Time Series Analysis: the Avocado Prices

Junyi Liao

20307110289@fudan.edu.cn

1. Introduction.

Dataset. For this project, I chose the Avocado Prices (2020 updated version) Dataset from Kaggle, which is publicly accessible via `https://www.kaggle.com/datasets/timmate/avocado-prices-2020`. This dataset is compiled from the Hass Avocado Board (HAB) data, and contains weekly collected data on avocado prices and sales volume in multiple cities, states, and regions of the US from 4 January 2015 up to 29 November 2020.

The analytical data I drew from the dataset is a time series of length 306, containing the average price and volume of conventional avocados at the national level. Baseline covariates of analytical data are listed below.

- `date`: Date of observation, from 4/1/2015 to 29/11/2020;
- `average_price`: Average price of a single avocado;
- `total_volume`: Total number of avocados sold;
- `type`: Type of avocados (conventional/organic). Only conventional is included in analytical data;
- `year`: The year of observation;
- `geography`: Geographic region of observation. Only total U.S. is included in analytical data.

Overview. The remaining part of the article is organized as follows. In section 2, I conduct some exploratory data analysis to study the overall characteristics. In section 3, I build ARIMA models to fit the avocado price series to grasp its change rule. Forecasting of avocado price is carried out in section 4, where different models including ARIMA and LSTM are applied. Some further discussions are given in section 5.

2. Exploratory Data Analysis. I first constructed a time plot of the analytical data. To see any trend or seasonal behavior of the data, I also applied Lowess smoothing in the plot. The result is shown in Figure 1. As is shown in the plot, from January 4, 2015 to November 29, 2020, the average price of a single conventional avocado varied in $[0.75, 1.65]$. Furthermore, no trend or seasonal pattern is observed in this plot. Lowess smoothing implied that the analytical data is approximately stationary.

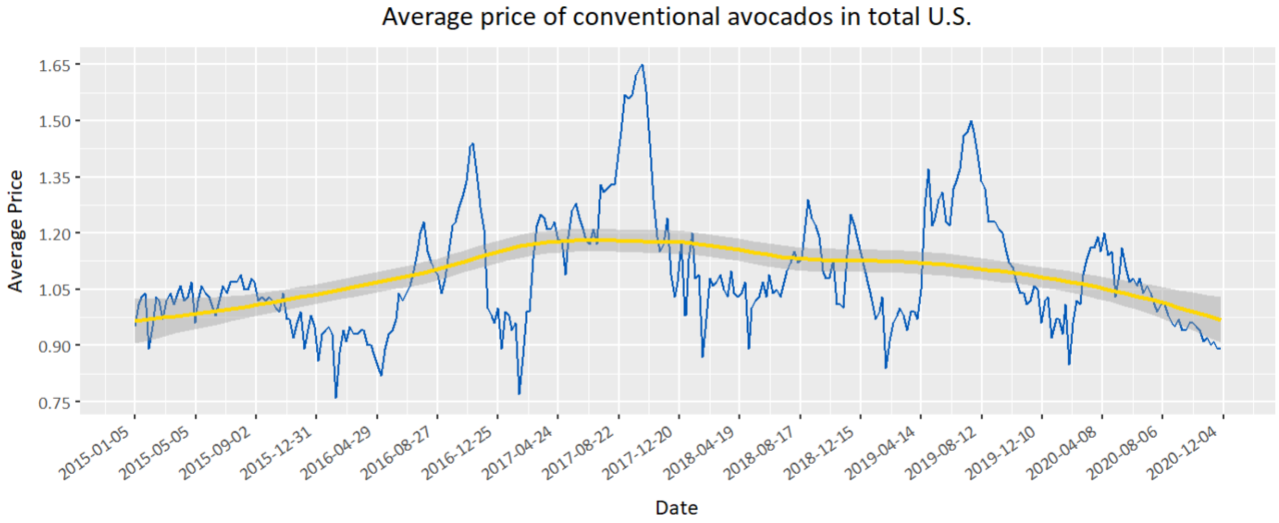


Figure 1. *Avergae price of conventional avocados in total U.S., from Jan 4, 2015 to Nov 29, 2020.*

The left-hand of Figure 2 is a lagged scatterplot that displays the values of the average avocado price X_t on the vertical axis plotted against X_{t-h} on the horizontal axis. This plot shows a strong autocorrelation within the price series. I also noted that the lowess fits are approximately linear, so that the sample autocorrelations displayed in the upper right-hand corner are meaningful.

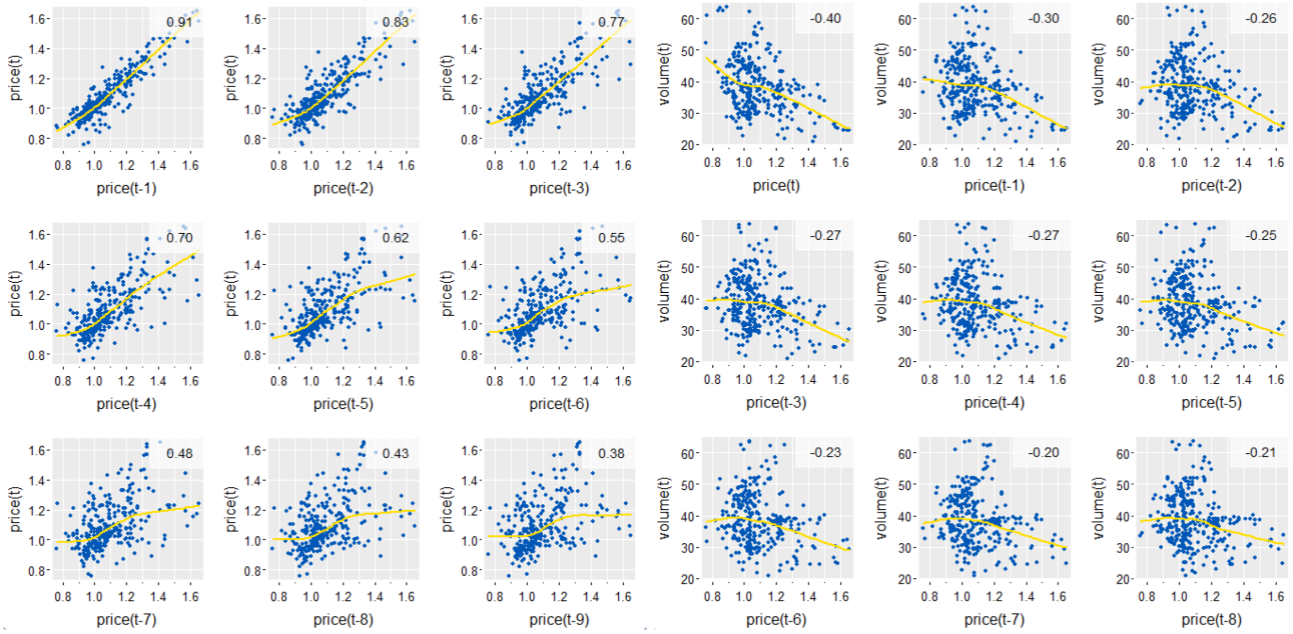


Figure 2. *Left: Scatterplot matrix relating current avocado prices, X_t, to past avocado prices, X_{t-h}, at lags $h = 1, \cdots, 9$; Right: Scatterplot matrix relating current avocado sales volumes, V_t, to past avocado prices, X_{t-h}, at lags $h = 0, \cdots, 8$. The values in the upper right corner are the sample autocorrelations and cross-correlations, and the lines are a lowess fit.*

The right-hand of Figure 2 is another scatterplot of the sales volume series V_t on the vertical axis against the price series X_{t-h} on the horizontal axis. It shows a fairly strong negative correlation between the avocado price X_t and the avocado sales volume V_t, which is consistent with a common sense that a price rise of a certain commodity often causes a decrease in its sales volume. Moreover, the plot of V_t against $X_{t-h}(h = 1, \cdots, 8)$ indicates that the avocado price series might chronologically lead the avocado sales volume series.

3. ARIMA model. Based on the observations in section 2, I applied autoregressive integrated moving average (ARIMA) models to fit the avocado price series. To choose the appropriate orders of ARIMA models, I calculated and plotted the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of the price data. The result is shown in Figure 3.

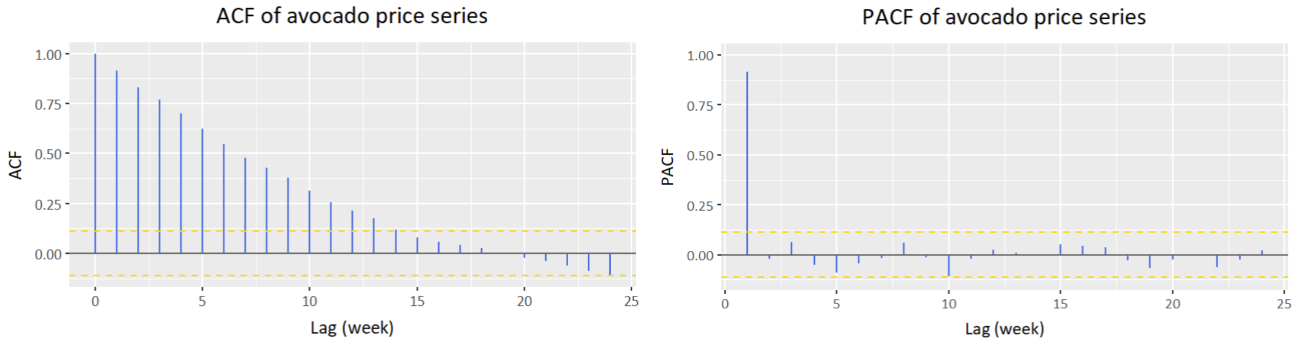


Figure 3. *Estimated ACF and PACF of the average price of conventional avocados in total U.S.*

In Figure 3, one can observe that the ACF of avocado price series decays slowly (tail off), and that the PACF is cut off after lag 1. Such a pattern is consistent with the theoretical properties of AR(1) model. Denote the avocado price series as X_t, the AR(1) model can be formulated as:

$$X_t - \mu = \phi(X_{t-1} - \mu) + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \text{WN}(0, \sigma^2), \tag{1}$$

where μ can be estimated by the first sample moment, and ϕ, σ^2 can be estimated via Yule-Walker equation, or they can be jointly estimated by MLE. I fitted two AR(1) models on the avocado price series using both `ar.yw` and `ar.mle` function in R. The result is shown in Table 1, where two methods gave similar estimations.

To check the rationality of fitted AR(1) model, I ran diagnostics over the model fitted by Yule-Walker estimation using residual analysis. The result is shown in Figure 4.

Table 1. *AR(1) models fitted on the average avocado price series*

Method	Sample size[a]	$\hat{\mu}$	$\hat{\phi}$ (std. error)[b]	$\hat{\sigma}^2$
Yule-Walker Estimation (Method of Moments)	261	1.101	0.9144 (0.0252)	4.472×10^{-3}
Maximum Likelihood Estimation	261	1.099	0.9176 (0.0246)	4.270×10^{-3}

[a] All models were fitted on the training set, including the first 261 observations of the avocado price series. The remaining 45 were used as test set in forecasting task.

[b] The standard error is given by the asymptotic-theory variance matrix of the coefficient estimates.

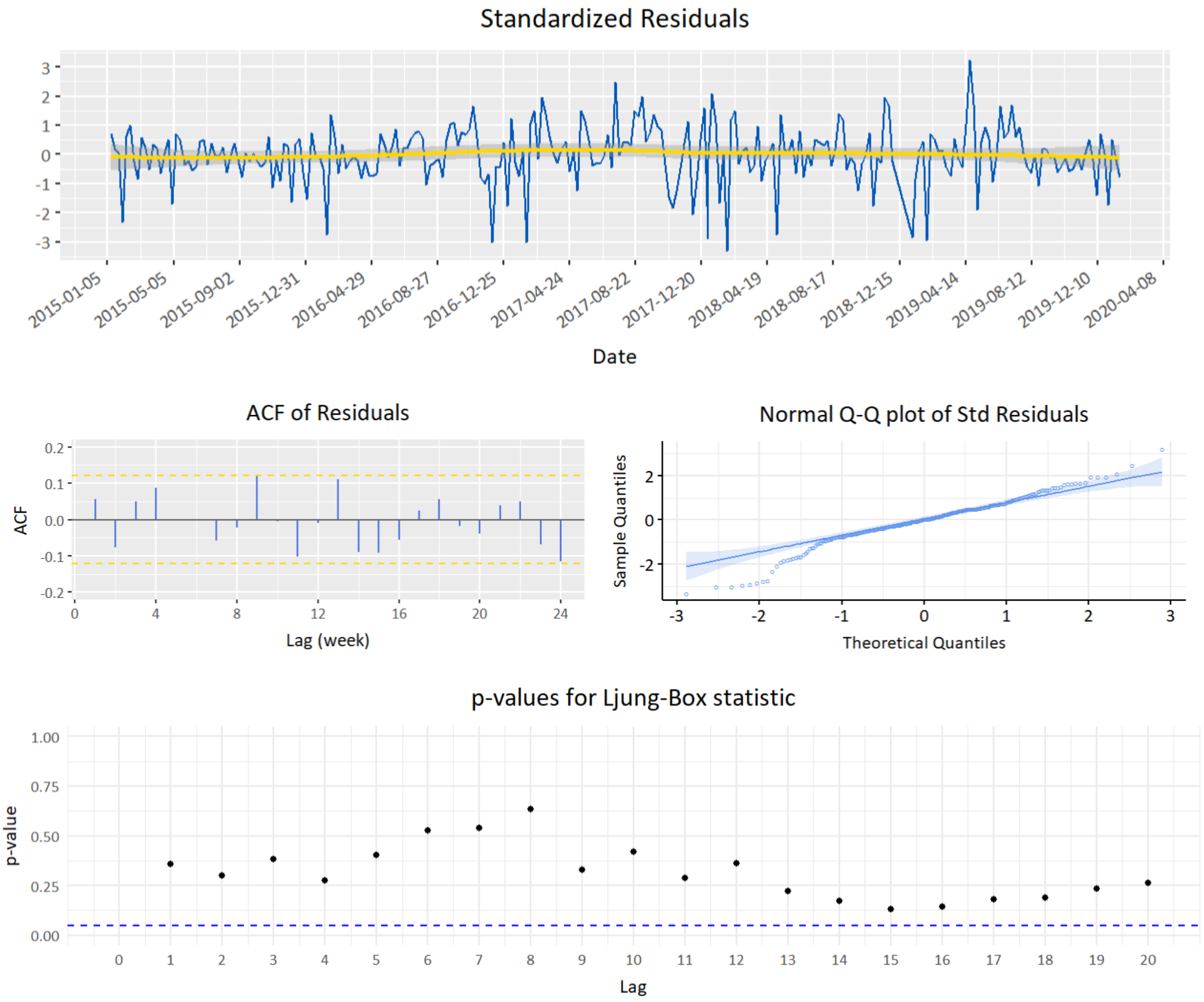


Figure 4. *Diagnostics of the residuals from AR(1) fit on average conventional avocado price*

Table 2. *Model comparison: ARIMA models fitted on the average avocado price series*

Model	No. of Params	SSE	df	MSE	R^2	AIC	AICc	BIC
AR(1)	2	1.114	259	4.303×10^{-3}	0.8424	-4.441	-4.433	-5.414
AR(2)	3	1.111	258	4.306×10^{-3}	0.8429	-4.436	-4.428	-5.395
ARMA(1, 1)	3	1.110	258	4.302×10^{-3}	0.8430	-4.437	-4.429	-5.396
ARIMA(1, 1, 0)	2	1.158	259	4.471×10^{-3}	0.8362	-4.403	-4.395	-5.375
AR(12)	13	1.061	248	4.280×10^{-3}	0.8499	-4.405	-4.391	-5.228

[a] Abbreviations: SSE, error sum of squares; df, degree of freedom; MSE, mean squared error; AIC, Akaike's Information Criterion; AICc, Bias Corrected AIC; BIC, Bayesian Information Criterion.

Figure 4 displays the time plot of the standardized residuals, revealing no discernible patterns. The ACF plot of the standardized residuals indicates no deviation from the model assumptions, and the Ljung-Box-Pierce Q-statistic, employed to detect residual autocorrelation, is non-significant at the depicted lags. The normal Q-Q plot of the residuals suggests reasonable adherence to normality, except for potential outliers.

Furthermore, I fitted different ARIMA models (using MLE) to check more potential candidates. Some metrics of fitted models are reported in Table 2. From this table, one can observe that adding parameters to the ARIMA model increases its complexity, while the goodness-of-fit does not remarkably improve. Among all models listed above, both AIC and BIC are minimized by AR(1), which was the final model I chose.

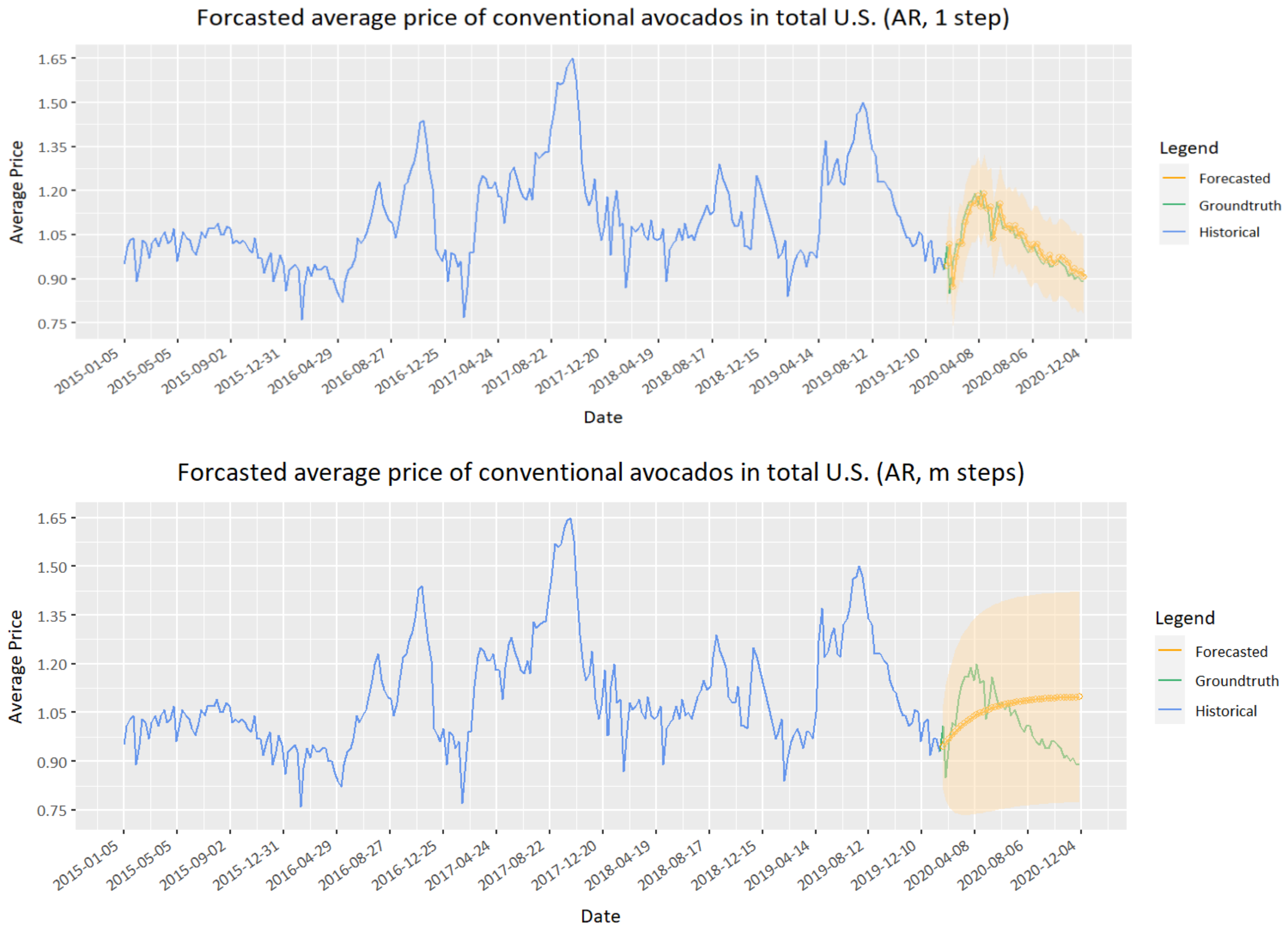
4. Forecasting. In the previous section, I divided the avocado price data to training set (the first 261 observations) and test set (the remaining 45 observations). I built an AR(1) model to fit the training set. The fitted result turned out to be satisfactory, and diagnostics detected no departure of model assumptions. In this section, I applied this fitted model to conducting forecasting. The test set were used as groundtruth to evaluate the accuracy of forecasting result.

To begin with, I investigated one-step forecasting, that is, predicting X_{t+1} conditioning on $\{X_1, \cdots, X_t\}$. In AR(1) model, the best linear unbiased prediction (BLUP) of X_{t+1} and its variance are given by

$$X_{t+1}^t = \phi X_t + (1-\phi)\mu, \quad P_{t+1}^t = \left(1-\rho^2(1)\right)\gamma(0) = \sigma^2, \tag{2}$$

and I plugged in $\{\hat{\mu}, \hat{\phi}, \hat{\sigma}^2\}$ to specify the unknown model parameters. The result is shown at the top of Figure 5, where the forecasted avocado prices appear to be a lagged version of groundtruth. Such a phenomenon can be interpreted by equation (2): since X_{t+1}^t is a linear combination of the previous observation X_t and the expectation μ, it must be close to X_t when ϕ is close to 1 (in this model, $\hat{\phi} = 0.9144$).

Next, I investigated m-step forecasting. Conditioning on $\{X_1, \cdots, X_t\}$, the BLUPs of $X_{t+m}(m = 1, 2, \cdots)$ can be iteratively solved via Innovations algorithm. In AR(1) model, the solution enjoys a very simple form similar to equation (2): $X_{t+m}^t = \phi^m X_t + (1-\phi^m)\mu$.


Forcasted average price of conventional avocados in total U.S. (AR, 1 step)

Forcasted average price of conventional avocados in total U.S. (AR, m steps)

Figure 5. *Top: One-step forecasting result using AR(1); Bottom: Multi-step forecasting result using AR(1), the avocado price in future 45 weeks were forecasted based on the previous 261 weeks. The orange shaded region is a 95% confidence interval of my prediction.*

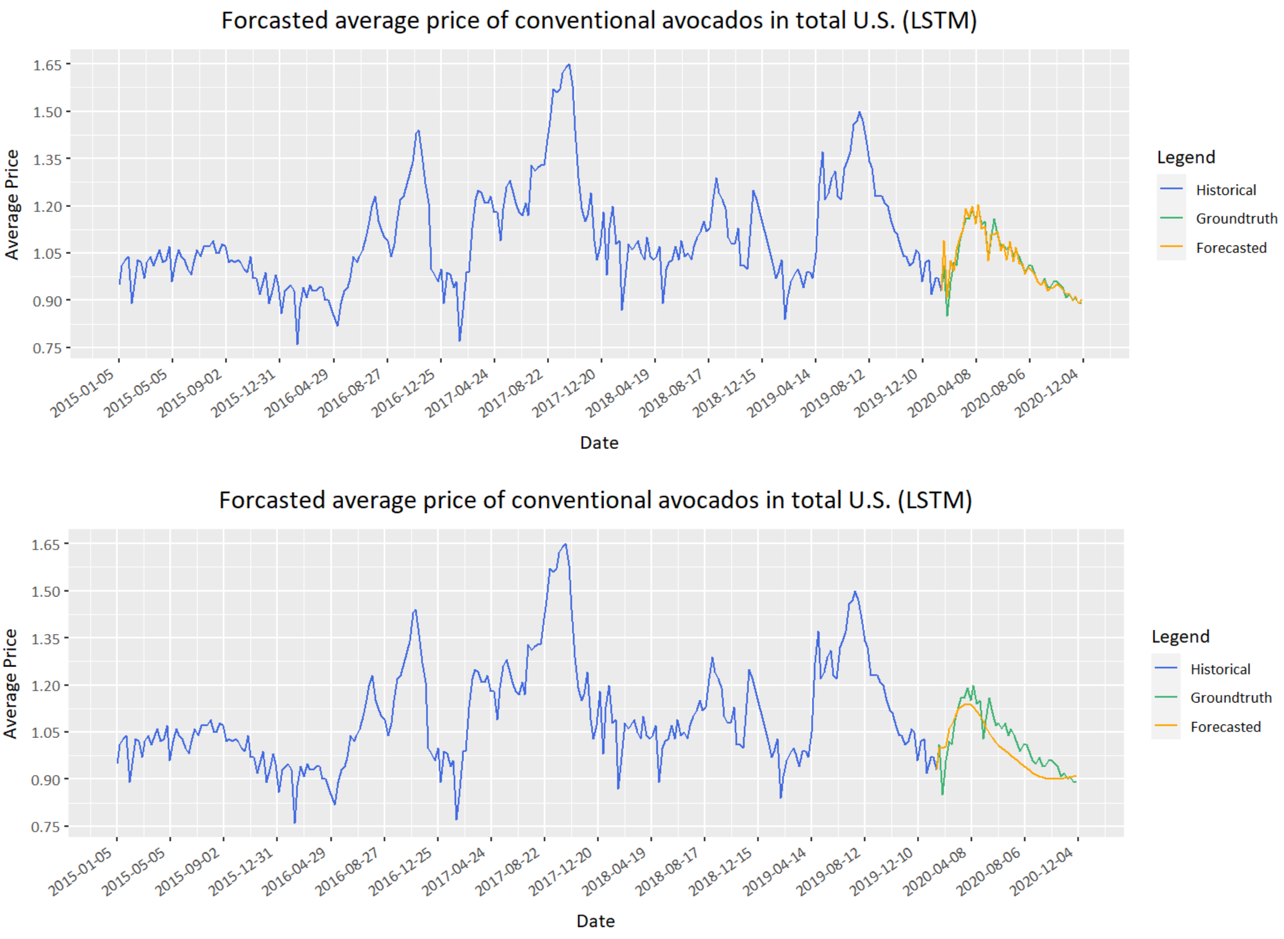
The result of m-step forecasting is shown at the bottom of Figure 5, where the prediction converges to $\hat{\mu}$ exponentially fast as the steps increase. In general, multi-step predictions given by ARIMA model reduce to the expectation μ exponentially fast, and their variance converges to the autocovariance $\gamma(0)$.

I also employed long-short term memory (LSTM) to forecast the avocado price series. I set the input dimension of LSTM to be 6, so that the avocado prices in past 6 weeks will be feed into LSTM in each step. I applied Adam optimizer to train LSTM, with learning rate set as 0.01 and weight decay 10^{-6} .

The one-step forecasting result is shown at the top of Figure 6, where the input of LSTM in each step was groundtruth data. The m-step forecasting result is shown at the bottom, where the input of LSTM in each step is generated by itself in previous steps. As shown in the figure, the avocado price series forecasted by LSTM is much more accurate than by ARIMA.

5. Discussion. In this project, I conduct my analysis on the Avocado Price Dataset. Exploratory data analysis revealed that the avocado price data were approximately stationary with a strong correlation. According to the estimated ACF and PACF, I built an AR(1) model to fit the average avocado price data, and the fitting result turned out to be satisfactory. However, the one-step forecast given by this model is inflexible, while the multi-step forecasts decay to the expectation monotonically and exponentially fast. Hence the predictive power of this model is weak. Possible reasons include: (i) the AR(1) model applied in this example is oversimplified, with only 2 learnable parameters (μ and ϕ), hence it is incapable to capture some complex features in the data; (ii) in the prediction stage, the BLUPs given by AR(1) only make use of information from the most recent previous observation and neglect a large proportion of historical data. Given these defects, AR(1) is not applicable for this forecasting task.

Additionally, I construct an LSTM network to predict the average avocado price in future weeks, and the result is much more precise than that given by ARIMA models. The LSTM has a cell state that memorizes information from historical inputs, and I also manually set the input to be the observations in the most recent 6 weeks, so that LSTM con exploit historical data. Moreover, the number of parameters in LSTM is much larger (about 200 in this example), hence it can learn more complex features in the avocado price data. With these improvement, the forecasting result becomes much more accurate.


Forcasted average price of conventional avocados in total U.S. (LSTM)

Forcasted average price of conventional avocados in total U.S. (LSTM)

Figure 6. *Forecasting result using LSTM; Top: One-step forecasting, Bottom: Multi-step forecasting.*