Forecasting Chickenpox Cases in Hungary

Group 7

Vidhath Raghavan, Kate Fang, Prasanth Chinta

Part 1: Exploratory Data Analysis

Dataset Overview

Dataset description:

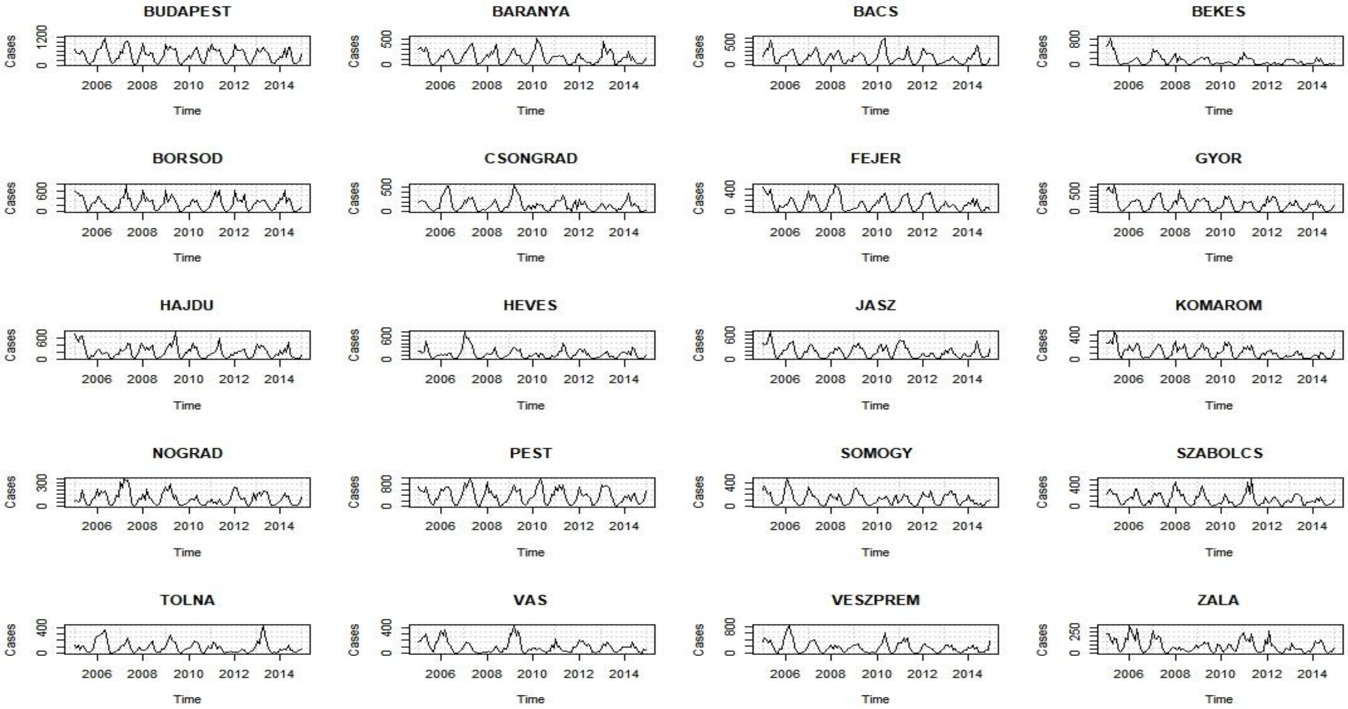
- This is a spatio-temporal dataset of weekly chickenpox cases from Hungary. The dataset consists of a time series of the county-level reported cases between 2005 and 2015. Attributes are weekly counts of chickenpox cases in Hungarian counties
- 522 records across 20 counties
- No missing data

Problem statement:

- Create a model that accurately predicts future chickenpox cases at the country-level and county-level of Hungary

Time Series Features

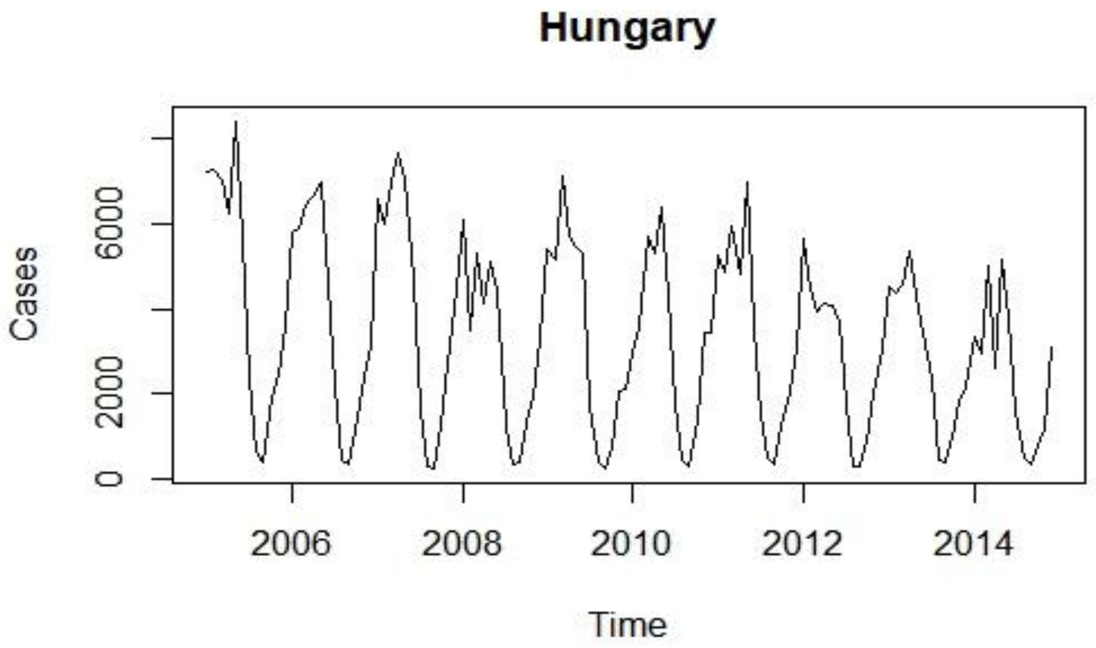
- ➢ Seasonality and cyclic (county level)



The chickenpox cases time series of the 20 counties show no trend but clear seasonality and cyclic.

Time Series Features

➢ Seasonality and cyclic (country level)

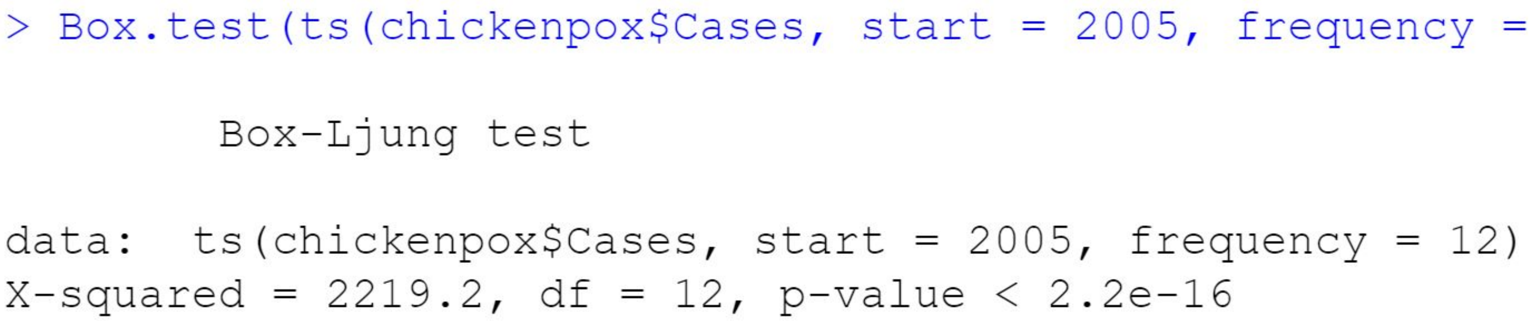


Considering Hungary is a small country with a population of only 9.71 million people (2021) and that all the counties contained the same time series features, we considered country level chickenpox cases by taking the sum of county level cases for time series analysis

Time Series Features

➢ Examination of white noise

Ljung-Box Test



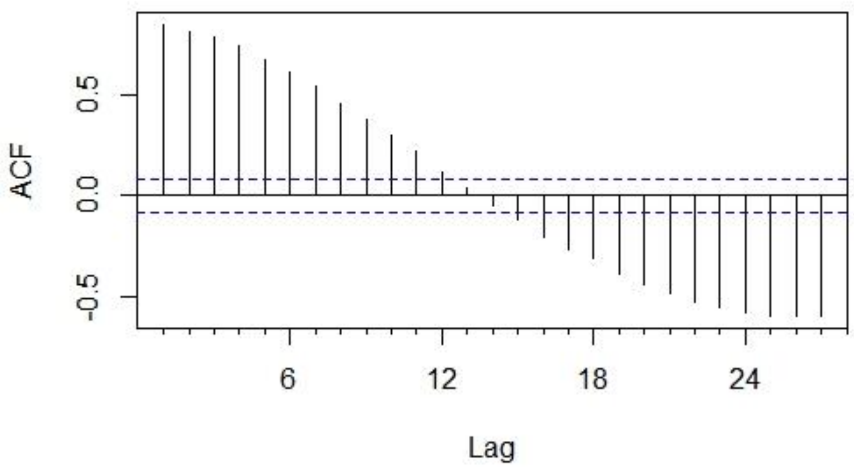
```
> Box.test(ts(chickenpox$Cases, start = 2005, frequency =

        Box-Ljung test

data:  ts(chickenpox$Cases, start = 2005, frequency = 12)
X-squared = 2219.2, df = 12, p-value < 2.2e-16
```

The noted p-value from the Ljung-Box test is smaller than the significance level, so we can reject the null hypothesis that the first 12 lags of autocorrelations equals to zero with 5% level. Therefore, the time series likely does not behave like white noises.

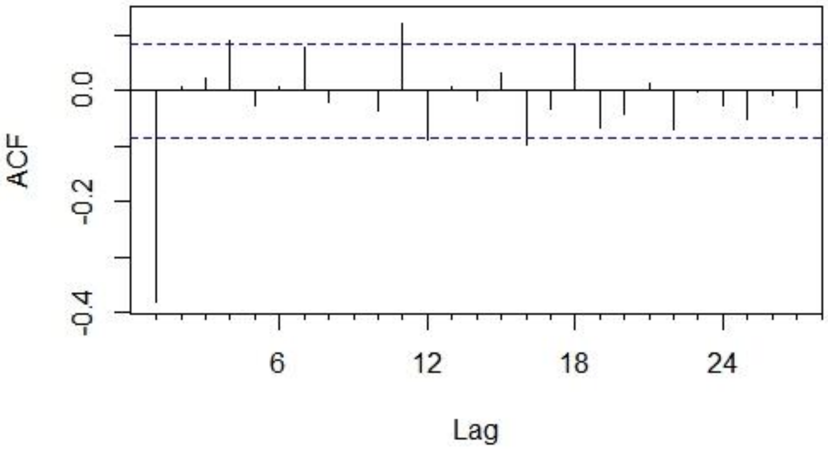
Time Series Features

➢ Examination of stationary

Original time series ACF



1st order differencing time series ACF

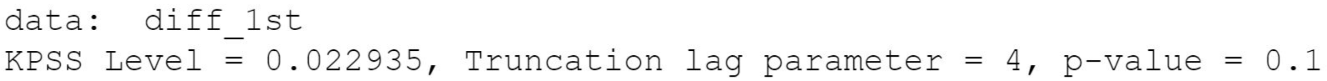


The ACF graphs show that the original time series is non-stationary, we need transform it to the first order differencing time series

Time Series Features

➢ Examination of stationary

KPSS Test



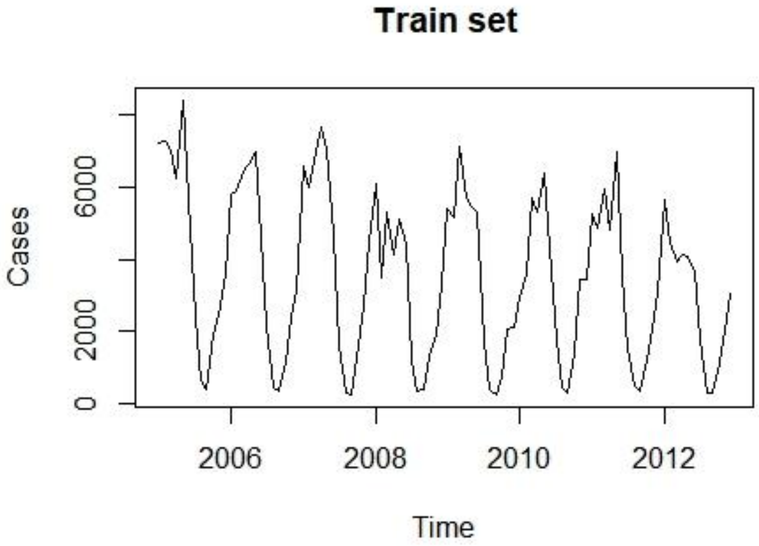
```
data:  diff_1st
KPSS Level = 0.022935, Truncation lag parameter = 4, p-value = 0.1
```

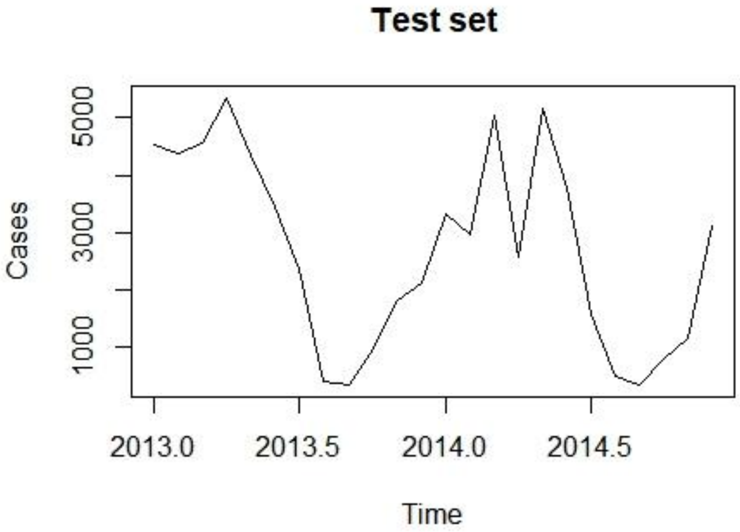
The p-value of first order differencing is 0.1 (which is greater than 0.05). Therefore, we can reject the null hypothesis of the KPSS test and assume the time series is stationary.

Part 2: Methodology

Train/Test Split

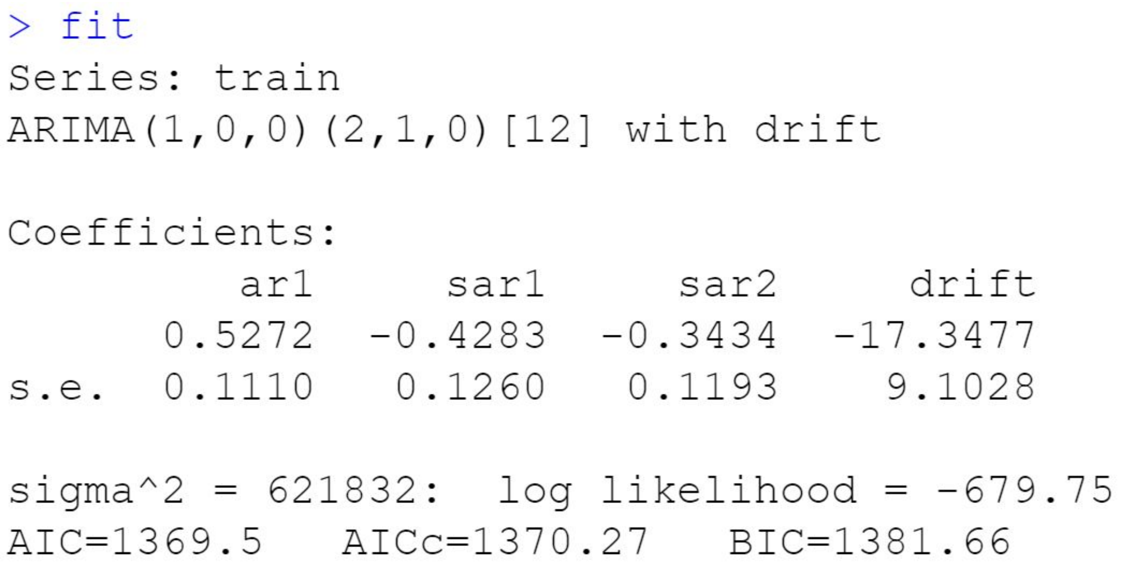
- ➢ We used window() function to split our dataset into train set and test set according to 80:20 principle.





Model 1: ARIMA

➢ Model Parameters

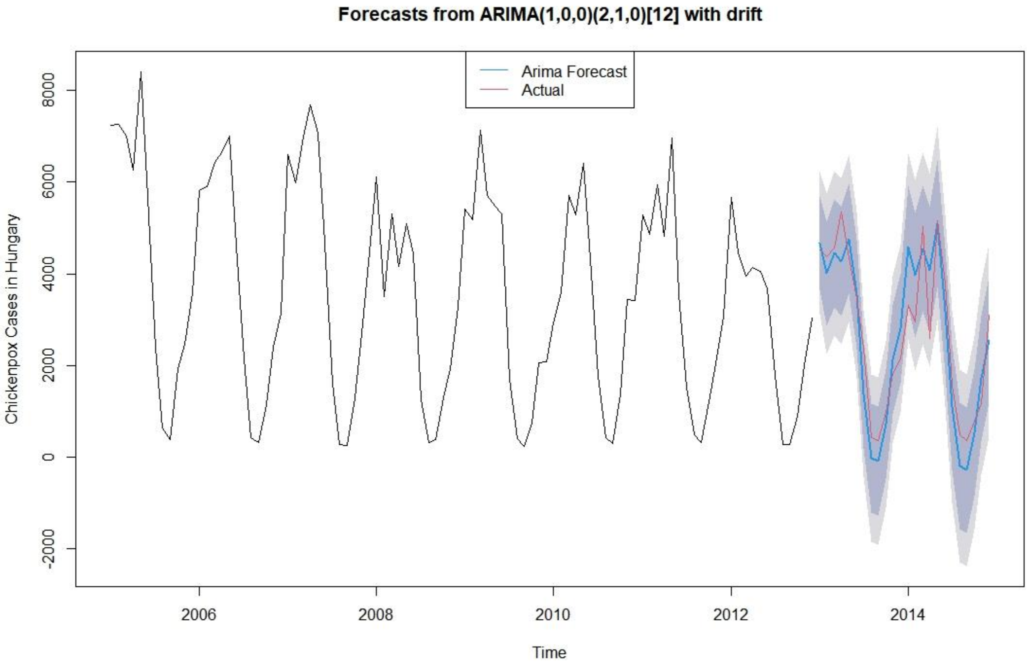


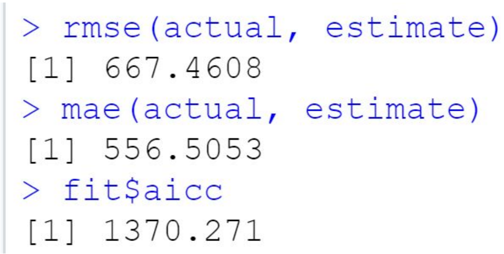
```
> fit
Series: train
ARIMA(1,0,0)(2,1,0)[12] with drift

Coefficients:
         ar1      sar1      sar2     drift
      0.5272   -0.4283   -0.3434  -17.3477
s.e.  0.1110    0.1260    0.1193    9.1028

sigma^2 = 621832:  log likelihood = -679.75
AIC=1369.5   AICc=1370.27   BIC=1381.66
```

After training, our best model is ARIMA(1,0,0)(2,1,0)[12], there also exist drift which indicate there is a change in level in the time series.

Model 1: ARIMA

➢ Model Performance

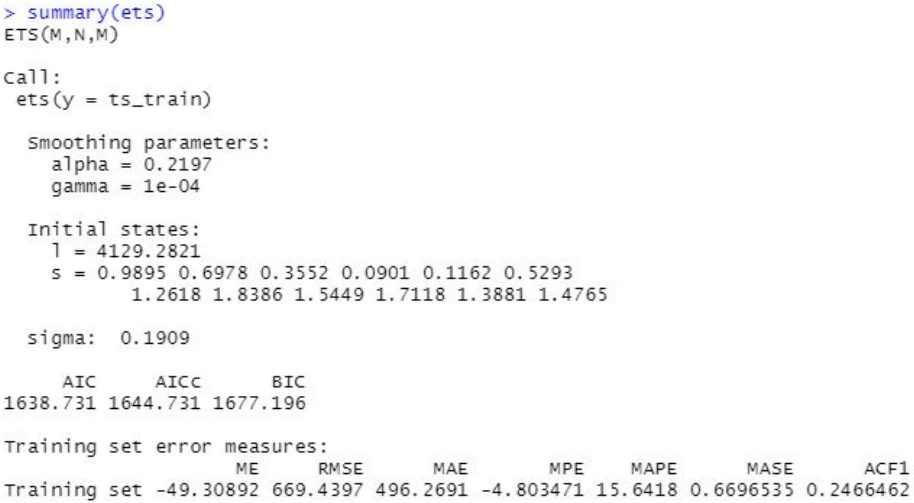



```
> rmse(actual, estimate)
[1] 667.4608
> mae(actual, estimate)
[1] 556.5053
> fit$aicc
[1] 1370.271
```

Model 2: Exponential Smoothing

Summary Statistics

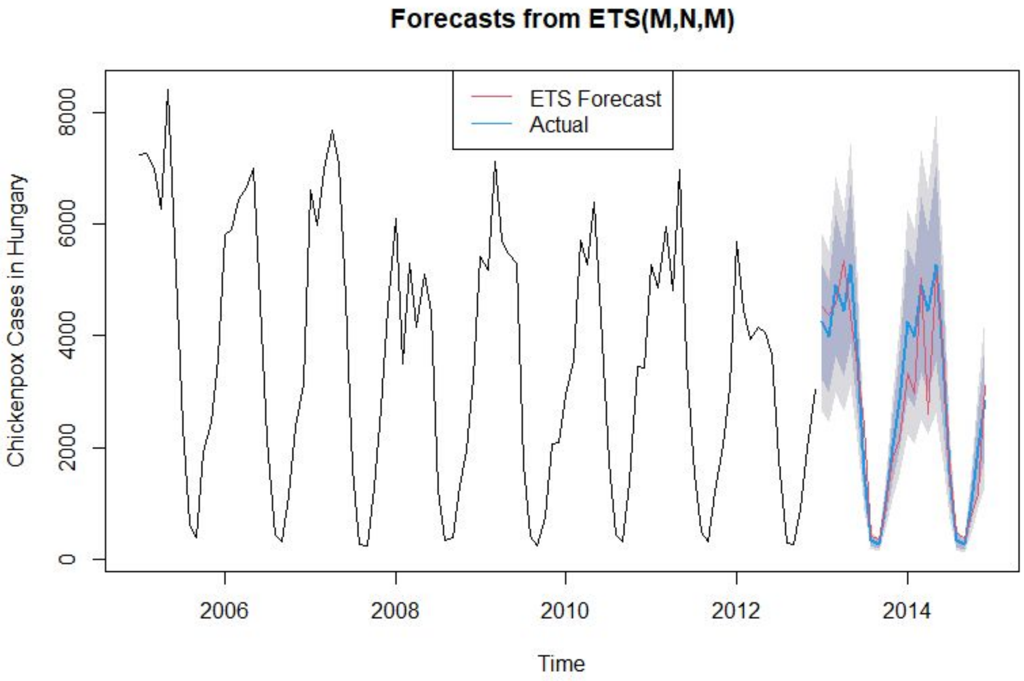
- The ets model was trained on data from 2005 - 2012
- The Exponential Smoothing model is (M,N,M) as:
    - 1. A multiplicative model is applied to error because the variability of the error decreases over time
    - 2. There's no trend
    - 3. A multiplicative model is applied to seasonality because the seasonality changes proportional to the level

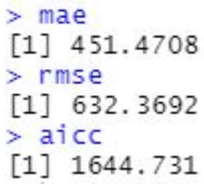


```
> summary(ets)
ETS(M,N,M)

Call:
 ets(y = ts_train)

  Smoothing parameters:
    alpha = 0.2197
    gamma = 1e-04

  Initial states:
    l = 4129.2821
    s = 0.9895 0.6978 0.3552 0.0901 0.1162 0.5293
           1.2618 1.8386 1.5449 1.7118 1.3881 1.4765

  sigma:  0.1909

     AIC     AICc      BIC
1638.731 1644.731 1677.196

Training set error measures:
                     ME     RMSE      MAE       MPE    MAPE      MASE      ACF1
Training set -49.30892 669.4397 496.2691 -4.803471 15.6418 0.6696535 0.2466462
```

Model 2: Exponential Smoothing

Model Evaluation





```
> mae
[1] 451.4708
> rmse
[1] 632.3692
> aicc
[1] 1644.731
```

RMSE is chosen over AICc:

- Goal is to minimize error
- Not high risk of overfitting due to the relatively low complexity of the data

Model 3: Primary Assumptions

- Assume one analyst has country-level data only, while another has county-level (granular) data that can be aggregated to region & country levels. Now, which analyst can provide better forecasts and help the Hungarian government

- The assumption is that the country-level forecast would be better when we add-up the forecasts at the granular level

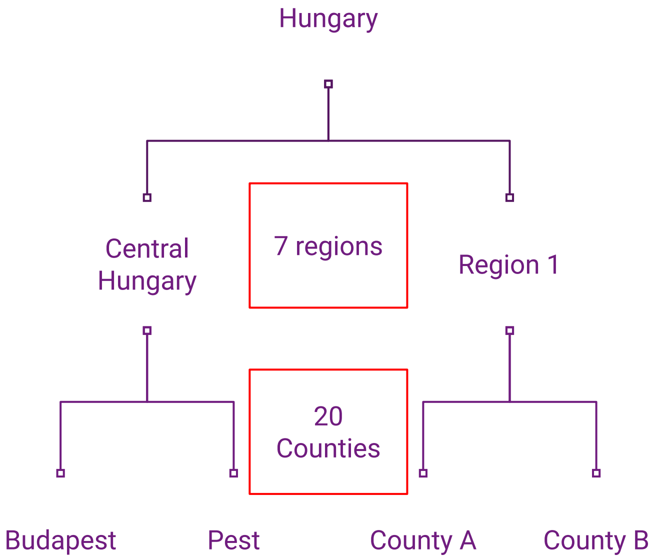
- Use Fable package to identify hierarchy and implement the hierarchical modeling

Model 3: Hierarchical Time Series

- Hungary Chickenpox data is a prime example of Hierarchical Time Series (HTS)

- At the granular level we have the *county level* chickenpox cases count that rolls-up to region-level, and eventually, to country-level

- The challenge is that we require forecasts to add up in a manner that is consistent with the aggregation structure of the hierarchy or group that defines the collection of time series (needs to be *coherent*)



Model 3: Fable package and Reconcile()

- Chapter 11 of Forecasting: Principles and Practice has details on the hierarchical time series, and forecasting reconciliation, and related Fable package

- HTS reconciliation is a process of readjusting the forecasts yielded by independent models on a set of hierarchically-linked time series, in order to improve the forecasts and ensure that they are coherent and sum up correctly

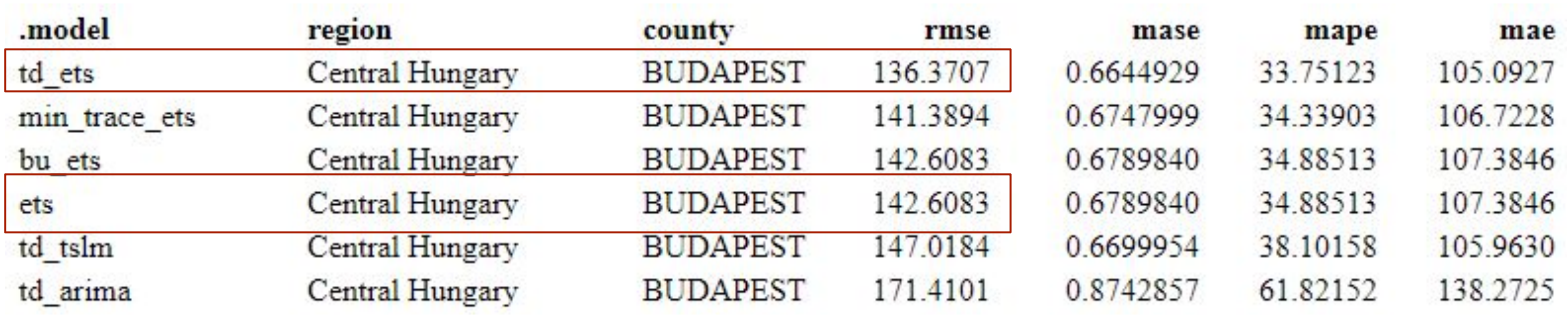
- Process involves -

```
data %>% aggregate_key() %>% model() %>% reconcile() %>% forecast()
```

Model 3: Reconciliation methodologies

- Four common methods of generating coherent forecasts-
  - **Bottom-up**: involves first generating forecasts for each series at the bottom level, and then summing these to produce forecasts for all the series in the structure
  - **Top-down**: involve first generating forecasts for the country-level series, and then disaggregating these down the hierarchy
  - **Middle-out**: combines bottom-up and top-down approaches. Again, it can only be used for strictly hierarchical aggregation structures.
  - **MinT (Minimum Trace)**: the optimal reconciliation forecasts are generated using all the information available within a hierarchical or a grouped structure.

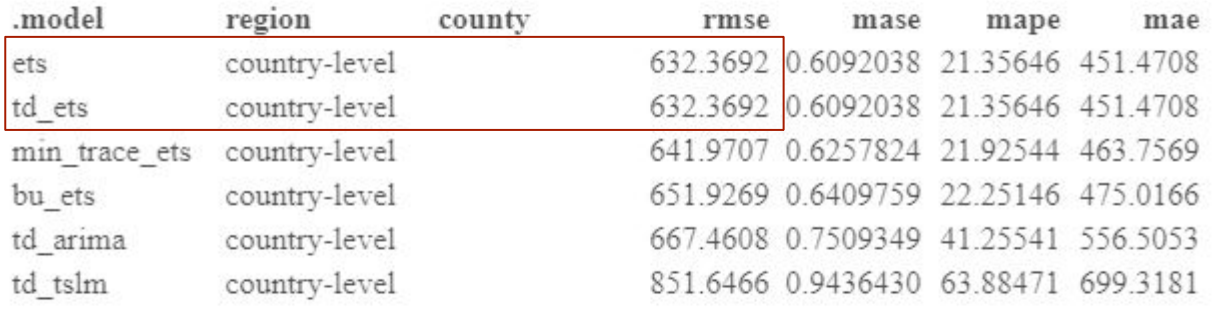
Model 3: Model Evaluation



.model	region	county	rmse	mase	mape	mae
td_ets	Central Hungary	BUDAPEST	136.3707	0.6644929	33.75123	105.0927
min_trace_ets	Central Hungary	BUDAPEST	141.3894	0.6747999	34.33903	106.7228
bu_ets	Central Hungary	BUDAPEST	142.6083	0.6789840	34.88513	107.3846
ets	Central Hungary	BUDAPEST	142.6083	0.6789840	34.88513	107.3846
td_tslm	Central Hungary	BUDAPEST	147.0184	0.6699954	38.10158	105.9630
td_arima	Central Hungary	BUDAPEST	171.4101	0.8742857	61.82152	138.2725

- As you can see that county-level forecast from the top-down reconciliation method is better than standard ets (ets_m) on the same county. In fact all the reconciliation outcomes have better results.

Model 3: Model Evaluation



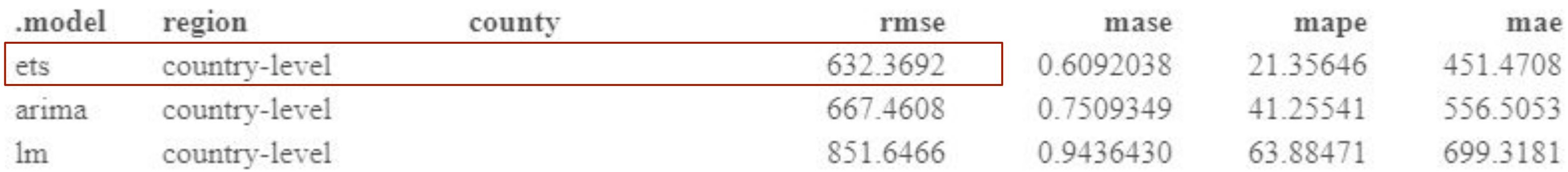
.model	region	county	rmse	mase	mape	mae
ets	country-level		632.3692	0.6092038	21.35646	451.4708
td_ets	country-level		632.3692	0.6092038	21.35646	451.4708
min_trace_ets	country-level		641.9707	0.6257824	21.92544	463.7569
bu_ets	country-level		651.9269	0.6409759	22.25146	475.0166
td_arima	country-level		667.4608	0.7509349	41.25541	556.5053
td_tslm	country-level		851.6466	0.9436430	63.88471	699.3181

- Surprisingly, against our assumption, at the total level, the country-level model and the related reconciliation approach have same accuracy metrics

- This shows that the reconcile() function works for getting better granular level forecasting, disaggregating from top to bottom levels, not the other way around

Part 3: Summary

Final Outcome



.model	region	county	rmse	mase	mape	mae
ets	country-level		632.3692	0.6092038	21.35646	451.4708
arima	country-level		667.4608	0.7509349	41.25541	556.5053
lm	country-level		851.6466	0.9436430	63.88471	699.3181

- As shown, exponential smoothing resulted in the best country-level Chickenpox prediction
- In a county-level approach, top-down reconciliation method is better than the standard ets forecasting

Future Work

Improving

- VARIMAX model:
  - More parameters:
    - Birth rate data
    - Vaccination data
- Optimizing Smoothing Parameters
- Recurrent Neural Network + LSTMs
- STL decomposition

Extension

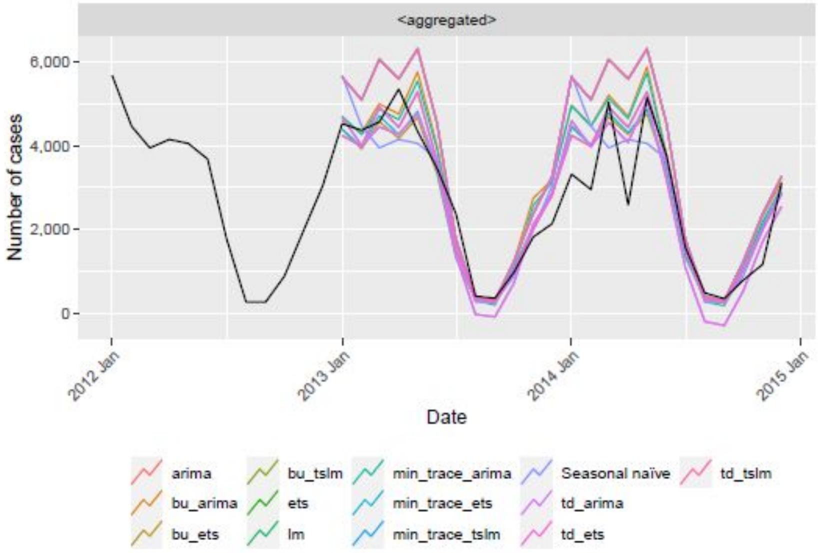
- Collect Chickenpox data during the pandemic to see how that affected the seasonal spread of Chickenpox

Questions

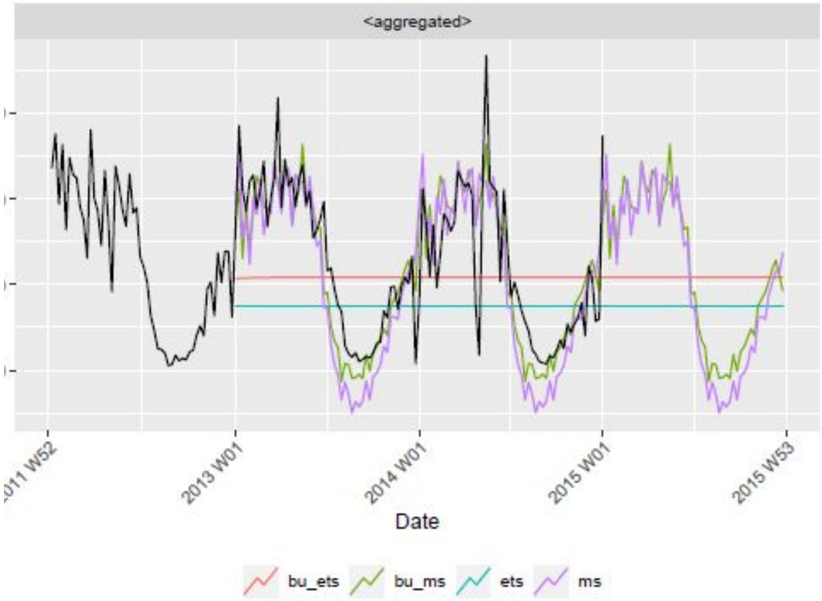
Part 4: Appendix

Forecast Plots - Country level

Monthly Data

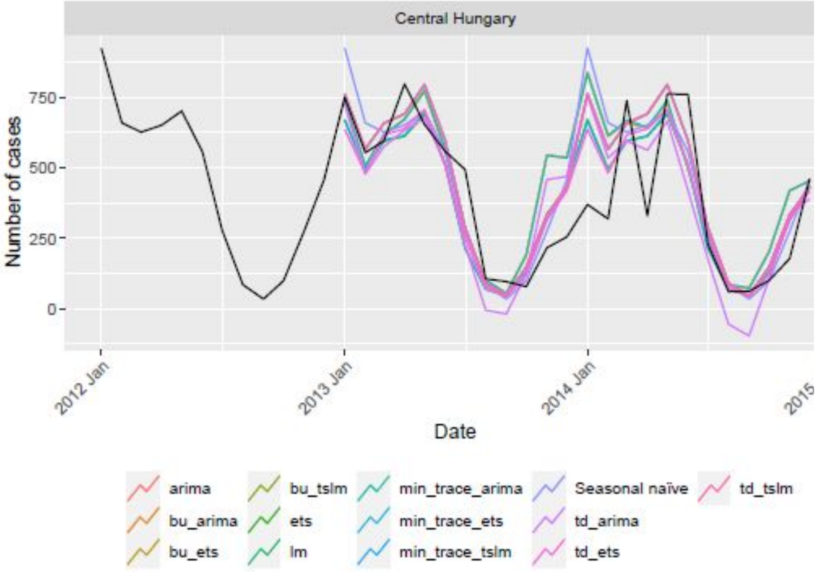


Weekly Data

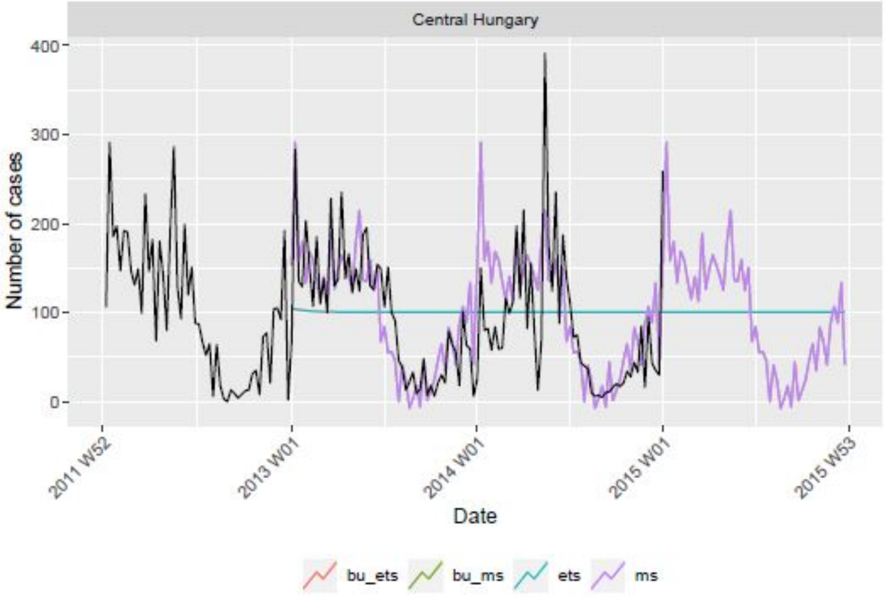


Forecast Plots - County level

Monthly Data



Weekly Data


Central Hungary
Number of cases
2011 W52
2013 W01
2014 W01
2015 W01
2015 W53
Date
bu_ets
bu_ms
ets
ms