Supplementary Material: Price Forecasting & Anomaly Detection for Agricultural Commodities in India

1 CONTEXT OF ONION & POTATO PRODUCTION IN INDIA

Table 1 shows the major onion producing states in India based on data from the MoSPI (Ministry of Statistics and Programme Implementation). Maharashtra ranks first in onion production with a total share of 28.32%, and forms the focus of most of our analysis in this paper.

The website also reports that India has exported 24,15,757.11 MT of fresh onions of worth Rupees 3,106.5 crores (464.02 USD Millions) during the year 2016-17. The major countries where onion is exported are Bangladesh, Malaysia, Sri Lanka, United Arab Emiratees, and Nepal.

State	*Production in Thousand Tonnes*
Maharashtra	5362.0
Karnataka	2985.8
Madhya Pradesh	2967.4
Bihar	1247.3
Gujarat	1126.5
Rajasthan	800.1
Haryana	667.1
Andhra Pradesh	575.6
Telangana	419.1
Uttar Pradesh	413.4

Table 1: Major onion producing states

Table 2 shows the major potato producing states in India. Uttar Pradesh is the largest producer with a 30.40% share in the total potato production, and is followed closely by West Bengal with a share of 26.07%.

Seasonality in the arrival, mandi and retail time series can be clearly observed in Figure 1, which follows (almost) similar cycles in sowing and harvesting every year. Unseasonal disturbances in prices can be attributed to a lot of reasons including weather disturbances, hoarding, demonetization, etc. We have discussed two specific instances in the paper - Weather and Hoarding. Weather disturbances are mainly initiated by cases of low or extreme levels of rainfall, where it can destroy the crops and consequently affect their produce. Autumn crops are mostly affected due to such rainfall occurrences, while on the other end it is beneficial for the winter and spring crops while sowing since they can utilize the moisture in the soil. Unseasonal rainfall mainly occurring in winter months can destroy the already harvested crops lying in the open due to lack of storage facilities by smallholder farmers. Such an issue can be exploited by traders who hoard the stock to artificially cause an increase in the prices in the market. We can find similarity in terms of time series, and arrival-price analysis for both onions and potatoes. Just like onions, the decline after the last harvesting season is quite gradual, and not a sudden dip. However, the problem of hoarding is also crucial to the price variations in the case of potatoes, and some states like Odisha have already started taking actions to curb this problem since 2018.

State	*Production in Thousand Tonnes*
Uttar Pradesh	14755
West Bengal	12652.5
Bihar	5719.5
Gujarat	3835.79
Madhya Pradesh	3144.01
Punjab	2571.04
Assam	1072.78
Haryana	813.8
Jharkhand	688.66

Table 2: Major potato producing states

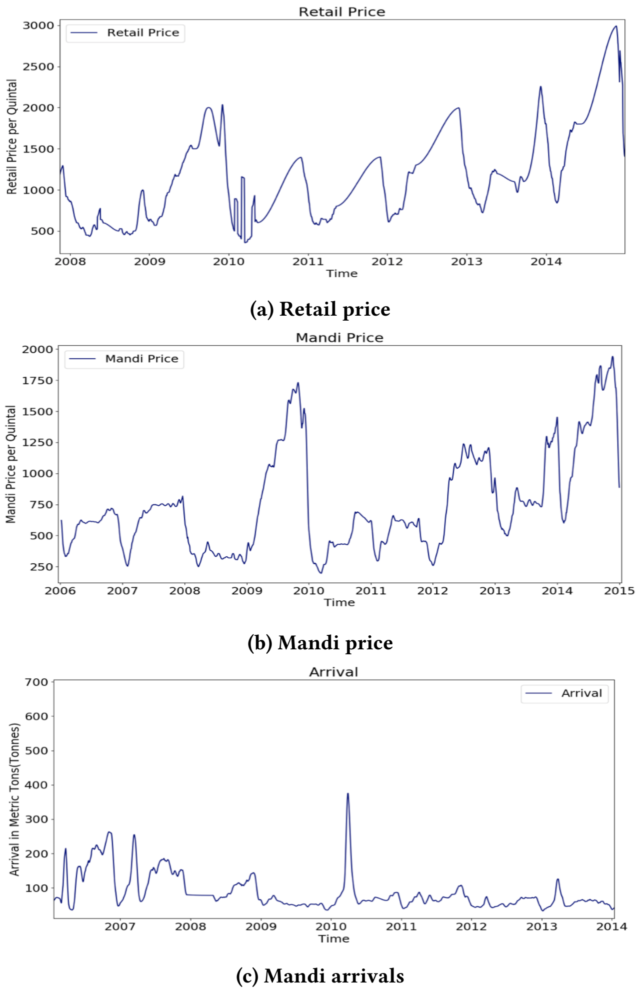
2 PRICE TRANSMISSION AND TRADING LINKAGES ACROSS GEOGRAPHIES: ONION

2.1 Coefficient of Variation(c_v)

We used the Coefficient of Variation in our analysis to identify Source and Terminal Mandis. The coefficient of variation (c_v) is a measure of relative variability. It shows the extent of variability in relation to the mean of the population. One of the key differences between *Source* and *Terminal* mandis is the volatility in the arrivals. We measure volatility in the arrivals by computing the c_v of daily arrivals. c_v is calculated by first computing the monthly means of arrivals for a given mandi for a given center, and then taking the ratio of standard deviation to the mean value. It gives us a measure of how the arrival stock volume fluctuates for a given mandi. Based on our study of the source and terminal mandis, the classification problem is essentially equivalent to placing a threshold value on this coefficient. Roughly, it is around 0.2 based on our study of these crops i.e. mandis with c_v value greater than 0.2 are source mandis and less than 0.2 are terminal mandis.

$$c_v = \frac{\sigma}{\mu}$$

COMPASS '19, July 3–5, 2019, Accra, Ghana
© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) Retail price

(b) Mandi price

(c) Mandi arrivals

Figure 1: Example of the 3 time series for Potato

Centers	Mandi	Mean Arrival (tonnes)	Coeff. of variation
Bengaluru	Bengaluru	2762	0.51 (source)
Mumbai	Pune	1167	0.42 (source)
Mumbai	Lasalgaon	1339	0.25 (source)
Lucknow	Bahraich	907	0.122 (terminal)
Delhi	Azadpur	110	0.032 (terminal)
Hyderabad	Karimnagar	762	0.126 (terminal)

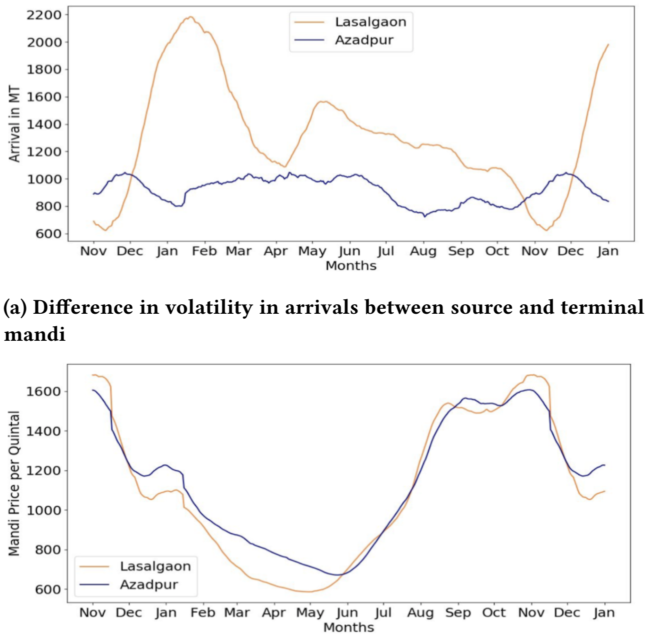
Table 3: Onion retail centers and mandis

Centers	Mandi	Mean Arrival	c_v
Lucknow	Bahraich	907	0.122
Delhi	Azadpur	110	0.032
Hyderabad	Karimnagar	762	0.126

Table 4: Mandis with lowest c_v values

As can be deduced from the tables 3 and 4, mandis with high c_v values are the source mandis because the arrivals in the source mandis do not depend on the demand. It is dependent on the produce brought by the farmers which can show large fluctuations on a daily basis because of seasonal variations due to cropping cycles. On the other hand, mandis with low c_v values are the terminal mandis where the arrival volumes in the mandis depend on the consumer demand which is not expected to vary much on an average. Interestingly, we can see that source mandis have a higher mean arrival than the terminal mandis in general.

It can also be visualized in a better way by plotting the mandi arrival graphs (Figure 2 a) for an average year. Here, we take Lasalgaon as a source mandi and Azadpur to be a terminal mandi. We can clearly see that there are large variations in the arrivals of Lasalgaon mandi (source mandi) whereas the Azadpur arrivals (terminal mandi) do not show such large variations. But this does not have any impact on the average mandi prices at the two places as can be clearly observed in Figure 2 b.



(a) Difference in volatility in arrivals between source and terminal mandi

(b) Similarity in the mandi prices of source and terminal mandi

Figure 2: Source and Terminal Mandi Comparison

Centers	Mandi	Mean Arrival	c_v
Kolkata	Nadia	306.25	0.1586
Kolkata	Chakdah	67.65	0.0843
Lucknow	Bijnaur	6.55	0.1557
Lucknow	Safdarjung	55.41	0.033

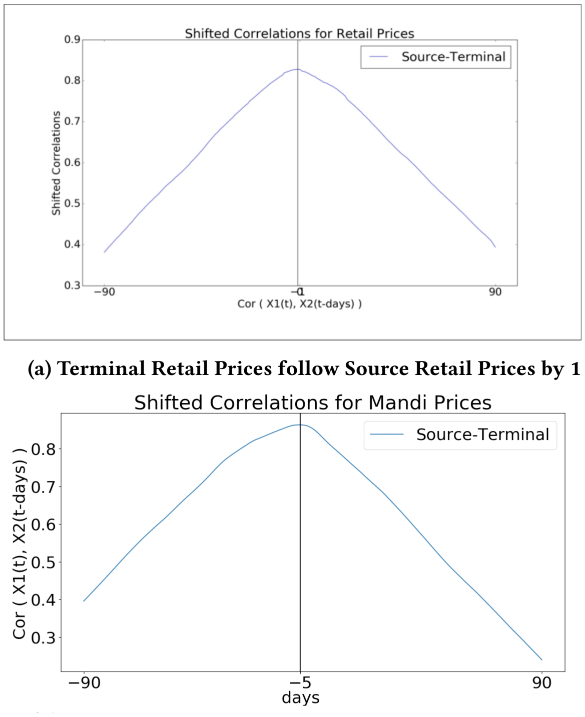
Table 5: Mandis with lowest c_v values

Due to trade linkages across mandis, and also simply price transmission through various communication channels, we can expect some correlation linkages across the mandis. We used the Pearson metric to compute a linear correlations between pairs of mandi time series, after smoothening. It might so happen that two time-series might not align perfectly with each other, and more often than not there is a certain *delay* associated between the time series which can be positive or negative. To align the time series, we *adequately shifted* one time-series keeping the other stationery at each shift-step. According to the Cauchy-Schwarz inequality, it has a value between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. We might not be interested in the cases when the correlation coefficient is very small since a weak correlation implies the time series under consideration are nearly independent of each other. Hence, peak correlations above a certain threshold are useful.

For two samples S_1 and S_2, Pearson correlation can be computed as:

$$r(S_1, S_2) = \frac{\sum_{k=1}^{n} (S_1(t_k) - \bar{S_1})(S_2(t_k) - \bar{S_2})}{\sqrt{\sum_{k=1}^{n} (S_1(t_k) - \bar{S_1})^2 \sum_{k=1}^{n} (S_2(t_k) - \bar{S_2})^2}}$$

where $S_1(t_k)$ is a particular value of the time series S_1 at time t_k, and $\bar{S_1}$ is the the average value of S_1.



(a) Terminal Retail Prices follow Source Retail Prices by 1 day

(b) Terminal Mandis follow Source Mandi prices by 5 days

Figure 3: Shifted correlation between a source and terminal pair of onion mandis

In Figure 3a, we have plotted shifted or delayed correlations between average source retail prices and average terminal retail prices. We observe that a peak occurs at X = -1 which simply means that retail prices in terminal mandi centers follow the source mandi retail prices by just 1 day. This is intuitive because any changes in the retail prices would initially begin at the source mandis and then will be quickly communicated to the terminal mandis within a day or two. This gives us some knowledge of how strong the network of the traders is in this nation.

Source Mandi	Terminal Mandi	Delay	Peak Correlation
Lasalgaon	Delhi(Azadpur)	-1	0.954
Lasalgaon	Lucknow(Bahraich)	-8	0.936
Bangalore	Delhi(Azadpur)	-7	0.753
Bangalore	Lucknow(Bahraich)	-13	0.754
Pune	Delhi(Azadpur)	-1	0.918
Pune	Lucknow(Bahraich)	-5	0.903

Table 6: Absolute values of peak correlations for various mandis of Onion

Similarly, we have the shifted correlations between average source mandi prices and average terminal mandi prices in Figure 3b. This time the peak occurs at X = -5. This means that the mandi prices at terminal mandis follow the trend at source mandi prices after a duration of about 5 days. This leads us to the conclusion that retail price information travels faster than the mandi price information because the information network of farmers may not be as strong as that of the traders.

The magnitude of actual correlations is also a strong factor in analyzing the lead-lag graphs for various mandis as shown in Table 6. We can see that Lasalgaon mandi has more impact on Delhi mandi prices since it is more correlated than Bangalore or Pune. Hence, for every terminal mandi, we can define an ordering for the source mandis in the order of their impact on the latter.

Figures 4a & 4b shows the cumulative & yearwise variation in the delay lag days and correlations with time between Mandi Prices. It is evident that with time, onion mandis have become more correlated between source and terminal regions with a lag of 21 days in 2006 to a lag of 3 days in 2017. This shows that markets for onions have become more synchronized in terms of price flow with much lesser lag.

No fixed pattern has been observed for anomalies as can be seen visually in Figure 5. This shows that anomalies are not clustered together but rather distributed over the entire time series.

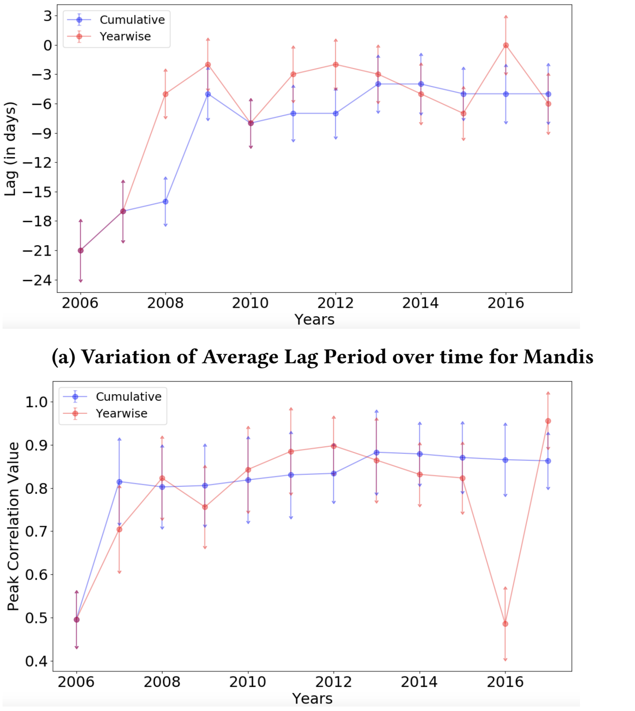
3 TIME SERIES FORECASTING

Given the time series up to a certain timestep t, our motive is to predict(forecast) the values of the time series after step t. We start with univariate modelling in which there is only one underlying variable.

3.1 Auto Regressive Integrated Moving Average(ARIMA)

The ARIMA is a combination of Auto Regressive(AR) and Moving Average(MA) time series models. It takes into account both the lag values of the underlying variable and also the lag error terms. $ARIMA(p, d, q)$ model is written as:

$$\phi^p(L)\nabla^d X_t = \theta^q(L)\epsilon_t$$



(a) Variation of Average Lag Period over time for Mandis

(b) Variation of Average Peak Correlation over time for Mandis

Figure 4: Cumulative & Yearwise Variation in the lead-lag pattern over time for Onion Mandis

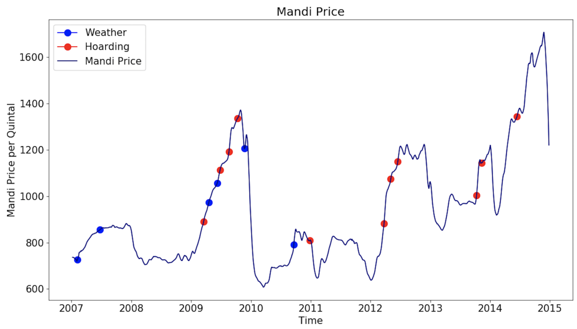


Figure 5: Distribution of Hoarding and Weather events over Lucknow Mandi Time Series for Potato

where L is the lag operator, i.e. $L^h X_t \equiv X_{t-h}$, ∇ is the differencing operator, i.e. $\nabla X_t \equiv X_t - X_{t-1} \equiv (1-L^1)X_t$. We also have $\phi^p(L) = 1 - \phi_1 L^1 - \cdots - \phi_p L^p$ and $\theta^q(L) = 1 - \theta_1 L^1 - \cdots - \theta_q L^q$. p is the number of lag values of the underlying variable X_t. d is the differencing factor which helps in converting a non-stationary time series to a stationary one. q is the number of lag error terms in X_t which are sampled from standard normal distribution. $\phi_i' s$ & $\theta_i' s$ are the parameters of the model obtained after fitting on the given time series. The optimum values of p, d & q are obtained using Box Jenkins method.

3.2 Seasonal Auto Regressive Integrated Moving Average(SARIMA)

It is an extension of ARIMA model to capture the seasonality in the time series. It is represented as $ARIMA(p,d,q)(P,D,Q)_m$ where P, D & Q are the seasonal hyperparameters defined similarly as p, d and q respectively and m is defined as the number of time periods until the pattern repeats itself in the time series, which in our case is 365(corresponding to one year). In SARIMA, seasonal AR and MA terms predict the values of the underlying variable using data values and error terms that are multiples of m. Mathematically, it is written as:

$$\Phi^P(L^m)\nabla^D \phi^p(L)\nabla^d X_t = \Theta^Q(L^m)\theta^q(L)\epsilon_t$$

where the non seasonal components are:
AR:

$$\phi^p(L) = 1 - \phi_1 L^1 - \cdots - \phi_p L^p$$

MA:

$$\theta^q(L) = 1 - \theta_1 L^1 - \cdots - \theta_q L^q$$

and the seasonal components are:
Seasonal AR:

$$\Phi^P(L^m) = 1 - \Phi_1 L^m - \cdots - \Phi_P L^{Pm}$$

Seasonal MA:

$$\Theta^Q(L^m) = 1 - \Theta_1 L^m - \cdots - \Theta_Q L^{Qm}$$

3.3 Modified SARIMA

The prices of commodities increase with time due to a number of factors. This increase has a negative effect in capturing the seasonality in the time series. So, we used linear regression to get the trendline in the time series, removed the trend from the original series and passed this modified series for training to SARIMA. The forecasted value is equal to the value given by the optimized parameters of SARIMA Model plus the trend value obtained from the equation of the trendline.

3.4 Custom Regression Model

Our custom regression model is mathematically described below:

$$(1 - \sum_{i=1}^{p} \alpha_i L^i)(1-L)X_t - (\sum_{i=1}^{r} \beta_i L^i)(1-L)Y_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\epsilon_t$$

Here, Y_t is the shifted time series of some other mandi where the the correlation is maximized. This shift is determined from section 4.2. We can also use inflation data for Y_t. For inflation, we use CPI data specifically vegetable index in CPI. The CPI data is monthly. To convert it into daily data, we assign the value of CPI index of a specific month to all the days in that corresponding month.

3.5 SARIMA with eXternal regressors

SARIMAX is an extension of SARIMA to incorporate more than one variable. Here, external regressors can be inflation, or inflation combined with shifted time series of some other mandi/retail center. It consists of 2 types of variables - endogenous and exogenous. Endogenous variable is the underlying variable of the base time series on which forecasting is done. Exogenous variables are the external regressors.

Actual ↓ Predicted →	**Anomalous**	**Non Anomalous**
Anomalous	91	37
Non-Anomalous	53	91
Precision	0.63	0.71
Recall	0.71	0.63

Table 7: Results of rule based classifier on Onion Data, Accuracy: 66.9%

Actual ↓ Predicted →	**Anomalous**	**Non Anomalous**
Anomalous	77	29
Non-Anomalous	60	73
Precision	0.56	0.72
Recall	0.73	0.55

Table 8: Results of rule based classifier on Potato Data, Accuracy: 62.8%

3.6 Long Short Term Memory(LSTM) Networks

LSTM units are units of a Recurrent Neural Network(RNN) which is referred to as a LSTM Network. LSTMs can be used for prediction and forecasting in sequence data for e.g. time series data. We used a LSTM network with 50 units in the first hidden layer and 1 neuron in the output layer for predicting the price at a given time step.

4 ANOMALY DETECTION & CLASSIFICATION ON RETAIL PRICES

We chose a 43-day long event horizon to build feature vectors in our anomaly detection and classification task, to capture a window of three weeks on either side of the anomaly date.

4.1 Anomaly detection

Before using machine learning methods like random forest classifier for the anomaly detection task, we first used a simple rule based method to detect the anomalies. In this method, we evaluated different statistics on the time series to differentiate between the anomalous and non anomalous events. Some of the statistics are rate of increase of retail price during the interval, or average difference between the actual and expected value of time series during that interval. The results reported in tables 7 and 8 are shown corresponding to the best statistic. The threshold was kept on the rate of increase of retail price during the first 21 days and if the slope of the regressed line was greater than 2, it was classified as an anomalous period. The accuracy in this case comes out to be 66.9% for onion data and 62.8% for potato data. We observe that accuracies achieved by machine learning methods are 9.2% and 5.4% better in case of onion and potato respectively.