Almond Prices, Production, & California Weather. (1980 to 2021)

Joe Mallonee (jwmallon@mtu.edu)
Doni Obidov (dobidov@mtu.edu)

Michigan Technological University
04/24/2023

Introduction

In this study, we analyze the potential relationship between weather patterns and almond prices and production in California using a time series approach. Inspired by Richard Roll's 1984 paper on the correlation between temperature and frozen orange juice concentrate futures, we adapt and expand his methodology to examine the contemporary almond industry (Roll). Focusing on California's significant almond production, we assess the impact of weather variables on almond yield and prices.

Roll's findings indicate that temperature, rather than precipitation, better explains variations in commodity markets for orange juice concentrate. Although we do not consider trading data, we similarly explore the almond industry, a crop that presents unique environmental preferences and challenges, utilizing improved data accessibility and distinct time series analysis methods.

In 2019, California produced 80% of the world's almonds and almost 100% of the U.S. commercial supply, with production concentrated in roughly 16 counties (California Department of Food and Agriculture, Almond Board of California). Between 2019 and 2021, almond production increased by a remarkable 28%, highlighting the crop's significance and the importance of understanding the impact of weather variables on this industry (United States Department of Agriculture).

Critically, the extreme geographic concentration of a certain crop's production enables us, as it did Roll, to conduct a feasible analysis with accessible data. For Roll, "98 percent of U.S. production [took] place in the central Florida region around Orlando" (Roll 1). For us, 2,898,126,359 pounds of edible almond kernels were produced across the 1,320,000 bearing acres of only 16 California counties in the 2021 crop year alone (Almond Board of California).

Figure 0. The 16 primary almond producing counties of California: Butte, Colusa, Glenn, Solano, Sutter, Tehama, Yolo, Yuba, Merced, San Joaquin, Stanislaus, Fresno, Kern, Kings, Madera, Tulare

Data Sets

Our data consists of five sets taken over an identical period stretching from January 1st, 1980, and December 31, 2021 (41 years). Importantly, our analysis will focus primarily on *two* distinct time series in terms of response: Grower's Price (cents/pound in USD), and Production (pounds).

Response

Grower's Price (cents/pound)	Average price (USD cents/pound) paid to farmers for almonds in a season, influenced by crop yields.
Production (pounds)	Shelled almond kernel production (millions of pounds) in a season.

Sourced from: https://www.ers.usda.gov/data-products/fruit-and-tree-nuts-data/fruit-and-tree-nuts-yearbook-tables/#Tree%20Nuts

Weather conditions significantly affect both almond production and grower's price, making an analysis of their impact on price and production essential. Extreme events like droughts can reduce crop yields, decreasing supply and increasing prices, while favorable conditions result in higher yields and lower prices. Similarly, factors such as lack of rain, water shortages, frost, or heat waves can impact kernel size, quality, and overall production. Therefore, analyzing the relationship between weather conditions, season-average grower's price, and almond kernel production offers valuable insights into the almond industry's health and the role of weather on crop yields.

To that end, we introduce the following weather covariate time series data sets. These will be further explored in the section "ARIMA Covariates," where we explore their potential effects on our chosen model information criteria (BIC[1]) and on overall model parsimony:

Covariates

Relative Humidity (%)	Air's moisture content (%) relative to max capacity at a given temperature, influencing plant growth.
Dry Bulb Temperature (°F)	Dry bulb temperature is the temperature of the air measured with a thermometer that is not affected by moisture, and is commonly used as the standard air temperature reported. Affects photosynthesis, plant growth, and evapotranspiration in crops.
Wind Speed (miles/hour)	Horizontal air movement (miles/hour) influencing transpiration, pollen dispersal, and disease spread.

Modesto, CA (Station WBAN:23155); Sourced from https://www.ncei.noaa.gov/cdo-web/datatools

[1] BIC is chosen as our goal is to build an explanatory model, rather than to forecast (AIC, AICc).

We speculate that weather conditions significantly impact almond production and grower's price, with variables such as relative humidity, temperature, and wind speed playing crucial roles. Low relative humidity reduces growth and yield, while high humidity increases the risk of plant diseases and reduces photosynthesis due to low light penetration through the canopy. Dry bulb temperature affects photosynthesis, plant growth, and evapotranspiration, with high temperatures causing heat stress, reduced yield and quality, and low temperatures slowing plant development or causing frost damage. Wind speed influences transpiration, pollen dispersal, and disease spread, where strong winds increase water stress and reduce yields, while gentle breezes aid pollination and prevent moisture buildup. These interconnected variables may collectively determine the health of the almond industry, impacting crop yields and, consequently, the grower's price.

Goals

The time series analysis that follows aims to examine the potential relationship between changes in weather conditions (relative humidity, temperature, and wind speed) and changes in almond kernel price and production over time. The primary research question at hand is:

> *"How are changes in the weather covariates and changes in price and production potentially related?"*

The focus is on understanding the impact of weather changes on price and production changes, rather than predicting future values. To that end: we are not interested in forecasting. This implies our selection of the Bayesian information criterion (BIC) for assessing relative model performance and conducting later diagnostics.

Additionally, this analysis contributes to more general time series research by exploring questions such as:

> "Can factors that consistently improve BIC be identified?"
> "Do relevant covariates consistently lead to more parsimonious models?"

While this particular study investigates the almond industry and weather, our ultimate goal is to explore these more broad and fundamental questions of time series analysis.

Process Summary

1. Data preparation (missing values, aggregation)
2. Stationarity:
    a. Deterministic trends (models, data transformations)
    b. Stochastic trends (differencing)
3. Assumption verification (residual analysis):
    a. Independence of error process (Runs test, ACF)
    b. Zero-mean and homoscedasticity (residual plots)
    c. Normality (Q-Q plots, histograms, Shapiro-Wilks test)
4. ARIMA(p,d,q) model selection:
    a. Identify candidate (p, d, q)'s (ACF, PACF, EACF)
    b. Estimate models (MLE)
    c. Select best performing model(s) (Information Criteria).
5. Parameter estimation:
    a. Determine AR(p) and MA(q) coefficients.
6. Model diagnostics (residual analysis):
    a. Verify independence (Ljung-Box), zero-mean, homoscedasticity, and normality.
    b. Verify that additional AR(p) and MA(q) parameters are not significant.

Modeling

Model Specification

Price

Time plot and stationarity

Figure 1 illustrates the yearly grower price of almonds produced in California from 1980 to 2021. The price appears to grow linearly without a seasonal pattern, which is not surprising given our use of yearly aggregation.

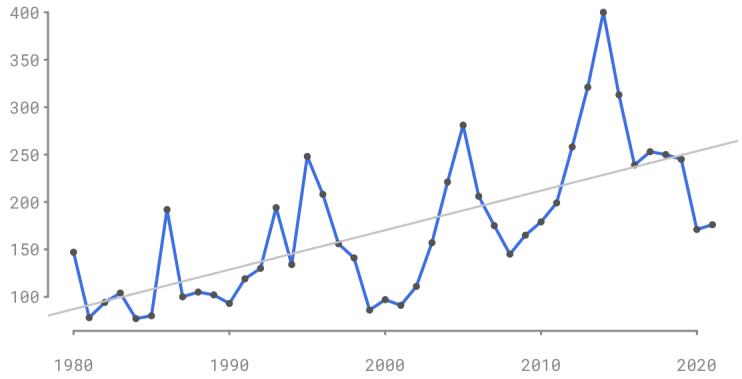

Figure 1. Time plot of grower price (cents/pound) of almonds produced in California

In this project, we are not interested in predicting the future grower price, but rather we are trying to explore how change in various weather conditions and price change are correlated. Therefore, we need to look at the difference in price. We decided to work with a log difference in price as it approximates percentage change and is therefore highly interpretable.

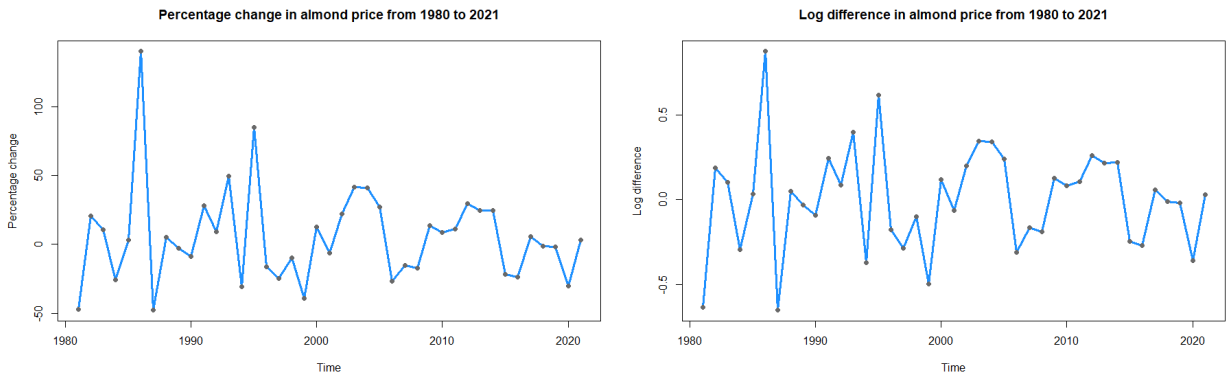

Figure 2. Log difference vs percentage change in almond price

Taking the log difference also turned a non-stationary price into a stationary series.

Table 1: Unit root tests for price

VARIABLE	METHOD	P-VALUE	STATUS
Price	ADF	0.049	On the border
Price	PP	0.045	On the border
Price	KPSS	0.013	Not stationary
log Price	ADF	0.014	Stationary
log Price	PP	0.012	Stationary
log Price	KPSS	0.010	Not stationary
diff Price	ADF	0.037	Stationary
diff Price	PP	0.010	Stationary
diff Price	KPSS	0.100	Stationary
diff log Price	ADF	0.038	Stationary
diff log Price	PP	0.010	Stationary
diff log Price	KPSS	0.100	Stationary

The assumption of a non-stationary price and a stationary difference in log price can also be confirmed by the corresponding ACF plots.

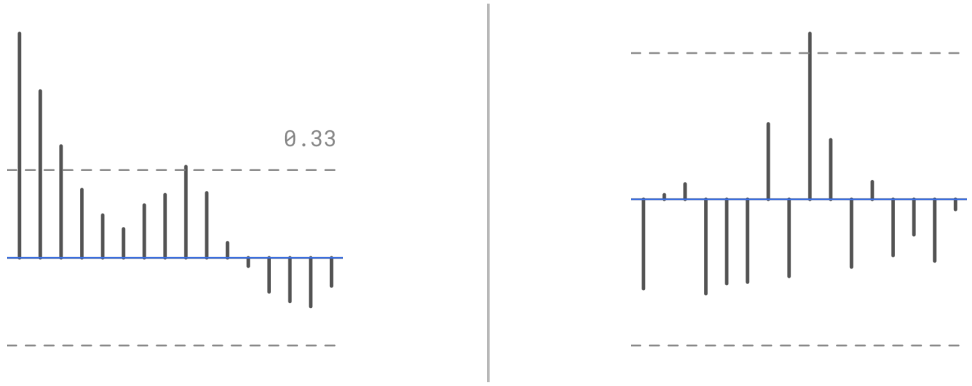

Figure 3. ACF plots of price (left) vs difference in log price (right)

Residual analysis

We have fit a linear model to the difference in log price. We will discuss the linear model in the following sections. In this section we will make sure that the residuals meet the assumptions of a linear model.

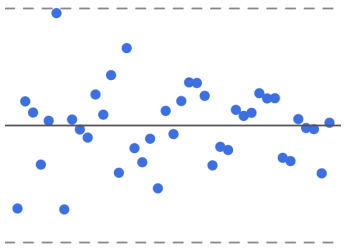

Figure 4. Residual plot for the linear model of difference in log price

Homoscedasticity and zero mean

The residual plot questions the homoscedasticity assumption of a linear model. However, if we ignore four outlier points, the residual plot starts looking like a random scattering of points. Therefore, we decided not to further investigate the violation of homoscedasticity assumption.

Normality

Both the Q-Q plot and the histogram of the residuals confirm the assumption of normality.

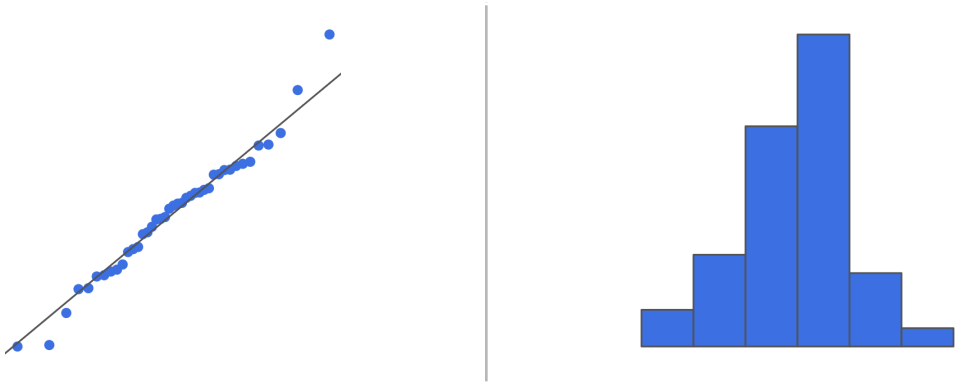

Figure 5. Q-Q plot (left) and histogram (right) of the residuals (linear model + diff log price)

The Shapiro-Wilk test has a p-value of 0.598, which further confirms the conclusions above.

Independence

The p-value of 0.780 for the runs test and the ACF plot do not indicate any violations of the independence assumption.

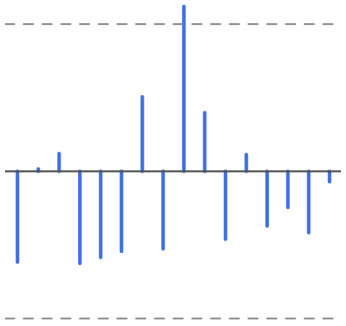

Figure 6. ACF plot of the residuals for the linear model for the difference in log sales

Production

Time plot and stationarity

Figure 7 shows that the production of almonds in California grew exponentially from 1980 to 2021. We cannot see any seasonal pattern in the data aggregated by years.

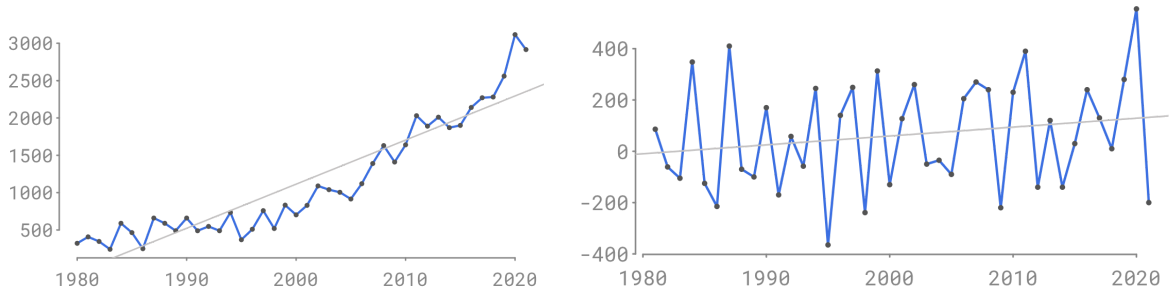

Figure 7. Time plot of production (million pounds) vs difference in production

Our primary goal is to see how the production change is correlated with the change in various weather regressors. Figure 7, above, suggests that the difference in production is stationary.

Table 2: Unit root tests for production

VARIABLE	METHOD	P-VALUE	STATUS
Production	ADF	0.990	Not stationary
Production	PP	0.704	Not stationary
Production	KPSS	0.010	Not stationary
log Production	ADF	0.675	Not stationary
log Production	PP	0.010	Stationary
log Production	KPSS	0.010	Not stationary
diff Production	ADF	0.017	Stationary
diff Production	PP	0.010	Stationary
diff Production	KPSS	0.083	Stationary
diff log Production	ADF	0.010	Stationary
diff log Production	PP	0.010	Stationary
diff log Production	KPSS	0.100	Stationary

Figure 8 further confirms that the production is not stationary (which is obvious given the exponential trend), while the first difference in production is stationary.

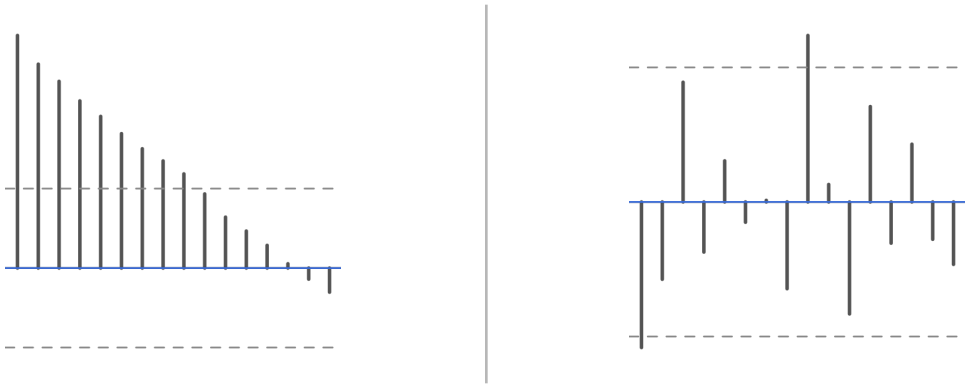

Figure 8. ACF plots of production (left) vs difference in production (right)

Residual analysis

We detrended the difference in production by fitting a linear model. We will come back to the linear model in the next sections. In this section, we want to make sure that the assumptions of a linear model are met.

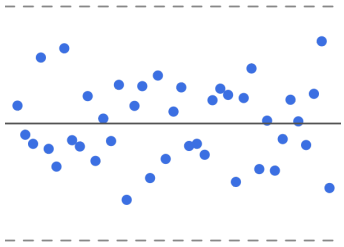

Figure 9. Residual plot for the linear model of difference production

Homoscedasticity and zero mean

Figure 9 does not suggest any violations in zero mean and homoscedasticity assumptions, as the residual points do not follow any clear pattern.

Normality

Both the Q-Q plot and the histogram suggest a normal distribution of the residuals.

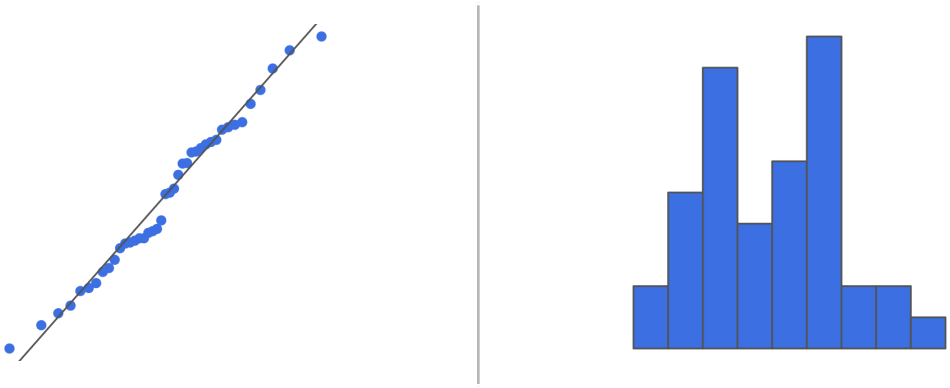

Figure 10. Q-Q plot (left) and histogram (right) of the residuals (diff production)

The Shapiro-Wilk test also supports the assumption of normality by having a p-value of 0.590.

Independence

The runs test does not reject the assumption of independence with a p-value of 0.056. The ACF plot indicates that there could be some violations of this assumption. We decided to move on with assuming the independence; however, in the future work, we could use a hack estimator during hypothesis testing.

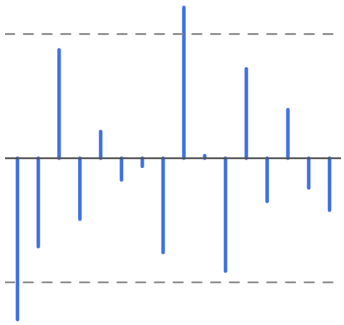

Figure 11. ACF plot of the residuals for the linear model for the difference in production

Model Fitting and Diagnostics

Selection of (p, d, q) orders

We have selected the orders of the ARIMA model by assessing plots for ACF, PACF, and ARIMA.subsets() of diff log Price and diff Production.

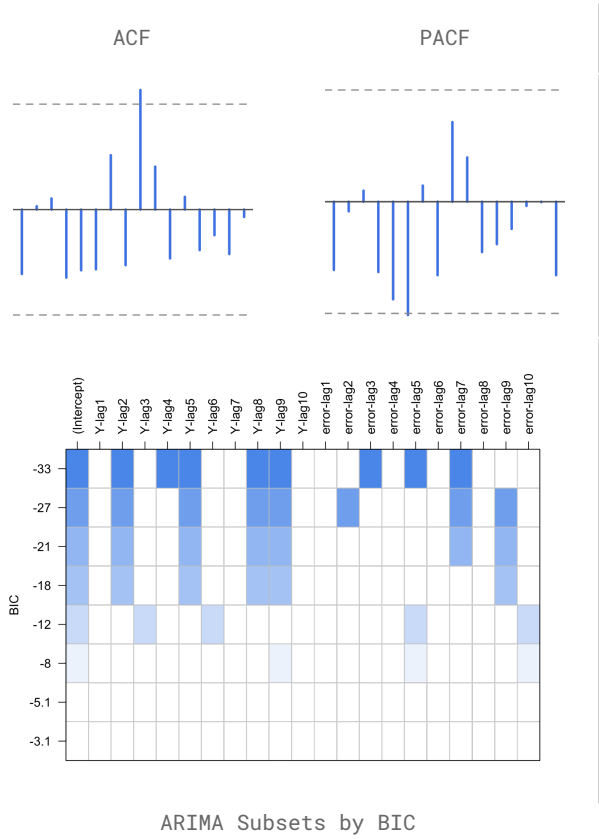

Figure 12. Selecting the ARIMA order for diff log price

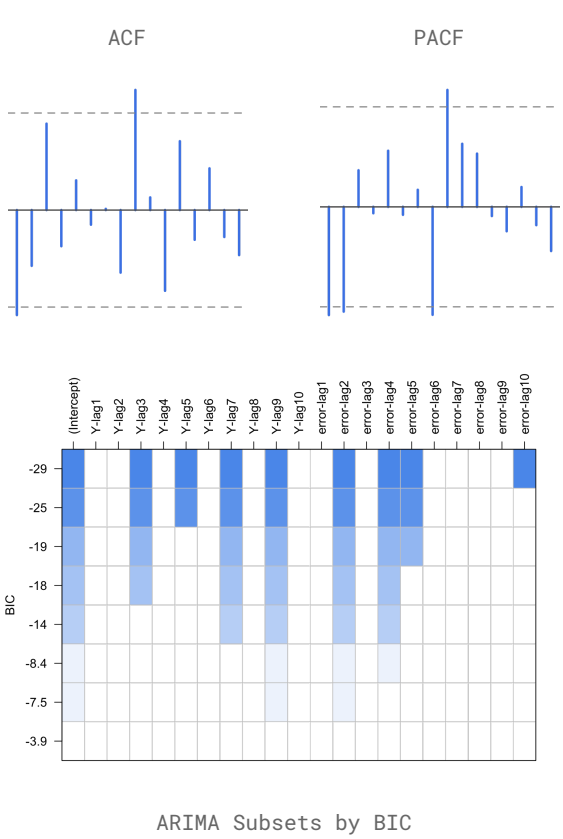

Figure 13. Selecting the ARIMA order for diff production

Based on the ACF, PACF, BIC ARIMA Subsets plots—and the BIC values of individual models themselves[2] for each of our response time series, we opt to assess the following ARIMA models:

[2] Please see appendix for a listing of the potential candidates by BIC.

Table 3: Candidate models for diff log price (left) and diff production (right)

MODEL	BIC	MODEL	BIC
ARIMA(0, 0, 0)	26.5878	**ARIMA(0, 0, 1)**	559.2685
ARIMA(9, 0, 7)	52.24535	**ARIMA(9, 0, 0)**	569.1044

We will now move on to verify these model selections through the overfitting process, conducted with respect to each of these four selected models.

Overfitting

Price

For the top candidate models for diff log price (p, d, q), we have examined the higher order models (p+1, 0, q) and (p, 0, q+1) to check whether the higher order coefficients are significant.

Table 4: Overfitting candidate models for diff log price

MODEL		OVERFIT LAG	95% C.I.	SIGNIFICANCE
p + 1	**ARIMA(1, 0, 0)**	AR(1)	-0.20 ± 0.32	**Insignificant**
q + 1	**ARIMA(0, 0, 1)**	MA(1)	-0.21 ± 0.33	**Insignificant**
p + 1	**ARIMA(10, 0, 7)**	AR(10)	-0.17 ± 0.60	**Insignificant**
q + 1	**ARIMA(9, 0, 8)**	MA(8)	-0.95 ± 1.29	**Insignificant**

Here, we can see that all of the overfit model coefficients are insignificant. This means that we have selected the correct models for diff log Price: ARIMA(0, 0, 0) and ARIMA(9, 0, 7). We then assess the standardized residuals of these two models to ensure they satisfy the necessary assumptions:

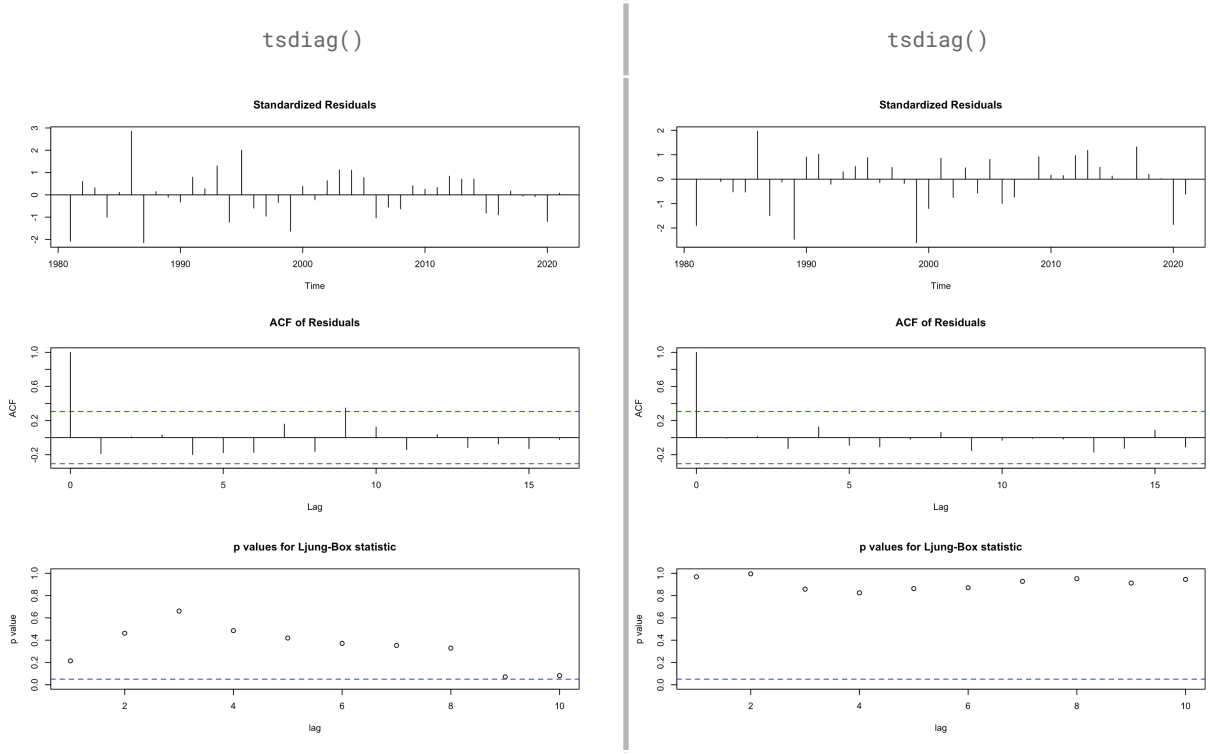

Figure 14. Residual analysis for ARIMA(1, 0, 0) (left) and ARIMA(9, 0, 7) (right) for diff log price

The residual plots, ACF plots of the residuals, and the Ljung-Box tests do not indicate any violations of stationarity and independence for any of the models.

Production

We have checked the higher order models for the candidate models of diff production to make sure that none of the higher order coefficients are significant.

Table 5: Overfitting candidate models for diff production

MODEL		OVERFIT LAG	95% C.I.	SIGNIFICANCE
p + 1	**ARIMA(1, 0, 1)**	AR(1)	-0.02 ± 0.59	**Insignificant**
q + 1	**ARIMA(0, 0, 2)**	MA(2)	+0.04 ± 0.49	**Insignificant**
p + 1	**ARIMA(10, 0, 0)**	AR(10)	-0.29 ± 0.33	**Insignificant**
q + 1	**ARIMA(9, 0, 1)**	MA(1)	-0.52 ± 0.37	**Significant**

Here, we can see that only the overfit model coefficients for the ARIMA(0, 0, 1) are insignificant. The BIC of ARIMA(9, 0, 1) is much higher than the BIC of MA(1), so we did not explore the ARIMA(9, 0, 1) model further. Once again, we assessed the standardized residuals for the candidate models to check the model assumptions.

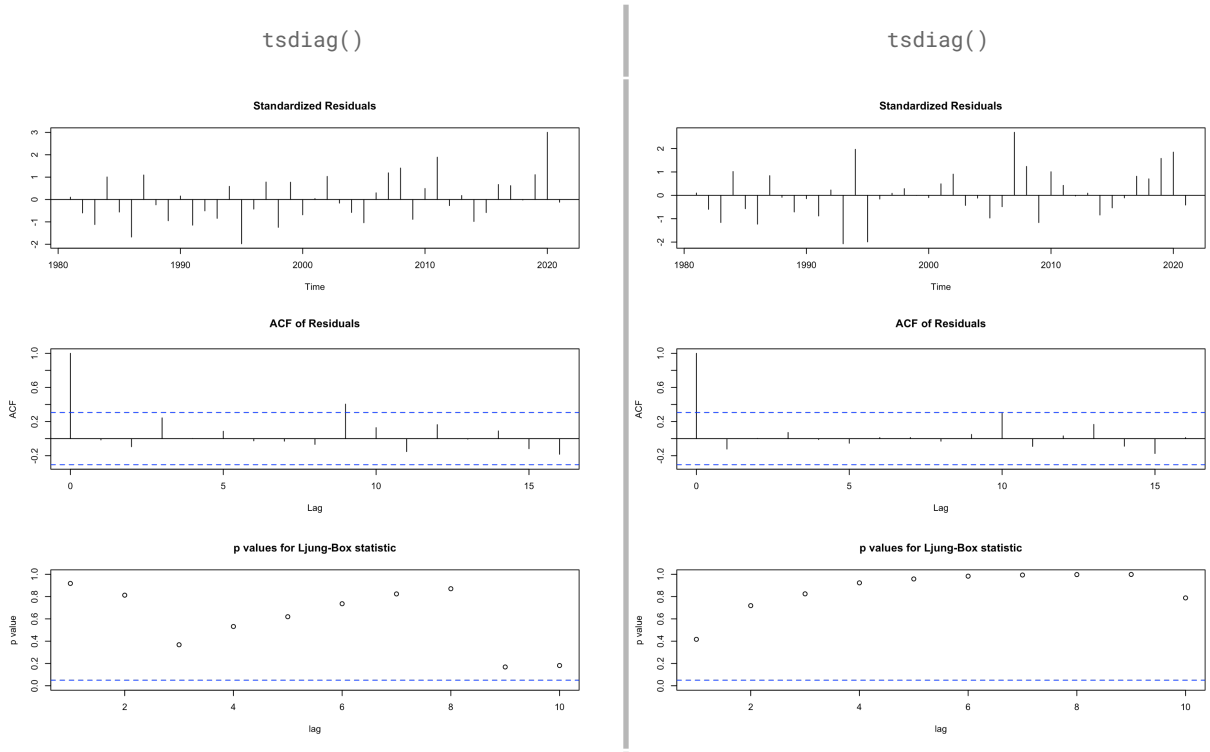

Figure 15. Residual analysis for MA(1) (left) and ARIMA(9, 0, 0) (right) for diff production

The residual plots, ACF plots of the residuals, and the Ljung-Box tests do not indicate any violations of stationarity and independence for any of the models.

ARIMA Regression/Covariates

We now introduce three weather covariates to the candidate ARIMA(p, d, q) models in order to assess the impact that each possible combination of them may have on the BIC performance of the diff log Price ARIMA(p, d, q) and diff Production ARIMA(p, q, q) models.

Humidity	Temperature	Windspeed
Relative %	Degrees F	Miles/Hour

As discussed in the introductory section, we feel there are three particular weather covariates worth exploring in this context: relative humidity, dry bulb temperature, and wind speed. As this source data was in a frequency of days, these observations were rolled up into yearly averages in order to match the frequency of our response data sets on almond price and production.

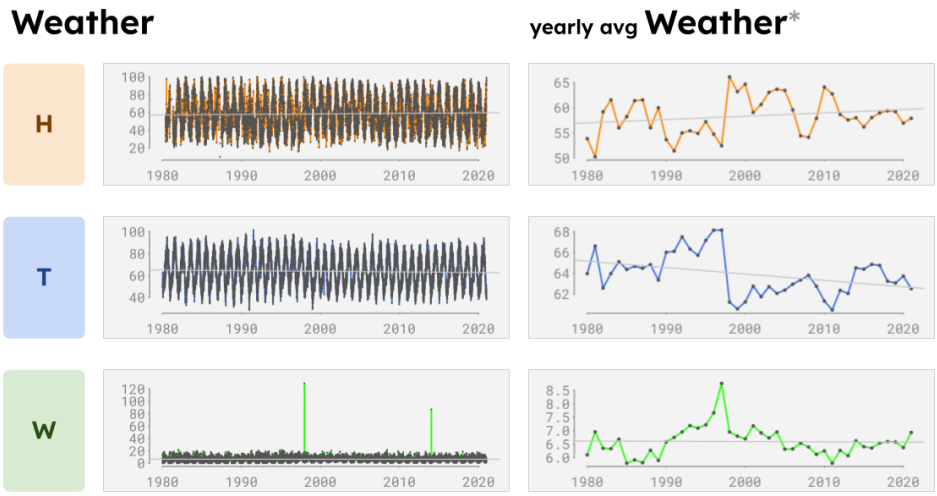

Figure 16. Weather covariates (relative humidity, temperature, windspeed) over the time period of 1980 to 2021. Daily averages were rolled up into yearly averages to match the cadence of the response variables (price, production)

Based on the best fitting models according to the diagnostics, we introduced three relevant weather covariates and conducted ARIMA(p,d,q) regressions. Below, one can see an interesting pattern both in terms of the relationships between these covariates and almond price and production, as well as in terms of time series analysis more broadly.

Specifically, notice in the following table that the introduction of some combination of covariates into an ARIMA(p,d,q) model tends to result in a noticeable improvement to BIC.[3]

[3] The sole exception to this is the diff log Price ARIMA(9,0,7) model, which shows no improvement in BIC across any combination of covariates.

Table 6: BIC of the candidate models with the weather regressors

MODEL		COVARIATES[4]			BIC
diff log Price	ARIMA(0,0,0)	H	T	W	29.48383
diff log Price	ARIMA(0,0,0)	H	T		25.96474
diff log Price	ARIMA(0,0,0)	H		W	27.95859
diff log Price	ARIMA(0,0,0)		T	W	33.75498
diff log Price	ARIMA(0,0,0)	H			25.74740
diff log Price	ARIMA(0,0,0)		T		30.22771
diff log Price	ARIMA(0,0,0)			W	30.05483
diff log Price	ARIMA(0,0,0)				26.58779
diff log Price	ARIMA(9,0,7)	H	T	W	55.94549
diff log Price	ARIMA(9,0,7)	H	T		58.31152
diff log Price	ARIMA(9,0,7)	H		W	54.72239
diff log Price	ARIMA(9,0,7)		T	W	58.15431
diff log Price	ARIMA(9,0,7)	H			55.92905
diff log Price	ARIMA(9,0,7)		T		59.17534
diff log Price	ARIMA(9,0,7)			W	57.97431
diff log Price	ARIMA(9,0,7)				52.24535
diff Production	ARIMA(0,0,1)	H	T	W	560.3088
diff Production	ARIMA(0,0,1)	H	T		556.7073
diff Production	ARIMA(0,0,1)	H		W	559.8764
diff Production	ARIMA(0,0,1)		T	W	566.5588
diff Production	ARIMA(0,0,1)	H			557.6343
diff Production	ARIMA(0,0,1)		T		562.9416
diff Production	ARIMA(0,0,1)			W	562.8474
diff Production	ARIMA(0,0,1)				559.2685
diff Production	ARIMA(9,0,0)	H	T	W	573.2787
diff Production	ARIMA(9,0,0)	H	T		570.4709

[4] Please see the appendix for analysis on the stationarity and standardized residuals of these covariates.

MODEL		COVARIATES[4]			BIC
diff Production	ARIMA(9,0,0)	H		W	570.3970
diff Production	ARIMA(9,0,0)		T	W	573.1860
diff Production	ARIMA(9,0,0)	H			566.9554
diff Production	ARIMA(9,0,0)		T		570.2323
diff Production	ARIMA(9,0,0)			W	569.7328
diff Production	ARIMA(9,0,0)				569.1044

The best ARIMA(p, d, q) models for price and production were then selected and confidence intervals for each of the model coefficients were then assessed for significance. The ultimate results being:

MODEL		COVARIATES		BIC
diff log Price	ARIMA(0,0,0)	H		25.74740
diff Production	ARIMA(0,0,1)	H	T	556.7073

Figure 17. Best-performing ARIMA(p,d,q) regression models for diff log Price; diff Production

MODEL		COVAR. I	95% C.I.	COVAR. II	95% C.I.
diff log Price	ARIMA(0,0,0)	Humidity	0.0251 ± 0.0115	n/a	
diff Production	ARIMA(0,0,1)	Humidity	-34.659 ± 10.335	Temperature	-53.455 ± 24.019

Figure 18. Significance checking of model coefficients through confidence interval construction

Discussion & Summary

We have discovered that an increase in humidity is positively correlated with a rise in price and decrease in production. We have also concluded that there is a negative correlation between the difference in temperature and the difference in production.

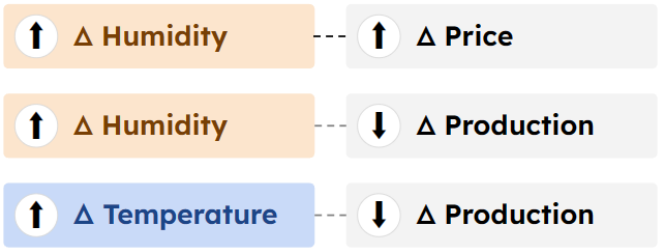

Figure 19. Correlations between the responses and the weather regressors

The introduction of relevant covariates into ARIMA(p, d, q) models improves BIC in many cases. Moreover, adding additional weather regressors simplifies models by reducing the number of components.

The study of covariates in this way might serve certain contexts as a feature selection approach. Notice that, in contrast with Roll's earlier study on orange juice concentrate in Florida, humidity is worth consideration, both individually and when paired with temperature.

There are some improvements that can be made over this work. We have only explored linear relationships between weather regressors and the variables of interest; however, nonlinear relationships are also worth investigating. Moreover, there are many other weather conditions that were not studied in this project.

Appendix

Weather Covariates Models' Residual Analysis

Stationarity

Table A1: Unit root tests for weather covariates

COVARIATE	METHOD	P-VALUE	STATUS
Humidity	ADF	0.336	Not stationary
Humidity	PP	0.022	Stationary
Humidity	KPSS	0.100	Stationary
diff Humidity	ADF	0.010	Stationary
diff Humidity	PP	0.010	Stationary
diff Humidity	KPSS	0.010	Not stationary
Temperature	ADF	0.460	Not stationary
Temperature	PP	0.079	Not stationary
Temperature	KPSS	0.090	Stationary
diff Temperature	ADF	0.032	Stationary
diff Temperature	PP	0.010	Stationary
diff Temperature	KPSS	0.100	Stationary
Windspeed	ADF	0.693	Not stationary
Windspeed	PP	0.093	Not stationary
Windspeed	KPSS	0.100	Stationary
diff Windspeed	ADF	0.078	Stationary
diff Windspeed	PP	0.010	Stationary
diff Windspeed	KPSS	0.100	Stationary

Homoscedasticity, Zero Mean, Normality, and Independence

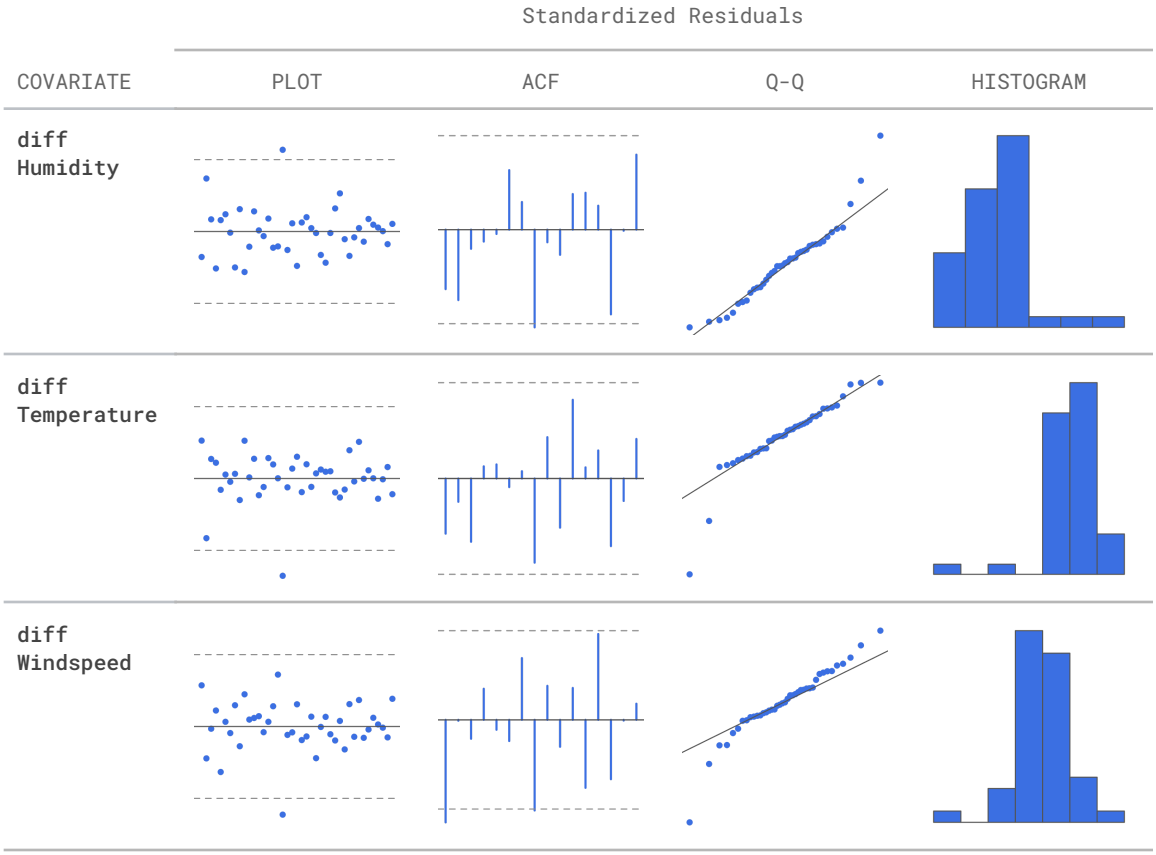

Figure A1. Residual analysis for models with the weather regressors

References

Almond Board of California. *Almond Industry Position Report, July 2019*. 2019. *Position Reports*, https://www.almonds.com/sites/default/files/2020-02/2019.07_PosRpt_EMCII.pdf. Accessed 1 April 2023.

California Department of Food and Agriculture. *California Agricultural Exports 2019–2020*. 2020. *California Department of Food and Agriculture*, https://www.cdfa.ca.gov/Statistics/PDFs/2020_Exports_Publication.pdf. Accessed 1 April 2023.

Chan, Kung-Sik, and Jonathan D. Cryer. *Time Series Analysis with Applications in R*. 2 ed., Springer, 2008.

Roll, Richard. “Orange Juice and Weather.” *The American Economic Review*, vol. 74, no. 5, 1984, pp. 861–80. *JSTOR*, http://www.jstor.org/stable/549. Accessed 1 April 2023.

United States Department of Agriculture. *2021 California Almond Acreage Report*. 2022. *United States Department of Agriculture, National Agricultural Statistics Service*, https://www.nass.usda.gov/Statistics_by_State/California/Publications/Specialty_and_Other_Releases/Almond/Acreage/202204almac.pdf. Accessed 1 April 2023.

R Markdown (abridged)

Full .Rmd and data set .CSVs can be downloaded from: https://github.com/jmallo/almonds-and-weather

````
---
title: "Almond Prices, Production, and California Weather (1980 to 2021)"
author: "Joe Mallonee (jwmallon@mtu.edu), Doni Obidov (dobidov@mtu.edu)"
---

```{r setup, include=FALSE}
alpha <- 0.05

# Set up ######################################################################
knitr::opts_chunk$set(echo = TRUE, warning = FALSE)
options(tinytex.verbose = TRUE)

packages <- c("dplyr", "extrafont", "forecast", "grDevices", "lmtest",
              "lubridate", "MASS", "sandwich", "tinytex", "TSA", "tseries")
for (package in packages) {
  if (!requireNamespace(package, quietly = TRUE)) {
    install.packages(package)
  }
}

library(dplyr)
library(extrafont)
library(forecast)
library(grDevices)
library(lmtest)
library(lubridate)
library(MASS)
library(sandwich)
library(tinytex)
library(TSA)
library(tseries)

# Helper functions ############################################################
# Keep track of assumptions, hypothesis test results/p-values, etc.
add_assumption <- function(assumption_matrix, data_name,
                           method, p_value, conclusion, other = "") {
  new_row <- data.frame(
    data.name = data_name,
    method = method,
    p.value = p_value,
    conclusion = conclusion,
    other = other,
    stringsAsFactors = FALSE)

  assumption_matrix <- rbind(assumption_matrix, new_row)
}

series_unit_root_tests <- function(series, assumptions_record) {
  data.name_ <- toString(substitute(series))

  adf_ <- adf.test(series)
  h0 <- "H_0: non-stationary"; hA <- "H_A: stationary"
  assumptions_record <- add_assumption(assumptions_record,
                                       data.name_,
````
````

```
                                      adf_$method,
                                      adf_$p.value,
                                      ifelse(adf_$p.value > alpha, h0, hA))
  pp_ <- pp.test(series)
  h0 <- "H_0: non-stationary"; hA <- "H_A: stationary"
  assumptions_record <- add_assumption(assumptions_record,
                                      data.name_,
                                      pp_$method,
                                      pp_$p.value,
                                      ifelse(pp_$p.value > alpha, h0, hA))
  kpss_ <- kpss.test(series)
  h0 <- "H_0: stationary"; hA <- "H_A: non-stationary"
  assumptions_record <- add_assumption(assumptions_record,
                                      data.name_,
                                      kpss_$method,
                                      kpss_$p.value,
                                      ifelse(kpss_$p.value > alpha, h0, hA))
}

resids_independence_tests <- function(series, assumptions_record) {
  data.name_ <- toString(substitute(series))

  series_lm_ <- lm(series ~ time(series))
  resids_lm_ <- rstandard(series_lm_)

  runs_ <- runs(resids_lm_)
  h0 <- "H_0: independent"; hA <- "H_A: dependent"
  assumptions_record <- add_assumption(assumptions_record,
                                      paste("std residuals,", data.name_),
                                      "Runs Test",
                                      runs_$pvalue,
                                      ifelse(runs_$pvalue > alpha, h0, hA))

  acf_ <- acf(resids_lm_)$acf
  lags_ <- "Excessive lag(s): "
  for (lag in 1:length(acf_)) {
    if (abs(acf_[lag]) > 1.96/sqrt(length(series))) {
      lags_ <- paste(lags_, toString(lag))
    }
  }
  assumptions_record <- add_assumption(assumptions_record,
                                      paste("std residuals,", data.name_),
                                      "ACF Plot",
                                      "n/a",
                                      lags_)
}

resids_normality_tests <- function(series, assumptions_record) {
  data.name_ <- toString(substitute(series))

  series_lm_ <- lm(series ~ time(series))
  resids_lm_ <- rstandard(series_lm_)

  par(mfrow = c(1, 2))
  qqnorm(resids_lm_)
  qqline(resids_lm_)
  hist(resids_lm_)
  print(shapiro.test(resids_lm_))

  sw_ <- shapiro.test(resids_lm_)
```

```
  h0 <- "H_0: normally distributed"; hA <- "H_A: not normally distributed"
  assumptions_record <- add_assumption(assumptions_record,
                                       paste("std residuals,", data.name_),
                                       sw_$method,
                                       sw_$p.value,
                                       ifelse(sw_$p.value > alpha, h0, hA))
}
```

```{r}
# Create assumption matrices
assumptions_price <- data.frame(
  data.name = character(),
  method = character(),
  p.value = numeric(),
  conclusion = character(),
  stringsAsFactors = FALSE
)

assumptions_production <- data.frame(
  data.name = character(),
  method = character(),
  p.value = numeric(),
  conclusion = character(),
  stringsAsFactors = FALSE
)

assumptions_humidity <- data.frame(
  data.name = character(),
  method = character(),
  p.value = numeric(),
  conclusion = character(),
  stringsAsFactors = FALSE
)

assumptions_temperature <- data.frame(
  data.name = character(),
  method = character(),
  p.value = numeric(),
  conclusion = character(),
  stringsAsFactors = FALSE
)

assumptions_windspeed <- data.frame(
  data.name = character(),
  method = character(),
  p.value = numeric(),
  conclusion = character(),
  stringsAsFactors = FALSE
)
```

\newpage
General Questions, illustrated through the "straight-forward/common-sense" belief that "weather SHOULD impact crop yield." Here, we explore almond production in California (geographic concentration makes such an analysis possible—compare to commodities such as maize, rice, wheat

````
which are grown all over the world).

General Questions:
1) Do additional features [assuming relevant to response] (weather points...) lead to more
parsimonious models?
2) Do additional features (weather points...) *consistently* improve information criteria?




# 0. Load and format data

```{r}
weather <- read.csv("local/data/weather_daily.csv")
almonds <- read.csv("local/data/almonds_yearly.csv")

weather[weather == 0] <- NA

# Convert relevant columns to time series objects
production <- ts(almonds$Production, start = c(1980), end = c(2021), frequency = 1)
price <- ts(almonds$GrowerPrice, start = c(1980), end = c(2021), frequency = 1)
temperature <- ts(weather$HourlyDryBulbTemperature, start = c(1980), end = c(2021), frequency =
365)
humidity <- ts(weather$HourlyRelativeHumidity, start = c(1980), end = c(2021), frequency = 365)
windspeed <- ts(weather$HourlyWindSpeed, start = c(1980), end = c(2021), frequency = 365)

# Aggregate weather to yearly cadence
weather_agg <- weather %>%
  mutate(Date = as.Date(DATE, format = "%Y-%m-%d")) %>%
  group_by(Year = year(Date)) %>%
  summarize(MeanTemperature = mean(HourlyDryBulbTemperature, na.rm = TRUE),
            MeanHumidity = mean(HourlyRelativeHumidity, na.rm = TRUE),
            MeanWindSpeed = mean(HourlyWindSpeed, na.rm = TRUE))

temperature_agg <- ts(weather_agg$MeanTemperature, start = c(1980), end = c(2021), frequency =
1)
humidity_agg <- ts(weather_agg$MeanHumidity, start = c(1980), end = c(2021), frequency = 1)
windspeed_agg <- ts(weather_agg$MeanWindSpeed, start = c(1980), end = c(2021), frequency = 1)

par(mfrow = c(2, 3))
plot(humidity_agg)
plot(temperature_agg)
plot(windspeed_agg)
plot(price)
plot(production)
```




# Analysis for Price
## 1. Plot, examine, and correct the as-is data for `price`

```{r}
# Plot data
price_lm <- lm(price ~ time(price))
plot(price)
abline(price_lm$coefficients)
````
````

```
```

1a, 1b. Examine deterministic and stochastic trends for `price`

```{r, paged.print=FALSE}
# Explore deterministic trends and apply data transformations as needed
acf(price)
assumptions_price <- series_unit_root_tests(price, assumptions_price)
assumptions_price <- series_unit_root_tests(log(price), assumptions_price)

# Explore stochastic trends and apply data transformations as needed
assumptions_price <- series_unit_root_tests(diff(price), assumptions_price)
assumptions_price <- series_unit_root_tests(diff(log(price)), assumptions_price)

assumptions_price
```

2. Analyze standardized residuals for detrended `price` (`diff log price`)
2a. Conduct independence tests for `diff log price`
2b. Check for zero-mean and homoscedasticity for detrended `diff log price`
2c. Check normality for `diff log price`

```{r}
# 2a. If independent, then you can skip the rest. If fails, may need HAC
assumptions_price <- resids_independence_tests(diff(log(price)), assumptions_price)

# 2b. Zero mean, homoscedasticity assumptions via standardized residual plots
detrend_price_lm <- lm(diff(log(price)) ~ time(diff(log(price))))
plot(rstandard(detrend_price_lm), ylim=c(-4,4))
abline(h=+3, col="red")
abline(h=0)
abline(h=-3, col="red")

# 2c. Approximately normal
assumptions_price <- resids_normality_tests(diff(log(price)), assumptions_price)
```

```{r, paged.print=FALSE}
# Output assumption checks summary
assumptions_price
```

3. Determine order of appropriate ARIMA(p,d,q) model for `diff log price`
3a. Check ACF, PACF, and EACF plots for `diff log price`
3b. Estimate candidate models for `diff log price`
3c. Choose the most reasonable model for `diff log price`

```{r}
# 3a. Create sample ACF, PACF, EACF plots to determine candidate pairs of (p,q)
acf(diff(log(price)))
pacf(diff(log(price)))
# eacf(diff(log(price)))

# 3b. Estimate the candidate models using MLE
plot(armasubsets(y = diff(log(price)), nar = 10, nma = 10, ar.method = "ols"))
# auto.arima(diff(log(price)))

# 3c. Select model with smallest value of relevant Information Criteria (BIC)
price_0.0.0 <- arima(diff(log(price)), order=c(0,0,0)) # Base case
price_9.0.7 <- arima(diff(log(price)), order=c(9,0,7))
```

```
price_9.0.9 <- arima(diff(log(price)), order=c(9,0,9))
price_6.0.10 <- arima(diff(log(price)), order=c(6,0,10))
BIC(price_0.0.0)
BIC(price_9.0.7) # Winner
BIC(price_9.0.9)
BIC(price_6.0.10)
```

4. Conduct parameter estimation for `diff log price`

```{r}
# Not relevant to us at this point in the analysis, comes with later regression
```

5. Conduct model diagnostics for `diff log price`
5a. Conduct residual analysis for `diff log price`

```{r}
# 5a. Perform residual analysis on the estimated error process
# Independence
assumptions_price <- resids_independence_tests(rstandard(price_9.0.7), assumptions_price)

# Zero-mean, homoscedasticity
plot(x=rstandard(price_9.0.7), ylim=c(-4,4))
abline(h=+3, col="red")
abline(h=0)
abline(h=-3, col="red")

# Normality
assumptions_price <- resids_normality_tests(rstandard(price_9.0.7), assumptions_price)

# General residual test for chosen model
tsdiag(price_9.0.7)
```

5b. Verify candidate model selection by overfitting for `diff log price`

```{r}
# Check if the additional AR(p) and MA(q) parameters are significant or not
price_1.0.0 <- arima(diff(log(price)), order=c(1,0,0)) # ar1 insignificant
price_0.0.1 <- arima(diff(log(price)), order=c(0,0,1)) # ma1 insignificant
price_10.0.7 <- arima(diff(log(price)), order=c(10,0,7)) # ar10 insignificant
price_9.0.8 <- arima(diff(log(price)), order=c(9,0,8)) # ma8  insignificant
```

6. Try xreg for other factors...

```{r}
# Function to generate xreg matrices
create_xreg_matrix <- function(h, t, w, include) {
  xreg <- cbind(
    if (include[1]) diff(h),
    if (include[2]) diff(t),
    if (include[3]) diff(w)
  )
  return(xreg)
}

# Define variables
variables <- list(
```

```
  h = humidity_agg,
  t = temperature_agg,
  w = windspeed_agg
)

# Generate all possible combinations of the variables
variable_combinations <- lapply(variables, function(x) c(TRUE, FALSE))
variable_combinations <- expand.grid(variable_combinations)

# Fit ARIMA models and calculate BIC for each model
models <- list()
model_bic <- c()

for (i in 1:nrow(variable_combinations)) {
  include <- as.logical(variable_combinations[i,])
  xreg <- create_xreg_matrix(variables$h, variables$t, variables$w, include)

  model_name <- paste0("price_0.0.0_", paste(names(variables)[include], collapse="."))

  if (all(!include)) {
    model <- arima(diff(log(price)), order = c(0, 0, 0))
  } else {
    model <- arima(diff(log(price)), order = c(0, 0, 0), xreg = xreg)
  }

  models[[model_name]] <- model
  model_bic[model_name] <- BIC(model)
}

for (i in 1:nrow(variable_combinations)) {
  include <- as.logical(variable_combinations[i,])
  xreg <- create_xreg_matrix(variables$h, variables$t, variables$w, include)

  model_name <- paste0("price_9.0.7_", paste(names(variables)[include], collapse="."))

  if (all(!include)) {
    model <- arima(diff(log(price)), order = c(9, 0, 7))
  } else {
    model <- arima(diff(log(price)), order = c(9, 0, 7), xreg = xreg)
  }

  models[[model_name]] <- model
  model_bic[model_name] <- BIC(model)
}

print(model_bic)
```

```
# Analysis for Production
## 1. Plot, examine, and correct the as-is data for `production`

```{r}
# Plot data
production_lm <- lm(production ~ time(production))
plot(production)
abline(production_lm$coefficients)
```
```

```
```

1a, 1b. Examine deterministic and stochastic trends for `production`

```{r, paged.print=FALSE}
# Explore deterministic trends and apply data transformations as needed
acf(production)
assumptions_production <- series_unit_root_tests(production, assumptions_production)
assumptions_production <- series_unit_root_tests(log(production), assumptions_production)

# Explore stochastic trends and apply data transformations as needed
assumptions_production <- series_unit_root_tests(diff(production), assumptions_production)
assumptions_production <- series_unit_root_tests(diff(log(production)), assumptions_production)

assumptions_production
```

2. Analyze standardized residuals for detrended `production` (`diff production`)
2a. Conduct independence tests for `diff production`
2b. Check for zero-mean and homoscedasticity for detrended `diff production`
2c. Check normality for `diff production`

```{r}
# 2a. If independent, then you can skip the rest. If fails, may need HAC
assumptions_production <- resids_independence_tests(diff(production), assumptions_production)

# 2b. Zero mean, homoscedasticity assumptions via standardized residual plots
detrend_production_lm <- lm(diff(production) ~ time(diff(production)))
plot(rstandard(detrend_production_lm), ylim=c(-4,4))
abline(h=+3, col="red")
abline(h=0)
abline(h=-3, col="red")

# 2c. Approximately normal
assumptions_production <- resids_normality_tests(diff(production), assumptions_production)
```

```{r, paged.print=FALSE}
# Output assumption checks summary
assumptions_production
```

3. Determine order of appropriate ARIMA(p,d,q) model for `diff production`
3a. Check ACF, PACF, and EACF plots for `diff production`
3b. Estimate candidate models for `diff production`
3c. Choose the most reasonable model for `diff production`

```{r}
# 3a. Create sample ACF, PACF, EACF plots to determine candidate pairs of (p,q)
acf(diff(production))
pacf(diff(production))
#eacf(diff(production))

# 3b. Estimate the candidate models using MLE
plot(armasubsets(y = diff(production), nar = 10, nma = 10, ar.method = "ols"))
# auto.arima(diff(production))

# 3c. Select model with smallest value of relevant Information Criteria (BIC)
production_0.0.0 <- arima(diff(production), order=c(0,0,0)) # Base & by auto
production_0.0.1 <- arima(diff(production), order=c(0,0,1))
```

```
production_9.0.10 <- arima(diff(production), order=c(9,0,10))
production_9.0.5 <- arima(diff(production), order=c(9,0,5))
production_9.0.4 <- arima(diff(production), order=c(9,0,4))
production_9.0.0 <- arima(diff(production), order=c(9,0,0))
production_9.0.1 <- arima(diff(production), order=c(9,0,1))
production_9.0.2 <- arima(diff(production), order=c(9,0,2))
production_8.0.0 <- arima(diff(production), order=c(8,0,0))
BIC(production_0.0.0)
BIC(production_0.0.1)
BIC(production_9.0.10)
BIC(production_9.0.5)
BIC(production_9.0.4)
BIC(production_9.0.0) # Winner
BIC(production_9.0.1)
BIC(production_9.0.2)
BIC(production_8.0.0)
```

4. Conduct parameter estimation for `diff production`

```{r}
# Not relevant to us at this point in the analysis, comes with later regression
```

5. Conduct model diagnostics for `diff production`
5a. Conduct residual analysis for `diff production`

```{r}
# Perform residual analysis on the estimated error process
assumptions_production <- resids_independence_tests(rstandard(production_9.0.0),
                                                    assumptions_production)

plot(x=rstandard(production_9.0.0), ylim=c(-4,4))
abline(h=+3, col="red")
abline(h=0)
abline(h=-3, col="red")

assumptions_production <- resids_normality_tests(rstandard(production_9.0.0),
                                                 assumptions_production)

tsdiag(production_9.0.0)
```

5b. Verify candidate model selection by overfitting for `diff production`

```{r}
# Check if the additional AR(p) and MA(q) parameters are significant or not
production_1.0.1 <- arima(diff(production), order=c(1,0,1)) # ar1 insignificant
production_0.0.2 <- arima(diff(production), order=c(0,0,2)) # ma2 insignificant
production_10.0.0 <- arima(diff(production), order=c(10,0,0)) # ar10 insignificant
production_9.0.1 <- arima(diff(production), order=c(9,0,1)) # ma1 significant
```

6. Try xreg for weather covariates

```{r}
# Function to generate xreg matrices
create_xreg_matrix <- function(h, t, w, include) {
  xreg <- cbind(
    if (include[1]) diff(h),
```

```
    if (include[2]) diff(t),
    if (include[3]) diff(w)
  )
  return(xreg)
}

# Define variables
variables <- list(
  h = humidity_agg,
  t = temperature_agg,
  w = windspeed_agg
)

# Generate all possible combinations of the variables
variable_combinations <- lapply(variables, function(x) c(TRUE, FALSE))
variable_combinations <- expand.grid(variable_combinations)

# Fit ARIMA models and calculate BIC for each model
models <- list()
model_bic <- c()

# Adding baseline ARIMA(0,0,0) for comparison, it's not particularly relevant
for (i in 1:nrow(variable_combinations)) {
  include <- as.logical(variable_combinations[i,])
  xreg <- create_xreg_matrix(variables$h, variables$t, variables$w, include)

  model_name <- paste0("production_0.0.0_", paste(names(variables)[include], collapse="."))

  if (all(!include)) {
    model <- arima(diff(production), order = c(0, 0, 0))
  } else {
    model <- arima(diff(production), order = c(0, 0, 0), xreg = xreg)
  }

  models[[model_name]] <- model
  model_bic[model_name] <- BIC(model)
}


for (i in 1:nrow(variable_combinations)) {
  include <- as.logical(variable_combinations[i,])
  xreg <- create_xreg_matrix(variables$h, variables$t, variables$w, include)

  model_name <- paste0("production_9.0.0_", paste(names(variables)[include], collapse="."))

  if (all(!include)) {
    model <- arima(diff(production), order = c(9, 0, 0))
  } else {
    model <- arima(diff(production), order = c(9, 0, 0), xreg = xreg)
  }

  models[[model_name]] <- model
  model_bic[model_name] <- BIC(model)
}


for (i in 1:nrow(variable_combinations)) {
  include <- as.logical(variable_combinations[i,])
  xreg <- create_xreg_matrix(variables$h, variables$t, variables$w, include)
```

```
  model_name <- paste0("production_0.0.1_", paste(names(variables)[include], collapse="."))

  if (all(!include)) {
    model <- arima(diff(production), order = c(0, 0, 1))
  } else {
    model <- arima(diff(production), order = c(0, 0, 1), xreg = xreg)
  }

  models[[model_name]] <- model
  model_bic[model_name] <- BIC(model)
}

print(model_bic)
```

```
# Weather covariates

```{r}
# Export original covariate time series plots
export_series(plot_dir, humidty)
export_series(plot_dir, temperature)
export_series(plot_dir, windspeed)
# Export aggregated and diff(aggregated) covariate time series plots
covariates <- list(humidity_agg, temperature_agg, windspeed_agg)
for (covariate in covariates) {
  export_series(plot_dir, covariate)
  export_series(plot_dir, diff(covariate))
  export_acf(plot_dir, diff(covariate))
  export_pacf(plot_dir, diff(covariate))
  export_stdresids_plot_acf_qq_hist(plot_dir, diff(covariate))
}

# Function to run the tests
conduct_tests <- function(data, assumptions) {
  # As-is data
  assumptions <- series_unit_root_tests(data, assumptions)
  assumptions <- resids_independence_tests(data, assumptions)
  assumptions <- resids_normality_tests(data, assumptions)

  # Box-Cox data
  lambda <- BoxCox.lambda(data)
  boxcox_data <- BoxCox(data, lambda)
  assumptions <- series_unit_root_tests(boxcox_data, assumptions)
  assumptions <- resids_independence_tests(boxcox_data, assumptions)
  assumptions <- resids_normality_tests(boxcox_data, assumptions)

  # log data
  assumptions <- series_unit_root_tests(log(data), assumptions)
  assumptions <- resids_independence_tests(log(data), assumptions)
  assumptions <- resids_normality_tests(log(data), assumptions)

  # diff data
  diff_data <- diff(data)
  assumptions <- series_unit_root_tests(diff_data, assumptions)
  assumptions <- resids_independence_tests(diff_data, assumptions)
  assumptions <- resids_normality_tests(diff_data, assumptions)
```
```

```
  # diff log data
  assumptions <- series_unit_root_tests(diff(log(data)), assumptions)
  assumptions <- resids_independence_tests(diff(log(data)), assumptions)
  assumptions <- resids_normality_tests(diff(log(data)), assumptions)

  return(assumptions)
}

# Run the tests for each variable
assumptions_humidity <- conduct_tests(humidity_agg, assumptions_humidity)
assumptions_temperature <- conduct_tests(temperature_agg, assumptions_temperature)
assumptions_windspeed <- conduct_tests(windspeed_agg, assumptions_windspeed)
```