Recalled Food Products distributed to North Carolina (NC)

Joyce Choe
Department of Biostatistics, UNC-Chapel Hill

BIOS 669: Working with Data in a Public Health Setting (SAS II)
Final Class Project
April 30, 2024

Abstract

The Food Enforcement Report created by the U.S. Food and Drug Administration (FDA) is an up-to-date record of recalled food product events in the United States. Each recall is dated and retrievable through an open-source FDA Application Programming Interface (API). The FDA has made the data available to the public and the data consist of recalled food products that failed to meet public safety and health regulations or standards. To examine factors related to recalled food products distributed to North Carolina (NC) between June 2012 through April 2024, I will analyze such data for this project using SAS 9.4. The primary goal is to explore the frequency of recalled food products distributed to NC from a food firm's U.S. location or abroad. A secondary goal is to calculate the number of days between initiation and termination recall dates to understand relationships between number of days and region, reason for recall, initial firm notification, or voluntary-mandated status using 1-way non-parametric ANOVA. This project will explore unique trends and patterns in the data, so that these initial discoveries will facilitate statistical testing of where and when recalled food products distributed to NC are likely to occur.

Data Sources

The data for this project were sourced from:

1) openFDA food enforcement report API
2) USDA infographic of food business center regions

Although the first source is updated weekly, the data were extracted between June 1, 2004 to April 30, 2024, inclusive. The second source was derived from an infographic, which was used to create a data set where a food firm's location was matched to a geographic region. For this project, I modified the look-up infographic data to exclude Intertribal region, and replaced this category with international region, since some recalled food products were sourced from outside of the U.S. Then, without making further changes to the look-up data set, I retained eleven U.S. regions: Northwest and Rocky Mountain, Southwest, North Central, Heartland, Rio Grande Colonias, Great Lakes Midwest, Delta, Appalachia, Northeast, Southeast, and Islands and Remote Areas (Figure 1). After merging these two data sources into one analysis data set, I cleaned the data set for recalls with initiation dates between June 1, 2012 – April 30, 2024, recalls with terminated status (completely investigated status), and recalls distributed to NC. In addition, I derived and validated custom variables (denoted as * in Table 1), restricted the number of observations to unique recall event IDs, and excluded observations with missing values from variables of interest.

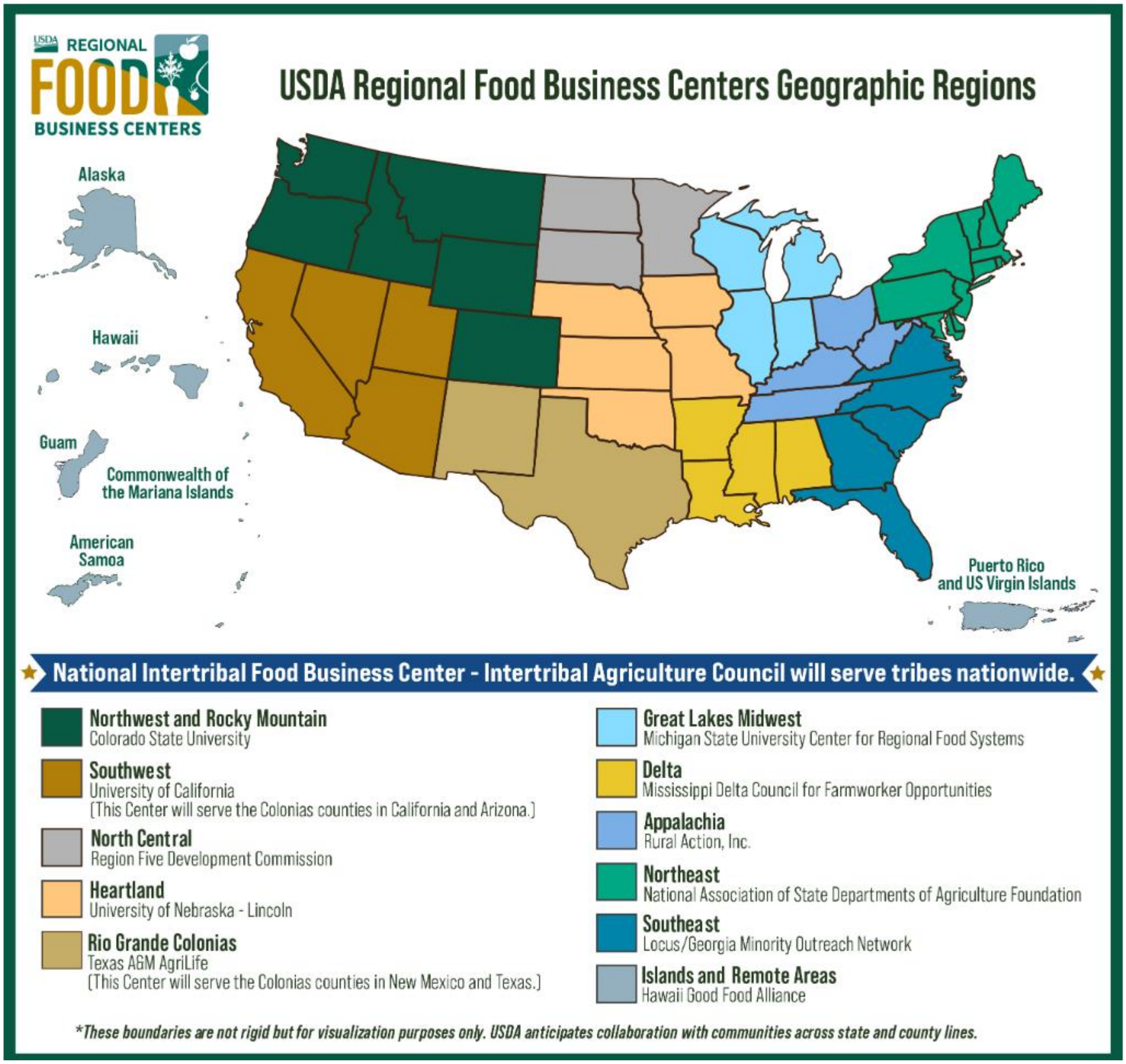
USDA REGIONAL FOOD BUSINESS CENTERS
USDA Regional Food Business Centers Geographic Regions
Alaska
Hawaii
Guam
Commonwealth of the Mariana Islands
American Samoa
Puerto Rico and US Virgin Islands
National Intertribal Food Business Center - Intertribal Agriculture Council will serve tribes nationwide.
Northwest and Rocky Mountain
Colorado State University
Southwest
University of California
(This Center will serve the Colonias counties in California and Arizona.)
North Central
Region Five Development Commission
Heartland
University of Nebraska - Lincoln
Rio Grande Colonias
Texas A&M AgriLife
(This Center will serve the Colonias counties in New Mexico and Texas.)
Great Lakes Midwest
Michigan State University Center for Regional Food Systems
Delta
Mississippi Delta Council for Farmworker Opportunities
Appalachia
Rural Action, Inc.
Northeast
National Association of State Departments of Agriculture Foundation
Southeast
Locus/Georgia Minority Outreach Network
Islands and Remote Areas
Hawaii Good Food Alliance
*These boundaries are not rigid but for visualization purposes only. USDA anticipates collaboration with communities across state and county lines.

Figure 1. Second Data Source: USDA infographic of food business center regions

Methods

The final analysis data set was composed of 1865 unique observed recalls (rows) and a set of 10 variables (columns) listed in Table 1. The variables with asterisks were derived through calculation or re-classification of existing variables. Altogether, these variables were used to visualize trends and to perform statistical tests.

For visualization, I plotted series and stacked bar plots to determine the frequency of recalls each year by region. Then, I plotted multiple boxplots showing number of days to process a recall to completion (response variable) by several categorical variables. The six categorical variables chosen for analyses include: voluntary mandated status, region, general reason, initial firm notification, classification, and year. For a set of boxplots, the median days to process a recall and total recalls per categorical level showed that the observations were positively skewed and inconsistent in sample size across different levels. Next, after assuming independence of the observed recalls and discovering a non-parametric distribution of the response variable, I performed 1-way Kruskal-Wallis ANOVA to test for the relationship between days and voluntary mandated status, region, general reason, initial firm notification, classification, or year. Kruskal-Wallis tests whether the median number of days is equal across level for a single categorical variable for the null hypothesis, while the alternative hypothesis is that at least one median is significantly different between levels of a categorical variable.

Results

Exploratory data analysis showed that the number of recall events were relatively low and extremely skewed across levels of categorical variables: voluntary mandated status and geographic food business center region. For example, 7 FDA-mandated recalls and 1879 voluntary firm-initiated recalls were recorded between June 2012 and April 2024. Also, the International region had 3 recalled food products compared to 23 recalled food products in Island & Remote areas and Delta (mainland) in the same time period.

Since discrepancies in frequency of recalls were present for each categorical variable, the median number of recalls for each level were compared using Kruskal-Wallis test. Kruskal-Wallis test showed that the number of days to process recalls differed significantly by year. Year 2015 had the highest median days (344 days), while 2024 has the lowest median days (69 days) as expected, since 2024 data was limited from January 1, 2024 to April 30, 2024, inclusive. Surprisingly, year 2012 was relatively high even though the earliest data point began on June 2012.

The Kruskal-Wallis test also showed at least one significant difference in the number of recalls by region. The international region had the highest median days (497 days) to process recall events, although with a sample size of 3, while the Southwest region had the lowest median days (132 days). All other categorical variables, voluntary status, reason, notification, and classification did not significantly differ in median days. For most grouped variables, the boxplots were positively skewed, which makes sense since we would hope for or expect most recalls in the U.S. to be processed quickly.

Conclusion

In conclusion, exploratory data analysis uncovered patterns in the data that would otherwise not have been known and generated interesting questions about the data. By using exploratory techniques, including frequency line plots, box plots, and descriptive statistics tables between variables, the most appropriate and rigorous statistical tests for comparing differences of the response variable can be determined. For example, for this project, I compared differences in median days to process a recall across different levels of a specific categorical variable. Different frequencies of recalls for each level by categorical variable showed that the number of recalls distributed to NC was not equal across level for year and geographic location of a food business. Using plots, tables, and formal hypothesis testing, we can understand factors related to recall occurrence and recall duration in days to potentially prevent recalled food products distributed to NC of similar profile with a recurring theme, place, or time. Ultimately, we can examine in detail where and when recalled food products distributed to N.C. generally occur by analyzing the data.

Limitations

The raw data set was derived from the openAPI food enforcement site, where data were ‘untidy’ prior to analysis. Several values in the data set were missing or mistyped under the reason_for_recall variable. Thus, I filtered and parsed through the text using regular expressions to generalize each food product to one of six reasons for recall: Unprepared, Precaution, Microbe, Mislabeled, Contaminant, and Other.

These levels are defined as follows: The ‘Unprepared’ reason includes undercooked or under washed foods that were not prepared safely. The ‘Precaution’ reason includes foods that were potentially unsafe or at risk of causing hazardous outcomes if consumed. The ‘Microbe’ reason includes foods linked with a toxic outbreak or pathogen resulting in a foodborne illness. The ‘Contaminant’ reason includes foods with unacceptable levels of contamination, not microbial in nature. Finally, the ‘Other’ category includes all other recalled food products not belonging to the first five categories.

Table 1. Variables of interest used in the final analysis data set

Name	**Description**
classification	Class I (adverse), Class II (somewhat adverse), Class III (unlikely adverse)
*date1year	Year of recall initiation date; calculated as substring from recall_initiation_date variable
*days	Number of days between termination-initiation dates; calculated difference between dates
distribution_pattern	Places in the U.S. where recalled food products were distributed
*general_reason	General reason for recall; parsed text from reason_for_recall variable to generalize reasons
initial_firm_notification	Recall notification method
reason_for_recall	Statements on reason for recall
recall_initation_date	Date when recall was first notified
recall_termination_date	Date when recall investigation is done
state	State location of firm
*usda_region	Food regional business center; classified state to a usda_region
voluntary_mandated status	Recall status indicating product was either voluntarily reported by firm or mandated by FDA

References

U.S. Department of Agriculture. (n.d.). *USDA Regional Food Business Centers Program.* https://www.ams.usda.gov/services/local-regional/rfbcp

U.S. Food and Drug Administration (FDA). (2024, May 6). *Enforcement Reports.* https://www.fda.gov/safety/recalls-market-withdrawals-safety-alerts/enforcement-reports

U.S. Food and Drug Administration FDA). (2023, May 12). *Enforcement Report Information and Definitions.* https://www.fda.gov/safety/enforcement-reports/enforcement-report-information-and-definitions#report_label

U.S. Food and Drug Administration (FDA). (n.d.). *Food Enforcement Overview.* https://open.fda.gov/apis/food/enforcement/