**The Implementation of Cellular Location Data Analytics on the Analysis of Food Accessibility Issues in the Washington, DC and Baltimore, MD Metropolitan Areas**

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirement for the Degree
Bachelor of Science, School of Engineering

Liam Billings
Spring, 2021

Technical Project Team Members
Natasha Foutz
Weiguang Wang

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signature_Liam Billings_____Date_05/09/2021

Approved_Natasha Zhang Foutz, McIntire School of Commerce_____Date_05/12/2021

The use of cellular location data in demographic studies offers a series of advantages in both responsiveness and accuracy. Given the high availability of user location tracking systems, whether they are associated with conventional marketing groups such as Google or Facebook or more specialized firms such as X-Mode enable the collection of activity associated with tens or hundreds of thousands of possible citizens ("About Us"). While many groups do use such data for advertisement tracking such as with Facebook's efforts (Jolly, 2018), there are some cases where academic institutions or other groups can obtain access to such cellular data for more humanitarian goals. This work which was performed with the assistance of the McIntire School of Commerce is one of a series of projects leveraging cellular location data to conduct demographic studies on American metro areas. The differences associated with this work in comparison to previous projects are the computing optimizations which were implemented for cellular location data analytics and the focus of this work, which was both the Washington, DC and Baltimore, MD metro areas in a food desert study. The original intent for this work was a food desert study focused on New York, NY leveraging current location data but because of other concurrent research projects on food deserts leveraging cellular location data analytics and the lack of current data on Washington, DC and/or Baltimore regarding food deserts, the study focus was changed to both metro areas. Through the implementation of novel computational approaches such as the infomap algorithm, the use of two food store case studies, and various confidential resources related to food desert/cellular location data analytics, highly descriptive information on the food environments of the Washington, DC and Baltimore, MD metro areas was captured. The result was a highly descriptive dataset which could be used to influence food accessibility policy in at least both the Washington, DC and Baltimore, MD metro areas.

## Washington, DC and Baltimore, MD Metro Areas: Data Preprocessing

Since September of 2020, the focus of the project was changed to April-September 2019 for both the Washington, DC and Baltimore, MD metro areas, as the associated business activity increases during that period could provide similar information to the direct study of 2020 data without creating redundancies with concurrent McIntire School research projects utilizing the same dataset. The starting dataset from April to September 2019 represented a population of 1,423,590 users for the Washington, DC metro area and 939,877 users for the Baltimore, MD metro area. This data was then filtered for residence within a specific coordinate bound representing each metro area (redacted for compliance with data source confidentiality agreement) with additional limitations placed on the speed of individuals (maximum speed 10 meters/second), the residency of individuals (at least 10 days per month for all months) and the amount of time an individual is considered idle (at least 60 seconds). Following the implementation of such restrictions, the user counts for both the Washington, DC and Baltimore, MD metro area datasets were reduced to 156,046 and 68,651 users, respectively. Then, a list of device identifications (IDs) for each metro area was created representing users who were present in the metro area after filtering for all six months for at least ten days of each month. A sample size of 11,700 for the Baltimore, MD metro area dataset and 29,105 for the Washington, DC metro area dataset were identified, with 297 device IDs being present in both datasets. These filtering methods were used to ensure that a critical analysis can be more useful: the infomap stop identification algorithm.

The infomap algorithm can effectively identify unique instances wherein a user stops and provide a series of quasi-unique numerical IDs to reference each stop (Aslak and Alessandretti, 2020, p. 1). This algorithm is also more efficient at identifying stop placement than other approaches such as density-based spatial clustering of applications with noise (DBSCAN)

meaning that more cellular location data can be processed with equivalent computing resources (Aslak and Alessandretti, 2020, p. 3). As these IDs can be inconsistent and running this model on subsets of ~1000 is necessary for computing speed, they are supplemented with the centroid for each label and the amount of time a person has stopped at the label per stop event. The centroid provides an alias of a numerical ID to a point with latitude and longitude coordinates which can be used to map the stopping zones identified with the infomap algorithm. Additional modifications were made to the implementation of the infomap algorithm to improve performance and place ID precision. The sizing of place ID zones was set to 50-100 meters, with the maximum number of significant figures for coordinates set to 5, which effectively limited the precision of the infomap algorithm's stop identification process to 1.1 meters. The time interval for which a stop could be valid was set to start at five minutes and end at 12 hours.

Following the infomap execution, home locations based on the mode of infomap place IDs within a specific date and time range (Tuesday, Wednesday, Thursday, and Friday, from 21:00 Eastern Standard Time (EST) on the previous day to 07:00 EST on the current day) were identified. Efforts to recover users which failed this process by having their mode place ID be an invalid ID were also conducted by reprocessing these users specifically through the infomap algorithm and conducting the same mode analysis procedure. Of the 29,105 users found to be six-month residents of the Washington, DC metro area, 17,332 were found to have identifiable home locations. For the Baltimore, MD metro area, of the 11,700 users found to be six-month residents of the Washington, DC metro area, 7,429 were found to have identifiable home locations. An effort was made to recover some additional users with identifiable home information from the users which previously lacked such information by grouping the users with no identifiable home location (11,799 for the Washington, DC metro area and 4,283 for the

Baltimore, MD metro area) and obtaining stop IDs by applying the infomap algorithm to these users exclusively, but very few users were recovered from these efforts. Only 31 users of the 11,799 Washington, DC metro area users could be recovered and only 12 of the 4,283 Baltimore, MD metro area users could be recovered. The primary reason this effort had a high failure rate was because the users which were reevaluated generally had significantly fewer location logs than users which had successfully identified home IDs. The quantity of logs impacts the success of the home identification process as users with fewer logs generally had less location data with valid place IDs in general. For the Washington, DC metro area data, a user with an identifiable home location had on average 1.44 MB of location data, while a user without an identifiable home location had on average 859.83 KB of location data, which is only 58.31% of the data associated with a home-identifiable user. For the Baltimore, MD metro area data, a user with an identifiable home location had on average 1.52 MB of location data, while a user without an identifiable home location had on average 508.93 KB of location data, which is only 32.70% of the data associated with a home-identifiable user. A small subset of users was found to have no data related to a home location (492 six-month residents of the Washington, DC metro area and 178 six-month residents of the Baltimore, MD metro area).

Given the small number of users available for Baltimore specifically, all users who were present in the Baltimore, MD and/or Washington, DC metro areas for at least one month were selected for cellular data processing with the infomap algorithm and subsets of this data were created for users present for two to five months. As the target user count for each metro area was 20,000, the six-month Washington, DC metro area user set consisting of 16,817 users and the 3-month Baltimore, MD metro area user set consisting of 20,979 users were used.

**Identifying User Demographics, Food Access, and Store Activity**

To rapidly determine the food access circumstances of both user activity and store placement, two different food access metrics were used. The first metric is an indicator for if a US Census tract is a food desert. Provided by the US Department of Agriculture (USDA), this metric indicates if there exists a grocer within half a mile of the general population in the census tract from the Food Access Research Atlas (FARA) in an urban area or if a grocer exists with ten miles of the general population for a rural area census tract (U.S. Department of Agriculture). The second indicator is a composite known as the modified Retail Food Environment Index (mRFEI) which is constructed for each census tract consisting of the quantity of healthy food retailers divided by the sum of all food retailers and is generally represented as a percentage (Gustafson et al., 2012, p. 3). Specifically, the mRFEI is constructed from the same data source as the cellular data. This metric can indicate with more granularity the quality of food access and is not bound to the same assumption that the USDA metric is presuming that the presence of a grocer alone will nullify the impact of unhealthy locations (Gustafson et al., 2012, p. 2). However, this metric is less complete than the USDA's food desert metric, which is present for every census tract region within both the Washington, DC and Baltimore, MD metro areas. Specifically, of the 418 census tracts for the Washington, DC metro area and 216 census tracts for the Baltimore, MD metro area, this food swamp metric is missing for 88 and 9 of the tracts for each metro area, respectively. This issue impacts 21.05% of users with home locations in the Washington, DC metro area and 4.17% of users with home locations in the Baltimore, MD metro area of the user sample sets of 16019 and 20039 users, respectively.

The cellular data of these users with movement speeds below 10 meters per second and an idle time of 60 seconds or more was then processed to identify the number of days the set of users for the Washington, DC and Baltimore, MD metro areas visited 462 stores, with 61 of

these stores being grocers and the remaining 401 stores being restaurants. This process assumed that the user visits a given location only once per day the user was identified within the store's radius and that the user was idle for long enough to have an idle time of 60 seconds or more. These limitations are why these user encounters with stores are being measured as visit days instead of visits. Different radii were used based on the store type, with restaurants using a 30-meter radius and grocers using a 70-meter radius which a user must be present in for a visit to a store to be counted, with this distance being calculated using a Euclidean approach given the small distances associated with the radii. All stores were chain restaurants common to the East Coast of the United States and/or the United States as a whole. A total of 701,384 visit days were identified from this user set, with a plurality of these visits being related to unhealthy restaurants at 333,678 visit days as other visit types such as healthy restaurant visit days (194,050) and grocer visit days (173,656) being less prominent.

Given the importance of the granularity of the mRFEI metric, the user count was reduced to remove all users with home locations in a US census tract without mRFEI data. With a new user count of 12,586 users for the Washington, DC metro area and 18,924 users for the Baltimore, MD metro area, respectively. Similarly, the total amount of unique store visits was reduced to 615,888 unique visits, with a plurality of these visit days being to unhealthy restaurants (294,545) and reduced prominence of healthy restaurant visit days (172,910) and grocer visit days (148,433). Many of the visits to all store types were also from users with home locations in food deserts, with 67.29% or 414,419 visit days being from such users. Specifically, the highest rate of user swamp visit days was with healthy restaurants (70.25% or 121,466 visit days of 172,910 visit days), with unhealthy restaurants (67.29% or 198,208 visit days of 294,545 visit days) having reduced prominence and grocers (63.83% or 98,745 visit days of 148,443 visit

days) having further reduced user swamp visit day margins. The reduction of users and user activity was found to be too severe at a total user count reduction of 16.63% with a user count reduction of 25.16% for the Washington, DC metro area, so the mRFEI metric was discontinued and greater emphasis was placed on extracting analytics relevant to the FARA food desert indicator.

## Initial Statistics Collection

With the foundational elements of the datasets complete, the collection of statistics which could better describe the dataset were collected. Firstly, the store count per capita was collected for all three food store types for each region. Secondly, metrics on visitor counts and visitor-days were collected with subdivisions for user type (food desert resident vs non-food-desert resident) and food store residency (food desert resident vs non-food-desert resident). Additional metrics were generated identifying the count of visitors and visitor-days per store. Percentages were then calculated for each food store and averaged on the visitor and visitor-day distributions of users from food deserts and users not from food deserts. Additional averages were generated using only food stores present in food deserts and food stores which were not present in food deserts. Statistics which were measured per citizen were also generated such as the stores visited and visitor-days which were generated with an average of aggregate visited stores or visitor-days. Like the previously covered percentages, metrics were also generated with both user-level and store-level subdivision.

A collection of minima, maxima, and median statistics of travel distances from users to food store. These metrics began with a set of minima, maxima, and median distances using the metrics of all users for each visited store type, including a visited store only once. Then, metrics were compiled using aggregates of minima, maxima, and median statistics generated for each

store, firstly identifying minima, maxima, and medians for each food store type using all of the distances users traveled to food stores. Additional statistics were computed using the minima, maxima, and medians computed for each store individually which were then averaged to obtain the minimum, maximum, or median distance for each food store type. These statistics were also computed using both food store subdivision (desert-resident vs non-desert-resident) and user subdivision (desert-resident vs non-desert-resident). Then, statistics which were averaged per visitor to each food store type were calculated with only user subdivision (desert-resident vs non-desert-resident). Lastly, statistics with user subdivision (desert-resident vs non-desert-resident) using the same per-user calculation method were generated with weights to each store proportional to the visit count for each visit distance.

**The Washington, DC Metro Area Case Studies**

With the conclusion of these analyses, more detailed metrics were collected on the Washington, DC metro area dataset regarding the two store opening cases. The first store was Trader Joes: a healthy food store which opened on May 2$^{nd}$, 2019 in an area in the Washington, DC metro area not considered a food desert. The second store was Chick-fil-A: an unhealthy food store which opened on July 19$^{th}$, 2019 in an area in the Washington, DC metro area considered to be a food desert. For both case studies, only the user activity of users within a three-mile radius of these stores was considered and activity to food stores was considered within both three-mile and ten-mile radii of these stores. These distance-based limits were used to measure impacts to food store activity for users local to these stores before and after these stores opened for groupings of stores which were at varying levels of locality with the opened store. This is similarly reflected by the choice of the two store opening cases, which do not overlap in date range for 30 days before and 30 days after each store opening. Such limits for user count

restricted analysis to a population of 2,683 users for the Trader Joes case study, with 77 of these users being resident in a food desert and 2,606 of these users not being resident in a food desert. Similarly, the user count for the Chick-fil-A case study was restricted to 1,958 users, with 1.037 of these users being resident in a food desert and 921 of these users not being resident in a food desert. The three-mile radius around each store opening restricted the store sample to eleven grocers (all were not in food deserts), 22 healthy restaurants (all were not in food deserts), and 31 unhealthy restaurants (two were in food deserts, 29 were not in food deserts) for the Trader Joes store opening. For the Chick-fil-A case study, this three-mile radius restricted the food store set to five grocers (two were in food deserts, three were not in food deserts), ten healthy restaurants (four were in food deserts, six were not in food deserts), and 25 unhealthy restaurants (seven were in food deserts eighteen were not in food deserts). The ten-mile radius for the Trader Joes case study enabled an expanded store set of 48 grocers (five were in food deserts, 43 were not in food deserts), 86 healthy restaurants (fifteen were in food deserts,71 were not in food deserts), and 158 unhealthy restaurants (32 were in food deserts, 126 were not in food deserts). Similar increases in food store counts were observed for the set of stores within ten miles of the studied Chick-fil-A, with 39 grocers (three were in food deserts, 36 were not in food deserts), 82 healthy restaurants (20 were in food deserts, 62 were not in food deserts), and 148 unhealthy restaurants (37 were in food deserts, 111 were not in food deserts).

For these sets of users, a discrete set of food store data had to be generated for the time range because the previous food store activity measurements were cumulative for the April to September 2019 period, while the current metrics required data over four date ranges. These date ranges matched with the 30-day buffer before and after each store opening and included April 2nd to May 1st, 2019 (30 days before the opening of Trader Joes), May 2nd to May 31st, 2019 (30 days

after the opening of Trader Joes including opening day), June 19th to July 18th, 2019 (30 days before the opening of Chick-fil-A) and July 19th to August 17th, 2019 (30 days after the opening of Chick-fil-A including opening day). For each food store radii set for each food store opening, various metrics were compiled. These include the total number of unique visitors and visitor-days before and after the opening of the store the store group is related to, with subdivision of both user data (desert vs non-desert-resident users) and store data (desert vs non-desert-resident stores). Additional metrics were generated without restricting the store radius, which include the quantity of visitors and visitor-days to the opened food store 30 days after opening (including the opening date), the number of food stores visited by nearby users before and after the case study food store opened and the number of visitor-days to food stores before and after the case study food store opened. Where applicable, results were divided between user residence (food desert or non-food-desert) and store residence (food desert or non-food-desert).

## Conclusions and Future Analytics Work

The computational optimizations which were achieved using the infomap algorithm along with the statistical processes applied to the dataset to infer trends about food store visitation behavior are potentially valuable for future studies on food access leveraging cellular location data. The utility of the dataset was further enhanced with the food store case studies, wherein the response of two different local demographics to the opening of two different food store types was effectively measured, which could indicate a trend with food store type openings. However, as this assessment was performed for only two store openings, repeatability over a much larger sample of food store openings must be conducted, which could take decades to realize, which is time that might not be available for population-wide cellular location data analytics. As mobile application developers crack down on the ability for third-party location

data to be collected such as with X-Mode (Sonnemaker, 2020), this will likely restrict at least the population size of future attempts at measuring population behavior if not prevent such studies from being conducted. While there is substantial uncertainty on the repeatability of this technical process for future timeframes or metro areas, the dataset which was developed over the course of this work can be expanded further and more analytics can be conducted. Specifically, leveraging the imminent improvement of address aliasing, user activity related to food stores can be more precisely determined. With such enhanced store data, a regression analysis which can be supplemented with a random tree machine learning model can be utilized, using FARA data as an alternative to the mRFEI data used in previous approaches and user behavior associated with food stores (Amin et al., 2020).

## WORKS CITED


About Us. *X-Mode*. Retrieved from https://xmode.io/about-us/

Amin, M. D., Badruddoza, D., & McCluskey, J. J. (2020). Predicting access to healthful food retailers with machine learning. *Food Policy*. doi:10.1016/j.foodpol.2020.101985

Aslak, U. & Alessandretti, L. (2020). Infostop: Scalable stop-location detection in multi-user mobility data. *DeepAI*. Retrieved from https://arxiv.org/pdf/2003.14370v1.pdf

Gustafson, A.A., Lewis, S., Wilson, C. & Jilcott-Pitts, S., 2012. Validation of food store environment secondary data source and the role of neighborhood deprivation in Appalachia, Kentucky. *BMC public health, 12*(1), pp. 1-12.

Jolly, J. (2018, April 20). How Facebook tracks your every move: Fact vs fiction. *USA Today*. Retrieved from https://usatoday.com/story/tech/columnist/2018/04/20/how-facebook-tracks-your-every-move-facts-vs-fiction/51613002/

Sonnemaker, T. (2020, December 9). Apple and Google have reportedly banned a major data broker from collecting location data from users' phones amid scrutiny over its national security work. *Business Insider*. Retrieved from https://www.businessinsider.com/apple-google-ban-developers-x-mode-collecting-location-tracking-data-2020-12?op=1

U.S. Department of Agriculture Economic Research Service. Documentation. Retrieved from https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/

## BIBLIOGRAPHY


About Us. *X-Mode*. Retrieved from https://xmode.io/about-us/

Amin, M. D., Badruddoza, D., & McCluskey, J. J. (2020). Predicting access to healthful food retailers with machine learning. *Food Policy*. doi:10.1016/j.foodpol.2020.101985

Aslak, U. & Alessandretti, L. (2020). Infostop: Scalable stop-location detection in multi-user mobility data. *DeepAI*. Retrieved from https://arxiv.org/pdf/2003.14370v1.pdf

Cookensey-Stowers, K., Schwartz, M. B., & Brownell, K. S., 2017. Food Swamps Predict Obesity Rates Better Than Food Deserts in the United States. *International Journal of Environmental Research and Public Health, 14*(11), pp. 1-20. doi:10.3390/ijerph14111366

Gustafson, A.A., Lewis, S., Wilson, C. & Jilcott-Pitts, S., 2012. Validation of food store environment secondary data source and the role of neighborhood deprivation in Appalachia, Kentucky. *BMC public health, 12*(1), pp. 1-12.

Jolly, J. (2018, April 20). How Facebook tracks your every move: Fact vs fiction. *USA Today*. Retrieved from https://usatoday.com/story/tech/columnist/2018/04/20/how-facebook-tracks-your-every-move-facts-vs-fiction/51613002/

Sonnemaker, T. (2020, December 9). Apple and Google have reportedly banned a major data broker from collecting location data from users' phones amid scrutiny over its national security work. *Business Insider*. Retrieved from https://www.businessinsider.com/apple-google-ban-developers-x-mode-collecting-location-tracking-data-2020-12?op=1

U.S. Department of Agriculture Economic Research Service. Documentation. Retrieved from https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/