**Machine Learning in Gravity Models: An Application to Agricultural Trade**

Munisamy Gopinath

Feras A. Batarseh

Jayson Beckman

**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 and one of the most successful empirical models in economics (Yotov et al., 2016; Anderson and van Wincoop, 2003). The gravity model has been developed over the years to incorporate different explanatory variables depending on the questions to be 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 often plague estimation of the gravity model. Additionally, the importance of accounting for observable and unobservable country-level heterogeneity and multilateral resistance terms through the use individual and pair-wise fixed effects (Feenstra, 2004).

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

(A.1)

where is the value of imports from country *i* to country *j*. and are importer and exporter fixed effects, GDP and population (*Pop*) are defined for importers and exporters, 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). Most of the variables in the econometric model are statistically significant, and at the 0.01 percent level, but similarities and differences are visible from the R-square (e.g. rice versus soybean). 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, but the coefficient on GDP importer is mixed. Similarly, the coefficients on population (both the exporter and importer) are mixed. 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).

As noted earlier, specifying gravity models remains a major challenge. Attempts were made to introduce pair-wise fixed effects, e.g. origin-time and destination time, as in Yotov at al. (2016) as well as Correia, Guimaraes and Zylkin (2019). With several thousand such effects, convergence and multicollinearity, added to the above specification challenge. Compounding these issues, is the difficulty in identifying the relative importance of explanatory variables as (World Economic Outlook, IMF, 2019). A deeper comparison of prescriptive ML models with high-dimension fixed effects specifications of PPML should be considered in future work.

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



|  |
| --- |
| Standard errors in parentheses, \* *p* < 0.10, \*\* *p* < 0.05, \*\*\* *p* < 0.01 |