altREU: Exploring Food Deserts and Environmental Impacts on Health in Chicago and Oregon
Podcast Transcription
Paulina Grzybowicz + Sivasomasundari (Sundari) Arunarasu
August 6th, 2020

[0:00]

Paulina

Hello, my name's Paulina, I'm a Computer Science Student at DePaul University

Sundari

And my name is Sundari, I'm studying Quantitative Science at Emory University

Paulina

And we've been working on a project as part of the altREU program, looking at the food desert impact on health outcomes using computational modeling.

Sundari

So my objective for this project was to explore the relationship between access to grocery stores, food services, and other environmental variables on health outcomes, focusing on diabetes. And I wanted to identify differences between urban and rural counties in Oregon that influence health outcomes

Paulina

So my objectives were also pretty similar: I wanted to showcase the relationship between food deserts and health outcomes in Chicago, specifically diet- related outcomes like diabetes. And further I wanted to explore the specific variables that are dictating food deserts, so grocery stores, income, food stamps, stuff like that, and their combined impacts on health outcomes, with the goal of understanding the significance of each variable, to then later analyze how different solutions, how effective they can be

So, just a little background on food deserts in general: A food desert is an impoverished area where residents lack access to healthy foods. Food deserts may exist in rural or urban areas and are associated with complex geographic and socioeconomic factors, as well as with poor diet and health disorders such as obesity.

This is a pretty widely accepted definition, but we think that a lot of the time, food access, people assume that that just means grocery store prevalence or how close you are to a grocery store, and we wanted to make clear that there's a lot more that goes into creating a food desert, different variables like public transport, income, cultural preferences, how much time you have to cook and to shop, nutritional literacy, and even policy, and we wanted to incorporate all these a lot of these different variables and analyze how food deserts are going to impact diet-related health

[01:50]

Though we had slightly different objectives, we used generally the same tools and techniques to conduct our analyze and create our models, which we'll go into right now

Sundari

So first we used R and Python to explore the different relationships between variables. We first used linear regression with each individual variable and several combinations were plotted against diabetes rate, the main health outcome we were looking at, as well as the obesity percentage, in order to observe the correlation between the variables and look for possible patterns and how strong different correlations were. In Python, the libraries we used were pandas, numpy, matplotlib and sklearn. For R there were many libraries used, as well as a specific one to create a heat map which was to view the strength of all the correlations in one diagram, which was very useful.

Paulina

After that, we moved onto conducting a mediation analysis, which just further explores the relationships between different variables. We used R to conduct the mediation analysis. And this basically is an evaluation of whether or not a third variable acts as an intervening variable or mediator between the independent and dependent variable. A mediation analysis consists of 3 regressions and we used Baron and Kenny's steps for mediation and then we used a bootstrap technique to test at the end the significance of the mediation effect.

This consists of creating a regression between the independent and dependent variable, And then the independent variable and mediator, and then finally the independent and mediator with dependent and just analyzing the p-values to see whether the mediation effect is significant.

We were recommended this technique by our advisor, Professor Wayne Wakeland from PSU. And we inserted here what he thinks about it.

Wayne

Why did you suggest mediation analysis? That's because it seemed to me that there could be logical causal pathways or sequences of what influences what, not just a pure independence model. So standard statistical analysis would start with this idea where you would say: "here a bunch of presumed independent factors and we want to figure out how they contribute to my outcome variable", and you have two outcome variables, sort of. And partly because you have two outcome variables, and partly because it felt like the ideas of food and food deserts and the presence or not of other conditions, they aren't independent. And that's why fundamentally, the suggestion of mediation analysis was because maybe there's a more logical pathway towards the outcome, rather than just an independent model.

Paulina

And then finally, we used Python to create several machine learning models with the goal of predicting the diabetes rate based of off a variety of environmental and food-related variables

We used several different machine learning algorithms, including a Support Vector Regression, Ridge. Lasso, Elastic Net Regressions, as well as the Random Forest method.

[5:04]

Sundari

Hi, this is Sundari, and my project focused on diabetes rates in different Oregon counties and trying to find the most influential factors that affect the prevalence. So, I decided to focus on Oregon since I thought looking at all the counties in the United States would be too much data and I wanted to kind of use a more manageable amount.

So at first, I started with only a few variables related to the environment and food access, which were a number of grocery stores, number of primary care providers and number of food services per person. And then I decided to expand my data search to collect a few more variables such as unemployment rate, high school graduation rate, percent of people physically inactive access to physical activity, the number of fitness centers per person, the number of people who are food insecure in a county, the percent of people who qualified for free and reduced lunch, percent obesity in a county and the average family income.

So I looked for which factors had strong relationships, relationships to diabetes rates, and I found some expected and some surprising results. So surprisingly, having limited access to healthy food and the number of grocery stores per person are not very strongly correlated to diabetes rates. On the other hand, the number of primary care providers per person was highly correlated, as well as the number of food services per person. One note to make here was that more food services in a county was actually correlated with lower diabetes rates, which seems a bit surprising, but this suggests that there are some underlying differences between rural and urban counties. So having more access to food services, which are generally thought of as unhealthy or fast food doesn't necessarily lead directly to poor health. And this is a topic that can be looked into further.

So I also look for differences between rural and urban counties as well as possible mediator variables, which are variables that the independent variable affects and then the mediating variable then affects the outcome, which was diabetes rate. So, I found that being physically inactive was part of a mediation analysis where percent obesity was the mediator. So this means that physical inactivity basically led to percent obesity, which then led to a higher diabetes, right. So there was also a significant difference in having limited access to healthy food between rural and urban counties with rural counties having a much higher percent of people who did not have access to healthy food. But this also needs a bit more research because it seems that just having not having access to healthy food is not actually strongly correlated with higher diabetes rates.

So a few things to mention is that I had a pretty small sample size because Oregon has 36 counties. So I do think that some of the data could have been, the data analysis, could have been more accurate if I had a larger sample size. And some counties also have quite small populations, less than 2000 people. So that could also have skewed some of the percent based data such as percent obesity or percent diabetes.

Okay, so then I used different machine learning models to find the most important variables and see if they matched up with what I had observed so far. So overall, I found that the number of primary care

providers per person, percent obesity in a county, and the number of food services per person had the most weight in the overall analysis.

So, when using machine learning, the overall process for each model was to split the data into a training and a test set with 25% of the data saved for the testing. So this means that the model first uses the training data as a base for its calculations, and then it uses this training data to make predictions on the test data, which is what we are trying to predict. And then I looked at the accuracy score, which is the R squared score between the actual data and the predicted value values from each model.

So, then I compared the accuracy of my different models. So, the first model I used was a simple linear regression which with the three variables mentioned, this was not as accurate as the following models I used, then I used ridge regression which minimizes the coefficients of irrelevant variables. This is usually so the model does not overfit the data and conform, conform too much to the training set and not make as good predictions for the test set. So, for there's a very similar regression called lasso regression, which actually reduces the irrelevant coefficient values to zero, which is useful for narrowing down variables when you have a large set of variables and you want to use only the most important ones.

So for ridge and lasso regression, they had very similar accuracy scores. And they were quite a bit more accurate than simple linear regression. So then I also use the random forest model, which is a bit more complicated, but basically it just constructs a number of decision trees. So it uses a lot of different samples from the data to try to come up with an accurate prediction and accurate model. So the random forest regression was also more accurate than the linear model, it was a little bit less accurate than the rich and lasso models.

So one other thing I wanted to mention was how overfitting and underfitting can really affect the models especially since I had a pretty small data set. So since the testing set, which was 25% of the data was only nine counties, it was a little difficult to see how accurate the models really were because the model only had basically nine chances to try and predict the diabetes rate.. So if I did have a larger sample size, the models accuracy would likely have been a bit higher.

So of the three most important factors that were found the number of physician primary care providers per person, and number of food services per person and the obesity rate. The correlation between obesity and diabetes rates is not that surprising, and the entire nation and other countries as well have seen large increases in obesity over time. And this is a pretty large scale issue not just in Oregon or Oregon rural counties. The number of primary care providers being quite a bit lower in rural areas, is also a pretty established problem. And there are some incentives for physicians to practice in rural or underserved areas, like the National Health Services core program pays for medical school for physicians that commit to practicing and health shortage areas. And for the final variable, the number of food services per person, I think that this is quite an interesting relationship because more food services actually lead to lower diabetes rates. And I think this relationship can definitely be explored further to try to find more differences between rural and urban counties and what exactly is another influential variable that affects both the number of food services and diabetes rates as well.

So I also found that the average family income, the percent of people in poverty and the unemployment rate all have weak to moderate relationships with obesity rates and being physically inactive, which both contribute to the overall diabetes rates. So I think this shows that there are many economic and possibly demographic factors also involved in overall health outcomes, and how simple judgments and simple conclusions can't really be made about the overall health of a county or state or even a group of people.

[14:32]

Paulina

Hey, this is Paulina speaking now, and I'm going to transition into talking about my work with Chicago communities.

Like we said earlier, my specific project is focused on communities in Chicago
Food deserts are very prominent in CHicago, and they correlate specifically with the socioeconomic status of a community, which is pretty common. But in Chicago, communities of color are disproportionately impacted by food deserts. The city has a very diverse population, with White, Black, and Hispanic/Latino populations making up 32.3, 26.7, and 30.9 percent of the total population respectively. This makes for an almost even split between the three groups, making it one of the most diverse cities in the United States.

But it's also extremely segregated, due to many factors, but primarily due to a long history of racist housing and investment policies. After decades of this kind of unfair investments in housing, education, medical resources and food systems, you just have a lot of populations in Chicago at a disadvantage to this day. And this divide is very unsurprisingly correlated to race.

So there's significant inequalities among Chicago community areas, and this obviously impacts food systems as well. This is why we see communities of color being disproportionately impacted by food deserts.

I collected several datasets to work on these projects and they were all sourced from Chicago Health Atlas and Chicago Data Portal, which are online sources that provide datasets about Chicago citizens All the data is divided by Chicago Neighborhood, which are the 77 communities the city is made up of.

Some of the variables I collected datasets for were: Population, Square Milage, Percent Active Transportation, Percentage Non-Hispanic White, Percentage Non-Hispanic African American or Black, Percentage Hispanic or Latino, Median Income, Percent Poverty Rate, Unemployment Rate, Food Stamp Usage, Percent with Easy Access to Fruits and Vegetables, Number of Grocery Stores, Adult Obesity Percentage and Diabetes Rate.

So initially I conducted several linear regressions, all variables considered to create a food desert were plotted against Diabetes rate to search for relationships between variables. Variables with a stronger correlation to Diabetes Rate include: **Limited Food Access, Active Transportation, Unemployment Rate, Percent Poverty.** Variables with unexpectedly low correlations include **Median Income, Percent Non-Hispanic White, Percent Non-Hispanic African American,** and in general the racial variables.

The stronger correlations were pretty evenly distributed between socioeconomic and food-related variables, meaning there wasn't an indication that one had more influence on diabetes rate than the other. This supports the general idea that a lot of variables go into creating a food desert and a lot of variables have an impact on diet-related health outcomes.

The lack of correlation between race variables and diabetes rates was initially surprising, considering the segregation in Chicago.But when I ran a regression against the combined percentage of Non-Hispanic African American or Black population and the Percent Hispanic/Latino population, this correlation was a lot stronger, and it just showcases the overall correlation between communities of color and higher diabetes rates.

The lack of strong correlation between median income alone and diabetes rate is possible, because socioeconomic status is a lot more complicated than just income. When you see the regression between a combination of these variables, like Poverty Rate and Unemployment Rate, the correlation becomes a lot stronger.

Following this initial exploration, we conducted a mediation analysis which was recommended by our advisor. Like mentioned previously, mediation analysis is going to further investigate the relationships between different variables. The hypothesis that I created and was being tested was that different food-related variables like access to fresh fruits and vegetables, number of grocery stores, things like that, act as a mediator between socioeconomic indicators and diabetes rates.

I was thinking that A community area's socioeconomic status influences the area's access to healthy foods, in numerous ways, which in turn influences health outcomes in the area.

So I tested several combinations of Socioeconomic status and food-related indicators. Some of these groupings, and I'll list them as the Independent Variable, Potential Mediator, and Dependent Variable, are as follows: Median Income, Percent Easy Access to Fruits and Vegetables, and Diabetes Rate; Poverty Rate, Percent Easy Access to Fruits and Vegetables, Diabetes Rate; Unemployment Rate, Percent Easy Access to Fruits and Vegetables, and Diabetes Rate; and then Median Income, Percent Easy Access to Fruits and Vegetables, and Obesity Rate. Out of these, the latter three were significant. The first one was not very significant, but upon evaluating the different p-values of all the regressions and also using a bootstrapping technique, the mediation effect did prove significant here. This further confirmed the relationship between the variables and was helpful when choosing the variables that went into our training set for the machine learning model.

So following the mediation analysis, I moved onto creating the machine learning model. The machine learning models aim was to predict the diabetes rate in a community based upon the variables I had collected.I created several models and used several training sets in search of the most accurate model. In the end, the variable importance was evaluated.

Like mentioned before, the models included: Support Vector Regression, with linear, rbf, poly, and sigmoid kernels, Ridge Regression, Lasso, ElasticNet, and then also Random Forest models. I also used

several training sets. These involved different combinations of variables in order to explore what variables were necessary to create an accurate model

Some of these combinations were actually chosen by an algorithm that selected the most important features, but some of these were also just chosen by me based upon the groupings I understand within the variables. So for example, I had a training set with all the variables I listed before. I also had a training set with all the variables except obesity rate, since that is so highly correlated with diabetes rate. I had a training set without any racial variables. I had a training set without any socioeconomic status variables. And then I also tried one without any food-related variables. The goal of this was just to see whether or not the model would be significantly more or less accurate based upon the inclusion or exclusion of these variables.

So in the end, after creating all these training sets, I did find that the most accurate one was the training set with all the variables included. The models that didn't include SES indicators and food-related variables both had significantly lower average accuracy rates, indicating the importance of these variables, which was helpful to learn. It's also important to note that all the training sets that included Percent Adult Obesity Rate yielded a higher average accuracy than those without. And again, this showcases the very large influence of this variable, and further puts to question the relevance of this variable, which we'll discuss in our conclusion.

I also analyzed the average accuracies of all the models I used, and in the end we found that the Ridge, Lasso, and ElasticNet regressions had the highest average accuracies, as well as the SVR with the linear kernel. The SVR models that used the rbf and sigmoid kernels had the lowest accuracy rates. The high accuracy of the Ridge Regression, Lasso and ElasticNet models, which was probably due to the fact that these models work well with datasets that have multicollinearity, meaning that the predictor variables have a lot in common with each other and are correlated to each other. In this case, most of my variables are, so it makes sense that this specific model worked best. In the case of the SVR with the linear kernel, that's just a simplest model, and oftentimes the simplest model, especially with linear regression, is the most accurate.

After running the machine models, I evaluated the importance of the variables. For simplicity, the importance of variables is going to be evaluated based on the first training set, which is the training set with all the variables.

Among all the models, a combination of Obesity Rate, Use of Active Transportation, Number of Grocery Stores, Use of Food Stamps generally have the largest importance among the variables in the training set. These variables are not concentrated in one grouping, rather, they represent a wide range of categories - health, infrastructure, socioeconomic status , and food-related variables.

This was exemplified by the linear regressions, the mediation analysis, and all the work done with machine learning models.

Some important things to note are that based upon the training set variation and machine learning model technique variations, the average accuracy rate of the models didn't vary drastically, as they all remained within the range of 20 to 40 percent. The highest individual accuracy rate we saw was around 60 percent, which in general is pretty low.

So because of that, one of the most important conclusions from this project does have to do with dataset size. The datasets that were considered were all dictated by Chicago Community Area, which there are 77 of. Though each data point represents a large population, the machine learning models were all trained on 77 different points, which is a very low amount. The accuracy and importance were somewhat inconsistent because of this, just because a small dataset means you have to compromise on both the size of the training and testing set. There is a lot of room for error with smaller size. So this definitely works to explain some of the inconsistencies in results.

Another thing that my partner Sundari and I, as well as our advisor Professor Wakeland, discussed at one point was the inclusion of the Obesity Rate variable. Obesity is so strongly linked with Diabetes Rate that its going to heavily improve the accuracy rate of any model it is included in. We questioned whether or not we should include this variable, since obesity is not a food-desert variable, but rather, a diet-related outcome, which is what we were trying to predict. It was controversial to include the variable in the first place, the fact that it so strongly influenced the models confirmed that it should likely not be used in the final machine learning model.

Moving on to an evaluation of potential solutions, knowing the food-desert variables that influence health outcomes like diabetes can help in the search for a solution in communities. Health is obviously an extremely complex subject, but it is evident there is a strong correlation between food deserts and diet-related health outcomes. Knowing which variables make up food-deserts can help in the evaluation of solutions.

For example, opening up grocery stores in communities that lack them is an overly simple solution when considering the other variables that contribute to food deserts. Factors that have to do with socioeconomic status, such as transportation access, time to cook and shop (due to jobs and/or childcare), food stamp eligibility at stores, etc have to be considered when discussing solutions.

Knowing that **Percent Active Transportation Use** contributes significantly to predicting diabetes rate in our model might mean that making online grocery shopping more accessible and affordable could be helpful to those in food deserts. Knowing that **Percent Food Stamp Usage** influences our model strongly could indicate that expanding the types and amounts of stores that accept food stamps could be beneficial to those living in food deserts.

These are just two small examples, but the underlying idea is that food deserts are a complex phenomenon made up of a wide array of related variables. Because these variables are so strongly correlated with several socioeconomic indicators, including race in Chicago, it's important to realize that it's impossible to reduce food deserts in Chicago without taking income and racial inequality under consideration.

[26:32]

Paulina

So when comparing our findings, we both found that the Ridge and Lasso models were most effective. We both had the highest accuracy rates when it came to those models. This makes sense as these models are best for datasets with collinearity.

In terms of variable importance I personally found that in Chicago, **Food Stamp Usage, Number of Grocery Stores, Obesity Rate, and Transportation,** were the most important variables.

Sundari:

In Oregon, I found that the most influential variables were the number of food services per person, the obesity rate, and the number of primary care providers per person.

Paulina

Though these are different variables, we found that both of these sets of variables showcase a variety of factors. This just confirms that the relationship between food deserts and health outcomes is complex, and involves the influence of several different indicators.

On top of that, something that we noted was that in CHicao, because it's more racially diverse and segregated, race just plays a larger role than it does in Oregon, since Oregon is less diverse racially.

The differences in our results can also somewhat be attributed to the general differences in our projects: I focused on one city, while Sundari focused on Oregon counties, which includes urban and rural areas. We also analyzed different variables based upon their availability.

[27:30]

Sundari:

So for moving forward, we had a few different ideas about what we could do to expand on the project.

So we considered using ensemble regression, which uses a lot of different machine learning models and averages the results to try to create a more accurate model and more accurate predictions. We were also thinking about figuring out a way to bring in more data which could be in the form of more variables or possibly data over time, which would be quite interesting to look at. We were also thinking about using different packages outside of the scikit learn models that we used in Python, and using machine learning models to consider different solutions or try to address the larger scale issues that we discovered in our project.

I also think That based on the variables that were found to be most important, it would be really interesting to try to analyze what type of public health programs are put into place or have been put into place in the past and see how effective or what type of impact these programs programs have had on a County's overall health or on a neighborhood or city's overall health.

So no matter what we do, we want to continue combining our background knowledge with any of our data findings. So something our advisor and cards throughout this process was to balance all data driven processes with theoretical knowledge. And here was his response when we asked him to expand this idea a little further.

Wayne:

Algorithms are essentially knowledge blind. Which is a mixed blessing. It's a good thing, because it means its going to go in directions that knowledge may have discouraged you from going, so there's real benefits to that. But on the other hand, we have paid dearly over centuries for the knowledge that we have about how things work, and to not take advantage of that seems inappropriate as well. That immediately leads us to - well maybe the other thought would be that human cognition is very limited by computational capability. Our brains aren't don't have - we've created computational devices that can essentially, in terms of speed, run circles around our brains, or so to speak. So we want to leverage those capabilities. That seems to me perfectly natural that we would want to figure out really smart hybrids of combining ways to use computational tools with human cognition which has been developed over millennia, many millennia. And it works, in very interesting ways. So it just seems obvious that we'd want to explore different hybrids.

Sundari:

We also asked our mentor, Professor Wakeland, his thoughts on our future work and possible ideas for the future. Here were his comments.

Wayne:

I would say first of all, I've been encouraging you and you've attempted to do some sort of combination. So that might make what you guys are doing moving in a direction that might not have been fully explored yet. So you're maybe going in that direction.

And then also the importance of your fresh eyes on this. A lot of the researchers, a lot of what people are doing, has been informed by years of experience, which like being knowledge-blind, you are knowledge-blind in that you are coming in fresh. So you might think of things or look at it in a dis fashioned way rather than a invested way like other researchers have.

Paulina:

After that helpful commentary on our project, Wayne provided us with some advice going forward.

Wayne:

How do you think our project can be expanded upon in the future? I think there's an interesting opportunity to develop feed-back oriented causal models in this domain, which is what I do. I never pushed you in this direction because it didn't seem important or relevant or time to do it. But I think there is a real opportunity to build a system-dynamics type of model, that's the method I use, to essentially create a couple of differential equations that link these things together and work over time and talk about, you know, if you really want to understand how we developed into this highly obese population in the United States, that story is a story that can be casually explained as a set of interacting equations or

increasing dependence on agriculture thats not human-scaled, that's massive scale, food production system that have had an effect on caloric intake and how people do it. And that we're moving away from our lifestyle that naturally includes getting a lot of exercise, and mixed results on figuring out how to put exercise as sort of an artificial part of the way we live, you know, taking part in exercise.

And the whole thing about diet and how that works - all of that is an interesting story. I think what you are working on has the opportunity to inform, and move into a direction of trying to create a more complex causal theory about how different local, or global, or national societies, populations work so you can create a model that can be tailored to a United States, a local, a Oregon model that at the same time can be plotted as a more general explanation for why there are all these huge differences in different regions and different countries.

[32:22]

Sundari:

So thank you for listening to our podcast, I hope you found it at least a little interesting and feel free to reach out to us with any questions or comments that you might have. Thank you!