Journal of Agricultural and Resource Economics 45(2):317–334 ISSN: 1068-5502 (Print); 2327-8285 (Online)
Copyright 2020 the authors doi: 10.22004/ag.econ.302457

Factors Influencing Gulf and Pacific Northwest Soybean Export Basis: An Exploratory Statistical Analysis

David W. Bullock and William W. Wilson

The response of U.S. soybean export basis (Gulf and Pacific Northwest) to changes in supply and demand (domestic and international), transportation costs, logistics conditions, and export activity variables was examined from both a market-year average and seasonal analog perspective. The market-year average results indicated that basis at both locations were highly correlated and influenced primarily by international and domestic competition. The seasonal analog results indicated a wide variation in seasonality across marketing years for both locations with transportation costs, logistic conditions, and export activity having the greatest influence on the seasonal analog grouping.

Key words: agglomerative hierarchal clustering (AHC), partial least squares regression (PLS-R), seasonal analogs, variable characterization z-test

Introduction

Basis values for agricultural commodities at nondelivery markets are typically less stable and more unpredictable than basis values at delivery markets. Arbitrage pressures and the fixed costs for delivery assure a high degree of stability and convergence in basis at futures delivery markets. Basis is less stable for nondelivery markets and is particularly volatile in nondelivery markets at export locations that compete directly with other exporting countries. In these markets, basis is affected by numerous factors, including the effects of competitor-country basis values, off-shore demand, and multiple complicated logistic variables. This volatility and understanding the factors that affect basis for these markets have important implications for market participants regarding decisions about risk management, trading, shipping, and storage. The importance of this interdependency is particularly apparent in the international soybean market, where China is by far the largest buyer, the United States and Brazil are vigorous competitors, and transportation plays an important role.

The purpose of this study is to analyze the effects of market and logistical variables on both the level (marketing-year average) and seasonality (by marketing year) of Gulf and Pacific Northwest (PNW) nearby soybean basis values. Explanatory variables include Brazilian basis, nearby futures spreads, rail-transportation costs (tariff plus fuel surcharge, secondary rail market values), barge and ocean rates, the number of ships at the port (Gulf and PNW), and a number of additional supply/demand variables.

David W. Bullock (corresponding author) is a research associate professor, and William W. Wilson is a University Distinguished Professor and the CHS Chair of Trading and Risk in the Department of Agribusiness and Applied Economics at North Dakota University.

We thank Siew Lim, David Roberts, and two anonymous reviewers for their helpful comments on an earlier version of this manuscript. We acknowledge the North Dakota Soybean Council for funding support. This project was also funded under USDA-AMS Cooperative Agreement #16-TMTSD-ND-0004.

Review coordinated by Darren Hudson.

Previous Studies

Grain and Oilseed Basis Behavior

Concepts of arbitrage and convergence are normally developed in reference to basis at delivery markets. However, some studies have examined factors that affect basis values at interior, nondelivery locations. These issues include the effect of fundamental (Zhang and Houston, 2005; Wilson and Dahl, 2011), time-series (Taylor, Dhuyvetter, and Kastens, 2006; Hatchett, Brorsen, and Anderson, 2010; Lee and Brorsen, 2017), and a combination of fundamental and time-series (Jiang and Hayenga, 1997) factors.

Only a few studies have analyzed basis variability at export locations. Tilley and Campbell (1988) analyzed the U.S. Gulf hard red winter wheat basis and found that weekly variability was mostly explained by exports, free stocks, and the 1980 Russian grain embargo. These influences were in addition to the selected monthly variables that we included to capture seasonality in basis. Notably absent was the inclusion of shipping costs as an explanatory variable in their model. A more recent study (Lakkakula and Wilson, 2020) analyzed the interdependency among the origin and destination (PNW) basis for soybeans and the impacts of rail shipping costs. Results indicated the origin and destination basis were determined simultaneously and changes in shipping costs had a greater impact on destination basis.

Several studies have examined the relationship between rail prices and basis levels to producers, which is important for nondelivery locations. In one of the first investigations to analyze the interrelationships between basis and shipping costs, Wilson and Dahl (2011) found that basis values have become more volatile over time and are affected by factors such as shipping costs, ocean rate spreads, export sales, and railroad performance. The econometric results indicated that the following variables were significant for explaining the variability of the origin basis values: shipping costs, Gulf–PNW ocean rate spreads, outstanding export sales, shipping-industry concentration, rail performance (measured as cars late), the ratio of stocks to storage capacity, futures prices, and varying measures for the futures and destination spreads. These results validated other studies about increased basis volatility and the importance of export sales' impact on interior basis values. The results also suggested that the performance of rail-car shipments was less of a determining factor for basis, whereas other studies found the influence of this factor to be much greater.

Seasonal Analog Analysis

Even though many commodity markets exhibit seasonal patterns, it is a commonly held belief that deviations from these seasonal patterns are driven by fundamental factors. In commodity analysis, unique seasonal patterns are often grouped into what are called seasonal analogs. These analogs are typically grouped based on a visual examination of seasonal year-on-year plots or through a rough application of correlation. The groupings may also be made based purely on a particular fundamental factor or set of factors. For example, one analog may be based on "large crop" years and another on "small crop" years. There is a common adage that "short crops have long tails" and that "long crops have short tails."

The use of formal analog forecasting methods has a long history in meteorology and climatology research, including Alexander et al. (2017), who used a method known as kernel analog forecasting to predict tropical oscillations. Djalalova, Delle Monache, and Wilczak (2015) used Kalman filtering and analogs to evaluate air quality forecasts. Finally, Comeau et al. (2019) used a prediction approach based on analog forecasting to analyze ice anomalies in the Arctic Ocean. The concept of analog years in weather has been used to correlate planting dates and corn yields (Elmore and Taylor, 2013).

In agricultural economics, most of the research related to seasonal analogs has focused on correlating weather analogs to crop yields and production (Hansen, Potgieter, and Tippett, 2004;

Menzie, 2007; Johansson et al., 2015; Irwin and Good, 2016). Some extension publications, such as Flaskerud and Johnson (2000), have published seasonal price indices based on crop fundamental analogs by grouping marketing years based on a fundamental factor (e.g., crop production). There have also been patents filed (Kolton, Gamboa, and Chimenti, 1996; Phillips et al., 2004) for systems that use analog techniques to forecast commodity prices. Some recent studies have examined the potential of using analog techniques in financial markets (Wanat, Śmiech, and Papież, 2016; Lahmiri, Uddin, and Bekiros, 2017).

Model Specification and Data

Model Specification

Basis for a spot market in a nondelivery location has specificity with respect to location, quality, and other delivery terms. An important factor that may affect export basis is shipping costs (Wilson and Dahl, 2011), which are volatile through time. It is common knowledge that specifications at export cash markets differ from those specified for par futures delivery. These variations often result in quality-related price differentials in export markets and have been fairly constant over time (Hertsgaard, Wilson, and Dahl, 2019).

For nonspot transactions calling for forward delivery, the supply of and demand for storage determines a market-equilibrium intermonth futures price spread, which has an effect on the intertemporal basis. The supply function for storage is affected, in part, by the convenience yield and is normally defined as the intermonth futures price differential. The concept of convenience yield originated in Working (1949) and was later described in most texts on futures markets (Hieronymus, 1977; Kolb and Overdahl, 2007; Hull, 2017). Importantly, the convenience yield is a component of the supply function for storage and, together with the demand for storage, determines the equilibrium interperiod futures price differential. The convenience yield relates to the value of holding grain in storage even if the market is inverted, due to the convenience of owning stocks and is a declining function of the stock level. The convenience yield differs from the spot basis at a nondelivery market, where basis reflects the cash price for immediate delivery and is specified as a period concurrent with the nearby futures. This is the basis value, which is the focus of this study. The exporters' value of the convenience may be related to the level of exports and the intermonth futures price differentials, which are included as explanatory variables in this study.

For an export market, in addition to the specificity indicated above, basis also responds to competition and market-specific costs, including competition from other export countries and internal competition from domestic users and internal and external logistics costs. For U.S. soybeans, the primary competing country is Brazil, where basis reflects the competitiveness of Brazilian soybean exports relative to competing export markets. Because soybeans are not stored at export locations, the spot basis mostly reflects shipping costs, the flow of the commodity through the marketing system, and quality differentials. In contrast, the futures carry is influenced by the supply of and demand for storage, including the convenience yield. A strong inverse carry in the futures-price spread is an indicator of strong nearby demand relative to current and anticipated flows in the marketing channel. Therefore, export basis strengthens in order to assure a sufficient flow of soybeans into the export channels when demand is strong and decreases in response to weak export demand.

Finally, export basis must exceed the basis value at domestic markets by at least the cost of shipping in order to attract grain into the export channels. These shipping costs include rail shipping (tariff plus the fuel surcharge and secondary rail car value) and barge rates. Export basis may also be influenced by the costs of ocean shipping, primarily to Asian destinations in the case of soybeans. Higher ocean freight costs are hypothesized to put downward pressure on export basis values in order to maintain competitiveness in the international market. However, this impact depends on the relative change in ocean shipping costs between U.S. export locations and international competitors.

This study's model specification builds on previous research by Wilson and Dahl (2011). Their models differ from our specification in that they focus on the origin basis, use pooled weekly data, do not examine interyear variability in basis, and implicitly assume that seasonality is homogeneous across years. However, Wilson and Dahl illustrated that logistical conditions, such as railcar shortages and secondary railcar market values, can have significant impacts on origin basis values. These variables also impact export basis and are therefore included in our analysis. Additionally, measures of current and anticipated export activity—such as export inspections, ships in port, and outstanding export commitments—may have a direct influence on export basis values and are included in our model specification.

This study specified and examined the following analytical models with regard to the annual average and seasonality of the export basis:

$$\begin{aligned} \bar{B}_{i,t} &= f(IntD_t, DomD_t, Trans_t, Logistic_t, ExpActivity_t), \text{and} \\ S(B_{i,t}) &= f(IntD_t, DomD_t, Trans_t, Logistic_t, ExpActivity_t), \end{aligned} \tag{1}$$

where i is a subscript for the particular export market (1 for Gulf and 2 for PNW), t represents the soybean marketing year beginning September 1 (2004/05–2015/16), $\bar{B}$ represents the marketing-year (MY) average export basis value, and $S(\cdot)$ represents a transformation of the monthly basis values into a particular seasonal-analog categorical variable. *IntD* is a set of explanatory variables representing international demand and competition for soybean exports; *DomD* is a set of explanatory variables representing the level of domestic demand and competition; *Trans* is a set of explanatory variables reflecting transportation costs for origin to export markets; *Logistic* is a set of explanatory variables representing the current state of logistical conditions; and *ExpActivity* is a set of explanatory variables measuring current levels of activity at the export ports.

Dependent Variables

We define and use the term export basis as our dependent variable. These values are for basis at the geographic markets, which are referred to as the U.S. Gulf and Pacific Northwest (PNW), areas that represent the majority (84.7% of total export volume in 2017) of soybeans exported from the United States. However, the values used need a technical clarification. Free-on-board (FOB) prices are not routinely reported for these markets. Therefore, we use what are referred to as "CIF NOLA" and "Track PNW" as the basis values delivered by barge to New Orleans, Louisiana, and by rail to Portland, Oregon, respectively. As such, these basis values are a near-perfect representation of the export FOB basis, omitting only the trader's margin. Further, we only use the "spot" basis for immediate shipment at both markets relative to the nearby futures delivery month (i.e., rollover occurs the business day prior to First Notice Day).

For the dependent variables, we obtained weekly nearby basis data from TradeWest Brokerage for the weeks covering the 2004/05–2015/16 marketing years for both the Gulf (NOLA) and Pacific Northwest (PNW) export markets. Missing values (64 weeks total between the two series) were interpolated using the non-linear iterative partial least squares (NIPALS) procedure (Wold, 1973). We then converted the data to monthly and marketing year (September through August) averages for use in the analytical model. The monthly data were used to derive the MY seasonal-analog classifications by applying agglomerative hierarchal clustering (AHC). Some of the explanatory variables used in the model were only reported on a MY total or average basis; therefore, the partial least squares regression (PLS-R) and seasonal-analog model specification z-tests were based on a MY time increment.

The full dataset covers the 2004/05–2015/16 marketing years. Growth of the PNW as a major U.S. soybean export market corresponds with the emergence of China as a major international importer of soybeans beginning in 2001/02. Price quotes for PNW soybeans are reported less frequently (due to market liquidity issues) prior to the 2004/05 marketing year, making it very

Table 1. Explanatory Variables Used in the Analysis

Variables	Description	Source	Aggregation
International demand (*IntD*)			
Basis-Brz	Brazil soybean basis, Paranagua, CME futures (US$/bu)	CME, Cepea	MY avg. of weekly values
SA-Prod	Total South American (Argentina–Brazil–Paraguay) soybean production (mmt)	USDA-WASDE	MY reported values
China-Import	China total soybean imports (mmt)	USDA-WASDE	MY reported values
World-SU	World soybean ending stocks-use ratio (%)	USDA-WASDE	MY reported values
Domestic demand (*DomD*)			
Futures-NB	Nearby CBOT soybean futures price (c/bu)	CME, DTN ProphetX	MY avg. of weekly values
FutSprd1	2nd NB futures minus NB soybean futures price (c/bu)	CME, DTN ProphetX	MY avg. of weekly values
FutSprd2	3rd NB futures minus 2nd NB soybean futures price (c/bu)	CME, DTN ProphetX	MY avg. of weekly values
SU-Ratio	U.S. soybeans ending stocks-use ratio (%)	USDA-WASDE	MY reported values
MealP	U.S. domestic soybean meal price ($/ton)	USDA-WASDE	MY reported values
OilP	U.S. domestic soybean oil price (c/lb)	USDA-WASDE	MY reported values
Crush	U.S. domestic soybean crush as percentage of total supply	USDA-WASDE	MY reported values
Transportation costs (*Trans*)			
Rail-Gulf	Total rail cost, shuttle trains (tariff plus fuel surcharge), Freemont, NE to Texas Gulf ($/car)	BNSF Railroad	MY avg. of weekly values
Rail-Sprd	PNW versus Gulf rail cost spread, shuttle trains, Freemont, NE ($/car)	BNSF Railroad	MY avg. of weekly values
Barge-Spot	Spot barge rate, St. Louis to Gulf ($/ton)	USDA-AMS	MY avg. of weekly values
Barge-3M	3-month forward barge rate, St. Louis to Gulf ($/ton)	USDA-AMS	MY avg. of weekly values
Ocean-PNW	Ocean freight rate from PNW to Japan ($/mt)	USDA-AMS	MY avg. of weekly values
Ocean-Sprd	Ocean freight spread to Japan, Gulf vs. PNW ($/mt)	USDA-AMS	MY avg. of weekly values
Logistic conditions (*Logistic*)			
Cars-Late	Average BN railcars placed late (cars)	BNSF Railroad	MY avg. of weekly values
DCV	Daily secondary market car values for shuttle trains ($/car)	TradeWest Brokerage	MY avg. of weekly values
FarmDel-Q1	Cumulative farmer delivery % through Q1 of MY	USDA-NASS	Cumulative sum of monthly values
FarmDel-Q2	Cumulative farmer delivery % through Q2 of MY	USDA-NASS	Cumulative sum of monthly values
FarmDel-Q3	Cumulative farmer delivery % through Q3 of MY	USDA-NASS	Cumulative sum of monthly values
Level of export activity (*ExpActivity*)			
Gulf-InPort	Average number of ships in Gulf ports (ships)	USDA-AMS	MY avg. of weekly values
PNW-InPort	Average number of ships in PNW ports (ships)	USDA-AMS	MY avg. of weekly values
Export-Gulf	FGIS export inspections at Gulf ports (1,000 bushels)	USDA-FGIS	MY avg. of weekly values
Export-PNW	FGIS export inspections at PNW ports (1,000 bushels)	USDA-FGIS	MY avg. of weekly values
Export-Out	U.S. soybean export sales-outstanding balance (1,000 bushels)	USDA-FAS	MY avg. of weekly values

difficult to establish a consistent weekly and monthly time series for the prior marketing years. Further, the tariff war with China that commenced in 2018 had a significant impact on world and U.S. soybean markets. These effects had the potential of causing a significant structural break in U.S. soybean export markets, basis, and trade flows. Therefore, the end of the dataset (2015/16) represents the last complete marketing year for soybeans prior to the 2016 election.

Independent (Explanatory) Variables

Table 1 reports a complete list of the potential explanatory variables, sources, and levels of aggregation. All the listed variables are rolled up to a MY average. We chose Freemont, Nebraska, as the interior point for the rail cost variables due to its position relative to the Gulf and PNW markets as a tributary shipper to both markets.

China represents the major source of international soybean demand; therefore, we included MY total soybean imports (*China-Import*) as the primary measure. Brazil is the major source of competition for U.S. soybean exports; therefore, we used its MY-average export basis for the port of Paranaguá (*Basis-Brz*) as a primary measure of international competition. Previous studies have used total South American production (*SA-Prod*) as a measure of international competition. We used the domestic soybean crush (*Crush*) to represent a dominant source of domestic soybean demand. The two primary by-products of the soybean crush are meal (*MealP*) for livestock feed and oil (*OilP*) for food and industrial (biodiesel) use; therefore, we included their prices as a measure of the domestic competition for the flow of soybeans. Table 1 defines the other variables.

Methodology

Marketing-Year Average Basis Models

The dataset contained 2 dependent variables and 27 potential explanatory variables observed over 12 marketing years. Due to the overidentification of the explanatory-variable matrix, an ordinary least squares regression cannot be applied. Additionally, we observed a high degree of correlation between individual variables in the explanatory dataset, indicating the presence of multicollinearity in the dataset.

We used the PLS-R model (Wold, 1966) to estimate the impact of the explanatory variable set on the MY average basis for both export markets. PLS-R is particularly useful when predicting a set of dependent variables from a large set of independent variables (Abdi, 2007) and has found numerous applications in chemometrics (Wold, 2001) and sensory evaluation (Martens and Næs, 1989). PLS-R has also been utilized frequently in the social sciences as a multivariate tool for examining both nonexperimental and experimental data within a structural equation modeling (PLS-SEM) framework (Hair Jr et al., 2014).

Traditionally, finding solutions for regression problems in the presence of multicollinearity and/or data sparsity follows one of two approaches: (i) removing the highly correlated predictors with one of several techniques or (ii) conducting principal component analysis (PCA) on the explanatory variables and regressing the dependent variables on the extracted principal components (Kuhn and Johnson, 2013), a technique referred to as principal component regression (PCR; Massy, 1965). PCR has been one of the most commonly used procedures in social science research, but a disadvantage of this method is that it is an unsupervised procedure in that it only considers information in the explanatory variables when constructing the principal components. If significant differences in variability between the explanatory variable and dependent variable space exist, then PCR has a high probability of not correctly identifying all the data's predictive relationships since it is completely focused on decomposing the variance of the explanatory variables only.

Partial least squares regression (PLS-R), on the other hand, is a supervised procedure because it considers information in both the dependent and explanatory variable sets. The PLS-R method extracts its components (called latent variables) in the direction of optimizing the covariance between the dependent variable(s) and the explanatory variables. PLS-R also has the advantage of considering multiple dependent variables because the components are directed toward explaining the covariance between the dependent ($\boldsymbol{Y}$) and explanatory ($\boldsymbol{X}$) data matrices.

The general underlying model for multivariate PLS-R simultaneously decomposes the explanatory ($\boldsymbol{X}$) and dependent ($\boldsymbol{Y}$) variable matrices as follows (Esbensen and Swarbrick, 2018):

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{T}\boldsymbol{P}^{T} + \boldsymbol{E}, \\ \boldsymbol{Y} &= \boldsymbol{U}\boldsymbol{Q}^{T} + \boldsymbol{F}, \end{aligned} \tag{2}$$

where $\boldsymbol{X}$ is an $n \times m$ matrix of explanatory variables with n equal to the number of observations and m equal to the number of explanatory (independent) variables; $\boldsymbol{Y}$ is an $n \times p$ matrix of dependent variables with p equal to the number of dependent variables; $\boldsymbol{T}$ and $\boldsymbol{U}$ are $n \times r$ matrices that contain the $\boldsymbol{X}$ and $\boldsymbol{Y}$ scores, respectively, with r equal to the number of retained latent (component) variables; $\boldsymbol{P}$ and $\boldsymbol{Q}$ are $m \times r$ and $p \times r$ orthogonal loading matrices, respectively; and $\boldsymbol{E}$ and $\boldsymbol{F}$ represent the $n \times m$ matrices of error terms, which are *i.i.d.* random normal variables.

The decompositions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ illustrated in equation (2) are directed with the goal of maximizing the covariance between the $\boldsymbol{T}$ and $\boldsymbol{U}$ matrices. The regression coefficients for the latent variables can then be derived by finding an $r \times r$ matrix of regression coefficients ($\boldsymbol{\beta}_{PLS}$) such that

$$\boldsymbol{U} = \boldsymbol{T}\boldsymbol{\beta}_{PLS}. \tag{3}$$

Equation (3) implies that $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{P}$; therefore, the individual PLS-R regression coefficients for each explanatory variable are the sum product of the variable's loadings (from $\boldsymbol{P}$) and the PLS-R latent-variable regression coefficients ($\boldsymbol{\beta}_{PLS}$). Given that PLS-R results are highly sensitive to differences in unit measurement of the explanatory variables, it is standard practice to convert all variables to a z-score equivalent by normalization prior to model estimation.

When applying PLS-R, it is common practice to use a cross-validation technique, such as jackknife leave-one-out (LOO), to determine the number of retained components (latent variables) in the model and to generate the sample goodness-of-fit statistics for the model. These techniques typically rely wholly or in part on the fitted model's out-of-sample prediction characteristics. Most cross-validation techniques typically divide the dataset observations into an estimation (or tuning) dataset to approximate the model and a testing dataset to compute the out-of-sample prediction statistics.

Cross-validation procedures can be either exhaustive or nonexhaustive; exhaustive methods utilize all possible dataset divisions, while nonexhaustive methods do not. The jackknife LOO cross-validation procedure utilized in this study is an exhaustive technique that sequentially removes each single observation from the testing dataset and then uses the remaining observations as the estimation dataset. This process is conducted until all observations have been used in the testing dataset.

In most PLS-R applications, the number of latent variables retained in the model is usually determined by taking the maximum value of a goodness-of-fit metric. This metric could be an in-sample measure such as the proportion of variance explained in the $\boldsymbol{Y}$ matrix by the model (the R^2Y statistic). Or the metric could be an out-of-sample measure (using cross-validation) such as taking the minimum of the predictive residual sum of squares (PRESS) statistic or by taking the maximum of the Q^2 quality index statistic. The Q^2 statistic is, essentially, the same as the traditional regression R^2 statistic, with the PRESS statistic substituting for the residual sum of squares (RSS). The individual, explanatory variable t-statistics are derived using the application of jackknife LOO sampling to derive the sample statistics. In this study, we used the maximum Q^2 quality index criteria to determine the optimal number of retained latent variables for the model.

A number of procedures are available to determine the optimal scoring and loading matrices for PLS-R regression as specified in equation (2), the most popular of which is the NIPALS algorithm originally developed by Wold (1973). The implementation used in this study was the standard PLS-R procedure from the *XLStat* software package (Addinsoft, Inc., 2019), which is based on the NIPALS algorithm.

A useful output of the PLS-R regression model is an index assigned to each explanatory variable, called the variable importance in projection (VIP), which can be expressed by the following equation

(Wold, Sjostrom, and Eriksson, 1993):

$$VIP_j = \sqrt{\frac{\sum_{i=1}^{h} R^2(y,t_i)(w_{ij}/\|w_i\|)^2}{(1/p)\sum_{i=1}^{h} R^2(y,t_i)}}, \tag{4}$$

where p is the number of predictor variables, h is the number of retained component variables, w_{ij} is the weight of the jth predictor variable in component variable i, and $R^2(y_,t_i)$ is the fraction of variance in $\boldsymbol{Y}$ explained by component i. The VIP score represents the proportion of explained variance for $\boldsymbol{X}_j$ relative to $\boldsymbol{Y}$ (through the component variables), divided by the average explained variance between all $\boldsymbol{X}$ variables and $\boldsymbol{Y}$. By definition, the average squared VIP score is equal to 1, therefore, a typical rule of thumb used for variable reduction and retention is to keep all variables with a VIP score of greater than 1 for subsequent PLS-R estimation. This is the variable reduction approach used in this study.

The PLS-R procedure utilized in this study provides a statistically valid and effective approach to determine the significance and relative influence of the explanatory variables for predicting the average export basis level in both markets, despite the large number of explanatory variables (27) relative to the number of observations (12). PLS-R is similar to PCR, an accepted approach to handle multicollinearity (Amemiya, 1985) in econometrics research, but PLS-R has the additional advantage of incorporating information from both the $\boldsymbol{Y}$ and $\boldsymbol{X}$ matrices in constructing the component variables. PLS-R is a widely accepted and validated statistical technique for handling the data sparsity issue that arises frequently in other scientific disciplines, such as chemometrics (Wold, 2001) and genomics (Tenenhaus et al., 2010).

Utilizing the jackknife LOO cross-validation procedure adds additional statistical rigor to the derivation of the PLS-R regression metrics (Q^2 and VIP indices) and the individual coefficient standard errors and t-statistics. Additionally, given the high level of correlation between the two export basis markets, the PLS-R procedure has the added advantage of supporting simultaneous estimation of regression equations for both markets by incorporating them into a single $\boldsymbol{Y}$ matrix.

Seasonal-Analog Models

To analyze basis seasonality, we converted the monthly basis values into additive seasonal indices by subtracting the monthly basis value from the MY average. We then grouped MY basis patterns into similar seasonal-pattern analogs by applying AHC (Ward, 1963) to the index data. The AHC procedure is an iterative classification method that starts by calculating the dissimilarity in seasonal patterns among the 12 marketing years. The proximity between marketing years is measured using the Euclidian distance metric. The first 2 marketing years are clustered based on the minimization of Ward's agglomeration criterion, which aggregates such that within-group inertia increases as little as possible. Then, the iterative process continues by calculating the dissimilarity between this first class and the remaining 10 marketing years. This process continues until all the objects have been clustered.

The successive clustering operations produce a binary clustering tree called a dendrogram, in which the root is the class that contains all marketing years. This tree graphically represents the partitions' hierarchy. We chose the final set of analogs by truncating the dendrogram at a determined level using the minimum-entropy criterion to determine the point of truncation. An index originally developed by Shannon (1948) was used as the measure of entropy.

To characterize the individual seasonal analogs, we applied a two-sample z-test originally proposed by Lebart, Morineau, and Piron (2000). The test statistic is similar to a classic t-test but is applied to test the difference in means between a particular set and subset of observations. Because the two estimates are correlated, a classic t-test cannot be used. The test statistic is

$$z_k(X) = \frac{\bar{X}_k - \bar{X}}{s_k(X)}, \tag{5}$$
$$\text{where } s_k^2(X) = \frac{n - n_k}{n - 1} \times \frac{s^2(X)}{n_k},$$

with $\bar{X}_k$ and $\bar{X}$ equal to the sample means from the subset and parent set, respectively; n_k and n equal to the number of observations in the subset and parent set, respectively; and $s^2(X)$ equal to the parent set's variance. The z_k statistic is distributed as asymptotically unit normal, so it can be treated as a standard z-statistic when determining statistical significance.

Results

Market-Year Average Basis for the Gulf and the PNW

When calculating the MY average basis levels for the Gulf and PNW soybean export markets, there was a statistically significant change in the mean basis level following the 2007/08 marketing year for both markets. To correct for this shift in the mean, we augmented the explanatory dataset (Table 1) using a dummy variable (*Prior to 2008?*) that is equal to 1 for marketing years prior to 2008/09, and 0 otherwise. We also added a linear trend variable (*Trend*) to correct for any gradual changes in basis. These observed changes in the mean basis level can be attributed, in part, to the radical change and increased volatility for all commodity markets following the 2008/09 marketing year along with the growing trend in U.S. soybean exports, particularly to China.

The mean basis values for marketing years 2004/05–2007/08 were \$0.38/bushel and \$0.55/bushel for the Gulf and the PNW, respectively. For the subsequent 2008/09–2015/16 period, the average basis levels were \$0.80/bushel and \$1.14/bushel for the Gulf and the PNW, respectively. We applied a two-sample, one-tail t-test with the null hypothesis of no difference between the means versus the alternate hypothesis that the mean in the first period (2004/05–2007/08) was significantly lower than the latter (2008/09–2015/16). The t-test rejected the null hypothesis in favor of the alternative hypothesis at the 99% confidence level for both the Gulf and PNW markets.

The observed volatility of export basis has escalated over time in both markets. The standard deviation of basis (derived using monthly data) for the period prior to the 2008/09 marketing year was \$0.0953/bushel and \$0.1858/bushel for the Gulf and the PNW, respectively. For the subsequent period (2008/09–2015/16), these values increased to \$0.2403/bushel and \$0.2672/bushel for the Gulf and the PNW, respectively. This result was similar to the results observed previously in Wilson and Dahl (2011). Application of Fisher's F-test to the basis variance (H$_a$: The first-period variance is less than that in the latter period.) indicated that only the Gulf basis variance was significantly lower (at the 90% confidence level) in the first period, while the variance difference for the PNW was not significant at the 90% level.

The MY average basis values for the Gulf and the PNW were highly correlated (93.0%) over the 12 MY observations; therefore, we included both basis series in the $\boldsymbol{Y}$ matrix for the PLS-R estimation. For the following analysis, we used a two-step procedure. First, the PLS-R model was applied to the full explanatory dataset. Variables with VIP scores greater than 1 were retained in the dataset for subsequent estimation, while variables with values less than 1 were removed. The second step of the procedure applied PLS-R to the reduced dataset for the final estimation of the regression coefficients and the interpretation of significance. We determined the number of PLS-R components retained by using the jackknife LOO cross-validation procedure, with the number of components (latent variables) determined by the maximum value of the Q^2 quality index statistic.

Applying the PLS-R procedure to the full explanatory dataset resulted in one retained component (latent variable) with an optimal Q^2 statistic of 0.6555 using the cross-validation procedure. The

Table 2. PLS-R Regression Results for Gulf Market-Year Average Soybean Basis

Variable	Coefficient	Standardized Coeff.	Std. Dev.	Lower Bound (90%)	Upper Bound (90%)	*T*-Statistic	Significance
Intercept	−1.475	n/a	14.770	−28.001	25.051	−0.100	0.922
Basis-Brz	0.072	0.092	0.024	0.029	0.116	2.973	0.013
FutSprd1	−0.180	−0.092	0.070	−0.305	−0.056	−2.594	0.025
FutSprd2	−0.167	−0.091	0.072	−0.295	−0.038	−2.331	0.040
MealP	0.026	0.090	0.006	0.014	0.037	4.037	0.002
Rail-Gulf	0.004	0.086	0.001	0.003	0.006	5.612	0.000
Prior to 2008?	−4.787	−0.082	0.987	−6.560	−3.014	−4.850	0.001
PNW-InPort	0.842	0.079	0.374	0.171	1.513	2.253	0.046
Export-Out	0.024	0.079	0.005	0.014	0.034	4.429	0.001
Trend	0.623	0.078	0.099	0.445	0.802	6.270	0.000
China-Import	0.112	0.077	0.021	0.075	0.149	5.408	0.000
SU-Ratio	−0.461	−0.076	0.170	−0.766	−0.156	−2.714	0.020
Rail-Sprd	0.009	0.075	0.003	0.005	0.014	3.627	0.004
Futures-NB	0.006	0.066	0.002	0.003	0.009	3.648	0.004

value of the Q^2 statistic indicated that the retained latent variable explained approximately 65.5% of the variability with the out-of-sample values for the ***Y*** matrix. The R^2Y statistic was equal to 0.788, indicating that the retained latent variable explained approximately 78.8% of the in-sample variability in the ***Y*** matrix.

The VIP index scores for the explanatory variables from the initial application of the PLS-R regression model to the Gulf and PNW average basis levels indicated that the Brazilian export basis (*Basis-Brz*) was the most important variable for projecting both Gulf and PNW basis levels. Following in importance were the nearby futures spreads (*FutSprd1* and *FutSprd2*), the domestic soybean meal price (*MealP*), and the total rail shipping costs from Freemont, Nebraska, to the Gulf (*Rail-Gulf*). Less than half (13 of 29) of the variables had a VIP score greater than 1 and were retained for the final regression-estimation procedure. The additional variables retained (in order of their VIP scores) were *Prior to 2008?*, *PNW-InPort*, *Export-Out*, *Trend*, *China-Import*, *SU-Ratio*, *Rail-Sprd*, and *Futures-NB*.

The second round of the PLS-R estimation regressed the ***Y*** matrix of MY average basis values for the Gulf and PNW export markets on the one retained latent variable component derived from optimizing the covariance between the ***Y*** matrix and the VIP-reduced explanatory variable set (***X*** matrix). This regression resulted in a quality index (Q^2) statistic of 0.7198, which indicated an improvement in the model's out-of-sample predictability (6.43% gain in Q^2 compared to the initial estimation). The in-sample R^2Y equaled 0.798, indicating that the retained latent variable accounted for almost 80% of the variability in the ***Y*** matrix containing both the Gulf and PNW average basis levels.

Tables 2 and 3 show the PLS-R regression equations for the Gulf and the PNW, respectively. The Gulf basis regression fit had an R^2 coefficient of 0.7989 with a root mean squared error (RMSE) of $0.123/bushel (both in-sample). The PNW basis regression fit had an R^2 coefficient of 0.7971 with an RMSE of $0.16/bushel. The tables present both the nonstandardized and standardized coefficient estimates, with the explanatory variables ordered by the absolute value of their standardized coefficient estimates.

The very high level of correlation (93%) between the average basis values for the Gulf and the PNW resulted in only one latent component variable retained in the model based upon the optimal value of the Q^2 quality index. Therefore, the regression results for the two export markets were highly similar, with a few exceptions. First, the intercept for the PNW equation had a premium of $0.0825/bushel over the intercept in the Gulf equation. However, the high standard deviations for both coefficients indicated that the difference was not statistically significant. Second, the regression

Table 3. PLS-R Regression Results for Pacific Northwest Market-Year Average Soybean Basis

Variable	Coefficient	Standardized Coeff.	Std. Dev.	Lower Bound (90%)	Upper Bound (90%)	*T*-Statistic	Significance
Intercept	6.778	n/a	14.803	−19.807	33.363	0.458	0.656
Basis-Brz	0.093	0.092	0.031	0.038	0.148	3.056	0.011
FutSprd1	−0.233	−0.092	0.084	−0.384	−0.083	−2.783	0.018
FutSprd2	−0.215	−0.090	0.085	−0.369	−0.062	−2.523	0.028
MealP	0.033	0.090	0.007	0.021	0.045	4.870	0.000
Rail-Gulf	0.005	0.086	0.001	0.004	0.007	6.868	0.000
Prior to 2008?	−6.186	−0.082	1.575	−9.014	−3.358	−3.928	0.002
PNW-InPort	1.088	0.079	0.454	0.274	1.903	2.399	0.035
Export-Out	0.031	0.079	0.006	0.020	0.042	4.964	0.000
Trend	0.805	0.078	0.120	0.589	1.022	6.682	0.000
China-Import	0.145	0.077	0.026	0.098	0.192	5.498	0.000
SU-Ratio	−0.595	−0.076	0.231	−1.009	−0.181	−2.583	0.025
Rail-Sprd	0.012	0.074	0.003	0.008	0.017	4.819	0.001
Futures-NB	0.008	0.065	0.002	0.004	0.012	3.978	0.002

coefficient estimates for the PNW equation were marginally higher in magnitude in absolute value but had the same signs as the Gulf equation, indicating a slightly higher impact for each variable in the PNW equation. However, this difference can be mostly attributed to a higher variability in the PNW basis since the standardized coefficients were nearly identical between the two equations. When examining the dependent variables' loadings on the one retained latent component, the Gulf (0.7989) had a higher loading compared to the PNW (0.7971). From an out-of-sample forecasting perspective, using the jackknife LOO cross-validation procedure, the PNW had a slightly higher Q^2 value (0.7263) versus the Gulf (0.7134). The coefficient standard errors and t-statistics for the PLS-R estimation procedure were estimated directly from the cross-validation procedure; therefore, the coefficients estimated for the PNW equation had slightly higher (in absolute value) t-statistics compared to the Gulf equation.

From these statistical results, the following observations can be made: First the ranking and impacts of the explanatory variables were nearly indistinguishable between the two markets. However, the magnitude of the impact of changes in the explanatory variables was slightly higher for the PNW compared to the Gulf. This partially explains the perceived higher volatility of basis in the PNW compared to the Gulf.

Second, the most important explanatory variables for both markets, based upon the magnitude of the standardized coefficient estimates, were export basis in Brazil (*Basis-Brz*) and variables reflecting competition from the domestic market (*FutSprd1*, *FutSprd2*, and *MealP*). Export basis responds to competitor basis, as noted previously, which indicates that Gulf and PNW basis primarily respond to maintain competitiveness in both the international and domestic markets. Basis in both markets also responds in a positive and statistically significant manner to the level of Chinese soybean imports (*China-Import*) but at a slightly lower magnitude based on its standardized coefficient value.

Third, the growth in Chinese export demand increased the relative importance of the PNW market over time, despite the fact that the volume of exports moving through the PNW was consistently less than the volume moving through the Gulf. Activity directly related to the export volume from the PNW (*PNW-InPort*) had more influence on the basis levels at both the Gulf and the PNW compared to a similar measure of activity (*Gulf-InPort*) from the Gulf.

Fourth, internal costs of logistics were of secondary importance and were mainly limited to rail costs (*Rail-Gulf* and *Rail-Sprd*). The export basis level responds positively to these costs in order to assure adequate export market flows and supplies. The absence of barge costs and the positive and significant signs of both rail cost variables in both export equations indicate that the average basis

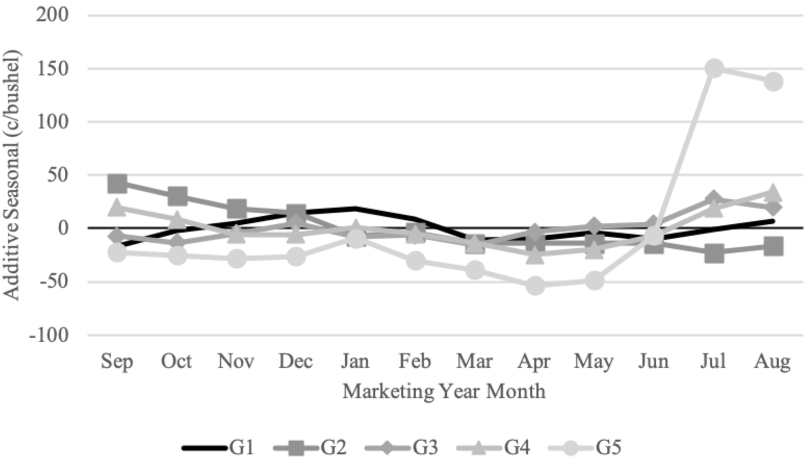

Figure 1. Seasonal Analogs for Gulf Soybean Basis

level in both markets primarily responds directly to changes in the total rail cost to the PNW market (i.e., *Rail-Gulf* + *Rail-Sprd*).

Both the early-period dummy variable (*Prior to 2008?*) and the MY trend variable (*Trend*) were highly significant (99% confidence level) in both equations and had the anticipated signs (negative for *Prior to 2008?* and positive for *Trend*). These results were important, confirming that the average export basis levels shifted permanently higher in the period following the 2007/08 marketing year.

Seasonal-Analog Derivation

For each basis market and marketing year, we calculated a monthly additive seasonal index by subtracting the MY average basis from the monthly average basis value. Visual examination of the plots of the additive seasonal indices by month and marketing year showed no distinct prevailing seasonal pattern in the data for either export market. Application of AHC with the minimum-entropy criterion produced five distinct seasonal-analog groupings for the Gulf basis market and four distinct groupings for the PNW basis market.

Figure 1 shows the average seasonal indices for each of the Gulf's five seasonal analogs. Analog G1 (2004/05, 2005/06, and 2011/12) has a typical pattern with a relatively stable basis level throughout the marketing year, with a slight seasonal increase from September through January and a decline through March, followed by a slight increase from June through August. Analog G2 (2006/07 and 2014/15) has a weakening (declining) pattern for basis throughout the marketing year. Analog G3 (2007/08, 2010/11, and 2012/13) has a relatively stable pattern from September through June, with a slight increase for basis in the final 2 months of the marketing year. Analogs G4 (2008/09, 2009/10, and 2015/16) and G5 (2013/14) are similar in that they show a general weakening of basis through April, with a strengthening through the end of the marketing year. The main difference is that G5 (an outlier) shows much more volatile swings in basis compared to G4. Generally, a cluster that contains one observation is considered a potential "outlier" observation. The 2013/14 marketing year, with its extreme fluctuation in monthly basis levels, is assigned to a singular analog (G5).

Figure 2 shows profiles for the PNW's four seasonal analogs. Analog P1 (2004/05, 2008/09, 2010/11, 2012/13, and 2015/16) is similar to G1 in that the basis becomes stronger through the first 4 months of the marketing year (through January), with a weakening through June and a strengthening pattern through the final 2 months of the marketing year. This behavior is probably the most typical pattern for this market. Analog P2 (2005/06, 2006/07, and 2009/10) has a general weakening of basis throughout the marketing year, with a slight uptick in the final month. Analog P3 (2007/08, 2011/12,

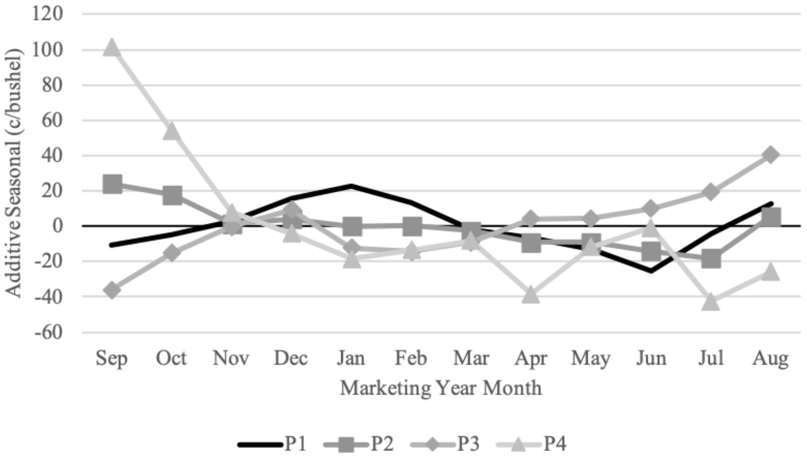

Figure 2. Seasonal Analogs for Pacific Northwest Soybean Basis

and 2013/14) does not correspond with any of the Gulf patterns, showing a substantial strengthening in basis through the first 3 months, followed by a relatively stable pattern before basis strengthens again in the final 3 months. Analog P4 (2014/15), as with G5, is a single-year outlier analog with a highly variable pattern; however, P4 is characterized by a sharp weakening of basis in the first 2 months followed by an uneven weakening pattern for the remainder of the marketing year.

Statistical Characterization of Basis Seasonal Analogs for the Gulf and the PNW

While the MY average basis levels were highly correlated between the two export markets, the seasonal analogs did not exhibit statistically significant correlation. We conducted a χ^2 contingency table test to measure the dependence between the analog categories for the two basis markets. The χ^2 test value of 12.53 was short of the critical value of 18.55 at the 90% confidence level.

Table 4 reports the Lebart z-scores for the Gulf and the PNW basis analogs. The sign for the z-score indicates whether the analog mean was less than (negative) or greater than (positive) the overall mean value for the explanatory variable. The test is two-tailed; therefore, z-scores exceeding 1.64 in absolute value are significant at the 90% confidence level. Absolute scores that exceed 1.96 are significant at the 95% level, and those that exceed 2.56 are significant at the 99% level.

For the Gulf basis, the results indicate that analog G1 was mostly associated with marketing years that exhibited lower-than-average weekly ships in port at the Gulf (*Gulf-InPort*), lower-than-average South American soybean production (*SA-Prod*), lower-than-average weekly outstanding exports (*Export-Out*), and lower-than-average weekly export inspections at the PNW (*Export-PNW*). Analog G2 was characterized by years with a higher-than-average world soybean stocks-use ratio (*World-SU*). Analog G3 was characterized by higher-than-average domestic soybean oil prices (*OilP*), higher-than-average nearby futures prices (*Futures-NB*), and higher-than-average Gulf ocean freight costs relative to the PNW (*Ocean-Sprd*). Analog G4 was characterized by lower-than-average weekly railcars placed late (*Cars-Late*). Analog G5 was characterized by higher-than-average secondary railcar values (*DCV*), lower-than-average second nearby futures carry spreads (*FutSprd2*), a higher-than-average weekly number of railcars placed late (*Cars-Late*), a higher-than-average weekly number of ships at the PNW port (*PNW-InPort*), a lower-than-average nearby futures carry spread (*FutSprd1*), a higher-than-average weekly number of ships at the Gulf port (*Gulf-InPort*), and a higher-than-average domestic soybean-meal price (*MealP*). All of these mean differences were significant at the 90% confidence level or higher.

Table 4. Variable Characterization Test Z-Scores (2-Tailed) for Seasonal Analogs

	Gulf Analogs					PNW Analogs			
Variable	**G1**	**G2**	**G3**	**G4**	**G5**	**P1**	**P2**	**P3**	**P4**
Futures-NB	−1.322	−0.967	1.984*	−0.462	0.991	0.174	−1.872*	1.814*	−0.219
FutSprd1	0.854	1.002	−0.356	−0.116	−1.951*	−0.796	1.417	−0.680	0.264
FutSprd2	0.756	1.153	−0.599	0.309	−2.286**	−0.308	1.480	−1.433	0.477
Basis-Brz	−0.522	−0.528	−1.062	1.061	1.530	0.350	−0.845	0.059	0.606
Rail-Gulf	−1.432	−0.189	0.695	0.101	1.252	0.081	−1.706*	1.060	0.868
Rail-Sprd	−1.628	0.624	0.715	−0.532	1.422	−1.050	−1.207	1.509	1.399
Barge-Spot	−0.609	0.447	1.077	−1.598	1.166	−1.304	−0.014	0.877	0.973
Barge-3M	−1.355	0.806	1.125	−1.021	0.873	−1.666*	0.203	1.108	0.917
Ocean-PNW	−0.314	−0.020	1.404	−0.868	−0.321	−1.280	0.533	1.484	−0.876
Ocean-Sprd	−0.689	−1.008	1.909*	−0.430	0.120	−0.683	−0.669	1.897*	−0.707
Gulf-InPort	−2.378**	0.105	−0.142	1.222	1.892*	−0.292	−0.345	0.238	0.687
PNW-InPort	−1.292	−0.781	0.581	0.059	2.075**	0.208	−1.235	0.756	0.379
Cars-Late	0.204	0.485	−0.239	−1.756*	2.153**	−1.481	−0.027	1.020	1.088
DCV	−0.529	0.146	−0.812	−0.728	3.045***	−1.216	−0.239	1.217	0.635
Export-Gulf	−1.429	0.516	−0.532	1.125	0.615	0.675	−1.334	−0.524	1.705*
Export-PNW	−1.789*	0.541	−0.280	1.352	0.395	−0.066	−1.270	0.267	1.690*
Export-Out	−1.836*	0.237	0.408	0.378	1.327	0.001	−1.444	0.580	1.352
FarmDel-Q1	−1.625	1.461	1.520	−1.415	0.411	−0.967	−0.786	0.472	2.217**
FarmDel-Q2	−0.931	0.687	1.020	−1.463	1.227	−1.155	−0.488	1.020	1.227
FarmDel-Q3	−0.235	0.494	−0.101	−0.772	1.070	−1.013	−0.369	1.243	0.438
Crush	1.214	−1.243	0.888	−0.665	−0.576	−0.347	0.254	1.048	−1.421
SU-Ratio	1.008	1.497	−0.613	−1.056	−0.982	−0.891	2.322**	−1.030	−0.435
MealP	−1.533	−0.623	1.101	−0.104	1.679*	0.083	−1.927*	1.562	0.423
OilP	−0.851	−0.865	2.422**	−0.872	0.073	−0.103	−1.332	1.804*	−0.556
World-SU	−0.593	1.676*	0.026	−0.524	−0.551	−0.522	1.604	−1.441	0.678
SA-Prod	−1.862*	0.911	0.208	0.266	0.947	−0.129	−0.838	−0.114	1.722*
China-Import	−1.378	0.178	−0.131	0.707	1.016	0.176	−1.581	0.461	1.442

Notes: Single, double, and triple asterisks (*,**,***) indicate significance at the 10%, 5%, and 1% level. *Z*-scores represent two-tail test of null hypothesis that analog mean equals population mean.

For the PNW basis, analog P1 was characterized by a lower-than-average weekly forward barge rate (*Barge-3M*). Analog P2 was characterized by a higher-than-average domestic soybean stocks-use ratio (*SU-Ratio*), a lower-than-average domestic soybean-meal price (*MealP*), a lower-than-average nearby futures price (*Futures-NB*), and a lower-than-average weekly rail cost from Freemont, Nebraska, to the Gulf (*Rail-Gulf*). Analog P3 was characterized by a higher-than-average weekly ocean freight cost for the Gulf relative to the PNW (*Ocean-Sprd*), a higher-than-average nearby soybean futures price (*Futures-NB*), and a higher-than-average domestic soybean oil price (*OilP*). Analog P4 was characterized by a higher-than-average percentage of farmer deliveries in quarter 1 of the marketing year (*FarmDel-Q1*), higher-than-average South American soybean production (*SA-Prod*), and higher-than-average export inspections at both the Gulf and the PNW (*Export-Gulf* and *Export-PNW*). All of these mean differences were significant at the 90% confidence level or higher.

For all seasonal analogs, the average Brazilian export basis and the volume of Chinese soybean imports were noticeably absent as statistically significant factors. This represents a significant difference compared to the reported PLS-R results applied to the market year average basis levels. From the *z*-test results, the primary factors influencing the choice of seasonal analog included all logistics costs (barge and ocean rates in addition to rail costs), logistics conditions (cars placed late, secondary railcar market values, and pace of farmers' marketing), and port-specific export activity (ships in port and export inspections).

Summary and Conclusions

Basis at export locations are not only volatile but highly seasonal, and the characteristics of this seasonality are not the same across marketing years. These characteristics have important implications for market participants in making trading and risk management decisions. This study examined the influence of market supply and demand and logistical variables on both the average level and seasonality of U.S. export basis values for soybeans. We developed and estimated models using a statistical regression technique called partial least squares regression (PLS-R) to estimate the impact of market and logistical variables on the marketing-year (MY) average level of basis. To explain seasonality, this study used agglomerative hierarchal clustering (AHC) to cluster similar marketing years by common seasonal patterns, which are called seasonal analogs. Additionally, we used a statistical variable characterization test (Lebart, Morineau, and Piron, 2000) that compared the means of a subset and its parent set to explain the explanatory variables' influence to predict the derived seasonal analogs.

There are four important results. First, the average level for basis in the two U.S. export markets (Gulf and PNW) is primarily driven by international and domestic competitive pressures. The U.S. export basis values adjust with changes in the competitor's basis values. The level of imports by China, the dominant importer, is also a very important variable.

Second, seasonality in Gulf and PNW export basis values is not consistent across marketing years. As a result, these varying market conditions create different seasonal basis patterns, called seasonal analogs. From the 12 MY periods analyzed, we derived 5 and 4 unique seasonal-analog patterns using statistical clustering applications for the Gulf and the PNW, respectively. For each market, one pattern was a unique year (outlier) characterized by extremely high basis volatility. The high number of analogs (5 for the Gulf and 4 for the PNW) for the short time frame indicates that the seasonal characterization of basis is highly unstable from year to year.

Third, unlike the average MY basis level, the seasonal-analog patterns are primarily driven by three categories of variables, which are mostly unique to each export market: (i) the level of export activity at a particular port, (ii) the pace of farmers' marketing throughout the marketing year, and (iii) the logistical conditions (lateness of railcar placement, cost of secondary railcars, and barge and ocean freight rates) present during the marketing year along with transportation-cost differentials (between the two ports, and primarily barge and rail). In particular, the extreme-outlier analogs are both characterized by strong nearby demand and complicated logistical conditions.

Finally, the results are reflective of market conditions embedded in the study period. Since then, the Chinese tariff war that began in 2018 has had a drastic impact on the international soybean market. Regarding basis, the observed impacts of this intervention have been a drastic increase in the Brazilian free-on-board basis; a decrease in shipments from Brazil to non-Chinese markets, particularly from November 2018 forward; a decrease in U.S. basis values to unprecedented low levels; and an increase in U.S. basis volatility to higher-than-normal levels. These changes were concurrent in a sharp reduction in U.S. exports to China and reductions in secondary market rail values. Our model includes each of these changes, and the results are as expected, specifically a sharp reduction in U.S. basis values and radical changes in storage and soybean flows, both internationally and within the United States.

This study makes several contributions. First, it explicitly analyzes basis in export markets, in contrast to other studies whose focus is on origin or futures delivery markets. As a result, international demand, competition, and logistics are very important. Second, the analytical tools we use have not been used previously in the agricultural marketing literature, to our knowledge, though they are commonly used in other sectors. Finally, while it is customary to speak of seasonal analogs in industrial commodity market analysis, this is less common in the academic literature. We develop a model to statistically categorize seasonal analogs and determine the factors impacting these characteristics.

[First submitted May 2019; accepted for publication November 2019.]

References

Abdi, H. "Partial Least Square Regression." In N. J. Salkind, ed., *Encyclopedia of Measurement and Statistics,* Thousand Oaks, CA: Sage, 2007, 741–744. doi: 10.4135/9781412952644.

Addinsoft, Inc. *XLSTAT Statistical and Data Analysis Solution.* New York, NY: Addinsoft, Inc., 2019.

Alexander, R., Z. Zhao, E. Székely, and D. Giannakis. "Kernel Analog Forecasting of Tropical Intraseasonal Oscillations." *Journal of the Atmospheric Sciences* 74(2017):1321–1342. doi: 10.1175/JAS-D-16-0147.1.

Amemiya, T. *Advanced Econometrics.* Cambridge, MA: Harvard University Press, 1985.

Comeau, D., D. Giannakis, Z. Zhao, and A. J. Majda. "Predicting Regional and Pan-Arctic Sea Ice Anomalies with Kernel Analog Forecasting." *Climate Dynamics* 52(2019):5507–5525. doi: 10.1007/s00382-018-4459-x.

Djalalova, I., L. Delle Monache, and J. Wilczak. "PM2.5 Analog Forecast and Kalman Filter Post-Processing for the Community Multiscale Air Quality (CMAQ) Model." *Atmospheric Environment* 119(2015):431–442. doi: 10.1016/j.atmosenv.2015.05.057.

Elmore, R., and S. E. Taylor. "Analog Years for Weather Forecasting and Correlating Corn Planting Dates with Yield in Iowa." *Integrated Crop Management News* 82(2013).

Esbensen, K. H., and B. Swarbrick. *Multivariate Data Analysis: An Introduction to Multivariate Analysis, Process Analytical Technology and Quality by Design.* Oslo, Norway: CAMO Software AS, 2018, 6th ed.

Flaskerud, G., and D. Johnson. "Seasonal Price Patterns for Crops." Extension Bulletin EB-61, North Dakota State University, Fargo, ND, 2000.

Hair Jr, J. F., M. Sarstedt, L. Hopkins, and V. G. Kuppelwieser. "Partial Least Squares Structural Equation Modeling (PLS-SEM): An Emerging Tool in Business Research." *European Business Review* 26(2014):106–121. doi: 10.1108/EBR-10-2013-0128.

Hansen, J. W., A. Potgieter, and M. K. Tippett. “Using a General Circulation Model to Forecast Regional Wheat Yields in Northeast Australia.” *Agricultural and Forest Meteorology* 127(2004):77–92. doi: 10.1016/j.agrformet.2004.07.005.

Hatchett, R. B., B. W. Brorsen, and K. B. Anderson. “Optimal Length of Moving Average to Forecast Futures Basis.” *Journal of Agricultural and Resource Economics* 35(2010):18–33. doi: 10.22004/ag.econ.53048.

Hertsgaard, D. J., W. W. Wilson, and B. Dahl. “Costs and Risks of Testing and Blending for Essential Amino Acids in Soybeans.” *Agribusiness* 35(2019):265–280. doi: 10.1002/agr.21576.

Hieronymus, T. A. *Economics of Futures Trading for Commercial and Personal Profit*. New York, NY: Commodity Research Bureau, 1977.

Hull, J. *Fundamentals of Futures and Options Markets*. Boston, MA: Pearson, 2017, 9th ed.

Irwin, S., and D. Good. “Forming Expectations for the 2016 U.S. Average Soybean Yield: What about El Niño?” *farmdoc daily* 6(2016):46.

Jiang, B., and M. Hayenga. “Corn and Soybean Basis Behavior and Forecasting: Fundamental and Alternative Approaches.” 1997. Paper presented at the NCR-134 Conference on Applied Commodity Analysis, Forecasting, and Market Risk Management, April 21–22, Chicago, Illinois. doi: 10.22004/ag.econ.285704.

Johansson, R., E. Luebehusen, B. Morris, H. Shannon, and S. Meyer. “Monitoring the Impacts of Weather and Climate Extremes on Global Agricultural Production.” *Weather and Climate Extremes* 10(2015):65–71. doi: 10.1016/j.wace.2015.11.003.

Kolb, R. W., and J. A. Overdahl. *Futures, Options, and Swaps*. New York, NY: Wiley-Blackwell, 2007, 5th ed.

Kolton, A. D., R. A. Gamboa, and D. S. Chimenti. “System for Forming Queries to a Commodities Trading Database Using Analog Indicators.” 1996. Available online at https://patents.google.com/patent/US5590325A/en. [Accessed May 15, 2019].

Kuhn, M., and K. Johnson. *Applied Predictive Modeling*. New York, NY: Springer-Verlag, 2013. doi: 10.1007/978-1-4614-6849-3.

Lahmiri, S., G. S. Uddin, and S. Bekiros. “Clustering of Short and Long-Term Co-Movements in International Financial and Commodity Markets in Wavelet Domain.” *Physica A: Statistical Mechanics and its Applications* 486(2017):947–955. doi: 10.1016/j.physa.2017.06.012.

Lakkakula, P., and W. Wilson. “Origin and Export Basis Interdependencies in Soybeans: A Panel Data Analysis.” *Journal of Agricultural and Resource Economics* forthcoming(2020). doi: 10.22004/AG.ECON.302464.

Lebart, L., A. Morineau, and M. Piron. *Statistique Exploratoire Multidimensionnelle*. Paris, France: Dunod, 2000.

Lee, Y., and B. W. Brorsen. “Permanent Shocks and Forecasting with Moving Averages.” *Applied Economics* 49(2017):1213–1225. doi: 10.1080/00036846.2016.1213368.

Martens, H., and T. Næs. *Multivariate Calibration*. London, UK: Wiley, 1989.

Massy, W. F. “Principal Components Regression in Exploratory Statistical Research.” *Journal of the American Statistical Association* 60(1965):234–256. doi: 10.1080/01621459.1965.10480787.

Menzie, K. “Methods of Evaluating Agrometeorological Risks and Uncertainties for Estimating Global Agricultural Supply and Demand.” In M. V. K. Sivakumar and R. P. Motha, eds., *Managing Weather and Climate Risks in Agriculture,* New York, NY: Springer, 2007, 125–140. doi: 10.1007/978-3-540-72746-0_9.

Phillips, G. M., W. P. Jennings, M. C. F. III, S. A. Klein, and M. E. Rice. “Combination Forecasting Using Clusterization.” 2004. Available online at https://patents.google.com/patent/US6792399B1/en. [Accessed May 15, 2019].

Shannon, C. E. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27(1948):379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Taylor, M. R., K. C. Dhuyvetter, and T. L. Kastens. "Forecasting Crop Basis Using Historical Averages Supplemented with Current Market Information." *Journal of Agricultural and Resource Economics* 31(2006):549–567. doi: 10.22004/ag.econ.8625.

Tenenhaus, A., V. Guillemot, X. Gidrol, and V. Frouin. "Gene Association Networks from Microarray Data Using a Regularized Estimation of Partial Correlation Based on PLS Regression." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7(2010):251–262. doi: 10.1109/TCBB.2008.87.

Tilley, D. S., and S. K. Campbell. "Performance of the Weekly Gulf-Kansas City Hard-Red Winter Wheat Basis." *American Journal of Agricultural Economics* 70(1988):929–935. doi: 10.2307/1241935.

Wanat, S., S. Śmiech, and M. Papież. "In Search of Hedges and Safe Havens in Global Financial Markets." *Statistics in Transition New Series* 17(2016):557–574. doi: 10.21307/stattrans-2016-038.

Ward, J. H. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58(1963):236–244. doi: 10.1080/01621459.1963.10500845.

Wilson, W. W., and B. Dahl. "Grain Pricing and Transportation: Dynamics and Changes in Markets." *Agribusiness* 27(2011):420–434. doi: 10.1002/agr.20277.

Wold, H. "Estimation of Principal Components and Related Models by Iterative Least Squares." In P. R. Krishnaiah, ed., *Multivariate Analysis,* New York, NY: Academic Press, 1966, 391–420.

———. "Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments." In P. R. Krishnaiah, ed., *Multivariate Analysis III: Proceedings of the Third International Symposium on Multivariate Analysis Held at Wright State University, Dayton, Ohio, June 19–24, 1972,* New York, NY: Academic Press, 1973, 383–407. doi: 10.1016/B978-0-12-426653-7.50032-6.

Wold, S. "Personal Memories of the Early PLS Development." *Chemometrics and Intelligent Laboratory Systems* 58(2001):83–84. doi: 10.1016/S0169-7439(01)00152-6.

Wold, S., M. Sjostrom, and L. Eriksson, eds. *PLS: Partial Least Squares Projections to Latent Structures*. Leiden, Netherlands: ESCOM, 1993.

Working, H. "The Theory of Price of Storage." *American Economic Review* 39(1949):1254–1262.

Zhang, R., and J. E. Houston. "Effects of Price Volatility and Surging South American Soybean Production on Short-Run Soybean Basis Dynamics." Paper presented at the NCR-134/NCCC-134 Applied Commodity Price Analysis, Forecasting, and Market Risk Management conference, April 18–19, St. Louis, Missouri, 2005.