Using Machine Learning to Capture Heterogeneity in Trade Agreements

Scott L. Baier and Narendra R. Regmi

MERCATUS WORKING PAPER

MERCATUS CENTER
George Mason University
3434 Washington Blvd., 4th Floor, Arlington, Virginia 22201
www.mercatus.org

Scott L. Baier and Narendra R. Regmi, "Using Machine Learning to Capture Heterogeneity in Trade Agreements," Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, March 2021.

Abstract

In this paper, we employ machine learning techniques to capture heterogeneity in free trade agreements. The tools of machine learning allow us to quantify several features of trade agreements, including volume, comprehensiveness, and legal enforceability. Combining machine learning results with gravity analysis of trade, we find that more comprehensive agreements result in larger estimates of the impact of trade agreements. In addition, we identify the policy provisions that have the most substantial effect in creating trade flows. In particular, legally binding provisions on antidumping, capital mobility, competition, customs harmonization, dispute settlement mechanism, e-commerce, environment, export and import restrictions, freedom of transit, investment, investor-state dispute settlement, labor, public procurement, sanitary and phytosanitary measures, services, technical barriers to trade, telecommunications, and transparency tend to have the largest trade creation effects.

JEL codes: F10, F13

Keywords: Free Trade Agreements, Machine Learning, Gravity Model

Author Affiliation and Contact Information

Scott L. Baier
John E. Walker Department of Economics
Clemson University
sbaier@clemson.edu

Narendra R. Regmi
Department of Economics
University of Wisconsin–Whitewater
regmin@uww.edu

This paper can be accessed at https://www.mercatus.org/publications/regulation/using-machine-learning-capture-heterogeneity-trade-agreements.

Using Machine Learning to Capture Heterogeneity in Trade Agreements

Scott L. Baier and Narendra R. Regmi

I. Introduction

The gravity model—often referred to as the workhorse model in international trade—has been widely used to study the effects of various determinants of trade flows across countries. Drawing from the analogy of physical science, Tinbergen (1962) first used the gravity equation to evaluate the impact of free trade agreements (FTAs) on bilateral trade flows. Since Tinbergen (1962), numerous papers have studied the role of various determinants of trade flows, such as adjacency, common language, presence of a bilateral agreement, and past colonial links, to name a few (cf. Head and Mayer 2014). However, challenges abound in properly estimating the impact of free trade agreements on trade volumes. Broadly, there are two challenges that researchers ought to address to quantify the effects of FTAs accurately. We will now summarize the challenges and discuss this paper's potential contributions in light of those challenges.

The first challenge is the potential endogeneity of trade policies. Countries do not randomly select into trade agreements. Without controlling for *selection* effects, empirical estimates are likely biased. It could be the case that two countries are more likely to enter into a trade agreement if they are already significant trading partners. The possible reverse causality implies that the trade policy variable is endogenous, thereby making identification challenging. This reverse causality may lead the researcher to conclude that trade agreements increased bilateral trade when, in fact, it was higher trade volumes that *caused* the trade agreement. Another possible source of endogeneity is when trade policies are

correlated to unmeasurable trade costs between the two countries, which may induce the two countries to "self-select" into a free trade agreement (see Baier and Bergstrand [2007] for a detailed analysis of the sources of endogeneity). If the trade agreements are correlated with unmeasured trade costs and the researcher does not account for the possible correlation, the empirical estimates of trade agreements' effects may be biased toward zero.

Several studies identify the issue of endogeneity and show that the estimates that do not allow for simultaneous determination of trade policy and trade flows are highly underestimated (cf. Trefler 1993; Lee and Swagel 1997; Baier and Bergstrand 2007). Trefler (1993) shows that when trade policy is modeled endogenously, allowing for the simultaneous determination of imports and nontariff barriers (NTBs) in US manufacturing, the restrictive impact of NTBs increases tenfold. Lee and Swagel (1997) also find that the exogenous treatment of trade flows and the presence of an FTA leads to an underestimation of the role of FTAs.

Before Baier and Bergstrand (2007), most studies did not control for the potential endogeneity of trade agreements; see, for example, Tinbergen (1962); Brada and Méndez (1985); and Frankel, Stein, and Wei (1995, 1997). Baier and Bergstrand (2007) show that the estimates obtained using cross-section instrumental variable and control function approaches are unstable in the presence of endogeneity. They show that using panel data with country-pair fixed effects accounts for endogeneity and leads to an unbiased estimation of the impact of FTAs. They find that an FTA will, on average, increase two member countries' trade by about 86 percent after 15 years, six times the effect using ordinary least squares (OLS).

The second challenge in properly estimating the impact of FTAs is that an extensive heterogeneity exists across FTAs with regard to treaty design, coverage areas of trade policies, legal enforceability, and even the overall objectives. Consider, for example, the India-Bhutan Free Trade Agreement and the North American Free Trade Agreement. The former does not go beyond commitments in tariff liberalization of goods. In contrast, the latter covers commitments in a wide array of topics, including goods and services liberalization, investment liberalization, and environmental and labor standards, to name a few. The motive of signing free trade agreements can also differ across country pairs. Rosen (2004) provides evidence that the US-Israel Free Trade Agreement and the US-Jordan Free Trade Agreement are means of using trade policy to pursue foreign policy objectives.

Unlike endogeneity, capturing heterogeneity in trade agreements remains a challenge in the literature. The most common approach is to treat the existence of an FTA between trading partners as an indicator variable to estimate the common average effect across agreements (cf. Baier and Bergstrand 2007; Anderson, Milot, and Yotov 2011; Anderson and Yotov 2016). However, this methodology cannot take into account the fact that FTAs differ extensively in the scope and the level of integration commitments between the parties.

Baier, Bergstrand, and Feng (2014) provide the first evidence of the differential effects of different types of trade agreements. They categorize a large number of trade agreements on the basis of their level of economic integration, namely, nonreciprocal preferential trade agreements, reciprocal preferential trade agreements, free trade agreements, customs unions, common markets, and economic unions based on the traditional definition by Frankel, Stein, and Wei (1997). They find that economic unions and common markets are associated with higher levels of bilateral trade compared with nonreciprocal and reciprocal

trade agreements. In short, Baier, Bergstrand, and Feng (2014) were the first to show that more comprehensive trade agreements result in more trade.

Employing a Melitz-style model, Baier, Bergstrand, and Clance (2018) account for heterogeneity in trade agreements that can be associated with lower fixed trade costs versus marginal trade costs. Baier, Yotov, and Zylkin (2019) use a two-stage approach to account for the heterogeneity in trade agreements. In the first stage, they estimate a gravity equation to measure the trade impact by agreement and then use these estimates to identify economic, geographical, and political factors associated with the agreements. Their second approach involves what Kohl, Brakman, and Garretsen (2016) call a "specialist" approach, in which researchers examine the effect of individual FTAs on the members' trade volumes.

At best, researchers restrict themselves to a small number of trade agreements in a geographical region with a shorter time horizon. As such, the generalizations from these studies can be difficult, and the policy implications might be limited.

However, the commitments in modern trade agreements go far beyond tariff barriers or factor market integration and incorporate numerous policy areas that may affect the overall bilateral trade costs among member countries. For example, Horn, Mavroidis, and Sapir (2010) examine the content of 14 European Commission and 14 US trade agreements by going through the 28 agreements in their entirety and identify up to 52 policy areas included in the trade agreements signed by the European Union and the United States. Their work suggests that examining the differential effects of trade agreements requires a detailed analysis of this extensive set of policy areas.

Kohl, Brakman, and Garretsen (2016) conducted the only study (referred to as KBG hereafter) that examines the coverage and restrictiveness of policy provisions in trade

agreements. They read through 296 trade agreements and quantified them on the basis of the 26 policy areas they cover and the legal enforceability of those provisions. For each provision, a score of 0 is assigned if the document does not include the provision, 1 if it does include the provision, and 2 if it includes the provision and the provision is also legally enforceable. They then add up those scores and obtain a composite score for each trade agreement.

Although highly informative and distinct in its approach, the KBG study suffers from a few problems, and our paper, using a machine learning approach, mitigates those problems.

First, the assignment of 0, 1, and 2 seems quite ad hoc, and a score of 2 does not necessarily imply that the provision is comprehensive. In the KBG study, a score of 2 is assigned as long as a policy provision is covered and if it contains a binding word such as *shall* or *must*, regardless of the provision's specificity and comprehensiveness. In this paper, we use techniques grounded in the distribution of words and phrases in trade agreements to examine the comprehensiveness of coverage areas and their legal enforceability.

Second, with the KBG approach, many provisions must be within the border between 0, 1, and 2. Using machine learning techniques, we will be able to keep judgment calls at bay. In addition, we will be able to capture nuances that the KBG approach might have missed.

Third, KBG adds up scores across all provisions and obtains a composite score to measure "depth" for each agreement. Adding up scores across all provisions may not be prudent because not all provisions are trade promoting; some provisions may restrict trade as well. Rather than add up the scores, we employ a simple form of machine learning that

identifies the *clusters* of provisions that are typically grouped together. The appeal of the clustering approach is that it identifies the patterns in the data that are commonly associated with different types of agreements.

In some sense, the KBG approach assumes that each provision is equally important for the pair of countries. However, some provisions may not be necessary for some agreements, and other agreements' provisions may be present, but they may be less complex. For example, for North-South trade agreements, we expect we are more likely to observe labor and environmental provisions that are more detailed than those we might observe in a North-North agreement.[1] Finally, the KBG study examines only the coverage and the enforceability of 26 trade policy areas, whereas our study examines 36 policy areas. Therefore, our study allows for a broader understanding of trade agreements.[2]

In this paper, we proceed in three steps. First, we classify trade agreements into distinct clusters using *k*-means clustering, an unsupervised learning method. Unsupervised learning is a machine learning technique that identifies patterns in datasets without users' having to provide any labels. In contrast, supervised learning methods train on labeled data to figure out patterns in new data.

Second, we use multilabel classification, a supervised learning method, to examine the nature of each cluster for the coverage and comprehensiveness of specific trade policies of

---

[1] We recognize that this implies that the determination and the complexity of the provisions are endogenously determined. However, the theoretical and empirical determination of what provisions are included in a trade agreement is beyond the scope of this paper.

[2] In this study, we include 10 additional provisions not identified by Kohl, Brakman, and Garretsen (2016) in their original paper and use a supervising learning technique to identify the presence of the provisions. In future work, we would like our algorithm to identify distinct provisions on its own.

interest. It turns out that the groupings of trade agreements obtained from clustering in the first stage carry economic interpretation. The clustering exercise can separate shallow agreements from deep agreements quite well.

Third, we then run the gravity model of trade regression and find evidence that trade agreements that cover a wide range of trade policy areas with high legal enforceability lead to the most substantial impact on trade flows. Our approach also allows us to identify the provisions that have the most substantial impact on trade creation. These provisions are antidumping, capital mobility, competition laws, customs harmonization, dispute settlement framework, e-commerce, environment, export and import restrictions, freedom of transit, investment, labor, public procurement, sanitary and phytosanitary measures, services, subsidies and countervailing measures, technical barriers to trade, telecommunications liberalization, and transparency. These results have far-reaching implications in shaping modern trade policy institutions.

The paper is structured as follows. Section II explains the data used in the analysis. Sections III and IV present discussions on clustering and classification, respectively. Section V discusses the gravity model of trade flows as applied in the context of our study. Section VI provides the main empirical results and findings, section VII provides robustness analysis, and section VIII concludes.

II. Database of Trade Agreements

We use a total of 280 FTA texts in our analysis. The trade agreement documents—as well as data on relevant bilateral country pairs—come from Baier and Bergstrand (2017).[3] The economic integration agreement (EIA) database is a panel dataset that includes bilateral agreements between pairs of countries annually from 1950 to 2012.[4] We combine these data with bilateral trade data from Comtrade so that we have 159 countries in our study, which implies that we have approximately 25,000 country pairs annually from 1970 to 2014. For each bilateral pair, the dataset maps the relevant EIA (along with the PDF document of the actual treaty) governing the bilateral relationship and when the EIA status of the bilateral pair changes.[5] We remove annexes, protocols, and schedules and focus on the text's main body to ensure that our clustering results are not driven by the mere presence of schedules, protocols, and annexes.

III. Clustering

The application of machine learning methods requires converting trade agreement texts to numerical vectors. We remove punctuation, symbols, numbers, and white spaces and segment trade agreement texts into single words and two-word phrases, also known as bigrams, in the clustering literature. We then count the frequency of single words and two-

[3] These data are similar to the database used in the KBG study. The main difference is that these data cover more periods. For each agreement, there are hyperlinks to associated PDF documents that contain the terms of the agreement or modifications to the agreement.
[4] Appendix A contains the list of all the countries used in the empirical analysis.
[5] Appendix B lists the bilateral agreements for each pair and the years in force.

word phrases within a trade agreement and normalize the frequency by a vector of normalized frequencies of single words and two-word phrases.[6]

We then count the frequency of single words and two-word phrases within a trade agreement and normalize the frequency by the document's size. Essentially, each trade agreement is uniquely represented by a vector of normalized frequencies of single words and two-word phrases.[7]

The goal of cluster analysis is to find natural groupings of objects based on a number of object features. The first stage of clustering involves determining the "appropriate" number of clusters, and the second stage involves grouping the trade agreements into the "appropriate" number of clusters. We use one of the most widely used clustering techniques, called the *k*-means algorithm owing to Hartigan and Wong (1979), and apply standard procedures in choosing the appropriate number of clusters by minimizing the sum of squared errors between the empirical mean of a cluster and the objects assigned to that cluster over all clusters.

We will now formally introduce the *k*-means algorithm. Let $X = \{x_i\}$, $x_i \in \mathsf{R}^{\mathsf{D}}$ be the set of trade agreements to be clustered into a set of K clusters, $C = \{c_k : k = 1,\ldots,K\}$, where D is the number of features per trade agreement. Assume a priori that there exist K clusters with cluster centers $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^D$.[8] K-means clustering solves

$$argmax_C \sum_{i=1}^{K} \sum_{x \in C_i} \|x_i - \mu_i\|^2. \qquad (1)$$

---

[6] We also conduct the exact analysis done in the main body of this paper in the robustness section by removing extremely rare and common words and two-word phrases, and the main results still hold.

[7] Appendix C discusses the methods in detail.

[8] Although we assume that there exist K centers, we will eventually update this on the basis of the value of our loss function and the application in hand.

The implementation of this optimization takes place in the following steps:

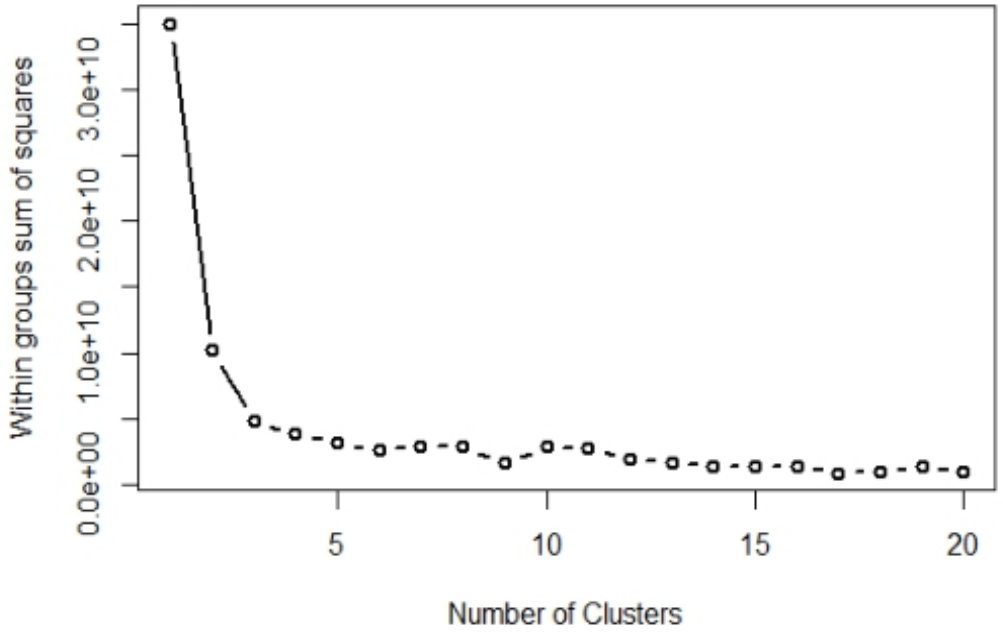
1) Choose an initial number of clusters, k, based on domain knowledge.
2) Initialize cluster centers $\mu_1, \mu_2, \ldots, \mu_k$ arbitrarily.
3) Given the fixed cluster centers, choose the optimal group assignment for each data point (trade agreements) x_i on the basis of the closest cluster center.
4) Update $\mu_1, \mu_2, \ldots, \mu_k$ on the basis of group assignments of x_i.
5) Repeat steps 3 and 4 until convergence—that is, until the cluster's centroids do not move.

We repeat the preceding steps for different values of K and choose an appropriate number of clusters, $k*$.

A. Optimal Number of Clusters

We use one of the most commonly used techniques in the literature to determine the appropriate number of clusters. The technique is known as the "elbow method," where within-groups sums of squares are plotted against the number of clusters. If the plot resembles an arm, then the "elbow" on the arm is the appropriate number of clusters. Figure 1 plots the within-groups sums of squares against the number of clusters. As suggested by the elbow method, the appropriate number of clusters is anywhere between four and five. The within-group sum of squares does not fall much after the fifth cluster, implying that the upper bound on the number of clusters is five. Throughout the rest of the paper, we will present the results when trade agreements are split into four and five clusters.

Figure 1. Within-Groups Sums of Squares



B. Five Clusters vs. Four Clusters

As we move from five clusters to four clusters, it is natural for a few trade agreements to change their cluster memberships. However, it would be a cause for concern if many agreements were switching their cluster memberships. To investigate the movement of country pairs between clusters, we report the correlation between the two sets of clusters. It turns out that the clusters appear to be relatively stable. Table 1 shows the correlation matrix between the two sets of clusters. As we move from five clusters to four clusters, most of the trade agreements stay in their "natural" groups as indicated by the perfect correlation (correlation = 1.00) in three of the cluster groups. The only time we see movements in cluster

assignments is when the pairs in clusters 1 and 2 (in the case with five clusters) merge into a single cluster, and the rest of the clusters' membership does not change.[9]

Table 1. Correlation between the Two Sets of Clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	**0.42**	–0.21	–0.16	–0.22
Cluster 2	**0.65**	–0.33	–0.25	–0.35
Cluster 3	–0.51	**1.00**	–0.13	–0.18
Cluster 4	–0.39	–0.13	**1.00**	–0.14
Cluster 5	–0.54	–0.18	–0.14	**1.00**

Note: This table shows the correlation matrix between the two sets of clusters. As we move from five clusters to four clusters, clusters 1 and 2 merge into one cluster, whereas the rest of the clusters do not change, as indicated by a correlation of 1.00.

C. Stability of Clustering Results

We ensure the stability of our clustering results by running 1,000 iterations of *k*-means clustering in the cases of both four and five clusters. We then compute the correlation between the 1,000 sets of clustering results against the base clustering result that we use in this paper.

In the case of four clusters, the mean correlation is 0.9796, and the standard deviation is 0.0656. This outcome indicates that clustering results are highly stable, and trade agreements do not change cluster membership erratically. The correlation is also relatively high in the case of five clusters. The mean correlation is 0.9294, and the standard deviation is 0.1513. Evidently, the lower correlation and higher standard deviation occur with five

[9] We could also use seven clusters as indicated by the elbow method; however, as we will discuss, the four and five clusters provide a cleaner economic interpretation.

clusters as some of the pairs switch between clusters 1 and 2. Therefore, this result makes us confident that the *k*-means clustering is a promising way to find natural grouping in the trade agreements.

IV. The Characteristics of the Identified Clusters

Thus far, we have gathered trade agreements into their natural groupings. However, we have not yet presented any insights into the actual content of the trade agreements in each cluster. We now turn to a discussion of the supervised learning method, which enables us to analyze the content of trade agreements in each cluster.

A. Supervised Learning: Classification

Supervised learning involves inferring the underlying function from labeled training data. The training data comprise a set of training examples and a label associated with each training example. On the basis of the training examples and the user-specified labels for these examples, the supervised learning method infers the label for test data. More formally, given training examples of the form $(x_1,y_1),\ldots(x_N,y_N)$, where x_i is a vector of features and y_i is an assigned label, the goal of a learning algorithm is to infer a function $g : X \rightarrow Y$ and thus to predict the output label for an unseen test sample.

In the context of our study, we train our model with instances of 36 trade policy provisions and allow the machine learning algorithm to "crawl" through each paragraph of the trade agreement to estimate the likelihood of the paragraph's being about one or more policy areas. The 36 policy areas we identify are agriculture, anticorruption, antidumping, capital mobility, competition, consumer protection, customs administration, dispute

settlement, e-commerce, education and training cooperation, energy, environment, financial cooperation, freedom of transit, export and import restrictions, industrial cooperation, institutional arrangements, intellectual property rights, investment, investor-state dispute settlement, labor, money laundering and illicit drugs, political dialogue, public procurement, safeguard procedures, sanitary and phytosanitary measures, science and technology, services, small and medium-sized enterprises, state aid, state trading enterprises, subsidies and countervailing measures, technical barriers to trade, telecommunications, transparency, and transportation infrastructure.

We train the model with examples of highly comprehensive and legally binding provisions for each policy area. For policy areas that are already covered in the World Trade Organization (WTO) agreements, we seek commitments above and beyond the WTO agreements. For policy areas that are not already a part of the WTO agreements, we compare the provisions in all trade agreements and choose the most comprehensive ones. What we mean by legally binding is that the provision is very specific; the provision contains at least one restrictive word such as *shall* or *must*. The provision specifies the course of action in case either party deviates from the commitment listed in the provision. This course of action may be different from a generic dispute settlement present in the trade agreements. The following provision is a training example for the *Investment* provision:

National Treatment

1. Each Country shall accord to investors of the other Country and to their investments treatment no less favorable than that it accords in like circumstances to its own investors and to their investments with respect to the establishment, acquisition, expansion, management, operation, maintenance, use, possession, liquidation, sale, or other disposition of investments (hereinafter referred to in this Chapter as investment activities). Each Country shall accord to investors of the other Country and to their investments treatment no less favorable than that it accords in like circumstances to

investors of a third State and to their investments, with respect to investment activities.

Article 9.6: Performance Requirements.
Neither Party may impose or enforce any of the following requirements, enforce any commitment or undertaking, in connection with the establishment, acquisition, expansion, management, conduct, operation, or sale or other disposition of an investment of an investor of a Party or of a non-Party in its territory to:

a) export a given level or percentage of goods or services;
b) achieve a given level or percentage of domestic content;
c) purchase, use, or accord a preference to goods produced in its territory, or to purchase goods from persons in its territory;
d) relate in any way the volume or value of imports to the volume or value of exports or to the amount of foreign exchange inflows associated with such investment;
e) restrict sales of goods or services in its territory that such investment produces or provides by relating such sales in any way to the volume or value of its exports or foreign exchange earnings;
f) transfer a particular technology, production process, or other proprietary knowledge to a person in its territory;
g) supply exclusively from the territory of the Party the goods that it produces or the services that it provides to a specific regional market or to the world market.

The preceding paragraph is highly comprehensive about investment commitments between the parties, and the language of the text is also highly enforceable. We feed training examples such as the one above for each of the 36 provisions and perform a multilabel classification on each paragraph of trade agreements to estimate the paragraph's likelihood of being associated with one or more than one category. Essentially, we find a similarity measure between an untrained paragraph and the trained examples. Because a paragraph can be related to more than just one policy domain, we also allow for multilabel classification with a technique called cross-training (Boutell et al. 2004). The idea is to use paragraphs with multiple labels more than once during training. This implies that each training example can be a positive instance for more than one category during training. We

then perform multilabel classification as 36 individual binary classification problems using a one-versus-rest strategy. This method has proved to be effective in classifying multilabel images in the classification literature. Therefore, for each unit of analysis (i.e., a paragraph), we estimate the probability of the unit's being relevant in one or more of the 36 categories. As is common in the literature, we put a threshold probability of 0.5 for the document to both be pertinent about the category and have a minimum level of legal enforceability in its language (McLaughlin and Sherouse 2016).

We experimented with the two popular classification algorithms: *k*-nearest neighbors and random forest classifier. We evaluated both models' performance using macro F1-scores averaged across all classes, the details of which will be presented in the next subsection. This process reveals that the *k*-nearest neighbors classifier performs superior to random forest classifier. We use *k*-nearest neighbors with five neighbors and uniform weights to classify our trade agreement paragraphs. With this, a brief discussion of the *k*-nearest neighbors classifier is thus warranted.

B. K-Nearest Neighbors Classifier

K-nearest neighbors classifier is one of the most popular nonparametric classification algorithms used in many machine learning applications. Given training data $D = (x_1,y_1),\ldots,(x_N,y_N)$ and a positive integer k where x_i is a vector of features and y_i represents self-assigned labels for that set of features, the class prediction for a new test point x_0 involves identifying the K observations in the training data that are closest to x_0. Therefore, the conditional probability for class j is given as

$$Pr(Y = j|X = x) = \frac{1}{K} \sum_{\Omega} I(y_i = j) \ , \quad (2)$$

where Ω represents the set of K observations closest to the x_0 (Friedman, Hastie, and Tibshirani 2001).

C. Model Evaluation

We evaluate our models and choose the appropriate parameters by performing classification on 5-fold cross-validated data. Cross-validation is a model validation technique to assess the generalizability of the classification results. The idea is that the training examples are further separated into training and test set, and the classification model is estimated only on the training set. The model is then used to predict the class of the test set.

The metric used to evaluate the classification model is the macro F1-scores averaged across all classes. F1-scores are the harmonic average of precision and recall. In a binary classification setting, precision is the percentage of selected items that are correct. So precision is given by

$$\frac{True\ Positives}{True\ Positives + False\ Positives.} \quad (3)$$

Similarly, recall is defined as the percentage of correct items that are selected. So, recall is given by

$$\frac{True\ Positives}{True\ Positives + False\ Negatives.} \quad (4)$$

Then F1-score is given by

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision\ +\ Recall\ .} \quad (5)$$

This measure strikes a balance between precision and recall. For our classification, we average F1-scores across all classes. Table 2 presents average F1-scores across all classes from 5-fold cross-validation for *k*-nearest neighbors and the random forest classifier with various parameter values. It can be seen that the *k*-nearest neighbors classifier performs better than the random forest classifier for every parameter value tested. Within the *k*-nearest neighbors classifier, the model with five neighbors and uniform weight performs the best classification, which we use to classify our test data.

Table 2. Macro F1-Scores by Model: Classification

Model	Average Macro F1-Scores
K-nearest neighbors, weight = uniform and *k* = 7	0.8366 (0.0696)
K-nearest neighbors, weight = distance and *k* = 7	0.8160 (0.0704)
K-nearest neighbors, weight = uniform and *k* = 6	0.8243 (0.0632)
K-nearest neighbors, weight = distance and *k* = 6	0.8214 (0.0635)
K-nearest neighbors, weight = uniform and *k* = 5	0.8373 (0.0649)
K-nearest neighbors, weight = distance and *k* = 5	0.8123 (0.0649)
Random forest classifier, number of trees = 5	0.5712 (0.0704)
Random forest classifier, number of trees = 10	0.5388 (0.0635)
Random forest classifier, number of trees = 15	0.5991 (0.0628)

Note: F1-scores are averaged across all classes using 5-fold cross-validation. For the *k*-nearest neighbors model, *k* represents the number of neighbors used. The weight parameter of distance indicates that within a class, closer neighbors will have a greater influence than neighbors that are farther away.

D. Provision Scores

Table 3 reports the relative frequencies of 36 provisions across the trade agreements in our sample. We place a threshold score of 0.5 for a provision to be counted in a trade agreement in this table. The most common provision is the liberalization of export and import restrictions, which is to be expected in free trade agreements, with more than 95 percent of the trade agreements containing at least a minimum level of content as well as legal enforceability. This result seems obvious because the purpose of most trade agreements is to remove tariffs and other import restrictions. The next four common provisions are on safeguard procedures, competition, intellectual property rights, and agriculture, respectively, with each present in more than half of the agreements. Similarly, the least common provision is consumer protection, with only about 5 percent of the trade agreements containing this provision. The next least common provisions are anticorruption, transport infrastructure, energy, and money laundering and illicit drugs, respectively. Each of them is present in 8 percent or fewer of the trade agreements in our sample.

We also perform correlation on all of our provision scores simultaneously to understand what provisions tend to co-occur in trade agreements. Figure 2 represents the correlation matrix heatmap. Dark blue squares represent a very high positive correlation, dark red squares represent a very high negative correlation, and white squares represent zero correlation. We reorganize the figure so that the provisions that are more likely to co-occur are grouped together in the boxes outlined in black. For instance, provisions on anticorruption, customs administration, freedom of transit, capital mobility, labor, investment, environment, investor-state dispute settlement, transparency, dispute settlement

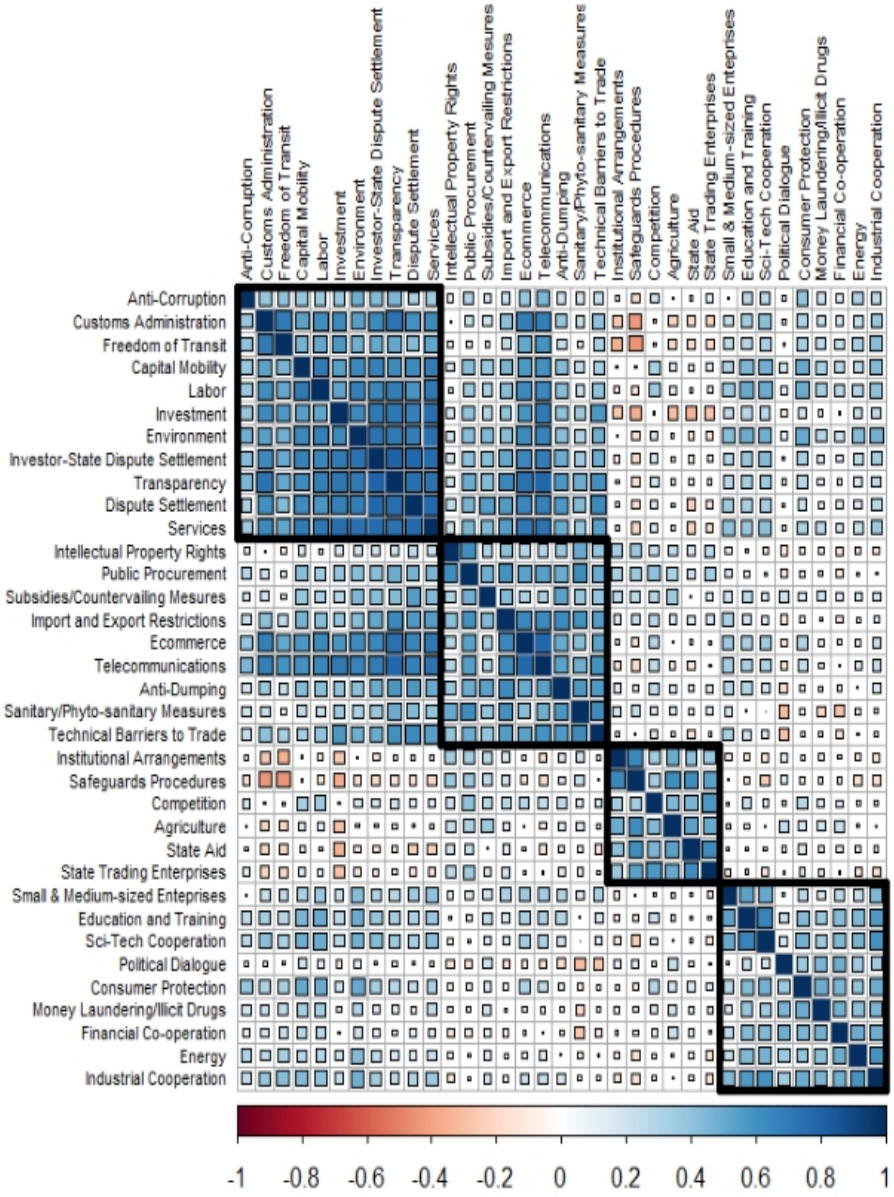
(state-state), and services are more likely to be present together in trade agreements, as seen in the topmost left square. It turns out that according to our gravity analysis, which is the topic of the next section, most of these provisions are highly trade promoting. Similarly, provisions on intellectual property rights, public procurement, subsidies and countervailing measures, export and import restrictions, e-commerce, antidumping, sanitary and phytosanitary measures, and technical barriers to trade do co-occur in trade agreements, as do provisions on institutional arrangements, safeguard procedures, competition, agriculture, state aid, and state trading enterprises.

Table 3. Most Common to Least Common Provisions

Provision	% of FTAs	Provision	% of FTAs
Export/import restrictions	95	Services	26
Safeguard procedures	71	Investment	23
Competition	57	Investor-state dispute settlement	19
Intellectual property rights	57	Political dialogue	18
Agriculture	53	Freedom of transit	17
Dispute settlement	49	Science and technology cooperation	17
State trading enterprises	47	Telecommunications	16
Institutional arrangements	44	Environment	14
Technical barriers to trade	42	E-commerce	13
Public procurement	41	Industrial cooperation	12
Customs administration	38	Education and training	11
Capital mobility	38	Financial cooperation	10
State aid	36	Small and medium-sized enterprises	9
Transparency	35	Money laundering/illicit drugs	8
Sanitary/phytosanitary measures	33	Energy	7
Antidumping	32	Transportation infrastructure	7
Labor	29	Anticorruption	6
Subsidies/countervailing measures	29	Consumer protection	5

Note: FTA = free trade agreement.

Figure 2. Heatmap: Correlation across Provisions



Furthermore, there is strong support for provisions on political dialogue, consumer protection, measures against money laundering and illicit drugs, financial cooperation, energy, and industrial cooperation to co-occur together, as seen in the bottom right square. These provisions are commonly present in trade agreements written by the European Union

with African economies and some eastern European countries before they joined the European Union. On an individual provision level, some relationships are worth mentioning. For example, provisions on transparency and harmonization of customs administration generally tend to feature together. The same goes for services and investment provisions. However, state aid and investment provisions tend not to feature in the same trade agreement.

Recall that the *k*-means cluster groups the trade agreements into different clusters without any intent to subscribe economic meaning to the clusters. However, given the provision scores for each of the clusters, it is easy to label the clusters for the depth of the agreements. Tables 4 and 5 show that these clusters can be organized by the depth of the trade agreement going from the least comprehensive (labeled "Shallowest") to the most comprehensive agreements (labeled "Deepest"). Thus, the *k*-means cluster groups by the depth of the agreement, whereby a deeper agreement implies that the cluster is more likely to have higher provision scores than a shallow agreement. One might be inclined to think that these clusters map directly into the type of trade agreements used in Baier, Bergstrand, and Feng (2014). Recall that in their study, the authors classify the agreements by FTAs, customs union, common market, and economic union. Tables 6 and 7 show the correlation between these definitions and the clusters identified in this paper. Although the clusters used in this paper, organized by depth, show some degree of correlation with the types of agreements in Baier, Bergstrand, and Feng (2014), the correlation is far from perfect.

Table 4. Provision Scores for Five Clusters

Provision	Shallowest	Shallow	Moderate	Deep	Deepest
Agriculture	0.25	0.70	0.66	0.67	0.49
Anticorruption	0.00	0.00	0.02	0.37	0.15
Antidumping	0.08	0.27	0.52	0.37	0.70
Capital mobility	0.09	0.25	0.55	0.87	0.89
Competition	0.19	0.51	0.49	0.57	0.70
Consumer protection	0.00	0.016	0.013	0.28	0.16
Customs administration	0.26	0.14	0.69	0.85	0.90
Dispute settlement	0.27	0.36	0.78	0.82	0.92
E-commerce	0.00	0.02	0.18	0.56	0.44
Education & training cooperation	0.02	0.04	0.21	0.16	0.30
Energy	0.03	0.06	0.13	0.17	0.17
Environment	0.05	0.09	0.22	0.65	0.51
Export & import restrictions	0.69	0.85	0.91	0.92	0.93
Financial cooperation	0.02	0.06	0.25	0.19	0.19
Freedom of transit	0.17	0.06	0.29	0.59	0.48
Industrial cooperation	0.09	0.10	0.24	0.24	0.30
Institutional arrangements	0.34	0.90	0.75	0.64	0.73
Intellectual property rights	0.28	0.77	0.57	0.70	0.70
Investment	0.10	0.12	0.34	0.74	0.70
Investor-state dispute settlement	0.04	0.07	0.29	0.78	0.62
Labor	0.03	0.06	0.46	0.84	0.87
Money laundering/illicit drugs	0.00	0.05	0.19	0.11	0.15
Political dialogue	0.14	0.22	0.29	0.25	0.19
Public procurement	0.05	0.45	0.47	0.77	0.64
Safeguard procedures	0.21	0.90	0.71	0.66	0.65
Sanitary & phytosanitary measures	0.13	0.38	0.45	0.68	0.64
Science & technology cooperation	0.10	0.07	0.27	0.40	0.43
Services	0.07	0.10	0.44	0.74	0.80
Small & medium-sized enterprises	0.03	0.05	0.19	0.15	0.20
State aid	0.16	0.51	0.39	0.48	0.40
State trading enterprises	0.19	0.45	0.68	0.63	0.52
Subsidies & countervailing measures	0.08	0.38	0.48	0.44	0.45
Technical barriers to trade	0.23	0.34	0.64	0.71	0.78
Telecommunications	0.01	0.00	0.18	0.70	0.56
Transparency	0.00	0.06	0.60	0.91	0.88
Transport infrastructure	0.01	0.04	0.15	0.18	0.19

Note: This table reports the average provision scores across trade agreements within each cluster for five clusters. Low provision scores across all clusters implies that the provision is only prevalent in a few trade agreements.

Table 5. Provision Scores for Four Clusters

Provision	Shallowest	Moderate	Deep	Deepest
Agriculture	0.54	0.66	0.67	0.49
Anticorruption	0.00	0.02	0.37	0.15
Antidumping	0.21	0.52	0.37	0.70
Capital mobility	0.19	0.55	0.87	0.89
Competition	0.40	0.49	0.57	0.70
Consumer protection	0.01	0.13	0.28	0.16
Customs administration	0.18	0.69	0.85	0.90
Dispute settlement	0.33	0.78	0.82	0.92
E-commerce	0.01	0.18	0.56	0.44
Education & training cooperation	0.03	0.21	0.16	0.30
Energy	0.05	0.13	0.17	0.17
Environment	0.07	0.22	0.65	0.51
Export & import restrictions	0.79	0.91	0.92	0.93
Financial cooperation	0.05	0.25	0.19	0.19
Freedom of transit	0.10	0.29	0.59	0.48
Industrial cooperation	0.10	0.24	0.24	0.30
Institutional arrangements	0.71	0.75	0.64	0.73
Intellectual property rights	0.60	0.57	0.70	0.70
Investment	0.11	0.34	0.74	0.71
Investor-state dispute settlement	0.06	0.29	0.78	0.62
Labor	0.05	0.46	0.84	0.87
Money laundering/illicit drugs	0.04	0.19	0.11	0.15
Political dialogue	0.20	0.29	0.25	0.19
Public procurement	0.31	0.47	0.77	0.64
Safeguard procedures	0.66	0.71	0.66	0.64
Sanitary & phytosanitary measures	0.30	0.45	0.68	0.64
Science & technology cooperation	0.08	0.27	0.40	0.43
Services	0.09	0.44	0.74	0.80
Small & medium-sized enterprises	0.04	0.19	0.15	0.20
State aid	0.39	0.39	0.48	0.40
State trading enterprises	0.51	0.45	0.63	0.52
Subsidies & countervailing measures	0.28	0.48	0.44	0.45
Technical barriers to trade	0.30	0.64	0.71	0.78
Telecommunications	0.01	0.18	0.70	0.56
Transparency	0.04	0.60	0.91	0.88
Transport infrastructure	0.03	0.15	0.18	0.19

Note: This table reports the average provision scores across trade agreements within each cluster for four clusters. Low provision scores across all clusters implies that the provision is only prevalent in a few trade agreements.

Table 6. Correlation Clusters and Types of Economic Integration Agreements

	FTA	CU	CM	EU
Shallowest	0.399	0.055	−0.007	0.078
Shallow	0.617	0.074	−0.011	−0.008
Moderate	0.274	0.332	0.248	0.110
Deep	0.126	−0.009	0.575	0.432
Deepest	0.239	0.431	0.004	−0.006

Note: This table shows the correlation matrix between the clusters and the EIAs used by Baier, Bergstrand, and Feng (2014). FTA = free trade agreement; CU = customs union; CM = common market; EIA = economic integration agreement; EU = Economic Union.

Table 7. Correlation Clusters and Types of Economic Integration Agreements

	FTA	CU	CM	EU
Shallowest	0.737	0.093	−0.013	0.040
Moderate	0.274	0.332	0.248	0.110
Deep	0.126	−0.009	0.575	0.432
Deepest	0.239	0.431	0.004	−0.006

Note: This table shows the correlation matrix between the clusters and the EIAs used by Baier, Bergstrand, and Feng (2014). FTA = free trade agreement; CU = customs union; CM = common market; EIA = economic integration agreement; EU = economic union.

V. Gravity Analysis with Cluster

In section III, we cluster trade agreements in their natural groupings. In the subsequent section, we describe the actual contents of trade agreements belonging to each cluster. Here, we combine the cluster membership information obtained from the first stage with gravity analysis of trade flows to obtain more insights into the heterogeneous impacts of FTAs on trade flows. The gravity regression results—along with the clustering and classification procedures performed in the previous sections—will enable us to answer policy-relevant questions, such as what provisions or set of provisions matters the most for trade flows and under what conditions.

The gravity model—often referred to as the workhorse model in international trade—is widely used to study the effects of various determinants of country pairs' goods and factor flows. The gravity model uses the metaphor of Newton's law of gravitation and predicts that the trade flow between two countries is directly proportional to the product of their economic mass, typically measured by GDP, and inversely proportional to the distance between the two countries. Tinbergen (1962) uses the gravity equation to evaluate the effects of FTA dummy variables on bilateral trade flows among European countries. Despite its empirical success, the initial applications were atheoretical. Anderson (1979) conducted the first study to provide microeconomic foundations to the gravity equation by employing a model where goods were specific to the country of origin and agents had "Armington" preferences over goods from each country. Since then other studies have shown that "gravity-like" relationships emerge in a variety of settings (Bergstrand 1985; Eaton and Kortum 2002; Melitz 2003; Anderson and van Wincoop 2003).

Following Feenstra (2006) and Baier and Bergstrand (2007), we present the following gravity specification:

$$\ln(X_{ij,t}) = \delta_{i,t} + \delta_{j,t} + \chi_{ij} + \sum_{k=1}^{K} FTA_{ij,t} * cluster_k + \epsilon_{ij,t} , \qquad (6)$$

where the $lnX_{ij,t}$ represents the logarithm of trade flows from country i to country j, $\delta_{i,t}$ is the vector of exporter-time fixed effects that captures any exporter-specific factors, $\delta_{j,t}$ is the vector of importer-time fixed effects that captures any importer-specific factors, and the vector χ_{ij} denotes time-invariant country-pair fixed effects.

These fixed effects also control for "multilateral resistance," first introduced by Anderson and van Wincoop (2003). Anderson and van Wincoop argue that trade flows

between two countries depend not only on the bilateral trade barriers but also on the trade resistance across all trading partners. As shown in Baier and Bergstrand (2007), country-pair fixed effects will take into account any possible endogeneity of free trade agreements. $FTA_{ij,t}$ is a dummy variable that takes a value of 1 if country *i* and country *j* have an FTA between them and a value of 0 otherwise. $Cluster_k$ is also a dummy variable and only takes a value of 1 if $FTA_{ij,t}$ belongs to cluster *k*, and *K* is the total number of clusters.

One of the potential drawbacks of specification (6) is that, in the presence of heteroskedasticity, the coefficient estimates are likely to be biased and inconsistent, as pointed out by Santos Silva and Tenreyro (2006), hereafter SST. They suggest using a Poisson pseudo-maximum likelihood (PPML) estimator. Following SST, we also estimate

$$X_{ij,t} = \exp\left[\delta_{i,t} + \delta_{j,t} + \chi_{ij} + \sum_{k=1}^{K} FTA_{ij,t} * Cluster_k\right] + \epsilon_{ij,t}. \quad (7)$$

We estimate specifications (6) and (7) for two different sets of clustering results: one for five clusters and one for four gravity results, where we do not remove overly common words and very rare words.

VI. Results and Discussion

After performing the preceding text mining exercises along with the gravity analysis of trade flows, we have a compelling way to answer the following questions: Do more comprehensive agreements result in more trade creation? What trade policy provision or set of trade policy provisions matters the most for trade flows? To compare our results to previous work, table 8 presents some baseline results. Columns 1 and 2 provide baseline estimates of the gravity equation when we use the definition of trade agreements in Baier and Bergstrand (2007) and Anderson and Yotov (2016). In column 1, the results from the standard baseline OLS model

with an indicator variable for a trade agreement indicate that the presence of a trade agreement increases trade by about 50 percent (=(exp(.401) − 1)*100). For the PPML specification, the trade agreements boost bilateral trade by about 11 percent. If we use KBG data for provisions, we see that the coefficient on depth is 0.406 for OLS; therefore, if a country pair had all of the provisions and they were all enforceable (depth = 1), trade would be expected to be about 50 percent higher. If they had a depth measure of 0.50, trade would be about 23 percent higher. For the PPML specification the coefficient on depth is notably lower at 0.077.

Table 8. Baseline Gravity Regressions

Cluster	OLS FE FTA	PPML FE 5 clusters	OLS FE 4 clusters	PPML FE 4 clusters
FTA	0.401*** (0.015)	0.108*** (0.015)		
KBG depth			0.406*** (0.020)	0.077*** (0.011)
R^2	0.8296	-	0.8297	-
N	611,948	611,948	611,948	611,948

Note: This table reports the coefficients on the gravity regressions with country-pair fixed effects, importer fixed effects, and exporter fixed effects. FE = fixed effects; FTA = free trade agreement; KBG = Kohl, Brakman, and Garretsen (2016); OLS = ordinary least squares; PPML = Poisson pseudo-maximum likelihood.
***1%, **5%, *10%.

Turning to our data with four and five clusters, the first column of table 9 presents the gravity trade regression results when trade agreements are grouped into five clusters where the model is estimated using OLS. This result indicates that the most comprehensive set of agreements, denoted by *deepest*, has the largest impact on trade flows. The coefficient for this cluster is about 0.630. It indicates that the most comprehensive trade agreements can increase two members' trade flows by nearly 88 percent (=(exp(.630) − 1)*100 = 87.8%).

Similar results are obtained for the PPML estimates, but the magnitude of the coefficient is smaller.

Table 9. Gravity Regressions for Four and Five Clusters

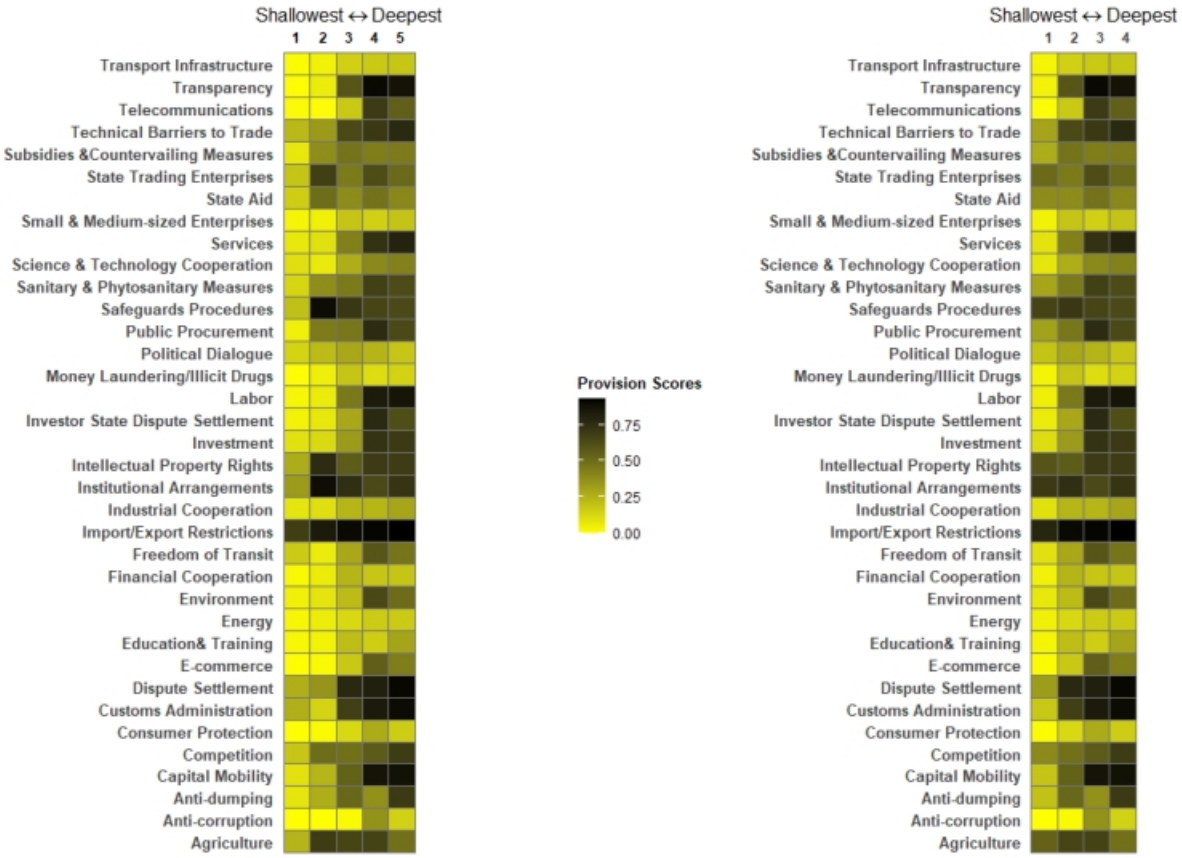
	OLS FE	PPML FE	OLS FE	PPML FE
Cluster	**5 clusters**	**5 clusters**	**4 clusters**	**4 clusters**
Shallowest	0.275*** (0.035)	0.118*** (0.015)	0.319*** (0.020)	0.096*** (0.011)
Shallow	0.336*** (0.022)	0.083*** (0.013)	-	-
Moderate	0.321*** (0.025)	0.045*** (0.013)	0.322*** (0.025)	0.046*** (0.013)
Deep	0.622*** (0.028)	0.138*** (0.119)	0.624*** (0.028)	0.137*** (0.012)
Deepest	0.630*** (0.031)	0.161*** (0.012)	0.632*** (0.031)	0.158*** (0.012)
R^2	0.8296	-	0.8297	-
N	611,681	611,681	611,681	611,681

Note: FE = fixed effects; OLS = ordinary least squares; PPML = Poisson pseudo-maximum likelihood. ***1%, **5%, *10%.

It is interesting to note that our research also enables us to identify the specific provisions that are predominantly present in these highly trade-generating agreements. Table 4 shows the average cluster scores for each provision across all five clusters. For better readability, we present a heatmap for cluster scores in figure 3. For consistency with gravity regression results, we have reordered clusters so that number *1* refers to the shallowest cluster and number *5* refers to the deepest cluster. For each provision, the darker the cell, the higher the cluster's performance in that specific provision. The prominent provisions in the *deepest* cluster are antidumping, capital mobility, competition, customs harmonization, dispute settlement mechanism, e-commerce, environment, export and import restrictions, freedom of transit, investment, investor-state dispute settlement, labor, public procurement, sanitary and phytosanitary measures, services, technical barriers to

trade, telecommunications, and transparency. There are some provisions—such as e-commerce, freedom of transit, and environment laws—for which the cells are not very dark because these provisions are relatively less frequent than the other provisions, and because the scores on these provisions are merely the average across the trade agreements in each cluster. One should interpret these provisions' scores relative to other clusters.

Figure 3. Heatmap of Provision Scores for Five Clusters and Four Clusters, Respectively, without Removing Extremely Common Phrases or Words



The most comprehensive set of trade agreements remains the most influential in increasing trade flows when we split trade agreements into only four clusters. The third and fourth columns of table 9 show the gravity trade regression results in the case of four clusters.

The FTA coefficient is about the same, 0.632 for OLS and 0.158 for PPML. This result should not be a surprise given the high degree of correlation between the clusters discussed above. The same provisions feature prominently in this cluster as well, as evidenced by high scores on those provisions in table 5 or the dark squares in the second column of figure 3.

The FTA coefficients for the next most comprehensive cluster, "deep," are also quite high for OLS and PPML. The coefficients are 0.624 and 0.137, respectively. Examining the provision scores reveals that the similar set of provisions that was prominently present in the most comprehensive agreements is also dominantly present, albeit with relatively less comprehensiveness and a lower degree of legal enforceability.

One crucial question that remains unanswered is whether the benefits of incorporating the "trade-inducing" provisions extend to an arbitrarily chosen pair of countries. In the context of our study, despite only covering commitments in export and import restrictions, why do some of the least comprehensive agreements generate a comparable level of trade flows to some of the clusters that were classified as "moderate," which tends to cover a wide range of policy areas, albeit with less comprehensive coverage than the "deeper" and "deepest" agreements? Upon further examination of the agreements in the "shallowest" cluster, it turns out that the majority of these agreements are between transition economies. Some examples include Armenia-Ukraine, the Eurasian Economic Community, Georgia-Ukraine, Kazakhstan-Kyrgyzstan, Russia-Azerbaijan, and Russia-Belarus. If we have agreements between transition economies, it might be the mere establishment of a trade agreement that strongly and positively affects bilateral trade. It may be the case that bilateral trade declined after the Cold War and the trade agreements helped rekindle some of the old trade patterns.

VII. Robustness Analysis

We also perform our clustering analysis by removing extremely common and extremely rare words and phrases, as well as overly used words and phrases. In particular, we ignore words and phrases that appear in fewer than 1 percent of the trade agreements. The idea is to avoid rare words and phrases, such as the names of the parties involved in the trade agreement, that may be driving our clustering results. Similarly, we ignore single words and two-word phrases that appear in more than 99 percent of the trade agreements. The idea is to avoid common words such as *chapter*, *annexes*, *party*, *parties*, and so forth that are quite prevalent in all trade agreements. The optimal number of clusters did not change in this case. The appropriate number of clusters as suggested by the elbow method remains the same, further validating our results.

As in the baseline results, we run the correlation between the two sets of clusters and present the robustness analyses for five and four clusters. Table 10 shows the correlation matrix between the two sets of clusters. The own-cluster correlation is pretty high for all clusters. The cross-cluster correlation is very low, indicating that most trade agreements stay in their "natural" groups as we change the number of clusters.

The gravity analysis of trade flows and the corresponding provisions that tend to have the largest impact on trade flows remain quite stable. The first column of table 11 shows the gravity trade regression results in the case of five clusters, and table 12 shows the average cluster scores for each provision. As before, we also present a heatmap for better readability in figure 4.

Table 10. Correlation between the Two Sets of Clusters for Robustness Analysis

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	**0.45**	−0.23	−0.17	−0.24
Cluster 2	**0.61**	−0.30	−0.24	−0.34
Cluster 3	−0.50	**0.99**	−0.13	−0.18
Cluster 4	−0.39	−0.13	**1.00**	−0.14
Cluster 5	–0.54	–0.18	–0.14	**1.00**

Note: This table shows the correlation matrix between the two sets of clusters. As we move from five clusters to four clusters, cluster 1 and cluster 2 merge into one cluster, while the rest of the clusters do not change as indicated by a correlation of 1.00 except for cluster 3.

Table 11. Gravity Regressions-Robustness Analysis

Cluster	OLS FE 5 clusters	PPML FE 5 clusters	OLS FE 4 clusters	PPML FE 4 clusters
Shallowest	0.435*** (0.028)	0.128*** (0.014)	0.320*** (0.020)	0.094*** (0.011)
Shallow	0.208*** (0.025)	0.077*** (0.013)	-	-
Moderate	0.351*** (0.026)	0.041*** (0.014)	0.321*** (0.025)	0.048*** (0.013)
Deep	0.622*** (0.028)	0.138*** (0.119)	0.623*** (0.028)	0.137*** (0.012)
Deepest	0.639*** (0.028)	0.162*** (0.012)	0.629*** (0.031)	0.158*** (0.012)
R^2	0.8296	-	0.8296	-
N	611,681	611,681	611,681	611,681

Note: This table reports the coefficients on the gravity regressions with country-pair fixed effects, country-import fixed effects, and country-export fixed effects. FE = fixed effects; OLS = ordinary least squares; PPML = Poisson pseudo-maximum likelihood.

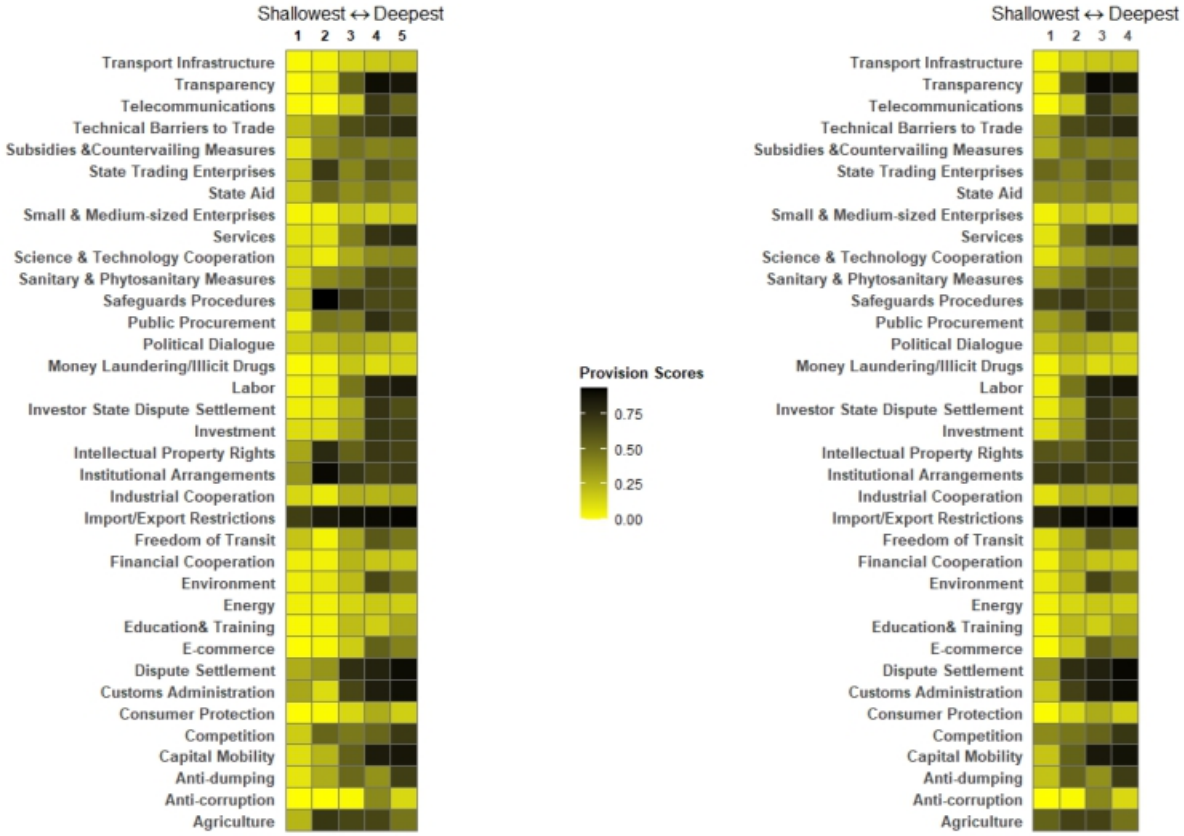
***1%, **5%, *10%.

Table 12. Provision Scores for Five Clusters for Robustness Analysis

Provisions	Shallowest	Shallow	Moderate	Deep	Deepest
Agriculture	0.25	0.70	0.66	0.67	0.49
Anticorruption	0.00	0.00	0.02	0.37	0.15
Antidumping	0.08	0.27	0.52	0.37	0.70
Capital mobility	0.09	0.25	0.55	0.87	0.89
Competition	0.19	0.51	0.49	0.57	0.70
Consumer protection	0.00	0.016	0.013	0.28	0.16
Customs administration	0.26	0.14	0.69	0.85	0.90
Dispute settlement	0.27	0.36	0.78	0.82	0.92
E-commerce	0.00	0.02	0.18	0.56	0.44
Education & training cooperation	0.02	0.04	0.21	0.16	0.30
Energy	0.03	0.06	0.13	0.17	0.17
Environment	0.05	0.09	0.22	0.65	0.51
Export & import restrictions	0.69	0.85	0.91	0.92	0.93
Financial cooperation	0.02	0.06	0.25	0.19	0.19
Freedom of transit	0.17	0.06	0.29	0.59	0.48
Industrial cooperation	0.09	0.10	0.24	0.24	0.30
Institutional arrangements	0.34	0.90	0.75	0.64	0.73
Intellectual property rights	0.28	0.77	0.57	0.70	0.70
Investment	0.10	0.12	0.34	0.74	0.70
Investor-state dispute settlement	0.04	0.07	0.29	0.78	0.62
Labor	0.03	0.06	0.46	0.84	0.87
Money laundering/illicit drugs	0.00	0.05	0.19	0.11	0.15
Political dialogue	0.14	0.22	0.29	0.25	0.19
Public procurement	0.05	0.45	0.47	0.77	0.64
Safeguard procedures	0.21	0.90	0.71	0.66	0.65
Sanitary & phytosanitary measures	0.13	0.38	0.45	0.68	0.64
Science & technology cooperation	0.10	0.07	0.27	0.40	0.43
Services	0.07	0.10	0.44	0.74	0.80
Small & medium-sized enterprises	0.03	0.05	0.19	0.15	0.20
State aid	0.16	0.51	0.39	0.48	0.40
State trading enterprises	0.19	0.45	0.68	0.63	0.52
Subsidies & countervailing measures	0.08	0.38	0.48	0.44	0.45
Technical barriers to trade	0.23	0.34	0.64	0.71	0.78
Telecommunications	0.01	0.00	0.18	0.70	0.56
Transparency	0.00	0.06	0.60	0.91	0.88
Transport infrastructure	0.01	0.04	0.15	0.18	0.19

Note: This table reports the average provision scores across trade agreements within each cluster in the case of five clusters for robustness analysis. Low provision scores across all clusters implies that the provision is only prevalent in a few trade agreements.

Figure 4. Heatmap of Provision Scores for Five Clusters and Four Clusters, Respectively, when Removing Extremely Common and Rare Words or Phrases (Robustness Analysis)



The gravity results indicate that the most comprehensive set of agreements, denoted by *deepest*, have the largest impact on trade flows. The coefficient for this cluster is about 0.639. This indicates that the most comprehensive trade agreements can increase two members' trade flows by up to 90 percent ($e^{0.639} = 1.895$). The prominent provisions in the "deepest" cluster are antidumping, capital mobility, competition laws, customs harmonization, dispute settlement mechanism, e-commerce, environment, export and import restrictions, freedom of transit, investment, investor-state dispute settlement, labor, public procurement, sanitary and phytosanitary measures, services, technical barriers to trade, telecommunications liberalization, and transparency.

The most comprehensive set of agreements remains the most influential with regard to increasing trade flows as we split trade agreements into only four clusters. The third column of table 11 shows the gravity trade regression results with four clusters. The FTA coefficient is 0.629. The prominent provisions present in these clusters are the same, as evidenced by high scores on the same set of provisions in table 13 and figure 4.

VIII. Conclusion and Future Work

In this paper, we capture heterogeneity in free trade agreements using the tools of machine learning. The tools of machine learning allow us to quantify several features of trade agreements, including volume, comprehensiveness, and legal enforceability. First, we employ unsupervised learning techniques to categorize agreements into four to five clusters. Second, we use supervised learning techniques to analyze the content in each cluster in relation to the coverage of policy areas and the legal enforceability of those provisions. Finally, assuming that the trade flow effects are common across agreements within each cluster, we run the gravity model of trade flows to estimate the differential effects of free trade agreements. We find that more comprehensive agreements result in larger trade creation. In addition, we also identify the provisions that are generally the most successful in driving trade flows. The provisions are antidumping, capital mobility, competition, customs harmonization, dispute settlement mechanism, e-commerce, environmental standards, export and import restrictions, freedom of transit, investment, investor-state dispute settlement, labor, public procurement, sanitary and phytosanitary measures, services, technical barriers to trade, telecommunications, and transparency.

Table 13: Provision Scores for Four Clusters for Robustness Analysis

Provisions	Shallowest	Moderate	Deep	Deepest
Agriculture	0.54	0.66	0.67	0.49
Anticorruption	0.00	0.02	0.37	0.15
Antidumping	0.21	0.52	0.37	0.70
Capital mobility	0.19	0.55	0.87	0.89
Competition	0.40	0.49	0.57	0.70
Consumer protection	0.01	0.13	0.28	0.16
Customs administration	0.18	0.69	0.85	0.90
Dispute settlement	0.33	0.78	0.82	0.92
E-commerce	0.01	0.18	0.56	0.44
Education & training cooperation	0.03	0.21	0.16	0.30
Energy	0.05	0.13	0.17	0.17
Environment	0.07	0.22	0.65	0.51
Export & import restrictions	0.79	0.91	0.92	0.93
Financial cooperation	0.05	0.25	0.19	0.19
Freedom of transit	0.10	0.29	0.59	0.48
Industrial cooperation	0.10	0.24	0.24	0.30
Institutional arrangements	0.71	0.75	0.64	0.73
Intellectual property rights	0.60	0.57	0.70	0.70
Investment	0.11	0.34	0.74	0.71
Investor-state dispute settlement	0.06	0.29	0.78	0.62
Labor	0.05	0.46	0.84	0.87
Money laundering/illicit drugs	0.04	0.19	0.11	0.15
Political dialogue	0.20	0.29	0.25	0.19
Public procurement	0.31	0.47	0.77	0.64
Safeguard procedures	0.66	0.71	0.66	0.64
Sanitary & phytosanitary measures	0.30	0.45	0.68	0.64
Science & technology cooperation	0.08	0.27	0.40	0.43
Services	0.09	0.44	0.74	0.80
Small & medium-sized enterprises	0.04	0.19	0.15	0.20
State aid	0.39	0.39	0.48	0.40
State trading enterprises	0.51	0.45	0.63	0.52
Subsidies & countervailing measures	0.28	0.48	0.44	0.45
Technical barriers to trade	0.30	0.64	0.71	0.78
Telecommunications	0.01	0.18	0.70	0.56
Transparency	0.04	0.60	0.91	0.88
Transport infrastructure	0.03	0.15	0.18	0.19

Note: This table reports the average provision scores across trade agreements within each cluster in the case of four clusters for robustness analysis. Low provision scores across all clusters implies that the provision is only prevalent in a few trade agreements.

Another question worth exploring is how the content and legal enforceability of a specific provision or a set of provisions affects trade flows in the relevant sector. For example, what aspects of provisions related to automobiles affect the trade flows in automobiles across the member countries? Is one method of rules of origin determination for tariff concessions more restrictive and trade hindering than the other? The answers to such questions may also be obtained by combining a standard empirical model of trade with the text analysis of provisions enabled by machine learning. We believe that this stream of research will have far-reaching implications in shaping modern trade institutions.

References

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

Anderson, James E., Catherine A. Milot, and Yoto V. Yotov. 2011. *The Incidence of Geography on Canada's Services Trade*. National Bureau of Economic Research, Cambridge, MA.

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

Anderson, James E., and Yoto V. Yotov. 2016. "Terms of Trade and Global Efficiency Effects of Free Trade Agreements, 1990–2002." *Journal of International Economics* 99: 279–98.

Baier, Scott L., and Jeffrey H. Bergstrand. 2007. "Do Free Trade Agreements Actually Increase Members' International Trade?" *Journal of International Economics* 71 (1): 72–95.

———. 2017. "Economic Integration Agreements: Historical Database of Entry into Economic Integration Agreements, 1960–2000." Ann Arbor, MI. https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/29762.

Baier, Scott L., Jeffrey H. Bergstrand, and Matthew W. Clance. 2018. "Heterogeneous Effects of Economic Integration Agreements." *Journal of Development Economics* 135 (C): 587–608.

Baier, Scott L., Jeffrey H. Bergstrand, and Michael Feng. 2014. "Economic Integration Agreements and the Margins of International Trade." *Journal of International Economics* 93 (2): 339–50.

Baier, Scott L., Yoto Yotov, and Thomas Zylkin. 2019. "On the Widely Differing Effects of Free Trade Agreements: Lessons from Twenty Years of Trade Integration." *Journal of International Economics* 116 (C): 206–26.

Bergstrand, Jeffrey H. 1985. "The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence." *Review of Economics and Statistics* 67 (3): 474–81.

Boutell, Matthew R., Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. "Learning Multi-Label Scene Classification." *Pattern Recognition* 37 (9): 1757–71.

Brada, Josef C., and José A. Méndez. 1985. "Economic Integration among Developed, Developing and Centrally Planned Economies: A Comparative Analysis." *Review of Economics and Statistics* 67 (4k): 549–56.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70 (5): 1741–79.

Feenstra, Robert C., 2006. *Advanced International Trade: Theory and Evidence.* Princeton, NJ: Princeton University Press.

Frankel, Jeffrey, Ernesto Stein, and Shang-Jin Wei. 1995. "Trading Blocs and the Americas: The Natural, the Unnatural, and the Super-Natural." *Journal of Development Economics* 47 (1): 61–95.

———. 1997. *Regional Trading Blocs in the World Economic System*. Washington, DC: Peterson Institute for International Economics.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Vol. 1, 10. Springer Series in Statistics. New York: Springer.

Hartigan, John A., and Manchek A. Wong. 1979. "Algorithm AS 136: A *K*-Means Clustering Algorithm." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28 (1): 100–108.

Head, Keith, and Thierry Mayer. 2014. "Gravity Equations: Workhorse, Toolkit, and Cookbook." In *Handbook of International Economics.* Vol. 4, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 131–95. Amsterdam: Elsevier.

Horn, Henrik, Petros C. Mavroidis, and André Sapir. 2010. "Beyond the WTO? An Anatomy of EU and US Preferential Trade Agreements." *World Economy* 33 (11): 1565–88.

Kohl, Tristan, Steven Brakman, and Harry Garretsen. 2016. "Do Trade Agreements Stimulate International Trade Differently? Evidence from 296 Trade Agreements." *World Economy* 39 (1): 97–131.

Lee, Jong-Wha, and Phillip Swagel. 1997. "Trade Barriers and Trade Flows across Countries and Industries." *Review of Economics and Statistics* 79 (3): 372–82.

McLaughlin, Patrick A., and Oliver Sherouse. 2016. "QuantGov: A Policy Analytics Platform." QuantGov, October 31.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71 (6): 1695–725.

Rosen, Howard. 2004. "Free Trade Agreements as Foreign Policy Tools: The US-Israel and US-Jordan FTAs." In *Free Trade Agreements: US Strategies and Priorities*, edited by Jeffrey J. Schott, 51–77. Washington, DC: Peterson Institute for International Economics.

Salton, Gerard, and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Santos Silva, J. M. C., and Silvana Tenreyro. 2006. "The Log of Gravity." *Review of Economics and Statistics* 88 (4): 641–58.

Tinbergen, Jan. 1962. *Shaping the World Economy*. New York: Twentieth Century Fund.

Trefler, Daniel. 1993. "Trade Liberalization and the Theory of Endogenous Protection: An Econometric Study of US Import Policy." *Journal of Political Economy* 101 (1): 138–60.

Appendix A

Countries Included in the Gravity Dataset

Albania, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Brazil, Brunei Darussalam, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Central African Republic, Chad, Chile, China, Colombia, Comoros, Congo, Costa Rica, Côte d'Ivoire, Croatia, Cyprus, Czech Republic, Democratic Republic of the Congo, Denmark, Djibouti, Dominica, Dominican Republic, Ecuador, Egypt, Equatorial Guinea, Estonia, Ethiopia, Fiji, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Grenada, Guatemala, Guinea, Guinea-Bissau, Honduras, Hong Kong SAR (China), Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Korea, Kuwait, Kyrgyzstan, Lao People's Democratic Republic, Latvia, Lebanon, Lesotho, Lithuania, Luxembourg, Macao, Macedonia, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Mauritania, Mauritius, Mexico, Moldova, Mongolia, Morocco, Mozambique, Namibia, Nepal, Netherlands, New Zealand, Niger, Nigeria, Norway, Oman, Pakistan, Panama, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Rwanda, St. Kitts and Nevis, St. Lucia, St. Vincent and the Grenadines, São Tomé and Príncipe, Saudi Arabia, Senegal, Sierra Leone, Singapore, Slovakia, Slovenia, South Africa, Spain, Sri Lanka, Sudan, Suriname, Swaziland, Sweden, Switzerland, Syrian Arab Republic, Tajikistan, Tanzania, Thailand, Togo, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Uganda, Ukraine, United Kingdom, United States, Uruguay, Uzbekistan, Venezuela, Vietnam, Yemen, Zambia, and Zimbabwe.

Appendix B

Trade Agreements by Year of Enforcement

Before 1970: European Economic Community (EEC) (1957), European Community (EC) (1958), European Free Trade Agreement (EFTA) (1960), EEC-Turkey (Ankara Agreement) (1963), Southern African Customs Union (1969)

1972: EFTA-EEC (Austria), EFTA-EEC (Norway), EFTA-EEC (Portugal), EFTA-EEC (Sweden), EFTA-EEC (Switzerland)

1973: Caribbean Community (CARICOM), EC-Iceland

1974: EC-Cyprus

1975: EEC-Israel

1981: Gulf Cooperation Council, EEC-Greece

1983: Australia–New Zealand FTA

1985: EEC–Spain, Portugal; USA-Israel

1988: Andean Community (Cartagena Agreement), Canada–US Free Trade Agreement (1988)

1991: Common Market for Eastern and Southern Africa (COMESA)

1992: Association of Southeast Asian Nations (ASEAN), Central European Free Trade Agreement (CEFTA), EC–Czech Republic, EFTA-Slovakia, EFTA-Turkey, European Union Treaty

1993: Armenia-Russia, Czech Republic–Slovakia, EC-Hungary, EFTA-Bulgaria, EFTA–Czech Republic, EFTA-Hungary, EFTA-Israel, EFTA-Poland, EFTA-Romania, EFTA-

Slovakia, EU-Poland FTA, Russia-Azerbaijan, Russia-Belarus, Russia-Kazakhstan, Russia-Tajikistan, Russia-Turkmenistan, Russia-Uzbekistan

1994: Baltic Free Trade Agreement–Industrial FTA, CARICOM-Colombia, EC-Bulgaria, European Economic Area, North American Free Trade Agreement, Russia-Georgia, Russia-Kyrgyzstan, Russia-Ukraine

1995: COMESA, EC-Israel, EC-Latvia, EC-Lithuania, EC-Turkey, EFTA-Slovenia, Mercosur (Argentina, Brazil, Paraguay, Uruguay), Mexico-Bolivia, Mexico-Colombia-Venezuela, Mexico–Costa Rica, West African Economic Monetary Union (WAEMU)

1996: Armenia-Kyrgyzstan, Armenia-Moldova, Azerbaijan-Ukraine, Bolivia-Chile, Canada-Chile, Czech Republic–Estonia, Czech Republic–Israel, EC-Morocco, EFTA-Estonia, EFTA-Latvia, Kazakhstan-Kyrgyzstan, Mercosur-Bolivia, Mercosur-Chile, Turkey-Israel, Turkmenistan-Ukraine, Uzbekistan-Ukraine

1997: Armenia-Turkmenistan, Armenia-Ukraine, Canada-Israel, Czech Republic–Latvia, Czech Republic–Lithuania, Czech Republic–Turkey, EFTA-Lithuania, EFTA-Morocco, Estonia-Slovenia, Estonia-Ukraine, Georgia-Azerbaijan, Georgia-Ukraine, Hungary-Israel, Israel–Slovak Republic, Kyrgyzstan-Moldova, Latvia-Slovenia, Lithuania-Slovenia, Macedonia-Slovenia, Poland-Israel, Poland-Lithuania, Slovak Republic–Estonia, Slovak Republic–Latvia, Slovak Republic–Lithuania, Turkey-Hungary

1998: Chile-Mexico, EC-Estonia, EC-Tunisia, India–Sri Lanka, Kyrgyzstan-Ukraine, Mercosur–Andean Community, Pan Arab Free Trade Agreement (PAFTA), Turkey-Bulgaria

1999: Armenia-Georgia, CEFTA-Bulgaria, EC-Slovenia, Egypt-Jordan, Egypt-Morocco, Hungary-Estonia, Israel-Slovenia, Kyrgyzstan-Uzbekistan, Lithuania-Turkey, Poland-

Latvia, Turkey-Estonia, Turkey-Macedonia, Turkey-Poland, Turkey–Slovak Republic, SICA (Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, Panama)

2000: Bulgaria-Macedonia, Central African Economic and Monetary Community (CEMAC), EC-Mexico, EC–South Africa FTA, EFTA-Morocco, Georgia-Kazakhstan, Georgia-Turkmenistan, Hungary-Latvia, Hungary-Lithuania, Mexico-Israel, New Zealand–Singapore, WAEMU

2001: Bosnia and Herzegovina–Croatia, East African Community (EAC), EFTA-FYROM, Guatemala-Mexico, Honduras-Mexico, Southern African Development Community, Turkey-Latvia, Turkey-Slovenia

2002: Armenia-Kazakhstan, Bulgaria-Israel, Central America–Dominican Republic, CARICOM–Dominican Republic, Chile–Costa Rica, EC-Croatia, EC-Jordan, EC-Macedonia, EFTA-Croatia, EFTA-Jordan, EFTA-Mexico, Eurasian Economic Community, South African Customs Union, Turkey–Bosnia and Herzegovina, Turkey-Croatia

2010: ASEAN-China, ASEAN-India, ASEAN-Japan, ASEAN–New Zealand–Australia, Canada-Peru, China–Costa Rica, China-Peru, EFTA-Albania, Eurasian Economic Community Customs Union, India-Korea, India-Nepal, India-Thailand, Japan-Vietnam, Peru-Singapore, Switzerland-Japan

2011: Malaysia–New Zealand, Turkey-Jordan

2012: Albania-Iceland, Albania-Norway, D-8 Preferential Trade Agreement, EFTA-Colombia, EFTA-Peru, EU-Korea, Japan-Peru, Korea-EU, Korea-Peru, Malaysia-Chile, Malaysia-India

Appendix C

From Text Documents to Numerical Feature Vectors

For either clustering or classification analysis, the text documents will first need to be converted to a vector of real numbers. We follow a three-step procedure that is commonly employed in natural language processing literature to transform text documents to numerical feature vectors. The first step involves assigning integer identification for each word or two-word combination, commonly referred to as tokenization. The trade agreement documents were tokenized using unigram (single word) and bigram counts (two-word phrases). The words for tokenization are defined as sequences of two or more alphabetic characters, excluding stop words, such as pronouns, articles, and prepositions that carry little meaning in differentiating one set of documents from another. We also remove punctuation, numbers, and white spaces. The second step is to count the number of occurrences of these tokens for each document in the collection of documents, commonly referred to as the corpus. The final step is to normalize each document so it has a feature matrix of fixed size and to weight tokens that occur in the majority of documents with diminishing importance. We use the tf-idf scheme developed by Salton and McGill (1983) to obtain weights for each token.