Household food retail environment and shopping behavior

Timothy KM Beatty[a]

Joel Cuffey[b]

[a] Department of Agricultural and Resource Economics

University of California-Davis

[b] Graduate student

Department of Applied Economics, University of Minnesota-Twin Cities

cuffey@umn.edu

Selected Paper prepared for presentation for the 2017 Agricultural and Applied Economics Association Annual Meeting, Chicago, Illinois, July 30-August 1

1 Introduction

Given well-documented nutritional and dietary disparities in the United States, recent research has highlighted the potential relationship between the food retail environment and diets (Caspi et al., 2012). This relationship has been a concern especially for areas thought to be food deserts, or low-income areas where grocery stores with healthier food options may be further away than other areas, but where convenience stores with less-healthy foods are readily available (Walker et al., 2010). The assumption is that, in food deserts, distance serves as an often-insurmountable hurdle, and that there is latent demand for grocery store shopping such that the same households would shop at grocery stores if they were closer.

Besides geographic access, nutritional disparities could result from consumer preference heterogeneity as well as income disparities (Handbury *et al.*, 2016; Allcott *et al.*, 2015). Understanding the relative roles of access, preferences, and income in store choice is therefore important for understanding the reasons for the food retail environment-diet relationship. Moreover, it is vital to take each of these determinants of store choice into account when analyzing potential policy solutions to the problem of nutritional disparities.

In this paper we summarize 1) the relationship between proxy variables for preferences and household expenditures at grocery stores and convenience stores, and 2) the relationship between preferences and geographic access, using a novel dataset of SNAP transactions from a large Midwestern city. The data provide us greater geographic detail than has previously been used to study geographic access and store choice.

2 Data

We use SNAP administrative data on all SNAP participants in a large Midwestern metropolitan area for the period October, 2007-September, 2010. The data comprise monthly information on the household composition, authorized benefits, and car ownership, as well as monthly metadata on each EBT transaction. Our analytic dataset collapses this transaction-level data to a monthly summary of transactions for each household.

Importantly, the household data include the household's current address, and the transaction metadata includes the name and address of the store. We geocode the addresses, and are able to geocode 97% of the household addresses and 100% of the store addresses within a two-county buffer of our metropolitan area. Our initial household-month sample size is 1,948,318. Of the geocodable household addresses, some households reported more than one address in a month. Given the uncertainty surrounding the location of these households, we do not calculate access for these households, and drop from the sample households who ever had an ungeocodable address or multiple addresses (694,744). To restrict our analysis to urban food deserts, we further drop from our sample households who live in Rural census tracts (68,999 of the geocodeable sample), resulting in a final total household-month sample size of 1,184,575.

Our transaction data include also the name of the store, which we use along with the location to classify stores into five categories. Convenience stores are those which are less likely to stock healthy food options such as fresh fruits and vegetables, and are defined as corner stores, gas-marts, pharmacies, and dollar stores. Grocery stores are more likely to stock healthier food options. We classify under the grocery store category all grocery stores, superstores, supercenters, warehouse clubs, and mass merchandisers. While mass merchandisers are not especially likely to stock healthier foods, it was often difficult in our data to distinguish between mass merchandisers and supercenters since many mass merchandise stores have been converted into supercenters. We do not have the necessary information on the date of conversion for each store. The discount store category is comprised of limited assortment and service stores such as Aldi. We identify ethnic stores as any Hispanic, Asian, or East African store or market. Ethnic stores often have a mixture of grocery store and convenience store properties: they are often smaller than grocery stores, and have more limited assortment, but may also sell fresh fruits and vegetables. All other stores such as cooperatives, natural food stores, butchers, bakeries, and farmers markets are classified as Other.

We can also use the transaction data to describe a household's food retail environment. Within the metropolitan area, the transactions occur at all stores that accept SNAP benefits and that

SNAP households choose to shop at. Since households can spend their benefits on stores outside of the metropolitan area, and indeed outside of the state, we restrict stores to be considered in describing the retail environment to all counties within a two-county buffer of our metropolitan area. With a great degree of accuracy, our monthly data then are able to identify the month in which a particular store opens or closes. There are three concerns with using the transaction data to define store availability. First, stores with low transaction frequencies may artificially appear to open or close throughout the period. To mitigate against this, we restrict our household sample to only urban residents - stores in urban areas are far more likely to have a large number of transactions per month, and have no transactions in a month in which the store is truly unavailable. Additionally, we categorize as available in all months any store with under 20 average transactions per month that is seen for at least 10 months in a row.[1] The second concern is that there may be other stores in a household's area that accept SNAP but where no one has ever used SNAP benefits. To address this, we compare the list of stores from the transaction data to a list obtained from the State of all stores authorized to accept SNAP in our metropolitan area at any point between October, 2007 and September, 2010. Only four percent of the total number of SNAP-authorized stores are not found in the transaction data. In order to most comprehensively describe the retail environment, we include these stores in our measurements of store access. The third concern is that there are other stores which do not accept SNAP that should be part of a household's food retail environment. We therefore provide the caveat that our food retail environment is properly the SNAP retail environment. SNAP benefits total on average about 67% of a household's total expenditures on food at home[2], so we expect substantial overlap between a household's SNAP retail environment and food retail environment in general.

Since the data are from an administrative database, we have both complete coverage of SNAP participants during our period, but also limited variables that can be used to proxy for preferences. Handbury *et al.*, 2016 use income and education. With education unavailable to us, we use income

[1] From inspection of the data, stores with 20 or more transactions per month have stable numbers per month, and have no transactions only when the store is not available.

[2] Authors' calculations based on data from the Current Population Survey Food Security Supplement for years 2004-2013.

and race. The administrative data identifies each individual in each household as any combination of black, white, or other. Hispanic households are not identified, though we would expect that to be an important confounder. We thus define households as "Hispanic" if, on average over all months that the household appears in our data, at least 5% of their SNAP expenditures are at Hispanic ethnic stores.[3] Black and white households are defined as non-Hispanic households where all members identify as that race. Finally, households of other race are households of mixed races, or households with self-identified mixed-race individuals, or Asian or Native American households. Income is not directly given in our dataset, but we do observe the household's authorized benefit amount. This amount is a function of the maximum benefit level and the household's monthly income net of SNAP deductions. Using this formula, we calculate the household's net income. About 29% of our sample has no net income; households with some income receive $538 in monthly net income.[4] Since so many households report no net income, we categorize income into four bins: $0, $1-200, $201-500, and $500 plus. All results are reported relative to no income.

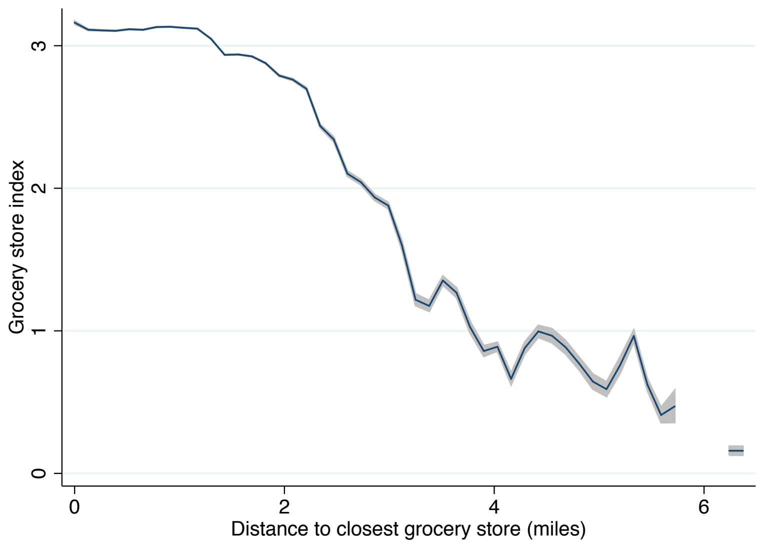
3 Measuring access

Households often do not shop at the closest grocery store (Ver Ploeg *et al.*, 2015). Thus a useful measure of household food retail access would consider more than just the closest grocery store, but also one in which stores further away are given lower weight. One possible such measure is an inverse distance weighted concentration index. We choose a Gaussian kernel to calculate the index (Handbury *et al.*, 2016). Since access can vary with the type of store, we calculate indices for all stores, and by store categories. For household h in month m, store s, and store category c, the concentration index is thus

[3] Four percent of our households are identified as Hispanic. This is broadly comparable to the 2010 Census numbers for the Minneapolis-St. Paul Metropolitan Statistical Area, which classifies 5.4% of the population as Hispanic or Latino of any race.

[4] This amount is net of allowable SNAP deductions. We do not have information on which deductions a household took, so we are unable to calculate the household's gross income, which would be substantially higher than the net income.

Figure 1: Kernel regression of household-month grocery store index on the distance to the closest grocery store



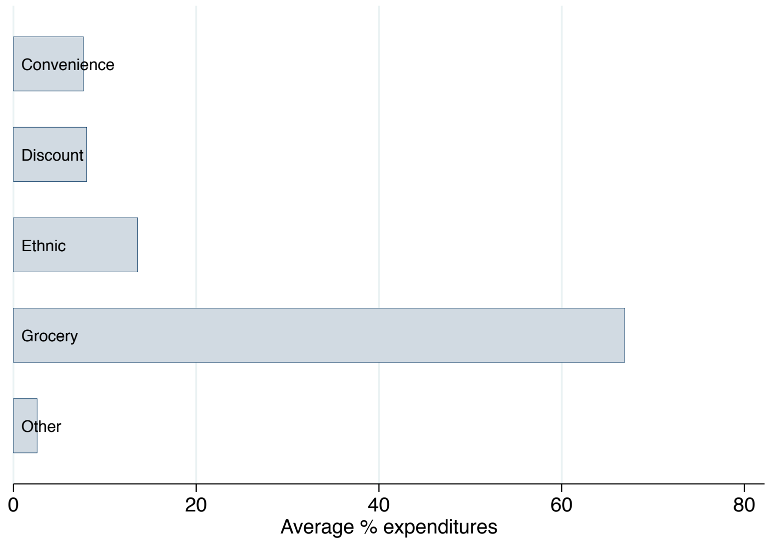
$$I_{hmc} = \sum_{s=1}^{S_{mc}} \frac{1}{2\pi} e^{(-\frac{1}{2})(\frac{d_{hs}}{3.7})^2}$$

where S_{mc} is the total number of relevant stores in category c in month m, and d_{hs} is the Euclidean distance between household h and store s. We choose a bandwidth of 3.7 miles, the average distance actually traveled in our dataset. We define "relevant" stores to be the 20 closest stores overall, and the 10 closest stores in each store category.[5]

Figure 1 shows how this index compares to a commonly-used measure of access: distance to the closest grocery store. As should be the case, the index is decreasing in distance. Note that higher levels of the index indicate higher concentrations of stores near that household.

[5]While Euclidean distance is not necessarily the same as travel time, in our data they are practically identical: the correlation coefficient between Euclidean distance and travel time as calculated by ArcGIS is 0.98.

Figure 2: Average household monthly % expenditures at each store category



4 Documenting differences in shopping patterns and access

4.1 Differences in shopping patterns

We first characterize shopping patterns and document a grocery store-race and grocery store-income gradient. As Figure 2 shows, the average household in our sample spends over 60% of their monthly budget at grocery stores (as defined above).

In Table 1, we regress the percent of a household-month budget spent at each store category on household characteristics.[6] Non-white households spend over 20% less at grocery stores than white households, but also just 1-5% at convenience stores. Ethnic stores make up a large portion of the rest of expenditures. While ethnic stores often have some characteristics of convenience stores, in general the healthfulness of ethnic store food options relative to grocery stores is ambiguous. Households in our sample with greater income tend to spend increasingly less at grocery stores, and split the rest between discount, ethnic, and convenience stores.

[6] Other variables used in the regressions but not shown in the table include household size, household age composition, whether a single parent household, the number of cars owned by the household, the month and year, and the length of the household's benefit month.

Table 1: Regressions of expenditure shares (%) in different store categories on household characteristics

	Convenience	Grocery	Discount	Ethnic	Other
Black	4.4***	-23.0***	5.3***	12.8***	-0.6***
	(0.2)	(0.3)	(0.2)	(0.2)	(0.07)
Hispanic	1.2***	-33.5***	2.1***	28.8***	1.2***
	(0.3)	(0.6)	(0.3)	(0.5)	(0.2)
Income $1-200	0.4**	-2.5***	0.3*	2.7***	-0.3***
	(0.1)	(0.3)	(0.1)	(0.2)	(0.08)
Income $201-500	1.3***	-2.1***	0.8***	1.3***	-0.4***
	(0.1)	(0.3)	(0.2)	(0.2)	(0.08)
Income $501+	4.9***	-5.5***	1.4***	0.9***	-0.7***
	(0.2)	(0.3)	(0.1)	(0.2)	(0.08)
Observations	1184575	1184575	1184575	1184575	1184575

Standard errors in parentheses

Standard errors clustered by household. Black, hispanic vs white; Inc vs no income.

Other controls: household size and age composition, single parent, number of cars owned, benefit month length, month-year dummies.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.2 Differences in access

The variation in shopping patterns can be compared with access differences between households. We use the household-month concentration index to measure access, and in Table 2 regress overall and store category-specific concentration indices against household characteristics. Non-white households live in areas with greater concentrations of stores overall, but especially ethnic stores. In general, households with no net income live in areas with greater concentrations of most store categories compared to households with some income.

Table 2: Regressions of household-month store concentration indices on household characteristics

	Convenience	Grocery	Discount	Ethnic	Other
Black	0.7***	0.3***	0.2***	1.0***	0.5***
	(0.009)	(0.005)	(0.003)	(0.01)	(0.008)
Hispanic	0.8***	0.3***	0.08***	1.1***	0.7***
	(0.02)	(0.01)	(0.005)	(0.02)	(0.02)
Income \$1-200	0.003	-0.04***	-0.03***	-0.004	-0.05***
	(0.009)	(0.005)	(0.003)	(0.01)	(0.008)
Income \$201-500	-0.06***	-0.09***	-0.06***	-0.1***	-0.2***
	(0.010)	(0.006)	(0.003)	(0.01)	(0.008)
Income \$501+	0.06***	-0.05***	-0.05***	0.04***	-0.06***
	(0.009)	(0.005)	(0.002)	(0.01)	(0.007)
Observations	1184575	1184575	1184575	1184575	1184575

Standard errors in parentheses

Standard errors clustered by household. Black, hispanic vs white; Inc vs no income.

Other controls: household size and age composition, single parent, number of cars owned, benefit month length, month-year dummies.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5 Conclusion

We measure the determinants of SNAP budget shares spent at different store categories, and the determinants of food retail store access. In so doing, we observe large impacts of proxy variables for preferences on differences in store category budget shares, but comparison with the relationship between preferences and access does not appear to explain these impacts. These results suggest that preferences may play at least as great a role as access to food stores in the shopping decisions of low-income households. While we do not argue here that our results are causal, they are suggestive, and future research will estimate the causal relationship between geographic access and store choice.