Food Deserts and Birth Outcomes

Dana Chermesh Reshef (dcr346@nyu.edu)
Emily Padvorac (ep2247@nyu.edu)
Rebecca Scheidegger (rls672@nyu.edu)

Machine Learning for Cities
May 3, 2018

Project Github: https://github.com/danachermesh/Food_Deserts_ML

I. Introduction

This research aims to identify the important factors impacting low birthweight and preterm/premature births on the zip code level, using random forests and Bayesian network approaches. Identifying if certain variables are important in determining negative birth outcomes can help decide what policy levers to pull or what demographic groups to focus intervention programs on. These findings are applicable to New York government at both the state and local level, and would be most relevant to the Department of Health. Additionally, the research is applicable to hospitals or healthcare providers who can help identify at-risk persons. This research seeks to understand to what extent living in a food desert may have on birth outcomes. In recent years, the topic of food deserts has become increasingly popular. In short, a food desert is an area that has low-access to healthy, nutritious food.

Population and socioeconomic variables are considered at the zip code level for New York state. Importantly, health variables are not considered (i.e. BMI). While individual health variables are certainly important, many medical studies have already considered these and data were not publically available. Additionally, it can be difficult to design effective policy interventions at scale for health variables.

Between the two approaches employed, the random forest models are found to have higher accuracy than the Bayesian networks. Findings from the random forest suggest that the percent of population holding no health insurance, as well as the share of population using the

Supplemental Nutrition Assistance Program (SNAP), commonly known as food stamps, has a relationship to negative birth outcomes.

Motivation

Preterm birth and low birthweight have serious short-term impacts, but can also have long-term impacts, such as neurological problems, autism, ADHD, anxiety, asthma, hearing loss, vision loss, and dental problems (March of Dimes, 2013). By identifying variables that contribute to adverse birth outcomes, these long-term negative impacts can be mitigated, thus ensuring higher quality of life for individuals, and potentially saving government resources that would have otherwise gone to programs or resources to help these individuals.

II. Literature review

Many reports have explored what constitutes a food desert. A key distinction made in the literature is that the definition of food deserts varies for urban versus rural areas. In urban areas, a radius of 1 mile to a grocery store is used, while in rural areas, it is a radius of 10 miles. Other key considerations to identifying food deserts area individual-level resources such as vehicle access, as well as neighborhood-level characteristics, such as the average income (Economic Research Service).

The percent of low weight births has been increasing in recent years, a concerning trend (Hamilton et al, 2018). A recent study explored low birthweight patterns across the US at the county level. The research found a pattern of disparity in low birthweight trends, suggesting that different groups or areas are disportionately at-risk (University of Wisconsin Population Health Institute, 2018).

The link between food stamps and birth outcomes has been explored as well, with results showing that exposure to the food stamp program resulted in increased birth weight. However, while interesting, this research was performed on data from the 1960s and 1970s (Almond et al, 2008).

Bearing these important studies in mind, this research seeks to use more recent data and the more granular geography of zip codes, which allows trends to be more nuanced and intervention strategies to be more finely targeted.

III. Data

Birth Outcome Variables Data[1]

Two adverse birth outcomes were considered: preterm birth and low birthweight. Data on these outcomes were collected from New York State Department of Health at the zip code level.

Population Variables Data[2]

At the beginning of research, a dataset containing variables on food deserts was considered. However, the variables had extremely high multicollinearity. As such, data was gathered on a subset of the variables most relevant to the aim of the research. Data was collected and derived from the American Community Survey (ACS) 5-year (2009-2016) at the zip code level. Variables included citizenship status, poverty rate, health insurance coverage, vehicle access, and SNAP (Food Stamp) Benefits.

A variable of teen birth rate was also considered, and was retrieved from the NYS Department of Health data. Additionally, each zip code was classified as either urban or rural, using classifications as defined by the US Census Bureau.[3]

Data limitations

- Mother's age: The age of the mother is a known factor related to negative birth outcomes. In particular, women under 18 or over 35 are considered higher risk for preterm birth (National Institutes of Health). However, data was not publically available on a zip code

[1] Please see Appendix B for more details.
[2] Please see Appendix C for more details.
[3] Please see Appendix D for more details.

level for the age of mothers. Using teen birth rate as a variable is a partial proxy for this, although data was still not available for older mothers.

- Multiple births: Another factor that is known to increase the risk of adverse birth outcomes is multiple births (i.e twins) ((National Institutes of Health). Again, data was not publically available on a zip code level. For this issue, it was assumed that multiple births occur at a similar rate in all areas.

IV. Methodology

This research aims to study the relationships between different social, geographical and health factors. All data that were included in this study are labeled, thus methodologies of supervised learning only have been taken into consideration. Two methodologies were used: random forest and Bayesian network. Using two approaches allowed results to be compared and contrasted to better understand the structure and accuracy of model choice. In selecting machine learning paradigms, the trade-off between accuracy and interpretability must be considered. This is particularly true when approaching public policy, where it is obviously important to find the right solution. However, it is also important that the output be interpretable and not a "black box" whose underlying logic cannot be deciphered. Transparency and accountability can sometimes be considered more than the highest possible accuracy. The issue of birth outcomes is obviously a sensitive issue, and a model that is not interpretable by humans could be seen negatively.

Bayesian Network

A Bayesian Network is a Directed Acyclic Graph (DAG), where each node (a variable) is conditionally independent and causal relationships can be understood. They can be used for prediction, diagnosis, and anomaly detection. Being that they are highly interpretable, they are natural candidates for public policy problems.

Random Forest

Random forests are a powerful supervised learning technique, especially distinguished for their high accuracy. In order to best inform decision makers in regards to efficient approaches needed

to be taken when addressing critical phenomena such as low birth weight and premature birth, accuracy must be an important consideration. For this sensitive and critical issue of policy planning in order to address negative birth outcomes, features importance must be taken into consideration and thus is necessary to be determined after formatting the forest.

Data Processing

The raw data from the Census comes as real numbers. To normalize this, the number of people for a given variable was divided by the entire population of that zip code. This created a percentage that allowed for better comparison among geographies.

For both the random forest and Bayesian network, the dependent variables of low birthweight and preterm birth were made into dummy variables based on standard deviations. For both approaches, the test/train split was 30%.

V. Results

Bayesian Network

The first approach used was Bayesian networks, on both preterm births and low weight births. The model was created using the pgmpy Python library, including using the Hill Climb Search and Bic Score methods. For both models, in sample accuracy was calculated on the training data, and out of sample accuracy was calculated on the test data.

For preterm birth, the highest in sample accuracy achieved was 25.92%, and the highest out of sample accuracy was 23.54%. When teen birth rate was added as an independent variable, in sample accuracy increased slightly to 26.33% and out of sample accuracy rose to 28.19%.

For low birth weight, in sample accuracy peaked at 26.72%, and out of sample accuracy at 29.89%. With the addition of teen birth rate, in sample accuracy declined slightly to 25.44% and out of sample accuracy remain relatively constant at 29.12%.

For the Bayesian Network, the learned structure was identical for both low birth weight and premature birth. However, because the accuracy was not strong for either birth outcome, the results have limited use and cannot be relied upon.

Random Forest

Seeking to produce a model with higher accuracy, random forests were employed. The random forest approach was applied for both low birth weight and preterm birth. Dummy variables were created for each of the predicted variables, based upon their means and standard deviations.

The dataset was split into a train set and a test set, keeping 30% of the data for validation of our models. In order to assess the best random forest model of our data, a model selection was conducted, tuning the maximum depth of the trees between a range of 1 to 11 and using an exhaustive grid search over the specified parameter values for 100 estimators. The best parameter value was estimated for each of the birth outcomes and was applied for the test sample in order to assess the out of sample accuracy of the model.

For low birthweight, the best parameter value was max depth of 8 nodes, giving the out of sample accuracy of 62.9%. For premature birth, the best parameter value was max depth of 9 nodes, achieving the out of sample accuracy of 60.4%. Overall, the results were found to be very similar for both predicted variables, which is not surprising, given that premature birth and low birth weight are related phenomena.

Feature importance (Table 1) was consistent for fairly similar for both birth outcomes, with uninsured percentages being the most prominent feature for both outcomes. In both instances, this was very closely followed by food stamps. Poverty and teen births were the third and forth features, respectively. In low birth weight, immigrant was fifth and vehicle access was sixth, while in premature births, these were swapped, with vehicle access as fifth and immigrant as sixth. In both instances, urban came in last at a very low feature importance.

Table 1. Feature Importance of Random Forests

Feature	Low birth weight	Premature Birth
Uninsured	0.256	0.241
Food stamps	0.247	0.240
Poverty	0.236	0.237
Teen birth	0.105	0.115
Immigrant	0.077	0.059
Vehicle access	0.071	0.073
Urban/rural	0.010	0.036

VI. Discussion + Conclusions

Discussion

Overall, the random forest performed significantly better than the Bayesian networks for either birth outcome. Random forests are known to have higher accuracy. However, they are less interpretable. As discussed earlier, this can be a problem in the context of such a sensitive issue.

For both the random forests and the Bayesian networks, low birthweight and preterm birth had very similar model results. It makes logical sense that low birthweight and preterm birth had similar outcomes. Very often, if a baby is born premature, she or he also has a low birthweight, and vice versa.

In the random forest feature importance, uninsured percentages rank first for both preterm and low birthweight, following closely by food stamps percentages across the zip codes. The first rank makes sense as it can be assumed that uninsured women will conduct minimal, late, or no prenatal care at all, leaving them at higher risk for having pregnancy problems that will not be identified nor treated. Medicaid does cover all pregnancy related costs, however, not all women may be eligible for medicaid or receive coverage in time to seek prenatal care. Additionally, it could be such that women are unaware they are eligible to receive pregnancy care through

Medicaid. Government could increase education campaigns so that women are aware. Additionally, the government at either the federal or state level could consider providing some basic level of prenatal care for all pregnant women, regardless of income or Medicaid eligibility.

Having the food stamps ranked at the second highest importance suggests that food deserts do indeed matter in the context of adverse birth outcomes. Food stamps are only eligible to be used for certain products, and vitamins, such as prenatal vitamins, are not an allowed purchase (SNAP Website). Additionally, it has been observed that food stamp recipients do not consume high-quality, nutritional food (Condon et al, 2015). Some potential policy proposals to be considered would be allowing the purchase of prenatal vitamins on food stamps or providing pregnant women with increased food stamp benefits during pregnancy so that they could access more nutritional food such as fresh fruits and vegetables, which tends to be more expensive. Additionally, there could be further education campaigns facilitated through the SNAP program that specifically alert women of childbearing age to the importance of nutrition throughout pregnancy.

The third most important feature for both outcomes was poverty. There are obvious linkages between poverty, food stamps, and being uninsured. Overall, it is apparent that poverty rates and indicators of poverty (uninsured, food stamps) are extremely relevant to predicting negative birth outcomes. While poverty as a whole is a complex issue and difficult to address, it is constructive to know that individual factors such as being uninsured or being on food stamps have high importance in determining negative birth outcomes, as there are concrete and practical steps that can be taken to improve these programs for pregnant women, and therefore potentially reduce negative birth outcomes.

Limitations + Future Research

As mentioned, data on the age of mothers and the rate of multiple births was not available at the zip code. These conditions are medically known to result in adverse birth outcomes. Having

these data would likely fine tune the results, by considering what percent of events was caused by these conditions.

Data on birth outcomes was not available country-wide at the zip code level. Given this data, the model could be further refined and better assessed. Similarly, data at a finer granularity such as zip codes could allow for an improved model.

Model accuracy could be further improved by considering more variables. As seen with adding teen birth rate to the model, additional variables should be considered in future research. Additional variables that could be considered include the mother's age group (pregnancy risks increase with age and very significantly after age 40), number of former pregnancies, late or no prenatal care, percentage of non-white population, crime data, and more.

The random forests indicate that being uninsured and being on food stamps are relevant factors. While some women are eligible for Medicaid, there are opportunities for government to further expand or offer prenatal care and coverage to pregnant women. Future research should specifically consider the impact food stamps have upon birth outcomes. Future work could explore if the accuracy of the random forest could be reproduced in a more interpretable model, as random forests have limited interpretability.

Conclusions

Overall, this research found that random forests were better than Bayesian networks at predicting adverse birth outcomes for the considered variables. In the random forest models, uninsured was the highest ranked feature for both models, with food stamps scoring a very close second. The high score of food stamps indicates that food deserts are relevant to understanding adverse birth outcomes. Future research should further consider this linkage by using data at a smaller geography or data from the entire country. This research could have numerous policy implications, including increased benefits or educational campaigns, particularly to those receiving food stamp benefits.

References

Almond, D., Hoynes, H., Schanzenbach, D. (2008) Inside the War on Poverty: The Impact of Food Stamps on Birth Outcomes. The Review of Economics and Statistics 2011 93:2, 387-403. http://www.nber.org/papers/w14306

Condon, E., Drilea, S., Jowers K., Lichtenstein, C., Mabli, J., Madden, E., & Niland, K. (2015) Diet Quality of Americans by SNAP Participation Status: Data from the National Health and Nutrition Examination Survey, 2007–2010. Prepared by Walter R. McDonald & Associates, Inc. and Mathematica Policy Research for the Food and Nutrition Service. https://fns-prod.azureedge.net/sites/default/files/ops/NHANES-SNAP07-10.pdf

Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Access Research Atlas. Accessed April 4, 2018. https://www.ers.usda.gov/data-products/food-access-research-atlas/.

March of Dimes (2013) Long-term Health Effects of Premature Birth. October 2013. Accessed April 15, 2018. https://www.marchofdimes.org/complications/long-term-health-effects-of-premature-birth.aspx

Martin J., Hamilton B., Osterman M., Driscoll A., Drake P. Births: Final data for 2016. National Vital Statistics Reports; vol 67 no 1. Hyattsville, MD: National Center for Health Statistics. 2018. https://www.cdc.gov/nchs/data/nvsr/nvsr67/nvsr67_01.pdf

National Institutes of Health. What are the risk factors for preterm labor and birth? Accessed April 6, 2018. https://www.nichd.nih.gov/health/topics/preterm/conditioninfo/who_risk

SNAP Website. Eligible Food Items. Accessed April 23, 2018. https://www.fns.usda.gov/snap/eligible-food-items

University of Wisconsin Population Health Institute (2018) County Health Rankings Key Findings 2018. http://www.countyhealthrankings.org/explore-health-rankings/rankings-reports/2018-county-health-rankings-key-findings-report

Appendix A: Team Contributions

All team members contributed equally and meaningfully to this project:

- Dana: Scrapped & cleaned birth data; Merged birth and food desert datasets; Made models; Edited report; Drafted presentation
- Emily: Found food desert dataset; Scrapped & cleaned birth data; Pulled census data; Made models; Edited report; Edited presentation
- Rebecca: Found birth data; Pulled & cleaned census data; Made models; Drafted report; Edited presentation

Appendix B: Birth Outcomes Data, NYS Department of Health

- Data on NY state birth outcomes was retrieved from New York State County/ZIP Code Perinatal Data Profile for the years 2012 - 2014 was accessed through this URL: https://www.health.ny.gov/statistics/chac/perinatal/county/2012-2014/
  More details can be seen in the project Github on how to retrieve this data.

Appendix C: American Community Survey 5-Year Data

Variable	Code	URL
Total Population	B05001_001E	https://api.census.gov/data/2016/acs/acs5/variables/B05001_001E.json
SNAP (Food Stamps)	S2201_C01_001E	https://api.census.gov/data/2016/acs/acs5/subject/variables/S2201_C01_001E.json
Poverty	S0601_C01_049E	https://api.census.gov/data/2016/acs/acs5/subject/variables/S0601_C01_049E.json
Immigrant	B05001_006E	https://api.census.gov/data/2016/acs/acs5/variables/B05001_006E.json
Vehicle Access	B08014_002E	https://api.census.gov/data/2016/acs/acs5/variables/B08014_002E.json
Uninsured	S2701_C04_001E	https://api.census.gov/data/2016/acs/acs5/subject/variables/S2701_C04_001E.json

More details can be seen in the project Github on how to retrieve this data.

Appendix D: US Census Geography Files

- The shapefile used to create maps was retrieved for the US Census Bureau: https://www.census.gov/geo/maps-data/data/cbf/cbf_zcta.html
- In addition, data on whether the area is considered urban or rural was retrieved from the US Census Bureau: https://www.census.gov/geo/reference/ua/urban-rural-2010.html