FORESEA: AUTOMATIC FORECASTING TOOL USING SEASONAL FACTOR MODEL

PALAKORN BURANASAMPATANON

APRIL 10, 2020

OUTLINE

1. Observed pain points
2. Solution and its abilities
3. Forecasting performance with CPI Disaggregate Data
4. Q&A

OBSERVED PAIN POINTS

- What model should I use?
- If I have several variables to forecast, do I have to manually run the same process repetitively?
- How do I know if the model performs well both in sample and out of sample?
- Can I trust the results from my model?

MY SOLUTION WITH THE FOLLOWING ABILITIES

A program with easy-to-use features

- Only one line of code to forecast multiple variables
- Accept an input as just a normal Excel file

The program can automatically find a time step with a potential structural break

Better ability to fit more recent data using automatic out-of-sample backtesting

Optimized code for the greater computing efficiency

WHAT DOES THE PROGRAM DO BEHIND THE SCENE? (FOR EVERY TIME SERIES)

1. Ensure the input fits proper time series properties
2. Find the most possible structural break in the data
3. Go over all predetermined model specifications and methods of using structural break information
    1. Some methods: remove data prior to the break, create dummy variable after the break
    2. Fit each model and forecast new values during periods that the model has not observed yet
4. Use the model with the best out-of-sample accuracy to forecast for future values

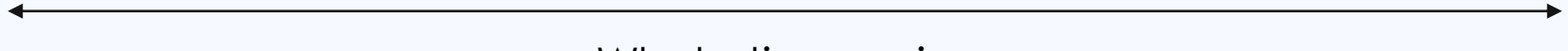
STRUCTURAL BREAK DETECTION AND USAGES

STRUCTURAL BREAK COMPONENT

- The program can detect a structural break time step and report the user
- So far, in literature, there does not exist any universally exact formal definition of structural break or how to detect it
  - Formal definitions rely on either model-based approach or data-characteristic approach
- Our chosen approach: data-characteristic approach
- For more comprehensive details: Troung et al. (2020)

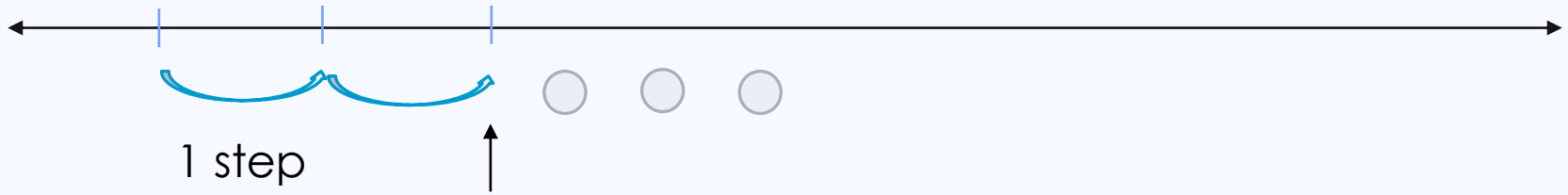
OUR STRUCTURAL BREAK DETECTION METHOD: BINARY SEGMENTATION

1) Input

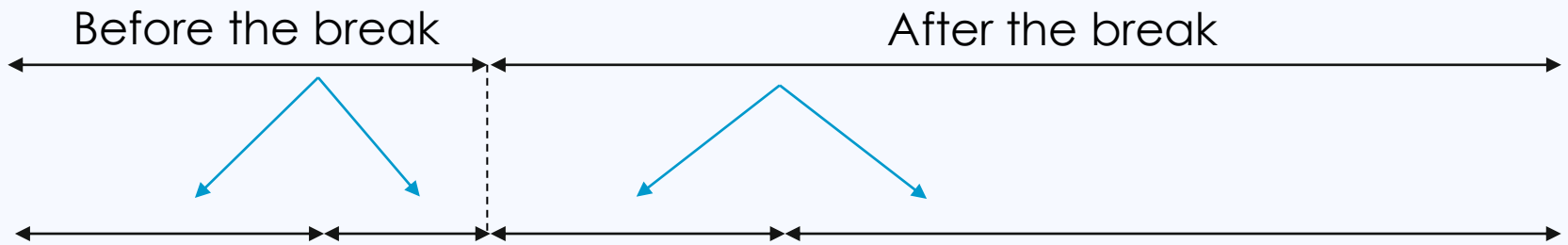


Whole time series

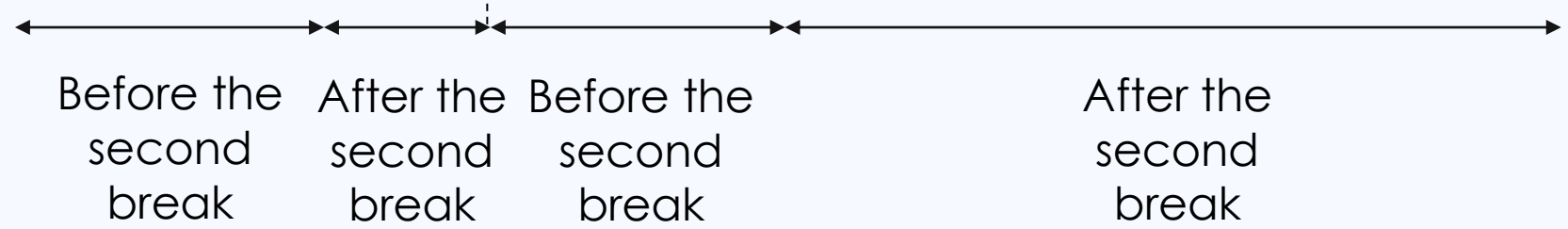
2) Find time step with the most 'optimal' break



3) Split the series into two



4) More splits? Do it again from each segment



HOW DO WE DEFINE THE 'OPTIMALITY'?

Let A – $(a_1, a_2, \dots, a_T)$ be a whole original series and B and C the series resulted from before and after the break respectively

Let our objective function be in a form of normal distribution

- We assume that B and C are generated from two different normal distributions
- With the idea of structural break, we assume that B and C are independent

Simplified objective function:

$$\underset{k \in \{1,2,\dots,T\}}{\operatorname{argmax}} \; P(B, C)$$

$$= \underset{k \in \{1,2,\dots,T\}}{\operatorname{argmax}} \; P(B)P(C) \; (independence)$$

$$= \underset{k \in \{1,2,\dots,T\}}{\operatorname{argmax}} \; P(a_1, \dots, a_k)P(a_{k+1}, \dots, a_T)$$

MORE ON STRUCTURAL BREAK OPTIMIZATION

$$\underset{k \in \{1,2,\dots,T\}}{\operatorname{argmax}} \; P(a_1, \dots, a_k) P(a_{k+1}, \dots, a_T)$$

$$= \underset{k \in \{1,2,\dots,T\}}{\operatorname{argmax}} \; \max_{\mu_{k_1}, \sigma_{k_1}, \mu_{k_2}, \sigma_{k_2}} \prod_{i=1}^{k} \frac{e^{-\frac{1}{2}\left(\frac{x_i - \mu_{k_1}}{\sigma_{k_1}}\right)^2}}{\sqrt{2\pi {\sigma_{k_1}}^2}} \prod_{j=k+1}^{T} \frac{e^{-\frac{1}{2}\left(\frac{x_j - \mu_{k_2}}{\sigma_{k_2}}\right)^2}}{\sqrt{2\pi {\sigma_{k_2}}^2}} \; (Normal\ distribution\ for\ B\ and\ C)$$

The optimization step can also be applied for other distributions and objective functions as well

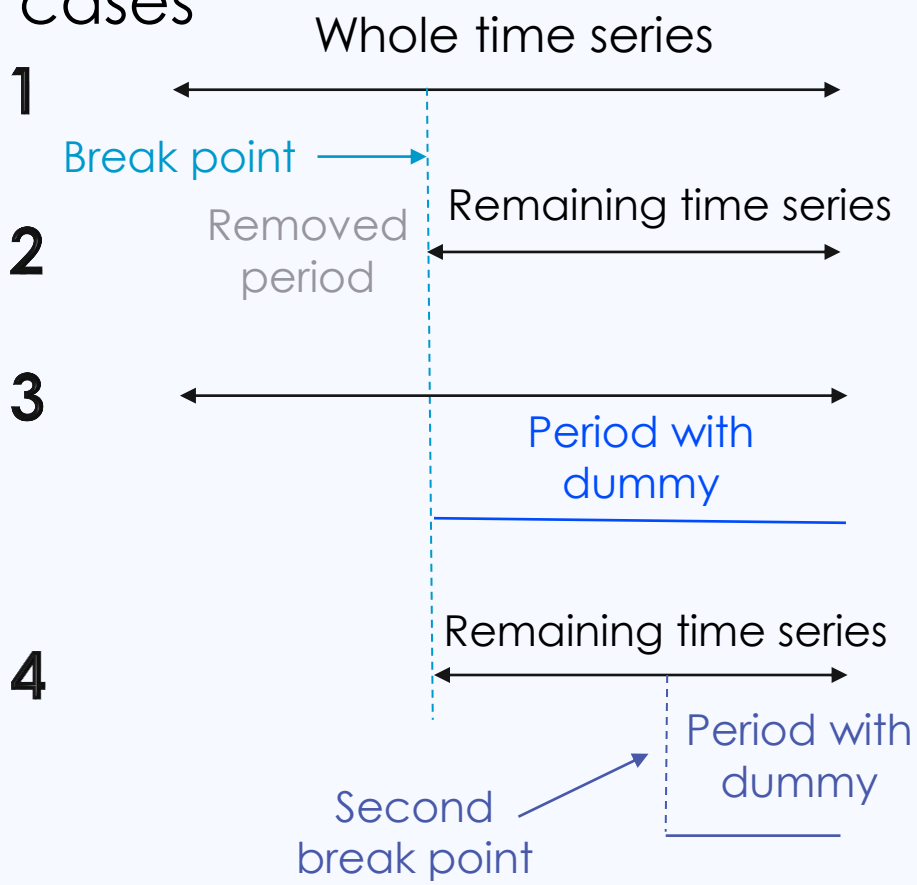
- Example: Laplace distribution, L2 norm (break in mean), L1 norm (break in median)

The user can also configure about the minimum number of observations – n - per each part of the series

USAGES OF STRUCTURAL BREAK INFORMATION

SARIMA, ARIMA – Run through the following cases

- Use only the original series
- Remove the prior-to-break data
- Keep the whole series while creating a dummy variable to indicate the post-break case
- Use the post-break data and find another break in mean from the remaining period



FORECASTING MODELS

THE CHOSEN MODEL: SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (SARIMA)

Model Structure (Monthly Case):

$$x_{t+1} = \underbrace{\rho_0 + \rho_1\theta_1 + \rho_2\theta_2}_{\text{Dummy and constant}} + \underbrace{\sum_{i=0}^{t_1}\alpha_i x_{t-i} + \sum_{j=0}^{t_2}\beta_j\varepsilon_{t-j}}_{\text{ARMA}} + \underbrace{\sum_{k=1}^{y_1}\gamma_k x_{t+1-12k} + \sum_{l=1}^{y_2}\delta_l\varepsilon_{t+1-12l}}_{\text{Seasonal ARMA}} + \varepsilon_{t+1};\ \varepsilon_{t+1}\sim\mathcal{N}(0,\sigma^2)$$

θ_1: First dummy variable with its value = 1 after the first break, and = 0 otherwise

- Appeared in the mode 3

θ_2: Second dummy variable with its value = 1 after the second break, and = 0 otherwise

- Appeared in the mode 4

Note: The program also accepts quarterly data with the terms -12k and -12l changed to -4k and -4l

BASELINE MODEL/METHOD

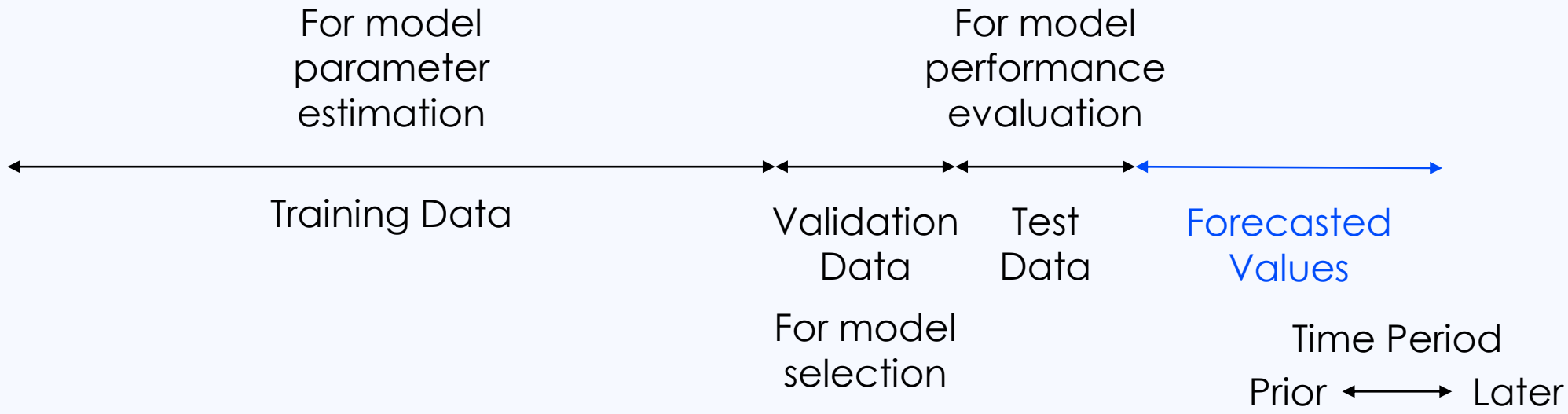
Autoregressive Integrated Moving Average (ARIMA)

- Included as the classic model for time-series forecasting

Historically average values

- Assumption: find the average values from the same month/quarter from the past n years
- Definition: $\widehat{y_t} = \frac{\sum_{i=1}^{n} y_{t-(i*f)}}{n}$
  - n: the number of historical periods taken for calculation
  - f: frequency of time period with f = 4 for quarterly data and f = 12 for monthly data

STANDARD PROCEDURE FOR DATASET SEPARATION: TIME SERIES CASE



Model selection is based on out-of-sample forecasting performance

- For this experiment, we use validation mean average error (MAE)
- Now, the program allows the user to use either MAE or RMSE

STRUCTURAL BREAK RESULT AND FORECASTING PERFORMANCE

FIRST PROTOTYPE: DISAGGREGATED CPI DATA

DATA INFORMATION

Topics	Details
MoM Series – 9 Major CPI Disaggregates	• Food in core • Raw food • Clothing • Entertainment • Housing • Health service • Tobacco • Transportation • Energy (Component of transportation)
Time period	2: 2000 - 10: 2019
Frequency	Monthly
Outlier removal	None

OUR FORECASTING MODEL SPECIFICATIONS

Training time steps: 02: 2000 – 11: 2016

Validation time steps: 12: 2016 – 5: 2018 (18 time steps)

Test time steps: 6: 2018 – 10: 2019 (18 time steps)

Model selection criterion: smallest validation mean average error (MAE)

Hyperparameter values					
SARIMA		ARIMA		Historical Average	
Terms	Number of lags	Terms	Number of lags	Terms	Number of lags
Autoregressive*	0, 1, 2, 3	Autoregressive*	0, 1, 2, 3	Number of years	5
Moving average*	0, 1, 2, 3	Moving average*	0, 1, 2, 3		
Seasonal autoregressive	1, 2				
Seasonal moving average	1, 2				

* The autoregressive and moving average terms must not be both zero

STRUCTURAL BREAK SPECIFICATIONS

Objective function:

- 1st break: normal distribution
- 2nd break: L2 norm (break in mean)

Number of break: 1,2, or None

Minimum number of observations per segment of series:

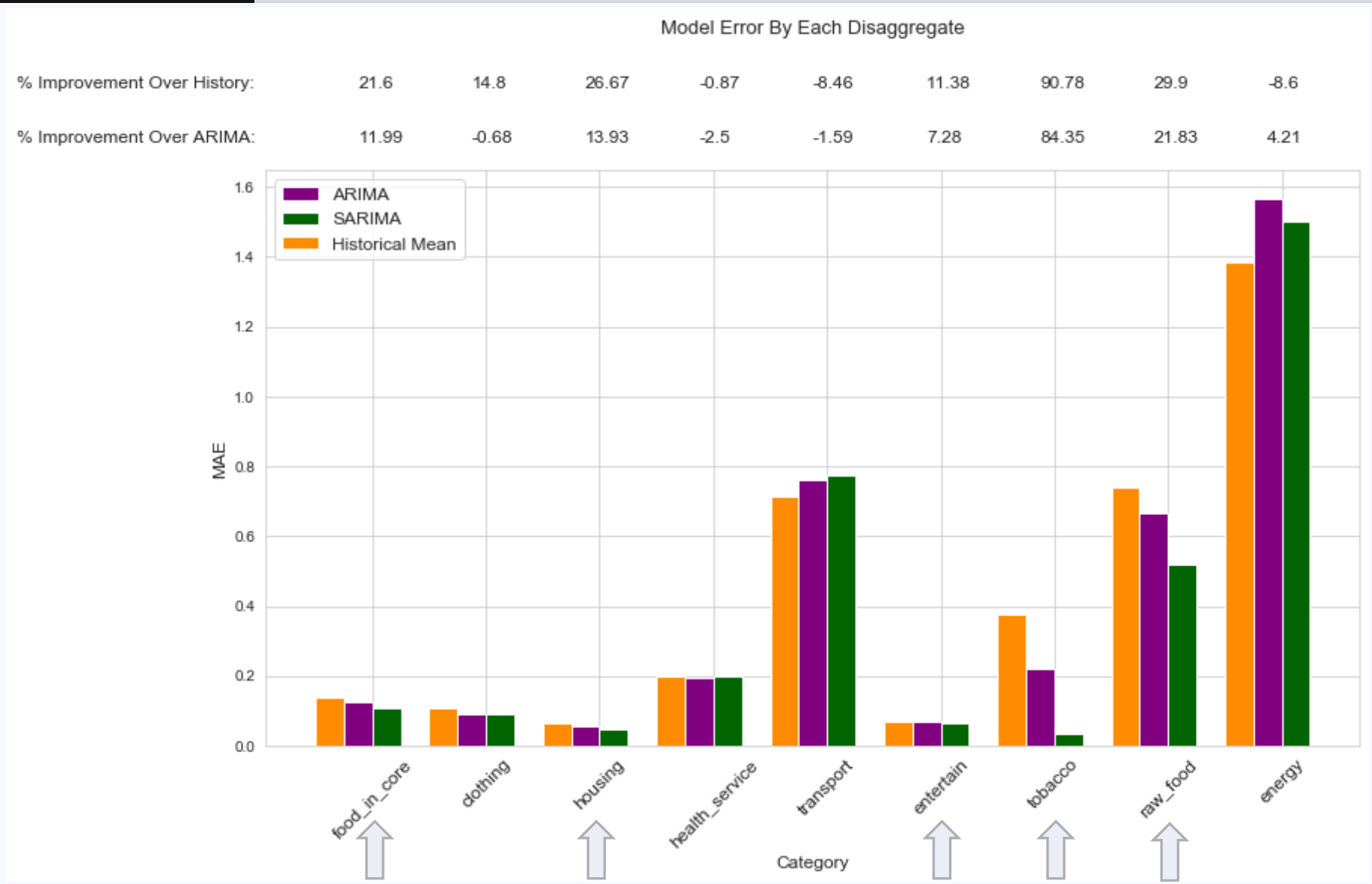
- First Break: 60
- Second Break: 36 (12 if the remaining series is shorter than 72 periods)
- Caution: the sample size has to be large enough at least to allow for adequate forecasting model fitting

STRUCTURAL BREAK RESULTS

Results from the structural break test		
Variable	First Break Period	Second Break Period
Health service	7: 2005	4: 2009
Food in core	3: 2007	8: 2012
Housing	6: 2008	10: 2011
Clothing	4: 2009	8: 2012
Tobacco	4: 2009	8: 2012
Transport	9: 2009	6: 2013
Entertainment	9: 2009	6: 2013
Energy	9: 2009	1: 2013
Raw food	5: 2011	6: 2013

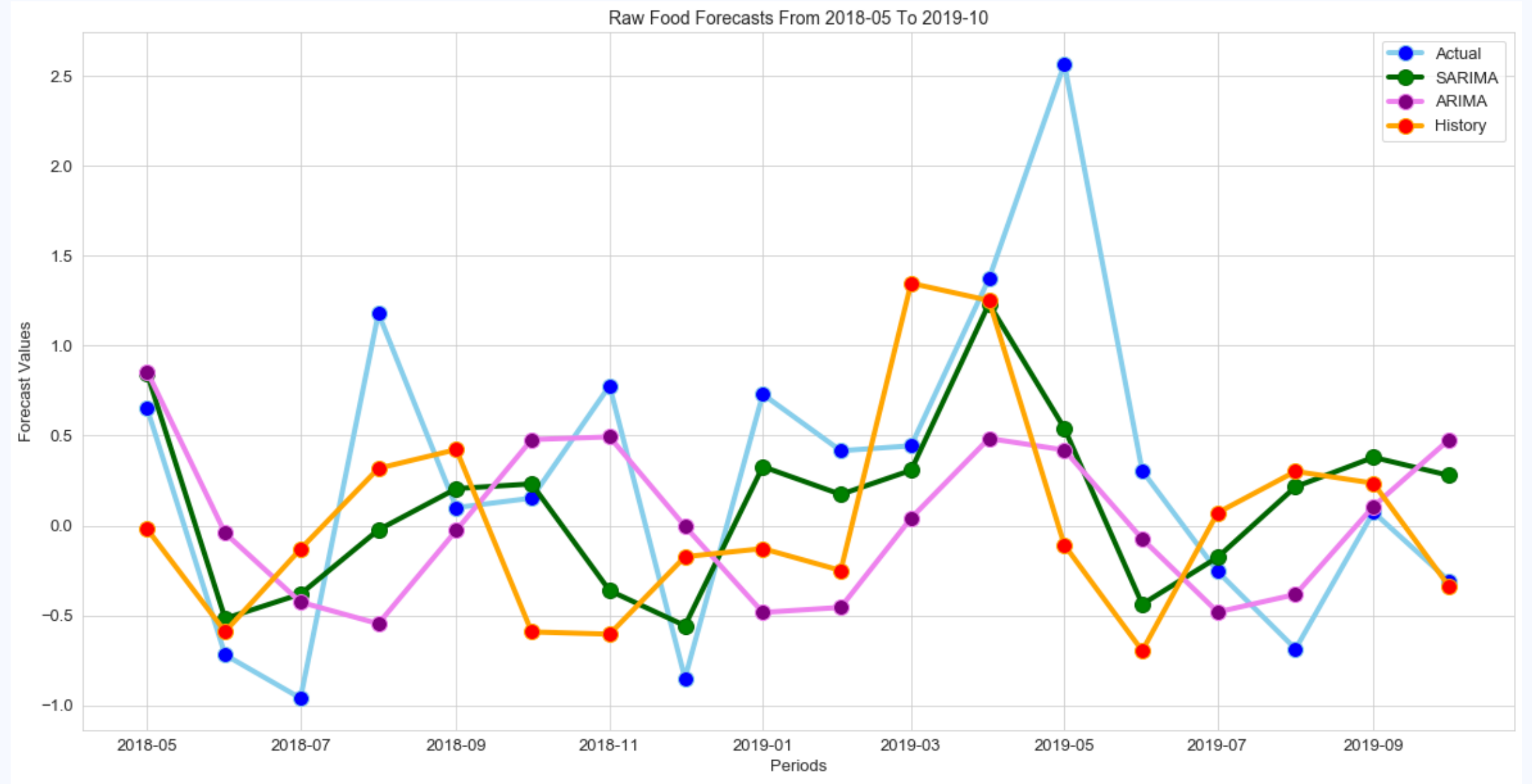
FORECASTING PERFORMANCE

FORECASTING MAE: TEST PERIOD

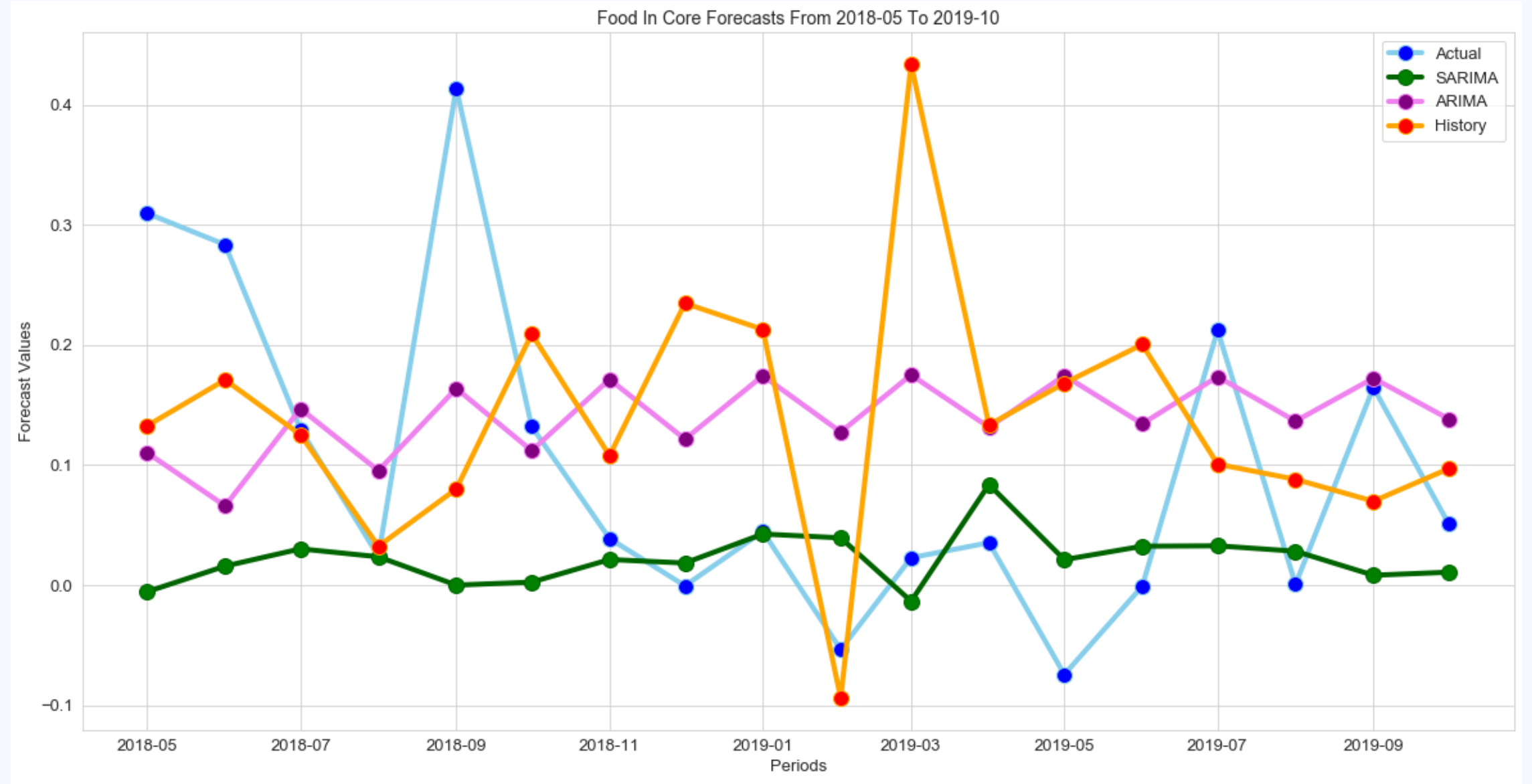


- Each variable with the arrow beneath it means SARIMA outperforms both historical mean and ARIMA

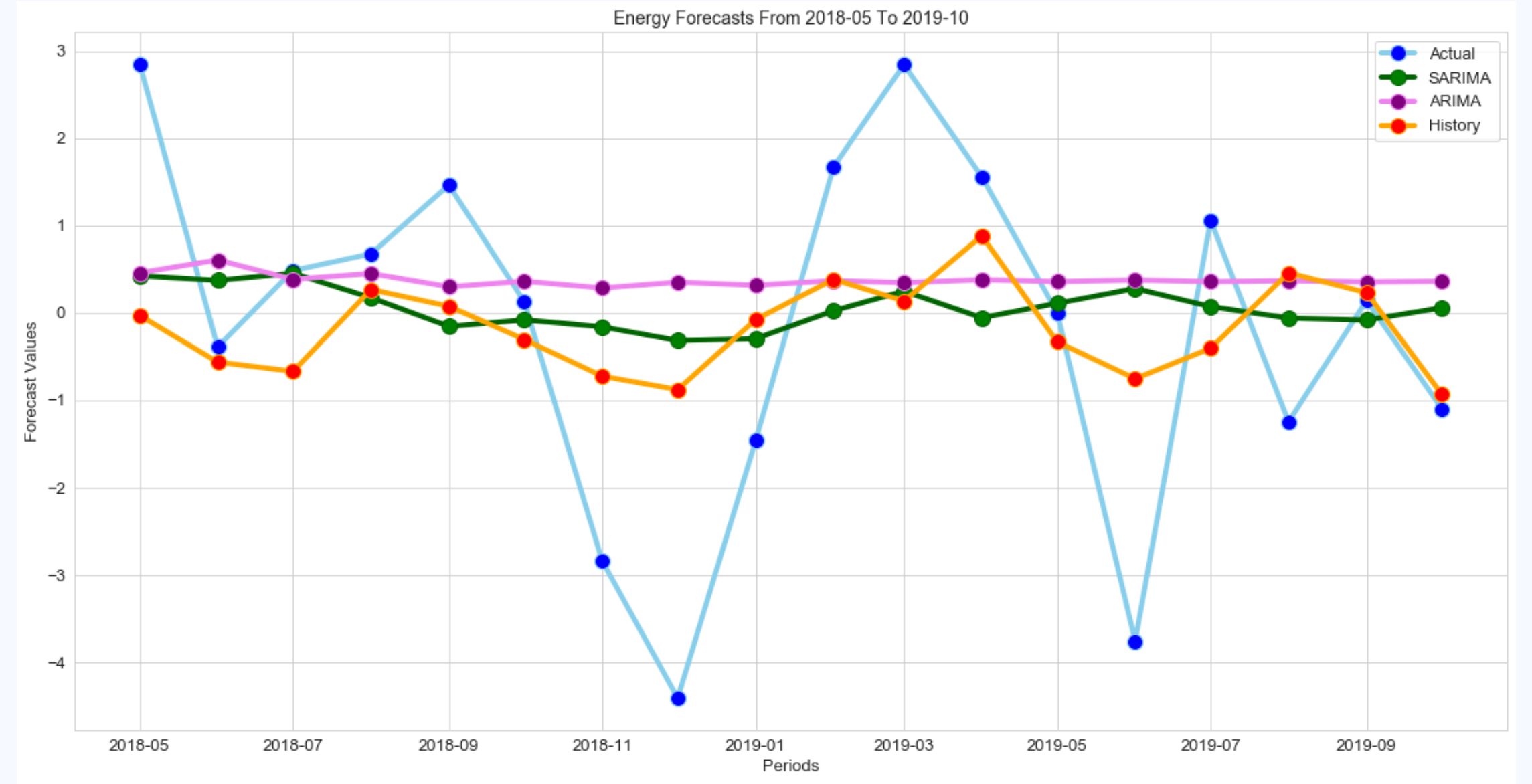
FORECASTING PLOTS: RAW FOOD, TEST PERIOD



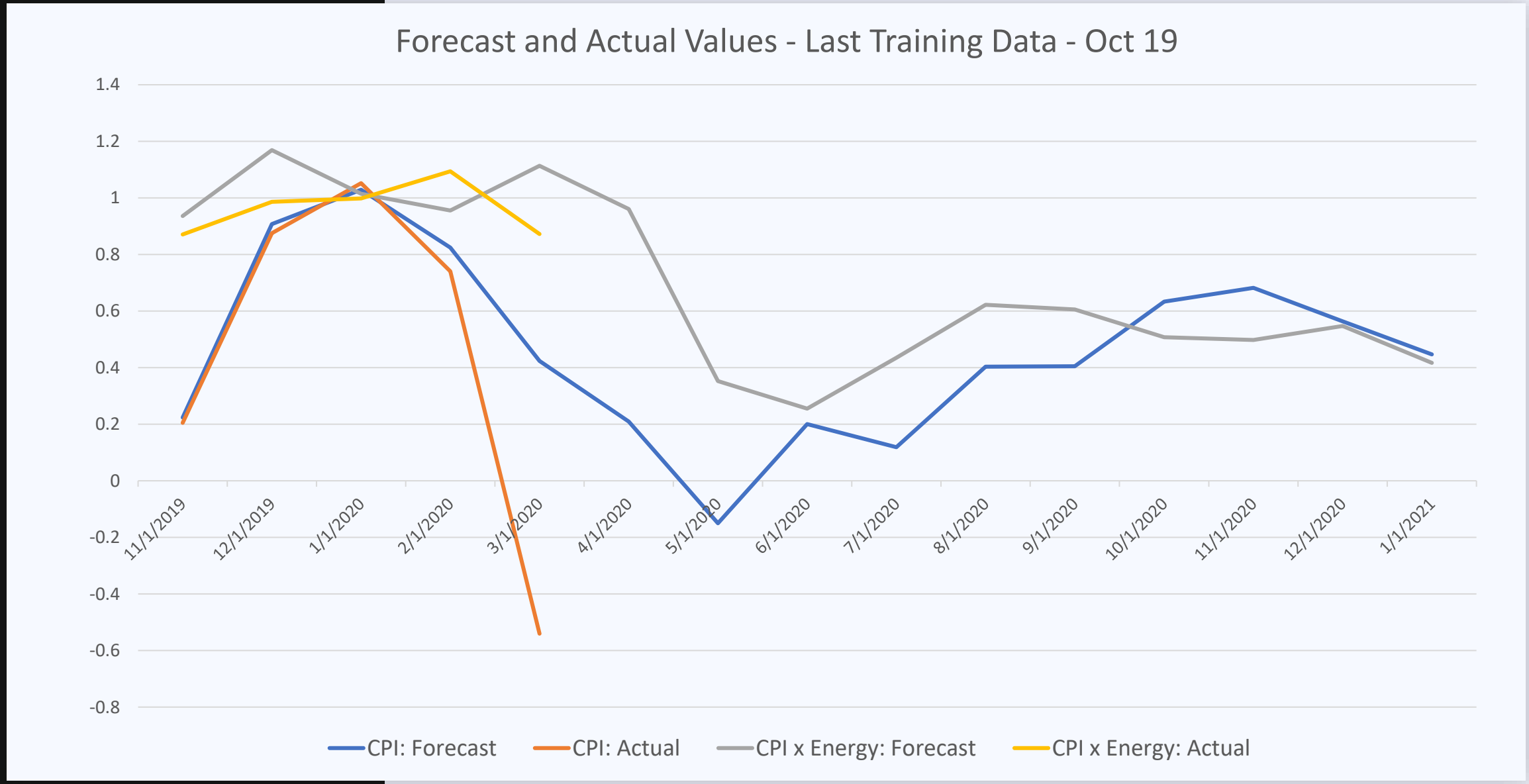
FORECASTING PLOTS: FOOD IN CORE, TEST PERIOD



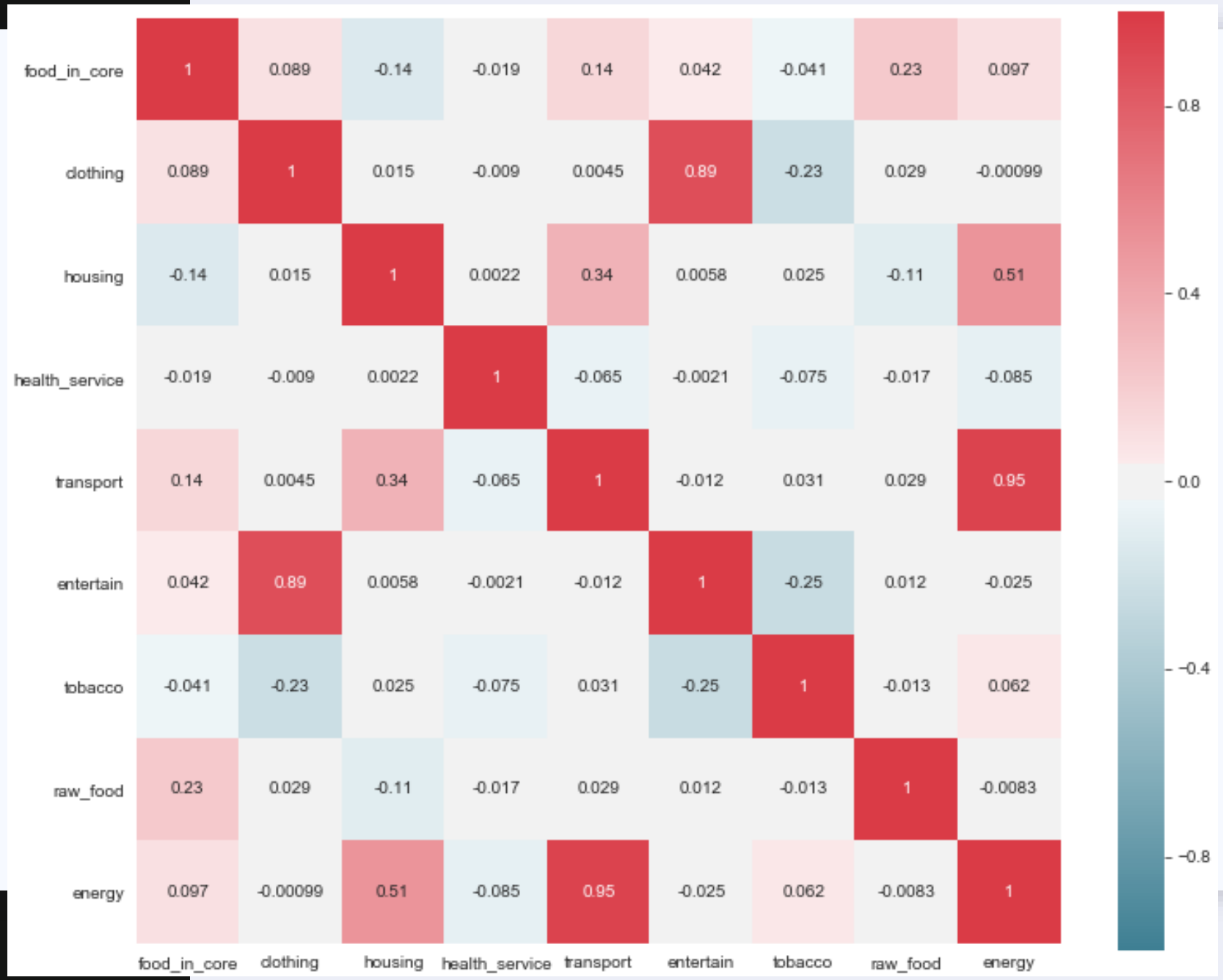
FORECASTING PLOTS: ENERGY, TEST PERIOD



FUTURE FORECASTING PLOTS: CPI



POSSIBLE REASON WHY IT WORKS?: LOW CORRELATION AMONG DISAGGREGATES



A CAUTION BEFORE USING THIS TOOL

The SARIMA model works mostly only when ARIMA models work as well

- If the data are very complicated and cannot be characterized by a sparse structure, it is very likely that none of univariate models will work

If the time series is quarterly or very short, the user must ensure that he or she does not overparameterize the model

LIST OF FEATURES CREATED FOR THIS TOOL

- Configurable forecasting time steps
- Configurable validation time steps
- Configurable model specifications
- Ability to deal with time series with different number of observations
- Two available validation metrics – MAE/RMSE
- Preemptive measure for a very small sample size
- Automatically test for unit root using ADF test both for the original series and the extracted one after the break
- Multicore processing for faster computation

QUESTION