Collecting & Analyzing Big Data for Social Sciences
Critical Analysis of Research Articles

Food Deserts in the United States

20 June 2022

Group 42
Anja Deric (r0873512)
Peter Day (r0866276)

Article 1 Summary

Lois Wright Morton et al[1], explore ways people make up for food insecurity in rural United States. Through consolidation, rural counties have lost many grocery stores, leaving many rural people without a source of food in their community. This paper investigates whether community and family/friends can act as a reliable source of food and make up for living in a food desert.

They selected two rural counties in Iowa with relatively high poverty levels. First, focus groups were conducted in the two counties, from which they constructed a mail survey, which got a 60.1% response rate. Their dependent variable is food insecure/food secure. They have four variables representing personal connections and local civic structure. These were controlled with age, income and education. These three variables have a well studied connection to food insecurity. The study asked six questions which were then translated to three classes of household food security: food secure, food insecure, food insecure with hunger.

A number of questions were then asked about alternative (not grocery stores, restaurant) sources of food. The first was whether the respondent had given food to family, friends or neighbors. They were then asked whether they had given food to people they did not know through a food bank, food drive or senior meal program. There were then three questions about receiving food from similar sources. Lastly there were seven questions about civic/governmental involvement in food distribution.

Three logistic regression models were fit to determine associations with food security. The first regressed upon age, income, education and whether the respondent lives in a town. The second model adds personal connections and the last model adds civic structure. None of the personal connection variables in model 2 showed an association with food security. The paper does find an association with perceptions of civic structure, with a higher perceived civic structure reducing the odds of food insecurity.

One limitation of this study is its narrow focus on two counties in one state. Other rural areas in the country should be included in a further study. In addition, the civic structure was a measure of individual perceptions and should be expanded to include other measures. Further research needs to be done on the precise association between food deserts, lack of quality food sources, and food insecurity on an individual level.

Article 2 Summary

The second article, [Widener and Li, 2014][2], explores twitter data in the context of food deserts. The main objective and research question of the article is to explore if and how social media data can be used in understanding topics relating to public health, specifically focusing on food deserts and how Twitter data could be leveraged to identify food deserts in the US by identifying areas where tweets about unhealthy foods are more prevalent.

The data for this study was collected using Twitter's API. A list of terms relating to healthy and unhealthy foods was used to extract tweets which mention those particular foods. Before they are

---

[1] LW Morton, EA Bitto, MJ Oakland, and M Sand. Solving the Problems of Iowa Food Deserts: Food Insecurity and Civic Structure. *Rural Sociology*, 70(1):94–112, MAR 2005. ISSN 0036-0112. doi: 10.1526/0036011053294628

[2] Michael J. Widener and Wenwen Li. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography*, 54(SI):189–197, OCT 2014. ISSN 0143-6228. doi: 10.1016/j.apgeog.2014.07.017.

entered into a database, the tweets are further filtered and any 'noisy' tweets which may not be related to food (such as those using the term 'apple' to refer to the technology company, Apple) are discarded. Sentiment analysis is then conducted using the AlchemyAPI on the remaining ~140,000 tweets and they are labeled with a sentiment score ranging from -1 (negative sentiment) to +1 (positive sentiment).

To analyze the data, the authors used some more simple methods such as creating frequency tables and using difference of proportions test to determine if the proportion of tweets about unhealthy or healthy foods differs between different areas of the US. For this analysis, areas of the US from which the tweets originate are classified as either low-income low-access (food desserts) or not based on data from the United States Department of Agriculture (USDA). The authors also used logistic regression to assess if variables like median age, proportion of males, proportion of black individuals, low income, and low access have a significant effect on the sentiment score of tweets.

Several data and method limitations are identified in the paper. Only around 2.5% of all tweets collected were georeferenced, while the rest had to be discarded. Additionally, Twitter usage is elective, which biases the sample and makes it difficult to get a representative data set (for instance, Twitter usage is lower in low-income areas). A method-related limitation includes the reliability of sentiment analysis as detecting sarcasm, for example, and correctly interpreting user-generated content is an imperfect process.

Critical Analysis of Articles

Both articles address the topic of food deserts in the United States, but take different approaches to research design. Article 1 takes a more traditional approach of collecting and analyzing survey data, while Article 2 takes a more liberal and creative approach by leveraging social media as a data source. The design methodology, though different, fits well with the aim/objective of each respective study. Survey data is used to answer a very specific/focused question regarding food access, while Twitter data is used for a more general public opinion analysis.

Though Article 2 has a more creative approach to data collection and analysis, a similar design would not be effective in answering a research question such as the one posed in Article 1. Twitter data often lacks formality, standardization, as well as georeferencing, so a similar design would not be appropriate for studies that need accurate and representative public opinions relating to a very specific set of questions. Given that Tweets have no formal requirements or structure apart from staying under a certain number of characters, gathering enough georeferenced and unbiased opinions relating to a very specific topic (such as food insecurity) is an unlikely scenario. Furthermore, when it comes to more private/personal topics such as food security (or even disclosure of race, age, income), an anonymous and private survey would not only be more ethically sound for data collection, but it would be nearly impossible to find such specific data on an online public social media platform.

While survey data would be more appropriate for answering very specific research questions, survey studies are more difficult and more costly to replicate. Survey data collection can be a very extensive process, and there is no guarantee that an identical sample can be obtained for two studies. People can move, pass away, or even just choose not to respond, so a replication of such a study would be a very complex endeavor. With Twitter data, on the other hand, replication would not only be easy, but also much cheaper. The Twitter API allows developers to specify a time frame for tweets they would

like to collect (as well as many other parameters), meaning that an exact replication of a study could easily be completed and the study could easily be expanded in the future without any overhead costs.

Another issue that both articles are affected by includes bias and sample selection. Article 1 collects survey data from two rural towns in Iowa, making it difficult to generalize the results to the rest of the US population, or even within the state of Iowa. Additionally, both towns included are in rural areas, so the results of the study would likely not be representative of people living in food deserts in urban areas. Though such a small sample simplifies the design greatly, it also makes it impossible to make claims for any territories outside of the two towns that were selected, which significantly narrows the impact of the study.

With Article 2, the data collection process is biased simply due to the fact that not everyone in the US has a Twitter account. In fact, certain races and age groups are significantly more likely not to own a Twitter account, resulting in them being excluded and not represented in the study. Even for individuals who are on Twitter and choose to engage with this social media platform, not every user tweets about the topic of our interest (food). It's additionally likely that only people with strong opinions (strong like or dislike of a food) would choose to create tweets about food, while those with neutral opinions would most likely not have dedicated food-related tweets. It seems possible that users are more likely to tweet about a restaurant dinner than the everyday salad they had for lunch. Furthermore, as the article mentions, only around 2.5% of Twitter data collected was georeferenced, meaning that a large portion of the data had to be scrapped just due to lack of location referencing.

Ultimately, the data collected in both studies face the issue of bias. However, one advantage of Article 1 is that this bias can be pinned down and claims can be made for the two towns that the data were collected for. Future studies could also expand to include other states and urban settings in order to reduce this bias. With Article 2, on the other hand, the presence of bias is definite but extremely difficult to pinpoint and define, so none of the results can be generalized to any specific populations or geographical locations. Expanding to include data from other social media platforms may slightly reduce this bias, but complete elimination is not possible given the data sources as previously described.

In terms of analysis techniques, both articles employ logistic regression as a means of predicting food insecurity and healthy/unhealthy food sentiment. The regression models, in both cases, include a small sample of predictors, making them very easy to construct and interpret. However, this also means that some significant demographic or economic indicators may not have been included as part of the analysis. Article 1's analysis, for instance, could potentially benefit from including additional indicators of civic structure that are more general and unbiased as opposed to relying on measures of individual perception. Article 2 could consider variables such as education level breakdown or health-related statistics (such as obesity rates) for the georeferenced location in the analysis to potentially further explain the variation in sentiment scores.

In addition to logistic regression, Article 2 also employs NLP as an advanced analysis technique. This is a more novel technique but does not guarantee 100% accuracy as it attempts to interpret human language. NLP algorithms can have very high accuracy, however some language tools such as sarcasm or metaphors may not be recognized, thus leading to data misinterpretation. Furthermore, Article 2 does not discuss whether all the collected tweets were in English or not, but a large part of the US population are Spanish speakers and performance of the same NLP technique on data in multiple languages is questionable. This also extends to cases where a lot of slang or regional terminology is used.