A Data-Driven Approach to Mitigate Risk in Global Food Supply Chains

by

Amine Anoun

Submitted to the Sloan School of Management in partial fulfillment of the requirements for the degree of Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Massachusetts Institute of Technology 2017. All rights reserved.

Signature redacted

Author ........................................

Sloan School of Management

May 19, 2017

Certified by .................................

Signature redacted ........

Taufhid Zaman
KDD Career Development Professor in Communications and Technology, Assistant Professor of Operations Management

Signature redacted

Certified by ............... ..................

Retsef Levi
J. Spencer Standish (1945) Professor of Management, Professor of Operations Management

Signature redacted

Accepted by ............... ..........................

Dimitris Bertsimas
Boeing Professor of Operations Research, Co-Director, Operations Research Center
A Data-Driven Approach to Mitigate Risk in Global Food Supply Chains

by

Amine Anoun

Submitted to the Sloan School of Management on May 19, 2017, in partial fulfillment of the requirements for the degree of Master of Science in Operations Research

Abstract

Economically motivated adulteration of imported food poses a serious threat to public health, and has contributed to several poisoning incidents in the past few years in the U.S. [1]. Prevention is achieved by sampling food shipments coming to the U.S. However, the sampling resources are limited: all shipments are electronically sampled [2], but only a small percentage of shipments are physically inspected. In an effort to mitigate risk in shipping supply chains, we develop a data-driven approach to identify risky shippers and manufacturers exporting food to the U.S., as well as U.S. based consignees and importers receiving imported products. We focus our analysis on honey and shrimp, two products that are routinely imported and frequently adulterated.

We obtain over 62,000 bills of lading of honey entering the U.S. between 2006 and 2015 from public sources, and over a million shipment records of shrimp entering the U.S. between 2007 and 2015 from the Food and Drugs Administration (FDA). We analyze these data to identify common patterns between high risk shippers, manufacturers, U.S. consignees and importers, and use them to determine structural features of shipping supply chains that correlate with risk of adulteration. In our analysis of shrimp manufacturers, we distinguish two types of adulteration: intentional (driven by economic motivation) and unintentional (due to negligence or poor sanitary conditions). We use a Bayesian approach to model both the sampling or inspection procedure of the FDA, and the risk of adulteration. Our model is able to predict which companies are at risk of committing adulteration with high out-of-sample accuracy. We find that both geographical features (e.g., travel route, country of origin and transnational paths) and network features (e.g., number of partners, weight dispersion and diversity of the product portfolio) are significant and predictive of suspicious behavior. These outcomes can inform various decisions faced by the FDA in their sampling policy for honey and shrimp shipments, and their site inspection policy for consignees and importers. This work can also extend to other commodities with similar mechanisms, and provides a general framework to better...
detect food safety failures and mitigate risk in food supply chains.

Thesis Supervisor: Tauhid Zaman
Title: KDD Career Development Professor in Communications and Technology, Assistant Professor of Operations Management

Thesis Supervisor: Retsef Levi
Title: J. Spencer Standish (1945) Professor of Management, Professor of Operations Management
Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisors, Prof. Tauhid Zaman and Prof. Retsef Levi for their help, guidance and support. Their expertise has contributed greatly to this thesis, and I would not have come this far without their constant encouragement. Tauhid and Retsef have been exceptional role models for me. Working with them during my years at MIT has helped me grow both professionally and personally.

I also feel very fortunate to have worked closely with Prof. Vivek Farias as a teaching assistant for the 15.778 course. He is a truly amazing teacher and inspiring researcher.

I would also like to thank Prof. Dimitris Bertsimas for his guidance, availability and support. I am also very grateful to Laura Rose and Andrew Carvalho for their administrative assistance.

I also want to acknowledge the sponsors of my research work, the Food and Drugs Administration (FDA). This project has been a tremendous learning experience and an enjoyable journey.

Many thanks go to my friends at the ORC who have made my years at MIT unforgettable, and made the ORC a home-like atmosphere. I'm so fortunate to have met you and I look forward to many years of friendship. I am also grateful to my friends outside of the ORC and back home for their constant support and encouragement.

Finally, I owe my deepest gratitude to my parents and sister for their love and support. I owe them everything, and I dedicate this work to them.
# Contents

1 Introduction

1.1 Background and Motivation ........................................... 21
1.2 Thesis Outline ........................................................... 23
  1.2.1 Data ................................................................. 24
  1.2.2 Supply Chain Features ............................................ 25
  1.2.3 Model ............................................................... 26
  1.2.4 Estimation Results ................................................ 27

2 Data

2.1 Introduction ............................................................. 29
2.2 Supply Chain Data ....................................................... 29
  2.2.1 Import Genius ....................................................... 30
  2.2.2 Shrimp Supply Chain Data ........................................ 31
  2.2.3 Honey Supply Chain Data ......................................... 35
2.3 Labeling Data .............................................................. 36
  2.3.1 Shrimp Labeling Data ............................................. 36
  2.3.2 Honey Labeling Data ............................................... 39

3 Shrimp Consignee and Importer Inspection Models

3.1 Introduction ............................................................. 41
3.2 Shrimp Data .............................................................. 43
  3.2.1 Consignee Product Data .......................................... 44
  3.2.2 Importers Products Data ......................................... 45
List of Figures

1-1 Increase in imports of FDA-regulated goods in the U.S. between 2006 and 2016. https://www.fda.gov/ForIndustry/ImportProgram/ ... 21

1-2 Shipping supply chain diagram. ........................................... 24

2-1 Shipping supply chain and data sources used in the analysis. ....... 30

2-2 Fraction of FDA shrimp shipment records entering the U.S. by air, road or sea between 2007 and 2015. ................................. 33

2-3 Histogram of air entries in the FDA shrimp data between 2007 and 2015. 33

2-4 Histogram of road entries in the FDA shrimp data between 2007 and 2015. ................................................................. 34

2-5 Histogram of sea entries in the FDA shrimp data between 2007 and 2015. ................................................................. 34

2-6 A honey bill of lading from Import Genius. ............................. 36

3-1 U.S. consignees with shrimp entries between 2007 and 2015. The radius of the orange bubbles (created using a Gaussian kernel) is proportional to the number of consignees within that area. ................................. 43

3-2 Pie chart of the most common products identified in the 1.4 million shipments of shrimp consignees, using Import Genius data between 2007 and 2015. ................................................................. 45

3-3 Number of entries to the U.S. in each product category between 2006 and 2015, from shrimp importers. ........................................... 46
3-4 Subgraph of 10 products of shrimp consignees found in the FDA database. A node indicates a product and an edge weight indicates the number of unique consignees that received both products at the end points, between 2007 and 2015. .......................... 52

3-5 Illustration of steps to compute the product diversity feature among consignees. ........................................... 54

3-6 Subgraph of 10 products of shrimp importers found in the FDA database. A node indicates a product and a edge weight indicates the number of unique importers that received both products at the end points, between 2007 and 2015. .......................... 55

3-7 Graphical model of the Bayesian model for predicting inspections and inspection outcomes among shrimp consignees and importers. . . . . 58

3-8 Significant features for predicting inspections in the shrimp consignee inspection models. The 90% posterior credibility interval is given between parentheses. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant. ........................................... 59

3-9 Significant features for predicting violative inspections in the shrimp consignee inspection models. The 90% posterior credibility interval is given between parentheses. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant. ........................................... 60

3-10 Significant features for predicting inspections in the shrimp importers inspection models. The 90% posterior credibility interval is given between parentheses. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant. ........................................... 62
Significant features for predicting violative inspections in the shrimp importers inspection models. The 90% posterior credibility interval is given between parentheses. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant.

ROC for the shrimp consignees inspection model, predicting inspections between 2012 and 2015. The AUC is shown in each plot.

Boxplot for shrimp consignees inspection model AUC across 10 random data splits for predicting inspections.

ROC for the shrimp consignees inspection model, predicting inspection outcomes between 2012 and 2015. The AUC is shown in each plot.

Boxplot for shrimp consignees inspection model AUC across 10 random data splits for predicting inspection outcomes.

Inspection DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 model trials.

Inspection outcome DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 model trials.

ROC for the shrimp importers inspection model, predicting inspections, 2012-2015. The AUC is shown in each plot.

Boxplot for shrimp importers inspection model AUC across 10 random data splits for predicting inspections.

ROC for the shrimp importers inspection model, predicting inspection outcomes, 2012-2015. The AUC is shown in each plot.

Boxplot for shrimp importers inspection model AUC across 10 random data splits for predicting inspection outcomes.
3-22 Inspection DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 model trials. . .

3-23 Inspection outcome DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 model trials. .

4-1 Top ten shrimp exporting countries (in terms of number of shipments) between 2007 and 2015. .......................... 79

4-2 Number and fraction of sampled shipments and refused shipments per country, for the top ten shrimp exporting countries (in terms of number of shipments) between 2007 and 2015. .......................... 79

4-3 FDA refusals between 2007 and 2015 due per refusal category in South America and Asia. .......................... 80

4-4 Number of refused shrimp shipments per refusal category and per year, between 2007 and 2015. .......................... 81

4-5 Fraction of refusal categories out of refused shrimp shipments per year, between 2007 and 2015. .......................... 81

4-6 Illustration of dispersion of total shipment weight per consignee. Variables w represent the weight fraction received by each consignee from the manufacturer. .......................... 84

4-7 Graphical model of the Bayesian model for predicting sampling and adulteration among shrimp manufacturers. .......................... 87

4-8 Significant features for predicting sampling in the shrimp manufacturer sampling model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant. .......................... 88
4-9 Significant features for intentional adulteration in the shrimp manufacturer sampling model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant. 89

4-10 Significant features for unintentional adulteration in the shrimp manufacturer sampling model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant. 91

4-11 Sampling DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 random data partitions. 93

4-12 Intentional Refusal DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 random data partitions. 94

4-13 Unintentional Refusal DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 random data partitions. 95

5-1 Number of honey refusals per refusal code description between 2006 and 2015. 103

5-2 Pie chart of the number of honey shipments per country (total number of shipments is 62,000) between 2006 and 2015, using Import Genius data. 104
Pie chart of the total imported weight of honey per country (total weight is 4.7 billion kilograms) between 2006 and 2015, using Import Genius data. ................................................................. 105
Illustration of the overall graph, high risk graph and low risk graph. 106
High risk graph of honey shipments. ................................................. 106
Comparison of the distributions of the number of consignees per high risk shipper and unknown shippers. .............................................. 107
Comparison of the distributions of the average weight per high risk shipper and unknown shippers. ................................................ 108
Fraction of duplicate shipments among the 50 shippers with highest number of shipments (Import Genius 2006-2015). ......................... 111
Graph of shippers used in duplicate shipments. An edge indicates that the two shippers names have been used in at least two duplicate shipments. ................................................................. 111
Graph of consignees used in duplicate shipments. An edge indicates that the two consignees names have been used for consignees in at least two identical shipments. .............................................. 113
Graphical model of the Bayesian model for predicting adulteration among honey shippers. ................................................................. 117
Model parameters' statistics for sampling in the honey shipper risk model. ....................................................................................... 118
Model parameters' statistics for adulteration in the honey shipper risk model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant. ........................................................................... 119
ROC for honey shippers risk model using one random data split. The Area Under the Curve is 0.84. ....................................................... 120
Boxplot for honey shippers risk model AUC across 10 random data splits. 120
DIC shift from the full model (including all features) when we remove one feature at a time, for predicting honey adulteration. The error bars represent the standard error over 10 random data partitions. The null model does not include any model feature. 

Boxplot of the average number of shipments and average number of consignees for honey shippers on FDA refusals, before and after being put on a refusal.

Boxplot of the average weight and average duplicate fraction for honey shippers on FDA refusals, before and after being put on a refusal.

Density of in-sample predicted risk scores of honey shippers (red line) showing three peaks. The blue line represents the estimated distribution using a Gaussian mixture model.

Model parameters’ for predicting sampling in the Chinese manufacturers risk model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.

Model parameters’ for predicting intentional adulteration in the Chinese manufacturers risk model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.

Model parameters’ for predicting unintentional adulteration in the Chinese manufacturers risk model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Posterior median and 90% credibility interval of the correlation coefficient $\rho$ for each year in the consignees inspection model.</td>
<td>61</td>
</tr>
<tr>
<td>3.2</td>
<td>Posterior median and 90% credibility interval of the correlation coefficient $\rho$ for each year in the importers inspection model.</td>
<td>64</td>
</tr>
<tr>
<td>3.3</td>
<td>Legend table for Figure 3-16 with feature names in the x-axis.</td>
<td>68</td>
</tr>
<tr>
<td>3.4</td>
<td>Legend table for Figure 3-17 with feature names in the x-axis.</td>
<td>69</td>
</tr>
<tr>
<td>3.5</td>
<td>Legend table for Figure 3-22 with feature names in the x-axis.</td>
<td>73</td>
</tr>
<tr>
<td>3.6</td>
<td>Legend table for Figure 3-23 with feature names in the x-axis.</td>
<td>74</td>
</tr>
<tr>
<td>4.1</td>
<td>Posterior median and 90% credibility interval of the correlation coefficient $\rho$ for each year.</td>
<td>90</td>
</tr>
<tr>
<td>4.2</td>
<td>Posterior median and 90% credibility interval of the correlation coefficient $\rho$ for each year.</td>
<td>91</td>
</tr>
<tr>
<td>4.3</td>
<td>Out of sample Spearman rank correlation for predicting sampling and refusals.</td>
<td>92</td>
</tr>
<tr>
<td>4.4</td>
<td>Legend table for Figure 4-11 with feature names in the x-axis.</td>
<td>94</td>
</tr>
<tr>
<td>4.5</td>
<td>Legend table for Figure 4-12 with feature names in the x-axis.</td>
<td>95</td>
</tr>
<tr>
<td>4.6</td>
<td>Legend table for Figure 4-13 with feature names in the x-axis.</td>
<td>96</td>
</tr>
<tr>
<td>4.7</td>
<td>Number of violative shipments caught under the optimial policy compared to the FDA policy, using different definitions of the risk score.</td>
<td>99</td>
</tr>
<tr>
<td>5.1</td>
<td>Five most common refusal categories of honey, based on FDA public refusals between 2006 and 2015.</td>
<td>104</td>
</tr>
</tbody>
</table>
5.2 Statistics of the number of consignees per shipper (shipper’s degree) in the overall graph, the high-risk graph and the low-risk graph.

5.3 Statistics of the average weight (in Kg) per shipper in the overall graph, the high-risk graph and the low-risk graph.

5.4 Statistical tests p-values to compares features’ distributions of the 30 shippers with shipments before and after FDA refusals, showing that we fail to reject the null hypothesis that the means are equal.

5.5 Statistics of shippers’ features in the three K-means clusters.

5.6 Statistics of shippers’ fractions of shipment from each country of origin that is significant in the model estimation, in the three K-means clusters.

B.1 Violation codes and descriptions in honey refusals between 2006 and 2015. www.fda.gov/ForIndustry/ImportProgram/ImportRefusals/ucm144864.htm

Chapter 1

Introduction

1.1 Background and Motivation

It is estimated that 15% of all food products consumed in the U.S. are imported [3]. In some product categories, such as seafood, up to 90% is imported [4]. In fact, the number of imported entries in the U.S. has steadily increased every year since 2006 and reached close to 40 million unique shipments in 2016 (see Figure 1-1 [5]).

![Figure 1-1: Increase in imports of FDA-regulated goods in the U.S. between 2006 and 2016.](https://www.fda.gov/ForIndustry/ImportProgram/)
The globalization of food supply chains exposes them to a wide range of food safety risks stemming from different types of food adulteration [6]. It is estimated that 48 million people get sick and 3,000 die each year from food-borne diseases globally [6]. Risks can be attributed to different types of adulteration, specifically unintentional and intentional adulteration. Unintentional adulteration can be defined as any food adulteration resulting from lack of sufficient measures and processes to maintain the safety of food products. Salmonella contamination and the existence of filth in the food are often caused by negligence or lack of proper equipment facilities and processes, and environmental conditions. It is not necessarily a willful act on the part of the adulterer. Intentional adulteration involves a deliberate act of adding or substituting a substance or mislabeling a product. Intentional adulteration is often economically motivated, in which case it's called economically motivated adulteration (EMA). Intentional adulteration includes the use of illegal animal drugs, food additives, color additives, and fraudulent misbranding.

The risks of food adulteration have become so severe that in response to a wave of incidents of food adulteration, the U.S. government passed the Food Safety Modernization Act (FSMA) [7] in 2011 to strengthen the food safety system. The FSMA allows the Food and Drug Administration (FDA) to regulate the way food is grown, harvested, processed and transported to and within the U.S. Under the FSMA, the focus of the FDA shifted from reactive mitigation strategies to preventative food safety measures. Food facilities are now required to implement a written Hazard Analysis and Risk-based Preventive Controls (HARPC) plan that evaluates the hazards that could affect food safety, specifies the preventive steps as well as the actions the facility will take to correct problems that arise [8].

The FSMA also establishes far more stringent standards with respect to shipment sampling and site inspection frequencies based on risk. For example, all high-risk domestic facilities must be inspected within five years of enactment and no less than every three years after [8].

There has been over 30 million food shipments each year to the U.S. which are subject to FDA regulation since 2014, but the FDA only has resources to sample less
than 1% of them [9]. Since prevention is achieved through sampling shipments at the port of entry and site inspections of local and foreign companies by the FDA, the high imbalance between the rapid growth of regulated food imports and the marginal increase in inspection and port sampling capabilities make it essential to develop systematic approaches to identify risky products and allocate the scarce shipment sampling and site inspection resources in the most effective way.

The FDA also implemented a new screening system for imports called PREDICT (Predictive Risk-based Evaluation for Dynamic Import Compliance Targeting) [10], which assists entry reviewers in targeting higher-risk shipments for examination. PREDICT scores each entry line on the basis of risk factors and surveillance requirements using results of field exams, facility inspections and sample analyses of previous entries [10]. Risk includes factors, such as the inherent health risk of the product, compliance risk associated with firms, facility inspection results, and broker history, among others, and is then compared to all other entry lines within a specified commodity over the past 30 days [11].

Particular emphasis was given to the development of science-based approaches to prepare and protect food supply chains at specific vulnerable points. With the massive amount of global shipment data available, there is great potential to develop such approaches using data-driven statistical methods. In fact, several attempts have been made to understand and reduce risk in the global food network [12], but none have mapped the supply chain structure to risk.

1.2 Thesis Outline

The goal of this thesis is to develop systematic and data-driven approaches to understand how the structure and dynamics of shipping supply chains are correlated with food safety related risks, and to predict risky companies involved at different stages of the shipping supply chain. The approach is based on the hypothesis that firms involved in adulteration, particularly EMA, have similar supply chain dynamics that are distinguishable from firms that are not.
We choose to study two of the most heavily imported and adulterated commodities; honey and shrimp. Specifically, we develop predictive risk models of food adulteration based on supply chain features for honey shippers, shrimp manufacturers, and shrimp consignees and importers.

1.2.1 Data

We describe here the shipping supply chain of shrimp and honey as well as the main data sources used in our analysis.

After the product is produced in a farm, it is processed by a manufacturer. The product is then handed by the manufacturer to the shipper at the place of receipt. After that, the product is carried to the foreign port, ready to be shipped to the U.S. The product is typically transported in containers on board of a vessel. It enters the U.S. through the U.S. port (or port of arrival), and is then distributed to the consignee.

The shipping supply chain for sea shipments can be summarized in Figure 1-2.

![Shipping supply chain diagram](image)

Figure 1-2: Shipping supply chain diagram.

For the analysis of worldwide shippers of honey, we collect 62,000 bills of lading from a public database called Import Genius (https://www.importgenius.com/),
between 2006 and 2015. In order to classify honey shippers safety levels, we use FDA import alerts and refusals [13, 14] for the same period.

For the analysis of U.S. shrimp consignees, importers and manufacturers, we obtain over a million shipment records between 2007 and 2015 from the FDA, and focus on the top seven shrimp exporting Asian countries (in terms of number of shipments), which amount to about 62% of the total worldwide shipments into the U.S. The reason we restrict the analysis to Asian countries is discussed in Chapter 3. We use FDA site inspection data [15] to classify consignees' and importers' safety levels, and we use FDA shipment records with sampling and refusal data to classify manufacturers' safety levels.

1.2.2 Supply Chain Features

From these data, we construct a variety of supply chain features that describe the geography and the complexity of the supply chain, and the shipping patterns of the different companies involved.

Geographical features include the traveled distance before the product is shipped to the U.S., an important measure for perishable products like shrimp. Another important geographical feature we consider in our analysis is called "transnational", which describes instances in which two ports or locations in the shipping supply chain are located in different countries. This feature is important because crossing a border to ship a product could be an attempt to disguise the origin of the product if the country of origin is flagged by the FDA as hazardous.

Other features are related to the supply chain network. We hypothesize that a complex supply chain structure can be an attempt to hide suspicious behavior, and can result in bad product quality. Therefore, we compute the number of U.S. consignees each shipper or manufacturer works with, and extract the variation of this number across years, the variation of the number of shipments across years, as well as the shipped weight to the consignees and its dispersion among them. We consider similar features for consignees and importers where we compute the number and variation of shippers and manufacturers.
In the consignees and importers inspection models, we also account for the diversity of their product portfolios. We obtain data on all products received by shrimp consignees and importers, both from Import Genius and internal FDA data, and we construct a distance metric between those products. We aim to show whether shipping unrelated commodities can be used as a proxy for high risk behavior among consignees and importers.

1.2.3 Model

In our risk models, we assign a risk score between zero and one to each shipper, manufacturer, consignee or importer. We model this risk score as a probit regression of the companies’ supply chain features. However, since there is sampling bias (some companies have never been sampled or inspected in the past), we explicitly model the FDA sampling and inspection policies using a Bayesian framework. Shipments are sampled by the FDA according to a sampling probability, which may be a function of the companies’ features and country of origin. For consignees and importers, the framework is similar. However, instead of predicting the fraction of sampled shipments and the fraction of refused shipments, we predict a binary outcome: is the company inspected at least once? And in this scenario, is the inspection outcome violative?

In order to correct for the sampling bias due to the FDA sampling and inspection policies of shrimp, we use the Heckman Selection Model [16] to solve the model in two steps. First, we predict the likelihood of sampling or inspection, and second, we predict the likelihood of a violative response among sampled or inspected companies. Details of this method are provided in the model estimation sections of Chapters 2, 3 and 4.

We use a Metropolis Hastings within Gibbs MCMC algorithm [17] to estimate model parameters and assess the companies’ risk scores. A higher risk score leads to a higher number of refused shipments for shippers and manufacturers, and a higher chance of violative inspections for consignees and importers. The Bayesian inference method provides us with significance of the different features using the posterior
credibility interval of estimated model parameters.

1.2.4 Estimation Results

Estimating the models using the honey and shrimp data yields several important insights. We find that companies with a more dispersed and complex supply chain are more likely to be high risk. Yearly changes in the supply chain structure also seem to correlate with high risk behavior. In addition, we find that high product diversity among consignees and importers correlates with more risk of failing FDA site inspections, conducted to determine companies' compliance with regulations. This finding shows that shipping unrelated commodities can indicate a higher risk of food adulteration.

Since our goal is to predict risk in new shipments, it is important to perform out of sample validation of the predicted risk scores. We use different validation methods and metrics such as Receiver Operating Characteristics (ROC) curves for models with binary outcomes, out-of-sample Spearman rank correlation tests, and likelihood tests using the Deviance Information Criterion (DIC) [18]. We find that all models exhibit good predictive power.

In summary, we are able to construct quantitative measures of which structural features and characteristics of shipping food supply chains are correlated with increased levels of safety risks in imported food. Our predictive models allow us to map shipment activity to risk at different points of the supply chain. The same framework can be applied to other commodities, in an attempt to catch adulteration and prevent food safety hazards in the U.S.

This thesis is organized as follows:

• Chapter 2: Data

• Chapter 3: Shrimp Consignees and Importers Inspection Models

• Chapter 4: Shrimp Manufacturers Risk Model
• Chapter 5: Honey Shippers Risk Model

• Chapter 6: Conclusion
Chapter 2

Data

2.1 Introduction

In this chapter, we present the general supply chain data that we use throughout this thesis for the shrimp consignees and importers models, the shrimp manufacturers model, and for the honey shippers model. We describe the supply chain data we use, including public sources as well as private data obtained from the FDA. We also describe the data that we use to label companies as high or low risk.

2.2 Supply Chain Data

In this section, we describe supply chain data for honey and shrimp. The supply chain data starts at the farm where the product is produced. The product is then processed by the manufacturer, and handed to the shipper that carries it from the place of receipt to the foreign port. The shipper then ships the product to the U.S. Finally, the product is delivered to the U.S. consignee.

Figure 2-1 summarizes the shipping supply chain and the data sources used in our analysis. These data sources are detailed in the following subsections.
2.2.1 Import Genius

We use a public data source called Import Genius (https://www.importgenius.com/) to obtain bills of lading for shrimp and honey shipments. A bill of lading is a legal document between the shipper and the carrier detailing the type, quantity and destination of the goods being carried. The bill of lading also serves as a receipt of the shipment.

Import Genius contains over 100 million bills of lading for sea shipments to different countries including the U.S. Each bill of lading provides the following detailed supply chain information [19]:

- **Product description**: The description of the product. It can include specifications about the product and information about the packaging.

- **Consignee**: The firm that takes final delivery of the merchandise. In our data, the consignee is located in the U.S.

- **Consignee address**: The address of the consignee in the U.S.

- **Shipper**: The company that ships the product. It could also be the manufacturer or a logistic company.

- **Shipper address**: The address of the shipper.
• Arrival date: The day, month and year of arrival of the shipment to the U.S.

• Gross weight (kg): The weight in kilograms of the shipment.

• Foreign port: The last overseas port in which the shipment is laden on board a vessel for transportation to the U.S.

• U.S. port: The U.S. port in which the shipment is discharged.

• Country of origin: The country in which an article being shipped is grown, produced, or manufactured.

• Place of receipt: The location where the shipper takes the shipment from the manufacturer.

• Bill of lading: A unique ID that identifies the bill of lading.

• Container number: The ID number of the container in which the product is being shipped.

• Voyage number: A number which is assigned to a round-trip sea voyage intended to transport trade goods along an existing trade route. These numbers are assigned by the ship owner or agent and are unique to both the ship and the trip.

• Carrier code: Four digits code referring to the party of the contract of carriage who has undertaken to perform the carriage.

2.2.2 Shrimp Supply Chain Data

For shrimp consignees, importers and manufacturers, we processed over 251,000 shrimp bills of lading that entered the U.S. from all countries between 2007 and 2015. These bills of lading are obtained by querying the words "shrimp", "vannamei" and "camaron" and their misspellings in the product description of the shipments. We then clean the data set by removing shipments that contain the following strings: "shrimp cakes", "shrimp balls", "shrimp crackers", "noodles", "chips", "chips",
For a more thorough cleaning, we go through 50,000 shipments where we manually mark shipments as relevant or not. This process was made faster by automatically marking shipments containing "black tiger", "headless shrimp", "white shrimp" or "vannamei shrimp" in the product description as relevant. We then split the data into a training set and a testing set of equal sizes. We use a Bag of Words approach to build a Classification and Regression Tree (CART) model that predicts if a shipment is relevant or not. The model has a 92\% out-of-sample accuracy on the 50,000 shipments that were manually marked. We then run the model on the entire data set to identify relevant shipments based on the product description text field. At the end of this process, we obtain 218,000 bills of lading from Import Genius.

We also clean companies' names in the Import Genius database using an open source tool called OpenRefine [20]. OpenRefine is a software for cleaning and transforming messy data, it allows to cluster company names based on their spellings or pronunciations.

In addition, we obtain shrimp shipments data between 2007 and 2015 from the FDA. The FDA data for shrimp has over a million shipments arriving to the U.S. through air, road or sea. We show the distribution of the three entry routes in Figure 2-2.
Figure 2-2: Fraction of FDA shrimp shipment records entering the U.S. by air, road or sea between 2007 and 2015.

Our analysis shows that manufacturers use almost exclusively one type entry route; either air, road or sea, as seen in Figures 2-3, 2-4 and 2-5.

Figure 2-3: Histogram of air entries in the FDA shrimp data between 2007 and 2015.
In order to leverage Import Genius data (which only has sea shipments), and given the high fraction of shrimp sea entries to the U.S., we restrict our analysis to sea shipments, which account for 852,000 bills of lading from all countries.

It is important to note that the FDA data has additional entries (such as the manufacturer name and address) that are not available in Import Genius, but at the same time misses some entries (such as the place of receipt and the foreign port) that are found in Import Genius. Therefore, we had to match shipments between the two data sets.
We consider two shipments from the two data sets to be identical if one of the following criteria is fulfilled:

- Same container number and date of arrival (we allow a 5 days window)
- Same bill of lading code and date of arrival (we allow a 5 days window)
- Same voyage number, carrier code and date of arrival (we allow a 5 days window)

The disproportionate number of shipment records between the two databases is due to the fact that Import Genius contains many master bills of lading which group shipments of different containers of shrimp, while the FDA data breaks them down to house bills of lading of single shrimp containers.

After matching FDA shipments to Import Genius bills of lading, we query shipment dates and container numbers of unmatched FDA shipments in Import Genius to expand the Import Genius database. We were able to identify many bills of ladings that did not include the words 'shrimp', 'vannamei' or 'camaron' in the product description, but a more generic word such as 'food', 'seafood' or 'foodstuff'. At the end of this process, we are able to match 91% of FDA shipments to Import Genius. The final matched data set has approximately 775,000 shipments, 1,916 unique manufacturers, 2,549 unique consignees, and 1,755 unique importers.

2.2.3 Honey Supply Chain Data

For the analysis of worldwide shippers of honey, we collect bills of lading from Import Genius between 2006 and 2015.

The honey database is obtained by querying the word "honey" in the product description field of the bills of lading in Import Genius. We then remove irrelevant data from the Import Genius database (such as bills of lading where the word "honey" is used for "honey dew", or to describe a color or a flavor of a different food product). We also correct shippers and consignees' names mispelling by clustering similar names using OpenRefine [20]. After processing the data, we are left with approximately 62,000 bills of lading.
2.3 Labeling Data

2.3.1 Shrimp Labeling Data

To label shrimp manufacturers as high risk, we use information about sampled and refused shipments available through the FDA shipment data. We identify sampled shipments using the 'Intermediate Activity Number'. If this variable has an entry in the data, it means that the shipment was sampled. We identify refused shipments using the 'Final Admissibility Activity Num'. Refused shipments are defined as instances where the "Final Admissibility Activity Number" is either "154: Refused Inform FDA Before Export" or "155: Refused Inform FDA After Export".

In addition to identifying refused shipments, we need to determine the reason for the refusals to distinguish intentional and unintentional adulteration. Therefore, we need to match FDA refused shipments with the FDA public refusal records, available in the following page: http://www.accessdata.fda.gov/scripts/importrefusals/.

The FDA public refusal data set is a list of refused shipments with the manufacturer name, the manufacturer country, the product code, the date of the refusal, and
the violation code. The translation of the different violation codes can be found in the link above. The descriptions for the most common refusals in shrimp are summarized in Appendix B.

We will focus on the two types of adulteration defined previously (intentional and unintentional) and run a separate model on each type.

To label consignees and importers as high risk, we use site inspection data. In particular, we use two inspection data sources for FDA site inspections of consignees and importers:

- **Public inspection database** [13]: List of firm site inspections from October 2008 to February 2016. It has over 83,000 inspections performed by the Center for Food Safety and Applied Nutrition (CFSAN) on manufacturers, consignees and importers. The public inspection database specifies the date of the inspection, the name of the inspected company and its address, as well as the project area and the outcome of the inspection. The project area gives an indication of the violation being tested for, such as 'Food-borne Biological Hazards' or 'Pesticides and Chemical Contaminants'. This information is used to understand the reason for failing inspections. The outcome of each inspection is either NAI (No Action Indicated), VAI (Voluntary Action Indicated) or OAI (Official Action Indicated). We label a consignee as failing an inspection if it resulted in VAI or OAI. These data include inspections for shrimp as well as other products, but do not indicate the product that was tested for.

- **FDA inspection database**: This database was provided by the FDA, and contains 5,933 inspections for consignees and importers between October 2005 and July 2016. The FDA inspection database provides information about the inspection date, the name of the company, as well as the product code and description, which allows us to restrict the analysis to companies being inspected specifically for shrimp. The FDA inspection database also has a list of Product Assignment Codes (PAC) for each inspection, but only one inspection outcome (NAI, VAI, OAI or RTS: Referred To State) that represents the "worst" outcome for all
the listed PAC codes. Product Assignment Codes provide a detailed reason for the type of inspection. The first 2 digits of the PAC code can be linked to the inspection project area.

We proceed to match the two inspection databases. The goal of the matching process is to restrict the analysis to inspections conducted specifically for shrimp, as well as to determine the inspection outcome corresponding to each Product Assignment Code in the FDA inspection database (and not just the worst outcome).

Our matching process is as follows:

- Filter the FDA inspection database on product description containing the strings "shrimp" or "mixed seafood" or "fish".

- Take the subset of inspections in the public inspection database where the date of the inspection, the name of the company and the city of the company match with an inspection in the filtered FDA database. It is worth mentioning that the number of unique PAC codes (the unique first two digits of PAC codes) in the FDA inspection database is the same as the number of unique project areas in the public inspection database, for inspections where \{date, name, city\} match.

- Add to the obtained data set any inspections that were unmatched from the FDA inspection database, with either a unique PAC code or multiple PAC codes with the same first two digits (for which we can determine the project area and the outcome for this specific project area), or inspections with different PAC codes but NAI as the inspection outcome (since we know that if NAI is the "worst" outcome for all these PAC codes, then it is the outcome for each one of the matched project areas.)

At the end of the matching process, we obtain 497 inspection records for shrimp (inspections are separated by project areas.)

Of the 2,549 consignees in the FDA shrimp database, we find that 158 have been inspected at least once, and 93 failed at least one inspection between 2007 and 2015. Of the 1,755 importers in the FDA shrimp database, we find that 188 have been
inspected at least once, and 130 had at least one failed inspection between 2007 and 2015.

2.3.2 Honey Labeling Data

In order to label honey shippers as high risk, we use FDA alert and refusal records, defined below:

- **Alerts**: violations of FDA regulation resulting in the detention of the product. Those violations appear on a 'red list alert'.

  - Red lists include firms, products or countries that are subject to Detention without Physical Examination (DWPE) [21]. The FDA may detain without physical examination certain products offered for entry from a specified country or area, a specific manufacturer, a shipper, or an entire product from all countries outside of the U.S. Some of the reasons a product or firm may be subject to DWPE are [22]:
    * FDA has sampled a product and it tested violative for a pathogen
    * FDA has sampled a product and it contains illegal colors or food additives
    * The product contains pesticides that are not allowed or do not meet tolerance levels
    * The firm has not provided sufficient evidence to support adding them to the green list
    * The product is an unapproved new drug
    * The foreign firm had a violative inspection by FDA

- Green list alerts are lists of firms, products and/or countries that have met criteria for exemption from Detention without Physical Examination (DWPE) [22]. In order for the FDA to consider removing a product and/or firm from DWPE and add it to the green list, the FDA must have evidence that the conditions that gave rise to the apparent violation have been
resolved and that gives FDA confidence future entries will be in compliance with laws and regulations. Note that there are no FDA green list alerts for honey.

- Refusals: failure to meet FDA standards resulting in refusal of admission of the product or part of the product.

These data are obtained from public records on the FDA website [13, 14]. They are used as a proxy to classify high risk shippers. Specifically, shippers that appear on at least one red alert or on at least one refusal are classified as high risk.
Chapter 3

Shrimp Consignee and Importer Inspection Models

3.1 Introduction

Many food poisoning incidents in the U.S. have been linked to imported seafood from Asian countries, such as Bangladesh and Malaysia [23]. Under the FSMA, consignees and importers are required to perform certain risk-based activities to verify that seafood imported into the U.S. has been produced in a manner that meets applicable U.S. safety standards. For this purpose, the FDA conducts site inspections of regulated facilities to assess compliance with regulations and protect public health. The only currently active program by the FDA is for shrimp [24].

For each regulatory action, the FDA provides a prioritized list of high potential risk seafood products based on the health hazard they may pose to consumers. The risk status is determined by a combination of the severity of the hazard, its likelihood of occurrence, and previous industry compliance data.

The products with a high priority for field examinations are [24]:

- Refrigerated seafood products packed in oxygen limiting packaging or reduced oxygen packaged (ROP). Seafood in such packages has increased likelihood for the growth and toxin formation by C. botulinum.
- Raw (fresh and fresh frozen) molluscan shellfish from un-certified shippers. The major concern is the harvest from unapproved waters, which may result in the presence of pathogens, marine toxins or heavy metals.

- Ready-to-eat fish or fishery products using any of the following processes: cooking or pasteurization process (e.g., cooked shrimp, crabmeat, cooked lobster, cooked crayfish, pasteurized crabmeat, surimi-based analogs, etc.) hot or cold smoking process.

The major concerns are inadequate cooking, pasteurization or smoking processes and insanitary post-processing conditions that allow the re-introduction of post-process pathogens, including Listeria monocytogenes and salmonella, that will not be destroyed by any subsequent cooking step prior to consumption by the consumer.

The FDA recommends performing 90% of the number of planned inspections at importers that have been identified as problematic importers, and have an average of 100 or more entry line items per year. The remaining 10% of the number of planned inspections should be made at importers that import less than 100 entry line items per year [24].

In this chapter, we use this site inspection data to assess the risk of failing at least one inspection for shrimp consignees and importers.

In addition to supply chain features, we introduce a new feature called "product diversity", based on a distance metric we build between different products. We hypothesize that consignees or importers that receive shipments of unrelated products are more likely to fail FDA site inspections. It is more suspicious to receive unrelated products, and this could also indicate a poor regulatory structure in place for the different products among non-specialized companies.

In this chapter, we present a framework for identifying risk at the consignee and importer levels, using global shrimp data. First, we present data on the consignees and importers in the FDA shrimp database for shipments, site inspections and received products. Then, we present the consignees' and importers' inspection model frameworks and the models estimation results. Finally, we validate the models using

42
ROC curves and the Deviance Information Criterion [18].

### 3.2 Shrimp Data

The consignee is the party in shipment documents to whose order a consignment will be delivered at the port of destination. The consignee is the buyer or the owner of the consignment. The importer is the entity responsible for ensuring the imported goods comply with local laws and regulations, filing a completed duty entry and associated documents and paying the assessed import duties and other taxes on those goods.

As mentioned in Chapter 2, we obtain companies’ data from the shrimp shipments records obtained from the FDA, for all global sea shipments of shrimp entering the U.S. between 2007 and 2015. These data contain 2,549 unique consignees, and 1,755 unique importers. Consignees and importers overlap in 64% of sea shipments in the FDA shrimp data.

All shrimp consignees in the FDA data set are located in the U.S. We represent in Figure 3-1 a map of these consignees.

![Figure 3-1: U.S. consignees with shrimp entries between 2007 and 2015. The radius of the orange bubbles (created using a Gaussian kernel) is proportional to the number of consignees within that area.](image)
3.2.1 Consignee Product Data

As mentioned previously, we will use the diversity of the consignees’ product portfolios as a predictive feature of failing an inspection. In order to assess product diversity among shrimp consignees, we need data about all products received by those consignees between 2007 and 2015. We obtain the list of all shipments of consignees appearing in the FDA shrimp database (global sea shipments between 2007 and 2015) from Import Genius. We query all consignee names for all shipments they received (including all products and from all countries) between 2007 and 2015. This query results in approximately 2 million shipments.

The biggest challenge is to identify product categories based on the description provided in each bill of lading in Import Genius only. The product description in Import Genius is a text description of the product, that sometimes contains a Harmonized Item Description and Coding System (HS). The HS code is an international standard maintained by the World Customs Organization (WCO) that classifies traded products. Items are identified by a 6-digit number that is recognized by countries that have adopted the harmonized system.

In order to identify products where the HS code is not available, we look for the most frequent strings in all product descriptions across all consignees. We then go through the most common strings manually and build a list of the 100 most relevant and frequent products (for example: shrimp, rice, pants, wood, plastic, soy.) Then, we identify product categories if the product description contained one or more of the 100 product categories. At the end of this process, 1.4 million product descriptions were linked to at least one product category, amounting for 71% of the entire data set. Most of the bills of lading that were left were vague or did not list a specific product, they were treated as a different category: other.

The pie chart in Figure 3-2 shows the most common products found in the 1.4 million identified shipments. The main products are shrimp, clothes, furniture, rice and wood.
Figure 3-2: Pie chart of the most common products identified in the 1.4 million shipments of shrimp consignees, using Import Genius data between 2007 and 2015.

3.2.2 Importers Products Data

We obtained the list of all products that shrimp importers had with respect to FDA regulated products from the FDA, in order to build the product diversity feature for the importers. In the consignee inspection model, we used Import Genius data. The importers' data is not available in Import Genius, but the FDA data for importers' products gives the exact product code, making product identification an easier task.

The list of products provided by the FDA has over 7 million entries for 69 unique products identified by industry code (e.g., 16 for shrimp, 36 for honey). This database has 1,696 unique importers. After matching them with the 1,755 importers in the shrimp database, we obtain products data for 1,681 importers (95.8% of all importers in the shrimp database), and we compute the product diversity feature for those matched importers. Details of this computation are explained in Section 3.4.2.
Figure 3-3 shows the number of entries per product category in the FDA products database. The most common products are fish, bakery products and vegetables.

![Bar Chart](image)

Figure 3-3: Number of entries to the U.S. in each product category between 2006 and 2015, from shrimp importers.

### 3.3 Predictive Features

#### 3.3.1 Consignee Inspection Model Features

The goal is to predict whether a consignee will fail an FDA site inspection or not, based upon the following supply chain features. There are multiple networks associated with the shipment supply chain, each of which provides a set of features for the consignees. There is the supply chain network itself which contains information about the physical route of the shipment, as well as the network of supply chain firms (manufacturers, shippers, and consignees). There is also the network of products shipped by the consignees.

We present below the features computed in the consignee inspection model:

1. **Number of manufacturers:** The total number of unique manufacturers the consignee works with. We hypothesize that working with many manufacturers can be an attempt to hide high risk behavior, or can result in bad product quality.
2. Number of shippers: The total number of unique shippers the consignee works with.

3. Manufacturer is also shipper: The fraction of the consignee’s shipments where the manufacturer is also the shipper.

4. Normalized yearly average shipment weight: The average shipment weight normalized by the yearly average across all consignees to account for any trends in shipment weight. Let $w_{csy}$ be the weight of shipment $s$ of consignee $c$ in year $y$. Let $w_y$ be the average of all shipments in year $y$. Let the normalized shipment weight $z_{csy} = \frac{w_{csy}}{w_y}$. Then the normalized yearly average shipment weight for consignee $c$ is the average of all $z_{csy}$ across all shipments and years for consignee $c$.

5. Number of U.S. ports: The number of unique U.S. ports in all consignee’s shipments.

6. Average distance from manufacturer to place of receipt: the average distance between the manufacturer city and the place of receipt, in km, over all consignee’s shipments. The distance is computed using Google Maps API and the Haversine formula, which is:

$$
a = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_1 - \phi_2}{2}\right) + \cos(\phi_1) \cos(\phi_2) \sin^2\left(\frac{\lambda_1 - \lambda_2}{2}\right)}\right)
$$

(3.1)

where $r$ is the radius of earth (in km), $\phi_1$ and $\phi_2$ are the latitudes of the shipper’s address and the place of receipt, and $\lambda_1$ and $\lambda_2$ are the longitudes of the shipper’s address and the place of receipt, obtained using Google Maps API.

7. Average distance from place of receipt to foreign port: The average distance between the place of receipt and the foreign port, in km, over all consignee’s shipments. The distance is computed using Google Maps API and the Haversine formula.
8. **Transnational manufacturer - place of receipt:** The fraction of shipments where the manufacturer city and the place of receipt are located in different countries, over all consignee’s shipments. This feature is computed using Google Maps API.

9. **Transnational place of receipt - foreign port:** The fraction of shipments where the place of receipt and the foreign port are located in different countries, over all consignee’s shipments. This feature is computed using Google Maps API.

10. **Coefficient of variation of number of manufacturers across years:** The coefficient of variation (standard deviation divided by mean) of the number of unique manufacturers the consignee works with each year. This feature captures variability in the yearly manufacturer number for the consignee.

11. **Coefficient of variation of number of shipments across years:** The coefficient of variation (standard deviation divided by mean) of the number of shipments the consignee received each year. This feature captures variability in the yearly number of shipments for the consignee.

12. **Dispersion of total shipment weight per manufacturer:** For each manufacturer of the consignee, we calculate the cumulative shipment weight sent by the manufacturer to the consignee. We then normalize these values by the total weight shipped to the consignee so they represent the fraction of volume from each manufacturer. Let the fraction of manufacturer $i$ be $f_i$ and assume the consignee has $M$ manufacturers. The dispersion of the total shipment weight per manufacturer is defined as $d = - \sum_{i=1}^{M} f_i \log (f_i)$. Dispersion is computed using entropy, similar to the dispersion feature in the shrimp manufacturer model, as described in Section 4.3.

13. **Number of sampled shipments:** The number of sampled shipments of each consignee.

14. **Number of refused shipments:** The number of refused shipments of each consignee.
15. Consignees with total weight below 10% percentile: Binary indicator of whether the total received weight of the consignee is in the bottom 10% percentile of all total weights received by consignees in the data set.

16. Fraction of shipments from China: The fraction of shipments where the manufacturer’s country is China.

17. Product diversity: This feature reflects the diversity in the consignee’s product portfolio. It captures how similar or unrelated the different products received by the consignee are. Details of how we compute this feature are given in Section 3.4.1.

### 3.3.2 Importer Inspection Model Features

Similar to the consignees inspection model, the goal here is to predict whether an importer will fail an FDA site inspection or not, based upon the following supply chain features:

1. **Number of manufacturers**: The total number of unique manufacturers the importer works with.

2. **Number of shippers**: The total number of unique shippers the importer works with.

3. **Manufacturer is also shipper**: The fraction of the importer’s shipments where the manufacturer is also the shipper.

4. **Normalized yearly average shipment weight**: The average shipment weight normalized by the yearly average across all importers to account for any trends in shipment weight. Let \( w_{isy} \) be the weight of shipment \( s \) of importer \( i \) in year \( y \). Let \( w_y \) be the average of all shipments in year \( y \). Let the normalized shipment weight \( z_{isy} = \frac{w_{isy}}{w_y} \). Then the normalized yearly average shipment weight for importer \( i \) is the average of all \( z_{isy} \) across all shipments and years for importer \( i \).
5. **Number of U.S. ports:** The number of unique U.S. ports in all importer’s shipments.

6. **Average distance from manufacturer to place of receipt:** The average distance between the manufacturer city and the place of receipt, in km, over all importer’s shipments. The distance is computed using Google Maps API and the Haversine formula.

7. **Average distance from place of receipt to foreign port:** The average distance between the place of receipt and the foreign port, in km, over all importer’s shipments. The distance is computed using Google Maps API and the Haversine formula.

8. **Transnational manufacturer - place of receipt:** The fraction of shipments where the manufacturer city and the place of receipt are located in different countries, over all importer’s shipments. This feature is computed using Google Maps API.

9. **Transnational place of receipt - foreign port:** The fraction of shipments where the place of receipt and the foreign port are located in different countries, over all importer’s shipments. This feature is computed using Google Maps API.

10. **Coefficient of variation of number of manufacturers across years:** The coefficient of variation (standard deviation divided by mean) of the number of unique manufacturers the importer works with each year. This feature captures variability in the yearly manufacturer number for the importer.

11. **Coefficient of variation of number of shipments across years:** The coefficient of variation (standard deviation divided by mean) of the number of shipments the importer received each year. This feature captures variability in the yearly number of shipments for the importer.

12. **Dispersion of total shipment weight per manufacturer:** For each manufacturer of the importer, we calculate the cumulative shipment weight sent by the man-
ufacturer to the importer. We then normalize these values by the total weight shipped to the importer so they represent the fraction of volume from each manufacturer. [25] Let the fraction of manufacturer $i$ be $f_i$ and assume the importer has $M$ manufacturers. The dispersion of the total shipment weight per manufacturer is defined as $d = - \sum_{i=1}^{M} f_i \log(f_i)$. Dispersion is computed using entropy, similar to the dispersion feature in the shrimp manufacturer model, as described in Section 4.3.

13. **Number of sampled shipments**: The number of sampled shipments of each importer.

14. **Number of refused shipments**: The number of refused shipments of each importer.

15. **Importers with total weight below 10% percentile**: Binary indicator of whether the total received weight of the importer is in the bottom 10% percentile of all total weights received by importers in the data set.

16. **Fraction of shipments from China**: The fraction of shipments where the manufacturer’s country is China.

17. **Product diversity**: This feature reflects the diversity in the importer’s product portfolio. It captures how similar or unrelated the different products received by the importer are. Details of how we compute this feature are detailed in Section 3.4.2.

### 3.4 Product Diversity Feature

In the consignee and importer inspection models, we use a feature that reflects the diversity of the product portfolio of these companies. We hypothesize that consignees or importers that receive shipments of unrelated products are more likely to fail FDA site inspections. This could be the case because more specialized consignees or importers are more likely to have a structure in place to ensure the safety and
compliance of the product they receive. Importing unrelated products can result in negligence and failure to comply with the regulatory instruction of the FDA.

In order to define such a feature, we need to build a distance metric between products. The main assumption here is that products that are being rarely shipped to the same companies have a higher distance between them. We expect more companies to receive both shrimp and squid for example, than to receive shrimp and chemicals.

### 3.4.1 Consignee Product Diversity

We build a product diversity graph, where nodes represent product categories received by shrimp consignees (100 nodes in this case, based on Import Genius data), and where an edge exists between two products if at least one consignee received both products. The weight of the edge is the number of consignees that imported both products based on the Import Genius database of sea shipments between 2007 and 2015. We represent in Figure 3-4 the subgraph of the product diversity graph showing the 10 most common product categories, and the number of consignees receiving them. In this graph, the weight on each edge is the number of consignees receiving both products at the endpoints of the edge.

![Figure 3-4: Subgraph of 10 products of shrimp consignees found in the FDA database. A node indicates a product and an edge weight indicates the number of unique consignees that received both products at the endpoints, between 2007 and 2015.](image-url)
After we build the entire product diversity graph, we are able to compute a value for the product diversity feature for each consignee. For this purpose, we use a graph metric called modularity [26], that measures the strength of division of a graph into "modules" or clusters. A high modularity of a subgraph reflects dense connections between the nodes within that subgraph, but sparse connections with nodes in different clusters. However, since we want to capture how diverse the product portfolio of each consignee is, we take the product diversity feature as the negative modularity of the consignee products' subgraph within the product diversity graph.

Let \( m_c \) be the modularity for consignee \( c \) (with product subgraph \( g^c = (V^c, E^c) \)) in the product graph \( g = (V, E) \). \( V \) and \( V^c \) are respectively the set of nodes (products) in the graphs \( g \) and \( g^c \), and \( E \) and \( E^c \) are respectively the set of edges (with weights representing the number of consignees importing both products at the endpoint) in the graphs \( g \) and \( g^c \).

The modularity \( m_c \) can be expressed as follows:

\[
    m_c = e_c - a_c^2
\]

where:

\[
    e_c = \text{Weighted fraction of edges in subgraph } g^c = \frac{\text{sum of edge weights between nodes in subgraph } g^c}{\text{sum of all edge weights in the graph}}
\]

And,

\[
    a_c = \text{Weighted fraction of edges with at least one end in subgraph } g^c = \frac{\text{sum of weights of edges with at least one end in subgraph } g^c}{\text{sum of all edge weights in the graph}}
\]

To illustrate how the product diversity is computed for each consignee, we represent in Figure 3-5 a bipartite graph of products-consignees with four products and three consignees. From this graph, we can build the products graph with four nodes representing the four products, and where an edge weight represents the number of consignees receiving both products at the end points of the edge. Consignee A for example receives shrimp and fish, and there are three consignees that receive both...
products. We then compute the modularity of the consignee A subgraph generated by the nodes \{shrimp, fish\} in the products graph.

Products  Consignees

![Diagram showing product diversity feature computation among consignees.]

\[
m_A = \frac{3}{7} - \left( \frac{3 + 1 + 1 + 1}{7} \right)^2 = -0.306
\]

Figure 3-5: Illustration of steps to compute the product diversity feature among consignees.

Recall that the product diversity feature is equal to the negative of modularity. In the illustration, consignee C has the highest product diversity, followed by consignee B and consignee A.

### 3.4.2 Importer Product Diversity

We compute the importers’ product diversity feature similar to the consignees’ product diversity, but instead using data on products provided by the FDA. We make the same assumption that products that are being rarely shipped to the same companies have a higher distance between them.

The FDA importer products data contains 69 unique product categories. We build the product diversity graph where nodes represent product categories (69 nodes in this case), and where an edge exists between two products if at least one importer received both. The weight of the edge is the number of importers that received both products based on the FDA data, between 2007 and 2015.

We show in Figure 3-6 the subgraph of the top 10 most common products in the FDA importers products data.
Figure 3-6: Subgraph of 10 products of shrimp importers found in the FDA database. A node indicates a product and an edge weight indicates the number of unique importers that received both products at the end points, between 2007 and 2015.

The entire graph has dense connections (with higher weights) around common products such as fishery products, but low weight connections around rare products such as animal drugs, dental devices, or cardiovascular devices. To capture this observation, we use modularity to compute importers’ product diversity. For each importer, the product diversity feature is equal to the importer's products subgraph negative modularity (recall that an importer’s subgraph is generated by the list of the importer’s products only). Subgraph modularity is defined as in Section 3.4.1, similar to the consignees’ subgraph modularity.

3.5 Consignees and Importers Inspection Model Framework

In order to estimate the model using past data only, we compute model features up to a certain year, between 2012 and 2015 (for the 2012 model, we compute feature values until December 31st, 2011). We then look at companies that had no inspections prior to that year, and label a consignee or importer as inspected if it had at least one...
inspection in the subsequent years. We label those consignees and importers as high risk if they failed at least one inspection in the subsequent years.

The model framework in the shrimp consignee and importer inspection models is identical for consignees and importers. We associate two probabilities with every consignee or importer. The first is the probability of a company being inspected, which is \( q_i \) for company \( i \). The second is the probability that a company’s site inspection results in a violative outcome (at least one failed inspection), which is \( p_i \) for company \( i \). We assume that inspections are perfect, that is, if a consignee or importer does not comply with FDA regulations, then the inspection always results in a violative outcome (VAI, OAI or RTS). We assume that each company, consignee or importer, has a set of features which we can calculate from its shipment data. For company \( i \), let these features be \( X_i = \{X_{i1}, X_{i2}, ..., X_{ik}\} \). We use a probit model for the inspection and inspection outcome probabilities. We specify the following model:

\[
\log \left( \frac{q_i}{1 - q_i} \right) = \sum_{j=1}^{k} \gamma_j X_{ij} + \epsilon_i 
\]

\[
\log \left( \frac{p_i}{1 - p_i} \right) = \sum_{j=1}^{k} \beta_j X_{ij} + \delta_i 
\]

where \( \beta_j \) is the effect of feature \( j \) on the inspection outcome probability, \( \gamma_j \) is the effect of feature \( j \) on the inspection probability, and \( \epsilon_i \) and \( \delta_i \) are zero mean Gaussian noise terms with standard deviations \( \sigma_\epsilon \) and \( \sigma_\delta \). We fix \( \sigma_\delta \) to one to simplify model estimation.

There may be factors related to the inspection policy of the FDA which we do not measure that impact both probabilities. This could affect our model estimation and may give false conclusions. To account for these unseen factors, we use the Heckman Selection Model, which requires the noise terms to have a correlation coefficient \( \rho \). We write \( \epsilon_i = \sigma_\epsilon \nu_i \), and \( \delta_i = \rho \nu_i + \sqrt{1 - \rho^2} \eta_i \), where \( \nu_i \) and \( \eta_i \) are independent standard normal random variables.

We observe for each consignee and importer whether they have at least one inspection, and if so, if the inspection had a violative outcome.
Let $S_i = 1$ if consignee or importer $i$ has at least one inspection, 0 otherwise. Let $Y_i = 1$ if consignee or importer $i$ has a violative inspection, 0 otherwise. Let $F$ be the cumulative distribution function of the standard normal distribution. Using this notation, we have:

$$P(S_i = 1|X_i) = P(\gamma X_i + \sigma_i \nu > 0|X_i) = F\left(\frac{\gamma X_i}{\sigma_i}\right)$$  \hspace{1cm} (3.5)

Let $S = \{S_1, S_2, \ldots\}$. The inspection log-likelihood is:

$$\log (P(S|X, \gamma, \sigma_i)) = \sum_i S_i \log \left(F\left(\frac{\gamma X_i}{\sigma_i}\right)\right) + (1 - S_i) \log \left(1 - F\left(\frac{\gamma X_i}{\sigma_i}\right)\right)$$  \hspace{1cm} (3.6)

The inspection outcome likelihood, conditioned on being inspected is:

$$P(Y_i = 1|S_i = 1, X_i) = P(\beta X_i + \rho \nu_i + \sqrt{1 - \rho^2} \eta_i > 0|\nu_i > -\frac{\gamma X_i}{\sigma_i}, X_i)$$

$$= \int_{-\frac{\gamma X_i}{\sigma_i}}^{\infty} F\left(\frac{\beta X_i + \rho \nu}{\sqrt{1 - \rho^2}}\right) f(\nu) d\nu$$

where $f$ is the standard normal probability density function.

Let $Y = \{Y_1, Y_2, \ldots\}$ The inspection outcome log-likelihood is then:

$$\log (P(Y|X, \gamma, \sigma_i, \beta, \rho)) = \sum_i Y_i \log \left(\int_{-\frac{\gamma X_i}{\sigma_i}}^{\infty} F\left(\frac{\beta X_i + \rho \nu}{\sqrt{1 - \rho^2}}\right) f(\nu) d\nu\right)$$

$$+ \sum_i (1 - Y_i) \log \left(1 - \int_{-\frac{\gamma X_i}{\sigma_i}}^{\infty} F\left(\frac{\beta X_i + \rho \nu}{\sqrt{1 - \rho^2}}\right) f(\nu) d\nu\right)$$  \hspace{1cm} (3.8)

The joint log-likelihood can be written as:

$$\log (P(S, Y|X, \gamma, \sigma_i, \beta, \rho)) = \sum_i S_i Y_i \log(\phi(\gamma X_i, \beta X_i, \rho))$$

$$+ (1 - S_i) Y_i \log(\phi(-\gamma X_i, \beta X_i, -\rho))$$

$$+ S_i(1 - Y_i) \log(\phi(\gamma X_i, -\beta X_i, -\rho)) + (1 - S_i)(1 - Y_i) \log(\phi(-\gamma X_i, -\beta X_i, \rho))$$  \hspace{1cm} (3.9)

where $\phi$ is the cumulative distribution function of the bivariate normal distribution.

We take a Bayesian approach to estimate model parameters. We place a flat prior
distribution on each model parameter and we calculate the posterior distribution given the data (the features). Priors on $\beta$ and $\gamma$ are normally distributed, centered in 0 with standard deviation 100 so the prior distribution is uninformative. The prior on $\sigma_{\epsilon}$ is uniform between 0 and 10, and the prior on $\rho$ is uniform between -1 and 1.

We present in Figure 3-7 the graphical model for the shrimp consignees and importers inspection models:

![Graphical model of the Bayesian model for predicting inspections and inspection outcomes among shrimp consignees and importers.](image)

Figure 3-7: Graphical model of the Bayesian model for predicting inspections and inspection outcomes among shrimp consignees and importers.

Using the same notation, the posterior distribution of the model parameters is:

$$P(\beta, \gamma, \sigma_{\epsilon}, \rho|X, S, Y) = \frac{P(S, Y|X, \beta, \gamma, \sigma_{\epsilon}, \rho|X)P(\beta, \gamma, \sigma_{\epsilon}, \rho|X)}{P(S, Y|X)}$$

We sample from the posterior distribution using a Metropolis Hastings within Gibbs sampler [17] (see Appendix A).

**3.6 Model Estimation and Discussion**

We present in this section results of the in-sample model estimation for the consignees and importers inspection models for predicting inspections and inspection outcomes.
3.6.1 Consignee Inspection Model Estimation

The tables in Figures 3-8 and 3-9 show the significant features for predicting inspections and inspection outcomes among shrimp consignees (in-sample). We consider a feature to be significant if the 90% posterior credibility interval does not contain 0. Features that were not significant in any of the models for four years are excluded from the tables.

We train the model on data up to each year separately (between 2012 and 2015) to use past features only to predict future inspections and violative inspections. Features are computed up to each year (December 31, 2011 for the 2012 model for example.)

In the first step of the model, we predict if a company had at least one inspection. Therefore, we train the model on all companies that did not have inspections up to a certain year, and predict whether they were inspected in the following year or after.

In the second step of the model, we predict whether those inspected companies failed at least one of their inspections.

![Figure 3-8](image)

Figure 3-8: Significant features for predicting inspections in the shrimp consignee inspection models. The 90% posterior credibility interval is given between parentheses. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.

We find that the number of manufacturers is significant for predicting inspections among shrimp consignees, consistently in the four years. This suggests that the FDA tends to inspect consignees with larger networks more.

Consignees with a high fraction of shipments where the manufacturer and the
shipper are the same are also more likely to be inspected, in all four models.

We find that consignees with a high variation of number of shipments across years are also more likely to be inspected, in three out of the four models.

We also find that the number of unique U.S. ports of a consignee is significant and negatively correlated with the likelihood of inspection in three out of the four models, which could be explained due to FDA inspections being focused around specific locations.

We find that consignees that are sampled more frequently are also more likely to be inspected by the FDA, in all four models.

Product diversity, however, is significant but negatively correlated with the likelihood of inspection. The FDA seems to focus inspections on more "specialized" consignees who have a less diverse product portfolio. This is supported by public data.

The table in Figure 3-9 shows model estimation results for predicting inspection outcomes (whether a company fails a site inspection or not) among consignees that have at least one inspection.

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of manufacturers</th>
<th>Variation of number of manufacturers across years</th>
<th>Variation of number of shipments across years</th>
<th>Number of U.S. ports</th>
<th>Number of refused shipments</th>
<th>Fraction of shipments from China</th>
<th>Product diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.55 (0.12,0.95)</td>
<td>0.43 (-0.11,0.87)</td>
<td>0.06 (-1.05,0.33)</td>
<td>-0.09 (-1.06,0.63)</td>
<td>1.42 (0.92,1.88)</td>
<td>-0.51 (-0.95,0.1)</td>
<td>1.14 (0.67,1.5)</td>
</tr>
<tr>
<td>2013</td>
<td>0.12 (-0.14,0.33)</td>
<td>0.43 (0.05,0.99)</td>
<td>-0.58 (-1.60,0.21)</td>
<td>0.50 (-1.11,0.20)</td>
<td>1.88 (1.23,2.43)</td>
<td>-0.71 (-1.42,0.14)</td>
<td>1.26 (0.76,1.9)</td>
</tr>
<tr>
<td>2014</td>
<td>0.61 (0.22,0.98)</td>
<td>0.11 (-0.45,0.99)</td>
<td>-0.06 (-0.44,0.68)</td>
<td>0.72 (-1.43,0.23)</td>
<td>2.01 (1.20,2.73)</td>
<td>-0.25 (-0.98,0.46)</td>
<td>0.88 (-0.11,1.6)</td>
</tr>
<tr>
<td>2015</td>
<td>0.34 (0.11,0.78)</td>
<td>0.24 (-0.61,1.02)</td>
<td>-0.59 (-0.95,0.13)</td>
<td>1.07 (-1.7,0.66)</td>
<td>2.22 (1.47,2.85)</td>
<td>-0.28 (-0.95,0.13)</td>
<td>0.76 (0.54,1.1)</td>
</tr>
</tbody>
</table>

Figure 3-9: Significant features for predicting violative inspections in the shrimp consignee inspection models. The 90% posterior credibility interval is given between parentheses. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant.

We find that inspected consignees that have a higher number of manufacturers are more likely to fail FDA site inspections in three models out of four.
Interestingly, consignees with a higher variation of the number of shipments across years are less likely to fail inspections in the 2012 and 2014 models, according to our model estimation results. Note that these consignees are more likely to be inspected, but among the inspected ones, less consignees fail inspections.

We find that consignees using more U.S. ports are less likely to fail FDA site inspections in three models out of four.

We also find that consignees with a higher fraction of refused shipments are more likely to fail FDA site inspections (in all four models), and those with a higher fraction of Chinese shipments are less likely to fail inspections (in three models out of four).

Finally, product diversity is significant and positively correlated with the risk of failed inspections in 2012, 2013 and 2015 models. This confirms our initial hypothesis that receiving unrelated products could indicate suspicious behavior among shrimp consignees.

It is interesting to note that consignees with a higher product diversity are less likely to be inspected, but when inspected, are more likely to have a violative outcome. This indicates that the FDA could benefit from sampling more consignees with diverse product portfolios.

Table 3-1 shows the estimated median and 90% credibility interval of the correlation coefficient $\rho$ between the noise terms in the two regression steps (inspection and inspection outcome) in the Heckman Selection Model, in the consignees inspection model.

<table>
<thead>
<tr>
<th>Year</th>
<th>Median correlation coefficient ($\rho$)</th>
<th>90% credibility interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.53</td>
<td>[0.48, 0.57]</td>
</tr>
<tr>
<td>2013</td>
<td>0.56</td>
<td>[0.52, 0.60]</td>
</tr>
<tr>
<td>2014</td>
<td>0.59</td>
<td>[0.55, 0.62]</td>
</tr>
<tr>
<td>2015</td>
<td>0.56</td>
<td>[0.51, 0.61]</td>
</tr>
</tbody>
</table>

Table 3.1: Posterior median and 90% credibility interval of the correlation coefficient $\rho$ for each year in the consignees inspection model.
The positive sign of the correlation coefficient indicates that the FDA is more likely to inspect consignees that have a higher chance of failing site inspections. However, there is still room for improvement as the median correlation coefficient varies between 0.53 and 0.59 in the four years models. A perfect inspection policy would result in a correlation coefficient equal to 1.

### 3.6.2 Importer Inspection Model Estimation

The tables in Figures 3-10 and 3-11 show the significant features for predicting inspections and inspection outcome among shrimp importers. We consider a feature to be significant if the 90% posterior credibility interval does not contain 0. Features that were not significant in any of the models for four years are excluded from the tables.

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of consignees</th>
<th>Number of shippers</th>
<th>Manufacturer is also shipper</th>
<th>Variation of number of shipments</th>
<th>Number of U.S. ports</th>
<th>Number of sampled shipments</th>
<th>Product diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>1.01 (0.64, 1.14)</td>
<td>0.64 (-0.45, 1.22)</td>
<td>1.14 (0.69, 1.75)</td>
<td>0.72 (0.42, 1.13)</td>
<td>-0.77 (-1.21, 0.12)</td>
<td>1.50 (1.08, 1.93)</td>
<td>-0.08 (-0.26, 0.73)</td>
</tr>
<tr>
<td>2013</td>
<td>1.14 (0.59, 1.28)</td>
<td>0.66 (0.13, 1.12)</td>
<td>1.34 (0.34, 1.94)</td>
<td>0.28 (0.09, 0.94)</td>
<td>-0.48 (-0.65, 0.35)</td>
<td>1.12 (0.87, 1.45)</td>
<td>-0.16 (-0.56, -0.09)</td>
</tr>
<tr>
<td>2014</td>
<td>0.87 (0.33, 1.37)</td>
<td>0.24 (-0.17, 1.00)</td>
<td>0.24 (-0.98, 1.10)</td>
<td>0.55 (0.24, 1.10)</td>
<td>-0.94 (-1.39, -0.29)</td>
<td>1.46 (1.01, 1.87)</td>
<td>-0.91 (-0.92, -0.48)</td>
</tr>
<tr>
<td>2015</td>
<td>1.06 (0.62, 1.55)</td>
<td>1.03 (0.45, 1.38)</td>
<td>0.26 (0.06, 1.67)</td>
<td>0.56 (0.16, 1.35)</td>
<td>-0.96 (-1.71, -0.58)</td>
<td>1.21 (0.92, 1.98)</td>
<td>-1.10 (-1.63, -0.67)</td>
</tr>
</tbody>
</table>

Figure 3-10: Significant features for predicting inspections in the shrimp importers inspection models. The 90% posterior credibility interval is given between parentheses. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.

Once again, we find that the number of manufacturers is significant and positively correlated with the likelihood of being inspected, in all four models.

We also find that importers with a high fraction of shipments where the manufacturer is also the shipper, and those with a high variation of number of shipments across years are more likely to be inspected by the FDA.

The number of unique U.S. ports of an importer is significant and negatively correlated with the likelihood of inspection in 2014 and 2015 models.

We also find that importers with a high fraction of sampled shipments are also
more likely to be inspected, in all four models.

Finally, we see that the FDA tends to sample importers with lower product diversity, (i.e., more specialized in shrimp). This is similar to the consignee’s model estimation results.

Note that many significant features overlap between the consignees and importers inspection models estimation. This makes sense given that consignees and importers are the same in 64% of shipments in the FDA global sea shipments database.

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of manufacturers</th>
<th>Variation of number of manufacturers across years</th>
<th>Number of refused shipments</th>
<th>Fraction of Refused shipments from China</th>
<th>Product diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>1.14 (0.48, 1.65)</td>
<td>-0.09 (-0.52, 1.18)</td>
<td>1.25 (0.98, 1.73)</td>
<td>-0.66 (-1.13, -0.16)</td>
<td>1.43 (1.02, 1.89)</td>
</tr>
<tr>
<td>2013</td>
<td>1.24 (0.92, 1.83)</td>
<td>0.14 (-0.57, 0.97)</td>
<td>1.09 (0.45, 1.87)</td>
<td>-0.52 (-1.18, -0.03)</td>
<td>1.27 (0.84, 1.87)</td>
</tr>
<tr>
<td>2014</td>
<td>1.29 (0.82, 1.75)</td>
<td>0.74 (0.08, 1.15)</td>
<td>1.32 (0.62, 1.95)</td>
<td>-0.33 (-1.01, 0.05)</td>
<td>1.82 (1.43, 2.21)</td>
</tr>
<tr>
<td>2015</td>
<td>1.05 (0.68, 1.71)</td>
<td>0.56 (0.14, 1.48)</td>
<td>0.99 (0.26, 1.23)</td>
<td>-0.57 (-0.88, 0.14)</td>
<td>2.03 (1.55, 2.79)</td>
</tr>
</tbody>
</table>

Figure 3-11: Significant features for predicting violative inspections in the shrimp importers inspection models. The 90% posterior credibility interval is given between parentheses. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.

We find that the number of manufacturers is significant and positively correlated with risk of failing a site inspection in all four years. We also find that importers with a high variation of manufacturers across years are more likely to fail inspections in the 2014 and the 2015 models.

Importers with a high fraction of refused shipments are also more likely to fail a site inspection, in all four models. We also find that importers with more shipments from China are less likely to fail inspections in 2012 and 2013 models.

Finally, product diversity is significant and positively correlated with risk of failed inspections. This supports the hypothesis that importers with more diverse product portfolios are more likely to be high risk.

Table 3-2 shows the estimated median and 90% credibility interval of the correlation coefficient ρ between the noise terms in the two regression steps in the Heckman Selection Model in the importers inspection model.
<table>
<thead>
<tr>
<th>Year</th>
<th>Median correlation coefficient ($\rho$)</th>
<th>90% credibility interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.54</td>
<td>[0.51, 0.58]</td>
</tr>
<tr>
<td>2013</td>
<td>0.55</td>
<td>[0.50, 0.60]</td>
</tr>
<tr>
<td>2014</td>
<td>0.58</td>
<td>[0.54, 0.65]</td>
</tr>
<tr>
<td>2015</td>
<td>0.59</td>
<td>[0.54, 0.63]</td>
</tr>
</tbody>
</table>

Table 3.2: Posterior median and 90% credibility interval of the correlation coefficient $\rho$ for each year in the importers inspection model.

Similar to the consignees model, the positive sign of the correlation coefficient indicates that the FDA is more likely to inspect importers that have a higher chance of failing site inspections. However, there is still room for improvement as the median correlation coefficient varies between 0.54 and 0.59 in the four years models.

3.7 Model Validation

We perform an out-of-sample analysis to assess the consignees and importers models’ predictive power. Since we run the models up to each year, and predict if a company had at least one inspection (and in that case, if it has at least one failed inspection), we train the model on half of the companies that did not have inspections up to a certain year and predict whether they were inspected in the following years. Then, we test whether the other half of companies were inspected in the following years, and in that case, whether there was at least one failed inspection.

3.7.1 Consignees Inspection Model Validation

Since the model prediction is binary (inspected at least once or not, at least one failed inspection or no failed inspections), we use ROC curves and the Area Under the Curve (AUC) measure to assess out of sample model performance. Figures 3-12, 3-13, 3-14 and 3-15 show the out of sample ROC curves for the shrimp consignees inspection model and the boxplots of the AUC values for 10 random data partitions.
Figure 3-12: ROC for the shrimp consignees inspection model, predicting inspections between 2012 and 2015. The AUC is shown in each plot.

Figure 3-13: Boxplot for shrimp consignees inspection model AUC across 10 random data splits for predicting inspections.
Figure 3-14: ROC for the shrimp consignees inspection model, predicting inspection outcomes between 2012 and 2015. The AUC is shown in each plot.

Figure 3-15: Boxplot for shrimp consignees inspection model AUC across 10 random data splits for predicting inspection outcomes.
We see that the model always performs better than random (AUC is always above 50%).

We also perform a likelihood test to assess the performance of the full model (using all features) versus the null model (with no features), and to assess the importance of different features by removing one feature at a time. In order to penalize for the number of features used in the prediction, we use the Deviance Information Criterion (DIC) [18]:

For predicting inspections, the DIC is

\[
DIC_s = -2 \mathbb{E}_{\gamma, \sigma_c} [\log(p(S|\gamma, \sigma_c)) + \log(p(S|\mathbb{E}[\gamma], \mathbb{E}\sigma_c))] + C
\]  

(3.11)

For predicting inspection outcomes, the DIC is

\[
DIC_r = -2 \mathbb{E}_{\beta, \rho, \gamma, \sigma_c} [\log(p(Y|\beta, \rho, \gamma, \sigma_c)) + \log(p(Y|\mathbb{E}[\beta], \mathbb{E}[\rho], \mathbb{E}[\gamma], \mathbb{E}\sigma_c))] + C'
\]  

(3.12)

where \(C\) and \(C'\) are constants. A lower DIC indicates that the model fits the data better.

We run 10 random data partitions of the consignees inspection model. In each partition, and for each year, we compute the difference between the full model DIC and the null model DIC (with no features), as well as the difference between the full model DIC and the DIC obtained after removing one feature at a time. We show in Figures 3-16 and 3-17 the average DIC shift from the full model DIC as well as the standard error over the 10 data partitions. A bigger shift means the feature has more predictive power.
Figure 3-16: Inspection DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 model trials.

<table>
<thead>
<tr>
<th>Number</th>
<th>Model Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of manufacturers</td>
</tr>
<tr>
<td>2</td>
<td>Manufacturer is also shipper</td>
</tr>
<tr>
<td>3</td>
<td>Shipments variation</td>
</tr>
<tr>
<td>4</td>
<td>Product diversity</td>
</tr>
<tr>
<td>5</td>
<td>Total weight below 10% percentile</td>
</tr>
<tr>
<td>6</td>
<td>Manufacturers variation</td>
</tr>
<tr>
<td>7</td>
<td>Transnational manufacturer - place of receipt</td>
</tr>
<tr>
<td>8</td>
<td>Weight</td>
</tr>
<tr>
<td>9</td>
<td>Distance manufacturer - place of receipt</td>
</tr>
<tr>
<td>10</td>
<td>Transnational place of receipt - foreign port</td>
</tr>
<tr>
<td>11</td>
<td>Distance place of receipt - foreign port</td>
</tr>
<tr>
<td>12</td>
<td>Manufacturers weight dispersion</td>
</tr>
</tbody>
</table>

Table 3.3: Legend table for Figure 3-16 with feature names in the x-axis.
We find that the number of manufacturers is the most important feature for model fit in both steps of the model, on average across all four years models. We also find that product diversity is the second most important feature in predicting inspection outcomes among shrimp consignees, on average across all four years models.
3.7.2 Importer Inspection Model Validation

Similar to the consignees inspection model, the prediction in the importers inspection model is binary (inspected at least once or not, at least one failed inspection or no failed inspections). Therefore, we can represent ROC curves and use the Area Under the Curve (AUC) to assess out of sample model performance.

Figures 3-18, 3-19, 3-20 and 3-21 show the out of sample ROC curves for the shrimp importers inspection model and the following boxplot of the AUC values for 10 random data partitions.

![ROC curves for shrimp importers inspection model](image)

Figure 3-18: ROC for the shrimp importers inspection model, predicting inspections, 2012-2015. The AUC is shown in each plot.
Figure 3-19: Boxplot for shrimp importers inspection model AUC across 10 random data splits for predicting inspections.

Figure 3-20: ROC for the shrimp importers inspection model, predicting inspection outcomes, 2012-2015. The AUC is shown in each plot.
All models perform better than random (AUC greater than 50%).

We run 10 random data partitions of the importers inspection model. In each partition, and for each year, we compute the difference between the full model DIC and the null model DIC (with no features), as well as the difference between the full model DIC and the DIC obtained after removing one feature at a time. We show in Figures 3-22 and 3-23 the average DIC shift from the full model DIC as well as the standard error over the 10 data partitions.

Figure 3-22: Inspection DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 model trials.
<table>
<thead>
<tr>
<th>Number</th>
<th>Model Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of manufacturers</td>
</tr>
<tr>
<td>2</td>
<td>Manufacturer is also shipper</td>
</tr>
<tr>
<td>3</td>
<td>Product diversity</td>
</tr>
<tr>
<td>4</td>
<td>Number of U.S. ports</td>
</tr>
<tr>
<td>5</td>
<td>Distance place of receipt - foreign port</td>
</tr>
<tr>
<td>6</td>
<td>Number of shippers</td>
</tr>
<tr>
<td>7</td>
<td>Transnational manufacturer - place of receipt</td>
</tr>
<tr>
<td>8</td>
<td>Shipments variation</td>
</tr>
<tr>
<td>9</td>
<td>Total weight below 10% percentile</td>
</tr>
<tr>
<td>10</td>
<td>Distance manufacturer - place of receipt</td>
</tr>
<tr>
<td>11</td>
<td>Transnational place of receipt - foreign port</td>
</tr>
<tr>
<td>12</td>
<td>Number of refused shipments</td>
</tr>
</tbody>
</table>

Table 3.5: Legend table for Figure 3-22 with feature names in the x-axis.

Figure 3-23: Inspection outcome DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 model trials.
We find that the number of manufacturers is, similar to the consignees inspection model, the most important feature for model fit in both steps of the model, on average across all four years models. The second most important feature is the fraction of shipments where the manufacturer and the shipper are the same, followed by product diversity.

We see that product diversity stands out as an important feature, not only in predicting the risk of failing inspections, but also in model fit.

### 3.8 Summary

In this chapter, we showed how shipping supply chain features of consignees and importers can be predictive of violative site inspections among these companies. In addition, we built a distance metric between products and introduced a product diversity feature to account for unrelated products that companies may receive. This feature was significant and positively correlated of risk of failing inspections in both
our models, for consignees and importers.

The findings can inform the FDA inspection policy by allocating resources to inspect consignees and importers with a high predicted risk score, in an attempt to enforce compliance with regulations and protect public health.
Chapter 4

Shrimp Manufacturer Sampling Model

4.1 Shrimp Adulteration

In this chapter, we focus on shrimp manufacturers and build a predictive model to estimate their risk of having refused shipments, due to either intentional or unintentional adulteration (as defined previously). We start by providing background on shrimp adulteration. Next, we present the manufacturer sampling model framework and the model estimation results. Finally, we validate the model using different approaches.

The risk of shrimp adulteration is not solely related to U.S. consignees and importers, it can also be traced to foreign manufacturers. In fact, over 90% of the raw shrimp available in the U.S. is farmed in Asian and South American countries, including Thailand, Ecuador, Vietnam, India and Indonesia, and a high fraction of adulterated shrimp in the U.S. seems to come from these countries [27].

Because of the crowded and polluted conditions that typically exist in fish-farming ponds or tanks, the shrimp are often given antibiotics. The use of some antibiotics in shrimp farms is not approved by the FDA and it is illegal to import shrimp raised with those antibiotics into the U.S. [28].

Shrimp can also grow in dangerous conditions that cause health hazards to con-
sumers. For instance, shrimp was found growing in toxic waters in China in 2007. Farmers cope with the toxic waters by mixing illegal drugs and pesticides into fish and shrimp in particular, which helps keep their stocks alive [29].

Furthermore, many incidents in the past years have had a negative effect on shrimp production levels, such as the Gulf oil spill in 2010 [30] and the Early Mortality Syndrome (EMS) outbreak in Asia in 2009 and 2013 [31]. These incidents have increased the economic incentives for adulteration of shrimp imported to the U.S.

EMS was first reported in China in 2009, and then spread to Vietnam, Malaysia and Thailand in 2010 and 2011. In particular, the production of farmed shrimp declined drastically in 2013 [32]. While EMS does not affect humans, shrimps can die as early as after 12 hours of exposure to the bacteria, and mortality rate can be 100% within the first 30 days.

There are at least three types of approaches to prevent and fight EMS. The first solution involves the use of antibiotics or pond additives to improve the health conditions of shrimps. For example, there has been evidence about farmers in Vietnam using Oxytetracycline (OTC), a type of antibiotics that can be readily attained with an affordable price. The second solution is polyculture, i.e., co-cultivation of shrimp and other types of fish such as tilapia. This helps to create an ecosystem that makes EMS bacteria less likely to multiply and increase its density to a critical point to initiate infection in shrimp. The third solution is the use of Biofloc Technology (BFT), which typically requires high investment, reliable power supplies and sufficient knowledge. Compared to more traditional methods such as using bleaches to improve water contamination, BFT involves a self-nitrification process within culture ponds, resulting in zero water exchange with microorganisms to control water quality [33].

4.2 Shrimp Manufacturer Data

The pie chart in Figure 4-1 shows the top ten shrimp exporting countries in terms of number of shipments, between 2007 and 2015.
Figure 4-1: Top ten shrimp exporting countries (in terms of number of shipments) between 2007 and 2015.

We restricted the analysis to these ten countries including Ecuador, Thailand, Indonesia, India, Vietnam, China, Malaysia, Peru, Honduras and Bangladesh.

The fractions of sampled and refused shipments vary considerably between different countries. Specifically, they tend to be high for Asian manufacturers, and low for South American manufacturers. Figure 4-2 shows the number and fraction of sampled shipments and refused shipments per country, for the top ten shrimp exporting countries (in terms of number of shipments) between 2007 and 2015.

Figure 4-2: Number and fraction of sampled shipments and refused shipments per country, for the top ten shrimp exporting countries (in terms of number of shipments) between 2007 and 2015.
Shipments coming from South America are also sampled by the FDA at a lower rate than those coming from Asia. The sampling outcomes are also different. In particular, refusals due to intentional adulteration amount for 33% of the refusals in the top seven countries in Asia, but only 6% of the refusals in the top three countries in South America, as shown in Figure 4-3.

![Figure 4-3: FDA refusals between 2007 and 2015 due per refusal category in South America and Asia.](image)

Furthermore, over the years, we observe trends in the shipping volumes between South America and Asia respectively. Specifically, shrimp volumes from Asia decreased from 2009-2015, while South American shrimp imports increased. This is primarily due to the outbreak of the Early Mortality Syndrome (EMS), which destroyed much of the shrimp supply in Asia.

This analysis suggests that the underlying dynamics that result in adulteration may be different for Asian and South American shrimp manufacturers. Therefore, we focus on shipments coming from Asia, specifically, Bangladesh, China, India, Indonesia, Malaysia, Thailand and Vietnam. These countries represent 62% of the total worldwide shrimp shipments into the U.S. between 2007 and 2015, based on FDA data. The data for these countries contain 506,000 sea shipments from 1,159 unique manufacturers in these seven countries. Of these manufacturers, 832 have at least one sampled shipment. The overall shipment sampling rate (number of sampled shipments to total number of shipments) is about 9%.

In addition to the different behavioral patterns between companies, the number
and type of FDA refusals has changed significantly year to year, as shown in Figures 4-4 and 4-5.

![Figure 4-4: Number of refused shrimp shipments per refusal category and per year, between 2007 and 2015.](image)

![Figure 4-5: Fraction of refusal categories out of refused shrimp shipments per year, between 2007 and 2015.](image)

We hypothesize that the yearly variation could be due to the Early Mortality Syndrome outbreaks in 2009 and 2013, since we observe an increase in the number of refusals in the following years: 2010 and 2014 (note that the total number of sampled shipments stayed about the same in this period). We also observe an increase in the
percentage of refusals due to intentional adulteration (animal drugs, food additives, color additives and misbranding) after the 2013 EMS outbreak.

Therefore, we build a year-by-year model where we map supply chain features up to each year between 2012 and 2015 to the risk of intentional and unintentional adulteration.

4.3 Predictive Features

The goal of the model is to predict which manufacturers have violative shipments using supply chain features. These features, detailed below, were used for predicting both intentional and unintentional adulteration.

1. **Number of consignees**: Total number of unique consignees the manufacturer works with.

2. **Number of shippers**: Total number of unique shippers the manufacturer works with.

3. **Manufacturer is also shipper**: Fraction of shipments where the manufacturer is also the shipper.

4. **Normalized yearly average shipment weight**: Average shipment weight normalized by the yearly average across all manufacturers to account for any trends in shipment weight. Let $w_{msy}$ be the weight of shipment $s$ of manufacturer $m$ in year $y$. Let $w_y$ be the average of all shipments in year $y$. Let the normalized shipment weight $z_{msy} = \frac{w_{msy}}{w_y}$. Then the normalized yearly average shipment weight for manufacturer $m$ is the average of all $z_{msy}$ across all shipments and years for manufacturer $m$.

5. **Average distance from manufacturer to place of receipt**: Average distance between the manufacturer city and the place of receipt, in km, over all shipments of the manufacturer. The distance is obtained using reverse geo-coding with Google Maps API and the Haversine formula.
6. **Average distance from place of receipt to foreign port:** Average distance between the place of receipt and the foreign port, in km, over all shipments of the manufacturer. This distance is also obtained using reverse geo-coding with Google Maps API and the Haversine formula.

7. **Transnational manufacturer - place of receipt:** Fraction of shipments where the manufacturer city and the place of receipt are located in different countries, across all shipments of the manufacturer. This feature is obtained using Google Maps API.

8. **Transnational place of receipt - foreign port:** Fraction of shipments where the place of receipt and the foreign port are located in different countries, across all shipments of the manufacturer. This feature is obtained using Google Maps API.

9. **Coefficient of variation of number of consignees across years:** coefficient of variation (standard deviation divided by mean) of the number of unique consignees the manufacturer works with each year. This feature captures variability in the yearly consignee number for the manufacturer.

10. **Coefficient of variation of number of shipments across years:** coefficient of variation (standard deviation divided by mean) of the number of shipments the manufacturer has each year. This feature captures variability in the yearly number of shipments for the manufacturer.

11. **Dispersion of total shipment weight per consignee:** For each consignee of the manufacturer we calculate the cumulative shipment weight sent to the consignee by the manufacturer on a given year. We then normalize these values by the total weight shipped by the manufacturer so they represent the fraction of volume each consignee receives. Let the fraction of consignee $i$ be $f_i$ and assume the manufacturer has $C$ consignees. The dispersion of the total shipment weight per consignee is defined as $d = -\sum_{i=1}^{C} f_i \log(f_i)$.

The dispersion of total shipment weight per consignee is based on the notion of
entropy, borrowed from information theory [25]. A higher value means a more dispersed network of consignees based on the total weights they receive. Figure 4-6 shows an illustration of entropy across consignees.

Figure 4-6: Illustration of dispersion of total shipment weight per consignee. Variables \( w \) represent the weight fraction received by each consignee from the manufacturer.

12. Fraction of shipments to consignees with total weight below 10\% percentile: fraction of shipments going to consignees whose total received weight is in the bottom 10\% percentile of all total weights received by consignees in the data set. This feature aims to capture manufacturers that work with less experienced consignees in shrimp.

We use the manufacturer's country of origin as a feature as well, to account for the disparity in sampling and refusal rates in companies from different Asian countries.

### 4.4 Shrimp Manufacturers Model Framework

In this section, we map shrimp shipping supply chain features to risk of intentional and unintentional adulteration at the manufacturer level, for intentional and unintentional adulteration. As mentioned before, we restrict the analysis to Asian manufacturers that sent shrimp sea shipments to the U.S. between 2007 and 2015, based on data provided by the FDA.
Our model is designed to predict a shrimp’s manufacturers fraction of refused shipments based on their supply chain features. The predicted refusal fraction could be used to identify new high risk manufacturers who have not been sampled before by the FDA.

Because the FDA sampling records identify the cause of the refusal of a shipment, we are able to estimate separate models for different types of adulteration, specifically, intentional and unintentional.

We associate two probabilities with every manufacturer. The first is the probability of a shipment of that manufacturer being sampled, which is $q_i$ for manufacturer $i$. The second is the probability of a shipment being adulterated, which is $p_i$ for manufacturer $i$. The assumption is that each shipment is adulterated independently and sampled independently, but the adulteration and sampling probabilities are the same across shipments of the same manufacturer. Each manufacturer is assumed to have a set of features which can be calculated from its shipment data. For manufacturer $i$, let these features be $X_i = \{X_{i1}, X_{i2}, \ldots, X_{ik}\}$. We use a probit model for the sampling and adulteration probabilities. We specify the following model:

$$\log \left( \frac{q_i}{1 - q_i} \right) = \sum_{j=1}^{k} \gamma_j X_{ij} + \epsilon_i$$  \hspace{1cm} (4.1)$$

$$\log \left( \frac{p_i}{1 - p_i} \right) = \sum_{j=1}^{k} \beta_j X_{ij} + \delta_i$$  \hspace{1cm} (4.2)$$

where $\beta_j$ is the effect of feature $j$ on the adulteration probability, $\gamma_j$ is the effect of feature $j$ on the sampling probability, and $\epsilon_i$ and $\delta_i$ are zero mean Gaussian noise terms with standard deviations $\sigma_\epsilon$ and $\sigma_\delta$. We set $\sigma_\delta = 1$ in order to simplify model estimation.

Since our observations are affected by the FDA sampling policy, there may be factors which we do not measure that impact both probabilities. This could affect our model estimation and may give false conclusions. To account for these unseen factors, we use the Heckman Selection Model, which requires the noise terms to have a correlation coefficient $\rho$. 
We can write \( \epsilon_i = \sigma_i \nu_i \), and \( \delta_i = \rho \nu_i + \sqrt{1 - \rho^2} \eta_i \), where \( \nu_i \) and \( \eta_i \) are independent standard normal random variables.

Let \( S_{ij} = 1 \) if shipment \( j \) of manufacturer \( i \) is sampled and 0 otherwise, and \( Y_{ij} = 1 \) if shipment \( j \) of manufacturer \( i \) is refused and 0 otherwise. Let \( S_i \) be the the fraction of sampled shipments of manufacturer \( i \), \( Y_i \) the fraction of refused shipments of manufacturer \( i \), and \( N_i \) the total number of shipments of manufacturer \( i \). Let \( F \) be the cumulative distribution function of the standard normal distribution. Using these notations, we have:

\[
P(S_{ij} = 1|X_i) = P(\gamma X_i + \sigma_i \nu_i > 0|X_i) = F\left(\frac{\gamma X_i}{\sigma_i}\right)
\]  

(4.3)

\[
P(S_i = s_i|N_i, X_i) = P\left(\sum_{j=1}^{N_i} S_{ij} = s_i N_i|N_i, X_i\right)
\]

\[
= \binom{N_i}{s_i N_i} F\left(\frac{\gamma X_i}{\sigma_i}\right)^{s_i N_i} \left(1 - F\left(\frac{\gamma X_i}{\sigma_i}\right)\right)^{(1-s_i)N_i}
\]

(4.4)

Let \( S = \{S_1, S_2, \ldots\} \). The sampling log-likelihood is then:

\[
\log (P(S|N, X)) = \sum_i S_i N_i \log \left(F\left(\frac{\gamma X_i}{\sigma_i}\right)\right) + (1 - S_i) N_i \log \left(1 - F\left(\frac{\gamma X_i}{\sigma_i}\right)\right) + C
\]

(4.5)

where \( C \) is a constant.

For predicting the fraction of refusals,

\[
P(Y_{ij} = 1|S_{ij} = 1, X_i) = P\left(\beta X_i + \rho \nu_i + \sqrt{1 - \rho^2} \eta_i > 0|\nu_i > -\frac{\gamma X_i}{\sigma_i}, X_i\right)
\]

\[
= \int_{-\frac{\gamma X_i}{\sigma_i}}^{\infty} F\left(\frac{\beta X_i + \rho \nu}{\sqrt{1 - \rho^2}}\right) f(\nu) d\nu
\]

(4.6)

where \( f \) is the standard normal probability density function.
And,

$$P(Y_i = y_i | S_i, N_i, X_i) = P\left( \sum_{j:S_{ij}=1} Y_{ij} = y_i S_i N_i | S_i, N_i, X_i \right)$$

$$= \left( S_i N_i \right) \left( \int_{-\infty}^{\infty} \frac{F(\beta X_i + \rho \nu)}{\sqrt{1-\rho^2}} f(\nu) d\nu \right)^y_i S_i N_i \times \left( 1 - \int_{-\infty}^{\infty} \frac{F(\beta X_i + \rho \nu)}{\sqrt{1-\rho^2}} f(\nu) d\nu \right)^{1-y_i} S_i N_i \quad (4.7)$$

Let \( Y = \{Y_1, Y_2, \ldots \} \). The refusal log-likelihood is then:

$$\log (P(Y|S, N, X)) = \sum_i Y_i S_i N_i \log \left( \int_{-\infty}^{\infty} \frac{F(\beta X_i + \rho \nu)}{\sqrt{1-\rho^2}} f(\nu) d\nu \right)$$

$$+ \sum_i (1 - Y_i) S_i N_i \log \left( 1 - \int_{-\infty}^{\infty} \frac{F(\beta X_i + \rho \nu)}{\sqrt{1-\rho^2}} f(\nu) d\nu \right) + C' \quad (4.8)$$

where \( C' \) is a constant.

In order to estimate model parameters \((\sigma_\epsilon, \rho, \beta, \gamma)\), we take a Bayesian approach. We place a flat prior distribution on each model parameter and we calculate the posterior distribution given the data (the model features). Priors on the \( \beta \) and \( \gamma \) are normally distributed, centered in 0 with a high standard deviation so the prior distribution is uninformative. The prior on \( \sigma_\epsilon \) is uniform between 0 and 10, and the prior on \( \rho \) is uniform between -1 and 1.

We present in Figure 4-7 the graphical model for the shrimp manufacturer model:

![Graphical model of the Bayesian model for predicting sampling and adulteration among shrimp manufacturers.](image-url)

Figure 4-7: Graphical model of the Bayesian model for predicting sampling and adulteration among shrimp manufacturers.
Using the same notations as above, we have:

\[
P(\beta, \gamma, \sigma_e, \rho|X, S, Y) = \frac{P(S, Y|X, \beta, \gamma, \sigma_e, \rho)P(\beta, \gamma, \sigma_e, \rho|X)}{P(S, Y|X)}
\]  

(4.9)

We sample from the posterior distribution using a Metropolis Hastings within Gibbs sampler [17].

### 4.5 Model Estimation and Discussion

In this section, we present results of the in-sample model estimation for the manufacturer sampling model for predicting sampling and refusals. In order to use past data only to predict future outcomes, we train the model on data up to each year between 2012 and 2015. This means that we compute features up to each year, and then compute the fraction of sampled and refused shipments of each manufacturer that had at least one shipment in the subsequent years. We then train the model on the those sampling and refusal fractions to estimate model parameters.

The table in Figure 4-8 shows the significant features for predicting sampling among shrimp manufacturers. We consider a feature to be significant if the 90% posterior credibility interval does not contain 0. Features that were not significant in any of the models for four years are excluded from the table.

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of consignees</th>
<th>Number of shippers</th>
<th>Manufacturer is also shipper</th>
<th>India</th>
<th>China</th>
<th>Bangladesh</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.66(0.31,1.04)</td>
<td>1.21(0.56,1.44)</td>
<td>1.78(1.12,2.29)</td>
<td>1.14(0.55,1.69)</td>
<td>1.85(1.12,2.32)</td>
<td>1.76(1.25,2.35)</td>
</tr>
<tr>
<td>2013</td>
<td>0.73(0.41,1.20)</td>
<td>1.11(0.64,1.71)</td>
<td>1.41(0.90,2.08)</td>
<td>1.06(0.49,1.36)</td>
<td>1.39(0.90,2.08)</td>
<td>1.45(1.16,2.06)</td>
</tr>
<tr>
<td>2014</td>
<td>1.13(0.56,1.69)</td>
<td>1.06(0.49,1.56)</td>
<td>1.41(0.90,2.08)</td>
<td>1.23(0.68,1.77)</td>
<td>1.39(0.90,1.68)</td>
<td>1.39(0.90,1.68)</td>
</tr>
<tr>
<td>2015</td>
<td>0.99(0.55,1.47)</td>
<td>1.23(0.72,1.89)</td>
<td></td>
<td>1.10(0.41,1.64)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4-8: Significant features for predicting sampling in the shrimp manufacturer sampling model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant.

We find that manufacturers with a high number of consignees and shippers are
more likely to be sampled by the FDA, consistently in the four model years. This suggests that the FDA focuses sampling efforts on manufacturers with larger networks.

We also find that manufacturers from China and those with a are more likely to be sampled in all four models. This finding matches the observation that China has the highest sampling fraction among all seven Asian countries in our model, as shown in Figure 4-2.

Manufacturers with a higher fraction of shipments where they are also the shipper were more likely to be sampled up to 2013.

The table in Figure 4-9 shows the significant features for predicting intentional refusals among shrimp manufacturers. We consider a feature to be significant if the 90% posterior credibility interval does not contain 0. Features that were not significant in any of the models for four years are excluded from the table.

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of consignees</th>
<th>Manufacturer is shipper</th>
<th>Fraction of shipments to &quot;unexperienced&quot; consignees</th>
<th>Transnational manufacturer — place of receipt</th>
<th>Malaysia</th>
<th>Thailand</th>
<th>China</th>
<th>Bangladesh</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.96(0.45,1.35)</td>
<td>1.26(0.70,1.83)</td>
<td>1.44(1.01,1.97)</td>
<td>1.02(0.63,1.51)</td>
<td>0.62(0.38,0.87)</td>
<td>0.59(0.45,0.74)</td>
<td>0.96(0.55,1.39)</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>0.83(0.61,1.18)</td>
<td>1.39(0.78,1.95)</td>
<td></td>
<td>1.03(0.81,1.26)</td>
<td>0.63(0.46,0.85)</td>
<td>0.55(0.40,0.72)</td>
<td>0.97(0.54,1.42)</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>0.88(0.59,1.12)</td>
<td>1.33(0.28,2.14)</td>
<td>1.00(0.66,1.51)</td>
<td>1.50(1.22,1.80)</td>
<td>0.91(0.70,1.12)</td>
<td>0.45(0.22,0.67)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>1.13(0.81,1.43)</td>
<td>0.99(0.75,1.44)</td>
<td>0.68(0.34,1.39)</td>
<td>0.94(0.56,1.62)</td>
<td>1.28(0.81,1.81)</td>
<td>0.54(0.40,0.72)</td>
<td>0.71(0.42,1.13)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4-9: Significant features for intentional adulteration in the shrimp manufacturer sampling model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.

We find that manufacturers with more consignees are riskier for intentional refusals in all four years. This suggests that a more complex supply chain could indicate a higher risk of adulteration.

Manufacturers with a high fraction of shipments where they are also the shipper are also more likely to be caught for intentional adulteration, except for the 2013 model.

We also find that manufacturers that work more with "inexperienced" consignees (with total weight below 10% percentile of total received weights) have a higher risk of intentional refusals in the 2014 and the 2015 models.
Manufacturers with higher fractions of transnational shipments between them and the place of receipt are more likely to be refused for intentional adulteration (except for the 2013 model). This could indicate an intentional effort to adulterate shrimp in order to maintain it during a long travel, since shrimp is a perishable product.

We also find that manufacturers from Thailand and China are less likely to be refused for intentional adulteration in three models out of four, while those from Malaysia and Bangladesh are more likely to participate in intentional adulteration. Recall that manufacturers from China have a higher sampling rate compared to the other Asian countries, but among their sampled shipments, intentional refusal rates tend to be lower.

Table 4-1 shows the estimated median and 90% credibility interval of the correlation coefficient \( \rho \) between the noise terms in the two regression steps (sampling and intentional refusals) in the Heckman Selection Model, in the manufacturer risk model.

<table>
<thead>
<tr>
<th>Year</th>
<th>Median correlation coefficient ( (\rho) )</th>
<th>90% credibility interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.63</td>
<td>[0.55, 0.74]</td>
</tr>
<tr>
<td>2013</td>
<td>0.66</td>
<td>[0.58, 0.74]</td>
</tr>
<tr>
<td>2014</td>
<td>0.67</td>
<td>[0.59, 0.75]</td>
</tr>
<tr>
<td>2015</td>
<td>0.69</td>
<td>[0.57, 0.79]</td>
</tr>
</tbody>
</table>

Table 4.1: Posterior median and 90% credibility interval of the correlation coefficient \( \rho \) for each year.

We find that the FDA is more likely to sample manufacturers’ shipments that have a higher chance of being refused for intentional. Note that that posterior median correlation coefficient is higher in the manufacturer model compared to the consignees and importers models, suggesting that the sampling policy is better at targeting manufacturers at a higher risk of refusals.

The table in Figure 4-10 shows the significant features for predicting unintentional refusals (salmonella and filth) among shrimp manufacturers. We consider a feature to be significant if the 90% posterior credibility interval does not contain 0. Features
that were not significant in any of the models for four years are excluded from the table.

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of consignees</th>
<th>Manufacturer is shipper</th>
<th>Fraction of shipments to &quot;unexperienced&quot; consignees</th>
<th>Variation Number of consignees</th>
<th>Variation Number of shipments</th>
<th>China</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.49 (0.26, 0.91)</td>
<td>1.10 (0.70, 1.76)</td>
<td>0.88 (0.41, 1.38)</td>
<td>0.66 (0.21, 1.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>0.66 (0.24, 1.09)</td>
<td></td>
<td>0.54 (0.23, 1.01)</td>
<td></td>
<td>0.33 (0.06, 0.72)</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>0.59 (0.31, 1.17)</td>
<td>0.76 (0.30, 0.93)</td>
<td>0.22 (0.03, 0.88)</td>
<td></td>
<td></td>
<td>0.92 (0.50, 1.27)</td>
</tr>
<tr>
<td>2015</td>
<td>0.59 (0.31, 1.17)</td>
<td></td>
<td></td>
<td>0.92 (0.50, 1.27)</td>
<td>1.11 (0.58, 1.62)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4-10: Significant features for unintentional adulteration in the shrimp manufacturer sampling model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.

We find that manufacturers with a high number of consignees are more likely to have their shipments refused because of unintentional adulteration, in three models out of four.

We also find that Chinese manufacturers are more likely to be refused for unintentional adulteration in three models, in opposition to intentional adulteration. Their shipments are often refused for salmonella contamination.

Table 4-2 shows the estimated median and 90% credibility interval of the correlation coefficient $\rho$ between the noise terms in the two regression steps in the Heckman Selection Model in the manufacturers risk model.

<table>
<thead>
<tr>
<th>Year</th>
<th>Median correlation coefficient ($\rho$)</th>
<th>90% credibility interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.61</td>
<td>[ 0.56, 0.73 ]</td>
</tr>
<tr>
<td>2013</td>
<td>0.62</td>
<td>[ 0.55, 0.75 ]</td>
</tr>
<tr>
<td>2014</td>
<td>0.63</td>
<td>[ 0.55, 0.72 ]</td>
</tr>
<tr>
<td>2015</td>
<td>0.62</td>
<td>[ 0.57, 0.69 ]</td>
</tr>
</tbody>
</table>

Table 4.2: Posterior median and 90% credibility interval of the correlation coefficient $\rho$ for each year.

We find that the FDA is also more likely to sample manufacturers’ shipments that have a higher chance of being refused for unintentional.
4.6 Model Validation

We perform a standard out-of-sample analysis where we partition the data in a stratified way (on the dependent variable). The model is trained on half of the data set and then tested on the other half.

Since we cannot validate the model using ROC (the outcome variable is not binary, it’s a continuous fraction between 0 and 1), we use the Spearman rank correlation test to compare the predicted rank to the actual rank of sampling fractions and risk scores, for each year and both types of adulteration.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson’s correlation assesses linear relationships, Spearman’s correlation assesses monotonic relationships. The sign of the Spearman correlation in this context indicates the direction of association between the actual sampling (or refusal) fractions, and the predicted sampling (or refusal) scores.

The results are presented in Table 4-3:

<table>
<thead>
<tr>
<th>Year</th>
<th>Sampling Rank Correlation</th>
<th>Sampling p value</th>
<th>Intentional Adulteration Rank Correlation</th>
<th>Intentional Adulteration p value</th>
<th>Intentional Rank Correlation</th>
<th>Intentional p value</th>
<th>Unintentional Adulteration Rank Correlation</th>
<th>Unintentional p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.58</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td>0.44</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td>0.47</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>0.49</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td>0.38</td>
<td>$7.73 \times 10^{-13}$</td>
<td>0.52</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>0.43</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td>0.42</td>
<td>$5.05 \times 10^{-16}$</td>
<td>0.49</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>0.59</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td>0.72</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td>0.44</td>
<td>$\leq 2.2 \times 10^{-16}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: Out of sample Spearman rank correlation for predicting sampling and refusals.

The out of sample rank correlation is significantly positive in all models at 5% confidence level.
We also perform a likelihood test to assess the performance of the full model (using all features) versus the null model (with no features), and to assess the importance of different features by removing one feature at a time. We use the Deviance Information Criterion (DIC) [18].

For predicting sampling, the DIC is

\[
DIC_s = -2 \mathbb{E}_{\gamma, \sigma_i} \left[ \log(p(S|\gamma, \sigma_i)) + \log(p(S|\mathbb{E}[\gamma], \mathbb{E}[\sigma_i])) + C \right] \quad (4.10)
\]

For predicting refusals, the DIC is

\[
DIC_r = -2 \mathbb{E}_{\beta, \rho, \gamma, \sigma_i} \left[ \log(p(S|\beta, \rho, \gamma, \sigma_i)) + \log(p(S|\mathbb{E}[\beta], \mathbb{E}[\rho], \mathbb{E}[\gamma], \mathbb{E}[\sigma_i])) + C' \right] \quad (4.11)
\]

where \( C \) and \( C' \) are constants. The expectation is taken over the posterior distribution of the model parameters. A lower DIC indicates that the model fits the data better. The DIC is computed out of sample, which lets us assess the model’s predictive power.

We run 10 random data partitions. In each partition, and for each year, we compute the difference between the full model DIC and the null model DIC (with no features), as well as the difference between the full model DIC and the DIC obtained after removing one feature at a time. We represent in Figure 4-11 the average DIC shift from the full model DIC as well as the standard error over the 10 data partitions, for predicting sampling.

Figure 4-11: Sampling DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 random data partitions.
<table>
<thead>
<tr>
<th>Number</th>
<th>Model Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Countries</td>
</tr>
<tr>
<td>2</td>
<td>Number of consignees</td>
</tr>
<tr>
<td>3</td>
<td>Manufacturer is also shipper</td>
</tr>
<tr>
<td>4</td>
<td>Consignees variation</td>
</tr>
<tr>
<td>5</td>
<td>Shipments variation</td>
</tr>
<tr>
<td>6</td>
<td>Shipments to inexperienced consignees</td>
</tr>
<tr>
<td>7</td>
<td>Transnational manufacturer - place of receipt</td>
</tr>
<tr>
<td>8</td>
<td>Distance place of receipt - foreign port</td>
</tr>
<tr>
<td>9</td>
<td>Distance manufacturer - place of receipt</td>
</tr>
<tr>
<td>10</td>
<td>Consignees weight dispersion</td>
</tr>
<tr>
<td>11</td>
<td>Weight</td>
</tr>
<tr>
<td>12</td>
<td>Transnational place of receipt - foreign port</td>
</tr>
</tbody>
</table>

Table 4.4: Legend table for Figure 4-11 with feature names in the x-axis.

Figure 4-12 shows the average DIC shift from the full model DIC as well as the standard error over the 10 data partitions, for predicting intentional refusals.

![Figure 4-12: Intentional Refusal DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 random data partitions.](image)
<table>
<thead>
<tr>
<th>Number</th>
<th>Model Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Countries</td>
</tr>
<tr>
<td>2</td>
<td>Transnational manufacturer - place of receipt</td>
</tr>
<tr>
<td>3</td>
<td>Consignees Variation</td>
</tr>
<tr>
<td>4</td>
<td>Transnational place of receipt - foreign port</td>
</tr>
<tr>
<td>5</td>
<td>Number of consignees</td>
</tr>
<tr>
<td>6</td>
<td>Distance place of receipt - foreign port</td>
</tr>
<tr>
<td>7</td>
<td>Shipments variation</td>
</tr>
<tr>
<td>8</td>
<td>Shipments to inexperienced consignees</td>
</tr>
<tr>
<td>9</td>
<td>Manufacturer is also shipper</td>
</tr>
<tr>
<td>10</td>
<td>Consignees weight dispersion</td>
</tr>
<tr>
<td>11</td>
<td>Distance manufacturer - place of receipt</td>
</tr>
<tr>
<td>12</td>
<td>Weight</td>
</tr>
</tbody>
</table>

Table 4.5: Legend table for Figure 4-12 with feature names in the x-axis.

Finally, Figure 4-13 shows the average DIC shift from the full model DIC as well as the standard error over the 10 data partitions, for predicting unintentional refusals.

Figure 4-13: Unintentional Refusal DIC of the full model (all features), compared to the null model (no features), and model with different features removed (x axis). Error bars represent the standard error for 10 random data partitions.
<table>
<thead>
<tr>
<th>Number</th>
<th>Model Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Countries</td>
</tr>
<tr>
<td>2</td>
<td>Consignees variation</td>
</tr>
<tr>
<td>3</td>
<td>Number of consignees</td>
</tr>
<tr>
<td>4</td>
<td>Manufacturer is also shipper</td>
</tr>
<tr>
<td>5</td>
<td>Shipments to inexperienced consignees</td>
</tr>
<tr>
<td>6</td>
<td>Consignees weight dispersion</td>
</tr>
<tr>
<td>7</td>
<td>Shipments variation</td>
</tr>
<tr>
<td>8</td>
<td>Distance place of receipt - foreign port</td>
</tr>
<tr>
<td>9</td>
<td>Weight</td>
</tr>
<tr>
<td>10</td>
<td>Transnational manufacturer - place of receipt</td>
</tr>
<tr>
<td>11</td>
<td>Transnational place of receipt - foreign port</td>
</tr>
<tr>
<td>12</td>
<td>Distance manufacturer - place of receipt</td>
</tr>
</tbody>
</table>

Table 4.6: Legend table for Figure 4-13 with feature names in the x-axis.

We see that the model performs the best (lower DIC) when we include all features, for predicting sampling, intentional and unintentional adulteration.

The features that have the highest impact on model fit for intentional adulteration are the manufacturer’s country of origin, the fraction of transnational features between the manufacturer and the place of receipt, and the variation of the number consignees across years. The features that have the highest impact on model fit for unintentional adulteration are the manufacturer’s country of origin, the variation of the number consignees across years and the number of consignees.

### 4.7 Optimization Formulation

In this section, we want to assess the improvement generated by the manufacturer risk model, in terms of targeting manufacturers in the sampling policy. We make the assumption that our model has a similar predictive power on all companies (both sampled and not sampled). This assumes that all companies behave in similar pat-
terns, not only the ones that were sampled by the FDA, for which we have available refusal data.

We build a distributional model of the refusal rates as a function of the risk model scores obtained using our model. We then use the empirical refusal rates to validate the model.

We assume we have $N$ manufacturers who can be sampled over a fixed time horizon. Manufacturer $i$ has $n_i$ total shipments, $s_i$ sampled shipments, and $r_i$ refused shipments within the time horizon. We define the total number of sampled shipments in the horizon as $N_s = \sum_{i=1}^{N} s_i$.

We assume that the sampling is done uniformly at random with each shipment of manufacturer $i$ sampled with probability $q_i$. We also assume each shipment of manufacturer $i$ is violative with probability $p_i$. Using this, for any given sampling policy $q = q_1, q_2, \ldots, q_N$, the average number of shipments refused is $N_b(q) = \sum_{i=1}^{N} n_i q_i p_i$.

Our goal is to select the sampling policy $q$ such that we maximize $N_b(q)$ while keeping the expected number of sampled shipments below some value $N_s$. We will use the value of $N_s$ set by the FDA over the given time horizon. In practice, we do not know $p_i$ for manufacturer $i$. Rather, we know some approximation for $p_i$ based upon our model. We will refer to this estimate as a risk score $r_i$. Our optimization formulation is then

$$\max_{q} \sum_{i=1}^{N} n_i q_i r_i$$

subject to

$$\sum_{i=1}^{N} n_i q_i \leq N_s$$

$$0 \leq q_i \leq 1, 1 \leq i \leq N.$$  

\[4.12\]

\[4.8\]

Evaluating Policy Performance with Historical Data

Given a policy $q$, we would like to evaluate its performance on real data. To do this, we require the actual refusal rate of manufacturer $i$, which is $p_i$. For many companies,
we do not know this value because they were not sampled. For these companies, we need an estimate for the refusal rate. We can build a distribution for the true refusal rate conditioned on the risk score $r_i$. We define this distribution as $f(p_i|r_i)$. We define the mean refusal rate conditioned on $r_i$ as $p_i' = E[p_i|r_i]$. Then for a given policy $q$, the expected number of shipments refused is simply

$$N_b(q) = \sum_{i=1}^{N} n_i q_i p_i'.$$  \hspace{1cm} (4.13)

### 4.8.1 Model for refusal rate conditioned on risk score

We find that a good model for the conditional distribution of the refusal fraction $p_i$ given the risk score $r_i$ is as follows. With probability $\alpha_0$, $p_i = 0$ and with probability $(1 - \alpha_0)$, logit($p_i$) is normal with mean logit($r_i$) and variance $\sigma^2$. We fit the risk score to refusal rates over a four year horizon to get a better estimate for the refusal rate. We estimate the model parameters using maximum likelihood. We find that the ML estimates of the parameters are $\alpha_0 = 0.72$ and $\sigma = 7.80$, meaning that the refusal fraction is 0 with probability 0.72.

### 4.8.2 Optimization Results

We use three month horizons for our analysis. This means every three months we recalculate the risk scores, then recalculate the sampling policy, and finally evaluate the expected number of refused shipments. We find that by using the intentional risk score for $r_i$ we have $N_b(q) = 1459$. Over the same time period, the FDA only refuses 1211 shipments. We beat the FDA by 20%. We can also use a variety of other risk scores in our approach.

- Intentional: estimated risk scores of intentional adulteration
- Unintentional: estimated risk scores of unintentional adulteration
- Max(intentional, unintentional): the maximum of the intentional risk score and the unintentional risk score for each manufacturer
• Intentional quantile: the quantile of the intentional risk score of each manufacturer in all intentional risk scores

• Unintentional quantile: the quantile of the unintentional risk score of each manufacturer in all unintentional risk scores

• Max(intentional quantile, unintentional quantile): the maximum of the quantile of the intentional risk score and the quantile of the unintentional risk score of each manufacturer

The results are summarized in Table 4.7.

<table>
<thead>
<tr>
<th>Risk score</th>
<th>Refused shipments (optimized policy)</th>
<th>Refused shipments (FDA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intentional</td>
<td>1,459</td>
<td>1,211</td>
</tr>
<tr>
<td>Unintentional</td>
<td>1,096</td>
<td>1,211</td>
</tr>
<tr>
<td>Max(intentional, unintentional)</td>
<td>1,379</td>
<td>1,211</td>
</tr>
<tr>
<td>Intentional quantile</td>
<td>1,494</td>
<td>1,211</td>
</tr>
<tr>
<td>Unintentional quantile</td>
<td>1,140</td>
<td>1,211</td>
</tr>
<tr>
<td>Max(Int. quantile, Unint. quantile)</td>
<td>1,372</td>
<td>1,211</td>
</tr>
</tbody>
</table>

Table 4.7: Number of violative shipments caught under the optimial policy compared to the FDA policy, using different definitions of the risk score.

4.9 Summary

We are able to predict the risk of intentional and unintentional adulteration in imported shrimp at the manufacturer level using shipping supply chain features for Asian manufacturers. We find that having a complex supply chain usually indicates a higher risk of adulteration.

The findings can be used to rank shrimp manufacturers based upon their risk scores (or some variation of it) and prioritize shipments from manufacturers with the highest risk levels in the sampling process. This can yield a higher number of refused
shipments than the current policy, providing for a better allocation of the sampling resources.
Chapter 5

Honey Shippers Risk Model

5.1 Introduction and Background

According to the USDA’s Economic Research Service, U.S. honey consumption reached over 410 million pounds in 2010 and has been estimated by Bee Culture Magazine to reach over 486.3 million pounds in 2015 [34]. Honey is produced in every state in the U.S., reaching production of over 82 million pounds of honey in 2013 [35]. However, honey demands in the U.S. far outweigh local honey production, and as a result, honey imports have been increasing rapidly. It is estimated by the National Honey Board that approximately 75% of honey consumed in the U.S. is imported [36].

Prior to 2000, China was a top exporter of honey to the U.S., but as a result of continuous adulteration and quality incidents as well as commercial dumping of surplus honey inventory in under cost prices, Chinese honey is currently banned or heavily taxed in many countries including the U.S. [36]. In particular, in 2001, due to the reported dumping of honey at low cost by Chinese exporters to the U.S., the U.S. government implemented high tariffs on Chinese honey that approximately triple the cost of honey from China.

Honey products are adulterated in various ways. China, in particular, has been a source of many adulteration incidents of honey. Honey contaminated with antibiotics originating in China has been documented back to at least 2006 [37]. Since most honey bees are imported from foreign countries into China for domestication and
honey nectar operations, the morbidity is relatively high. Due to the high morbidity rate, it is very common to use antibiotics to treat the honey bees diseases. Some Chinese beekeepers have been noted to spray their honey bees with antibiotics (e.g., chloramphenicol) to increase the health of the bees and to avoid Foulbrood disease, which is caused by bacteria that is common in bee colonies [38]. Chloramphenicol has been found to be toxic in children, causing DNA damage and carcinogenicity and its presence in foods has been banned by the FDA the US. Furthermore, the honey contaminated with antibiotics is reported to have an acrid or unpleasant taste, which is sometimes masked by the addition of granulated sugar and syrup solutions. The antibiotics in use in China are banned in the U.S. and the EU. When these drugs are used in honey nectar production they can seep into the honey and cause the finished product to become adulterated when exported from China [39].

The imposed U.S. tariffs since the early 2000s have also driven Chinese and other importers to engage in various adulteration activities. In particular, honey was imported to the U.S. from China but the country of origin was disguised by using ‘looping’ (shipping through other countries) and faking the respective shipping documents [40]. This practice has been highly correlated with other types of adulteration of the product, including the use of illegal antibiotics and pesticides at bee farming and the introduction of low quality honey with sugars and syrup substitutes.

Economically motivated adulteration of honey can also occur when pure honey is diluted with either sugar and syrup solutions and/or antibiotic containing honey. The same dilution concept has been used with other compounds such as high fructose syrup, corn syrup or sweeteners in recent years to avoid the rising sugar prices. [37]

In this chapter, we investigate honey adulteration and demonstrate the predictive power of shipping supply chains features to identify honey shippers that engage in high risk behavior. The shipper is the company that ships the product from the foreign country to the U.S.; it could be the manufacturer itself, a distributor or a logistics company. First, we present the data and investigate some observed inconsistencies. Then, we build supply chain predictive features. Finally, we discuss the model framework for predicting high risk behavior among shippers, estimate model
parameters, and use different out of sample validation methods to assess the model performance.

5.2 Honey Data

The analysis for honey shippers is based on detailed mapping of honey shipping supply chains using data obtained from the publicly available database called Import Genius (https://www.importgenius.com/). The obtained honey database contains approximately 6,200 different shippers of honey.

Honey products are adulterated in various ways. The total number of honey refusals between 2006 and 2015 is 365 (involving 145 unique companies). Note however that Import Genius data include shippers and not manufacturers, while FDA alerts and refusals list manufacturers. This means that the identified high risk shippers are ones that are also manufacturers.

Figure 5-1 and Table 5-1 summarize the main reasons for honey refusals between 2006 and 2015 (see Appendix B for a description of the most common refusal codes in honey):

![Number of Honey Refusals per Violation Code, 2006 - 2015](image)

Figure 5-1: Number of honey refusals per refusal code description between 2006 and 2015.
Table 5.1: Five most common refusal categories of honey, based on FDA public refusals between 2006 and 2015.

<table>
<thead>
<tr>
<th>ASC ID</th>
<th>CHRG CODE</th>
<th>Refusal Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>482</td>
<td>NUTRIT LBL-482</td>
<td>Misbranding</td>
</tr>
<tr>
<td>238</td>
<td>UNSAFE ADD-238</td>
<td>Food additives</td>
</tr>
<tr>
<td>265</td>
<td>SUBSTITUTE-265</td>
<td>Food substitutes</td>
</tr>
<tr>
<td>473</td>
<td>LABELING-473</td>
<td>Misbranding</td>
</tr>
<tr>
<td>321</td>
<td>LACKS N/C-321</td>
<td>Misbranding</td>
</tr>
</tbody>
</table>

We present in Figures 5-2 and 5-3 the top ten countries with the highest number of shipments and total honey weight sent by sea to the U.S. between 2006 and 2015. China, Mexico, India are the top exporting countries in terms of honey shipments to the U.S. between 2006 and 2015. Brazil, Guatemala and China have the sent the highest weight of honey to the U.S. in the same time frame.

Figure 5-2: Pie chart of the number of honey shipments per country (total number of shipments is 62,000) between 2006 and 2015, using Import Genius data.
There are four countries that overlap between the top ten countries with highest number of refusals and top ten countries with highest total shipped weight or number of shipments: New Zealand, Mexico, India and Taiwan. New Zealand also happens to have the highest number of honey refusals between 2006 and 2015.

5.3 Explanatory Analysis

In order to identify common patterns between high risk shippers, we build the "overall graph", the "high risk graph" and the "low risk graph" as illustrated in Figure 5-4. In these bipartite graphs, the nodes represent consignees or shippers, and the edge between a consignee and a shipper means that they appear together in at least one bill of lading.
Figure 5-4: Illustration of the overall graph, high risk graph and low risk graph.

- **Overall graph:** The graph that connects all manufacturers and consignees that work together.

- **High risk graph:** The graph of manufacturers on FDA alert or refusal, connected to the consignees they work with.

- **Low risk graph:** The complementary graph of the high risk graph. This is the graph obtained when we remove all high risk shippers and all the consignees they work with from the overall graph.

The high risk graph for honey is represented in Figure 5-5.

Figure 5-5: High risk graph of honey shipments.

We run a preliminary analysis where we compare features of companies in the three different graphs. These features are listed below:
1. Degree: The degree of the shipper in the shipper-consignee bipartite graph. Its the number of consignees the shipper works with. We believe that a company working with a high number of consignees in the U.S. may be trying to avoid detection with more complex supply chain, and that a complex supply chain can result in bad product quality. We have found that shippers that are on current FDA alerts or refusals (shippers in the high risk graph) tend to have a larger number of U.S. consignees in their supply chains, as shown in table 5-2:

<table>
<thead>
<tr>
<th>Graph</th>
<th>Min</th>
<th>10% quantile</th>
<th>Median</th>
<th>90% quantile</th>
<th>Max</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Graph</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>102</td>
<td>1.9</td>
<td>3.0</td>
</tr>
<tr>
<td>High risk Graph</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>102</td>
<td>9.1</td>
<td>16</td>
</tr>
<tr>
<td>Low risk Graph</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>96</td>
<td>1.8</td>
<td>3.1</td>
</tr>
</tbody>
</table>

Table 5.2: Statistics of the number of consignees per shipper (shipper's degree) in the overall graph, the high-risk graph and the low-risk graph.

Figure 5-6 compares the distribution of the number of consignees among high risk shippers and unknown shippers (i.e., not on FDA alert of refusal):

Figure 5-6: Comparison of the distributions of the number of consignees per high risk shipper and unknown shippers.
These results suggest that a high degree could indicate a higher risk of adulteration.

2. Weight: The average weight shipped by each shipper. We believe that a company may try to avoid detection by shipping higher volumes less often.

We perform a similar preliminary analysis and observe that the average weight in the high risk graph is significantly higher compared to the overall or low risk graph:

<table>
<thead>
<tr>
<th>Graph</th>
<th>Min</th>
<th>10% quantile</th>
<th>Median</th>
<th>90% quantile</th>
<th>Max</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Graph</td>
<td>1</td>
<td>1615</td>
<td>17585</td>
<td>40741</td>
<td>6652178</td>
<td>26855</td>
<td>151355</td>
</tr>
<tr>
<td>High risk Graph</td>
<td>860</td>
<td>6442</td>
<td>19874</td>
<td>47049</td>
<td>6118350</td>
<td>96034</td>
<td>665268</td>
</tr>
<tr>
<td>Low risk Graph</td>
<td>1</td>
<td>1583</td>
<td>17535</td>
<td>40720</td>
<td>6652178</td>
<td>25785</td>
<td>128132</td>
</tr>
</tbody>
</table>

Table 5.3: Statistics of the average weight (in Kg) per shipper in the overall graph, the high-risk graph and the low-risk graph.

![Unknown Shippers](image1)

![High-risk Shippers](image2)

Figure 5-7: Comparison of the distributions of the average weight per high risk shipper and unknown shippers.
Figure 5-7 compares the distribution of the average weight among high risk shippers and unknown shippers (i.e., not on FDA alert of refusal).

3. **Shipping company**: Binary feature that indicates if the shipper seems to be a logistics company. It is obtained by searching specific strings in the company’s name, such as “logistics”, ”transport”, ”freight”, ”maritime” and ”shipping”. We identify 632 shipping companies in the Import Genius honey database, out of which only 6 had an FDA refusal. That is 1% high risk shippers among shipping companies, compared to 2.3% overall. This suggests that shipping companies are more likely to appear on an FDA alert or refusal.

4. **Transnational shipper - place of receipt**: The fraction of shipments of each shipper where the shipper and the place of receipt are located in different countries. The underlying hypothesis is that crossing a border may be an attempt to disguise the origin of the product.

   This feature is computed by comparing the shipper’s country (available in the shipper’s address in the Import Genius bill of lading) to the place of receipt country, obtained through reverse geo-coding of the place of receipt. Specifically, we use Google Maps API to find the latitude and longitude of the place of receipt (available in the Import Genius bill of lading). We then use Google Maps API again to obtain the address from the geo-codes. We extract the country name from the address and compare it to the shipper’s country.

5. **Transnational place of receipt - foreign port**: The fraction of shipments of each shipper where the place of receipt and the foreign port are located in different countries, for each shipper. The underlying hypothesis is that crossing a border may be an attempt to mask the origin of the product. Similar to the previous feature, we compare the place of receipt country to the foreign port country using reverse geo-coding with Google Maps API.

6. **Distance shipper - place of receipt**: For each shipper, the average distance between the shippers address and the place of receipt.

7. **Distance place of receipt foreign port**: For each shipper, the average distance between the place of receipt and the foreign port. This feature is obtained similarly
to the previous one, using geo-coordinates of the place of receipt and the foreign port and the Haversine formula.

8. **Duplicate shipments**: fraction of duplicate shipments among all shipments of each shipper. Duplicate shipments are defined as shipments that share all entries except for the consignees’ and shippers’ names. More details are provided in section 5.3.1.

9. **Fraction of Country of origin**: this is a set of ten features that represent the fraction of shipments of each shipper that comes from each of the top ten countries of origin (in terms of number of shipments) in the global honey data: Argentina, Brazil, China, Guatemala, India, Honduras, Mexico, New Zealand, Taiwan and Singapore.

### 5.3.1 Data Inconsistencies

The analysis of the Import Genius global honey data reveals the existence of seemingly different shipments that share the exact same entries (i.e., identical bill of lading code, container number(s), product description, weight) except for shipper and/or consignees name. Note that the bill of lading code is supposed to be unique to each shipment. We hypothesize that this could be indicative of suspicious or even adulteration-related activities, as companies might be trying to change their names to avoid detection.

Duplicate shipments are defined as shipments sharing identical entries except for the shipper or consignee names. In particular, these shipments have the same arrival date. We looked at the number of duplicate honey shipments in Import Genius from 2006 to 2015 (62,000 bills of lading) and found the following:

- 2,548 unique shipments are duplicated.

- Duplicate shipments involve 729 shipper names and 660 consignee names.

- Most of the time, when a company’s name is changed, the address is different too, but it seems to be in the same country.
To understand the scale of this observation, we plot in Figure 5-8 a histogram of the fraction of duplicate shipments for the 50 companies with the highest number of shipments in the honey Import Genius database between 2006 and 2015.

![Histogram of fraction of duplicate shipments among the 50 companies with highest number of shipments (Import Genius 2006-2015).](image)

Figure 5-8: Fraction of duplicate shipments among the 50 shippers with highest number of shipments (Import Genius 2006-2015).

![Graph of shippers used in duplicate shipments. An edge indicates that the two shippers names have been used in at least two duplicate shipments.](image)

Figure 5-9: Graph of shippers used in duplicate shipments. An edge indicates that the two shippers names have been used in at least two duplicate shipments.
Figure 5-9 is a graph representing the shippers used in the duplicate shipments. An edge between two shippers means that the two shippers names have been used under the same bill of lading code and on the same arrival date.

In the shipper duplicate graph, there are 18 shippers on FDA Alerts or refusals, out of which 7 are only on FDA alerts and 15 are only on FDA Refusals. Out of all the shippers with duplicate shipments 2.3% are known to be high risk shippers, compared to 1.2% of all shippers. (Note that this does not take into account potential bias in sampling by the FDA). This suggests that the use of duplicate shipments is potentially a statistically significant feature to predict risk of shippers.

Assuming that shippers are sampled uniformly (we acknowledge that this assumption might not be true), we perform a one-sided z-test with the null hypothesis that the proportion of high risk shippers among shippers with duplicate shipments is the same as the proportion of high risk shippers among those with no duplicate shipments. The alternative hypothesis is that the proportion of high risk shippers is higher among shippers with duplicate shipments compared to those with no duplicate shipments. We find that the p-value is 0.003, below the 5% confidence level. Therefore, we reject the null hypothesis and find that shippers with duplicate shipments are more likely to be high risk than shippers with no duplicate shipments.

The median degree of the graph in Figure 5-9 is 1, the average degree is 1.58 and the maximum degree is 40. This means that most of the time, a shipper’s name gets changed to only one other name or a few other names, but there are shippers who use a higher number of alternative names.
Figure 5-10: Graph of consignees used in duplicate shipments. An edge indicates that the two consignees names have been used for consignees in at least two identical shipments.

Similarly, Figure 5-10 is a graph representing the consignees used in the duplicate shipments. An edge between two consignees means that the two consignees' names have been used under the same bill of lading code.

The median degree of the graph in Figure 5-10 is 1, the average degree is 1.7 and the maximum degree is 24. This means that, similar to the duplicate shippers' graph, a consignee’s name gets changed to one other name most of the time.

### 5.4 Shippers Risk Model Framework

In this section, we discuss the shippers risk model framework. Note here that we use potentially speculative data to predict past events (i.e. features account for potential shipments that happened after a company is put on an alert or a refusal, if any).
We consider all shipments that a company would send to the U.S. within a fixed time window if there were no sampling performed by the FDA. We refer to this number of shipments for shipper \( i \) as \( N_i \). In reality, there is sampling by the FDA, and we make the assumption that if a shipper is caught shipping adulterated products it must cease further shipments. Therefore, we define \( n_i \) as the number of shipments of shipper \( i \) that reach the U.S. port on their shipping path. Note that only \( n_i \) is observable, \( N_i \) is unknown.

Clearly, we must have \( n_i \leq N_i \) and for shippers that are never refused, \( n_i = N_i \). We will assume that for each shipper, the probability mass function for \( N_i \) is \( f_{N_i}(\cdot) \).

Though \( N_i \) may seem like an unnecessary variable, it is required in order to properly specify our model. However, we will show that the actual form of \( f_{N_i}(\cdot) \) is not needed to estimate the model parameters.

Each shipment is subject to a random sampling by the FDA. We assume that all shipments of shipper \( i \) are sampled independently with probability \( q_i \). Formally, we assume that the sampling probability \( q_i \) depends upon a set of features \( X_i = \{X_{i1}, X_{i2}, ..., X_{if}\} \) through the standard logistic relationship:

\[
\log \left( \frac{q_i}{1 - q_i} \right) = \sum_{t=1}^{f} \gamma_t X_{it}
\]  

(5.1)

where \( \gamma_t \) measure the importance of the features \( (X_{it}) \) to the sampling fraction of shipper \( i \).

We further assume that the sampling is perfect. That is, if a shipment contains adulterated products and it is sampled, then the adulteration is detected. In this case, the shipper cannot ship any further shipments.

Finally, we define for shipper \( i \) a Bernoulli random variable \( Y_i \) which is one if the shipper ships adulterated products and zero otherwise. \( Y_i = 1 \) means that shipper \( i \) is high risk. If \( Y_i = 0 \), then the shipper is not known to be on alert or refusal. This does not mean that the shipper is not high risk, just that we do not know its status. Therefore, we have a censored data problem.

Let \( p_i = P(Y_i = 1) \). This is the risk score between zero and one that we assign
to each shipper. We assume that if a shipper is high risk, then every one of its shipments contains adulterated products and will be detected by the FDA if any one of the shipments is sampled. We assume that the high risk probability $p_i$ depends upon a set of features $X_i = X_{i1}, X_{i2}, ..., X_{if}$ through the standard logistic relationship:

$$\log \left( \frac{p_i}{1 - p_i} \right) = \sum_{i=1}^{f} \beta_i X_{ii}$$

(5.2)

where $\beta_i$ measure the importance of the feature $(X_{ii})$ to the risk score of shipper $i$.

For $k \geq 2$, the likelihood of the observations is given by

$$P(Y_i = 1, n_i = k, N_i = l|p_i, q_i) = p_i q_i (1 - q_i)^{k-1} \mathbb{1}_{k \leq l} f_{N_i}(l)$$

(5.3)

$$P(Y_i = 0, n_i = k, N_i = l|p_i, q_i) = (1 - p_i + p_i (1 - q_i)^k) \mathbb{1}_{k = l} f_{N_i}(l)$$

(5.4)

and for $k = 1$, the likelihood is given by

$$P(Y_i = 1, n_i = 1, N_i = l|p_i, q_i) = p_i q_i \mathbb{1}_{1 \leq l} f_{N_i}(l)$$

(5.5)

$$P(Y_i = 0, n_i = 1, N_i = l|p_i, q_i) = (1 - p_i + p_i (1 - q_i)) \mathbb{1}_{l = 1} f_{N_i}(l)$$

(5.6)

The intuition behind these expressions is the following:

- A shipper is labeled low risk if it is actually low risk or is high risk but has never been sampled in all its $k$ shipments.

- A shipper is labeled high risk if it is actually high risk, was never caught on its $k - 1$ first shipments, but was caught on its last shipment $k$, for $k \geq 2$, or has one shipment and is on an FDA alert or refusal.

To show that the model is properly specified, we calculate the marginal probabilities $P(Y_i = 0; p_i, q_i)$ and $P(Y_i = 1|p_i, q_i)$ in the case where the sampling probability is the same across all shipments of a shipper and show that they sum to one.

$$P(Y_i = 1|p_i, q_i) = \sum_{l=1}^{\infty} \sum_{k=1}^{l} P(Y_i = 1, n_i = k, N_i = l|p_i, q_i)$$
\[
= p_i q_i \sum_{l=1}^{\infty} \sum_{k=1}^{l} (1 - q_i)^{k-1} \mathbb{1}_{k \leq l} f_{N_i}(l)
\]
\[
= p_i \sum_{l=1}^{\infty} (1 - (1 - q_i)^l) f_{N_i}(l)
\]
\[
= p_i - p_i \sum_{l=1}^{\infty} (1 - q_i)^l f_{N_i}(l)
\]

\[
P(Y_i = 0|p_i, q_i) = \sum_{l=1}^{\infty} \sum_{k=1}^{l} P(Y_i = 0, n_i = k, N_i = l) |p_i, q_i|
\]
\[
= p_i q_i \sum_{l=1}^{\infty} \sum_{k=1}^{l} (1 - p_i + p_i (1 - q_i)^k) \mathbb{1}_{k \leq l} f_{N_i}(l)
\]
\[
= 1 - p_i + p_i \sum_{l=1}^{\infty} (1 - q_i)^l f_{N_i}(l)
\]

\[
P(Y_i = 1|p_i, q_i) + P(Y_i = 0|p_i, q_i) = p_i - p_i \sum_{l=1}^{\infty} (1 - q_i)^l f_{N_i}(l) + 1 - p_i + p_i \sum_{l=1}^{\infty} (1 - q_i)^l f_{N_i}(l)
\]

We obtain: \( P(Y_i = 1|p_i, q_i) + P(Y_i = 0|p_i, q_i) = 1. \)

This model has a few limitations. The assumption that shippers cease any shipping activity after being put on an FDA alert or refusal does not always hold. In fact, out of the 145 high risk shippers in the honey data set, 30 shippers continue to ship honey to the U.S. after being put on alert by the FDA. However, we will show in section 5.6.2 that their shipping behavior does not change after they are caught.

The likelihood function that we wish to maximize is:

\[
L \propto \prod_{i:Y_i=0} \left( 1 - p_i + p_i (1 - q_i)^k \right) \times \prod_{i:Y_i=1} \left( p_i q_i (1 - q_i)^{k-1} \right) \tag{5.7}
\]

\[
\log(L) \propto \sum_i (1 - Y_i) \log \left( 1 - p_i + p_i (1 - q_i)^k \right) + Y_i \log \left( p_i q_i (1 - q_i)^{k-1} \right) \tag{5.8}
\]

We use a Bayesian approach to our model estimation. We assume we have data for \( s \) shippers. For shipper \( i \), we have a set of features \( X_i = \{X_{i1}, X_{i2}, ..., X_{if}\} \), a status label \( Y_i \in \{0, 1\} \), and the number of shipments it sent to the U.S. \( n_i \). We will
refer to the set of all shippers features in our data set as \( X = \{X_1, X_2, ..., X_s\} \), the set of all status labels as \( Y = \{Y_1, Y_2, ..., Y_s\} \), and the set of all shipment counts as \( n = \{n_1, n_2, ..., n_s\} \). We also define the (possibly) unobserved total possible shipments \( N_i \) of shippers \( i \). We wish to calculate the posterior distribution of the model parameters \( \beta = \{\beta_1, \beta_2, ..., \beta_m\} \), \( \gamma = \{\gamma_1, \gamma_2, ..., \gamma_y\} \) and the total shipments \( N = \{N_1, N_2, ..., N_s\} \). That is, we wish to calculate:

\[
P(\beta, \gamma, N|X, Y, n) = \frac{P(Y, n, N|X, \beta, \gamma)P(\beta, \gamma|X)}{P(Y, n|X)}
\] (5.9)

We present in Figure 5-11 the graphical model for the honey shippers risk model.

![Graphical model of the Bayesian model for predicting adulteration among honey shippers.](image)

Figure 5-11: Graphical model of the Bayesian model for predicting adulteration among honey shippers.

We use a Metropolis-Hastings within Gibbs sampler [17] to sample from the posterior distribution of the model parameters. The Metropolis-Hastings algorithm simulates samples from a probability distribution by using the joint density function and proposal distributions for each of the variables of interest. Details of this method are presented in Appendix A.

We define a flat prior distribution on each model parameter and we calculate the posterior distribution given the data (the model features). Priors on the \( \beta \) and \( \gamma \) are normally distributed, centered in 0 with standard deviation 100, so the prior distribution is uninformative.
5.5 Model Estimation and Discussion

In this section, we present results of the in-sample model estimation. We run the sampler for 5,000 iterations and discard a burn-in period of 5,000 iterations.

We consider a feature to be significant if the 90% posterior credibility interval does not contain 0. Results of the median posterior and 90% posterior credibility interval for each feature are shown in Figures 5-12 and 5-13.

We find that the degree (number of consignees) is significant, and positively correlated with risk. This supports our hypothesis that complex supply chains with a high number of consignees are a proxy for suspicious behavior.

We also find that weight is significant and positively correlated with risk of adulteration. This supports our hypothesis that some companies may be trying to avoid detection by shipping higher volumes in fewer shipments.

We find that shipping companies are less likely to adulterate honey. We also find that crossing a border to deliver the product to the shipping company in the place of receipt is correlated with high risk.

Finally, we find that shippers with a high fraction of honey coming from Brazil are less likely to be on an FDA refusal, while those with a high fraction of honey coming from India, Mexico, New Zealand, Taiwan or Singapore, are more likely to be caught on an FDA refusal.

<table>
<thead>
<tr>
<th>Features</th>
<th>Intercept</th>
<th>Degree</th>
<th>Weight</th>
<th>Shipping Company</th>
<th>Transnational SC-PR</th>
<th>Transnational PR-FP</th>
<th>Distance SC-PR</th>
<th>Distance PR-FP</th>
<th>Duplicate shipments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>0.0189</td>
<td>0.0281</td>
<td>-0.3031</td>
<td>0.0928</td>
<td>-0.2982</td>
<td>-0.0572</td>
<td>0.1094</td>
<td>-0.0639</td>
<td>-0.1100</td>
</tr>
<tr>
<td>90% confidence interval</td>
<td>[0.0182, 0.0195]</td>
<td>[-0.1004, 0.1209]</td>
<td>[-0.1778, 0.0778]</td>
<td>[-0.0042, 0.2603]</td>
<td>[-0.5402, 0.1393]</td>
<td>[-0.2049, 0.3984]</td>
<td>[-0.1553, 0.1840]</td>
<td>[-0.0924, 0.1529]</td>
<td>[-0.4447, 0.5372]</td>
</tr>
</tbody>
</table>

Figure 5-12: Model parameters' statistics for sampling in the honey shipper risk model.
Figure 5-13: Model parameters' statistics for adulteration in the honey shipper risk model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant.

It is interesting to see that neither the distance between the manufacturer and the place of receipt, not the distance between the place of receipt and the foreign port are significant, and the reason could be that honey is not a perishable product.

5.6 Model Validation

5.6.1 Out of Sample Analysis

We perform a standard out-of-sample analysis where we partition the data in half in a stratified way on the dependent variable.

We assess the performance of our model by plotting the Receiver Operating Characteristic (ROC) and computing the Area Under the Curve (AUC). Note that due to the censored nature of our data (i.e., some unknown companies may also be high risk
but they were never sampled), the false detection rate is an overestimate. Therefore, the AUC is an underestimate of the predictive power of our models. The AUC has a maximum value of 1 (perfect model) and random guessing has a value of 0.5.

Figure 5-13 shows the out of sample ROC curve for one data partition, and Figure 5-14 shows the boxplot of the AUC values for 10 random data partitions:

![ROC curve and boxplot](image)

Figure 5-14: ROC for honey shippers risk
Figure 5-15: Boxplot for honey shippers model using one random data split. Therisk model AUC across 10 random data splits.

The model performs significantly better than random. We are able to predict with a high accuracy high risk behavior among honey shippers based on supply chain features.

Note that due to the censored nature of our data (i.e., some unknown companies may also be high risk but they were never sampled), the false detection rate is an overestimate. Therefore, the AUC is an underestimate of the predictive power of the model.

To obtain an additional predictive measure, we perform an out of sample likelihood test to assess the performance of the model versus the null model (with no features), and to assess the importance of different model features. In order to penalize for the number of features used in the prediction, we use the Deviance Information Criterion (DIC) [18].

\[
DIC = -2 \mathbb{E}[\log(p(Y|\beta, \gamma))] + \log(p(Y|\mathbb{E}[\beta, \gamma])) + C
\]  

(5.10)
where $C$ is a constant. A lower DIC indicates that the model fits the data better. The Deviance Information Criterion is computed out of sample.

We run 10 random data partitions. In each partition, we compute the difference between the full model DIC and the null model DIC (with no features), as well as the difference between the full model DIC and the DIC obtained after removing one feature at a time. We represent in Figure 5-15 the average DIC shift from the full model (taken as a reference at $y=0$) as well as the standard error over the 10 data partitions.

![ DIC Difference Chart ]

Figure 5-16: DIC shift from the full model (including all features) when we remove one feature at a time, for predicting honey adulteration. The error bars represent the standard error over 10 random data partitions. The null model does not include any model feature.

We find that the model fits the data better when using all features, and always performs better than the null model when we remove different features, one at a time. We also find that the five most important features in model fit are: 1. number of consignees, 2. fraction of shipments with country of origin in New Zealand, 3. fraction of shipments with country of origin in Singapore, 4. transnational fraction between the place of receipt and the foreign port, and 5. the average shipment weight.
5.6.2 Model Robustness Check

In order to check that the model is well specified even though we may be using future data to predict past events, we need to check that high risk shippers don’t change their behavior after being caught, as this is an assumption of our model.

Recall that the honey risk model assumes that shippers cease any shipping activity after they are put on an FDA alert or refusal. In reality, of the 145 shippers on FDA alerts or refusals, 30 shippers continue to ship honey to the U.S. after being caught.

We analyze the difference between the features’ distribution of those shippers before and after being caught. We find that there is no proof that their shipping behavior changes after their refused shipments. This justifies the use of features spanning from 2006 until 2015, as the value of these features up to the date of the refused shipments does not vary significantly after.

We compute feature values of all 30 shippers on FDA alerts or refusals that continued shipping honey to the U.S. after being caught, before and after their refused shipment. Then, we compare the distributions of the two populations (before and after being caught) using a t-test and a Kolmogorov-Smirnoff test. The findings are shown in Table 5.4.

<table>
<thead>
<tr>
<th>Features</th>
<th>Number of Shipment</th>
<th>Number of Consignees</th>
<th>Average Weight</th>
<th>Duplicate Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-test p-value</td>
<td>0.88</td>
<td>0.96</td>
<td>0.32</td>
<td>0.55</td>
</tr>
<tr>
<td>Kolmogorov-Smirnoff Test p-value</td>
<td>0.79</td>
<td>0.58</td>
<td>0.95</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 5.4: Statistical tests p-values to compares features’ distributions of the 30 shippers with shipments before and after FDA refusals, showing that we fail to reject the null hypothesis that the means are equal.

We show the boxplots of shippers’ features before and after being put on an FDA refusal in Figures 5-16 and 5-17. We find indeed that there is no sufficient evidence to reject the null hypothesis that the distributions of features before and after a refused
shipments are similar.

Figure 5-17: Boxplot of the average number of shipments and average number of consignees for honey shippers on FDA refusals, before and after being put on a refusal.

Figure 5-18: Boxplot of the average weight and average duplicate fraction for honey shippers on FDA refusals, before and after being put on a refusal.
In order to check that the model assigns high risk scores to shippers previously caught by the FDA, we plot in Figure 5-18 the density of the in-sample predicted risk scores. We observe three peaks, suggesting three distinct clusters of shippers.

We use a Gaussian mixture to find the underlying distribution of the predicted risk scores, using three Gaussians, initialized respectively at $\mathcal{N}(0, 1)$, $\mathcal{N}(0.5, 1)$ and $\mathcal{N}(1, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution of mean $\mu$ and variance $\sigma^2$.

We perform an unsupervised k-means clustering on all model features with $k = 3$. We call the resulting clusters of shippers "group 1", "group 2" and "group 3", in increasing average risk score.

We analyze each cluster of shippers by computing average values of the significant features in the in-sample model estimation, and then we compare the values obtained between the three clusters. Tables 5.5 and 5.6 summarize the findings:
### Table 5.5: Statistics of shippers’ features in the three K-means clusters.

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of shippers</th>
<th>Average Number of Shipments</th>
<th>Average Number of Consignees</th>
<th>Average Weight (kg)</th>
<th>Average duplicate fraction</th>
<th>Fraction of high risk shippers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>5418</td>
<td>5.78</td>
<td>1.53</td>
<td>11,175</td>
<td>0.085</td>
<td>0</td>
</tr>
<tr>
<td>Group 2</td>
<td>625</td>
<td>56.52</td>
<td>4.83</td>
<td>68,312</td>
<td>0.126</td>
<td>0.016</td>
</tr>
<tr>
<td>Group 3</td>
<td>161</td>
<td>34.91</td>
<td>4.12</td>
<td>13,335</td>
<td>0.104</td>
<td>0.838</td>
</tr>
</tbody>
</table>

The majority of high risk shippers (83.8%) are captured in the third cluster (with high predicted risk scores), and none of them are present in the first cluster. This shows that our model performs well.

Features of shippers in groups 1 (lowest predicted risk scores) have low average feature values compared to other clusters. This observation matches the results of the model estimation that a high number of consignees, a high average weight and a high average duplicate fraction are correlated with higher risk of adulteration.

Features of shippers in cluster 3 (highest predicted risk scores) also match the findings of the model estimation (high number of consignees, high average weight and high duplicate fraction correlate with higher risk of suspicious behavior). The FDA could benefit from sampling the shippers in group 3 that did not have a high risk status (if they weren’t already sampled.) These are the shippers with the highest predicted risk scores based on our model, that were not caught on an FDA alert or
refusal.

Only a small fraction of shippers in group 2 (1.6%) were caught on FDA alerts or refusals, but this could be due to infrequent sampling in this group. However, these shippers have the highest average number of consignees, average weight and average duplicate fraction among the three groups. The difference in predicted risk scores between groups 2 and 3 is explained by the fraction of shipments from the different countries of origin used as predictive features. In particular, shippers with a high fraction of shipments coming from New Zealand belong to group 3, with the highest predicted risk scores. This observation matches the model estimation results that show that a higher fraction of shipments from New Zealand correlates with more risk. This is not surprising, since New Zealand has the highest number of honey refusals between 2006 and 2015.

5.7 Summary

We were able to find a predictive relationship between shipping supply chain structure and risk of adulteration in imported honey among foreign shippers. Some of these predictive features were a high number of consignees and a high fraction of duplicate shipments, supporting our initial hypotheses. These insights can be used to rank honey shippers based upon their risk level, and better inform sampling policies for honey shippers.
Chapter 6

Conclusion

6.1 Summary of the Results

In this thesis, we presented a framework to predict risk of food adulteration in imported products, specifically honey and shrimp, based on public and internal FDA data over the past ten years. We identified features predictive of risk, based on the supply chain structure, the traveled route, and the product portfolio of companies. We used a Bayesian approach to estimate the model, coupled with a two-step approach to correct for sampling bias in the shipments data.

While different features are significant for different products and types of adulteration, there seems to be a consensus between models that complex and diverse supply chain networks correlate with high risk of adulteration. This is reflected by the number of partners (shippers, manufacturers or consignees) that a company works with.

We also find that yearly variations in the supply chain structure indicates higher risk levels. Companies with high variation of number of shipments and number of partners across years were more likely to participate in adulteration.

Crossing a country border was correlated of a higher risk of intentional adulteration of shrimp, indicating that manufacturers could attempt to adulterate perishable products to maintain them in longer routes, or to disguise the country of origin if that country has a known history of food adulteration.
Another important insight was that U.S. shrimp consignees and importers that received unrelated products were more likely to fail FDA site inspections, indicating that non-specialized companies have a higher risk level.

The different models always performed better than random guessing. Therefore, they could be utilized to improve sampling and inspection policies to better allocate resources and catch more adulteration before entering the U.S.

6.2 Extensions and Future Directions

There are several directions to extend this work. First, similar models can be applied to different companies in the shipping supply chain. One natural direction is to predict risk of intentional and unintentional adulteration among South American shrimp manufacturers, using an identical framework to the manufacturer sampling model detailed in Chapter 3. Another direction is to analyze farmers behavior, and map farmer specific features to risk of food adulteration later in the supply chain.

Second, we could analyze types of adulteration at a more granular level, potentially with more data. Can we use supply chain structural features to predict presence of antibiotics, pesticides, food additives, or salmonella? This approach can be applied to different companies as well.

The honey shippers risk model was censored since we did not have data on sampled shipments. If these data become available, they could be utilized to build a two-step model, similar to the shrimp models, to model both sampling and refusals.

Another extension of this work is to base risk prediction on supply chain features computed from road and air entry data. These features would be different from what we presented in this work, but the general framework would be similar.

Finally, this work can be applied to commodities other than honey and shrimp. Other features may be correlated with risk of adulteration, but we would expect complexity and volatility of supply chain networks, as well as high diversity of the product portfolio, to still be indicative of risk.
Appendix A

Details of MCMC Sampler

We use a Metropolis-within-Gibbs scheme to sample from the posterior distribution of the model parameters. We define the set of model parameters as $\Theta$:

- Honey shipper risk model: $\Theta = \{\gamma, \beta\}$
- Shrimp manufacturer model: $\Theta = \{\gamma, \beta, \sigma_c, \rho\}$
- Shrimp consignee/importer inspection model: $\Theta = \{\gamma, \beta, \sigma_c, \rho\}$

For any parameter $\alpha \in \Theta$, we define the set of parameters excluding $\alpha$ as $\Theta_{-\alpha}$. We also define the set of observable data as $X$, and the response variable as $Y$. In the honey model, we observe the set of features and the number of shipments $n$. In the shrimp models, we observe the set of features $X$, the response variable in the first step of the Heckman selection model is $S$ (fraction of sampled shipments for manufacturers, or binary indicator of inspections for consignees and importers). The response variable in the second step of the model is $Y$ (fraction of refused shipments for manufacturers, binary indicator of failed inspections for consignees and importers). We must sample from the conditional distribution $P(\alpha|X, Y, \Theta_{-\alpha})$ for each model parameter.

Let $P(Y|X, \Theta)$ be the likelihood function. The posterior distribution of parameter $\alpha$ in $\Theta$ is:

$$P(\alpha|X, Y, \Theta_{-\alpha}) = \frac{P(Y|X, \Theta)P(\alpha|X, \Theta_{-\alpha})}{P(Y|X, \Theta_{-\alpha})}$$
In the honey shipper risk model, $\gamma$ and $\beta$ are initialized from a uniform distribution $U([-10, 10])$, and their priors are generated from a normal distribution: $\gamma \sim \mathcal{N}(0, 10^2)$ and $\beta \sim \mathcal{N}(0, 10^2)$. We use a symmetric random walk proposal $g = s \mathcal{N}(0, 1)$, where $s = 0.1$ is a fixed step size. We draw $\epsilon \sim g$ and set $\gamma = \gamma^{(t-1)} + \epsilon$ and $\beta = \beta^{(t-1)} + \epsilon$ at iteration $t$ of the MCMC sampler. We then compute the acceptance probability $\alpha_\gamma = \min\{1, \frac{P(\gamma|X,Y,\beta)}{P(\gamma^{(t-1)}|X,Y,\beta)}\}$ for $\gamma$, and $\alpha_\beta = \min\{1, \frac{P(\beta|\gamma,Y)}{P(\beta^{(t-1)}|\gamma,Y)}\}$ for $\beta$. Then we set $\gamma^{(t-1)}$ to $\gamma$ with probability $\alpha_\gamma$, and set $\beta^{(t-1)}$ to $\beta$ with probability $\alpha_\beta$. We run the sampler for 5,000 iterations and discard a burn-in period of 5,000 iterations. We then assess convergence visually using trace plots of the posterior distributions of model parameters.

In the shrimp models, the framework is similar. For each step of the Heckman selection model, we initialize model parameters: $\gamma^0 \sim U([-10, 10]), \beta^0 \sim U([-10, 10]), \sigma^0_\epsilon \sim U([-10, 10])$ and $\rho^0 \sim U([-1, 1])$. Then, we compute the posterior of each model parameter using the likelihood function and the prior. We use the following priors for model parameters: $\gamma \sim \mathcal{N}(0, 10^2), \beta \sim \mathcal{N}(0, 10^2), \sigma_\epsilon \sim U([-10, 10])$ and $\rho \sim U([-1, 1])$. We generate a random walk proposal $g = s \mathcal{N}(0, 1)$, where $s = 0.1$ is a fixed step size, draw $\epsilon \sim g$, and set $\gamma = \gamma^{(t-1)} + \epsilon$, $\beta = \beta^{(t-1)} + \epsilon$, $\sigma_\epsilon = \sigma_\epsilon^{(t-1)} + \epsilon$ and $\rho = \rho^{(t-1)} + \epsilon$. Similarly, we run the sampler for 5,000 iterations and discard a burn-in period of 5,000 iterations, and assess convergence visually using trace plots of the posterior distributions of model parameters.
Appendix B

Refusal Codes Tables

In this appendix, we present the description of the most common FDA refusals in honey and shrimp.

In Table B.1, we list the violation numbers, the charge codes and the corresponding charge statement texts in honey. The violations are ranked from the most frequent to the least frequent between 2006 and 2015.

Table B.2 lists violation codes for shrimp and the corresponding refusal descriptions. These description were used to map each violation code in FDA refusals to a refusal category, such as salmonella or animal drugs. This is particularly important in determining whether a shrimp refusal is intentional or unintentional. The violations are ranked from the most frequent to the least frequent between 2007 and 2015.
<table>
<thead>
<tr>
<th>ASC ID</th>
<th>CHRG CODE</th>
<th>CHRG STMNT TXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>UNAPPROVED-75</td>
<td>The article appears to be a new drug without an approved new drug application.</td>
</tr>
<tr>
<td>218</td>
<td>LIST INGRE-218</td>
<td>The article is subject to refusal of admission pursuant to Section 801(a)(3) of the FD&amp;C Act in that it appears to be misbranded within the meaning of Section 403(i)(2) of the FD&amp;C Act in that it is fabricated from two or more ingredients and the label fails to bear the common or usual name of each such ingredient and/or the article purports to be a beverage containing vegetable or fruit juice, but does not bear a statement with appropriate prominence on the information panel of the total percentage of such fruit or vegetable juice contained in the food. [Misbranded, Section 403(i)(2)].</td>
</tr>
<tr>
<td>238</td>
<td>UNSAFE ADD-238</td>
<td>The article is subject to refusal of admission pursuant to Section 801(a)(3) in that it appears to be adulterated because it appears to contain a poisonous or deleterious substance which may render it injurious to health [Adulteration, Section 402(a)(1)].</td>
</tr>
<tr>
<td>249</td>
<td>FILTHY-249</td>
<td>The article appears to consist in whole or in part of a filthy, putrid, or decomposed substance or be otherwise unfit for food.</td>
</tr>
<tr>
<td>251</td>
<td>POISONOUS-251</td>
<td>The article is subject to refusal of admission pursuant to Section 801(a)(3) in that it appears to contain a poisonous or deleterious substance which may render the article injurious to health [Adulteration, Section 402(a)(1)].</td>
</tr>
<tr>
<td>265</td>
<td>SUBSTITUTE-265</td>
<td>It appears that a substance has been substituted wholly or in part for one or more of the article’s ingredients.</td>
</tr>
<tr>
<td>315</td>
<td>ADDED BULK-315</td>
<td>The food appears to have a substance added to, mixed or packed with it so as to increase its bulk or weight, or reduce its quality or strength, or make it appear better or of greater value than it is.</td>
</tr>
<tr>
<td>320</td>
<td>LACKS FIRM-320</td>
<td>The article is subject to refusal of admission pursuant to Section 801(a)(3) of the FD&amp;C Act in that it appears to be misbranded within the meaning of Section 403(e)(1) of the FD&amp;C Act in that the food is in package form and the label fails to bear the name and place of business of the manufacturer, packer, or distributor [Misbranded, Section 403(e)(1)].</td>
</tr>
<tr>
<td>321</td>
<td>LACKS N/C-321</td>
<td>The article is subject to refusal of admission pursuant to Section 801(a)(3) of the FD&amp;C Act in that it appears to be misbranded within the meaning of Section 403(e)(2) of the FD&amp;C Act in that the food is in package form and the label fails to bear an accurate statement of the quantity of the contents in terms of weight, measure, or numerical count in accordance with Section 403(e)(2) of the FD&amp;C Act [Misbranded, Section 403(e)(2)].</td>
</tr>
</tbody>
</table>
The article is subject to refusal of admission pursuant to Section 801(a)(3) of the FD&C Act in that it appears to be misbranded within the meaning of Section 403(f) of the FD&C Act in that any word, statement, or other information required by or under the authority of the FD&C Act to appear on the label or labeling is not prominently placed thereon with such conspicuousness (as compared with other words, statements, designs, or devices, in the labeling) and in such terms as to render it likely to be read and understood by the ordinary individual under customary terms of purchase and use (for example, label contains information in two or more languages but fails to repeat all required information in both languages in accordance with 21 CFR 101.15(c)(2), or label fails to include all required information in English in accordance with 21 CFR 101.15(c)(1), except in the case of articles distributed solely in the Commonwealth of Puerto Rico or in a Territory where the predominant language is one other than English)) [Misbranded, Section 403(f)].

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Violation</th>
</tr>
</thead>
<tbody>
<tr>
<td>473</td>
<td>LABELING</td>
<td>The article appears in violation of FPLA because of its placement, form and/or contents statement.</td>
</tr>
<tr>
<td>473</td>
<td>473</td>
<td></td>
</tr>
<tr>
<td>482</td>
<td>NUTRIT</td>
<td>The article appears to be misbranded in that the label or labeling fails to bear the required nutrition information.</td>
</tr>
<tr>
<td>482</td>
<td>LBL-482</td>
<td></td>
</tr>
<tr>
<td>2860</td>
<td>VETDRUGRES</td>
<td>The article appears to contain a new animal drug (or conversion product thereof) that is unsafe within the meaning of section 512.</td>
</tr>
<tr>
<td>2860</td>
<td>2860</td>
<td></td>
</tr>
<tr>
<td>2920</td>
<td>PESTICIDES</td>
<td>The article is subject to refusal of admission pursuant to section 801(a)(3) in that it appears to be adulterated because it contains a pesticide chemical, which is in violation of section 402(a)(2)(B).</td>
</tr>
<tr>
<td>2920</td>
<td>2920</td>
<td></td>
</tr>
</tbody>
</table>

Table B.1: Violation codes and descriptions in honey refusals between 2006 and 2015. 
[www.fda.gov/ForIndustry/ImportProgram/ImportRefusals/ucm144864.htm](http://www.fda.gov/ForIndustry/ImportProgram/ImportRefusals/ucm144864.htm)

134
<table>
<thead>
<tr>
<th>Violation Code</th>
<th>Refusal Description</th>
<th>Refusal Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>The article appears to contain Salmonella, a poisonous and deleterious substance which may render it injurious to health.</td>
<td>Salmonella</td>
</tr>
<tr>
<td>2860</td>
<td>The article appears to contain a new animal drug (or conversion product thereof) that is unsafe within the meaning of section 512.</td>
<td>Animal Drugs</td>
</tr>
<tr>
<td>249</td>
<td>The article appears to consist in whole or in part of a filthy, putrid, or decomposed substance or be otherwise unfit for food.</td>
<td>Filth</td>
</tr>
<tr>
<td>3220</td>
<td>The article is subject to refusal of admission in that it appears to bear or contain a food additive, namely nitrofurans, that is unsafe.</td>
<td>Food Additives</td>
</tr>
<tr>
<td>238</td>
<td>The article is subject to refusal of admission pursuant to Section 801(a)(3) in that it appears to be adulterated because it appears to contain a poisonous or deleterious substance which may render it injurious to health.</td>
<td>Food Additives</td>
</tr>
<tr>
<td>482</td>
<td>The article appears to be misbranded in that the label or labeling fails to bear the required nutrition information.</td>
<td>Misbranding</td>
</tr>
<tr>
<td>11</td>
<td>The article appears to be, or to bear or contain a color additive which is unsafe within the meaning of Section 721(a).</td>
<td>Color Additives</td>
</tr>
<tr>
<td>218</td>
<td>The article is subject to refusal of admission pursuant to Section 801(a)(3) of the FD&amp;C Act in that it appears to be misbranded within the meaning of Section 403(i)(2) of the FD&amp;C Act in that it is fabricated from two or more ingredients and the label fails to bear the common or usual name of each such ingredient and/or the article purports to be a beverage containing vegetable or fruit juice, but does not bear a statement with appropriate prominence on the information panel of the total percentage of such fruit or vegetable juice contained in the food.</td>
<td>Misbranding</td>
</tr>
</tbody>
</table>

Table B.2: Translation of FDA violation codes in shrimp refusals refusals between 2007 and 2015. [www.fda.gov/ForIndustry/ImportProgram/ImportRefusals/ucm144864.htm](http://www.fda.gov/ForIndustry/ImportProgram/ImportRefusals/ucm144864.htm)
Appendix C

Chinese Manufacturers Risk Model

Similarly to the manufacturers risk model, presented in Chapter 5, we present in this appendix a model for Chinese manufacturers only, in order to predict risk of intentional and unintentional adulteration in shrimp coming from China to the U.S., based on the same data sources described in Chapter 5. There are 361 unique Chinese manufacturers in the shrimp database, including manufacturers from Hong Kong.

In addition to supply chain features, we leverage socio-economic data available through the Chinese annual industrial census (http://www.stats.gov.cn/english), which provides information on provincial value added tax (VAT) payments. The VAT is a consumption tax that is placed on a product whenever value is added at a stage of production and at final sale.

We compute the median VAT of each Chinese province, and use the obtained value for all manufacturers based in that province. If the value is missing for a province, we impute the median VAT across all provinces. A higher VAT reflects a stricter regulatory policy. Therefore, we expect a lower rate to be correlated with more risk of intentional adulteration.

We also two additionnal features for Chinese manufacturers:
• AQSIQ (the Administration of Quality Supervision, Inspection and Quarantine): Binary variable that indicates if a Chinese manufacturer is on the AQSIQ list of manufacturers certified to export to the US. There are 18 manufacturers on the AQSIQ list out of the 361 Chinese manufacturers in our data set. The AQSIQ list can be found in the following link: http://english.aqsiq.gov.cn/

• HACCP (Hazard Analysis and Critical Control Points): Binary variable that indicates if a Chinese manufacturer is on the HACCP list of manufacturers certified to export to the US. There are 31 manufacturers on the HACCP list out of the 361 Chinese manufacturers in our data set. The HACCP list can be found in the following link: http://english.cnca.gov.cn/

The model we estimate is identical to that described in Chapter 5, year by year between 2012 and 2015, for intentional and unintentional adulteration. The only differences are that the model is trained on Chinese manufacturers only, that we don’t use the manufacturer’s country as a feature for predicting sampling or refusals, and that we use the median provincial VAT as an additional feature.

The tables in Figures C.1, C.2 and C.3 show the significant features in the Chinese manufacturers risk model, for sampling, intentional adulteration, and unintentional adulteration respectively.

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of consignees</th>
<th>Number of shippers</th>
<th>Normalized Weight</th>
<th>Transnational PR-FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.44 (0.27,0.96)</td>
<td></td>
<td>-0.22 (0.04,0.53)</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>0.58 (0.28,1.09)</td>
<td>0.17 (0.05,0.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>0.55 (0.34,0.86)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>0.55 (0.24,1.12)</td>
<td></td>
<td>0.15 (0.05, 0.49)</td>
<td>-0.26 (-0.52,-0.07)</td>
</tr>
</tbody>
</table>

Figure C-1: Model parameters’ for predicting sampling in the Chinese manufacturers risk model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it’s negatively correlated with risk. No sign means that the feature is not significant.
Figure C-2: Model parameters' for predicting intentional adulteration in the Chinese manufacturers risk model. A positive sign (orange cell) means that the feature is positively correlated with risk, a negative sign (blue cell) means that it's negatively correlated with risk. No sign means that the feature is not significant.

We find that Chinese manufacturers with a higher number of consignees are more likely to be sampled by the FDA in all four year models. Sampled Chinese manufacturers with a higher number of consignees are more likely to be caught for intentional adulteration. Those shipping a high weight compared to the yearly average across Chinese manufacturers are also more likely to have their shipments refused because of intentional adulteration.

Perhaps the most interesting finding here is that the VAT is significant and negatively correlated with the risk of intentional adulteration among Chinese manufacturers, in the 2012 and the 2015 models. This finding matches the hypotheses that a lower regulatory structure (reflected by a low VAT) is correlated with a higher risk of intentional adulteration.
We find that a high number of consignees and a high number of average yearly normalized weight correlated with a higher risk of unintentional adulteration among Chinese manufacturers. However, the VAT is not significant for unintentional adulteration.

Finally, AQSIQ and HACCP lists were not significant in either sampling, intentional or unintentional adulteration.

The final results suggest that a lower VAT could indicate a higher risk of economically motivated adulteration among Chinese manufacturers.
Bibliography

[1] CDC research shows outbreaks linked to imported foods increasing. www.cdc.gov/media/releases/2012/p0314_foodborne.html.


[32] The situation of farmed shrimp production in asian countries facing ems.


