Machine Learning in Gravity Models: An Application to Agricultural Trade

Munisamy Gopinath*
Feras A. Batarseh
Jayson Beckman

Abstract: Predicting agricultural trade patterns is critical to decision making in the public and private domains, especially in the current context of trade disputes among major partners. Focusing on seven major agricultural commodities with a long history of trade, this study employed data-driven and deep-learning processes: supervised and unsupervised machine learning (ML) techniques – to decipher patterns of trade. The supervised (unsupervised) ML techniques were trained on data until 2010 (2014), and projections were made for 2011-2016 (2014-2020). Results show the high relevance of ML models to understanding trade patterns relative to traditional approaches, which are often subjective assessments or time-series projections. Moreover, supervised ML techniques quantified key economic factors determining agricultural trade flows over the past six decades.

Keywords: trade flows, gravity model, machine learning, supervised and unsupervised techniques, agriculture

JEL Classification: F14, C45, Q17

Gopinath is Professor in the Department of Agricultural and Applied Economics, University of Georgia; Batarseh is Assistant Professor in the Graduate School of Arts and Sciences (Data Analytics) at Georgetown University; and Beckman is Senior Economist at the Economic Research Service of the U.S. Department of Agriculture.

*Corresponding author: m.gopinath@uga.edu

Machine Learning in Gravity Models: An Application to Agricultural Trade

Introduction

Over the past decade, the availability of big data and advances in computer technologies have challenged conventional statistical and econometric techniques in modeling complex economic relationships (Varian, 2014). The challenges include dealing with the sheer volume of data (going from spreadsheets to SQL to Hadoop clusters and distributed datasets), the lengthy list of variables available to explain such relationships (and associated collinearity issues), and the need to move beyond simple linear models. Machine learning (ML) has been offered as an alternative to address many of these challenges (Bajari et al., 2015; Mullainathan and Spiess, 2017; Athey and Imbens, 2019). Several authors including Chief Economists of Google and Amazon have strongly advocated the use of big data and ML to uncover increasingly complex relationships even in an analysis as simple as fitting a supply or demand function. The economics community is catching on, but the speed of ML advances, i.e. new techniques emerge every month, can make an academic study stale by the time peer reviews are completed. Nonetheless, the academic community facing seismic shocks from advances in data and ML has been called on to revisit time-tested theories and relationships (Mullainathan and Spiess, 2017). This study takes on this challenge in the context of international trade and offers a ML application using agricultural trade data spanning several decades.

Many international institutions and government agencies project economic variables including trade flows to inform decisions in national and multilateral contexts (World Economic Outlook – International Monetary Fund, 2019; Trade in Goods and Services Forecast, Organization for Economic Cooperation and Development, 2019; World Trade Organization,

2019; U.S. Department of Agriculture, 2019). Since these projections are based on a combination of model-based analyses and expert judgment, several sources have pointed to their limitations, e.g. forecast accuracy under 35% (U.S. Department of Agriculture, 2019) and quantifying the contribution of underlying economic factors (Chapter 4, World Economic Outlook – International Monetary Fund, 2019). With ongoing trade disputes such as Brexit, U.S.-China/Japan-South Korea tariffs, the need for alternative approaches to understanding and predicting (modeling) trade flows is greater now than ever before. The initial and retaliatory tariffs between/among major trading economies, especially in the context of agricultural trade, have created a level of uncertainty and complexity unknown over the past several decades. Compounding the situation is the static nature of most trade models, which often conduct comparative static analysis of trade outcomes from deterministic trade policy changes. Little guidance exists on theoretical modeling of trade policy uncertainty and its implications for producer and consumer preferences or behavior.

In this study, ML techniques are applied to the gravity model of bilateral (aggregate or industry) trade flows. The gravity model is often referred to as the workhorse in international trade due to its popularity and success in quantifying the effects of various determinants of international trade. Originally due to Anderson (1979) and applied to data in Anderson and van Wincoop (2003), the gravity model provides the causal association needed to implement ML algorithms in the predictive domain (Athey and Imbens, 2019, Yotov et al., 2016, Santos Silva and Tenreyro, 2006). In doing so, this study offers an alternative to time-series projections and expert judgment analyses by relying on data-driven and deep learning approaches that allow for alternative and robust specifications of complex economic relationships (Baxter and Hersh,

2017). ML models can also provide accurate predictions, a priority of many economists in recent months given the trade disputes among the major economies of the world.

Machine Learning Methods

Machine learning, a set of algorithms for advanced statistical analysis and intelligent problem solving, offers a novel and flexible approach to model relationships, i.e. quantify *Y*'s response with or without a set *X* of possible predictors, either supervised or unsupervised, respectively (James et al., 2013). Supervised learning, which also includes regression and classification approaches, relate the response of *Y* to *X* for either better understanding of their relationship or predicting the response of *Y* to a potential/future set of *X*. In contrast, unsupervised learning usually does not have a pre-defined response variable (*Y*) and aims to understand the relationships between/among *X* or observations. In both settings, often relationships are derived from one or more "training" data sample and applied to a "test" data sample to compute prediction accuracy. The sheer volume of data available in recent times allows such a partition of data for cross-validation, i.e. training and testing. However, a major trade-off arises between prediction accuracy and model interpretability between conventional (econometric) analysis and machine learning. As applicable techniques become non-linear or multi-layered with the repeated feedbacks between training and test data, machine learning techniques often leave interpretability and inference behind to focus more on prediction accuracy. This study applies a variety of supervised and unsupervised ML techniques to the trade setting noted earlier and also presents the inner working and challenges of ML in these settings.

A comprehensive review of all ML techniques available at this time is beyond the scope of this study. Excellent sources for that information include James et al. (2013), Batarseh and

Yang (2017) and Athey and Imbens (2019). In the following, the generic optimization problem as in Athey and Imbens (2019) is presented along with an outline of techniques employed in this study. Consider a simple example where the outcome *Y* depends on a set of features *X*. Assume:

$$Y_i|X_i \sim N(\alpha + \beta^T X_i, \sigma^2) \qquad (1)$$

where $\theta = (\alpha, \beta)$ are parameters of interest, Y_i has a conditional normal distribution with variance σ^2. Conventional econometric estimation, i.e. least squares, would suggest:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} (Y_i - \alpha - \beta^T X_i)^2. \qquad (2)$$

In the ML setting, the goal is usually to make a prediction for the outcome from a new set of values for *X* i.e. predicting Y_{N+1} from X_{N+1}. Let that prediction, regardless of the actual specification of the relationship between *Y* and *X* be $\hat{Y}_{N+1}$. Then, the squared loss associated with this prediction would be:

$$(Y_{N+1} - \hat{Y}_{N+1})^2. \qquad (3)$$

While least squares is an approach to minimize the loss function, other estimators exist that dominate least squares (Athey and Imbens, 2019). However, ML-based estimators for equation (3) have a tendency to over or under-fit, which can be corrected by regularization and tuning parameters through out-of-sample cross-validation.[1] The following provides a brief overview of ML techniques considered in this study:

- LASSO, Ridge regression and elastic nets are some basic extensions to the least squares minimization problem in equation (2) to impose a penalty for increasing the dimensionality of X (i.e. regularization). A major concern with these approaches is the subjective choice on the penalty, often referred to as λ, but big data allows for its potential out-of-sample cross validation.

- Decision trees and their extensions have become extremely popular in recent years. They are referred to as trees since data are stratified or segmented into branches (splits) and leaves (nodes). The stratification is based on the number of predictors and cut-off values for predictors. For instance, if *X* contains two column vectors, then stratification will be based on both constituents, sequentially or in random, for all possible cut-off values for each of these 2 predictors, e.g. $X_1 < c_1$. To make a prediction for new values of *X*, trees typically use the mode or median of the outcome *Y* in the region to which the new *X* belongs. To illustrate, as in Athey and Imbens (2019), the total-sample sum of squared errors for outcome *Y* is given by:

$$Q = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \qquad \bar{Y} = \sum_{i=1}^{N} Y_i. \tag{4}$$

After a split based on one of the predictors (X_k) using the threshold $X_k < c$, the sum of total-sample squared errors is:

$$Q(k, c) = \sum_{i=X_{ik} \le c} (Y_i - \bar{Y}_{k,c,l})^2 + \sum_{i=X_{ik} > c} (Y_i - \bar{Y}_{k,c,r})^2, \tag{5}$$

where *l* and *r* denote left and right of X_k using the cut-off *c* and

$$\bar{Y}_{k,c,l} = \frac{\sum_{i=X_{ik} \le c} Y_i}{\sum_{i=X_{ik} \le c} 1}, \qquad \bar{Y}_{1,c,r} = \frac{\sum_{i=X_{ik} > c} Y_i}{\sum_{i=X_{ik} > c} 1}.$$

The objective of decision-tree based learning is to minimize $Q(k, c)$ for every k and every $c \in (-\infty, +\infty)$, and the process is repeated for subsamples or leaves. As noted earlier, there is a tendency in this approach to over-fit, which can be corrected by adding a penalty for the number of leaves or by pruning the tree using cross-validation. A single tree is often the preferred outcome from this approach for its interpretability. However,

prediction accuracy has been significantly improved (at the expense of interpretation) by procedures such as bagging, random forests and boosting:

- Bagging involves repeated sampling of the (single) training data to fit a tree each. Then, averaging across trees chosen independently (like in bootstrapping) improves its prediction by lowering the variance.
- The random forests approach is similar to bagging in the sense that bootstrapped training samples are used to generate decision trees to average out. However, at each split of the decision tree the choice on predictors (or a subset) is random.
- Boosting is also similar to bagging, but the decision tree for each training sample is not independent of previous trees, instead they are chosen sequentially. Each tree is grown using information from previously grown trees and thus, boosting does not involve bootstrap sampling. Unlike bagging and random forests which are applicable to decision trees only, boosting can be employed for any base learner.

In the ML literature, the boosting algorithms are popular as they convert a weak base learner (often a single decision tree) into a strong learner (by averaging over training sub-samples as well as tuning hyper-parameters).

- Neural networks and deep learning methods are often employed in situations with large volumes of unexplored data to predict an outcome or analyze a pattern. These techniques emulate the human brain's neural system by building layers of functions between the outcome and predictors. These layers include predictors in linear form, but then translate them into latent variables and use non-linear functional forms to relate to the outcome.

To illustrate a single-layer network learning, consider the latent variables Z defined as follows:

$$Z_{ik}^{(1)} = \sum_{j=1}^{k} \beta_{kj}^{(1)} X_{ij}, \quad k = 1, \dots, K_1. \tag{6a}$$

Then, a non-linear function $g(.)$ relates the outcome Y to Z:

$$Y_i = \sum_{k=1}^{K_1} \beta_k^{(2)} g\left(Z_{ik}^{(1)}\right) + \varepsilon, \tag{6b}$$

The objective of the estimation again is minimizing squared errors, but the layering allows for millions of functional possibilities and parameters. Note that such unsupervised ML techniques often do not have both outcome Y to X. In the context of equations 6(a) and 6(b), the X can be thought of some transformation of Y, e.g. lagged values.

In this study, ML methods are considered for the standard specification of the gravity model:

$$Y_{ijt} = g(X_{it}, X_{jt}, i, j, t), \tag{7}$$

where Y_{ijt} is bilateral trade between country i and country j at time t, (response variable) and $X_{it(jt)}$ is the set of possible predictors from both countries and the set $\{i, j, t\}$ refer to a variety of controls on all three dimensions (Anderson and van Wincoop, 2003; Yotov et al., 2016). The major ML techniques applied include decision trees including bagging, random forests and boosting, and neural networks. While it is tempting to compare ML models with econometric approaches, e.g. the popular Poisson Pseudo-Maximum Likelihood (PPML) method commonly used to estimate trade flows, a word of caution is in order. ML techniques often encompass three paradigms – descriptive, predictive and prescriptive – and the focus of this study is in the predictive domain. Nonetheless, we do present results from applying PPML estimates of the gravity model in the Appendix.

The Agricultural Trade Setting and Data

References to agricultural trade abundantly appear in literature dating back to 1000 BCE. One of the earliest pathways connecting Mediterranean to Arabia, Indian sub-continent and Far East was the Incense Route. As the name suggests, incense made of aromatic plants and oils was a major traded commodity, but spices, silk and precious stones were also major transactions along this route. Then came Spice and Silk Routes and numerous other inter- and intra-continental routes facilitating trade in agricultural products and other goods. The industrial revolution of 18th century favored agricultural industries, but that in late 19th and early 20th century expanded rapidly into transport and energy sectors. Fast forward to the later parts of the 20th century, agriculture still accounted for at least a quarter of exports (or imports) of major trading nations. For instance, the 1960 edition of the State of Food and Agriculture from United Nations' Food and Agriculture Organization noted the significant share of agriculture in merchandise exports from the new world (North America and Oceania) to the rest of the world.

Anderson (2016), reviewing the evolution of food trade patterns over the past six decades, notes that agriculture's share of global merchandise trade was about 27 percent in 1960. While that share has fallen to 11 percent in 2014 (due in part to lower prices of agricultural goods relative to industrial goods) the volume of agricultural exports has continued the strong upward trend of the early 20th century. More importantly, Anderson (2016) notes the concentration in both country and commodity shares of global exports of farm products. In particular, less than 10 items made up half of that trade in agricultural products and two-thirds of the world's exports of farm products are accounted by just a dozen trading economies. Despite that concentration, the set of commodities that stand out for having most countries involved in world trade are shown in Table 1. The seven commodities – wheat, maize, rice, sugar, beef, milk

powder and soybean – have not only been traded for long but also most countries engaged on the export or import side. Hence, this study chose to apply machine learning techniques to understand the patterns of agricultural trade where the longest time series and most country pairings exist, i.e. the seven commodities in table 1.

In terms of sources of data for the gravity model application, this study employs bilateral trade (import data) from UN Comtrade for the seven commodities noted above.[2] UN Comtrade provides data using several nomenclatures, we use the Standard International Trade Classification (SITC) Revision 1 classification as this classification features data with the longest possible time-frame, i.e. from 1962 for some commodities. Specific codes are: 0111 for beef, 0221 and 0222 for milk powder, 041 for wheat, 042 for rice, 044 for corn, 0611 for sugar, and 2214 for soy. Tariff data are obtained from the UNCTAD Trade Analysis Information System (TRAINS) database. For our purposes, we use the simple average for each bilateral trade pair across each of the seven commodities. To account for missing tariff data (countries often only report a single year across a decade or so), the approach of Jayasinghe et al. (2010) is employed to derive implied average tariffs for missing observations. Note that tariff data are only available from 1988.

Data for (35) gravity variables such as GDP, population, contiguity, distance, common language, WTO membership, preferential trade agreements, and colonial ties are from the dynamic gravity dataset constructed by Gurevich and Herman (2018). Their work built a new gravity dataset that improves upon existing resources (e.g. CEPII data) in several ways: first, it was constructed to reflect the dynamic nature of the globe by closely following the ways in which countries and borders have changed between 1948 and 2016. Second, they increased the time and magnitude of variation within several types of variables. All three sources of data are

merged to arrive at a data set featuring imports, tariffs, and economic variables from 1962-2016.[3] The data are in cubical form, country pairs and time, and the number of observations range from 6050 (sugar) to 19930 (milk powder).

Machine Learning Model Selection and Cross-Validation

Recall that multiple ML approaches are considered to learn and predict bilateral trade for each of the seven agricultural commodities. Within decision trees, in addition to random forests and bagging (extra trees), three types of boosting – Gboost, LightGBM, XGBoost – are considered (Ke et al., 2017):

- Gboost creates an initial tree (base learner) with equal weights on all observations. Errors from incorrect predictions of the base learner are given greater weights as subsequent learning occurs.
- LightGBM scans all data instances to estimate the *Gain*, measured in terms of the reduction in the sum of squared errors, from all the possible *Split* points. Instead of changing the weights for every incorrectly predicted observation at every iteration like GBoost, LightGBM tries to fit the new predictor to the errors made by the previous predictor.
- LightGBM splits the tree level-wise with the best fit, whereas XGBoost algorithms split the tree leaf-wise. The leaf-wise algorithm can reduce more loss than the level-wise algorithm, and hence can lead to better accuracy.

Each of these models employed all 35 gravity variables noted earlier along with tariffs in the bilateral trade context. Gurevich and Herman (2018) provided 70 gravity variables, but we chose to employ the 35 variables based on a correlation analysis. The first step here was to avoid near

perfect collinearity among the 70 variables. Then, the data had alternative representations for the size of the economies, e.g. GDP total and per capita, and in current and constant dollars. For this second step, measures of feature importance for each of the seven models under alternative variables representing the same phenomenon, e.g. size, were obtained. The splits and gains then determined the variables to be included in the model. However, in this study, the high cardinality (correlation) of X and alternative representations of X led variable selection by way of information gains from splits (using out-of-the-box python libraries). After selecting variables, boosting algorithms was deployed. Main parameters tuned for all three boosting methods are: maximum depth of the tree, learning rate, number of leaves, and feature fraction. The choice among these supervised models was also dictated by the adjusted R-square, the most commonly used statistical measure, as shown below (Ke et al., 2017). Both supervised and unsupervised ML methods allow for cross-validation, where predictions can be compared to actuals.

Available data spanned 1962-2016, and the training data cut-off point was set at 2010.[4] This partitioning of data allowed for a longer time series to learn as well as have enough data to compare predictions to actuals (2011-2016). Major challenges for many of these algorithms was a large number of zeros in bilateral trade matrices as well as time invariant variables in the gravity context such as distance, language and contiguity. Thus, each commodity had a different data distribution, cardinality and time frame for fitting the model and obtaining predictions for 2011-2016.

Table 2 presents the three best fitting models among the supervised ML approaches and the training and test sample sizes for each of these models.[5] Note that milk had the most observations for training, while wheat's test data was the largest among the seven commodities. As can be seen in table 2, LightGBM (XGBoost) had the best performance for corn and rice

(soy), while extra trees- bagging provided the best fit for the other commodities: beef, sugar, milk and wheat. However, the adjusted R-square in the best-fitting bagging models ranged between 27 and 38% only except for beef. Lower adjusted R-square values are due to high variance in the inputs used for the models and incomplete data from older years. However, under-fitting was not found for any of the seven commodities and therefore, adjusted R^2 metric appears to represent the model's quality. Moreover, each of the commodity models were deployed on a global scale, which places all countries on the same level of abstraction when using the response variable.[6]

Neural networks, a form of deep learning, can take several paradigms: recurrent and convolutional neural networks, hybrids and multilayer perceptron (MLP). This study employs MLP given the tabular data context (along with its cardinality) and the goal of obtaining better predictions. MLP is a feed forward algorithm with the following steps:

- #1, Forward pass: here, the values of the chosen variable are multiplied with weights and bias is added at every layer to find the calculated output of the model.
- #2: Calculate the error or the loss: the data instance is passed, the output is called the *predicted output*; and that is compared with real data called the *expected output*. Based upon these two outputs, the loss is calculated (using Back-propagation algorithm).
- #3: Backward pass: the weights of the perceptrons are updated according to the loss.

In Python, the MLP algorithm uses Stochastic Gradient Descent (SGD) for the loss function (Scikit, 2019). SGD is an iterative method for optimizing an objective function with smoothness properties.[7] As noted earlier, neural networks do not have validation statistics, unlike supervised models.

In addition to employing ML techniques, this study considered traditional approaches to estimating the gravity equation (7). As noted by Disdier and Head (2008), many econometric techniques have been used in estimating equation (7), but the more recent approach that accounts for zeros in Y_{ijt}, heteroscedasticity in additive errors to equation (7) and other issues is Poisson Pseudo-Maximum Likelihood estimation (Santos Silva and Tenreyro, 2006). The results from estimating equation (7) using PPML and the chosen 35 gravity variables for the seven commodities of this study are reported in the Appendix. With a focus on predictions and economic significance of gravity variables in bilateral trade flows, the following section presents results from ML methods only.[8]

Results and Discussion

Figures 1-3 and tables 3-4 present the results from the supervised learning models, while figure 4 details those from the neural networks application to capturing the gravity trade relationship in equation (7).

Supervised ML Model Results

Recall that the cut-off year for the training data was 2010 and so, projections from the supervised models are made through 2016. The seven panels of figure 1 show the 2010-16 aggregate (total) trade values, actual and predicted, for each of the chosen commodities. For consistency, we present the predicted values from LightGBM in figures 1-3.[9] Multiple factors contribute to the fit (adjusted R-square in table 2), but chief among them are the pruning of the decision trees, data incompleteness across commodities (training versus test data size), cardinality of the predictors and boosting to improve the successive fit of decision trees.

For beef, corn, rice and wheat, the predictions from LightGBM closely track that of actual trade values during 2010-14. Note from table 2 that beef and corn have some of the highest adjusted R squares. However, predictions of aggregate trade values for soybeans, sugar and milk powder significantly deviate from actuals during 2010-14. Again, table 2 shows that the fit of sugar and milk models, i.e. adjusted R-square, is significantly lower than beef, corn and rice. Soybeans and sugar models also have the fewest observations available for training, partly due to the delayed expansion of countries trading soybeans as shown in table 1. Much of sugar trade is also highly concentrated among a few countries (Anderson, 2016). The case of wheat is unique since it was the first of the seven commodities tested with extensive tuning of the hyper parameter for prediction quality (lessons learned were employed in modeling other commodities). Despite the lowest fit among the seven commodities, predictions of aggregate wheat trade values for 2010-14 track well the actuals including two point projections (2012 and 2014) that nearly match actuals.[10]

Note, however, all models' predictive ability considerably deteriorates for 2015-2016. A closer examination of actuals and predicted values for each bilateral pair indicates that the zero values, prevalent in the gravity models, are at the core of the falling predictive abilities of supervised ML in 2015-16 (figure 1). For instance, the zero values of trade data are often initially augmented by boosting to be a small positive or negative number, which upon iteration expands further to widen the gap between actuals and predictions. Thus, the above results suggest that boosting techniques might be better at near- to medium-term projections relative to those over a longer horizon.

Figures 2 and 3 present predictions for the two top bilateral pairs, in terms of trade value, for each of the seven commodities. As with the aggregate predictions in figure 1, the best

fitting models – beef, corn and rice – also produce bilateral predictions that closely track actuals for 2010-14. Predictions for other commodities' top trading pairs do no track well the actuals as noted earlier due to lower fit, data incompleteness and cardinality of predictors. Each of the models are deployed on a global scale, but as observed in the data and noted by Anderson (2016), agricultural commodity trade is concentrated in few countries in the early years of the sample. Grouping or categorizing countries and hyper-tuning variables may yield better model fits, but at the risk of losing data and employing arbitrary cut-offs. Again, for all commodities, the deterioration of predictive ability during 2015-16 is attributable to zero trade values and the associated boosting.

We compare supervised with unsupervised (neural networks) models later, but an advantage of the former is its ability to identify from among the set of predictors the features that have the greatest importance. Decision tree methods, particularly in the context of boosting, randomize among predictors for splitting the tree into nodes (or leafs) and the repetitive process searches for predictors and cut-off values that offer the greatest decline in the total-sample sum of squared errors (equation 5). In doing so, these methods identify the information gain, i.e. the reduction in total-sample sum of squared residuals, from each of the predictors. Recall that the gravity model, equation 7, in this study employed 35 variables as predictors. Tables 3 and 4 show both the ranking of predictors by information gain and their relative importance, respectively.

Economic Significance of Gravity Variables in Supervised ML Models

Table 3 presents the top 9 variables which provided the largest information gains from the 35 predictors included in these models. Alternatively termed as feature importance, they point to which of the variables are most indicative of the response variable (bilateral trade) in the model.

Removing any of these top predictors would drastically change the model results. Across commodities, the top 9 predictors remained the same, but their ranking and relative importance were different. As a review of gravity models predict, the size of the two economies engaged in trade has the largest influence in reducing the total-sample sum of squared errors. In particular, the destination country's population and GDP per capita are the top 2 information providers in the learning of agricultural trade flows, except in the case of wheat. This result is largely consistent with commodity trade, which is significantly determined by the importing country's size. The origin country's size also matters for all the reasons noted in gravity models, i.e. large countries tend to trade more with other large countries (Yotov et al., 2016; Chapter 4, World Economic Outlook – International Monetary Fund, 2019). Distance and time are the next set of predictors offering substantial information gains in the learning and prediction of trade flows. Noteworthy here is that the ranking of distance had the highest deviation ranging from 2 in wheat to 6 in corn, rice and sugar, a theme revisited in the context of table 3 below. In gravity models, latitude and longitude are often employed to represent the remoteness of a country (spatially and temporally).[11] The supervised models indicate the high relevance of both latitude and longitude to predict commodity trade flows. Political stability, mostly that of the destination country, was ranked the 9th highest information provider to these learning models.

The rankings in table 3 confirm the significance of commonly used gravity predictors but do not completely capture the relative importance of these variables. Table 4 normalizes the information gain of each variable with that of the top predictor (ranked #1 in table 3) for each commodity. Similar to table 3, most information gains in the supervised learning models arise from the size of the two economies. The variation in the gain associated with the distance and time predictors across commodities likely capture not only policy-induced differences but also

influences of extreme events associated with specific time periods (e.g. droughts or floods affecting particular growing regions). Interestingly, distance provides the second (third) highest information gains in the case of wheat (milk powder), where the top 5 exporters are developed countries. Given the North-South nature of trade in wheat and milk powder, distance likely captures significant barriers to such trade. Moreover, such information gains vary by the size of the training sample, which indicates that the impact of distance or any other feature can vary over sub-samples, offering a potential solution to some puzzles, e.g. the distance puzzle, commonly observed in econometric estimation of gravity models (Yotov et al., 2016). Together, the remaining 26 variables accounted for 84 to 126 percent of information gains relative to the top predictor for each commodity. Factors such as tariffs, common language or border, FTA or WTO membership and others matter collectively, but each of their effects is not large relative to the economies' size.

Unsupervised ML Model Results

Turning now to unsupervised models (neural networks), figure 4 presents predictions for the top exporter for each of the seven commodities. Recall from the discussion of equation 6(a) and 6(b) that these models do not necessarily have a response variable. In that sense, learning here happens primarily with the bilateral trade data. The training sample cut-off is set at 2014 and projections are made until 2020. Figure 4 shows that all commodities' predictions, with the exception of Netherlands' milk powder exports, closely track respective actuals for 2014-2017. In fact, there is at least one projection that almost mimics actual in all projections except in the case of milk powder.

It is not straightforward to compare supervised and unsupervised models, but each has its advantages and disadvantages. Likewise, comparisons to traditional/econometric approaches are

tenuous given causality issues noted earlier. Nonetheless, an attempt is made here to explore the merits of each approach and relevance in specific contexts. The supervised machine learning models, primarily of the decision-tree kind with bagging, random forests or boosting, have a structure relating a response variable to set of predictors. The estimated structure is often heterogeneous, non-linear and based on repeated learning, i.e. minimizing the total-sample sum of squared errors. The supervised techniques are straight forward to implement particularly in uncovering complex relationships and can be compared among themselves in terms of validation statistics such as error sum of squares. They also provide information on the most relevant predictors including the relative strength of alternative predictors. However, this technique cannot provide standard errors on predictors' contribution or a coefficient capturing the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, due in part to the non-linear and repeated learning process noted above. These models are a great fit for problems primarily focused on near- to medium-term predictions, e.g. prices, trade flows, especially when a large volume of data are available. Unsupervised models such as neural networks carry similar advantages in prediction problems, but do not have predictors or validation statistics. As demonstrated using bilateral trade data predictions, they appear most suited to longer-term projections, e.g. climate change, but supervised techniques with rolling training samples can also generate a sequence of medium-term projections that can be integrated for longer-term projections.[12] The pruning and regularization involved in supervised methods may limit the amount of data used in learning (as shown in table 2 on training and test data), whereas unsupervised approaches attempt to use all data. In contrast, econometric approaches offer structure and inference with carefully chosen causal relationships, often linear and homogeneous. The later are seldom cross-validated and

also suffers also from specification or variable selection bias like (unlike) supervised (unsupervised) models.

Conclusions

This study introduced ML models to the international trade setting and posed questions on their applicability and prediction quality. The basic specification of the popular gravity model of trade flows was subjected to data driven and deep learning processes in the form of supervised and unsupervised ML techniques using data from 1962-2016. A loss function that summed the squared error between actual and predicted bilateral trade flows for all available country pairs was minimized with supervised ML models including decision tress such as random forests, bagging and alternative types of boosting. Supervised ML models employed a set of predictors, commonly used gravity variables such as size of economies, the distance between them, and associated frictions. The validation statistics along with the data properties (distribution, cardinality and completeness) helped explain predictions and the relative importance of gravity variables from supervised ML models. Unsupervised models, also known as neural networks, are also employed in this study to uncover relationships of a single variable (trade) without the need for predictors (gravity variables). Both models are cross-validated, i.e. in supervised models the training data was set to 1962-2010 and the testing data to 2011-16, while unsupervised model's training and test data were set to 1962-2014 and 2015-2020, respectively.

Results from supervised ML show that the models fit well in the near- to medium-term (2011-2014), i.e. predictions closely track actuals, when the models have a high adjusted R-square and trade data encompass a large number of countries and years. However, the large presence of zero trade values and the use of boosting to transform weak supervised learners into

strong ones cause predictive quality to fall 3-4 years from the training data cut-off. A major advantage of the supervised ML model is the ability to identify which variables among the set of predictors provides more information to understanding bilateral trade flows. A ranking of top 9 variables by information gain, i.e. reduction in the total-sample sum of squared errors of the loss function attributable to a predictor, and their relative importance show that economies size, distance between them, location of countries and time are more important than other gravity variables in explaining trade flows. While these results are consistent with the trade and gravity model literature, ML's strengths are in variable selection, prediction and economic significance. In addition, the supervised ML models have opened up an opportunity to address time- and space-varying effects, e.g. the distance puzzle. Varying the training sample size likely yields different contributions by features, but other model criteria need to be carefully considered to fully unlock such puzzles. Unsupervised models appear to be better suited for long-term forecasting with predictions closely tracking actuals across commodities. However, they are known to be black boxes and are often difficult to validate.

Predicting agricultural trade patterns is critical to decision making in the public and private domains, especially in the current context of trade wars with tit-for-tat tariffs. For instance, farmers likely consider the potential demand from alternative foreign sources before deciding to plant crops, especially in large exporters. Similarly, countries setting budgets for farm programs need better predictions of prices and trade flows for assessing domestic production and consumption needs and instruments employed to achieve those outcomes. This study demonstrates the high relevance of ML models to predicting trade patterns with a greater accuracy than traditional approaches for a range of time periods. Existing forecasts of trade such as those by WTO, OECD and USDA are a combination of model-based analyses and expert

judgement and tend to have high variability. The ML models here offer a more accurate alternative by relying on data and deep learning to allow for alternative and robust specifications of complex economic relationships. Moreover, the ML models are cross-validated and easily allow for simulation of trade outcomes in alternative policy scenarios including their uncertainty in recent times. Future work focusing on data/matrix completeness, multi-variate response variables and alternative ML techniques would greatly aid in public and private decision making.

Footnotes

[1]For qualitative data, a similar contrast exists between conventional logistic regression and approaches such as discriminant analysis and K-Nearest Neighbors. This study does not discuss qualitative data techniques in the following since the application employs quantitative data.

[2]As pointed out in Harrison (2013) almost all countries use the standard free on board (FOB) basis (prices that include all charges up to placing the goods on board a ship at the port of departure), for exports; but some use FOB for imports while others use CIF ('Cost, Insurance and Freight', and includes the costs of transportation values for imports). Further, import data is not as reported as often as export data. https://www.imf.org/external/pubs/ft/bop/2012/12-30.pdf.

[3]Not all commodities had data from 1962-2016. Table 2 ahead shows the variation in number of observations across commodities. The economic and commodity data are merged into a SQL database. An R code is used to merge on *country-to-country* trade transactions, as well as year of economic variables. The data is merged using an **Inner Join**.

[4]Gurevich and Herman's (2018) gravity data are not available for 2017 and beyond.

[5]Although Random Forests method was employed, the results are not reported in Table 2 because it consistently underperformed relative to LightGBM.

[6]Zeros were used to fill missing bilateral trade data for all commodities. This can lead to the oversampling of *zeros* in these datasets. Oversampling increases the count of minority class instances to match it with the count of majority class instances, i.e. "upsizing" the minority class. Additional exploration of undersampling or correlation imputations for example could improve the models results, a topic for further research.

[7]For the MLP method, the MLPClassifier function is used along with the following parameters: early_stopping, epsilon, hidden_layer_sizes, learning_rate, learning_rate_init, max_iter, momentum, power_t, validation_fraction.

[8] As noted earlier, a direct comparison of gravity results can only be made with prescriptive ML models, which are under further development.

[9]LightGBM yielded the best overall results because of its recursive boosting performance, but the model was tuned correctly to stop at the right *validation level* and at a good *number of cycles* that proved successful for most commodities. The two mentioned parameters were selected automatically by the model (deployed using the standard boosting python function).

[10]Note that for other commodities more hyper parameters can be tuned further to achieve high predictive accuracy, but results are presented as such to highlight the advantages and disadvantages of alternative ML approaches.

[11]As Anderson (2014) notes, longitude matters for trade by capturing time and communication differences. In other instances, remoteness has often been used to capture multilateral resistance in gravity models (Yotov et al., 2016).

[12]For example, changing the training data cut-off to 2009 or 2011 would likely generate better projections for 2010-13 and 2012-15, respectively.

References

Anderson, Edward, "Time Differences, Communication and Trade: Longitude Matters" *Review of World Economics* 150:2 (2014), 337-369.

Anderson, James E. "A Theoretical Foundation for the Gravity Equation." *American Economic Review* 69:1 (1979), 106-16.

Anderson, James E., and Eric van Wincoop, "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review* 93:1 (2003), 170–92

Anderson, Kym, "Agricultural Trade, Policy Reforms, and Global Food Security" (pp 61-83), in Palgrave Studies in Agricultural Economics and Food Policy, (New York: Palgrave MacMillan 2016).

Athey, Susan and Guido W. Imbens, "*Machine Learning Methods Economists Should Know About*" Working Paper, (2019), Stanford University.

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang, "Machine Learning Methods for Demand Estimation." *American Economic Review* 105:5 (2015), 481–85.

Batarseh, Feras and Ruixin Yang. *Federal Data Science: Transforming Government and Agricultural Policy Using Artificial Intelligence*. (New York: Elsevier's Academic Press 2017).

Baxter, Marianne, and Jonathan Hersh. Robust Determinants of Bilateral Trade. Paper Presented at *Society for Economic Dynamics*, 2017.

Disdier, Anne-Celia, and Keith Head, "The Puzzling Persistence of the Distance Effect on Bilateral Trade." *Review of Economics and Statistics* 90:1 (2008), 37–48.

Feenstra, R.C. (2004). Advanced International Trade: Theory and Evidence. Princeton, NJ: Princeton University Press.

Gurevich, Tamara and Peter Herman, "The Dynamic Gravity Dataset: Technical Documentation" (Washington DC: U.S. International Trade Commission 2018).

Harrison, Anne, "FOB/CIF Issue in Merchandise Trade/Transport of Goods in BPM6 and the 2008 SNA" in Twenty-Fifth Meeting of the IMF Committee on Balance of Payments Statistics (Washington, D.C: International Monetary Fund 2013).

James, Gareth, Daniela Witten, Trevor Hastie and Rob Tibshirani, *An Introduction to Statistical Learning* (New York: Springer 2013).

Jayasinghe, Sampath, John C. Beghin, GianCarlo Moschini, "Determinants of World Demand for U.S. Corn Seeds: The Role of Trade Costs" *American Journal of Agricultural Economics*, 92:4 (2010), 999–1010.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye and Tie-Yan Liu (pp. 1-9), "LightGBM: A Highly Efficient Gradient Boosting" in (Long Beach, CA: 31st Conference on Neural Information Processing Systems, NIPS, 2017)

Liapis, Peter, "Structural Change in Commodity Markets: Have Agricultural Markets Become Thinner?" Food, Agriculture and Fisheries Paper No.54 (Paris: Organization for Economic Cooperation and Development 2012).

Mullainathan, Sendhil and Jann Spiess, "Machine Learning: An Applied Econometric Approach" *Journal of Economic Perspectives* 31:2 (2017), 87–106.

Organization for Economic Cooperation and Development, Trade in Goods and Services Forecast. doi: 10.1787/0529d9fc-en (Accessed on 30 September 2019)

Santos Silva, J.M. and Silvana Tenreyro, "The Log of Gravity" *Review of Economics and Statistics* 88:4 (2006), 641-658.

Scikit https://scikit-learn.org/stable/modules/sgd.html (Accessed on 30 September 2019)

U.S. Department of Agriculture, "Outlook for U.S. Agricultural Trade" (Washington DC: Economic Research Service, August 2019) https://www.ers.usda.gov/publications/pub-details/?pubid=94836 (Accessed on 30 September 2019)

Varian, Hal, "Big Data: New Tricks for Econometrics" *Journal of Economic Perspectives* 28:2 (2014), 3-28.

World Economic Outlook (Washington DC: International Monetary Fund 2019)

World Trade Organization https://www.wto.org/english/news_e/pres19_e/pr837_e.htm (Accessed on 30 September 2019)

Yotov, Yoto, Roberta Piermartini, José-Antonio Monteiro and Mario Larch, *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model* (Geneva: World Trade Organization 2016).

Figure 1: Supervised Model Predictions of Aggregate Trade Values, 2010-16

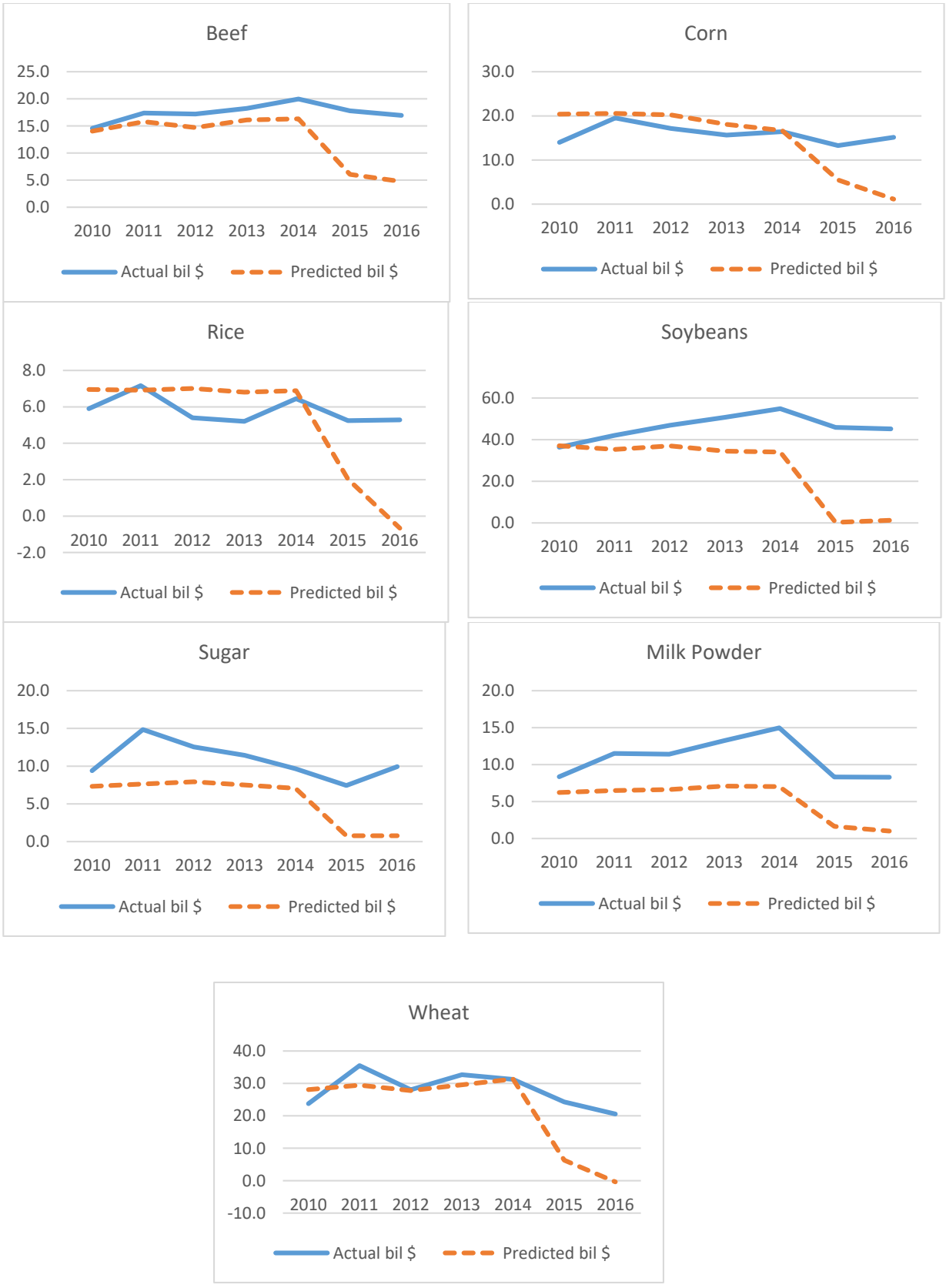


Figure 2: Supervised Model Predictions of Top Partners' Trade Values, 2010-16

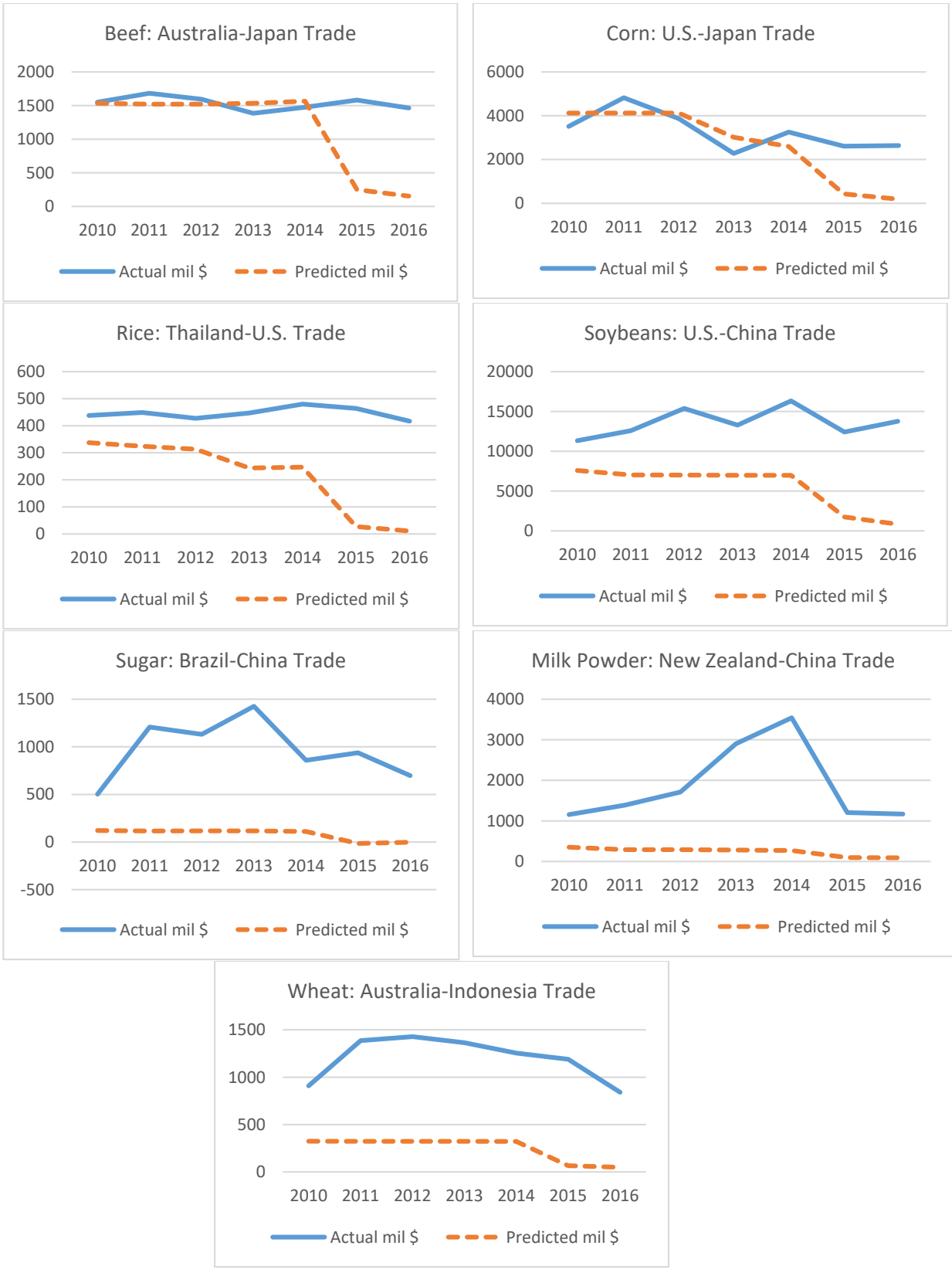


Figure 3: Supervised Model Predictions of 2nd Top Partners' Trade Values, 2010-16

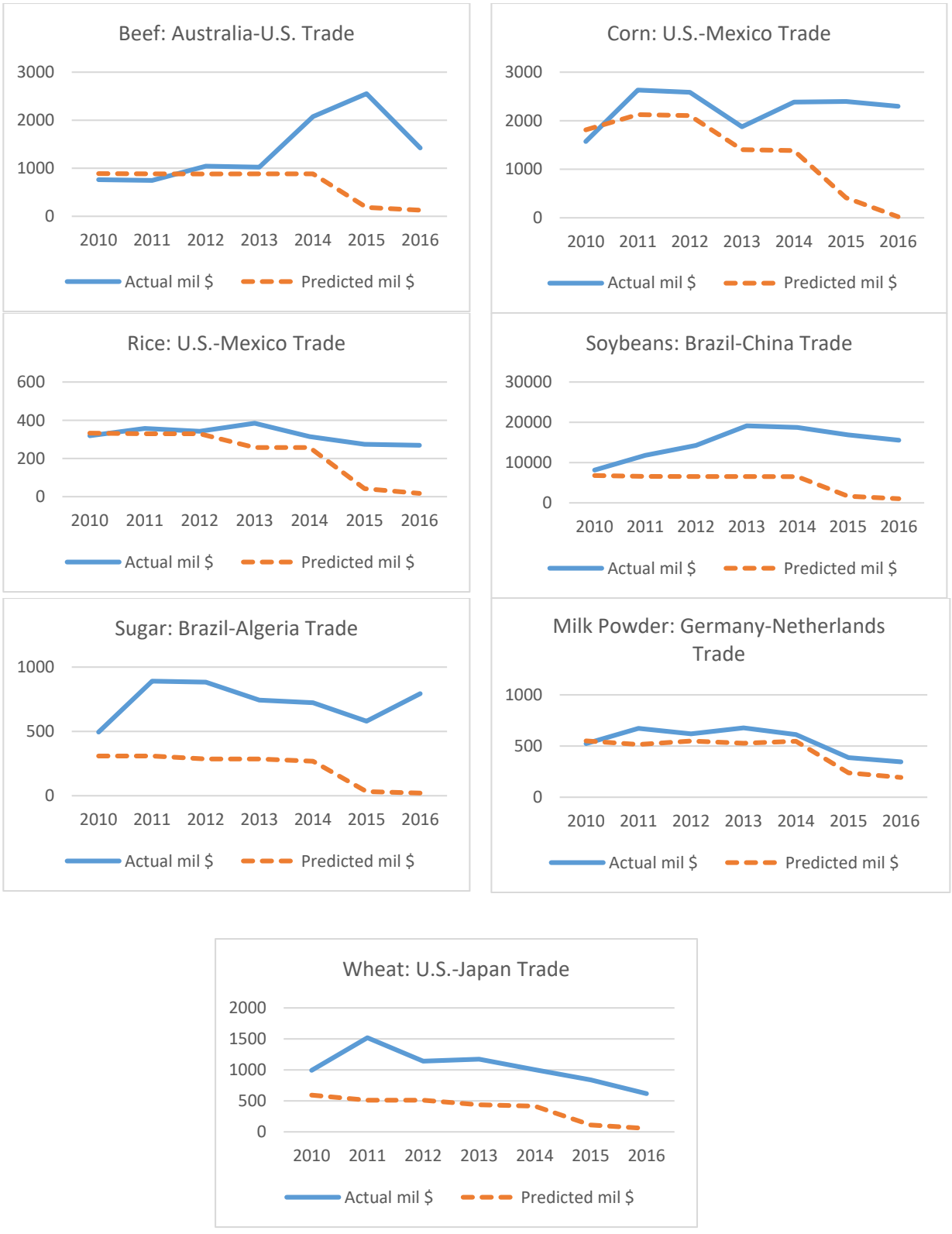


Figure 4: Neural Networks Prediction of Top Country's Aggregate Exports, 2014-2020

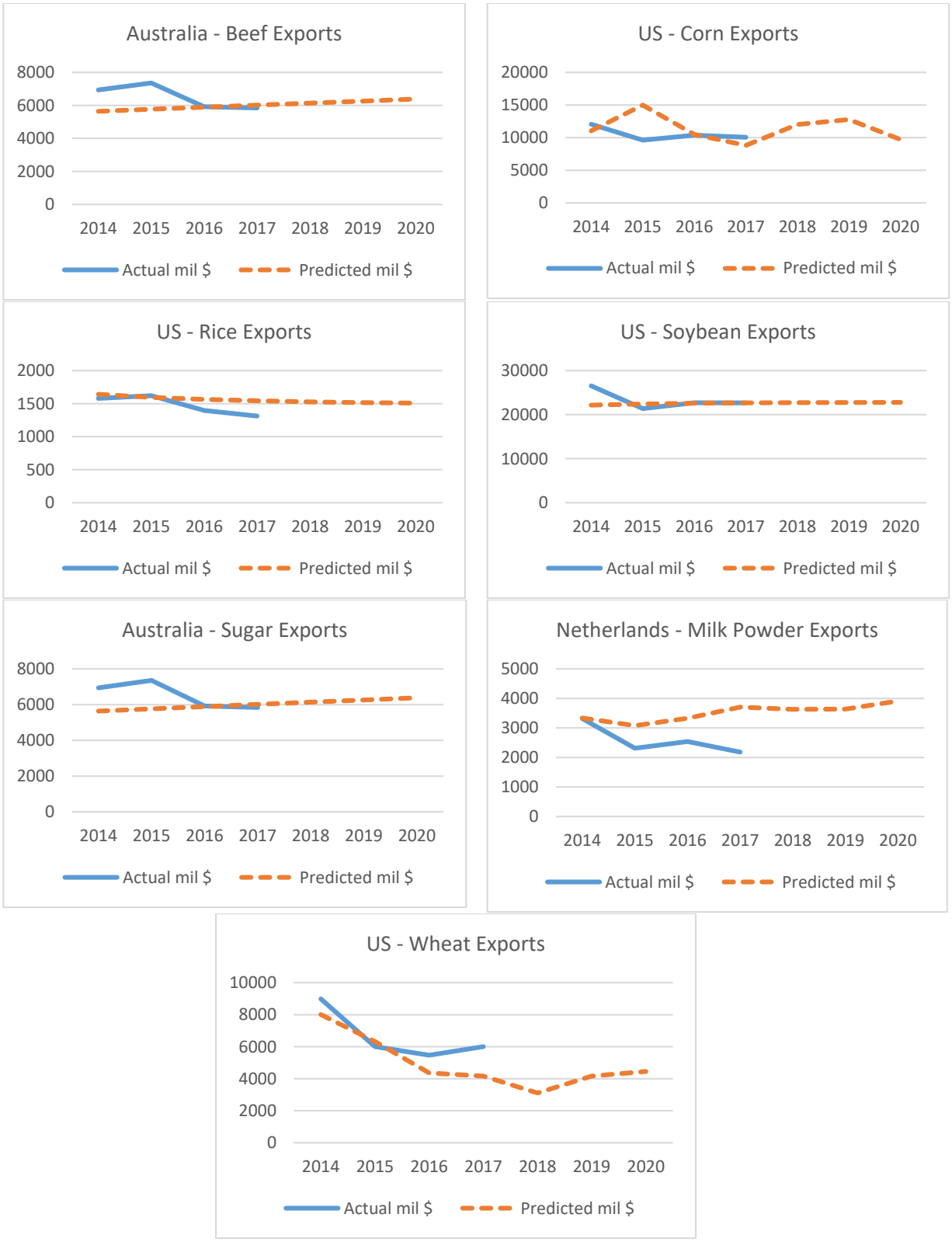


Table 1: Number of Exporting and Importing Countries

for Key Agricultural Products, 1970-2009

	1970s		*1980s*		*1990s*		*2000s*	
	Exporters	*Importers*	*Exporters*	*Importers*	*Exporters*	*Importers*	*Exporters*	*Importers*
Wheat	36	136	40	146	61	162	91	177
Maize	58	142	55	149	80	169	102	196
Rice	63	175	61	175	90	202	114	219
Sugar	60	165	56	174	81	207	111	222
Beef	62	159	64	175	82	202	109	216
Milk	48	184	49	186	81	206	116	219
Soybean	30	71	38	91	63	118	87	161

Source: Liapis (2012)

Table 2: Supervised Models' Validation Measures

Commodity	Observations	LightGBM	XGBoost	Extra Trees-Bagging
Beef	*Training-11385* *Test-3562*	0.623	0.510	**0.659**
Corn	*Training-9921* *Test-2706*	**0.680**	0.611	0.642
Rice	*Training-14259* *Test-4118*	**0.514**	0.380	0.450
Soy	*Training - 5421* *Test - 1669*	0.537	**0.541**	0.517
Sugar	*Training-4434* *Test-1616*	0.323	0.217	**0.340**
Milk	*Training - 16141* *Test - 3789*	0.290	0.281	**0.379**
Wheat	*Training - 10648* *Test - 5307*	0.262	0.195	**0.272**

Table 3: Ranking Variables by Information Gain (normalized values)

Variables	Beef	Corn	Rice	Soybeans	Sugar	Milk Powder	Wheat
Population_Destination	2	1	1	1	1	1	1
GDP Per Capita_Destination	1	2	2	2	2	2	3
Population_Origin	3	3	4	3	5	4	4
GDP Per Capita_Origin	4	5	7	4	3	5	5
Distance	5	6	6	5	6	3	2
Time	6	4	3	7	4	6	6
Longitude_Destination	7	7	5	6	7	7	7
Latitude_Destination	8	8	8	8	8	8	8
Political Stability	9	9	9	9	9	9	9

Table 4: Relative Importance of Variables based on Information Gain (percent)

Variables	Beef	Corn	Rice	Soybeans	Sugar	Milk Powder	Wheat
Population_Destination	99	100	100	100	100	100	100
GDP Per Capita_Destination	100	97	98	100	93	91	83
Population_Origin	82	88	69	75	67	78	76
GDP Per Capita_Origin	80	65	54	74	81	71	64
Distance	70	64	58	72	61	88	91
Time	63	71	72	59	74	64	63
Longitude_Destination	56	59	63	65	53	63	63
Latitude_Destination	50	53	51	54	47	60	60
Political Stability	20	26	36	43	36	33	32

*100 indicates the variable with the highest information gain

**All other variables (26) together accounted for another 84 to 126 % of information gains across the seven commodities

Appendix: PPML Estimation of the Gravity model

Examining factors that determine trade patterns has largely been accomplished through the use of the gravity model—described as the workhorse of international trade (Yotov et al., 2016) and one of the most successful empirical models in economics (Anderson and van Wincoop, 2003). The gravity model has been developed over the years to incorporate different types of variables depending on the questions answered, and to address some of the empirical issues such as the large amount of zeros in trade flows. In particular, the Poisson Pseudo Maximum Likelihood (PPML) laid out in Santos Silva and Tenreyro (2006) allows for the inclusion of zero trade flows and corrects for heteroscedasticity which can plague estimation of the gravity model. Additionally, work (e.g., Feenstra, 2004) noted the importance of accounting for observable and unobservable country-level heterogeneity and multilateral resistance terms through the use of importer and exporter fixed-effects.

To complement the ML application, the PPML regression considered belwo included imports in levels as the dependent variable, while the non-dummy independent variables are specified in log terms. Given problems in variable selection (Yotov et al., 2016), the ML application was used as the basis for the following specification:

$$\text{(A.1) } v_{ij}^{k} = exp[\alpha_i + b_j + GDP_i + GDP_j + Pop_i + Pop_j + \beta_1 \text{distance}_{ij} + \beta_2 \text{time} + \beta_3 \text{longitude}_j + \beta_4 \text{latitude}_j + \beta_5 \text{poltical}_j$$

where v_{ij}^{k} is the value of imports from country *i* to country *j*. α_i and b_j are importer and exporter fixed effects, GDP and population (*Pop*) are defined for importers and exporters, $distance_{ij}$ is the logged distance between countries, *time* is a time trend variable, *longitude* is the longitude for the exporting country, *latitude* is the latitude for the exporting country, and political represents

the political stability of the exporting country (except for sugar, which has political stability for the importing country as in the ML application). These variables come from the same source as the ML application, i.e., Gurevich and Herman (2018).

The econometric results from using the variables suggested by ML indicate high R-square for most commodities (Appendix table 1). The PPML model does a better job explaining variation in some commodities more than others (Table 2 versus Appendix Table 1). Soy has the highest R-square, followed by corn and beef. The fit in PPML and ML are lower for rice and sugar. Most of the variables in the econometric model are statistically significant, and at the 0.01 percent level. Distance and the year trend are statistically significant at the 0.01 percent level for every model. The coefficient on the distance variable is negative as expected since longer distance involves more costs, and thus, decreasing the amount of trade that occurs. The magnitude of the distance coefficient is largest for corn, indicating that commodity is most affected by distance; while the -1.06 coefficient on sugar is the smallest. The time trend variable is positive and is between 0.04 and 0.06, indicating that trade in these commodities is increasing over time. GDP is positive and statistically significant for many of the commodity exporting countries, indicating that richer countries are more likely to export these agricultural commodities. The coefficient on GDP importer is mixed—beef, rice, and sugar has a positive coefficient; while corn, milk, and wheat have a negative coefficient. Similarly, the coefficients on population (both the exporter and importer) are mixed. This indicates that trade for each commodity is mixed, thus any agriculture aggregation is missing this heterogeneity. The political stability coefficient is positive and statistically significant for most commodities. I.e., countries that are more political stable are more likely to be exporters. The coefficients on

latitude and longitude (exporters) are also mixed. The largest beef, milk, and soy (beef) exporters tend to be in lower latitudes (longitudes).[1]

Appendix table 1. Results from PPML Estimation of the Gravity Model

	Beef	Corn	Milk	Rice	Soy	Sugar	Wheat
gdp_x	0.43***	0.36***	0.71***	0.01	1.33***	0.36***	-0.15**
	(0.09)	(0.12)	(0.15)	(0.09)	(0.11)	(0.10)	(0.08)
gdp_m	0.35***	-0.21**	-0.23*	0.14*	-0.31	0.99***	-0.19*
	(0.12)	(0.10)	(0.13)	(0.08)	(0.19)	(0.23)	(0.10)
pop_x	-0.14	0.43***	0.32*	-0.11	0.23	-0.41***	0.52***
	(0.10)	(0.15)	(0.17)	(0.08)	(0.30)	(0.14)	(0.13)
pop_m	-0.46***	-0.13	-0.23**	-0.08	-0.61***	-0.27**	0.17**
	(0.10)	(0.20)	(0.09)	(0.12)	(0.12)	(0.11)	(0.07)
distance	-1.68***	-1.93***	-1.28***	-1.61***	-1.51***	-1.06***	-1.85***
	(0.03)	(0.05)	(0.02)	(0.03)	(0.08)	(0.04)	(0.03)
political*	0.04***	0.07***	0.06***	0.00	0.12***	0.06***	0.08***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
$latitude_x$	-0.57***	-0.03	-0.11***	0.09***	-0.05*	0.01	0.12***
	(0.04)	(0.02)	(0.02)	(0.01)	(0.03)	(0.01)	(0.03)
$longitude_x$	-0.11***	-0.01***	0.00	-0.01***	0.03***	-0.03***	-0.02***
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)
year	0.05***	0.05***	0.06***	0.05***	0.05***	0.05***	0.04***
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)
R^2	0.60	0.78	0.57	0.42	0.88	0.35	0.60
N	213,330	205,353	316,140	284,745	132,633	199,250	172,764

Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[1] While the ML application identified key variables explaining the variation in import data, PPML offered the direction of those effects. That is, ML offers options to select variables, and make and explain predictions, while PPML offers more structure in the context of gravity models. Future work should consider a deeper comparison of prescriptive ML models with high-dimension fixed effects specifications of PPML.