Data-Driven Management of International Trade Policy during Outlier Events

Feras A. Batarseh (George Mason University)
Gopinath Munisamy (University of Georgia)
Ruixin Yang (George Mason University)

1. Research policy question

Waves of international trade and food consumption dictate policy and farming options that lead to production practices of commodities that are then priced differently or exported to other countries – but what happens when those waves are majorly disrupted by an outlier event such as trade wars, economic shocks, embargoes, or pandemics?

2. Research methodology and data

Outlier events lead to uncertainty. Uncertainty is caused by multiple aspects within an outlier event, such as: food market shifts, trade shocks and shipping shortage of agricultural commodities, changes in consumer behavior, differing prices and demand, farm workers routines and altering supply, financial health of farms and farmers, and perishable commodities (such as fruit and vegetables) stuck in harbors and ports.

Data science methods presented in this abstract aim to mitigate uncertainty that is caused by outliers and provide guidance to policy makers, farmers deploying precision agriculture, and decision makers along the supply chain.

In recent years, many countries are concerned about rising trade deficits and their implications on employment, farming, markets, and wages. Real-time policy and farming decisions during outlier events are challenging tasks. For instance, during pandemics, premature deaths reduce the labor force; illness leads to absenteeism from work and eventually to reduced overall productivity of a nation. Economic resources flow to treatment and control measures to reduce disease spread to avoid a serious disruption to economic activity. During the Covid-19 outbreak, most models from the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) were way off (predictions were very different than actual values). The Reason being they were modeled without looking into the predictions through the lens of multivariate *outliers*. Albeit most traditional economic models deploy causality analysis to infer policy, regular models don't account much for outliers, and most analysts deal with outliers based on human experience or elimination of such events in models.

In our work; we present Deep Learning (DL), Genetic Algorithms (GA), and outlier detection models that aid in retraining models for better predictions. Additionally, we deploy Reinforcement Learning (RL) methods to extract more accurate economic causalities and prescriptive measures during outlier events. Data used are collected from the United States Department of Agriculture (USDA), as well as General Agreement on Trade Services (GATS). The data are cleaned, and injected into a SQL database for the analysis. During these outlier events, analysis is performed in uncharted waters. Issues arise such as the need to use daily if not hourly data (but not monthly data) for pattern recognition and predictions. Additionally, decisions become timelier and need to be executed in a quick manner using real time analysis and on-demand analytics.

The main goal is to present data-driven models that evaluate outliers (such as pandemics) and provide real time analysis for policy makers. An important dimension of the cost of a pandemic lies in its impact on income, consumer behavior, unemployment, mortality rates, production, and multiple other effects. Such factors ought to be modeled so better real time decisions are presented with more evidence and data-driven reasoning. However, the results of data mining endeavors are majorly driven by data quality. Throughout these deployments, serious show-stopper problems are usually unresolved, such as: data collection ambiguities, data imbalance, hidden biases in data, the lack of domain information, and data incompleteness. In a traditional data science lifecycle, outliers, bias, variance, ensemble machine learning, boosting, over and under sampling, and data wrangling are measures to create *context* and mitigate output quality issues and misinformation. As an output, dashboard development is applied for two categories of end users: 1. policy makers, educators and researchers, and 2. farmers deploying precision agriculture technologies. Using the datasets mentioned, we present a big data framework that considers such challenges and provides economic data analysis in real time.

3. Approach and findings (context-driven data analytics)

Deploying context to represent outlier events is not a straight forward task. The context within a dataset can be represented as ***features***. Features in general fall into three categories: primary features, irrelevant features, and contextual features. Primary features are the traditional ones which are pertinent to a particular domain. Irrelevant features are features which are not helpful and can be safely removed, while contextual features are the ones to pay attention to. That categorization helps in eliminating irrelevant data but doesn't help in clearly defining context.

Data science algorithms that don't realize their context could have an opacity problem. This can cause models to be (for example) ignorant, racist, or sexist (such as the famous Google model). Biases which arise in the data are independent of the sample size or statistical significance, and they can directly affect the context of the results of the model. They also affect the association between variables, and in extreme cases, they can even reflect the opposite of a true association or correlation. Based on reviewing multiple works in data science, the most commonly observed bias is class imbalance due to covariate shifts. Class imbalance is represented by the un-equal ratio of categories which can occur due to changes in the distribution of data (covariate shifts). Class imbalance depends on four factors: 1) degree of class imbalance 2) the complexity of the concept represented by the data 3) the overall size of the training size and 4) the type of classifier. Datasets with imbalance create difficulties in information retrieval, filtering tasks, and knowledge representation – which (if not accounted for) may lead to misinformation in the agricultural domain.

This proposed system is beneficial for short term decisions (during the outlier event), as well as long term goals, such as decisions on crops to farm/grow, farm financial decisions, what exports to deploy tariffs on, what agricultural measures to take during outlier events, forecasting prices, and predicting breaks in the agricultural supply chain (through cognitive and DL/RL methods). Data analytics and intelligent tools for real time decision making can help farmers in deploying snap decisions for improved precision agriculture and creating an overall healthy economy.