Special Projects in Econ Research

Comparing Econometrics and Machine Learning Models for Agricultural Price Prediction

Arif Akkan- NYU

1 Introduction and Literature Review

This project is based on the agricultural prices from the city of Antalya in Turkey. Antalya is an agriculturally important city with a big population that supplies both local, national, and international demand for certain agricultural products. The project uses price time series, combined with exogenous variables to predict prices from future periods. Exogenous variables are selected based on their potential effects on supply. Analyzed products are selected based on the production quantity, commercial importance, and data availability. One of the characteristics of this project is that it focuses on cash crops that are traded commonly in markets. Staple agricultural goods prices are commonly studied in the literature. However, focusing the analysis on cash crops are important from two aspects. Firstly, Turkey does not have future markets for cash crops, whereas there are future trading markets for other goods such as Wheat and Barley. Therefore, there is no common tool that market participants are able to use to forecast prices. Secondly, combined with the lack of future markets, there are no important regulations for the price variations in these goods, and market dynamics are important in setting prices. Therefore, supply-side based strategy of this project is expected to be able to achieve a certain prediction capability given the current settings.

Predictions with time series data is a common area of study in Economics. While the analysis objectives are centered around inference in econometrics, prediction problems are also studied for their practical and theoretical contributions. Studies surrounding agricultural prices are also a part of this literature. Some studies include Mutwiri (2019) that uses time series data with SARIMA models for tomato prices, using SARIMA models for Luo et al. (2013) for Cucumber prices. There is also an important literature of agricultural price prediction using weather based exogenous variables. Yoo (2015) and Yang et al. (2022) uses Vector Auto Regression methods with weather variables to forecast agricultural prices. This project takes roots from these econometric papers for the ARMA based baseline model.

Time Series prediction problems are also commonly studied with machine learning models for theoretical and practical conclusions. There is a growing literature using new methods for feature engineering and new models that were able to achieve favorable performance. Studies such as Shengwei et al. (2017) and Ribeiro and dos Santos Coelho (2020) have used models based on support vectors, random forests, and neural networks to predict agricultural prices.

There is also an important literature in agricultural economics that analyzes supply factors. While works of Welch et al. (2010); Fox et al. (2011); Potopová et al. (2017) are not similar to this project in terms of objectives and settings, they provide an important guidance with the methods for engineering weather variables such as temperature metrics and precipitation.

Aim of this project is to compare conventional econometric methods with machine learning models and find the best hypothesis for the model that can give the best prediction results. Models will be compared based on their validation test scores. Machine learning models are developing rapidly and used increasingly within the econometric studies. The settings used in this study aims to test the performance of commonly used models and how they can improve conventional methods used in econometrics. Paul et al. (2022) and Chen et al. (2021) were influential papers that guided the comparison process shows the performance of the machine learning methods, especially neural

network and random forest based models.

This paper aims to contribute to the literature with two aspects. Firstly, while similar studies exists in different part of the worlds, agricultural prices in Turkey is not an explored area in econometrics and machine learning literatures. This project aims to bring practical insights for the settings, data collection, and results in this context. Secondly, compared to the literature, this model aims to explore aspect of regularization by including a high number of exogenous variables. Results of the analysis process not only can help with model selection in this context, but also can give ideas about the important variables with predictive power.

2 Data

Data used in this project spans from 2011 to the end of 2019. This range is selected based on the data quality and availability.

2.1 Price Data

This study is based on the agricultural product price data from Antalya province. Monthly average prices are collected from Turkish Statistical Institute(TSI) for the 2011-2022 period. Datasets are publicly available and can be accessible through Biruni interface of TSI.

Three products are selected for the project: 'Tomato', 'Cucumber', 'Eggplant'. These products are supplied across the year even though with seasonality, commercially important for Antalya city, and have the best price data quality.

2.2 Weather Data

This study utilizes exogenous variables about weather and other important factors in agricultural production. The weather datasets are purchased from OpenWeather as 'History Bulk' packages for the locations. Raw weather datasets include hourly temperature, precipitation, humidity, and weather type variables. For the settings of this project, these variables are aggregated over each month.

Data used in the analysis contain weather variables for 4 different cities. On top of the analysis location of Antalya, temperature variables for the cities of Bursa, Izmir, and Manisa are also used. The justification for these exogenous variables comes from the supply factors. These cities are selected for the similarity of agricultural production patterns and geographical proximity of these cities. Within the competitive markets of cash crops in Turkey, these cities are in a distance to have different weather variables, while their supply can compete with the supply of Antalya. Therefore, a variation in their production is expected to affect the prices in Antalya. Model results will also be an evaluation for this expectation.

For temperature, several variables are created to represent the important aspects for agricultural production. Firstly, mean temperatures are calculated using hourly observations, and their comparison to the monthly expectations are kept. Secondly, maximum and minimum temperatures within

each day are averaged for each month and compared to their expectations. These variables are important for agricultural supply as crops have certain ranges of ideal temperatures, and deviations can hinder supply.

For precipitation measures, hourly rain per m2 values are summed over for months, and transformed with yearly differencing for stationarity.For humidity, monthly averages based on hourly values are calculated. Humidity is a percentage value.

2.3 Extreme Months Dummy

To account for unusual events in weather and geography, a dummy variable named extreme is created for each month. This value takes the value of 1 if that month was unexpected and can create chaotic affects due to this. The extremities are defined with 2 ways, temperature and precipitation extremities and exogenous shock.

2.4 Other Exogenous variables

Labor is an important factor in agricultural production, while it is hard to estimate the actual cost to producers as informal labor relations are common. TSI provides average labor costs for Antalya seasonal daily workers who mainly works in agricultural sector. However, this dataset goes back until only 2016. To tackle this problem, mandated minimum wages and seasonal worker wages are compared. Minimum wages predict the official data for seasonal daily worker wages nearly perfectly. Therefore, monthly minimum wages are used to represent labor costs for agriculture. Labor data is used after its seasonal decomposition(See Appendix). This method filters out trend and seasonality from the data, which aligns with our target variable which is log-differenced. Therefore, labor variable effectively represents distributed effects of labor cost increases.

Another source of variation can be changes in the input costs for agricultural products. Monthly producer prices index for agricultural producers is collected from TSI to represent this. However, the data is not available beyond a certain date and TSI resetted the index for calculation during the timespan of this project. To overcome this issue, PPI is compared to USD/TL currency data that is collected from Turkish Central Bank. For the available time frame, USD/TL is a good predictor for the index values. This is reasonable since inputs in Turkey such as machinery, pesticides, fertilizers, and seeds depends on global markets. In line with this result, USD/TL currency rate that is aggregated over each month is used as an exogenous variable.

Lastly, gas Prices corresponding to Antalya province are collected from Turkiye Petrolleri website, and after accounting for missing values, monthly gas price time series are created.

Further details about data collection and transformation can be found in Appendix.

2.5 Descriptive Statistics

- Data Frequency is monthly.

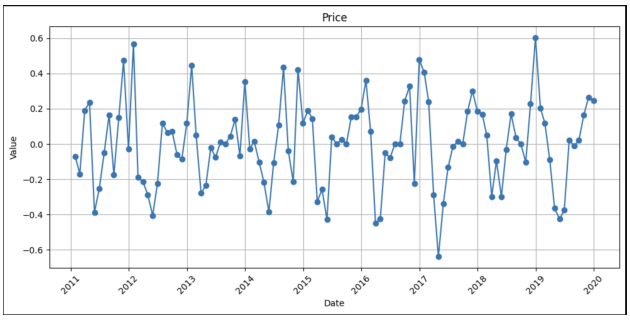


Figure 1: Cucumber Price Variation

	tempdiff	mindiff	maxdiff	raind	hum
count	108.000000	108.000000	108.000000	108.000000	108.000000
mean	0.049725	0.108137	−0.008841	5.518426	61.317920
std	0.078602	0.137654	0.061690	112.734813	7.131347
min	−0.129706	−0.318495	−0.177226	−358.650000	42.244624
25%	0.004410	0.044126	−0.053739	−28.717500	55.882426
50%	0.039769	0.089521	−0.014059	0.365000	61.943408
75%	0.088842	0.166306	0.027821	34.652500	65.662348
max	0.324728	0.650769	0.184472	454.300000	81.068314

Figure 2: Descriptive Statistics for Temperature Variables

- Price units are average monthly prices collected by the official authorities in Turkish Lira. Their final version is monthly change in their logged levels.
- Temperature units are celcius, Precipitation units are mm per sq, humidity is percentage. Other variables such as gas prices, currency rate, and labor wages are in Turkish Liras.

There are 108 observations within the range. For the analysis 106 observations are used. The analysis is based on time series prediction with lagged exogenous variables, therefore, first two months of data is discarded for the analysis. The number of observations are low due to the monthly structure and lack of quality data for longer periods.

Target variable is price. Price variable is created separately for all three agricultural products that are analyzed. Other variables are used as exogenous variables and they do not differ for agricultural products.

The target variable of prices shows significant variance that persists after data transformations. This variations may be reflecting the seasonal nature of the data, combined with the market dynamics.

Temperature variables before the scaling shows significant variation in minimum and maximum values. This variation might be useful with the prediction of prices.

2.6 Settings

This project is designed to forecast future values of agricultural prices, given the last observed price data and exogenous variables. Therefore, all exogenous variables are used as lagged variables corresponding to time t. 2 is selected as the number of lag for exogenous variables. Optimal lag selection is done with domain knowledge.

During the machine learning process, dataset also includes lagged variables for the target variable of price. This is with respect to the autoregressive nature of this project and mirrors the base model based on time series.

3 Methodology

3.1 Error Measures

MSE is selected as the main metric for evaluation. Due to the train validation split, the best estimate of E_{out} will be E_{val}.

The data used in this project is transformed and scaled to have better performance with the models used here. While this does not hinder the quality of results, it is also valuable to understand the practical value of the results, if this model were used. For this goal, a second set of metrics are calculated with best model predictions and scaling and log differencing on the target variable is reversed before error metrics are applied.

3.1.1 Mean Squared Error(MSE)

is the metric for model selection in validation for E_{in} and E_{out} estimates, as it is standard in Machine learning model selection with its sensitivity to outliers and predictive accuracy

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE is Sensitive to Large Errors compared to MAE. For estimating E_{out} error with test sets. However, main output metric is MAE. Data transformations on our target variable caused low variance, which can introduce a bias for MSE-based models.

An additional set of results will also be reported with Mean Absolute Error(MAE), Root Mean Squared Error, (RMSE) Mean Absolute Percentage Error(MAPE), to increase interpretability of the results and provide a comparison for the MSE scores.

3.2 Data methods

3.2.1 Scaling

Standard scaling is applied to all numerical variables, to ensure consistency among scales of different variables and allow models like SVR and XGBoost to converge better.

$$z = \frac{(x - \mu)}{\sigma}$$

3.2.2 Train Test Split

Data is split to Train and Validation sets for this study with a 85-15 % ratio. The time structure is kept during the split. This study does not contain a test set, however, a next step would be collecting data from further periods to test the best model selected here.

While having one validation set approach is not robust to bias as well as moving window validation, due to the limited number of observations it is observed that models cannot be trained well with moving window approach.

Moreover, validation set approach allowed better computational performance for all models to expand parameter space, as well as easy implementation to the base econometric model in this study. Thanks to this, all models can be compared among their validation set metrics directly.

3.3 Model Methods

3.3.1 Regularization

One of the objectives of this study is evaluate how well different models behaves against overfitting in the presence of high number of predictor variables. L1 regularization aims to limit L1 and L2 regularization terms are used with Lasso and Ridge regressions. The regularization terms of SVR and XGBoost models are also used within their model definitions.

3.3.2 Model Selection and Grid Search

Model selection is done with Validation set approach using Grid Search algorithms. The possible parameter space is defined for each model, models are trained with training set and evaluated in the validation set. Parameters with the lowest error metric in validation set is selected as the optimal parameters. Both manual grid search and python packages are used for this process.

Then, Optimal model is selected through comparing validation set errors of different models. Main metric is commonly used Mean Squared Error with its sensitivity to outliers. Other metrics are also kept for each model to see if metrics differ a lot based on parameters.

3.4 Models

Aim of this project is to find the best model hypothesis with respect to settings and data of this project. Econometric models are commonly used in academic literature to forecast future prices.

Machine Learning methods have been shown to perform well with time series data thanks to their ability to regularize and use non-linear forms that can explain the variation in target variables

3.4.1 SARIMAX

Due to the time series nature of this problem, SARIMAX is selected as the baseline model. SARIMAX is a form of Auto-regressive Moving Average (ARMA) model, extended with seasonality terms and exogenous variables. ARMA models and its extensions are used commonly in time series prediction problems. SARIMAX models and ARIMAX models are commonly used in forecasting

problems, with their ability to model time series data while accounting for exogenous variables within linear models. Therefore, it is selected as the baseline model.

Hypothesis 1: Functional form is adapted from the SARIMAX model of Liachoviius et al. (2023) and models in Luo et al. (2013), Alharbi and Csala (2022) and Mutwiri (2019)

$$y_t = \beta X_t + \left(\sum_{i=1}^{p} \phi_i y_{t-i}\right) + \left(\sum_{j=1}^{P} \Phi_j y_{t-js}\right) + \left(\sum_{k=1}^{q} \theta_k \epsilon_{t-k}\right) + \left(\sum_{l=1}^{Q} \Theta_l \epsilon_{t-ls}\right) + \epsilon_t$$

which can also be shown as

$$\Phi_p(B^s) \cdot \Phi_p(B) \cdot (1 - B^s)^{\frac{D}{n}} \cdot (1 - B)^d \cdot y_t = \Theta_Q(B^s) \cdot \Theta_q(B) \cdot \epsilon_t + \sum_{i=1}^{n} X_{i,t} \cdot \beta_i \qquad (1)$$

where:

- X_t represents the exogenous variables and lags at time t with β as their coefficients,
- $\phi_1, \ldots, \phi_p$ are the non-seasonal autoregressive coefficients
- $\Phi_1, \ldots, \Phi_P$ are the seasonal autoregressive coefficients,
- $\theta_1, \ldots, \theta_q$ are the non-seasonal moving average coefficients
- $\Theta_1, \ldots, \Theta_Q$ are the seasonal moving average coefficients
- s is the seasonal cycle = 12,
- Optimal values are searched with grid search based on MSE, to make comparisons with ML models suitable. Range for tuning p,P,q,Q is [0,1,2] which keeps most likely values while limits computational burden. The number of exogenous variables are high for the model, so it is expected to experience overfitting with SARIMAX. Regularization methods are not very well implemented for ARMAX and SARIMAX models.

3.4.2 Lasso Regression

Least Absolute Shrinkage and Selection Operator is a linear regression that includes a regularization term L1. The regularization term effectively zeroes coefficients for the least predictive variables. Minimizing the functional form:

$$(1/(2 * noofsamples)) * ||y - Xw||_2^2 + alpha * ||w||_1$$

Parameter space used for Lasso Regression is alpha: [0.0001, 0.001, 0.01, 0.1, 0.5, 1, 10, 100].

3.4.3 Ridge Regression

Ridge Regression is L2 regularization integrated to linear regression formula. It effectively minimizes the size of coefficients to prevent overfitting. The dataset contains a lot of variables that are correlated, therefore Ridge Regression can provide insights.
Functional form is minimizing

$$||y - Xw||_2^2 + alpha * ||w||_2^2$$

Parameter space used for Ridge Regression is alpha: [0.001, 0.01, 0.1, 0.5, 1, 10, 100,250, 500, 1000]

3.4.4 Support Vector Regression

Support Vector Regression is a form of Support Vector Machines that shows good performance in higher dimensional data. Ability of SVR to use different kernel functions allows model to use different nonlinear functional forms of Radial Basis Functions and Polynomial Regressions as hypotheses. SVR models are also commonly used in forecasting exercises with their practical usage and flexible model space.

Parameter Space used for Support Vector Regression. C: [0.00001,0.0001, 0.001, 0.01, 0.1, 1, 10], epsilon: [0.001, 0.01, 0.1, 0.5, 1, 10], kernel: ['rbf','linear','poly'], gamma : ['scale','auto', 0.1, 1, 10, 100]

- Kernel determines the type of model SVR uses.
- C to effect regularization of the model, finding the balance between performance and overfitting.
- Epsilon ϵ very important for SVR. Determines the boundaries that model will try to achieve. Directly effects overfitting and underfitting.
- Gamma γ to determine the degree of non-linearity allowed for the model.

3.4.5 XGBoost Regression

XGBoost is an ensemble tree-based model, integrated with gradient boosting algorithm. Extreme Gradient Boosting is an ensemble method that combines tree based models with gradient boosting algorithm. Tree Based models have been shown to have good performance with price prediction tasks in the literature reviewed in this report. This model is an extended version of them that is more suitable for a limited data setting.

Parameter Space $maxdepth$: [1,2,3, 5], $minchildweight$: [1, 3, 5, 7], $learningrate$: [0.5], $nestimators$: [10, 20,30, 50,70,100, 150], $alpha$:[0,0.1, 1, 10] $lambda$: [0, 1, 10, 100, 250, 500]

- alpha is Lasso regularization parameter. lambda is Ridge regularization parameter
- number of estimators controls how many trees does the model use. More trees can lead to overfitting. The range is kept conservative, considering the data availability.

- max depth and min child weight determines how deep the trees can go to fit the data. They are also a part of the regularization process.
- Learning rate is kept 0.5 as the mean of its possible range. Because of small number of observations and cross validation approach, finding ideal learning rate is not feasible as algorithm would select the max learning rate.

4 Results

4.1 SARIMAX

4.1.1 Tomato

Best model for Sarimax model is selected with the manual grid search on validation set test errors. Best parameters for the validation set Mean Squared Errors is ((1, 0, 2), (0, 0, 0, 12)) which indicates autoregressive term of order 1, moving average terms of order 2, and no seasonality orders. 1.4556069339018551

4.1.2 Cucumber

Best parameters for the validation set Mean Squared Errors is ((0, 0, 0), (1, 0, 0, 12)) with the 1.3274930114786598.

4.1.3 Eggplant

Best parameters for the validation set Mean Squared Errors is ((0, 0, 0), (1, 0, 0, 12)) 1.4715179419855748

Compared to tomato model, cucumber and eggplant prices use no moving average and one seasonal autoregressive term.

4.2 Lasso Regression

4.2.1 Tomato

Best L1 regularization parameter for the Lasso is was 0.5 with the 0.485 as the validation test error. However, these results were seen to make all coefficients zero in the tomato model.

4.2.2 Cucumber

Lasso regression showed the best performance on validation set with 0.1 alpha parameter, and achieved 1.092054.

4.2.3 Eggplant

Lasso regression showed the best performance on validation set with 0.1 alpha parameter, and achieved 1.128852 Lasso results suggest that the exogenous variables that are used in this study have better predictive power for cucumber and eggplant prices, comparing to tomato prices.

	param_alpha	Training Set MSE	Cross Validation MSE	Training Set MAE	Cross Validation MAE
0	0.000100	0.237110	1.261840	0.382859	0.799867
1	0.001000	0.241313	1.135509	0.393557	0.744964
2	0.010000	0.333966	1.171090	0.468298	0.839485
3	0.100000	0.760865	1.092054	0.682256	0.790700
4	0.500000	0.985831	1.169839	0.774687	0.834821
5	1	0.985831	1.169839	0.774687	0.834821
6	10	0.985831	1.169839	0.774687	0.834821
7	100	0.985831	1.169839	0.774687	0.834821

Figure 3: Lasso CV Results for Cucumber Prices

param_alpha	Training Set MSE	Cross Validation MSE	Training Set MAE	Cross Validation MAE
0.000100	0.231275	2.377062	0.382452	1.250976
0.001000	0.234683	2.172321	0.380399	1.179314
0.010000	0.348120	1.621624	0.440676	1.020963
0.100000	0.818443	1.128852	0.657434	0.838383
0.500000	0.977807	1.221055	0.706021	0.846613
1	0.977807	1.221055	0.706021	0.846613
10	0.977807	1.221055	0.706021	0.846613
100	0.977807	1.221055	0.706021	0.846613

Figure 4: Lasso results in validation set with different regularization

4.3 Ridge Regression

4.3.1 Tomato

Ridge regression showed the best performance on validation set with 1000 alpha parameter, and achieved 0.473671. Considering that 1000 was the maximum value allowed for alpha, the best results are achieved with the maximum regularization.

4.3.2 Cucumber

Ideal alpha parameter for Ridge Regression based on the Validation Set MSE scores is 100.

4.3.3 Eggplant

Ridge regression showed the best performance on validation set with 100 alpha parameter, and achieved 0.954805.

The magnitude of the optimal alpha parameter may be reflecting the multicollinearity among the X columns.

	param_alpha	Training Set MSE	Validation MSE	Training Set MAE	Validation MAE
0	0.001000	0.229440	1.229842	0.383884	0.786014
1	0.010000	0.230073	1.209302	0.383207	0.776827
2	0.100000	0.245547	1.165107	0.397224	0.775476
3	0.500000	0.280979	1.232056	0.426475	0.861671
4	1	0.302804	1.282292	0.441887	0.908135
5	10	0.412844	1.209578	0.512268	0.902057
6	100	0.648424	0.953554	0.633479	0.751859
7	250	0.761923	1.006876	0.684490	0.782087
8	500	0.832141	1.068703	0.714275	0.792897
9	1000	0.884151	1.116452	0.734735	0.805199

Figure 5: Ridge Validation Results for Different Alphas

param_alpha	Training Set MSE	Validation MSE	Training Set MAE	Validation MAE
0.001000	0.221122	2.414300	0.369308	1.266482
0.010000	0.221706	2.361175	0.369167	1.245131
0.100000	0.236730	2.152752	0.377349	1.158966
0.500000	0.271915	1.939541	0.397433	1.065019
1	0.294878	1.807964	0.410292	1.023122
10	0.457488	1.194701	0.496414	0.846327
100	0.729901	0.954805	0.606294	0.771649
250	0.823834	1.033546	0.646528	0.777576
500	0.877951	1.101306	0.669903	0.802259
1000	0.915904	1.152817	0.686608	0.824072

Figure 6: Ridge Regression for Eggplant Prices

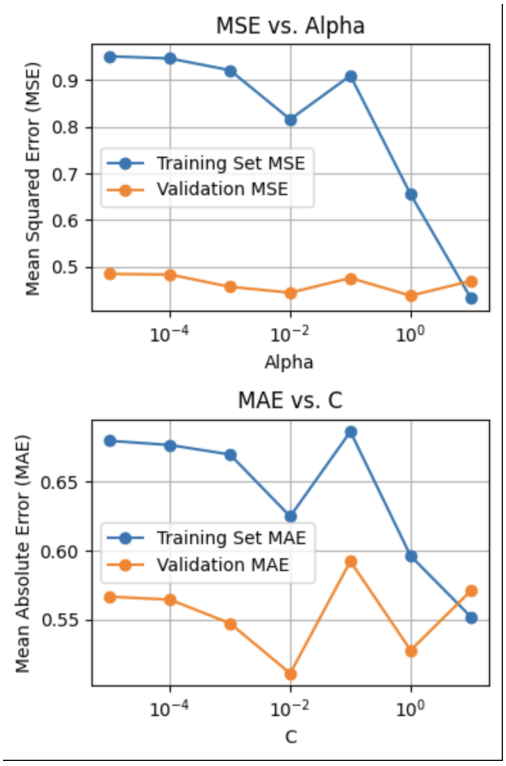


Figure 7: Tomato Prices Training-Validation Errors in Support Vector Regression

4.4 Support Vector Regression

4.4.1 Tomato

Support Vector Regression models showed the best performance with parameters C (1), epsilon (1), gamma ('scale') and kernel ('rbf') with a score of 0.4387539586705724

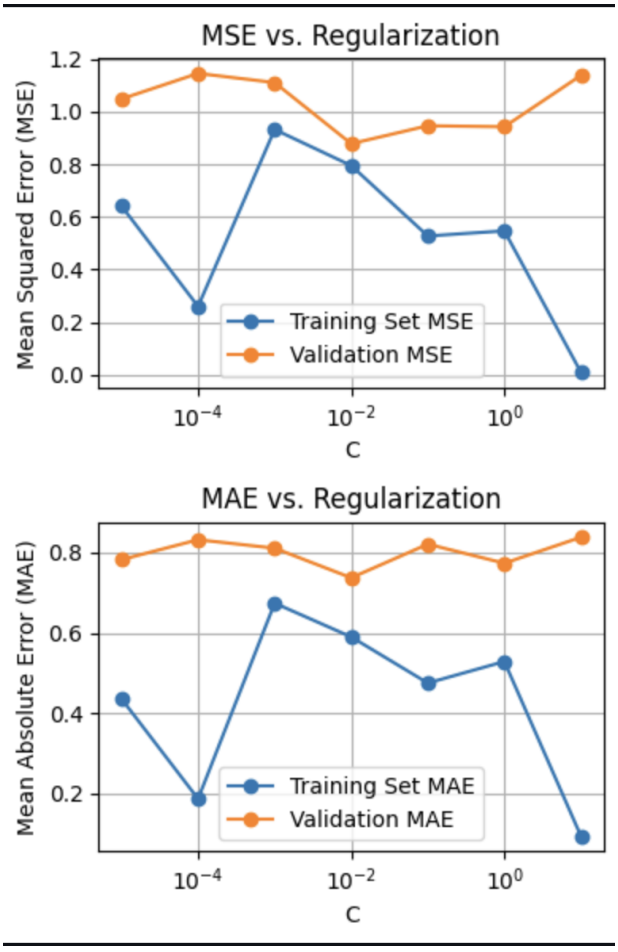
4.4.2 Cucumber

Best parameters for SVR was selected with Validation set MSE errors as C (0.01), epsilon (0.1), gamma(scale), kernel ('linear') with the score: 0.8451137430648522

Figure 8: Eggplant Prices in SVR

param_C	Training Set MSE	Cross Validation MSE	Training Set MAE	Cross Validation MAE
0.000010	0.640936	1.048160	0.435760	0.782182
0.000100	0.259446	1.145251	0.187087	0.831707
0.001000	0.932010	1.110275	0.673973	0.810685
0.010000	0.794735	0.877864	0.589855	0.737002
0.100000	0.527230	0.946169	0.475141	0.820709
1	0.546591	0.942351	0.528728	0.773245
10	0.009226	1.137634	0.094386	0.837899

(a) Eggplant Prices Training-Validation Errors in Support Vector Regression



(b) Eggplant Prices Training-Validation Errors in Support Vector Regression

4.4.3 Eggplant

Best parameters for SVR was selected with Validation set MSE errors as C (0.01), epsilon (0.1), gamma(scale), kernel ('linear') with the score 0.8778639534709995.

It is important to note the selection of linear kernel for Eggplant and Cucumber prices, compared to Tomato. In the graph and table, we can see the tradeoff between training errors and validation errors with regularization parameter for the SVR.

4.5 XGBoost Regression

4.5.1 Tomato

As the optimal model for Tomato Prices, XGBoost model with parameters: alpha (1), learning rate (0.5), max depth (5), min child weight(1), 10 estimators, and lambda regularization parameter(1) is selected with a MSE score of 0.366 on validation test.

4.5.2 Cucumber

As the optimal model for Cucumber Prices, XGBoost model with parameters: alpha (0.1), learning rate (0.8), max depth (5), min child weight(5), 70 estimators, and lambda regularization parameter(100) is selected with a MSE score of 0.697 on validation test.

Best parameters for XGBoost Regression based on the validation set MSE scores are alpha (1), learning rate (0.8), max depth (2), min child weight(5), 20 estimators, and lambda regularization parameter(1) is selected with a MSE score of 0.697 on validation test.

4.5.3 Eggplant

As the optimal model for Eggplant Prices, XGBoost model with parameters: alpha (0.1), learning rate (0.8), max depth (5), min child weight(5), 70 estimators lambda regularization parameter(100) is selected with a MSE score of 0.697 on validation test.

4.6 Model Selection based on Validation Scores

Optimal models are selected by comparing validation MSE scores. To have consistency, SARIMAX models are also evaluated for their predictive performance, rather than the quality of fit, within the validation set. Only difference in the settings are machine learning models also use lagged price variables, which are inherently used in base SARIMAX models.

4.6.1 Tomato

Results show that for the tomato prices time series, XGBoost models shows the best performance, followed by Support Vector Regressions. Baseline SARIMAX models have the worst performance.

The evaluation may be complicated if the data tranformations are reversed, with Support Vector regression having slightly better results than XGBoost in 3 metrics. However, results are close and

Figure 9: Tomato Validation Errors

	Validation MSE	Validation MAE
SARIMAX	1.455607	1.094550
SVR	0.438754	0.527990
XGBoost	0.365987	0.483507
Lasso	0.484526	0.569608
Ridge	0.473671	0.563730

(a) Validation Errors in Tomato Price Prediction

	Val_mse_actual	Val_rmse_actual	Val_mae_actual	Val_mape_actual
SARIMAX	1.060933	1.030016	0.863146	0.386229
SVR	0.362417	0.602011	0.449888	0.195931
XGBoost	0.566325	0.752546	0.531471	0.213518
Lasso	0.447602	0.669031	0.452663	0.175787
Ridge	0.436144	0.660412	0.456680	0.180848

(b) Reverse-Transformed Validation Errors in Tomato Price Prediction

and we can expect XGBoost to achieve similar performance without tranformed data, while support vector regression is expected to have less performance.

As the optimal model for Tomato Prices, XGBoost model with parameters: alpha (1), learning rate (0.5), max depth (5), min child weight(1), 10 estimators lambda regularization parameter(1) is selected.

To evaluate model performance from a practical point of view, we can look at reverse transformed MAPE values, which shows an average of 19.6$(\%)$ absolute percentage error.

4.6.2 Cucumber

Results show that for the cucumber prices time series, XGBoost models shows the best performance, followed by Support Vector Regressions. Baseline SARIMAX models have the worst performance. This performance is sustained when the values are reverse transformed.

As the optimal model for Cucumber Prices, XGBoost model with parameters: alpha (1), learning rate (0.8), max depth (2), min child weight(5), 20 estimators lambda regularization parameter(1) is selected. To evaluate model performance from a practical point of view, we can look at reverse transformed MAPE values, which shows an average of 16.9$(\%)$ absolute percentage error.

4.6.3 Eggplant

Results show that for the Eggplant prices time series, XGBoost models shows the best performance, followed by Support Vector Regressions. Baseline SARIMAX models have the worst performance.

Figure 10: Cucumber Validation Errors

	Validation MSE	Validation MAE
SARIMAX	1.327493	0.964953
SVR	0.845114	0.679446
XGBoost	0.476455	0.530669
Lasso	1.092054	0.790700
Ridge	0.953554	0.751859

(a) Validation Errors in Cucumber Price Prediction

	Val_mse_actual	Val_rmse_actual	Val_mae_actual	Val_mape_actual
SARIMAX	0.382189	0.618214	0.461795	0.254642
SVR	0.365696	0.604728	0.436138	0.237393
XGBoost	0.168504	0.410492	0.301152	0.168641
Lasso	0.612827	0.782832	0.520247	0.244866
Ridge	0.546709	0.739398	0.492920	0.226557

(b) Reverse-Transformed Validation Errors in Cucumber Price Prediction

Figure 11: Eggplant Validation Errors

	Validation MSE	Validation MAE
SARIMAX	1.471518	1.049803
SVR	0.877864	0.737002
XGBoost	0.696942	0.659385
Lasso	1.128852	0.838383
Ridge	0.954805	0.771649

(a) Validation Errors in Eggplant Price Prediction

	Val_mse_actual	Val_rmse_actual	Val_mae_actual	Val_mape_actual
SARIMAX	2.559669	1.599897	1.384276	0.655758
SVR	2.123910	1.457364	0.936240	0.282655
XGBoost	1.669720	1.292177	0.810563	0.238865
Lasso	2.557385	1.599183	1.104192	0.358332
Ridge	2.128267	1.458858	0.958856	0.295853

(b) Reverse-Transformed Validation Errors in Eggplant Price Prediction

This performance is sustained when the values are reverse transformed.

As the optimal model for Eggplant Prices, XGBoost model with parameters: alpha (0.1), learning rate (0.8), max depth (5), min child weight(5), 70 estimators lambda regularization parameter(100) is selected. To evaluate model performance from a practical point of view, we can look at reverse transformed MAPE values, which shows an average of 23.9(%) absolute percentage error.

Optimal models and performances have differences for Tomato and Cucumber and Eggplant

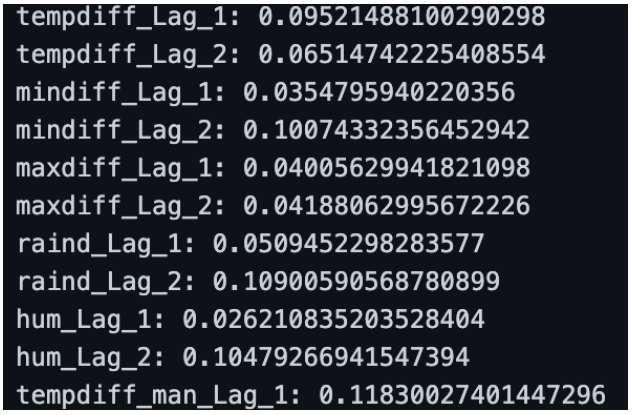


Figure 12: Feature importance for the XGBRegressor model with tomato prices

prices. An important difference is on optimal regularization parameters and number of estimators. This results may suggest that exogenous variables used in this project has significantly different relations with the different products.

A consistent result among models is the success of XGBoost models.

4.7 Results Evaluation

Optimal model and validation error for tomato prices are significantly different than others, which may indicate a different pattern of relations. Results suggest that even though these products are similar and produced in the same area, their price trends are significantly different to alter optimal parameters.

A number of insights can be gathered from these results. This project was a performance comparison against the problems of higher dimensional data and multicollinearity. One of the results was high values for regularization parameters, specifically L2 regularization, hinting to the multicollinearity present in the X variables.

A second insight is about the inclusion of weather variables for comparative cities. When feature importances are examined for the optimal model of XGBRegressor, these features showed an average importance combined with the high importance of the Antalya variables. Therefore, they are still valid predictors to a certain extent regardless of the similarities to the Antalya variables.

5 Discussion

This project aimed to find the best models to be used in agricultural price prediction with supply side variables, in the case of Antalya city prices. Conventional ARMA-based econometric method is compared with machine learning model. The striking difference for this project is utilizing a high number of exogenous variables.

Model Selection period showed the importance of regularization terms and other parameters against overfitting and generalization/approximation tradeoff. Validation test results indicated models that are better designed to handle with higher number of variables while allowing nonlinear relations performed the best. Given the current settings and datasets, it can be concluded that XGBoost model form is the optimal hypothesis for this prediction problem.

Additional comparisons of reverse-transformed prediction errors bring the advantage of interpretability for the models. Optimal models were able to predict actual prices with an average percentage error between 16.9 $(\%)$ and 23.9 $(\%)$. These results most likely suffers from the bias of the validation approach and out of sample errors would be higher. While these prediction errors are not perfect, they are much favorable compared to the baseline models. It is also important to notice focus on only supply side variables for the agricultural price prediction. Moreover, these results can guide a future project that has access to more granular datasets with daily and weekly frequencies and higher number of observations.

Conventional methods are well-studied and have good performance with time series problems. However, advancements in the digitalization and computational power not only allows more detailed datasets, but also allows the use of machine learning methods that can utilize these datasets to achieve better performances.

These results are economically important to show how exogenous variables can be used to forecast agricultural prices. Machine Learning methods can be applied by policymakers in agricultural policy to forecast prices regularly. Models that rely on more granular data in longer timespans can provide good forecasts which is favorable for the policy decisions about any regulations, storage decisions and any policy that can hedge agricultural producers against price fluctuations. Moreover, with the growing concerns of climate change, unexpected weather patterns are expected to be more

common. Insights from these models can be used with synthetic climate change scenarios to guess short to long term effects of climate change.

5.1 Limitations and Further Steps

This study faces several limitations that depends on the data availability and time constraints. While the range of exogenous variables can be extended, finding consistent price data about agricultural products is challenging. Official records for price data are not consistent until 2011. Moreover, the Covid-2019 pandemic, starting in 2020 is a major disruption to supply and demand of agricultural products that would create a lot of noise for this project. Therefore, the project was limited to the 2011-2019 timespan.

These limitations create several shortcomings in the analysis. Firstly, due to limited data, using time-series based validation approaches or leaving aside a test set compromise the performance in model selection. Therefore, the estimates for Eout may suffer from bias.

Another shortcoming was about model space. In the recent years, researchers showed that Neural Network models and various ensemble methods have good prediction performance for time series with exogenous variables. While these developments are not ignored, this project was limited by computational power and time availability to effectively use those models.

A possible shortcoming may be the settings for this research question. Capturing price variation is an important problem. However, it is possible that the markets fluctuate less in monthly periods compared to daily and weekly periods as the correction mechanisms may work fast. Therefore, model performance may increase with a more detailed price dataset.

A possible solution to these problems is increasing data granularity and redoing analysis in a longer period. Daily minimum and maximum price data from wholesale markets are available, however, the data is not well organized and processing it requires time that is beyond the limits of this project.

6 Bibliography

Akbas (2024). Important information. Akbas. Accessed: 2024-05-11.

Alharbi, F. R. and Csala, D. (2022). A seasonal autoregressive integrated moving average with exogenous factors (sarimax) forecasting model-based time series approach. *Inventions*, 7(4).

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica*, 47:1287–1294.

Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., and Liew, X. Y. (2021). Automated agriculture commodity price prediction system with machine learning techniques. *ArXiv*, abs/2106.12747.

Fox, J. F., Fishback, P. V., and Rhode, P. W. (2011). The Effects of Weather Shocks on Crop Prices in Unfettered Markets: The United States Prior to the Farm Programs, 1895-1932. In *The Economics of Climate Change: Adaptations Past and Present*, NBER Chapters, pages 99–130. National Bureau of Economic Research, Inc.

Hektaş (2024). Domates yetiştiriciliğine dair her Şey. Hektaş. Accessed: 2024-05-11.

Liachoviius, E., abanovi, E., and Skrickij, V. (2023). Freight rate and demand forecasting in road freight transportation using econometric and artificial intelligence methods. *Transport*.

Luo, C. S., Zhou, L. Y., and Wei, Q. F. (2013). Application of sarima model in cucumber price forecast. 373:1686–1690.

Mutwiri, R. M. (2019). Forecasting of tomatoes wholesale prices of nairobi in kenya: Time series analysis using sarima model. *International Journal of Statistical Distributions and Applications*, 5(3):46–53.

Paul, R. K., Yeasin, M., Kumar, P., Kumar, P., Balasubramanian, M., Roy, H. S., Paul, A. K., and Gupta, A. (2022). Machine learning techniques for forecasting agricultural prices: A case of brinjal in odisha, india. *PloS one*, 17(7):e0270553.

Potopová, V., Zahradníček, P., Štěpánek, P., Türkott, L., Farda, A., and Soukup, J. (2017). The impacts of key adverse weather events on the field-grown vegetable yield variability in the czech republic from 1961 to 2014. *International Journal of Climatology*, 37(3):1648–1664.

Ribeiro, M. H. D. M. and dos Santos Coelho, L. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied Soft Computing*, 86:105837.

Shengwei, W., Yanni, L., Jiayu, Z., and Jiajia, L. (2017). Agricultural price fluctuation model based on svr. In *2017 9th International Conference on Modelling, Identification and Control (ICMIC)*, pages 545–550.

Sputnik (2015). Russian sanctions killed tourist business in antalya. Sputnik Globe. Accessed: 2024-05-11.

Türkiye Meteoroloji Genel Müdürlüğü (2024). Weather forecast for antalya. Türkiye Meteoroloji Genel Müdürlüğü. Accessed: 2024-05-11.

Türkiye Petrol Ürünleri İşverenler Sendikası (2024). Geçmiş akaryakıt fiyatları. Türkiye Petrol Ürünleri İşverenler Sendikası. Accessed: 2024-05-11.

Welch, J. R., Vincent, J. R., Auffhammer, M., Moya, P. F., Dobermann, A., and Dawe, D. (2010). Rice yields in tropical/subtropical asia exhibit large but opposing sensitivities to minimum and maximum temperatures. *Proceedings of the National Academy of Sciences*, 107(33):14562–14567.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Yang, H., Cao, Y., Shi, Y., Wu, Y., Guo, W., Fu, H., and Li, Y. (2022). The dynamic impacts of weather changes on vegetable price fluctuations in shandong province, china: An analysis based on var and tvp-var models. *Agronomy*, 12(11).

Yoo, D.-i. (2015). Developing forecasting model of vegetable price based on climate big data. (330-2016-13943).

7 Appendix

7.1 Data Sources

Price datasets are collected from the Turkish Statistical Institute's geographical data service of Biruni. Data is publicly accessible and contains officially collected price averages for each month and city combination for commercial agricultural goods.

Weather Datasets for Antalya and other cities of Izmir, Manisa, and Bursa are collected from Open Weather with History Bulk Buy. These datasets provide consistent weather variables in a period spanning more than 40 years and provide hourly information about temperatures, precipitation, humidity, and weather events. This historical dataset needs to be purchased, but it is accessible with small costs. Other free weather data services are checked for this project, however, data quality was not sufficient.

Weather expectation data are collected from the Official Meterology Service of Turkey.(Türkiye Meteoroloji Genel Müdürlüğü (2024)) They include monthly normals for mean, daily maximum and daily minimum values for each city.

Currency exchange prices are collected from Turkish Republic Central Bank. Minimum wage dataset is collected from The Ministry of Labor and Social Security. Seasonal labor cost averages are collected from Turkish Statistical Institute.

Gas Prices corresponding to Antalya province are collected from Turkiye Petrolleri website, one of the major oil companies, for the period of 2011-2023. (Türkiye Petrol Ürünleri İşverenler Sendikası (2024)) The data is publicly available. End-user gas prices in Turkey are highly regulated and market is competitive, therefore, it is assumed that trends in TP data reflects the trends in the average oil prices in Antalya province. To support this assumption, oil prices are regressed against global crude oil prices and USD/TL currency, where nearly all variation can be predicted.

7.1.1 Extreme Months Dummy

To account for unusual events in weather and geography, a dummy variable named extreme is created for each month. This value takes the value of 1 if that month was unexpected and can create chaotic affects due to this. The extremities are defined with 2 ways, temperature and precipitation extremities and exogenous shock. For weather related extremities, very hot, cold, dry, and wet months are found compared with the full weather dataset spanning 40 years. This method has some similarity with the expectation comparisons for weather variables. However, an important aspect for this study justify it, the expectation difference method calculates a linear value in a scale. This method is used to represents very unusual months, that has a particular importance for agricultural production. Exogenous shock that is added for this study is the international events and tourism ban of Russia in 2015 Summer. (Sputnik (2015)). Antalya is a major tourist hub, and Russian tourists are an important source of demand during the summer, effectively doubling the population in some parts of the city. This shock is selected because it affects mainly the Antalya province with the main destination for this tourist group, as other international events are reflected with currency exchange rate.

One of the initial ideas for extreme month dummy was to use weather types observed. However, having a threshold to determine a month with extreme events based on categorical weather description involved a number of assumption, and ignored for this project.

7.2 Missing Data

The span for this project is selected to have the best quality of data with minimum missing values. However, some information is hard to retrieve historically. With the lack of official commercial gas price data, gas prices are collected from a later privatized major gas company. The price data has missing values as the date goes back. Forward filling, which means the missing values are filled with the last observed data point, is used to accommodate for this problem. The assumption here was that due to the regularized nature of the gas price industry, main price changes occur after the changes in regulations and global oil prices. This assumption is tested with crude oil prices and currency exchange rate, and it was shown that nearly all variation can be explained. Using an interpolation method with a consistent trend would not be more suitable as we would assume prices change regularly. This assumption seemed less inconvenient, given that monthly data is used for this project and nearly each month has an observation.

7.3 Data Tranformations

7.3.1 Log Differencing

For target variable of price, precipitation measures, currency exchange rate and gas prices, log differencing is used. This method first takes the logarithms of the values to effectively scale them, then takes their difference from the nth observation. For price, exchange rate, and gas prices, the difference is monthly. For precipitation metrics, the difference is yearly with the 12 month period, which accommodates the seasonality better.

7.3.2 Seasonal Decompose

Labor price data has unusual characteristics, as minimum wage is used as a proxy for seasonal wages. Minimum wage changes every 6 or 12 months in Turkey. This pattern makes it harder to have stationary labor cost series. To accommodate this problem, seasonal decomposition method is used with python packages. This method estimates the trend with a convolution filter, a weighed moving average method, and then estimates the seasonality from detrended values. The residuals are the final form.

7.3.3 Expectation Differencing

For the temperature variables, we are utilizing the expected temperature values. The expected weather data is collected from the Official Meterology Service of Turkey for each city and contains average mean, average minimum, average maximum temperatures for each month. Using this dataset, each temperature variable is compared to its monthly expectation by taking the difference and dividing by the expectation. This process standardizes the variables, makes time series stationary as it addresses seasonality and may increase the predictive quality by providing a better variation tool for temperatures.

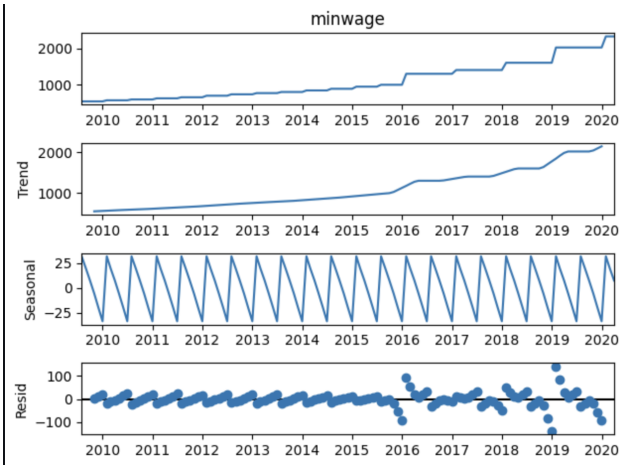


Figure 13: LaborSeasonality

7.4 Stationarity

All X and Y variables that are used in this project are stationary. Non-stationary variables are transformed to have stationary forms. Stationarity checks are done with Augmented Dickey-Fuller Test. ADF tests the null hypothesis of whether a unit root is present in the time series.

7.4.1 Optimal Lags-Domain

Optimal lag for variables outside of SARIMAX parameters was selected as two. While this decision is not exhaustive of all options, it is reasonable and convenient given the computational and time constraints. Sources referred includes Hektaş (2024) Akbas (2024), and an interview with an agricultural engineer.

7.5 Heteroscedasticity

One of the concerns for fitting Ridge and Lasso regressions is the possibility of heteroskedastic errors. These errors may persist even though data was transformed due to the time series nature and necessitate the use of methods like Generalized Least Squares. Two tests are employed to check for heteroskedasticity, White test and Breusch-Pagan that was laid out on the work of White (1980) and Breusch and Pagan (1979). Both tests utilize an auxiliary regression on the residuals of the data. Test results showed no heteroskedasticity with large p values for the data used in this project, with full rank matrices with no perfect multicollinearity. In the light of test results, homoskedastic errors are assumed for Lasso and Ridge Regression models.

7.6 Descriptive Stats Appendix

Name	X or Y	Variable	Transformations	Stationary	Sources
Average Monthly Oil Prices	X	gasprice	Data goes until 2011. Missing days are filled with forward fill method, then averaged for each month. Stationary its ensured with log differencing	Stationary after transformation	Turkiye Petrolleri website
Average Labor Costs	X	labor	Minimum wage is used as a proxy for agricultural wages, as it explains seasonal agricultural worker wages well. Value for each month is created. For Stationarity, seasonality and trend is removed to get residuals	Stationary after transformation	TR Labor Municipality
Antalya Temperature	X	tempdiff	Monthly temperature averages are calculated. Then, data for the baseline temperature for each month is collected from official meteorology authority and difference is calculated and normalized for each month	Stationary after transformation	OpenWeather, Turkish State Meteorological Service
Measures for Hot and Cold Days	X	mindiff, maxdiff	Max and Min Temp For everyday is averaged for each month. Then, data for the baseline temperature for each month is collected from official meteorology authority and difference is calculated and normalized for each month	Stationary after transformation	OpenWeather
Other Cities Temperature	X	tempdiff_i (i for cities)	Agriculturally important cities with similar product structures to Antalya. Same procedure as the Antalya temperature	Stationary after transformation	Open Weather, Turkish State Meteorological Service
Antalya Rain	X	rain	Monthly precipitation totals are calculated, then their yearly change is calculated, which is stationary	Stationary after transformation	OpenWeather
Other Cities Rain	X	rain_i (i for cities)	Monthly precipitation totals are calculated, then their yearly change is calculated, which is stationary	Stationary after transformation	OpenWeather
Humidity	X	hum	Monthly average for humidity is calculated from hourly data,	Stationary after transformation	OpenWeather
Dollar/TL Currency	X	usd	Rather than using producer price index with extrapolating, Dollar/TL is used as a proxy as it mostly explains variation.	Stationary after transformation	Turkish Central Bank, Turkish Statistical Institute
Extreme Month Dummy	X	extreme	Months are assigned dummies based on extreme conditions such as very hot or rainy months, months that extreme weather types occurred		OpenWeather
Monthly Prices for Tomato, Eggplant, Cucumbers	Y	price_j (J for goods)	Stationary is ensured with log differencing	Stationary after transformation	Turkish Statistical Institute

Figure 14: Data Table with explanation, transformations, and sources

	tempdiff	mindiff	maxdiff	raind	hum
count	108.000000	108.000000	108.000000	108.000000	108.000000
mean	0.049725	0.108137	-0.008841	5.518426	61.317920
std	0.078602	0.137654	0.061690	112.734813	7.131347
min	-0.129706	-0.318495	-0.177226	-358.650000	42.244624
25%	0.004410	0.044126	-0.053739	-28.717500	55.882426
50%	0.039769	0.089521	-0.014059	0.365000	61.943408
75%	0.088842	0.166306	0.027821	34.652500	65.662348
max	0.324728	0.650769	0.184472	454.300000	81.068314

Figure 15: Descriptive Statistics for Temperature Variables

	tempdiff_bur	mindiff_bur	maxdiff_bur	raind_bur
count	108.000000	108.000000	108.000000	108.000000
mean	0.102845	0.333162	-0.006107	-3.351944
std	0.189764	0.513694	0.120422	45.398875
min	-0.378790	-0.872126	-0.318934	-156.780000
25%	0.007609	0.102445	-0.063372	-25.240000
50%	0.069568	0.176559	-0.018410	-1.070000
75%	0.192590	0.481579	0.047161	25.360000
max	0.808192	2.378178	0.424834	100.180000

Figure 16: Descriptive Stats for Bursa City

	tempdiff_izm	mindiff_izm	maxdiff_izm	raind_izm
count	108.000000	108.000000	108.000000	108.000000
mean	-0.001330	-0.094531	0.029868	-1.858889
std	0.114474	0.181080	0.080115	69.986281
min	-0.440611	-0.796900	-0.211608	-210.190000
25%	-0.037267	-0.140466	-0.009997	-30.907500
50%	0.012141	-0.054297	0.041697	0.625000
75%	0.051446	-0.005399	0.078180	26.490000
max	0.355213	0.327900	0.290369	217.600000

Figure 17: Descriptive Stats for Izmir City

	tempdiff_man	mindiff_man	maxdiff_man	raind_man
count	108.000000	108.000000	108.000000	108.000000
mean	0.049188	0.017231	0.029434	-1.918056
std	0.153804	0.271909	0.105662	43.619865
min	-0.395699	-0.853123	-0.220711	-154.520000
25%	-0.022684	-0.060490	-0.028163	-18.685000
50%	0.019227	0.001410	0.017300	-1.035000
75%	0.110782	0.123553	0.071920	22.675000
max	0.595284	0.831290	0.404918	140.290000

Figure 18: Descriptive Stats for Manisa City

	labor	gasprice	usd
count	108.000000	108.000000	108.000000
mean	-1.895625	0.005831	0.012501
std	33.208373	0.028752	0.035497
min	-140.162236	-0.075764	-0.086562
25%	-12.745286	-0.010099	-0.008762
50%	-3.492161	0.006973	0.011853
75%	11.018086	0.023416	0.027667
max	141.988339	0.059717	0.188025

Figure 19: Descriptive Stats for Input Variables

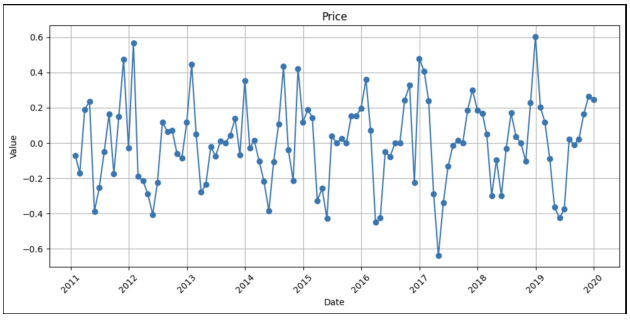


Figure 20: Cucumber Price Variation

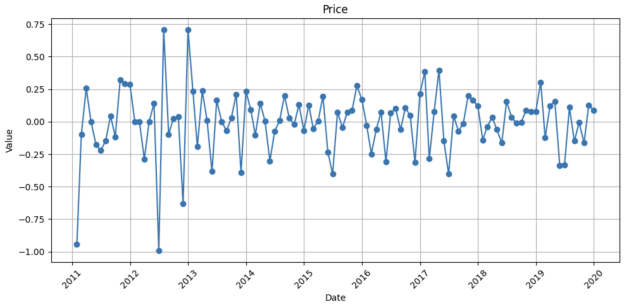


Figure 21: Tomato Price Variation

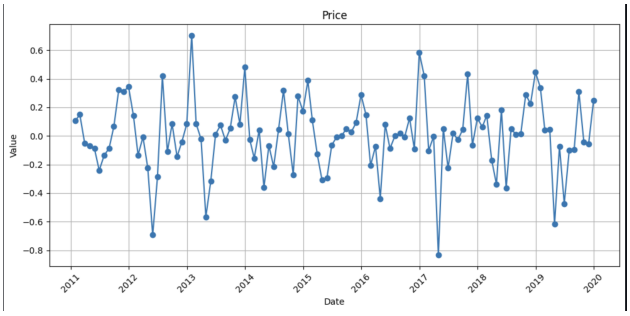


Figure 22: Eggplant Price Variation

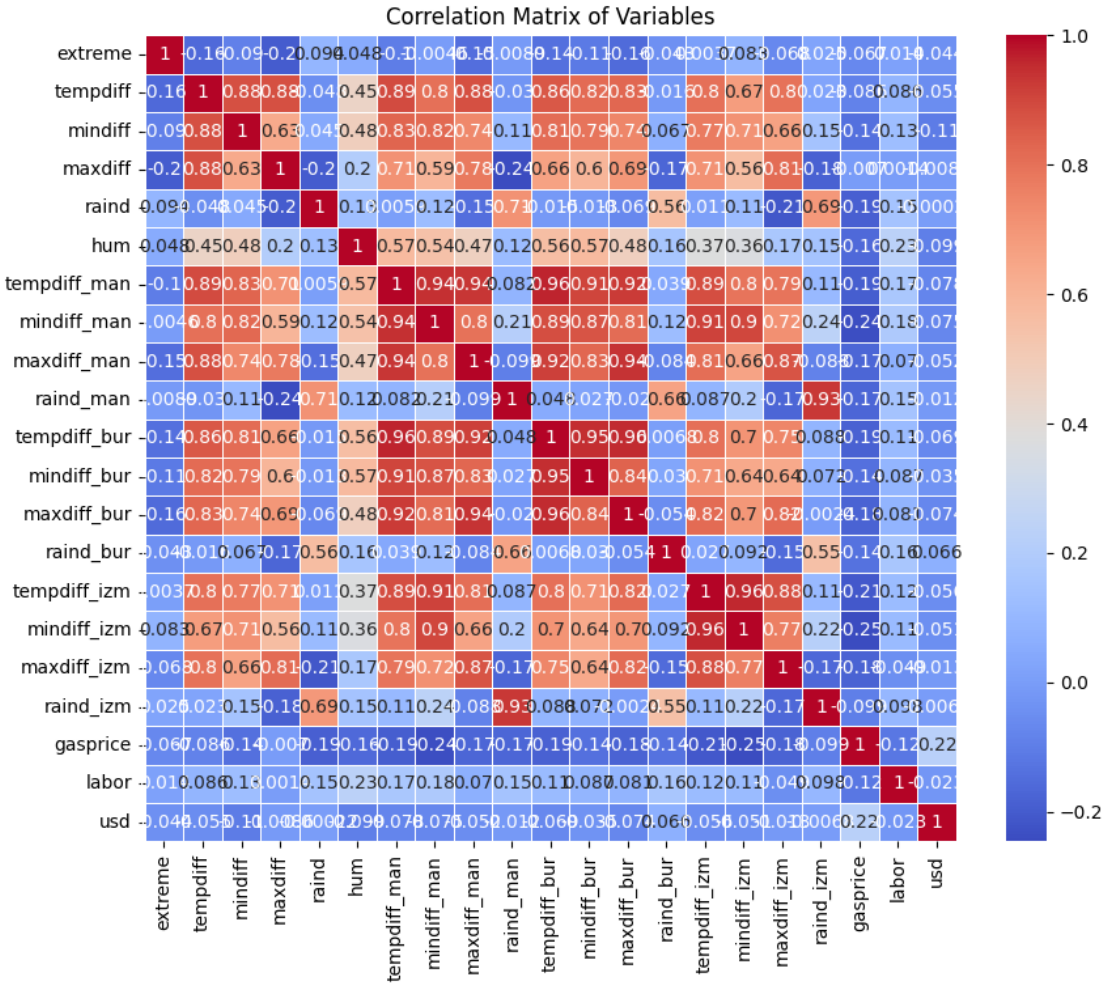


Figure 23: Correlationmatrix