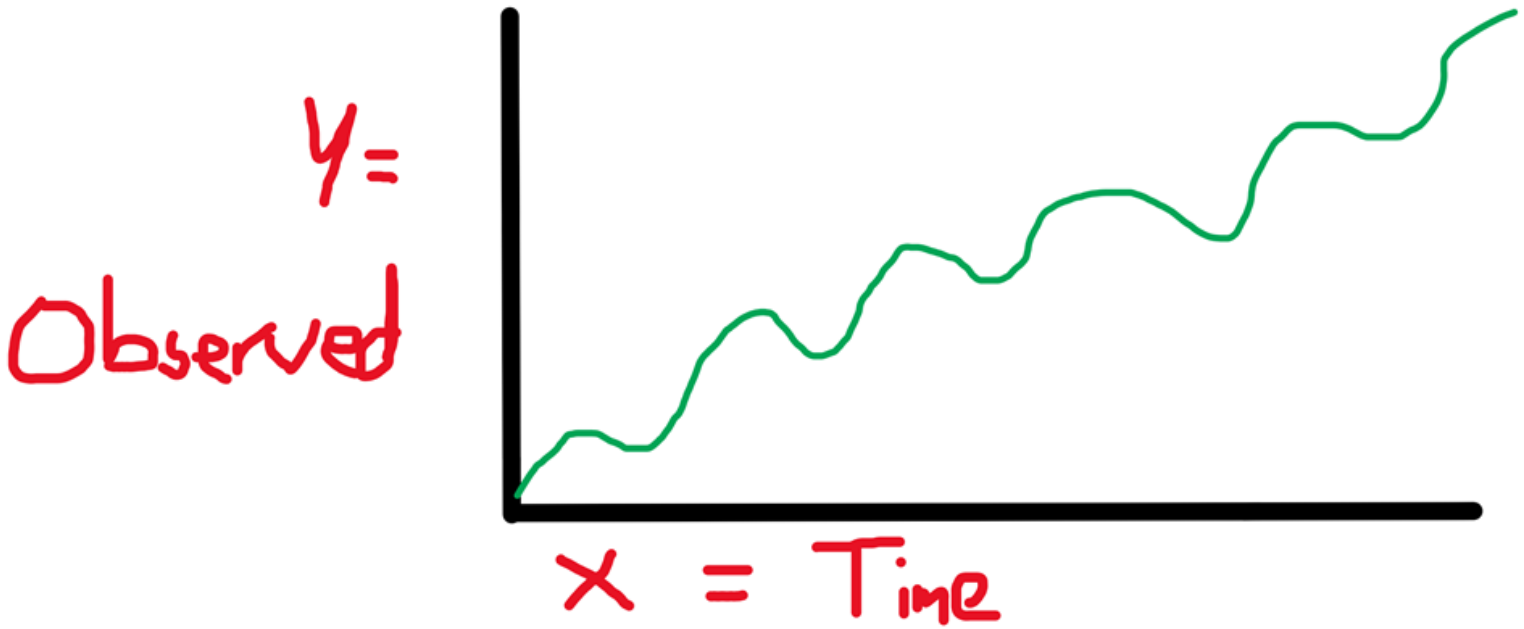
- Using police-recorded crime data, we want to explore Burglary rates from Detroit from 2015- 2020

- Aim:
  - **A) Explore the long-term trend and seasonality in ‘Burglary’ across the city of Detroit**
  - **B) How the frequency of Burglary changed in Detroit in 2020 (the start of the pandemic)**

Steps in Time Series Analysis

1. Explore your data
2. Identify and graph patterns
3. Model the data
4. Predict

Time Series Analysis

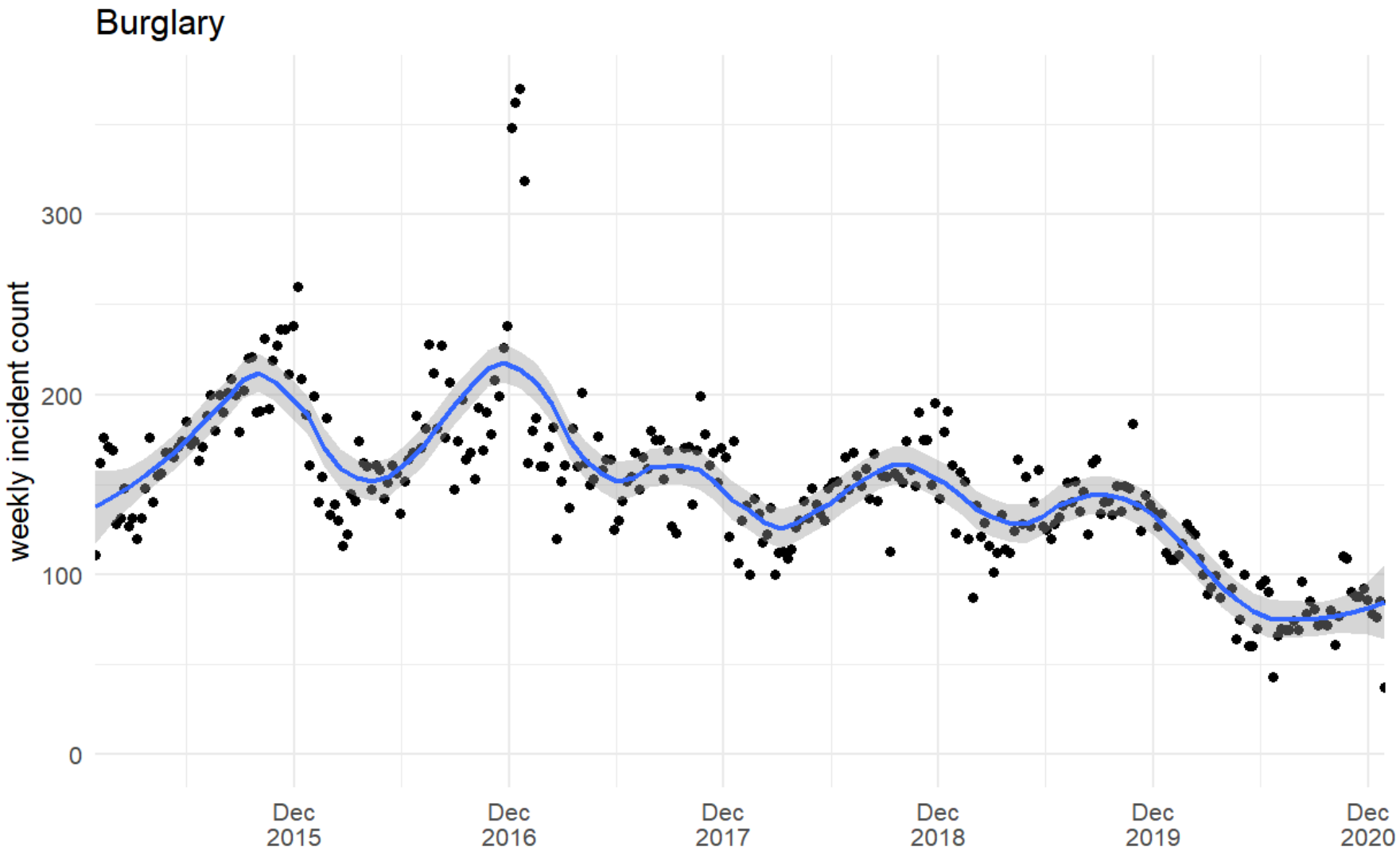


What time interval would you use to study police recorded crime data?

- Yearly
- Monthly
- Weekly
- Hourly
- Minutes
- Any Others?!

Aim *A*

- *A)* Explore the long-term trend and seasonality in 'Burglary' across the city of Detroit



Components of Time Series Analysis

- **Trend** – The linearity (increasing – decreasing)
- **Cyclic** – repeated patterns of non-periodic fluctuations
- **Seasonality**– Repeating patterns of behaviour over time
- **Random or irregular movements (noise)** – Variation that cannot be explained

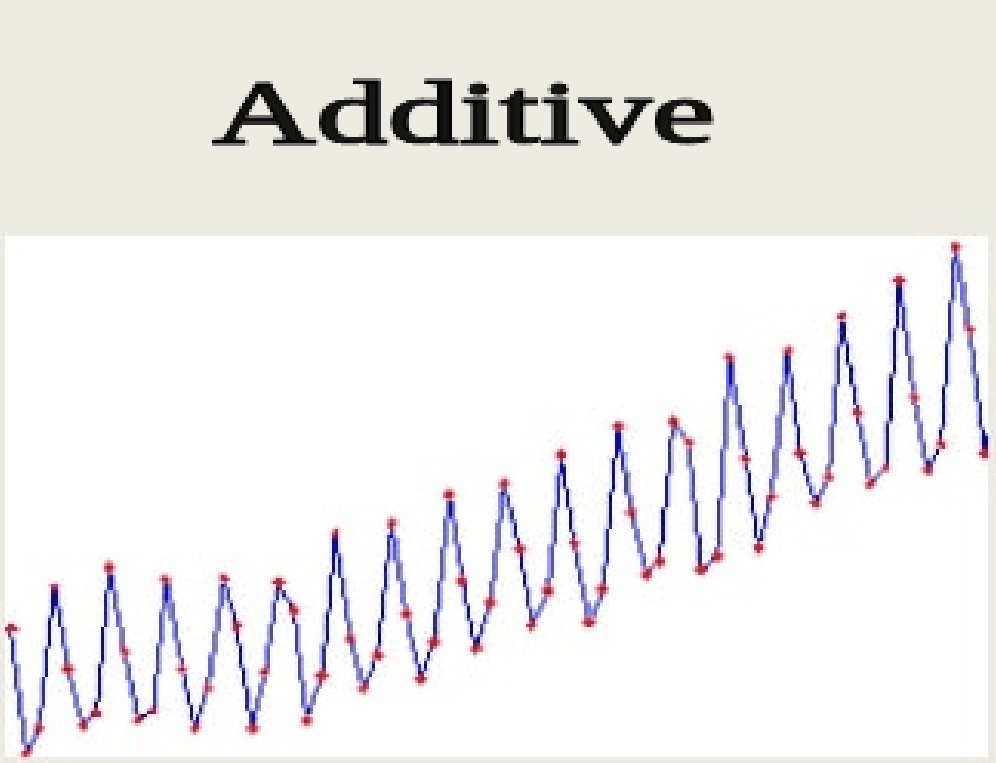
- The combination of these components can either be
  - Additive
  - Multiplicative

Additive vs Multiplicative

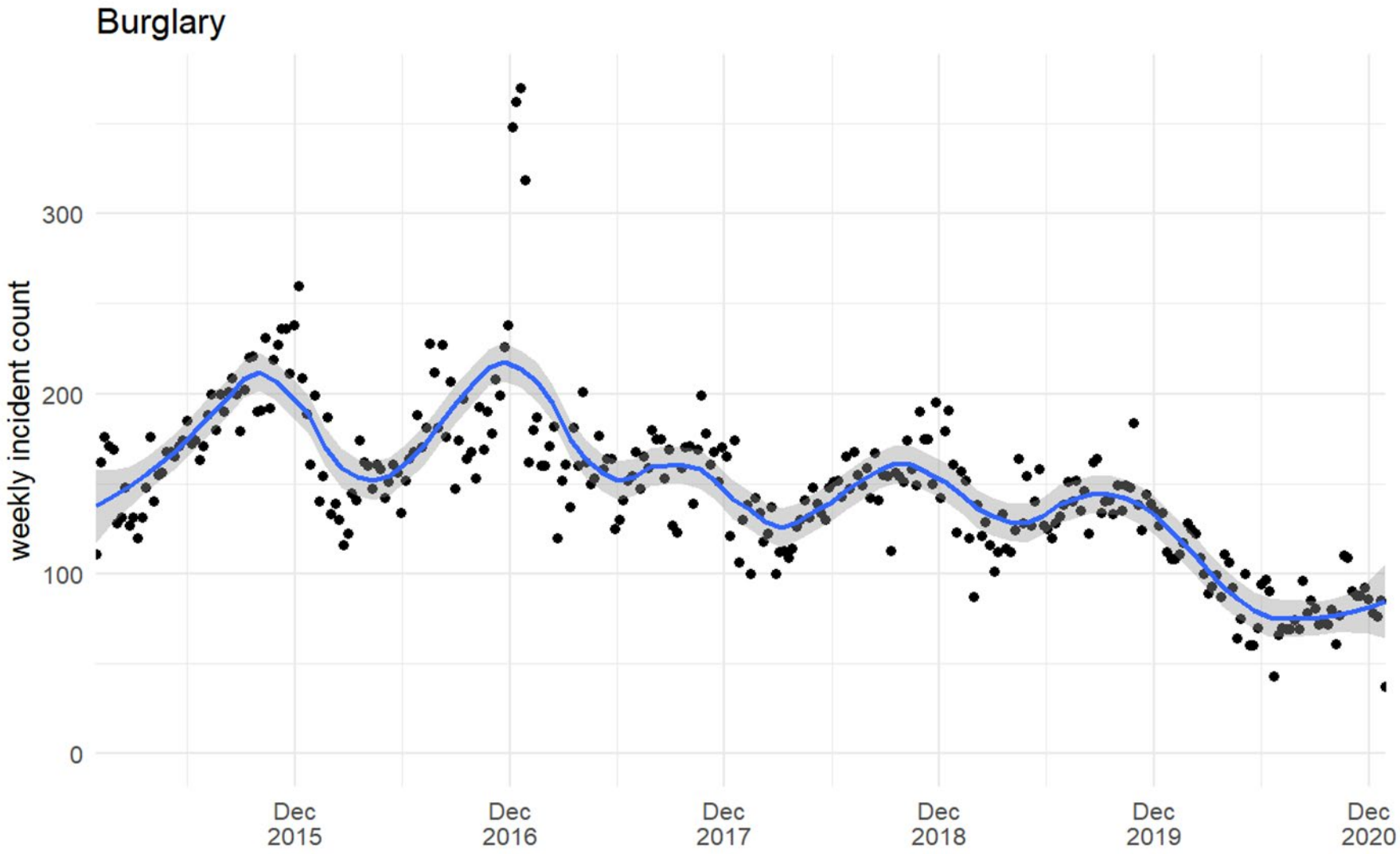
- Additive = the increasing or decreasing pattern of the time series is similar throughout the series.
  - ***Y = trend + cyclic + seasonality + noise (additive model)***

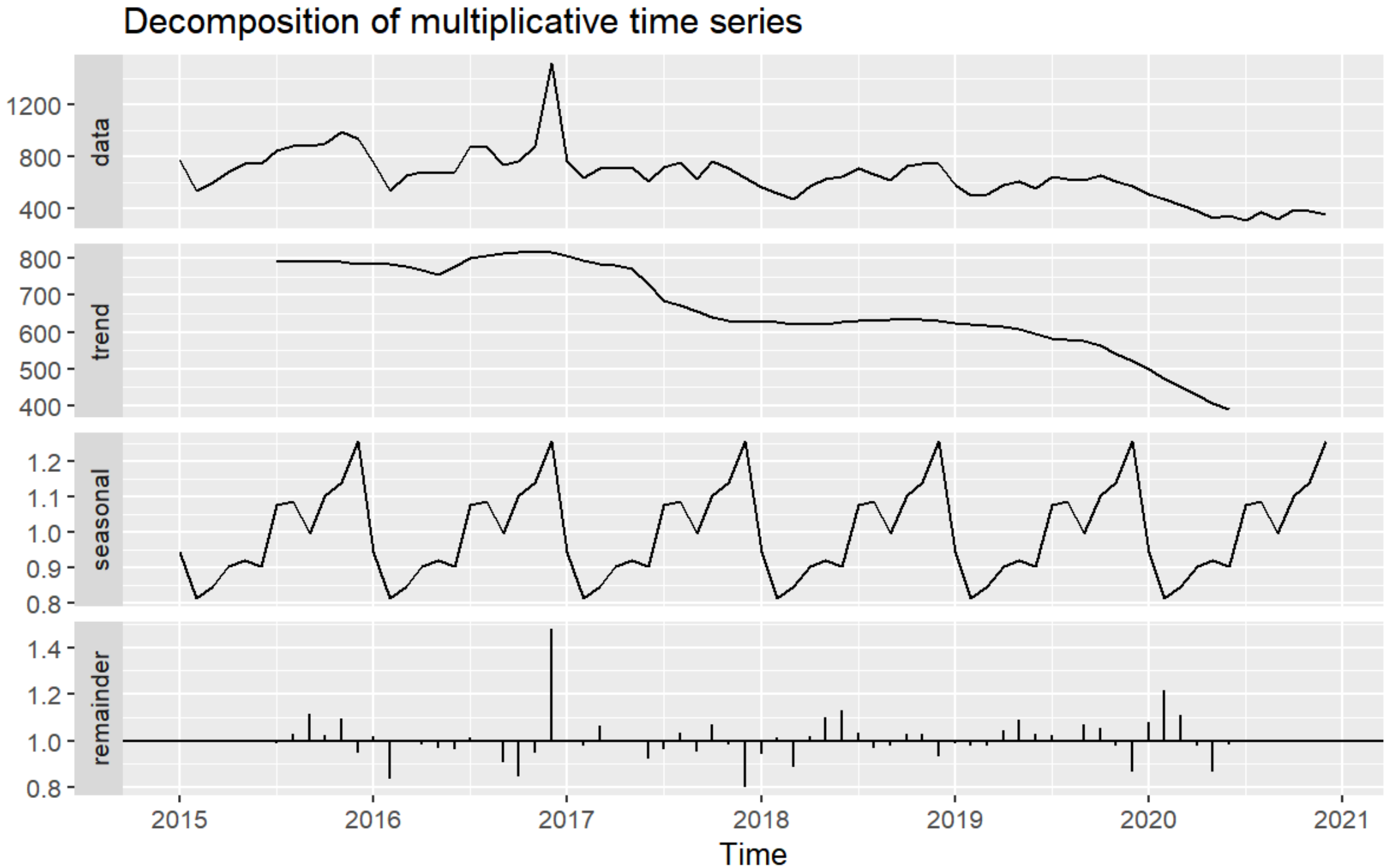
- Multiplicative = if the time series is having exponential growth or decrement with time
  - ***Y = trend *cyclic * seasonality * noise (multiplicative model)***

Additive vs Multiplicative Plot



What about with our example?




Decomposition of multiplicative time series
data
trend
seasonal
remainder
1200
800
400
700
600
500
1.2
1.1
1.0
0.9
0.8
1.4
2015
2016
2017
2018
2019
2020
2021
Time

Check for Stationarity

Stationarity = if its statistical properties such as mean, variance & covariance remain constant over time

We can examine this both;

- Visually (the decomposition plot)
- Statistically (the kppss, dickey-fuller test)

Make series stationary via *differencing.*

The Three Main TS Models

- ***Moving Averages***: useful for identifying trend and trend cycles
- ***Single Exponential Smoothing:*** used for time series with data with no trend or seasonality
- ***ARIMA/SARIMA Models:*** suitable for multivariate non-stationary data

More on S/ARIMA models

- Seasonal / Autoregressive Integrated Moving Average

- Used to predict future trends for time series
- ***"It is a form of regression analysis that evaluates the strength of the dependent variable relative to other changing variable****s"*

More on ARIMA/SARIMA models

- *Seasonal **(S)** / Autoregressive **(AR)** Integrated **(I)** Moving average (**MA)***

- ARIMA is characterised by three trend element; (p,d,q)
  - AR [p] = trend order (number of lagged observation)
  - I [d] = trend differencing (to make the series stationary)
  - MA [q] = trend order of moving average model

ARIMA(p,d,q)

- SARIMA is characterised by an additional four elements; (P,D,Q,m)
  - P = seasonal autoregressive order
  - D = seasonal differencing
  - Q = seasonal moving average
  - m = the number of time steps for a single period

SARIMA(p,d,q)(P,D,Q)m

How do you choose the values for p,d,q and P,D,Q?

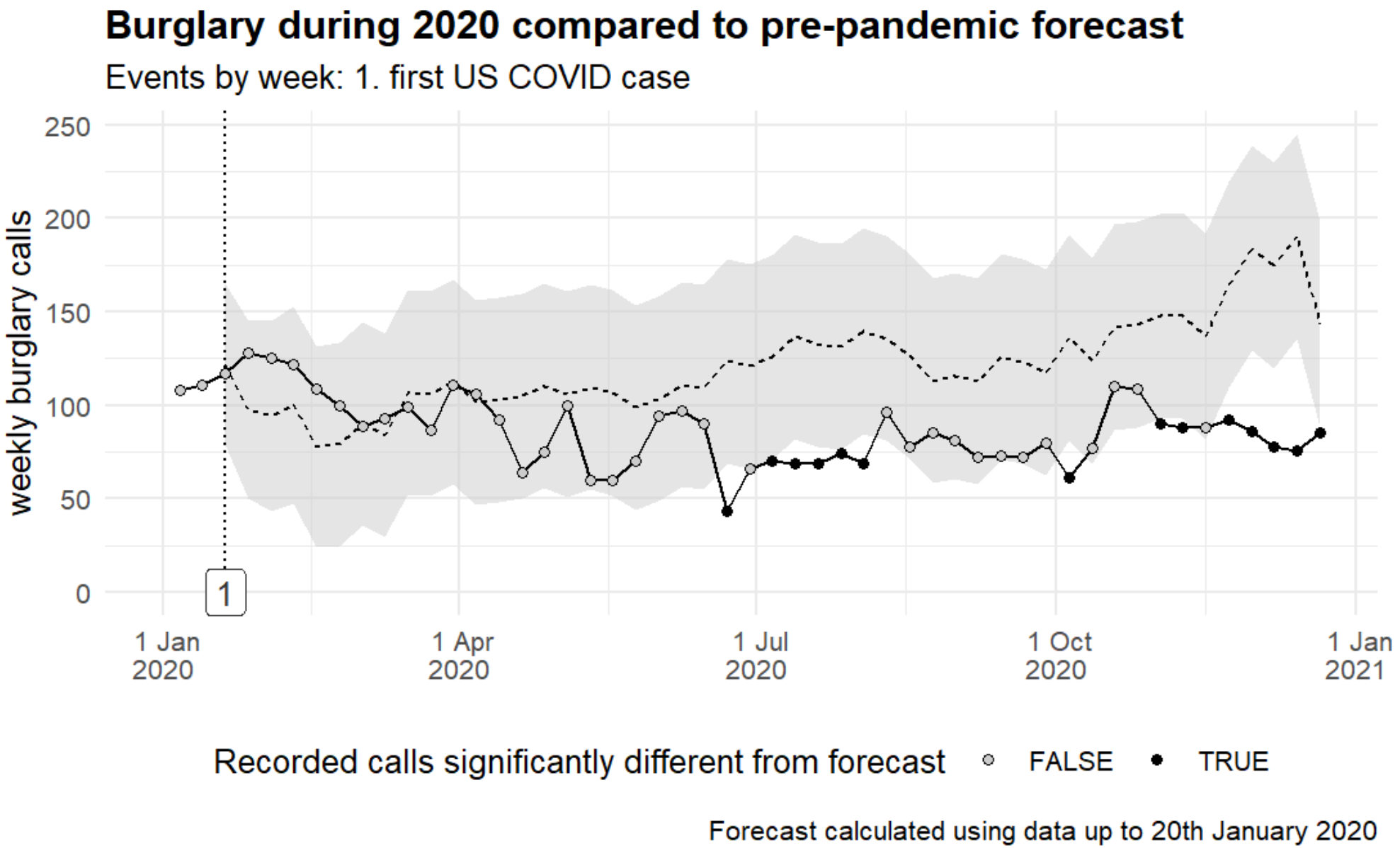
- The autocorrelation function (ACF)
- The partial autocorrelation function (PACF)

Steps to our second aim;

B) “Explore how the trend in 2020 would have looked in the absence of the pandemic”

- 1. Count the weekly crime
- 2. Model the weekly calls (e.g, using the function ‘ARIMA’ from the “fable” package in R)
- 3. Generate the forecast (e.g, using the ‘forecast’ function from the ‘forecast’ package in R)
- 4. Plot the forecast

S/ARIMA(2,0,0)(1,0,0)[52]



Software

- Python
  - Libraries: pandas. statsmodel, scikit-learn
- R
  - Packages: fable, forecast, tseries
  - Functions: ARIMA, forecast, ts, auto.arima
- FBProphet
  - https://facebook.github.io/prophet/

Any Questions?

Contact Details:

Email: Nadia.kennar@manchester.ac.uk
Twitter: @NadiaKennar

Resources:

- *Matt Ashby: The initial evidence on the relationship between the coronavirus pandemic and crime in the U.S https://link.springer.com/content/pdf/10.1186/s40163-020-00117-6.pdf*
- *Git-Hub for the code Code: https://github.com/mpjashby/covid19-crime*
- *Data found in the R package 'crimedata'*
  Citation; Ashby, M P J. (2018, August 12). Studying crime and place with the Crime Open Database. doi: http://doi.org/10.31235/osf.io/9y7qz

Economic and Social Research Council