World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:10, No:6, 2016

Spatially Random Sampling for Retail Food Risk Factors Study

Guilan Huang

***Abstract*—**In 2013 and 2014, the U.S. Food and Drug Administration (FDA) collected data from selected fast food restaurants and full service restaurants for tracking changes in the occurrence of foodborne illness risk factors. This paper discussed how we customized spatial random sampling method by considering financial position and availability of FDA resources, and how we enriched restaurants data with location. Location information of restaurants provides opportunity for quantitatively determining random sampling within non-government units (e.g.: 240 kilometers around each data-collector). Spatial analysis also could optimize data-collectors' work plans and resource allocation. Spatial analytic and processing platform helped us handling the spatial random sampling challenges. Our method fits in FDA's ability to pinpoint features of foodservice establishments, and reduced both time and expense on data collection.

***Keywords*—**Geospatial technology, restaurant, retail food risk factors study, spatial random sampling.

I. Introduction

ENSURING food safety is an important public health priority for our nation. The U.S. FDA has developed several programs applying risk-based methods. To quantitatively measure program effectiveness, it is necessary to collect sample data to first calculate a baseline, and then take additional samples during the study-timeline [1], [2]. Using the 1998-2008 ten-year study as a foundation, in 2013, FDA obtained approval to initiate the first phase of the study, which focuses on data collection within the restaurant segment. This study was designed to examine patterns of the occurrence of foodborne illness risk factors within the restaurants using multiple data collection periods. The study data collection focuses on the control of foodborne illness risk factors in fast food restaurants and full service restaurants. It is not intended to be a comprehensive assessment of compliance with all Food Code requirements. The data collector's priority is to observe food safety practices and behaviors associated with risk factors that have been epidemiologically linked to the occurrence of foodborne illness outbreaks at the retail level.

The entire population of restaurants in USA is too large for us to attempt to collect data from all restaurants. There are no strict rules to follow, and we need to choose samples which could accurately represent the population of the establishments, and at the same time be concerned of any potential limitations [3], [4]. We face two major challenges: (1) no existing ready-to-use database sources which conform to our definitions of full service restaurant and fast food restaurant; (2) our financial and logistical constraints limit our ability to perform a random selection of all facility types in the United States within the same timeframe. We need to find an alternative method that is cost-effective while still sufficient to the extent that the facilities in the sampling zones are representative of the overall restaurants industry.

Guilan Huang is with the U.S. Food and Drug Administration, College Park, MD 20740 USA (phone: 240-402-2904; e-mail: guilan.huang@fda.hhs.gov).

II. Data Preparation and Sample Size

A. Restaurants Definitions

This study was intended to examine food safety practices in restaurants that conduct a significant amount of on-site food preparation. In each restaurant data collectors need to observe and record food safety practices and behaviors, and make assessment of food product temperatures, employee health policies, handwashing frequency, food protection manager certification, and food safety management systems. Restaurants that serve only pre-packaged food items or only operated seasonally, or otherwise conducted low-risk food preparation activities were excluded from our selection.

For the purposes of this study, the restaurant segment is sorted into fast food and full service restaurants. We define full service restaurants as establishments where customers place their order at their table, are served their meal at the table, receive the service of the wait staff, and pay at the end of the meal. Buffets are considered as full service restaurants, though most buffets in USA are serving meals in which food is placed in a public area where the diners generally serve themselves, and diners pay before they eat the food. Fast food restaurants are also referred to as quick service restaurants and are defined as any restaurant that is not a full service restaurant.

B. Geospatial Data of Restaurants

In order to know how many fast food and full service restaurants there are, and where they are located. After online searching and contacting various data providers, we decided to use U.S. Business Location data from Esri as our primary source of restaurant geodatabase. The data is extracted from a comprehensive list of businesses licensed from Infogroup. Infogroup conducts annual telephone verifications with each business listed in the database. The business list contains name, address, franchise code, industrial classification code, number of employees, and sales. If businesses have street address, Esri geocoded the addresses, and assigned latitude and longitude coordinate to the business site; otherwise, Esri use other geographical unit (such as zip code, census block

unit) center latitude and longitude coordinate to the business site. The quality of the local address system varies; address matching is better in urban areas that use street-level address systems than in rural areas [5]. The U.S. Business Location data is a point-level data set, and it classifies businesses by the SIC and NAICS industry classifications. In addition to the typical SIC and NAICS codes, the database also includes Infogroup's proprietary six-digit SIC and eight-digit NAICS industry codes and a special industry code for some select industries.

After conducting geospatial analysis on the 2010-2012 U.S. Business Location data, we found that restaurants are sometimes misclassified, and restaurants change (e.g.: close, new open, change name, move to a new location…) frequently. Additionally, our definitions for fast food and full service restaurants do not exactly fit the SIC and NAICS classification criteria.

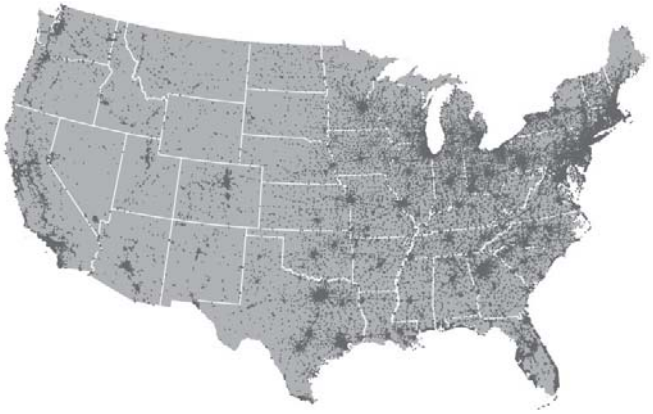

Fig. 1 Distribution of fast food restaurants (each dot represent a fast food restaurant establishment)

Fig. 2 Distribution of full service restaurants (each black square represent a full service restaurant establishment)

For extracting listings for fast food restaurants and full service restaurants, we created 3-tier filters to extract fast food restaurants (Fig. 1) and full service restaurants (Fig. 2) from the U.S. Business Location data. The 3-tier filters are derived from (1) industry classification and sub-classification, such as four-digit SIC code, six-digit SIC and eight-digit NAICS industry codes; (2) keywords in the restaurant name, such as buffet, express; and (3) our name sets from our previous three data collection periods (1998, 2003 and 2008) and our team members' knowledge accumulation. Our filters are continuously assessed and improved throughout study.

C. Sample Size

There are no strict rules to follow to determine the sample size, though sample size is an important feature of any study. Our survey is designed to use sample size results to estimate population status. In this case, we need *t*-score to calculate the confidence interval, but *t*-score depends on both the degrees of freedom and the desired confidence level, and sample size affects the degrees of freedom. However, when sample size is larger than 30, the value of *t*-score is quite close to the value of *z*-score, so often we ignore the distinction between the normal and *t*-distribution.

The minimum sample size of probability sampling can be calculated if we have the margin of error, the confidence level, the population size, and the response distribution. The sample size doesn't change much for populations larger than 200,000. For this study, we decided that we need at least a 95% confidence level, and we could accept a 5% margin of error, so the sample size is 384 based on sample size calculator of Raosoft [6]. In other words, we need to collect data from at least 384 fast food restaurants and at least 384 full service restaurants.

III. Customize Spatial Random Sampling Method for Our Study

We usually pay more attention to the sample sizes than spatial distribution of samples. The purest form of probability sampling is simple random sampling. It is a method that each eligible establishment has an equal chance of being selected. Each restaurant facility of the population has a known non-zero probability of being selected. Simple random sampling is not the most statistically efficient method of sampling because it treats the entire area as a whole and as a result, certain subgroups (some states may have no sample) may be left out.

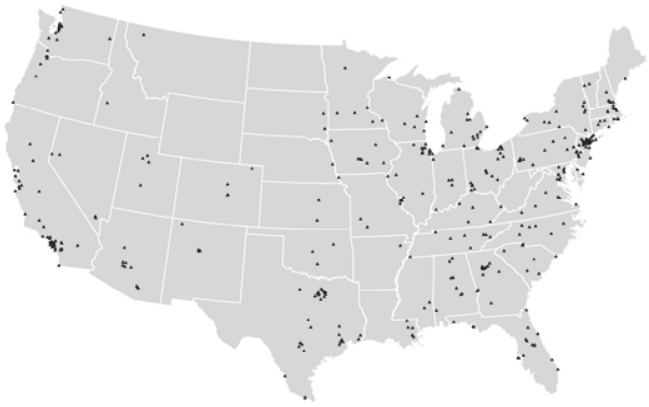

Fig. 3 Distribution of 384 samples of fast food restaurant (each black triangle represents a selected fast food restaurant)

In Fig. 3, we sampled 384 fast food restaurants via simple random sampling. We found that states such as North Dakota, West Virginia, and Wyoming were left out; in addition to that, states such as Arkansas, Idaho, Maine, Montana, and Nebraska only have one sample from the entire state.

In Fig. 4, we sampled 384 full service restaurants via simple random sampling. We found that states such as Mississippi, North Dakota, Vermont and Kentucky were left out; in addition to that, several states only have one sample from the entire state.

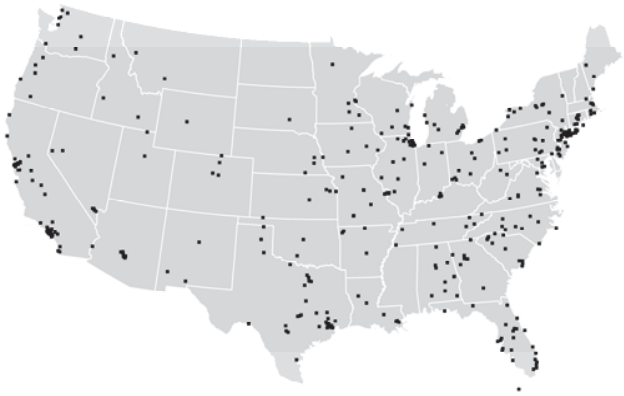

Fig. 4 Distribution of 384 samples of full service restaurant (each black square represents a selected full service restaurant)

For this study, 22 FDA regional retail food specialists conducted data collections from the restaurant facilities. The specialists are geographically dispersed in the United States; especially in relatively high density population centers (Fig. 5). To estimate the travel cost, we used geo-technology tool (e.g.: ArcToolbox) to measure the spatial distances between samples and specialists, we found that simple random spatial sampling will cost huge amount of money than we could have. Besides simple random sampling, systematic sampling (either on grids or on transects [7], [8]) method and stratified sampling method (subset: state-level) also cost huge amount of money and time than we could afford. Convenience sampling method is an inexpensive but only can get a gross estimate of the results which could not meet our study goal.

As we all know, the actual sampling method and sample size are frequently influenced by the cost of data collection. During 2013-2014, we had a limited travel budget. With the help of geospatial analysis, we developed a cost-effective (aka no-flight) sampling method. Generally speaking, specialist can travel approximately 240 kilometers (150 miles) to collect data, and be back home on the same day. After performing spatial analysis, we found that about 62% of restaurants are within 240-kilometer buffer zones around specialists' home zip code center (to protect specialists' privacy, we use their home zip code instead of home address). Therefore, our spatial coverage was based on the specialists' geographical areas of responsibility and provided a reasonably convenient method of estimating national risk-related behaviors and practices.

Our restaurant listings have governmental units (such as state, city and zip code) attribute. In the open data era, we could easily obtain the boundaries of country, state/province, and city, but there are no ready-to-use non-overlap 240-kilometer buffer zones for our 22 specialists.

Fig. 5 Distribution of 22 specialists (each circle represents a specialist)

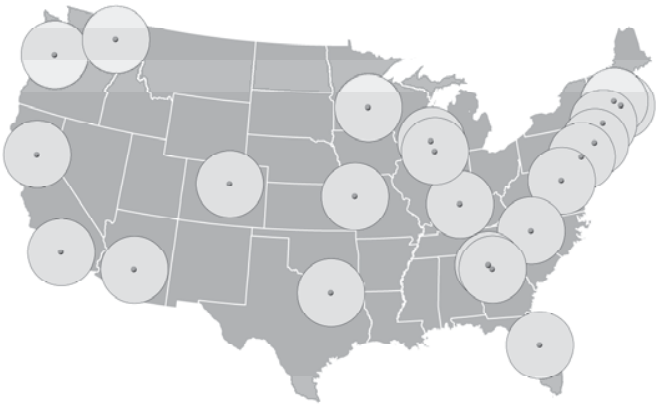

Fig. 6 Distribution of 22 specialists and 240 kilometers buffer zone around their home zip code centers

If we simply buffer 240-kilometer around each specialist's home zip code center, we will find that there are several overlaps between specialists (see Fig. 6). In order to create non-overlap 240-kilometer buffer zones, we used ArcGIS to create Thiessen polygons based on 22 specialists' home zip code centers [9]. Thiessen polygons boundaries define the area that is closest to each specialist relative to all other specialists (Fig. 7). We split overlapped 240-kilometer buffer zones via these Thiessen polygons, and thus creating 22 non-overlap 240-kilometer zones (Fig. 8).

Once we have the geometrical boundary of the buffer zones, we add specialist's name and home zip code as attributes, and then we spatially join with our fast food restaurants and full service restaurants, respectively. We are ready to perform random selection within each non-overlap 240-kilometer zone.

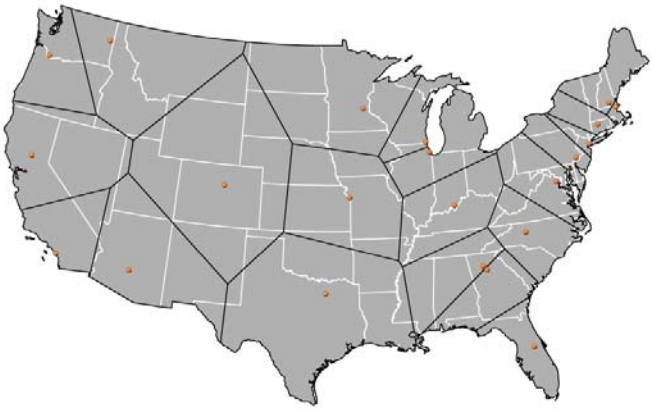

Fig. 7 Thiessen polygons for 22 specialists

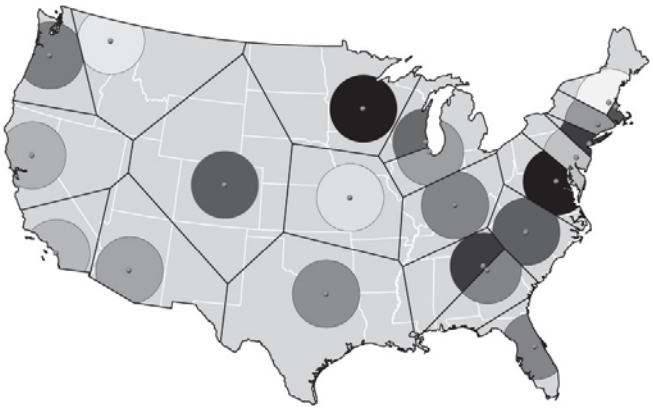

Fig. 8 Distribution of 22 specialists and their non-overlap 240-kilometer buffer zones

IV. The Results

To balance each specialist's workload and avoid data bias, the sample establishment inventory was evenly distributed among the specialists. Each of the 22 specialists was expected to collect data in 19 full service and 19 fast food restaurants for a total of 38 data collections per specialist. This sample size provides sufficient observations of food safety practices to be 95% confident that compliance percentages derived from the data collections are within 5% of their actual occurrence. Considering that randomly selected establishment may be misclassified, closed, or otherwise unable or unwilling to participate this study, we also sampled substitute establishments for each specialist.

Within fast food restaurants listing, we used the name of each specialist as unique identifier to do random selection and selected 19 samples for each specialist with the help of Hawth's tool [10]. Then we repeat the same spatial random sampling method for full service restaurants. Fig. 9 displayed the spatial distributions of 19 selected fast food restaurants, and 19 selected full service restaurants assigned to each specialist.

Geo-technology allows specialists to make work plan more efficiency. For example, he/she could combine data collection with other official tasks. When the specialists finish their data collection, we geo-visualize our results in map format to detect any potential errors.

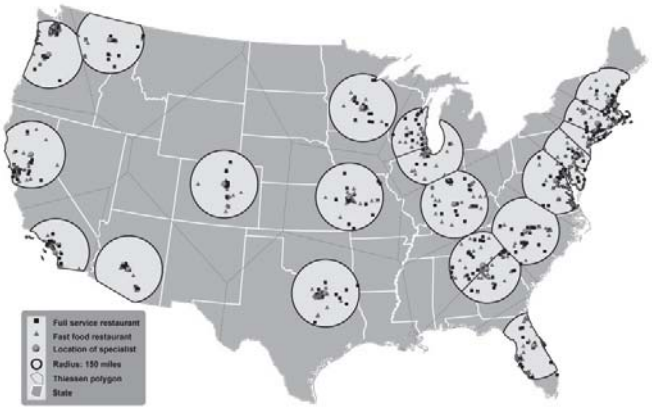

Fig. 9 Distribution of 22 specialists and their 19 fast food restaurants and 19 full service restaurants

V. Conclusion

Geospatial technology played an important role in customizing spatial random sampling for this study. This method will deliver more precise and accurate baseline information for the planned 2017 and 2021 restaurant data collection. Geospatial technology will help us quantitatively measure the progress of food safety policy/program, and it also optimizes the data-collector's work plan.

The data from this period and future restaurant data collection periods planned for 2017 and 2021 are expected to provide input into the Health People 2020 Food Safety Objective FS-6. This objective is designed to improve food preparation practices and food employee behaviors in restaurants. This study will be used to provide FDA research information that will assist the agency in developing retail food safety initiatives and policies focused on the control of foodborne illness risk factors.

Acknowledgment

This study is part of the FDA 2013-2024 National Retail Food Risk Factors Study Project. The suthor would like to thank my colleagues: Marc Boyer, Kevin Smith and John Marcello, who provided invaluable insight and expertise that greatly assisted the study and this paper.

References

[1] FDA National Retail Food Team, "FDA Report on the Occurrence of Foodborne Illness Risk Factors in Selected Institutional Foodservice, Restaurant, and Retail Food Store Facility Types (2009)", http://www.fda.gov/downloads/Food/FoodSafety/RetailFoodProtection/FoodborneIllnessandRiskFactorReduction/RetailFoodRiskFactorStudies/UCM224682.pdf

[2] FDA National Retail Food Team, "FDA Trend Analysis Report on the Occurrence of Foodborne Illness Risk Factors in Selected Institutional Foodservice, Restaurant, and Retail Food Store Facility Types (1998-2008)", http://www.fda.gov/downloads/Food/GuidanceRegulation/RetailFoodProtection/FoodborneIllnessRiskFactorReduction/UCM369245.pdf

[3] D.M. Theobald, D.L. Stevens, Jr., D. White, N.S. Urquhart, A.R. Olsen, and J.B. Norman, "Using GIS to generate spatially-balanced random survey designs for natural resource applications," in *Environmental Management*, 40(1), pp. 134-146, 2007.

[4] D. Walonick, "Survey Sampling Methods", http://www.statpac.com/surveys/sampling.htm
[5] An Esri White Paper, "Business Location and Business Summary Data", http://doc.arcgis.com/en/esri-demographics/data/business.htm
[6] "Sample size calculator", http://www.raosoft.com/samplesize.html
[7] R.M. Lark, "Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood", in Geoderma, 98(1-2), pp. 35-59, 2000
[8] Z. Zhu, M.L. Stein, "Spatial sampling design for parameter estimation of the covariance function", Journal of Statistical Planning and Inference, 134(2), pp. 583-603, 2005.
[9] Environmental Systems Research Institute, Inc, "Create Thiessen Polygons", http://pro.arcgis.com/en/pro-app/tool-reference/analysis/create-thiessen-polygons.htm
[10] H Beyer, "Hawth's Analysis Tools for ArcGIS: Random Selection within Subsets", http://www.spatialecology.com/htools/rndselss.php