R&A | DS Task | APMC/Mandi Advanced - Report

Author : Ramaravind Kommiya Mothilal
Date : 23-04-2018 to 27-04-2018

This file contains a brief description of the methodology and discusses the important results. It is recommended to go through the .Rmd files first before reading this report for better understanding.

1. Selection of APMC-Commodity groups:

    For most time series models/analysis, the recommended sequence size is at least two complete cycles of the period [1-2]. Hence, for the monthly data in our case, we need at least 24 observations. In addition, it is also difficult to detect seasonality for time series with less than 24 observations. Hence, I have considered only the APMC-Commodity groups which have at least 24 observations in them for analysis.

2. Choice of Decomposition method

    In order to decompose a time series into trend, seasonal and residual components, the “classical decomposition” method is not a suitable choice when there are very few observations. This is because, no trend estimate exist for the first six and last six observations (for monthly data). Consequently, there is also no estimate of the residual component for the same time periods. It is important to note that in our case, there is a only a maximum of 27 observations (from Sep 2014 to Nov 2016) in any group.

    Hence, I have used STL (Seasonal and Trend decomposition using Loess) decomposition which is more sophisticated than the classical method. Further, as STL supports only additive decomposition, in order to obtain the multiplicative decomposition, I first took logs of the data, and then back-transformed the components

3. Deciding between Additive and Multiplicative decomposition:

    The ACF (Autocorrelation Function) on the residuals for additive and multiplicative decompositions are compared to decide the nature of the time series. Since the correlation between the data points within the residuals should be as low as possible, the decomposition that gives minimal value is chosen. However, the results suggest that they don't differ much - there is only 1%-2% difference between the two types of residuals for almost all the time series. In other words, the raw and the de-seasonalised time series do not differ significantly. Hence, only the de-seasonalised time series are used for the subsequent tasks.

4. Highest and Lowest deviations of modal prices from MSPs:

    Figure 1 and Figure 2 displays the highest and lowest deviations of modal prices of different APMC-commodity groups from its corresponding commodity’s MSPs over the given time period. Various observations can be made from these plots, some of the important ones are discussed below.

    Overall, it can be observed from Figure 1 that the commodity “rice(paddy hus)” from APMC “Pen” in the “Raigad” district and the commodity “split black gram” from APMC “Mumbai” in the Mumbai district are often sold at a price 150% more than their corresponding MSPs from September 2014 to November 2016. Similarly, it can be observed from Figure 2 that the commodity “sorgum(jawar)” is frequently sold at less than 20% of its MSP across various APMCs in different periods. In an extreme scenario, the commodity “pigeon pea(tur)” was sold at less than 45% of its

MSP in October 2014.

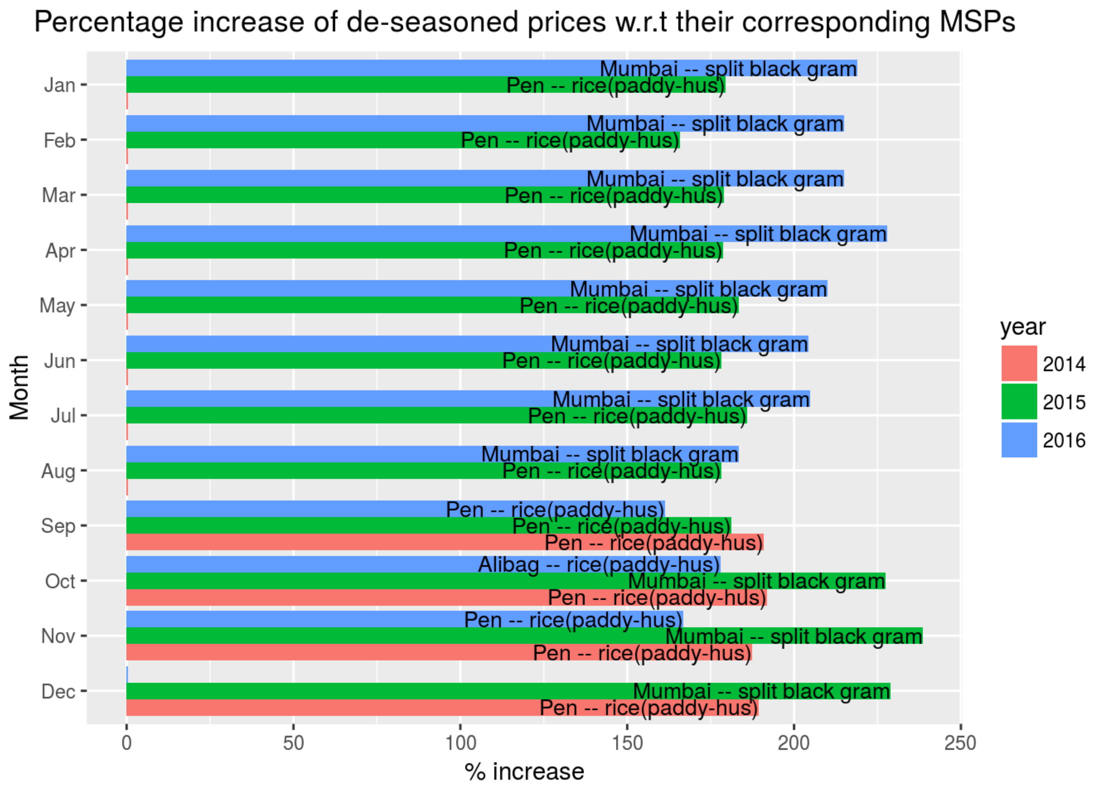


Figure 1

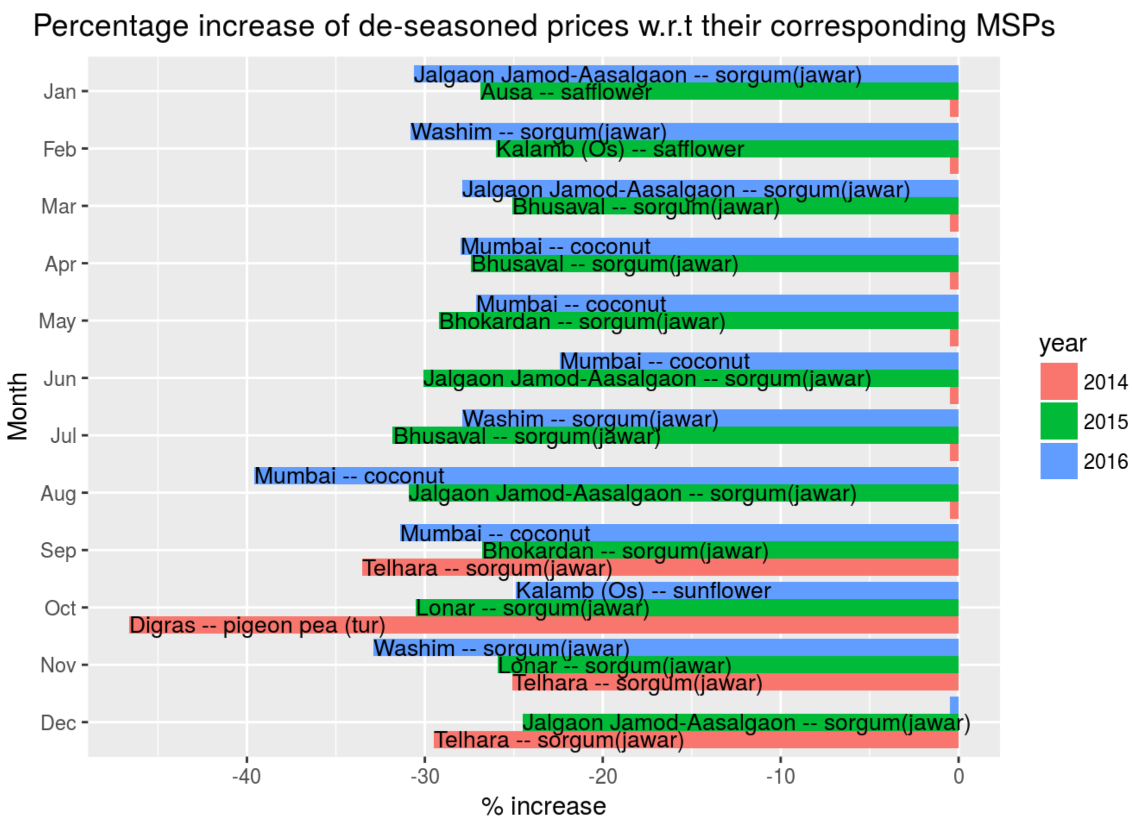


Figure 2

5. <u>Benchmarks for price fluctuations:</u>

Standard deviation (SD) from the mean of a time series could provide us information about the fluctuation present in the series. Hence, I first calculated the SDs of the prices of different commodities across different APMCs to get an idea of their overall variation. Let's call this sd1. Then, I computed the SDs of each APMC-Commodity group (each a time series) to understand how much a particular commodity is varying in a particular APMC. Let's call this sd2. Now, the ratio sd2/sd1 gives as an idea of "how much the price of a commodity fluctuates in an APMC relative to its overall fluctuation across different APMCs". This simple ratio is better than statistical tests such F-test in our case because out time series violates most of the assumptions of these test, such as normality, independence between observations etc.

I first computed the fluctuations in each year and the following tables summarises the results.

Year	**Ratio = 1**	**Ratio = 1.5**	**Ratio = 2**
2014	15	5	2
2015	32	0	0
2016	42	0	0

Table 1: Number of APMC-Commodities in each year for different benchmarks (ratio = sd2/sd1)

Year	**APMC**	**Commodity**	**Max ratio**
2014	Parli-Vaijnath	pigeon pea (tur)	2.33
2015	Digras pigeon	pea (tur)	1.39
2016	Majalgaon	bajri	1.32

Table 2: APMC-Commodities with highest fluctuations (max ratios) in each year

Then, in order to study fluctuations across seasons, our time series need to exhibit somewhat clear seasoning effects. However, there are no clear seasons in our data, perhaps due to its small size. Hence, I manually divided the time period into sub-periods to analyze the fluctuations in each of these sub-periods. I divided the time period from Sep 2014 to Nov 2016 into 5 seasons and considered the months from June to December of each year as rainy season and the months from January to May of each year as dry season. This is just a simple and quick segregation to analyze any possible fluctuations in different sub-periods. The table below summarises the results.

Season	Ratio = 1	Ratio = 1.5	Ratio = 2
rainy_2014	15	5	2
dry_2015	14	0	0
rainy_2015	33	0	0
dry_2016	7	0	0
rainy_2016	43	0	0

Table 1: Number of APMC-Commodities in each season for different benchmarks (ratio = sd2/sd1)

Season	APMC	Commodity	Max ratio
rainy_2014	Parli-Vaijnath pigeon	Pea (tur)	2.33
dry_2015	Malkapur	Maize	1.35
rainy_2015	Digras	Pigeon pea (tur)	1.45
dry_2016	Malkapur	Maize	1.45
rainy_2016	Malegaon	Maize	1.36

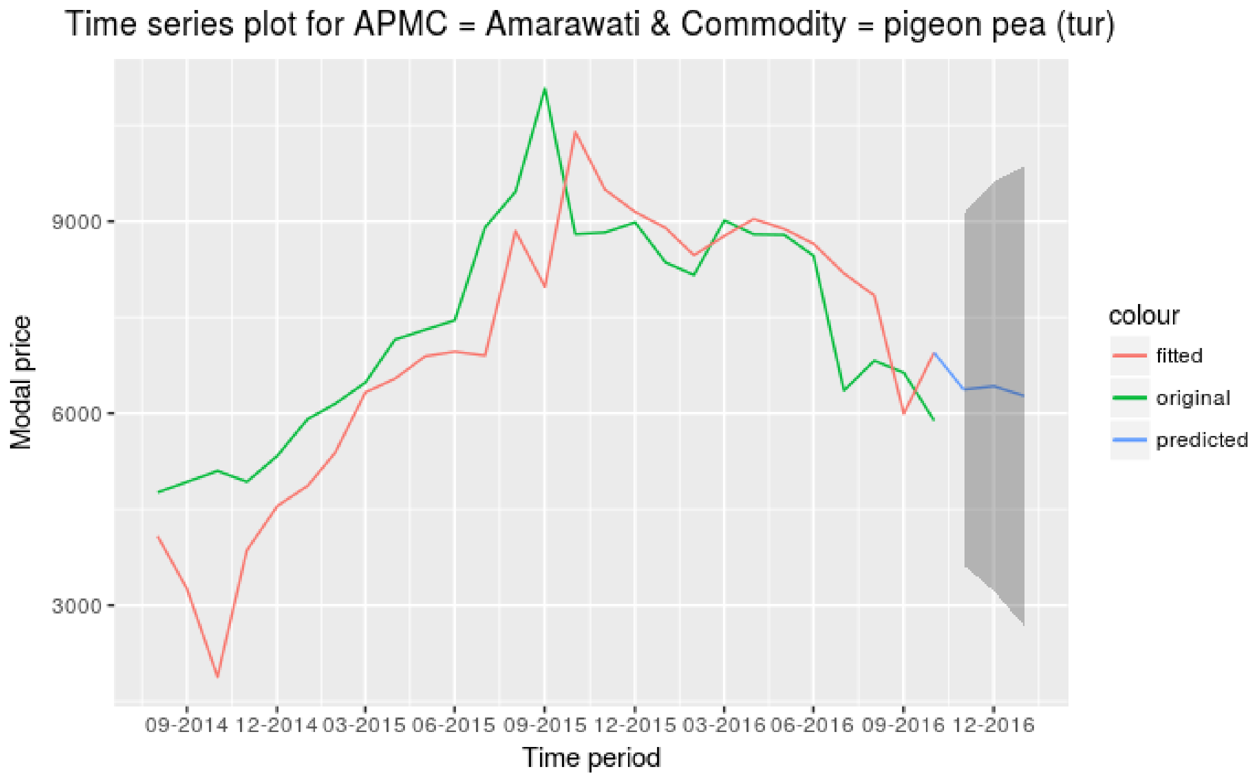
Table 2: APMC-Commodities with highest fluctuations (max ratios) in each season

6. Assumptions for Prediction:

Regarding the selection of models, first, it is difficult to choose optimal parameters for ARIMA models for 278 time series (the total number of APMC-Commodity groups) without plotting ACF and PACF (Partial ACF) for each of them. Considering this, the STL model is chosen for our task. However, STL model handles only additive decomposition (if the series is not log-transformed).

The results obtained so far indicates that there does not exist any significant difference between additive and multiplicative decomposition of (almost all of) our time series. Hence, all the time series are assumed to have only additive components so that they can be used with STL models. This is my assumption for prediction.

The following plot shows the prediction (with 95% confidence interval) of modal prices for a single APMC-Commodity group for the next three months. The plot also compares the fitted values with the original values.



References:

[1] https://robjhyndman.com/hyndsight/short-time-series/
[2]https://stackoverflow.com/questions/37691885/r-ts-error-in-stl-series-has-less-than-two-periods-erroneous