Price Forecasting & Anomaly Detection for Agricultural Commodities in India

Lovish Madaan
IIT Delhi
cs5150286@iitd.ac.in

Ankur Sharma
IIT Delhi
cs5150278@iitd.ac.in

Praneet Khandelwal
IIT Delhi
praneetkhandelwal1996@gmail.com

Shivank Goel
IIT Delhi
goelshivank12@gmail.com

Parag Singla
IIT Delhi
parags@iitd.ac.in

Aaditeshwar Seth
IIT Delhi
aseth@iitd.ac.in

ABSTRACT

Fluctuations in food prices can cause distress among both consumers and producers, and are often exacerbated by trading networks especially in developing economies where marketplaces may not be operating under conditions of perfect competition for various contextual reasons. We look at onion and potato trading in India and present the evaluation of a price forecasting model, and an anomaly detection and classification system to identify incidents of hoarding of stock by the traders. Our dataset is composed of time series of wholesale prices and arrival volumes of the agricultural commodities at several village-level marketplaces, and retail prices of the commodities at the city centers. We also provide an in-depth qualitative analysis of the effect on these time series of events such as hoarding, weather disturbances, and external shocks. Our results are encouraging and point towards the possibility of building pricing models for agricultural commodities which can be used to reduce information asymmetries and to detect anomalies that can help regulate agricultural markets to operate more fairly.

ACM CLASSIFICATION KEYWORDS

• **Information systems → Data analytics**; **Information retrieval**; • **Computing methodologies → Machine learning**; • **Applied computing → Agriculture**.

COMPASS '19, July 3–5, 2019, Accra, Ghana
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6714-1/19/07…$15.00
https://doi.org/10.1145/3314344.3332488

AUTHOR KEYWORDS

Agriculture; Commodities; Time Series; Anomaly; Analysis; Prices

1 INTRODUCTION

Price fluctuations in agricultural commodities is an important area of study in economics and development. High prices increase the expenses of retail consumers while low prices reduce the incomes of farmer producers. In India, rainfall is a significant source of price variation since the majority of agricultural production is rain-fed rather than reliant on robust irrigation systems[1]. Poor or excess or erratic rainfall can destroy crops and is especially detrimental to smallholder farmers who have severely strained cash flows with little cushion to manage such disruptions [5]. Several commodities for export such as cotton and oilseeds are also affected by global dynamics including speculation, as what happened during 2007-2008 when rising prices prompted cultivators to grow these crops [13]. Governments have typically responded to events of low prices by increasing the MSP (Minimum Support Price) to procure some commodities themselves through state-owned enterprises so that farmers get a decent price for their produce [28], or by offering debt waivers to farmers [43]. Similarly, responses of the government to events of high prices have been to restrict exports by imposing a high-enough Minimum Export Price (MEP) so that exporters are forced to sell locally and bring down domestic prices [14], or to import commodities and sell them at a subsidized price [37]. These measures often tend to be delayed and reactive, and have their own sets of limitations in systemically addressing the problem of price fluctuations. Market-based solutions like commodity exchanges have also been initiated in India to enable both farmers and buyers to get more predictable prices, but the reach of these exchanges and the reliability in their functioning remains suspect [38].

[1] 68% of the net sown area in India is rain-fed [data from the Union Ministry of Agriculture website]

Price fluctuations in the domestic market therefore continue to be frequent, and are in fact accentuated by the actions of local traders for whom weather disruptions or other events present profit-making opportunities. The agricultural marketplace in India is built of a large network of over 7500 government-regulated local marketplaces (called *mandis*) where farmers sell their produce to traders. These traders transport the produce to other states or city centers, and sell it to retailers. The reality of these supply chains is however very complex. First, smallholder farmers are often unable to bring their produce to mandis themselves due to transport costs, and for most commodities over 60% of the produce is sold by farmers to local traders who then bring it to the mandis [21]. This implies additional middlemen leading to a loss of margin for the farmers. Second, local or wholesale traders who have access to storage facilities, and often own their own facilities, are able to hoard stocks to create supply shortages and release it when retail prices are high, to get much higher profits [20]. Events like weather disruptions or global price fluctuations provide opportunities for the traders to hoard surreptitiously without getting caught. These higher prices do not benefit the smaller farmers though since the hoarding happens further upstream in the supply chain, and in fact it hurts the low-income consumers who have to purchase at higher retail prices. Third, capital in rural areas is heavily interlocked where often the farmers raise debt to purchase agricultural inputs from the same local elite to whom they sell their produce, and hence they command little bargaining power to be able to get better prices from the local traders. These farmers who are also predominantly poor, thus not only bear the greatest risk in agricultural production but also get smaller profit shares, and inequality is able to perpetuate itself relentlessly [25]. Several market-based mechanisms sometimes co-funded by the government[2] are probably slowly changing this deeply unequal feudal setup but they have a long way to go.

In this complex ecosystem, we try to solve two problems aimed at empowering smallholder farmers and low-income consumers. First, to help the farmers get a better price for their produce, we build a price forecasting model that can predict daily prices 30 days into the future, and can help farmers make a better decision of when to sell their produce. Second, to help low-income consumers who are affected badly by high prices, and who also tend to be smallholder farmers themselves, we build a hoarding detection model to strengthen regulatory mechanisms in the operation of agricultural markets in the country. For both these systems, we look at the agricultural commodities of onions and potatoes. These are both important crops with a large nation-wide domestic consumption, and are also probably simpler crops to analyze. Both are not covered under MSPs provided by the government, and both have long storage lifetimes; potato requires cold storage beyond a few weeks, but onions can be stored for longer even in temporary shelters that provide a dry and cool environment. The prices are therefore expected to be affected only by rainfall, productivity and area under production, manipulations by local traders, and other sporadic events. Using daily data for retail and mandi prices, and arrival quantities at mandis, collected for a period of more than a decade, we experiment with different price forecasting models and are able to finally achieve good performance using a multi-variate regression model. We then obtain a dataset of news articles carrying information about hoarding related incidents; treating this as a positive set, we train classifiers to spot hoarding using the time series data of prices and arrivals. We do the classification at two levels: after the event has occurred to see if we can correctly classify whether hoarding happened or not, and while the event might be underway to see if we can provide an early-warning for a likely incidence of hoarding. We believe that a technology platform to do such price forecasting and hoarding detection on a regular basis, and extended to other commodities, can help tame rural power structures and make it easier for fairer and responsible stakeholders to grow. The fact that more ethical capitalism is needed is highlighted by growing farmer suicides [9] and violent protests [1] due to farmers allegedly not being able to meet even their input costs for production, and the government itself having stretched fiscal constraints to not be able to offer anything meaningful through MSPs or other subsidy mechanisms [2]. The mandi system channels 60-80% of all agricultural produce, it is therefore important to strengthen the regulatory mechanisms and address information asymmetries for these markets to function better. Our proposed method can benefit both the farmer producers as well as the retail consumers.

We next describe related work in this area, followed by an introduction to the context of onion and potato cultivation and marketing in India. We then present an evaluation of our forecasting and classification methods, and conclude with a discussion of promising future work in this area.

2 RELATED WORK

Time-series modeling for price forecasting has been an active area of research. Standard techniques include the Auto

[2] This includes the setting up of storage facilities and cold-storage chains, growth in formal sources of credit such as microfinance institutions, setting up direct farmer-to-consumer marketplaces, sale of insurance products for farmers as a substitute for debt, opening up to contract farming, electronic marketplaces such as e-NAM supplemented with logistic networks for transportation of produce, aggregation of smallholder farmers into cooperatives and producer organizations, and setting up of local processing units to move up the value chain, among others.

Regressive (AR) and Moving Average (MA) models, the AR-Integrated-MA (ARIMA model), and seasonal ARIMA [39]. We evaluate these techniques as a baseline, and extend them to multivariate time-series modeling with exogenous variables, similar to [7]. We also compare our results with an LSTM approach, similar to one used for forecasting food prices in India [27]. Another recent study in the Indian context [31] adapted the collaborative filtering approach of recommendation systems to both fill missing data and to forecast price movement (increase, decrease, stay same) by one time step using data from neighbouring mandis. While we use statistical methods and aim to forecast the actual price, we plan to use such collaborative filtering methods in the future especially for imputation to be able to study mandis which have significant missing data.

Similar to our goal to detect malpractices in agricultural trading markets, others have build methods to detect insider trading in stock markets [10], and to identify instances of market manipulations [8]. They use statistical techniques to test for significance of relationships between features of different time series during incidences of market problems. ICA is another technique that has been applied actively, for example on cash flows of different branches of a retail chain [35], to identify hidden factors that might have influenced some branches but not others. Machine learning techniques have also actively been used, such as [17, 22, 44]. Our work too uses machine learning techniques in combination with statistical methods, and in the future we plan to extend to graph-based modeling of multi-variate time series for anomaly detection.

Certain commodity specific characteristics may also present opportunities for anomaly detection. In an interesting demand model for onion consumption [15], it is argued that when prices are low then onion consumption (proxied via mandi arrivals) does not increase, pointing to a fixed amount of onion requirement in the diets of Indians. When prices are high then consumption does decrease, pointing to a standard negative elasticity in demand. Deviations from this demand model could also be used as an anomaly detection technique.

Our work is also related to research in using text data such as news articles to improve price forecasting in time series. [11] applied ARIMA on data for food prices in India, and were able to improve the predictability of their model by modifying ARIMA to incorporate shocks caused due to events that could be identified through news articles. In a similar way, [19] were able to improve predictability of retail sales data of several products by incorporating information from Google trends. We plan to work on similar ideas in the future to model external shocks in time series models, but currently we only use newspaper data to build a ground-truth for our machine learning classifier.

Analysis of commodity food prices of mandis in India is actively pursed by economists. *Kapur et al* [16] have initiated a project to analyze variations in prices of commodities across mandis in India, and relate it to MSP and other initiatives by the government to determine policy effects and make recommendations for price management. Similarly, studies have examined the extent of seasonality in prices in different African markets [23]. The Competition Commission of India ran an extensive study to understand the reasons behind price fluctuations of onions [18]. Price and arrival movements of several commodities during the demonetization event were examined in detail [4]. While these studies have shaped our methods, our goal is different: We are building models to identify hoarding and to do price forecasting, rather than only characterize variations in prices and arrivals across different regions and times.

While one of our areas of focus is commodity price forecasting, there has been extensive work in making current commodity price information available to farmers, with a similar goal of bridging information asymmetries to give more bargaining power to smallholder farmers. The acclaimed Jensen study demonstrated the effect of mobile phones in helping fishermen in Kerala get access to information about prices at different markets along the coast, and which resulted in less deviations in prices across the markets [26]. This was found to not be a generalizable outcome though [41], being affected by numerous contextual factors including power structures in markets, risk taking ability, logistic issues, access to technology and capability to use it, etc. This has also been documented in other studies in India [45] and China [36]. However, strong evidence is also available about the positive effects of deploying ICT solutions for price transparency, such as in India [6] and Sri Lanka [30]. Encouraged by these experiments, and further motivated by the increased market power available to stakeholders with better access to resources for market prediction, we feel our contributions in making price prediction information accessible easily to farmers may help them make better decisions about when and where to sell their produce.

3 DATASET FOR ONION & POTATO PRICES AND ARRIVALS

The Agmarknet (Agricultural Marketing Information Network) website[3] run by the Government of India makes publicly available the daily data on mandi prices and arrival volumes of many commodities, including onion and potato, from across 1514 mandis in the country. We scraped all the data for onions & potatoes from all mandis, for almost 11 years from January 1st 2006 to November 30th, 2017. This

[3] http://www.agmarknet.nic.in/agnew/NationalBEnglish/Datewise-CommodityReport.aspx

data contained many missing values though, and therefore our analysis is restricted to only a few retail centers and mandis for which we had enough data. The selection was done to ensure that there was no missing data for more than 60 continuous days, and at least 65% of data for all the days was available.

The National Horticulture Board runs a portal[4] which provides retail prices from across 30 district centers across the country. We crawled these retail prices as well for all the above years. We also mapped the mandis to their nearest district centers using a simple Voronoi diagram approach, and although this is not an accurate assumption about the nearest district center being the primary destination for its neighbourhood mandis, it does help us analyze price movements in nearby geographies.

Additionally, we obtained monthly rainfall data for western Maharashtra for this period.

Finally, we manually identified all news reports from the Times of India (a leading English daily newspaper) archive about anything to do with onion and potato prices. This helped us create ground-truth labels for hoarding or weather or other events related to onion and potato production and marketing in the country. We obtained over 2000 articles from the newspaper archive and manually selected only the ones relevant to onion and potato commodity pricing, from these we further filtered out articles that did not reference statements made by government officials to weed out rumours or speculation about hoarding. This finally left us with 350+ articles. Multiple articles could be talking about the same event, and we further clubbed together articles describing events at the same location written within 14 days of each other. Finally, we were left with 128 events about hoarding or weather related aspects that affected onion prices, and 106 events that affected potato prices. We also found additional 20+ events about strikes and transport problems, but since there were very few such events we have not considered them for the event classification analysis in this paper.

4 CONTEXT OF ONION AND POTATO PRODUCTION

4.1 Seasons and main production centers

The major onion producing states in India are Maharashtra, Karnataka, and Madhya Pradesh, together making up almost 70% of the onion production of the country. Maharashtra alone has a share of 28.32% and forms the focus of most of our analysis in this paper.

In Figure 1, the green curve shows the average daily arrival (quantity in tonnes) and the blue curve shows the average daily mandi price (rupees per quintal) of onions in a mandi in Maharashtra for a prototypical year. Onion arrivals are healthy during Kharif and Rabi harvest seasons and show low prices. During other times, stored onions are sold which leads to higher prices.

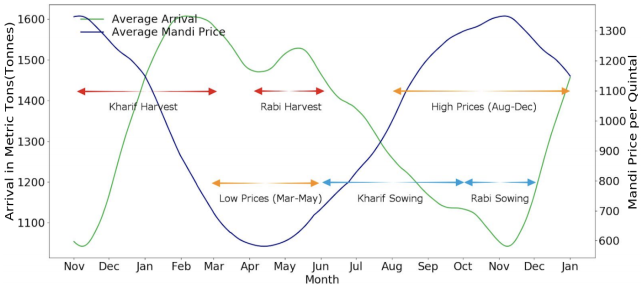


Figure 1: Average Arrival and Prices of Onion

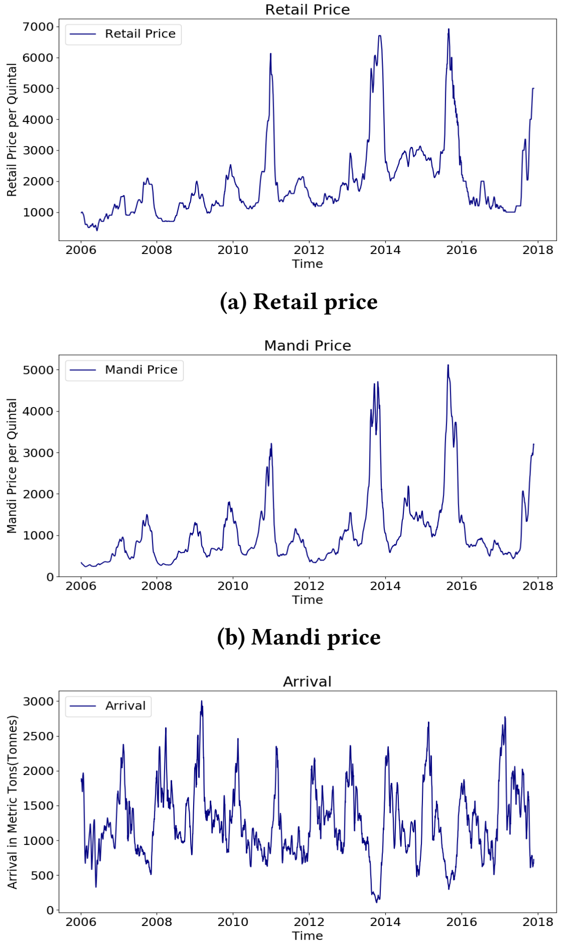
	Sowing	*Harvesting*
Rabi	Oct–Nov	Apr–May
Kharif	Jun–Sep	Nov–Feb

Table 1: Harvesting seasons of onions

Onions are grown in two main crop seasons in India, Kharif and Rabi, as shown in Table 1. The Kharif crop constitutes 40% of onion production and is sown during the months of June to September which is also the time of monsoon rains in India. It is then harvested from November to February. The Rabi crop constitutes the rest of the 60% of onion production and is sown after the monsoons during the months of October and November. Rabi harvesting is done from April to May. Figure 1 shows that the Kharif harvest starting in November leads to a rapid increase in mandi arrivals of onions, which continues into the Rabi harvest until May. Beyond May, until the following November for the next Kharif harvest, very little onion harvesting happens and mostly stored produce is released in the mandis. Typically, smallholder farmers who do not have storage capabilities, sell their produce as and when it is harvested during the months of November to May, and get fairly low prices for their produce because the markets have a glut of onions at that time. Since onions can be stored for as long as six months under appropriate conditions, traders with access to storage facilities procure these onions cheaply from smallholder farmers during the harvesting months and sell the stored onions in the mandis or retail markets when prices start climbing after the harvesting is over. Mandi price movements follow the inverse trend. They start dropping as the Kharif harvest hits the market, and start rising after the Rabi harvest. The highest prices are just before the Kharif harvest when the stored onions from the previous harvest are almost

[4] http://nhb.gov.in/

exhausted, and the lowest prices are during the Rabi harvest when abundant onion production has flooded the market.



(a) Retail price

(b) Mandi price

(c) Mandi arrivals

Figure 2: Onions: Retail price at Mumbai. Mandi price and Mandi arrivals at Lasalgaon.

Figure 2 shows that this seasonal pattern of onion mandi arrivals, mandi prices, and retail prices, recurs every year - prices drop during the harvesting months of November to May, and rise during the months of June to November. The significant price surges that occurred in 2013 and 2015 were mostly initiated by rainfall disturbances. An excess monsoon can destroy Kharif crops, but can be useful for Rabi crops which are sown after the rains are over and can utilize the moisture in the soil. Unseasonal rainfall during the winter months however can be harmful because it can destroy harvested crops which might be lying in the open due to a lack of access to storage facilities by smallholder farmers. Such events of weather disturbances raise alarms, and as we will explain later, they have often been leveraged by traders who exaggerate the problems and surreptitiously hoard onions to push the prices further. This has sometimes also led to raids by law enforcement officials [34], and in 2012 the Competition Commission of India conducted an exhaustive study to understand the structure and conduct of onion markets in the states of Maharashtra and Karnataka [18]. The study found significant cartelization among onion traders to manipulate the prices and prevent the entry of new players into the network. Other than hoarding as a mechanism to influence prices, the study also observed that prices do not drop sharply when the Kharif harvest hits the market, but follow a gradual decline, and cited that as evidence of price rigging by the traders in conjunction with the mandi commission agents (government appointed agents who assess the quality and quantity of the produce, and conduct mandi auctions).

The states of Uttar Pradesh and West Bengal are 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%.

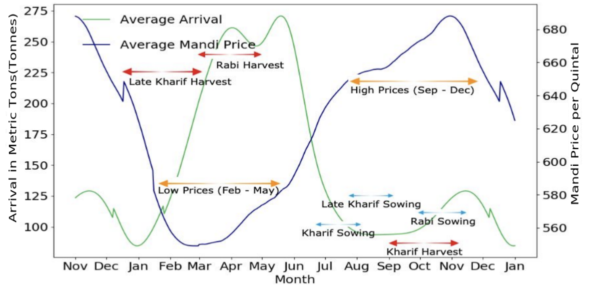


Figure 3: Average arrivals and prices of potato

Similar to the corresponding figure about onions, Figure 3 shows the average arrivals of potato in green and average daily mandi prices in blue, for mandis in Uttar Pradesh. In Uttar Pradesh and West Bengal, potato cultivation happens in three seasons: Kharif, Late Kharif and Rabi. Late Kharif and Rabi harvests comprise the majority of potato production, with harvesting happening during the months of January to April. During this harvesting season, prices are low and arrivals are large. This changes after the arrivals start falling and prices begin to increase in June. Stored potato is released in the market during this time; potatoes can be stored for three to five months in cold storage, the facilities being availed mostly by large farmers and traders.

4.2 Price transmission and trading linkages across geographies

As explained earlier, farmers or local traders bring their produce to mandis, where it is purchased by larger traders who

	Sowing	*Harvesting*
Kharif	July-Aug	Sept-Nov
Late Kharif	Aug-Sept	Dec-March
Rabi	Oct-Nov	March-April

Table 2: Harvesting seasons of potato

sell it to wholesalers and eventually to retailers. However, since major production of the commodities is localized to only a few regions, the commodities are primarily sold at *source* mandis close to the key production centers, from where traders take them to *terminal* mandis in other parts of the country. We were able to identify whether a mandi is a source or terminal mandi for a commodity, by calculating the coefficient of variation of the daily arrival volumes of the commodity at the mandi. Source mandis see considerable variation due to the seasonal production cycles as shown in Figures 1 and 3, but terminal mandis see more or less flat arrival volumes since both onions and potatoes are consumed all year round with a flat demand. Tables 3 and 4 show a few source and terminal mandis for onions and potatoes, on which we focus for subsequent analysis.

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 (tonnes)	Coeff. of variation
Kolkata	Kalyani	1015	1.607 (source)
Kolkata	Kalna	82	0.25 (source)
Lucknow	Mohammdi	79	1.1159 (source)
Lucknow	Lucknow	143	0.3538 (source)
Kolkata	Nadia	306	0.1586 (terminal)
Kolkata	Chakdah	68	0.0843 (terminal)
Lucknow	Bijnaur	7	0.1557 (terminal)
Lucknow	Safdarjung	55	0.033 (terminal)

Table 4: Potato retail centers and mandis

Due to these trade links between source and terminal mandis, and interactions between source mandis in different states themselves, we also observe price correlations between various mandis and retail centers. We develop a simple method of studying correlations between a pair of price time-series: We shift one of the time-series forward or backward by a certain number of days and find out when the two time-series align most closely with each other, ie. have the maximum correlation. We use Pearson correlation and as an example, Figure 4 shows the correlation between a source and terminal mandi pair for onions, where the terminal mandi follows the source mandi by five days.

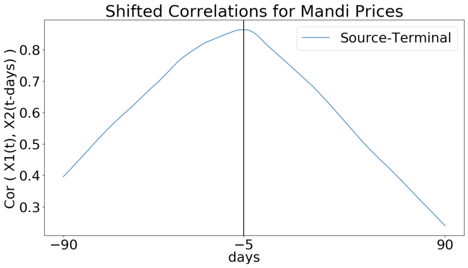


Figure 4: Shifted correlation between a source and terminal pair of onion mandis

We are able to use this method to build a lead-lag dependency graph between pairs of mandis and retail centers, that show a high correlation between their prices. Figure 5 shows two source regions for onions (the Mumbai center with two mandis, and the Bangalore center with two mandis) and two terminal regions (the Lucknow center and the Delhi center). Edges are drawn between pairs of centers that have a high correlation (more than 0.8), directed from the leading center to the lagging center. Shown also is the number of days of lead or lag between the pair of centers. Similar edges are shown between mandis and their closest retail centers. Although there are some exceptions, mandis are seen to typically lead their retail centers by a few days, and the source retail centers similarly lead the terminal centers. Such relationships can help improve price forecasting, as we show later, and also point towards fast transmission of price information across the entire country.

A similar method is followed for potatoes, and Figure 6 shows the lead-lag relationships between the Lucknow and Kolkata source regions. We were unable to identify any terminal regions for potatoes due to significant missing data in the time-series.

We also study changes in lead-lag relationships over the years, to see if the price movements in different mandis have become more or less synchronized with one another. We find that the peak correlation values have indeed increased over the years, especially between 2006 to 2009, after which they have held steady. Similarly, the lead-lag periods have reduced from almost 21 days in 2006 to between 3 to 6 days

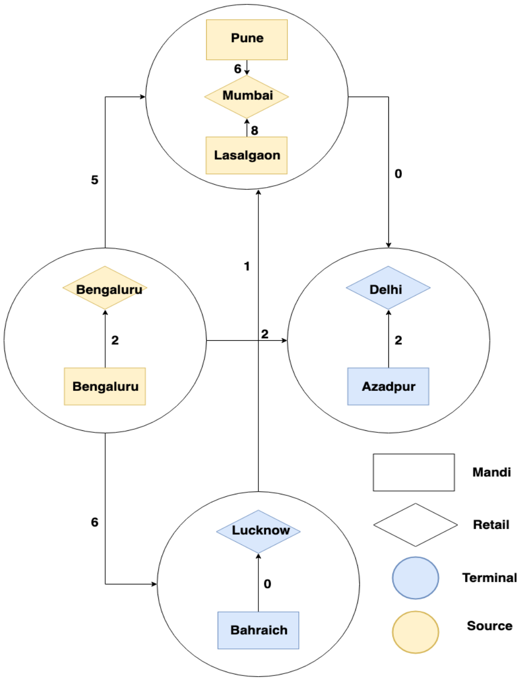


Figure 5: Onions: Lead-Lag graph showing the relationship between retail centers and mandis

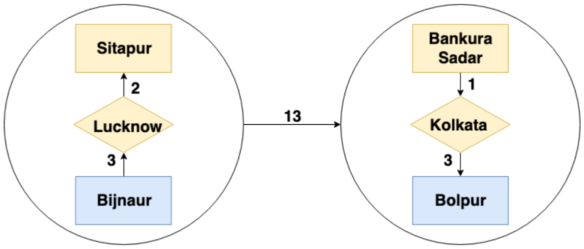


Figure 6: Potatoes: Lead-Lag graph showing the relationship between retail centers and mandis

after 2009. This is shown in the supplementary material [32] and seems to point towards a steady increase in price synchronization across markets over the years, possibly due to the growing penetration of mobile phones which can lead to faster price transmission as also noticed by Jensen in the case of fishermen [26].

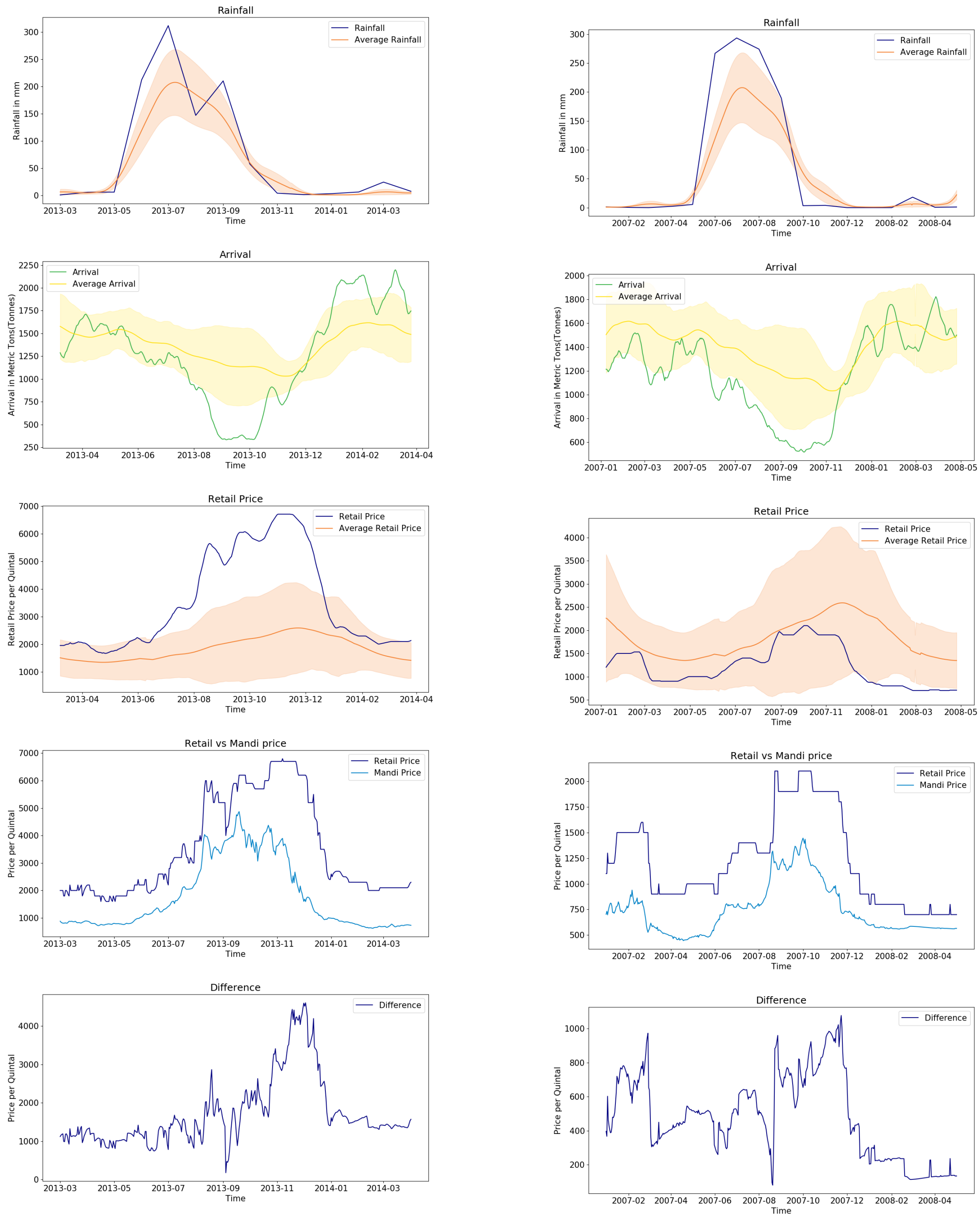
4.3 Qualitative introduction to pricing dynamics

To give a deeper view of the pricing dynamics, we describe a few major events in more detail. In 2013 (Figure 7a), the monsoons started somewhat early and were erratic, and newspapers reported that this destroyed significant Kharif crop of onions. This may seem to be the case with low arrivals reported in September and October which led to a steep price increase earlier than usual, but the arrivals climbed up steeply soon after in December and exceeded the annual averages significantly. If a lot of crop would have been destroyed as per the reports, then this would not have been the case. Newspapers around this time started reporting that traders had deliberately withheld release of the previous season's crop during September and October, in anticipation that the rainfall disturbances would raise alarms and lead to seemingly legitimate price rises. Several raids by law enforcement officials were also conducted during December and January in Maharashtra, and this seems to be a clear case of hoarding of the previous season's crop by opportunistically leveraging negative weather reports.

This can be compared with 2007 (Figure 7b) where too the rainfall was early, leading to reduced arrivals during October and November, and higher prices in this period. The arrivals in the later months however remained within one standard deviation of the annual averages unlike in 2013, and prices adjusted back to normal levels soon too. The newspapers reported weather as having caused problems with the Kharif crop, but no hoarding activities were reported. 2007 therefore seems to be a year when traders did not play foul and the markets reacted normally to weather problems.

An interesting point to notice in both of these years is also the movement of the daily retail and mandi prices, which are the bottom two figures in Figure 7. The difference between the retail and mandi price seems to increase during episodes of price rise, but the mandi prices begin to drop somewhat sooner than the retail prices. In general during normal periods though when no alarming activities have been reported, the relationship between retail and mandi prices is linear with a steady margin of around 100% between the two. Retail prices therefore seem to move with mandi prices closely in the usual course of activity, but they increase more than mandi price increases during anomalous events, and decrease with a delay after mandi prices start to decrease. This shows that traders gain at the expense of both the smallholder farmers as well as the retail consumers. Smallholder farmers are forced to sell their crops at low prices during the glut season when harvesting happens because they do not have access to storage facilities, and traders make standard margins during this time. They however often leverage weather disturbances and hoard the stock to push the prices further, leading to consumer-facing inflation, and while some farmers may benefit from these increased prices in the mandis as well, but traders benefit more and for a greater amount of time from the increased retail prices.

These patterns point to the possibility of being able to spot malpractices such as hoarding, by looking at the different



(a) Hoarding of the previous season's crop, initiated by weather

(b) Weather event but without hoarding

The green curve shows the mandi arrival volumes (left vertical axis) for a particular year and the yellow curve shows the arrivals averaged over the years. The shaded region marks one standard deviation from the average arrivals across the years. The other figures for prices are plotted similarly. The shaded region in the rainfall figure marks 0.5 standard deviation from the annual average. The two figures at the bottom show the difference between the retail and mandi prices.

Figure 7: Weather events with hoarding

price and arrival movements. We turn to this next, to model the time-series for price forecasting and anomaly detection.

5 MANDI PRICE FORECASTING

We next describe the results from different mandi price forecasting models we built for onions and potatoes. We tried seven different models on the data in the following manner. We have 4352 data points (Jan 1, 2006 - Nov 30, 2017) for each of mandi price series, retail price series, and arrival amounts for both the commodities. We first start with the 2006 data and fit a model on the 365 data points for the year, then we forecast on the next 30 days and evaluate the forecasting error. We then fit a model on the 365 plus 30 days, and forecast for the next 30 days. Thus, in a step-wise manner we keep updating the model.

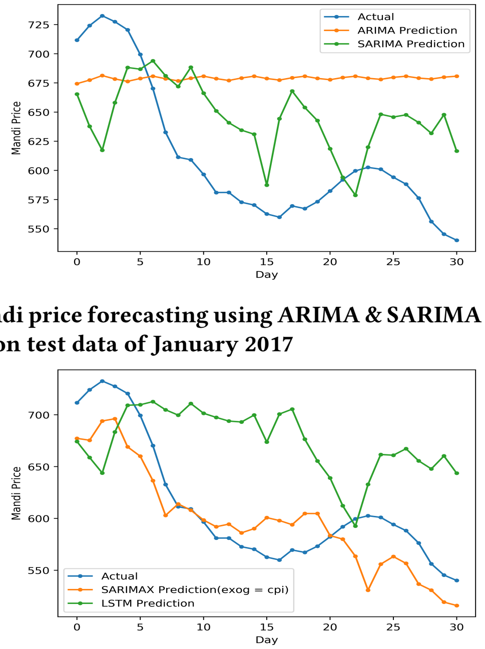
We first tried the univariate ARIMA (Auto Regressive Integrated Moving Average) model as a baseline, which takes into account lag values of the underlying variable and the lag error terms. Optimal parameter values were obtained using the Box-Jenkins method. We next evaluated the seasonal variation model, SARIMA (Seasonal ARIMA), which takes seasonal terms into account as well. SARIMA was able to capture the broad annual variations but as shown in Figure 8a for a sample 30-day window at the Lasalgaon (Mumbai) mandi, these simple regression based univariate models were not able to estimate the time-series accurately. A modified SARIMA model was also evaluated to remove the trend but it did not improve the results.

To make use of the price correlations with neighbouring mandis as observed in the previous section, we next built a custom regression model by modifying the SARIMA equation to include terms from a neighbouring mandi or retail center. We chose a mandi or retail center which showed a high shifted correlation with the target mandi. This model improved the results but further improvement came from the multivariate models explained next.

We evaluated two multivariate models, first using the series of a neighbouring mandi or retail center as an additional exogenous time-series, and then also incorporating the vegetable CPI (Consumer Price Index) series. Figure 8b shows the forecasted values for a sample 30-day window. The multivariate model with two exogenous series performed the best, as shown in Figure 9. The improvement noticed by incorporating CPI into the model indicates that the trends in the overall agricultural market do influence the prices of individual commodities as well.

Finally, we also evaluated a simple LSTM (Long Short Term Memory) model, to study how sequential neural models fare against linear regression based models. This model took as input the last 30-day price time-series for the mandi, and the price time-series for the highest correlated neighbouring mandi as well. A hidden layer of 50 LSTM neurons was then connected to an output layer of one neuron to carry the forecasted price for the next day. The model performed better than the ARIMA and SARIMA models, but not as well as the multi-variate model with two exogenous series. Further improvement should be feasible with using more layers, more extensive hyper-parameter tuning, and also incorporating the time-series of neighbouring mandis and of CPI.

Thus, we were able to build a 30-day mandi price forecasting model yielding average RMSE values of 754.6 for 30-day periods(25.15 for one day). The mean normalized deviation is 0.041, indicating a reasonable performance. This model can be used to provide price forecasting information to farmers, to help them decide whether to hold on to their produce for a few weeks or to sell it right away, and can give more bargaining power to the farmers. We also make use of the model in the anomaly detection step, explained next.



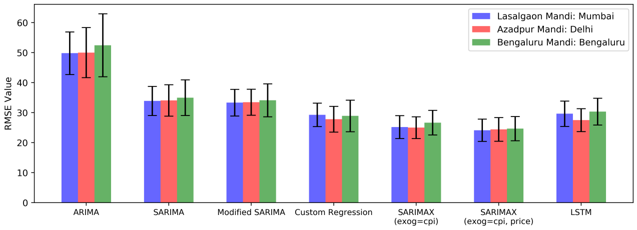
(a) Mandi price forecasting using ARIMA & SARIMA models on onion test data of January 2017

(b) Mandi price forecasting using multivariate models on onion test data of January 2017

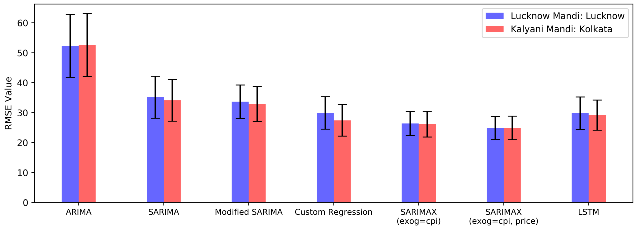
Figure 8: Price forecasting in a 30-day window

6 ANOMALY DETECTION & CLASSIFICATION ON RETAIL PRICES

We now turn to the second problem of detecting trading malpractices like hoarding. We first identify cases of hoarding and weather related anomalies using newspaper reports. We isolate the reports into hoarding incidents irrespective of a weather event, and weather events when no hoarding



(a) Comparison of RMSE of different models on Onion mandi data



(b) Comparison of RMSE of different models on Potato mandi data

Figure 9: Price Forecasting Model Comparison (RMSE Values reported are the average of RMSE values over the whole time series)

incidents were reported. We also club together reports about the same event that happened within 14 days of each other. Since the newspaper reports may have appeared some days after the events actually occurred, and the events themselves may have stretched across several days, we identify the point at which the prices were the highest in a 14 day period before the news report. We then choose a 43 day window around this peak (21 days before and 21 days after), as the period during which the anomalous event happened. These windows serve as samples of anomalous events.

To assemble a set of normal periods as samples when no anomalies happened, we choose periods when no newspaper report was published about any anomalous events, and the difference between the maximum and minimum onion retail price during that period did not exceed Rs. 300 per quintal (this value was approximately one fourth of the average difference between the maximum and minimum retail price for all the events, and hence we assume the threshold to be low enough to not select anomalous events). Using this method, we were able to identify 128 anomalous (58 hoarding and 70 weather) events and 144 normal periods for onions. Similarly, we were able to identify 106 anomalous (47 hoarding and 59 weather) events and 133 normal periods for potatoes. Note that several other anomalous events also took place over the years, such as strikes of transport companies, fuel price increase, even religious festivals when certain foods are avoided, etc, but there were too few incidences of such events for us to be able to add more classes to our anomaly classifiers.

This method for assembling a dataset using newspaper reports of anomalous events and normal periods is not perfect. It is possible that newspapers may not have reported hoarding incidents during some weather events, leading to such cases as being labeled only as weather anomalies. It is also possible that the normal periods identified by us using a hard-crafted rule, may have some unreported anomalies as well. This is however potentially the best we can do given the available data sources, and we proceed with the dataset. We then build two sets of classifiers: to first classify whether a 43-day window appears to be anomalous or not, and then to classify the type of anomaly as weather or hoarding.

6.1 Anomaly detection

We evaluated a random forest binary classifier to operate on different sets of features built on the 43-day event windows. Tables 5 and 6 show the cross-validation accuracies obtained for the different sets of features. Validation sets are formed by dividing the entire duration into 6 month periods; each of these periods has a number of anomalies falling into it, and the classifier is evaluated in a cross validation manner by leaving out in each iteration a 6-month period for testing. We validated that both hoarding and weather anomalies seem to be evenly spread across all planting seasons, and hence taking 6-month validation periods is justified [32]. Simple models using just the daily retail and/or mandi prices as features worked better than hand-crafted features like mean, standard deviation, skewness, kurtosis, etc. The best performance for both onions and potatoes was obtained when residuals from the multivariate price forecasting model (with two exogenous time-series) described in the previous section, were also incorporated as features.

The performance obtained above was when the entire 43-day data is used, ie. a post-hoc detection of anomalies. We also evaluate the models for an early-warning system, by taking fewer and fewer days from the start of the 43-day event window. Table 7 shows the accuracies obtained by taking the first 35 days, 28 days, 21 days and 14 days. The accuracies see a drop of 6-7%. As part of future work, we plan to improve this using graphical models, and also build a larger dataset by crawling regional media newspapers to spot more news reports. We also plan to include a new class of anomalies for low prices, which is a significant problem plaguing the farmers since input costs keep increasing for them but consumer-friendly policies (such as export restrictions) suppress the prices [24].

Feature Vector Set	Accuracy
Retail prices	66.9%
Mandi prices	70.6%
Retail and Mandi prices	71.3%
Mandi prices and Forecasting residuals	76.1%
Retail and Mandi prices, Arrivals	69.8%

Table 5: Accuracy with different sets of feature vectors on Onion data

Feature Vector Set	Accuracy
Retail prices	56.9%
Mandi prices	62.6%
Retail and Mandi prices	61.7%
Mandi prices and Forecasting residuals	68.2%
Retail and Mandi prices, Arrivals	59.3%

Table 6: Accuracy with different sets of feature vectors on Potato data

Days	**Accuracy** (Onions)	**Accuracy** (Potatoes)
43	76.1%	68.2%
35	75.4%	66.8%
28	73.9%	65.1%
21	71.5%	63.8%
14	70.2%	61.7%

Table 7: Accuracies for early warning system

6.2 Anomaly classification

Our next step is to classify the type of anomaly, ie. whether it is indeed a hoarding event, or a weather event when hoarding did not occur. Tables 8 and 9 show the confusion matrices for onions and potatoes, using the same feature set of mandi prices and the forecasting model residuals. The accuracies are not very high and we hope that with a larger dataset and different classifiers, we may be able to improve this. As part of future work, to improve the precision for hoarding classification we plan to build a verification service through which we can contact groups of registered farmers in different locations, and survey them for any reports of ongoing hoarding activities. We feel that such a system of data-driven red flags verified through a community of users can become a powerful tool for empowerment of the farmers.

7 DISCUSSION AND CONCLUSIONS

Almost all of onion and potato procurement flows through the mandis, and hence it is important to regulate these markets and reduce information asymmetries.

Actual ↓ Predicted →	Weather	Hoarding
Weather	55	15
Hoarding	29	29
Precision	0.66	0.66
Recall	0.79	0.5

Table 8: Anomaly classification for onions. Accuracy: 65.6%, F1 Score: 0.71

Actual ↓ Predicted →	Weather	Hoarding
Weather	37	22
Hoarding	21	26
Precision	0.64	0.54
Recall	0.63	0.55

Table 9: Anomaly classification for potatoes. Accuracy: 59.4%, F1 Score: 0.63

The mandis we have described throughout the paper, are operated by APMCs (Agriculture Produce Market Committee) constituted of elected members with the objective to conduct mandi trade in a transparent manner so that farmers and consumers alike are not exploited by middlemen and traders [12]. The APMCs grant licenses to commission agents to operate shops where farmers can bring their produce, have it assessed, and auctioned to traders. Traders too need licenses to operate. In theory, farmers can go to any commission agent and the auctions conducted by the agent should get the best price to the farmers, but this is typically violated through cartelization between the commission agents and traders, and also because the farmers may have obligatory relationships with the agents, often through loans taken by the farmers from the agents themselves [25]. Ties of the agents and traders with regional politics further hamper the goal of APMCs to regulate markets fairly because of likely rent-seeking practices that might exist locally[5].

De-regulating the APMCs to open up trading networks for competition is often discussed [40], but Harriss-White [25] argues that it will not lead to perfect competition and suggests strengthening of the state systems to regulate the markets as a more appropriate approach. Indeed, a model APMC Act was proposed in 2003 to abolish commission agents, have mandis provide paid services to assess the quality and quantity of the produce being sold, allow cross-mandi trading, etc., but the states have only adopted the model act in a piecemeal manner and no state has abolished commission agents so far. The state of Bihar repealed its APMC act altogether but there has been no change in the trading

[5] Such practices have been reported in sugarcane markets [42]

practices for farmers - the commission agents earlier aligned with a regional political party were only replaced by larger trading firms aligned with a different political party [29]. Karnataka too brought some changes but so far has not seen any significant outcomes - cross-mandi trade did not pick up and no new players joined the trading network, further traders prefer visual inspection of the produce themselves than rely on certification by the mandi operators, and the farmers do not want to use the electronic trading platform because they prefer getting paid in cash instead of through bank transfers [3].

Given these contextual factors which will continue to exist in the Indian agricultural market in the foreseeable future, with or without mandis, we see a strong role to use trading data for monitoring and regulating the markets, and to reduce information asymmetries. Vulnerable farmers and unsuspecting consumers may otherwise always get the short end of the stick in unregulated capitalism and its nexus with politics. An experiment with potato farmers in West Bengal indeed showed that when farmers were informed of wholesale prices, they were able to bargain more effectively with local traders and get a higher price [33]. Our methods of using machine learning to spot hoarding incidences and do price forecasting are just preliminary examples of the possibilities, but with access to more granular data, possibly even per-trade, it might be feasible to build a general pricing model which can help impose a check on market operations and lead to fair practices. At the same time, we hope that farmer collectives will become more empowered, farmers themselves will become more aware with greater access to information and communication technologies, government sponsored or market-based social enterprises will improve access to formal credit and logistics such as storage and transportation, and electronic marketplaces will bring transparency, to impose bottom-up checks that can build more equitable markets. Our key contribution in this paper is to emphasize on the role that forecasting and classification methods can play in the operation of agricultural markets in India.

8 ACKNOWLEDGEMENT

We thank IIT Delhi HPC Facility for the computational resources.

REFERENCES

[1] Bina Agarwal. 2017. The seeds of discontent. Retrieved from The Indian Express https://goo.gl/6wxeF1.

[2] Kabir Agarwal. 2018. Why MSP at 1.5 Times Cost Is Another Empty Promise for Farmers. Retrieved from The Wire https://goo.gl/1pcGtR.

[3] Nidhi Aggarwal, Sargam Jain, and Sudha Narayanan. 2016. *The Long road to transformation of agricultural markets in India: Lessons from Karnataka.* Technical Report. Indira Gandhi Institute of Development Research, Mumbai, India.

[4] Nidhi Aggarwal and Sudha Narayanan. 2017. *Impact of India's demonetization on domestic agricultural markets.* Technical Report. Indira Gandhi Institute of Development Research, Mumbai.

[5] Jamie Anderson and Wajiha Ahmed. 2016. Smallholder diaries: building the evidence base with farming families in Mozambique, Tanzania, and Pakistan. *Consultative Group to Assist the Poor (CGAP)* (2016).

[6] Bimal Arora and Ashley Metz Cummings. [n. d.]. Reuters Market Light, Creating Efficient Markets. Retrieved from UNDP Growing Inclusive Markets http://growinginclusivemarkets.org/media/cases/India_RML_2010.pdf.

[7] Nari Arunraj, Diane Ahrens, and Michael Fernandes. 2016. Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry. *International Journal of Operations Research and Information Systems* 7 (04 2016), 1–21.

[8] ASM Sohel Azad, Saad Azmat, Victor Fang, and Piyadasa Edirisuriya. 2014. Unchecked manipulations, price–volume relationship and market efficiency: Evidence from emerging markets. *Research in International Business and Finance* 30 (2014), 51–71.

[9] Nilanjan Banik. 2017. Farmer suicides in India and the weather god. *Procedia Computer Science* 122 (2017), 10–16.

[10] Emilio Barucci and Emanuele Squillantini. 2006. Market Abuse Detection: A Methodology Based on Financial Time Series. *Statistica Applicata* 18, 4 (2006).

[11] Sunandan Chakraborty, Ashwin Venkataraman, Srikanth Jagabathula, and Lakshminarayanan Subramanian. 2016. Predicting Socio-Economic Indicators using News Events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1455–1464.

[12] Ramesh Chand. 2012. Development policies and agricultural markets. *Economic and Political Weekly* (2012), 53–63.

[13] CP Chandrasekhar and Jayati Ghosh. 2008. Global Crisis and Commodity Prices. *Network Ideas, URL https://goo.gl/pU1qTX, last accessed* 15, 09 (2008), 2009.

[14] G Chandrashekhar. 2018. Why the minimum export price is irrelevant? Retrieved from https://goo.gl/nmDGXb.

[15] Devlina Chatterjee. 2016. A simple example for the teaching of demand theory: Aggregate demand estimation for onions in India. *IIMB Management Review* 28, 1 (2016), 20–24.

[16] Shoumitro Chatterjee and Devesh Kapur. 2016. Understanding Price Variation in Agricultural Commodities in India: MSP, Government Procurement, and Agriculture Markets. (2016). National Council of Applied Economic Research.

[17] Haibin Cheng, Pang-Ning Tan, Christopher Potter, and Steven Klooster. 2009. Detection and characterization of anomalies in multivariate time series. In *Proceedings of the 2009 SIAM International Conference on Data Mining.* SIAM, 413–424.

[18] PG Chengappa, AV Manjunatha, Vikas Dimble, and Khalil Shah. 2012. Competitive assessment of onion markets in India. *Institute for Social and Economic Change. Competition commission of India* 1 (2012), 86.

[19] Hyunyoung Choi and Hal Varian. 2012. Predicting the present with Google Trends. *Economic Record* 88, s1 (2012), 2–9.

[20] Wikipedia contributors. 2017. 2010 Indian onion crisis- Wikipedia, The Free Encyclopedia. Retrieved from https://goo.gl/AiAhY6.

[21] Harish Damodaran. 2016. Agricultural marketing: For the Kisan, it's the Bania who still calls the shots. Retrieved from Indian Express https://goo.gl/JrFW3t.

[22] R Drenth. 2014. The signs are there, now predict the future! Predicting System Failure and Reliability. (2014). Artificial Intelligence Radboud University Nijmegen.

[23] Christopher Gilbert, Luc Christiaensen, and Jonathan Kaminski. 2016. Price seasonality in Africa: Measurement and extent. (2016).

[24] Ashok Gulati and Carmel Cahill. 2018. Resolving the farmer-consumer binary. (9 July 2018). Retrieved from Indian Express https://goo.gl/TPdrhZ.

[25] White Harriss. 1996. *A political economy of agricultural markets in South India: masters of the countryside.* Sage Publications.

[26] Robert Jensen. 2007. The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector. *The Quarterly Journal of Economics* 122 (02 2007), 879–924. https://doi.org/10.1162/qjec.122.3.879

[27] Girish K Jha and Kanchan Sinha. 2013. Agricultural Price Forecasting Using Neural Network Model: An Innovative Information Delivery System. *Agricultural Economics Research Review* 26, 2 (2013).

[28] Elumalai Kannan. 2017. Why a price increase alone won't help farmers. Retrieved from The Hindu https://goo.gl/1RrHK7.

[29] Devesh Kapur and Mekhala Krishnamurthy. 2014. Understanding mandis: market towns and the dynamics of India's rural and urban transformations. Center For The Advanced Study of India, University of Pennsylvania.

[30] Sriganesh Lokanathan, Harsha De Silva, and Iran Fernando. 2011. *2. Price transparency in agricultural produce markets: Sri Lanka.* 15–32. https://doi.org/10.3362/9781780440361.002

[31] Wei Ma, Kendall Nowocin, Niraj Marathe, and George H. Chen. 2019. An Interpretable Produce Price Forecasting System for Small and Marginal Farmers in India Using Collaborative Filtering and Adaptive Nearest Neighbors. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development (ICTD '19)*.

[32] Lovish Madaan, Ankur Sharma, and Aaditeshwar Seth. 2019. Supplementary Material: Price Forecasting & Anomaly Detection for Agricultural Commodities in India. http://bit.ly/2FaTP7F

[33] Sandip Mitra, Dilip Mookherjee, Maximo Torero, and Sujata Visaria. 2013. Asymmetric Information and Middleman Margins: An Experiment With Indian Potato Farmers. *The Bureau for Research and Economic Analysis of Development (BREAD) Working Paper* (2013).

[34] Press Trust of India. 2017. Crackdown on hoarding: Taxmen raid 7 onion traders in Nashik. Retrieved from Business Standard https://goo.gl/Q6qwDv.

[35] Erkki Oja, Kimmo Kiviluoto, and Simona Malaroiu. 2000. Independent component analysis for financial time series. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000.* IEEE, 111–116.

[36] Elisa Oreglia. 2013. When Technology Doesn'T Fit: Information Sharing Practices Among Farmers in Rural China. In *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers - Volume 1 (ICTD '13).* ACM, New York, NY, USA, 165–176. https://doi.org/10.1145/2516604.2516610

[37] PTI. 2017. Onion export jumps 56 percent in Apr-July, but India now importing. Retrieved from The Economic Times https://goo.gl/oCc4pM.

[38] KG Sahadevan. 2012. Commodity Futures and Regulation: A Vibrant Market Looking for a Powerful Regulator. *Economic and Political Weekly* (2012), 106–112.

[39] Robert H Shumway and David S Stoffer. 2000. Time series analysis and its applications. *Studies In Informatics And Control* 9, 4 (2000), 375–376.

[40] Nath Srinivas. 2014. The solution to India's onion price inflation is an obvious one. Hint: it's not the hoarders. Retrieved from Quartz India https://goo.gl/TVyYec.

[41] Janaki Srinivasan and Jenna Burrell. 2013. Revisiting the fishers of Kerala, India. *ACM International Conference Proceeding Series* 1, 56–66. https://doi.org/10.1145/2516604.2516618

[42] Sandip Sukhtankar. 2012. Sweetening the deal? political connections and sugar mills in India. *American Economic Journal: Applied Economics* 4, 3 (2012), 43–63.

[43] A. Vaidyanathan. 2016. Farm loan waiver: a closer look and critique. Retrieved from The Hindu-BusinessLine https://goo.gl/6y4Vao.

[44] Owen Vallis, Jordan Hochenbaum, and Arun Kejariwal. 2014. A Novel Technique for Long-Term Anomaly Detection in the Cloud. In *HotCloud 2014*.

[45] Matt Ziegler, Lokesh Garg, Shailesh Tiwary, Aditya Vashistha, and Kurtis Heimerl. 2019. Fresh insights: user research towards a market information service for bihari vegetable farmers. 1–11. https://doi.org/10.1145/3287098.3287115