THREE ECONOMIC ISSUES IN HEALTH AND/OR SPACE

by

Jeffrey Stephen Young

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Agricultural Economics

West Lafayette, Indiana

December 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. James K. Binkley, Chair
Department of Agricultural Economics
Dr. Joseph V. Balagtas
Department of Agricultural Economics
Dr. Michael S. Delgado
Department of Agricultural Economics
Dr. Heather A. Eicher-Miller
Department of Nutrition Science

Approved by:

Dr. Nicole J. Olynk Widmar
Head of the Graduate Program

ACKNOWLEDGMENTS

I wish to express my deep gratitude to my wife, Ashlee, and my daughter, Leela, for all their patience, love, prayers, and encouragement. Additionally, I thank my parents, John and Beth, for their indispensible and unique roles in seeing me through to this point in my educational career, as well as each of my siblings – Mary, Matthew, Alexander, and Laura.

My church family members, both near and far, gave more counsel, prayer, encouragement, guidance, wisdom, direction, and assistance than I will ever be able to recount, and for that, I am indebted to each of them and hope to be a blessing to each of them in return.

My fellow graduate students in my PhD cohort provided extensive tutoring, guidance, discussion, entertainment, and assistance in my journey, for which I am ever grateful. The faculty of Purdue University's department of Agricultural Economics provided excellent counsel and mentorship, molding me to be become the researcher I am today, as well as teaching me how to provide an education to my students. Foremost of the faculty is my major professor, James K. Binkley. His guidance throughout my graduate career and beyond was invaluable and enlightening, for which I am inexpressibly thankful. My committee, both past and present, shaped me into a more rigorous contributor to our field of study through their patience and expertise: Raymond J.G.M. Florax, Heather A. Eicher-Miller, Michael S. Delgado, and Joseph V. Balagtas. I am forever indebted to each of them for his or her unique role in my education.

The faculty of Murray State University openly welcomed me back as family before my degree had even been officially scheduled for completion. Their enthusiasm to provide opportunity for me to begin my academic career imbued me with much confidence in finishing my degree, for which I am very grateful. While the duration of my full-time employment at Elanco Animal Health's team of Sales Reporting and Analytics was too short, the amount I learned and the skills I developed will serve me well for the rest of my life. The patience and flexibility you provided me to finish my degree was invaluable to my family and me. I wish to express my gratitude to all who contributed to this time of both professional and personal growth – may our friendship continue henceforth.

TABLE OF CONTENTS

LIST OF TABLES 7
LIST OF FIGURES 9
ABSTRACT 11

PART I. GEOGRAPHIC VARIATION IN LAND VALUE GROWTH THROUGHOUT THE CORN BELT: EFFECTS OF GROWTH IN THE ETHANOL INDUSTRY

CHAPTER 1. INTRODUCTION 14
CHAPTER 2. THEORY 17
CHAPTER 3. METHODS 19
3.1 Data 19
3.2 Groundwork for Model Design 26
3.3 Model Identification 30
3.4 Model Estimation 34
CHAPTER 4. RESULTS 38
4.1 Econometric Estimation (OLS) 38
4.2 Loan Rate & Ethanol Selection on Observables 41
4.3 Generalized Propensity Score Estimation 43
4.4 Econometric Estimation (GPS) 49
CHAPTER 5. CONCLUSIONS 59
REFERENCES 62
APPENDIX 64

PART II. TRACKING SPATIAL HEALTH PATTERNS WITH GEOGRAPHIC VARIATION IN GROCERY PURCHASING

CHAPTER 6. INTRODUCTION 71
CHAPTER 7. METHODS 74
7.1 Data 74
7.1.1 Food Purchasing. 75
7.1.2 Demographics, Behavior, and Environment. 80
CHAPTER 8. RESULTS 84

8.1 Regression Analysis 85
8.2 Elastic Net 94
8.3 Discrete Market Groupings 98
CHAPTER 9. CONCLUSIONS 105
REFERENCES 107
APPENDIX 109

PART III. DEFICIENT DIETARY BEHAVIOR IN LOW-INCOME AMERICANS: ASSESSING THE ROLE OF DIET COSTS

CHAPTER 10. INTRODUCTION 114
CHAPTER 11. LITERATURE REVIEW 117
11.1 Literature for/against the Affordability Axiom 117
11.2 Validation of Dissenting Literature 119
CHAPTER 12. THEORY 125
12.1 Income and Constrained Utility 125
12.2 Expected Utility and Health as a Means to Longevity 126
CHAPTER 13. THEORY 1 METHODS 128
13.1 Data 128
13.1.1 Diet Cost. 129
13.1.2 Diet Quality. 130
13.1.3 Palatability. 131
13.1.4 Diet Variety. 137
13.2 Models: Choosing Healthiness, Tastiness, and Costliness of Diets 138
13.3 Results 140
CHAPTER 14. THEORY 2 METHODS 146
14.1 Foundations 146
14.2 Data 148
14.2.1 Low-Fat Milk Purchasing. 148
14.2.2 Household Income. 149
14.2.3 Milk Prices. 150
14.2.4 Demographics. 150

14.3 Model 151
CHAPTER 15. CONCLUSIONS 155
REFERENCES 158

LIST OF TABLES

Part I:

Table 1: *Average Corn County Loan Rate (cents/bushel) Summary Statistics* 20

Table 2: *2007-2012 Land Value Average Annual Change (Percent) Summary Statistics* 22

Table 3: *Ethanol Market Influence Summary Statistics* 25

Table 4: *Estimation Results (pre-Staggers Act vs. post-Staggers Act)* 29

Table 5: *OLS (modified Equation 6)* 39

Table 6: *Selection Stage (Loan Rate, Equation 2)* 41

Table 7: *Selection Stage (Ethanol Market Influence, Equation 3)* 42

Table 8: *Outcome Equation Estimation Results (Equation 4)* 50

Table 9: *Average slope of "blue" outcome function (Low Loan Rate, Weak Ethanol Market)* 57

Table 10: *Average slope of "green" outcome function (Low Loan Rate, Strong Ethanol Market)* 57

Table 11: *Slope of "blue" vs. "green" outcome functions t-test* 57

Part II:

Table 12: *Summary Statistics of Food Expenditure Shares* 79

Table 13: *Market-Level Health Measure Summary Statistics* 80

Table 14: *Demographics and Environment Summary Statistics* 82

Table 15: *Univariate OLS Estimation Results* 85

Table 16: *OLS Estimation Results for Obesity (Univariate Food: Soft Drinks)* 87

Table 17: *OLS Estimation Results for Diabetes (Univariate Food: Vegetables)* 87

Table 18: *OLS Estimation Results for Colon Cancer (Univariate Food: Bacon/Sausage)* 88

Table 19: *OLS Estimation Results for Stomach Cancer (Univariate Food: Beef)* 89

Table 20: *OLS Estimation Results for Stroke (Univariate Food: Seafood)* 90

Table 21: *OLS Estimation Results for Heart Disease (Univariate Food: Nuts)* 91

Table 22: *Food + Non-Food OLS Estimation Results* 92

Table 23: *Output from Elastic Net Algorithm* 96

Table 24: *Food Purchasing Differences between the Top and Bottom 10 Markets* 99

Table 25: *Correlating Health and Posterior Probabilities* 100

Table 26: *Correlations of Posterior Probabilities 1 and 3 with Food Purchasing, Mean Differences, and T-Tests* 101

Table 27: *Product Module Purchasing Differences* 103

Part III:

Table 28: *Price Regressed on Weighted Nutrient Density (OLS)* 120

Table 29: *Price Regressed on Calorie Density* 122

Table 30: *Average Diet Cost ($/100g)* 130

Table 31: *Nutrient Intakes as Predictors of Importance of Tastiness Relative to Healthiness* ... 135

Table 32: *Examples of Palatability & Nutritional Quality (NHANES 2007-2010)* 136

Table 33: *Diet Variety* 138

Table 34: *OLS estimation of Equation (3)* 140

Table 35: *Descriptive Statistics Underlying Figure 4* 144

Table 36: *Low Fat Milk's Average Share of Total Milk Purchases by Income Range* 149

Table 37: *Milk Prices by Fat Content* 150

Table 38: *Estimation Results* 152

LIST OF FIGURES

Part I:

Figure 1: *1998-2001 Average Corn County Loan Rate (cents/bushel)* 20
Figure 2: *2007-2012 Land Value Average Annual Change (Percent)* 22
Figure 3: *2007-2012 Ethanol Market Influence* 25
Figure 4: *Central Hypothesis* 27
Figure 5: *Marginal Effect of Loan Rate with 95%CI* 40
Figure 6: *Unadjusted Covariates Balancing* 44
Figure 7: *GPS-adjusted Covariates Balancing (Bivariate Normal)* 45
Figure 8: *Map of Bivariate Normal GPS* 46
Figure 9: *GPS-adjusted Covariates Balancing (FMM Estimation)* 48
Figure 10: *Map of FMM-Estimated GPS* 49
Figure 11: *3D Dose-Response Surface* 51
Figure 12: *3D Dose-Response Surface* 52
Figure 13: *3D Dose-Response Surface* 53
Figure 14: *3D Dose-Response Surface* 54
Figure 15: *Cross-Sections of 3D Dose-Response Surface* 55

Part II:

Figure 16: *Layout of the 52 Nielsen Scantrack Markets* 75
Figure 17: *2010 Grocery Expenditure Share of Bacon & Breakfast Sausage* 77
Figure 18: *2010 Grocery Expenditure Share of Nuts* 78
Figure 19: 2014-2016 *Average Stroke Death Rate per 100k population* 80
Figure 20: *2012 Average Annual Per Capita Expenditure on FAFH as a Percentage of Household Income* 82

Part III:

Figure 21: *Scatterplot of Food Price and Nutritional Value* 120
Figure 22: *Scatterplot of Food Price and Calorie Density* 121
Figure 23: *Distribution of Importance of Taste relative to Importance of Nutrition* 133

Figure 24: *Marginal Effect of Healthfulness on Cost*.....141
Figure 25: *Marginal Effect of Tastiness on Cost*.....141
Figure 26: *Distributions of Diet Costs for Healthiest, Unhealthiest Quartiles*.....143

ABSTRACT

Author: Young, Jeffrey, S. PhD
Institution: Purdue University
Degree Received: December 2018
Title: Three Economic Issues in Health and/or Space
Committee Chair: James K. Binkley

This dissertation covers three separate topics. The first and second are related, the second and third are related, the first and third share no commonalities. Essay number one was previously investigated and therefore builds off the previous work to apply the methods and framework to a recent issue facing agriculture in the US. Essay number two is purely exploratory and offers methodological insights for the purpose of tracking health patterns at a regional level, as well as investigating new hypotheses regarding food and health. Essay number three contains two parts: one original, the second building off previous work and more rigorously investigating a widely discussed economic question.

The first essay addresses the spatial variation in land value growth in the Corn Belt. It is well established that land values in remote regions (a result of higher transport costs) are more vulnerable to volatility and thus incur risk for landowners. Two historical examples of this are the Great Depression and the 1980s Farm Financial Crisis. The study finds that the Staggers Act may have contributed to lower transport costs in the remote parts of the Corn Belt enough to stabilize land values a little. Similarly, the Ethanol Boom stabilized land values by way of bringing the market to farmers in these regions, thereby lowering their transport costs.

The second essay is an ecological study investigating links between food purchasing and health outcomes at a market level. Attention is given to links established at an individual level from findings in longitudinal and cohort studies, as well as meta-analyses and cross-sectional studies. The results indicate that many of these previously investigated diet-disease links appear in food purchasing patterns by region, and that regional food marketing data can be useful for nutritional epidemiological studies – within the limitations of an ecological study.

The third essay tests the hypothesis of whether healthy eating is necessarily more expensive. If so, then the lack of compliance with the Dietary Guidelines for Americans (DGA) common in low-income adults comes as no surprise. If not, then alternative explanations must be

offered and validated. A review of the literature finds that studies purporting the notion that healthy foods are more costly tend to use flawed cost metrics, and that there is a growing body of dissenting literature. Two alternative theories are proposed and tested. The findings generally support the theory in both cases. Thus, the study recommends that emphasis be placed on measures intended to improve diets through other avenues than cost.

PART I. GEOGRAPHIC VARIATION IN LAND VALUE GROWTH THROUGHOUT THE CORN BELT: EFFECTS OF GROWTH IN THE ETHANOL INDUSTRY

CHAPTER 1. INTRODUCTION

In the 1980's, there was a widespread financial crisis among American farmers. Interest rates reached historically high levels, commodity prices bottomed out, and land values began to decline. One interesting aspect of this familiar "Farm Financial Crisis" was that the prevalence of bankruptcies declared by farmers varied systematically throughout space (Benirschka & Binkley, 1994; Archer & Lonsdale, 1997). Specifically, there were more frequent and increasingly severe cases in the West and Central US.

During the 1960's and 1970's, land values, in general, rose steadily. During this period, farmers were aggressively expanding their operations and taking out large loans against their land to facilitate such expansions (Bultena et al., 1986), with land values being higher each subsequent year. However, in the western end of the Corn Belt and in the Great Plains, this rise in land values was much faster than elsewhere in the US.

Land values plateaued and then began to crash throughout the 1980's. Farmers who were in the process of paying back their massive loans quickly realized that their land was worth less than the loans borrowed against it shortly before (Bultena et al., 1986). Just as the previous period of land value growth, the drop in land values was dramatically more pronounced in the western end of the Corn Belt and in the Great Plains, and thus, it was farmers in that area who were more vulnerable to the crashing land values. It was this region where land values fell the most rapidly and widespread bankruptcies ensued.

From the last of the 1980s on through the 1990s, land values had stabilized and began to grow slightly after bottoming out during the Crisis (Archer & Lonsdale, 1997). In 2014, land values in much of the US had peaked and then began to fall for the first time since the Crisis. Observationally there appears to be minimal concern regarding a repeat of the Crisis, although

there have been isolated occurrences of bankrupted farming operations after commodity prices fell dramatically and land values began to slide (Huffstutter, 2016). What remains to be seen is whether widespread bankrupted farming operations might occur in a similar fashion in the near future as they did previously.

The motivation behind Benirschka & Binkley (1994) was to show empirical evidence of geographic disparity in land value changes, and connect the evidence of this phenomenon to the spatial trend of farm bankruptcies. One factor determining changes in land value is the land's location relative to terminal markets for the commodities produced on that land. In more remote regions, there are greater percentage changes in land values because the residual value[1] of land is discounted by cost of transport to market (Alonso, 1964). In the case of US farmers in the Corn Belt and Great Plains in recent years, it is reasonable to expect either the corn market to have changed or cost of transport to market to have changed, if not both.

One obvious candidate for causing these changes is the Staggers Act of 1980, which deregulated rail freight rates in the US, leading to the expanded use of unit trains in the Western Corn Belt in subsequent years (and, consequently, lower rail rates for grain transport - see Fuller et al., 1983; Koo et al., 1993; Caves et al., 2010). MacDonald (1989) also shows evidence that the Staggers Act led to rail rates declining the most in the Central Plains, and very little in the rest of the Corn Belt.

Shortly after the New Millennium, farmers experienced a boost in commodity prices stemming from the period commonly known as the "Ethanol Boom". The emergence of ethanol plants began in the eastern end of the Corn Belt, but moved west to the more remote parts of the Corn Belt and the Central Plains (Sarmiento et al., 2012). This is where transport costs to terminal markets tend to be highest (MacDonald, 1989; Benirschka & Binkley, 1994; Archer &

[1] In economics, this is referred to as "rent" or "producer surplus".

Lonsdale, 1997). With ethanol plants effectively bringing the market closer to this region, the significance of transport costs for corn likely diminished (Henderson & Gloy, 2009; Hofstrand, 2009; Miller, 2015).

Besides the Staggers Act and Ethanol Boom, there have been other events and changes to infrastructure in the US since the initial time period examined by Benirschka & Binkley (1994) from 1969 to 1982. Examples include the Inland Waterway Trust Act of 1978, the US Interstate Highway System from the 1960's to 1990's, and increased size and horsepower of trucks transporting grain over short distances.

Hence, it is reasonable to expect the magnitude and importance of grain transport costs to have diminished over time. Even if this is not the case, revisiting the Benirschka & Binkley (1994) model can speak to the severity of the recent downturn in the US Ag Economy.

CHAPTER 2. THEORY

The underlying theory behind the geographic heterogeneity in land value changes comes from the works of Ricardo's essay on rent and Von Thünen's work on market distance. Let the value of land at time t be V_t. Land value depends on discounted returns, taking the form $V_t - d$ where d includes all discounting associated with production costs and transport costs. For the purposes of this study, the analysis will address the cost of transport to market, holding costs of production related to land quality constant. That is, the value of land at time *t* is denoted as $V_t - c$, which monotonically declines with transport costs c. Thus the farther from market a land parcel is located, the lower is its surplus value as a direct result of incurred transportation cost a producer located at that parcel faces. This has implications for how the value of a parcel of land changes over time. Mathematically, the percentage change in land value over two time periods *t* and *t* - 1 is calculated as a function of cost *c* of transporting grain to market, taking the form

$$\%\Delta V = \frac{(V_t - c) - (V_{t-1} - c)}{V_{t-1} - c} = \frac{(V_t - V_{t-1})}{V_{t-1} - c} \qquad (1)$$

If transport costs, *c*, reduced following an event such as the Staggers Act, then the overall fraction in Equation (1) will diminish due to the larger base. Due to data limitations discussed later, any model can identify only the effect on this relation stemming from the Ethanol Boom, but the same principle is generalizable to any other events like the Staggers Act.

The chief contribution of this study is two-fold: first, build a model capable of testing for evidence of the relation shown in Equation (1) between market distance and land value growth in the US Corn Belt; secondly, use that model to provide insight into the current economic climate faced by US farmers. While the Staggers Act did lower transport costs faced by corn farmers in

remote regions, due to limitations in the data described below, it is not possible to identify any more than a difference in the estimates before and after the Staggers Act, rather than saying that the difference is entirely attributable to this event. Hence, the main event in focus for the study is on the impact on market distance and transport costs brought about by the Ethanol Boom. To do so, I test the following hypothesis:

- ***Hypothesis*:** Land values in areas near corn ethanol refineries are more stable than areas removed from a local ethanol market because of lower transport costs as a direct result of the Ethanol Boom.

CHAPTER 3. METHODS

In order to test this hypothesis, I construct an econometric model estimating causal effects of changes in transport costs on land value growth. The outcome variable is the average annual change in a county's average farmland value. This is a function of a bivariate treatment: cost of transport and presence of a local ethanol market. The first, cost of transport, cannot feasibly be measured below the county level. Hence, an alternative is to exploit the spatial variation in the corn county loan rate — which is appropriate for this problem because it reflects local corn prices discounted by cost of transport to markets, as per the Law of One Price under spatial arbitrage — in the same fashion as Benirschka & Binkley (1994). The second treatment measures the influence of ethanol plants. This is computed as the sum of the ratios of operating capacity to distance from the county centroid for each ethanol plant in operation as of 2007-2012. An attractive feature of measuring ethanol market influence with this plant-size-relative-to-plant-distance formula is that the result declines with distance and increases with capacity at a plant, thus the largest change in influence would arise from a change in capacity at a nearby plant.

3.1 Data

Figure 1 shows the county loan rate, which reflects corn transport costs (Benirschka & Binkley, 1994; Westcott & Price, 2004). Since the counties in Iowa and the Great Plains states are the most remote in the sample, the cost of transporting the grain produced there to terminal export markets is highest, and therefore the loan rates are the lowest. Computing the change in loan rate from 2007-2012, the counties in which at least one ethanol plant was constructed had the largest increase in loan rates. There are at least two possible reasons why this happened: (i) supply/demand equilibrium readjusted to a new, local source of demand for corn, and (ii) as

posited by Miller (2015) and Hofstrand (2009), the final corn market became local for the corn farmers, making location less important. Because supply and demand conditions altered local corn prices and therefore loan rates, I choose the pre-Ethanol Boom loan rate, averaged from 1998-2001. This represents market distance in a world changed by the Staggers Act, but still untouched by the Ethanol Boom. Thus, any changes in land values as a function of market distance would be most reliably estimated using this period.

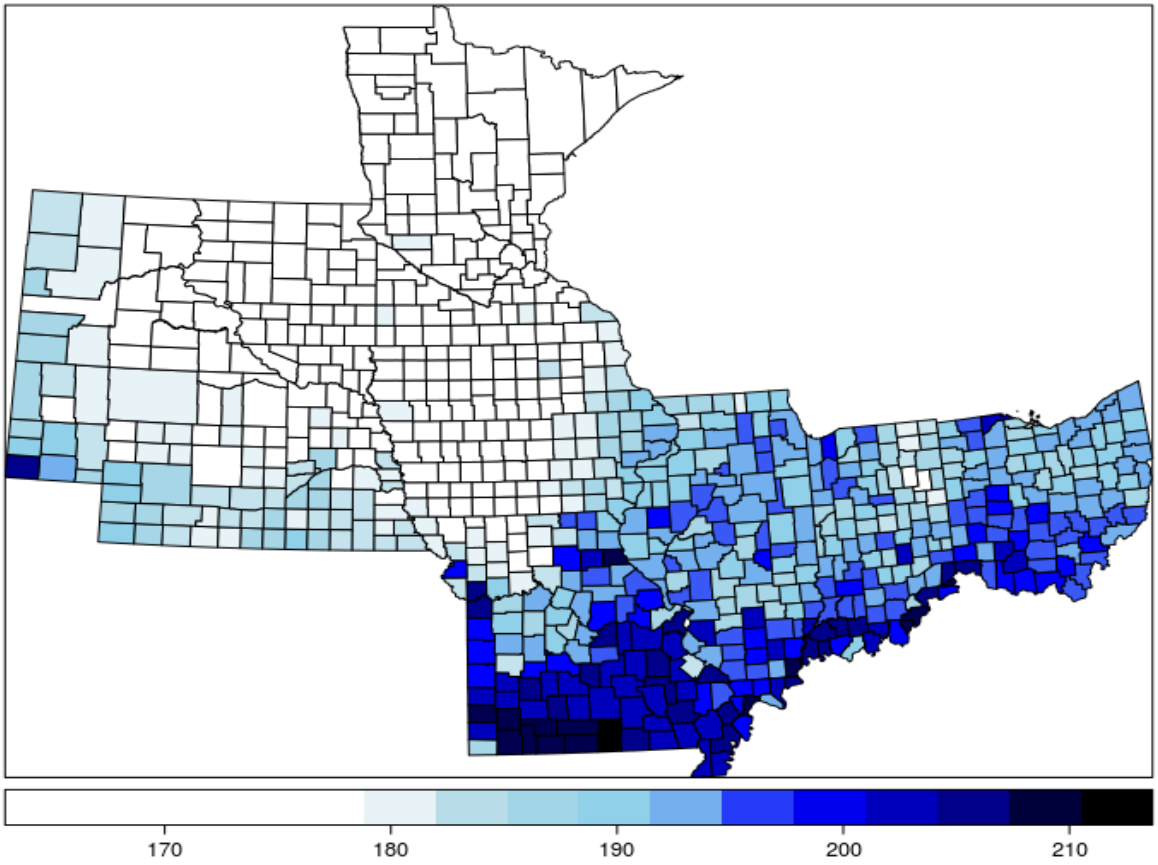


Figure 1: *1998-2001 Average Corn County Loan Rate (cents/bushel)*

Table 1: *Average Corn County Loan Rate (cents/bushel) Summary Statistics*

Mean	StdDev	Min	Max	Median
186.20	10.82	166.10	210.60	186.80

Below in Figure 2 is the average annual percentage change in county land values from 2007 to 2012, which is the most recently available data from the USDA Economic Research Service's Ag Census, administered every 5 years. I compute growth in each county as the natural logarithm of county land value in 2012 and subtract the natural logarithm of county land value in 2007, dividing the difference by the number of years between the two. Land values generally grew during this period. As found in Benirschka & Binkley (1994), the largest increases occurred in Iowa and the Great Plains states. This is the same pattern observed previously in the 1970s, and is consistent with the underlying microeconomic theory in that these counties are the most remote in the sample. Thus, the highest transport costs in the sample would fall to farmers in those counties.

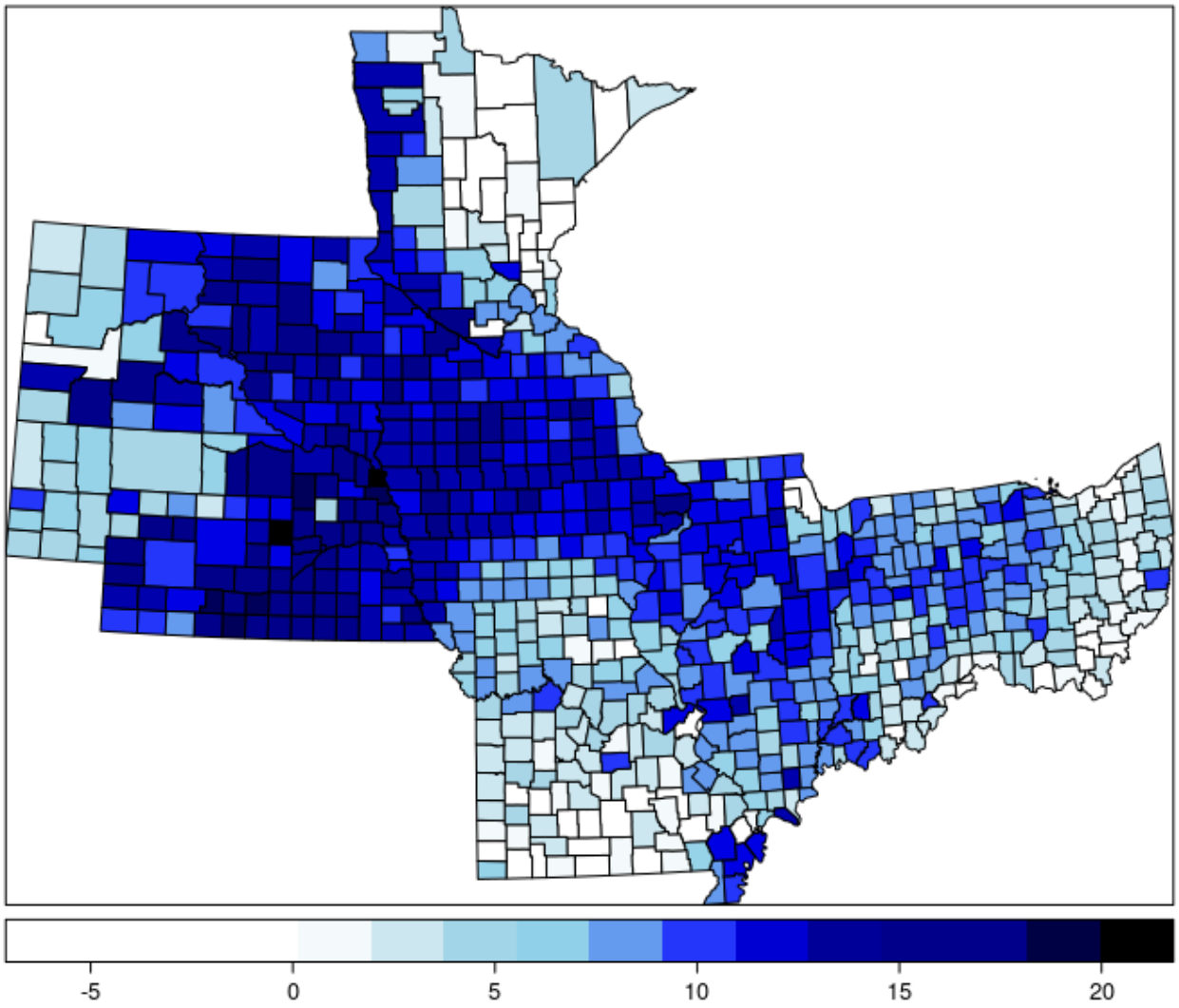


Figure 2: *2007-2012 Land Value Average Annual Change (Percent)*

Table 2: *2007-2012 Land Value Average Annual Change (Percent) Summary Statistics*

Min	Max	Mean	StdDev	Median
-5.33	20.01	8.47	4.91	8.72

Depicted in Figure 3 is the growth of influence from a county's local corn ethanol market. In its simplest form, this variable could be the difference of two dummy variables of whether or not a corn ethanol plant is in operation in that county in two different time periods – a difference of 1 indicating a county increased in the number of plants by 1 over period 1 to period 2.

However, this has three shortcomings. First, not all plants are of equal operating capacity. Second, counties themselves are not equal in terms of size and shape. Third, distance from a plant to a farm in a county is important. For example, an ethanol plant one mile from a county's center has more power to influence farmland values in that county than a plant 100 miles away could have.

Hence, a more meaningful measure of ethanol market influence is required. One possible method adopted by Motamed et al. (2016) is to disaggregate the geography into a grid of small square units, compute the market procurement area of a plant based on local corn prices and plant capacity. Their purpose was to model the response of local corn production to the growth of a local ethanol market. However, my question concerns average land value patterns between counties, not micro-level corn production. Furthermore, this data not publicly available. Another method is a method in a PhD dissertation by Wang (2017) calculates a market procurement area of each plant as a function of plant capacity and local corn production relative to total crop production. The formula for this market area defines a circular market procurement area from a draw radius of each ethanol plant – a two-dimensional case of the well-known Hotelling line. This measure focuses on ethanol plants themselves and how far away they are from their corn supply, which is valuable in the appropriate context, but is of limited relevance in my case because I am focusing on land and its nearness to many plants of varying capacities. In my context, using the 2D Hotelling line to measure ethanol market presence would assume that land values in a county on the edge of a plant's circle is equally "treated" by the ethanol market as those in a county at the epicenter, as well as portions of counties falling under multiple circles. To examine empirically whether this is too strong an assumption to make, I enlisted the GIS expertise of Alexander Young to compute these market procurement area circles in ArcGIS

using capacity and location data obtained from the Renewable Fuels Association, and found that there were multiple questionable observations arising in the sample. For example, there appeared multiple counties on the fringe of corn-producing areas that were still counted as being inside a market procurement area, or even in the overlap of several of these circles (as many as 12 in some counties). It is not at all clear what kind of weighting scheme is warranted by the prevalence of overlapping circles – given that county area falling under one circle should receive a lower weight than that under two circles and so on.

This study builds off of an alternative strategy by constructing a measure of the strength of a county's local ethanol market based on plant distance and capacity. The problem studied here is one concerning changes in county average land values, therefore I assume that the centroid (the geometric average) is the best choice of measuring average location in a given county. The calculation for the final ethanol market influence E_i in county i takes the form

$$E_i = \frac{C_1}{D_{i1}} + \frac{C_2}{D_{i2}} + \cdots = \sum \frac{C_j}{D_{ij}}$$

where C_j is the 2007-2012 average capacity of the j^{th} plant in millions of gallons and D_{ij} is the distance in miles from the centroid of county i to plant j. Figure 3 below shows the geographic pattern of this variable and Table 3 shows descriptive its statistics

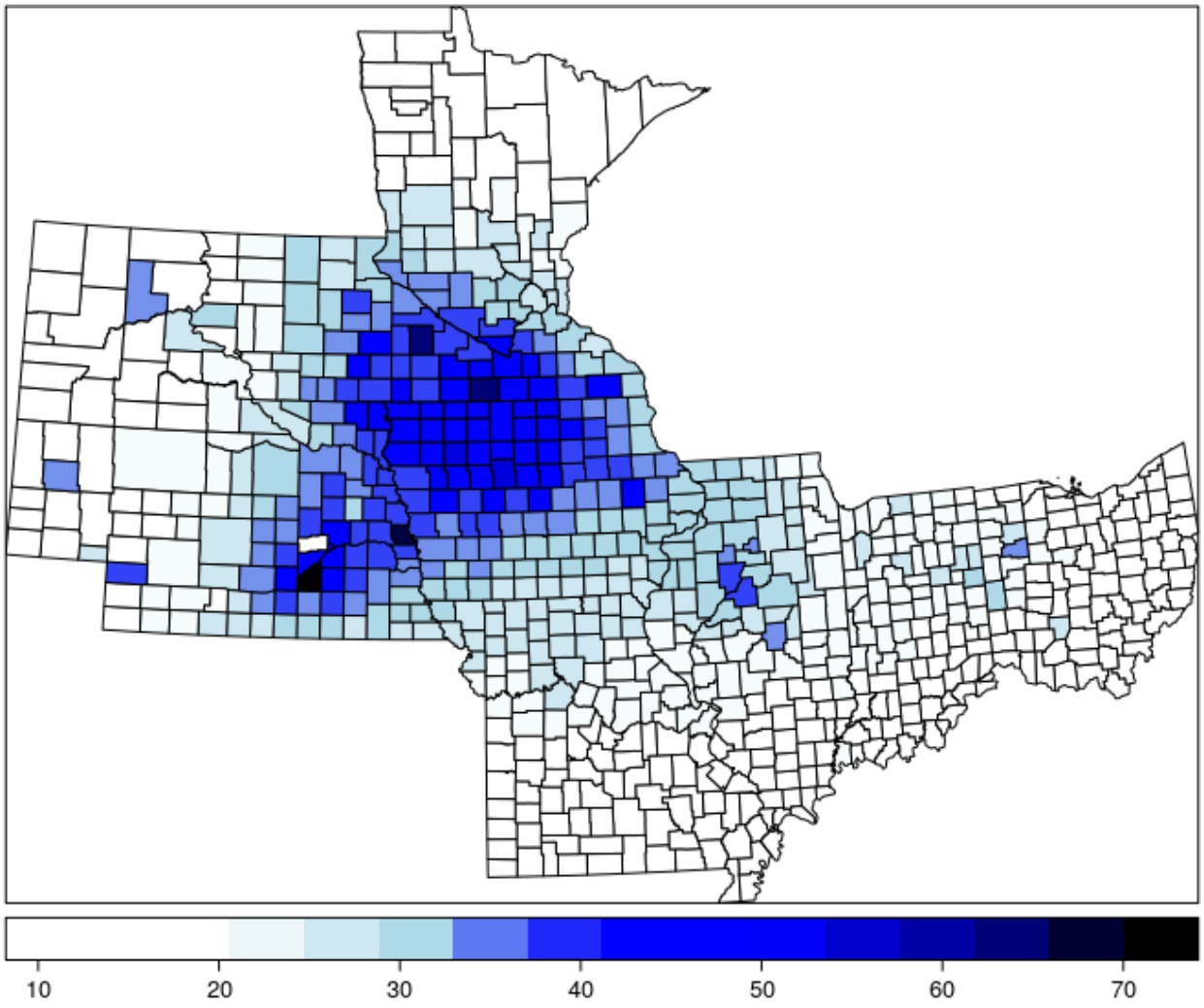


Figure 3: *2007-2012 Ethanol Market Influence*

Table 3: *Ethanol Market Influence Summary Statistics*

Mean	StdDev	Min	Max	Median
25.69	9.47	12.26	70.10	23.34

At the mean over 2007-2012, there are 25 million gallons of ethanol plant's operating capacity for every mile of Euclidean distance from a plant to a county's centroid. It is clear that this ratio is at its highest values in the western Corn Belt in Iowa and Nebraska. This serves as a mark of validation, since the map in Figure 3 showing counties with high values of influence is consistent

with a 2013 USDA report mapping corn production and ethanol plant locations, shown in Appendix A-1.

3.2 Groundwork for Model Design

It is reasonable to expect land value's responsiveness to market distance to have changed along with the market itself following the Ethanol Boom. As stated in the introduction, the majority of ethanol plants set up operations in the more remote parts of the US, such as the edge of and interior to the Great Plains. Corn growers in remote regions receive lower prices due to higher transport costs, but a local ethanol plant functions as terminal delivery point for locally produced corn.

It is expected for market distance to decline, effectively, for farmers located near a corn ethanol plant. If this is the case, and transport costs faced by these farmers improves, then the base of the fraction in Equation (1) increases and the percentage change in land value decreases proportionately. In the regression context, the "Benirschka Effect" of a negative marginal effect of loan rate on land value growth is likely to have been mitigated; that is, we should observe a flatter estimated slope on the county loan rate, if, indeed, the transport costs faced by corn farmers have been reduced. This is illustrated in Figure 4

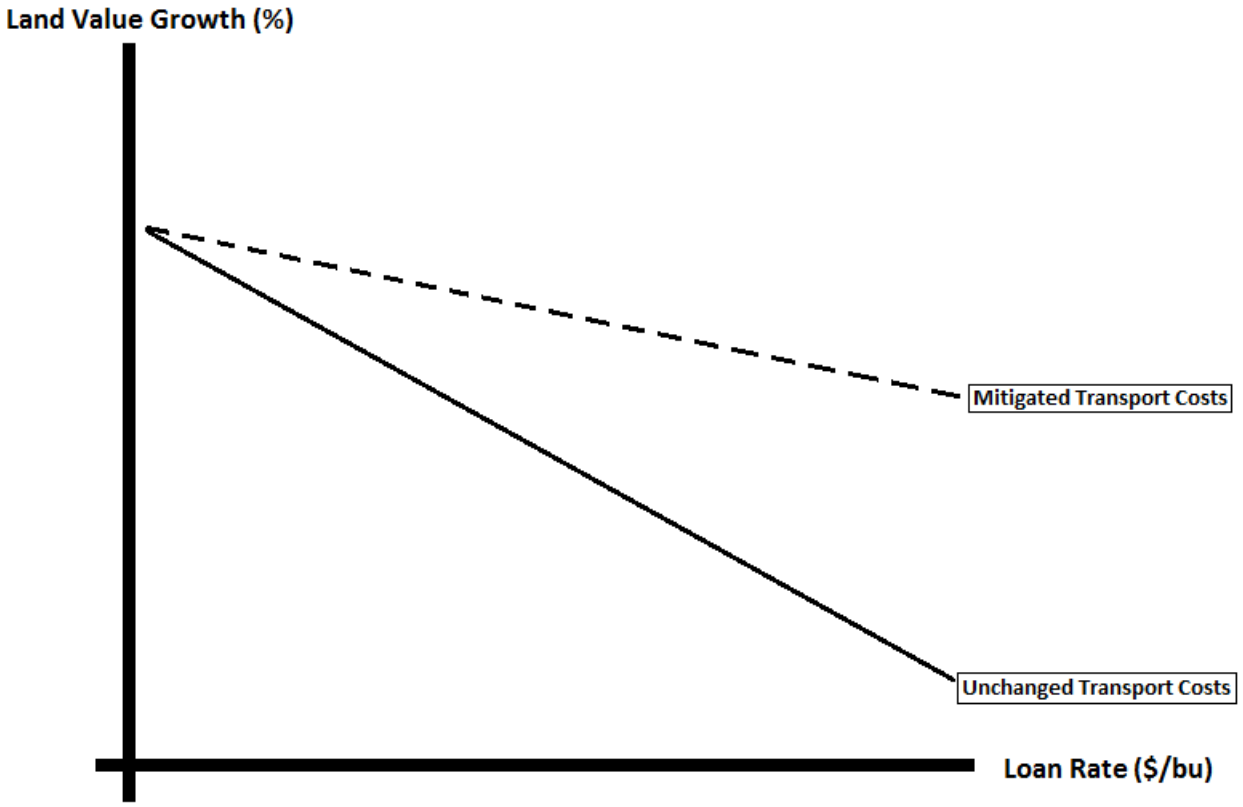


Figure 4: *Central Hypothesis*

In Benirschka & Binkley (1994), the result was a downward sloping line, indicating that as loan rates improved (transport costs dwindled) land values grew at a more stable rate. If the importance of market distance is less, then the slope should be less steep, as seen in Figure 4 where the slope of the solid line more negative than that of the dashed line.

As mentioned in the hypothesis motivation, a limitation of this study is the availability of data, specifically, county loan rates. Only county loan rate data beginning in 2007 are publicly available. Barry Goodwin provided these same data for the years 1998-2001. The original data cards for Martin Benirschka's PhD dissertation appear to be lost or destroyed, although his dissertation document did have these data for each of the USDA Crop Reporting District from 1970-1980, each district being a cluster of counties in the same state as the district. I use these

data to attempt replication of Benirschka & Binkley (1994) in order to test for differences arising from the Staggers Act, the Inland Waterway Trust Act and similar events. As a result, rigorous investigation can address only the effect of the Ethanol Boom of the early 2000s. There have been other factors such as the deregulation of rail rates, user charges on barge transportation, and the interstate highway system. For these, a limited investigation is still feasible.

The 1998-2001 data are the loan rates I use in order to avoid bias arising from any equilibrium effects in the post-Ethanol Boom loan rates. The extent to which I could address any impacts from the Staggers Act is to replicate the model estimated in Benirschka & Binkley (1994) which covers 1970-1980, estimate the same model using the 1998-2001 data, and test for differences in the estimated slopes on county loan rate. However, any differences found may not be completely due to lowered rail rates via the Staggers Act – or any other of the multiple infrastructural improvements, for that matter. It is for this reason that the hypothesis and the econometric model of this study concern the Ethanol Boom by itself – although, I contend that Equation 1 lends external validity to my findings.

Before estimating the full model and evaluating the impact of the Ethanol Boom on the "Benirschka Effect", I first address the impact of the Staggers Act and other infrastructural changes, which occurred at the end of and after the period evaluated by Benirschka & Binkley (1994). To do so requires replicating the results found in their paper using the USDA Crop Reporting District level data printed in Martin Benirschka's doctoral dissertation and then compare to the same model evaluated at post-Farm Financial Crisis and pre-Ethanol Boom data under the same econometric model. Rather than 495 counties as in Benirschka & Binkley (1994), the sample is aggregated up to those counties' respective 45 Crop reporting Districts for both pre and post-1980s. The model takes the form

$$y = \beta_0 + \beta_1 L + \boldsymbol{X\alpha} + \mu \qquad (6)$$

where L is the 1970-1980 county loan rate[2] and $\boldsymbol{X}$ contains for each county the average corn yield (bushels/acre), population density (persons/square mile), average annual change in population (percent), average farm size (acres/operation), and the percentage shares of soil capability classes 1-4. If the Staggers Act, interstate highways, and general advancement and improvement of infrastructure affected transport costs enough to stabilize land values, the estimated slope, $\hat{\beta}_1$, should be closer to zero after these events occurred. Of course, how much of the difference is attributable to the Staggers Act or any other event/policy cannot be identified using these data, but the difference itself is more informative than none at all. Table 4 shows the coefficient estimates slopes with a binary intercept-slope shifter, *D*, denoting before and after this time period, or 1970-1980 and 1998-2001, respectively

Table 4: *Estimation Results (pre-Staggers Act vs. post-Staggers Act)*

Variable	Estimate	t-statistic
Intercept	13.5310	2.20
Loan Rate	-0.0420	-0.84
Farm Size	-0.0034	-0.74
Corn Yield	0.0047	0.23
Pop. Growth	0.0000	0.00
Pop. Density	-0.0022	-1.73
Class1 Soil %	-0.0223	-0.47
Class2 Soil %	-0.0267	-1.80
Class3 Soil %	-0.0288	-1.29
Class4 Soil %	-0.0299	-0.88
D	-5.0568	-0.54
*Loan*D*	0.0373	0.50
*FarmSize*D*	0.0006	0.11
*CornYield*D*	-0.0087	-0.37
*PopGrowth*D*	-0.0461	-0.11
*PopDensity*D*	0.0019	1.08
*Class1*D*	-0.0944	-1.44

[2] I use this "pre" loan rate for the same reasons mentioned previously concerning any potential confounding of the coefficient if, indeed, transport costs themselves have changed.

Table 4 continued

*Class2*D*	-0.0111	-0.50
*Class3*D*	0.0347	1.14
*Class4*D*	0.0542	1.08
$\boldsymbol{R^2}$	0.77	

The pre-Staggers effect is quite close to the county-level estimates obtained in Benirschka & Binkley (1994), although it is slightly closer to zero and is not significant. As expected, the slope shifter on loan rate is positive, 0.0373, making the post-Staggers loan rate effect -0.0047. While the direction of this effect is consistent with the theory, the significance is found lacking – an F-test of the joint effect yields a p-value of 0.48. In terms of economic significance, these results would appear to fit with expectations of mitigated transport costs reflected by a change in the estimated "Benirschka Effect". As emphasized, which portions of this statistically insignificant difference can be attributed to the various underlying causes in the 1980s and 1990s cannot be identified, but the conclusion still stands concerning the mitigation of transportation costs' impact on land value growth patterns, or at least a null hypothesis of no worsened land value volatility following these developments.

3.3 Model Identification

The goal is to estimate average causal effects of a change in county loan rate in remote parts of the Corn Belt and examine differences in this causal effect arising from changes in the presence and impact of local ethanol markets. There are two potential (arguably rectifiable) obstacles in the analysis as I have defined it: first, both treatment variables — loan rate and presence of ethanol production — are continuous and a model computing estimates from discrete treatment variables (e.g. a difference-in-differences estimator) is not appropriate. Second, loan

rates are not randomly assigned to counties (Wescott & Price, 1999; Goodwin & Mishra, 2006) and neither are ethanol plants (Lambert et al., 2008; Tigges & Noble, 2012; Motamed et al., 2016). With only three time periods, a more traditional method of estimating causal effects of a continuous treatment (e.g. fixed or random effects) may not sufficiently eliminate bias associated with imbalanced covariates; that is, units (counties) may not be directly comparable to one another (Rosenbaum, 1984) and thus estimated treatment effects obtained from such comparisons may not be unbiased. Another approach would be to assume that treatment assignments are ignorable given past outcomes, i.e., conditional on previous land value growth — which is unaffected by current changes in treatment. This is desirable when the parallel trends assumption is likely violated (O'Neill et al., 2016). However, because my data is limited to three time periods (before, during, and after the Ethanol Boom), this approach used by Ashenfelter (1978), while appealing given more time observations of the data, may not appropriate for identification.

Because of these obstacles to estimating the causal effects of interest, I choose to follow the generalized propensity score approach developed by Imai & van Dyk (2004), Hirano & Imbens (2004), and Egger & Von Erhlich (2013). In short, this controls for differences in the conditional probability of a county possessing a given combination of treatment values. As pointed out by Rosenbaum (1984), this probability is only equal (or close to equal) when two units are directly comparable to one another, in which case average treatment effects computed from such a comparison are credible. If not, this probability of treatment assignment must be controlled for, which led to the development of the estimation of and controlling for the propensity function in Rosenbaum & Rubin (1983), which was later generalized in the papers

listed above. Therefore, I choose this as my strategy for identifying the effects of interest, and justify this choice in the following paragraphs.

For this approach to be valid and causal effects to be identified, there are two assumptions which must hold: (i) the Stable Unit Treatment Value Assumption (SUTVA), which claims that the outcome (land value growth) in county i is independent of the treatment assignment (county loan rate and ethanol market influence) in county $j \neq i$; (ii) the weak unconfoundedness or conditional ignorability assumption. This asserts that the combination of levels of treatment in a county and the outcome (land value growth) in county i are independent conditional on key observable covariates. As proven in Hirano & Imbens (2004), (i) and (ii) imply that selection bias arising from differences in the covariates is substantially reduced or eradicated.

I argue that (i) holds for the loan rate because the land value changes in a county are independent of neighboring counties' relative locations, and therefore loan rates. Of course, transport costs are spatially correlated, but that does not violate this assumption because farm location is fixed. In other words, even if one farmer has a neighboring farmer who is more advantageously located and faces lower transport costs, the more remote farmer cannot pack up and move to a more favorable location, and has no choice but to continue facing the transport costs reflective of the farm's distance to market. I also argue that the assumption holds for my ethanol treatment variable. Paraphrasing the findings in Henderson & Gloy (2009), if a farm at a county's center lies within some distance of an ethanol plant of a particular capacity, then the neighboring county's land values will respond only to the extent that its own land falls under its own distance to that same plant or some other plant.

To ensure that (ii) holds, I carefully select my county-level covariates following the literature concerning loan rates and location decisions of ethanol plant (e.g. Westcott & Price, 1999; Lambert et al., 2008). I include variables correlated with the loan rate concerning the local demand for corn such as location of grain elevators (classified as large and small in the County Business Patterns data reports) and their spatial lags. The spatial lags are computed by the matrix product of the variable, X, and a spatial weights matrix, W, whose entries denote whether the county in row i is a neighbor of the county in column j, where the spatial weighting scheme is immediately contiguous county edges and vertices[3]. This matrix is row-standardized to get an average measure/proportion when spatially lagging a variable. I include these because of their function as either a shipping source to export markets or commercial storage holdings for local corn users including food processors and livestock feedlots as well as ethanol plants. Also included are county proportions of soil capability classes 1 through 4 – the only classes suited for row crop production, class 1 being ideal, class 4 being limited – as well as average corn yield in the county and average size of farming operations in the county.

Key observables for the ethanol treatment variable include the Euclidean distance from nearest ethanol plant to second-nearest ethanol plant in miles, county average corn yield, county population growth and population density and their spatial lags, large and small grain elevators and their spatial lags, county size in square miles[4], and state fixed effects to capture state-level producer tax credits and exemptions.

Besides that obtained from the USDA Ag Census and the county loan rate data from Barry Goodwin, the remainder of my data was collected from the USDA-ERS's National

[3] "Queen" neighbors, referring to moves in the game of chess.
[4] Given that county sizes in my sample are highly variable which could influence the denominator of the ratio (miles to a plant) for a county – although the correlation of that final variable with county square mileage is less than 0.01 – I include county size measured by square mileage as an explanatory variable in the selection stage of my model.

Agricultural Statistics Service, the Natural Resource Conservation Service, the US Census Bureau, the County Business Patterns Database, and the Renewable Fuels Association

3.4 Model Estimation

Apart from loan rates, data from 2007-2012 is used to estimate my model and ultimately test the hypothesis.[5] The estimation takes place in three steps: first, I assume each treatment variable to be a function of observed covariates X^L and X^E for loan rate and ethanol, respectively. Following Imai & van Dyk (2004), Eggers & Von Ehrlich (2013), and Requena-Silvente et al. (2014), I estimate the first-stage selection equations

$$L_i = f(X^L) \qquad (2)$$

$$E_i = h(X^E) \qquad (3)$$

where L_i is the value of the loan rate in county *i,* E_i is the influence of the local ethanol market on that county, and X^L, X^E represent the key observable covariates affecting the two treatment variables. The functions *f* and *h* are estimated by OLS as reduced-form linear regressions, that is, for $X_i = [X^L, X^E]$, $(L_i|X^L) \sim N(X^L\beta^L, \sigma^2)$ and $(E_i|X^E) \sim N(X^E\beta^E, \sigma^2)$. Expected levels of treatment, $\hat{L}$ and $\hat{E}$, are used to estimate the conditional probability of a county having a given combination of loan rate and ethanol presence, *l* and *e,* that is, $\hat{G}(l, e; X) = \widehat{Pr}[L_i = l, E_i = e|X]$.

Upon obtaining $\hat{L}$ and $\hat{E}$, the next step is to plug the residuals $L - \hat{L}$ and $E - \hat{E}$ into a bivariate normal density function, thereby obtaining the general propensity score $\hat{G}(l, e; X) = \widehat{Pr}[L_i = l, E_i = e|X] = \phi(L - \hat{L}, E - \hat{E}|X)$, or the conditional probability of a county having any combination of loan rate and ethanol market presence given observables *X*. In order to verify my chosen form and distributional assumption of the general propensity score (GPS), I will perform

[5] Land value growth rates are computed as the average annual change over this period, others are averages across the period. In other words, while the period spans the 5 year USDA Ag Census timeframe, the calculations of the variables will such that there is one observation per unit.

balancing checks on each variable contained in X^L and X^E for different values of $G(L, E; X)$. This is the procedure suggested in the Hirano & Imbens (2004). However, if the errors from the first stage selection equations are not normally distributed, then I must choose another form of propensity function. In this study, I will first assume that a bivariate normal density is without any violation of the theory presented in Hirano & Imbens (2004) or Imai & van Dyk (2004) and is the "correct" propensity function, but will follow up by estimating a mixture model as my density function. According to the statistical theory underlying the Expectation-Maximization (EM) Algorithm, the joint distribution of the first-stage residuals can be estimated by a weighted combination of finitely many normal distributions, which would include the unknown "true" GPS.[6] I will compare the balancing of the covariates in *X* under both forms of the GPS, the rigid case of a bivariate normal density $\hat{G}(l, e; X) = \phi(L - \hat{L}, E - \hat{E}|X)$ followed by the more flexible mixture density, $\hat{G}(l, e; X) = \boldsymbol{\lambda} * \boldsymbol{\phi}(L - \hat{L}, E - \hat{E}|X)$ where the vector $\boldsymbol{\lambda}$ contains the final mixing proportions λ_i among the mixture components chosen by the EM Algorithm, each component ϕ_i having its own mean and variance vectors.

The next stage is to estimate the outcome equation by regressing land value growth on county loan rate, L, the presence of a local ethanol market, E, the (estimated) conditional probability of receiving that particular level of loan rate and ethanol market influence, G, and interaction terms. In Hirano & Imbens (2004) and Requena-Silvente et al. (2014), a flexible, nonlinear, interactive, linearly parametric functional form is estimated using OLS. The equation takes the form

$$y_i = \beta_0 + \beta_1 L_i + \beta_2 L_i^2 + \beta_3 E_i + \beta_4 E_i^2 + \beta_5 L_i E_i + \beta_6 L_i^2 E_i^2 + \beta_7 G_i(l, e; X) + \beta_8 G_i^2$$
$$+\beta_9 G_i^2 L_i^2 + \beta_{10} G_i^2 E_i^2 + \beta_{11} L_i E_i G_i + \beta_{12} L_i^2 E_i^2 G_i^2 + \epsilon_i \qquad (4)$$

[6] A rough sketch of a proof is in Appendix A-2.

Note that in the midst of this model is the interactions that capture the direct once-and-for-all change in land value due to a change in influence from an ethanol plant and the interaction between loan rate and ethanol. Finally, I compute a dose-response function at pre-specified levels of the two treatments in order to estimate the corresponding average potential outcomes $E[y_i|L_i, E_i, G_i(l, e; X)]$, estimated by the form in Equation 5

$$\hat{E}[y_i(l,e)] = \frac{1}{n}\sum\{\hat{\beta}_0 + \hat{\beta}_1 l + \hat{\beta}_2 l^2 + \hat{\beta}_3 e + \hat{\beta}_4 e^2 + \hat{\beta}_5 le + \hat{\beta}_6 l^2 e^2 + \hat{\beta}_7 \hat{G}_i(l,e;X) +$$

$$\hat{\beta}_8 \hat{G}_i^2(l,e;X) + \hat{\beta}_9 \hat{G}_i^2(l,e;X) l^2 + \hat{\beta}_{10} \hat{G}_i^2(l,e;X) e^2 + \hat{\beta}_{11} \hat{G}_i(l,e;X) le + \hat{\beta}_{12} \hat{G}_i^2(l,e;X) l^2 e^2\} \quad (5)$$

As stressed by Hirano & Imbens (2004), the expected conditional outcome $\widehat{\mathrm{E}}[y_i(l,e)]$ by itself does not have a causal interpretation since the marginal effect of *L* or *E* does not represent an average effect on *y* from changing the level of *L* or *E* for any particular unit(s) of interest. However, comparing $\widehat{\mathrm{E}}[y(\, l_1, \boldsymbol{e_0})]$ to $\widehat{\mathrm{E}}[y(\, l_2, \boldsymbol{e_0})]$ for a fixed value of a local ethanol market's presence in a county, $\boldsymbol{e_o}$, and two different values of loan rate, l_1 and l_2, does have a causal interpretation, an attractive and useful result of my use of treatment effect identification via GPS. The same procedure applies to compute a different average causal effect for two different values of ethanol market influence under a fixed value of loan rate if estimating the treatment effect of ethanol market influence was of interest.

To test the hypothesis of whether the growth of land values following the "Benirschka Effect" stabilized post-Ethanol Boom, that is, whether the growth of the ethanol industry reduced land price volatility, the first step is to calculate the "Benirschka Effect" by itself for a given level of the presence of ethanol. Let $\boldsymbol{e_o}$ be some fixed value of the distribution of the ethanol treatment variable, which can be arbitrarily chosen value depending on which region of the Corn Belt is of interest. Then let l_1 be an arbitrarily low value of loan rate, and l_2 some arbitrarily high

value of the same variable. Then the average causal effect of a change in the loan rate holding fixed the presence of a local ethanol market for a county in the Corn Belt is computed as

$$ACE = (\widehat{\mathrm{E}}[y(l_2, \boldsymbol{e_0})] - \widehat{\mathrm{E}}[y(\, l_1, \boldsymbol{e_0})])$$

The interpretation of this causal estimand is the change in average annual land value growth from 2007-2012 given a change in loan rates by a magnitude of $l_2 - l_1$. Given the underlying economic theory mathematized in Equation (1), this effect is negative since an increase in loan rates implies a reduction in transport cost. Now, in order to test my hypothesis, I compute two causal estimands and compare them. Specifically, the two estimands take the form

$$ACE_{low} = (\widehat{\mathrm{E}}[y(l_1, \boldsymbol{e_{low}})] - \widehat{\mathrm{E}}[y(\, l_2, \boldsymbol{e_{low}})])$$

$$ACE_{high} = (\widehat{\mathrm{E}}[y(l_1, \boldsymbol{e_{high}})] - \widehat{\mathrm{E}}[y(\, l_2, \boldsymbol{e_{high}})])$$

The first of this pair, ACE_{low}, represents the "Benirschka Effect" under a sparse or uninfluential local ethanol market, similar to what would have been the situation before the Ethanol Boom. The second, ACE_{high}, is representative of the world after the Ethanol Boom. Also note that the values of the loan rate l_1 and l_2 should be chosen according to the distribution of loan rates in the more remote counties of the Corn Belt, since that region has been shown to be more vulnerable to volatile land values and therefore subsequent financial turmoil in economic downturn. To test my hypothesis of whether or not the Ethanol Boom helped reduce this volatility, I compute the difference

$$ACE_{high} - ACE_{low}$$

If the hypothesis holds, then this difference will be negative and statistically significant, indicating that transport costs' reduced as a direct result of the Ethanol Boom for land values to have grown more stably in the remotest parts of the Corn Belt.

CHAPTER 4. RESULTS

4.1 Econometric Estimation (OLS)

The use of GPS might present a different estimate of the difference in treatment effects than OLS if, indeed, there is bias associated with imbalanced covariates. However, given that the selection-on-observables assumption of GPS is not empirically testable, I first estimate a modified version of the structural econometric model in Equation 6 using OLS, which tests for any differences in the "Benirshcka Effect" arising from ethanol under the assumption of no bias associated with imbalanced covariates. This is essentially the equivalent to replicating the model in Benirshcka & Binkley (1994) with two key differences. The first is the expansion of sample size from 495 counties to 741 counties, thus including data from three more states in and around the Corn Belt. The second difference is the inclusion of ethanol and its interaction with the loan rate. The marginal effect of loan rate on land value growth is therefore a linear function of the size of a local ethanol market.

The modified form of Equation 6 is estimated with OLS. The estimated loan rate slope in Benirschka & Binkley (1994) was -0.0747 and was statistically significant. The estimated response in counties more influenced by ethanol is less negative than for counties with less impacted by ethanol. Table 5 displays the OLS estimation results

Table 5: *OLS (modified Equation 6)*

Variable	Estimate	t-statistic
Intercept	31.11	5.01
Loan Rate	-0.1966	-5.49
Ethanol	-0.5343	-3.12
*Loan*Ethanol*	0.0035	3.70
Proportion Class 1 Soil	0.1037	5.15
Proportion Class 2 Soil	0.0640	8.21
Proportion Class 3 Soil	0.0378	3.51
Proportion Class 4 Soil	0.0166	0.97
Corn Yield	0.0171	3.26
Average Farm Size	0.0008	8.15
Population Growth	-0.1801	-1.23
Population Density	-0.0021	-5.82
R^2	0.55	

The increased presence of a local ethanol market reduces transport costs and is reflected in the "Benirschka Effect", which is now a function of ethanol. Practically speaking, land values in the more remotely located counties of the Corn Belt are more stable in light of changes in county loan rate for corn because, as consistent with the assertions in Hofstrand (2009), transport costs faced by local corn famers improved following the Ethanol Boom. To visualize this mitigation of the effect of loan rate, Figure 5 shows the marginal effect of loan rate on land value change as a linear function of ethanol (red line) with 95% confidence bands (dashed blue lines)

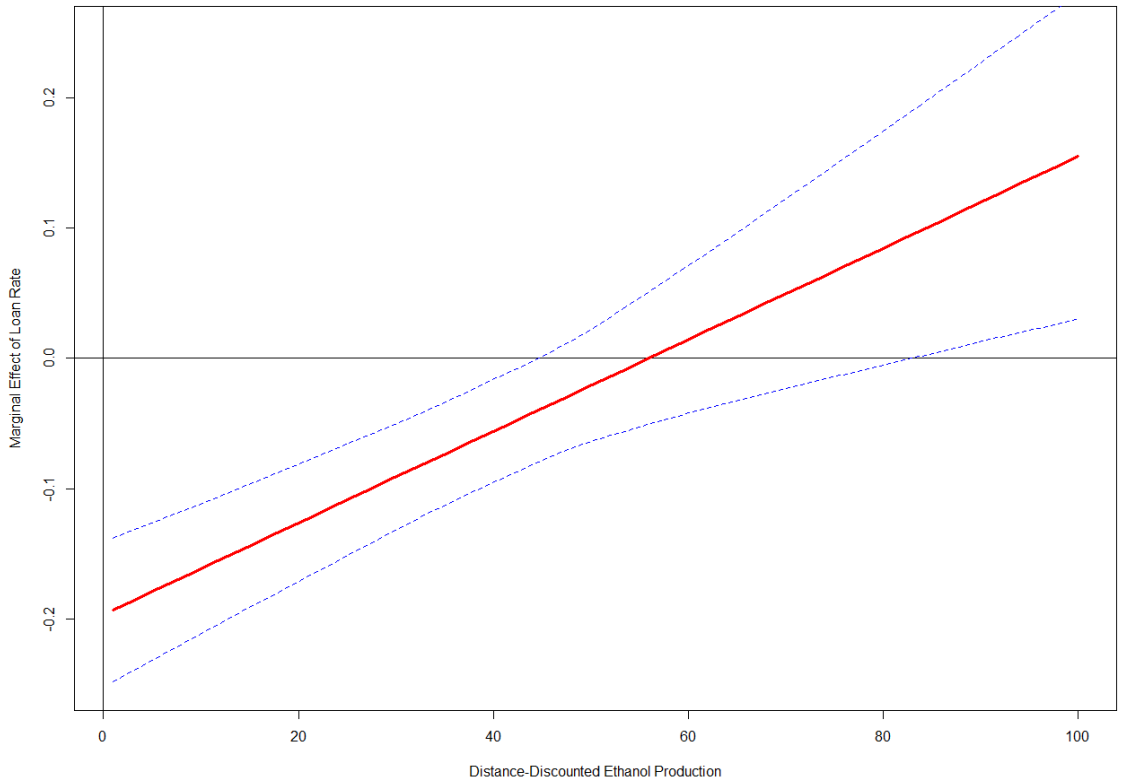


Figure 5: *Marginal Effect of Loan Rate with 95%CI*

The loan rate effect of -0.1966+0.0045*Ethanol is equal to -0.0747 when the ethanol variable is 34.5, about 35% above the mean. It is worth noting that once the ethanol variable exceeds a value of 83, the marginal effect of loan rate is significantly greater than zero at the 5% confidence level, and the positive axis falls outside the 95% confidence interval shown in Figure 5. However, because the maximum value of the ethanol variable is 70.10, the effect of a change in loan rate cannot be significantly positive given these data. These results support the hypothesis that increased influence from the ethanol market stabilizes land values in a county.

4.2 Loan Rate & Ethanol Selection on Observables

It is possible that controlling for the covariates in Equation 6 linearly does not adequately remove bias arising from the nonrandom values of loan rate and ethanol. Hence, the hypothesis is tested using GPS, the output of which is compared to the result of using OLS. To begin, I estimate Equations 2 and 3. The first selection equation for county loan rate includes the key regressors for the selection stages of county loan rate and ethanol market influence, and the estimates are displayed in Table 6.

Table 6: *Selection Stage (Loan Rate, Equation 2)*

Variable	**Estimate**	**t-statistic**
Intercept	207.70	82.53
Small Grain Elevators	0.08	0.43
Large Grain Elevators	0.10	0.24
W(Small Elevators)*	0.92	2.98
W(Large Elevators)*	-4.08	-4.77
Proportion Class 1 Soil	-0.20	-3.43
Proportion Class 2 Soil	-0.11	-4.68
Proportion Class 3 Soil	-0.11	-3.52
Proportion Class 4 Soil	-0.17	-3.57
Corn Yield	-0.07	-4.84
Average Farm Size	-0.00	-11.45
R^2	0.24	

It appears that, for an increase in the average number of large grain elevators in surrounding (spatial lag denoted with spatial weights matrix, *W*) counties – holding fixed the number of small grain elevators – loan rates tend to be lower. Similarly for average corn yield, a negative sign is possibly a reflection of excess corn supply exerting downward pressure on local prices. Using the parameter estimates in the table above, the loan rate fitted values are calculated by the matrix product $\hat{\boldsymbol{L}} = X^{loan}\hat{\boldsymbol{\beta}}$. For the second selection equation, that is, for the selection on observables

concerning ethanol market influence, Equation 3 is estimated and the results are displayed in Table 7

Table 7: *Selection Stage (Ethanol Market Influence, Equation 3)*

VARIABLE	Estimate	t-statistic
Intercept	29.69	23.42
1*{Iowa}*	25.86	19.86
1*{Minnesota}*	18.45	14.56
1*{Missouri}*	9.16	7.79
1*{South Dakota}*	14.12	10.02
1*{Illinois}*	4.44	3.37
1*{Indiana}*	6.27	5.19
1*{Nebraska}*	14.91	11.70
Small Grain Elevators	0.30	1.88
Large Grain Elevators	0.48	1.34
W(Small Elevators)*	1.37	4.98
W(Large Elevators)*	0.48	1.53
Distance from Nearest Plant to 2nd Nearest Plant	-0.13	-10.26
Population Change	-0.36	-1.01
Population Density	-0.00	-0.20
W(Population Change)*	-3.03	-3.87
W(Population Density)*	-0.00	-0.41
County SQMI	-0.01	-8.07
Corn Yield	0.09	5.95
R^2	0.72	

It is worth noting that for an increase in distance from a county's nearest plant to the next-to-nearest plant, the ethanol influence ratio decreases significantly. Also interesting is that the more grain elevators – large or small – there are either in the county or in neighboring counties, the higher the influence. One reason for this is because ethanol plants tend to utilize elevators as contracted storage, particularly in the common case of both the ethanol plant and the elevator being owned by a local cooperative. Possibly a result of state-level biofuels policy, Iowa has the largest effect above the omitted groups captured by the intercept. The ethanol presence fitted

values are calculated by the matrix product $\widehat{\boldsymbol{E}} = X^{ethanol}\widehat{\boldsymbol{\beta}}$**.** Using the parameter estimates from Tables 5 and 6, I finally obtain $\hat{L}$ and $\hat{E}$ and proceed to the estimation of the propensity function.

4.3 Generalized Propensity Score Estimation

Having now obtained $\hat{L}$ and $\hat{E}$, two propensity functions are computed: one using a simple bivariate normal density, and the other using the EM Algorithm to estimate parameters governing finitely many mixture components of the estimated density function. To begin, I verify whether there is a concerning imbalance in the covariates listed in Tables 5 and 6 which would indicate associated bias in the estimated treatment effects. To do this, I follow Hirano & Imbens (2004) by segmenting my data into 4 subgroups: counties with low values of loan rate and low values of ethanol market influence, those with high loan rates and low ethanol, those with low loan rates and high ethanol, and those with high loan rates and high ethanol. Checking for balancing involves performing a t-test between each covariate within a group and the same covariate in the remaining groups aggregated together. Then, a weighted average of the t-statistics (each covariate will have 4 such t-statistics, or one t-test for each of the groups) is computed, weighted by the number of observations in that group. This is the t-statistic shown below in Figure 6.

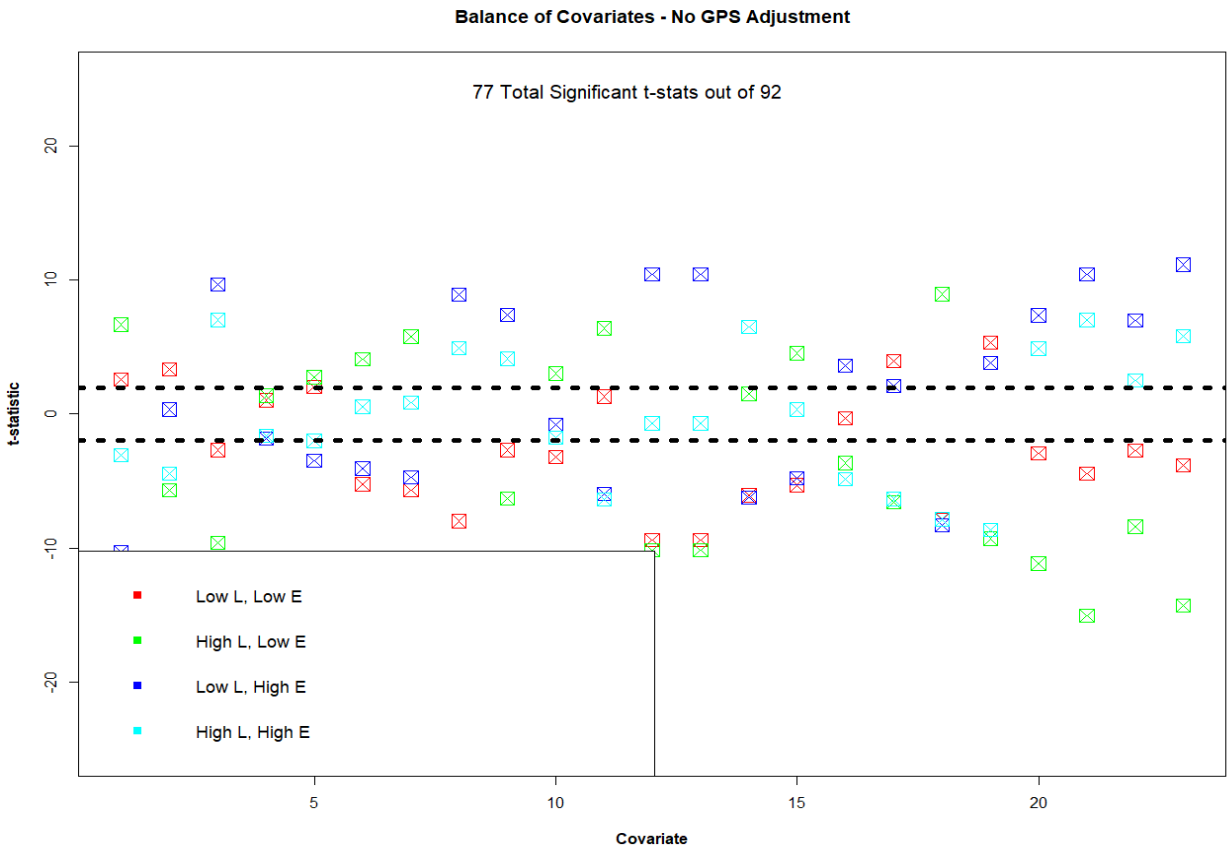


Figure 6: *Unadjusted Covariates Balancing*

There is evidence of a strong imbalance of covariates with 77/92 t-tests rejecting a null difference between groups. This makes land value changes from counties in the sample not directly comparable to one another (Rosenbaum, 1984), and balancing of the covariates likely improves if such a comparison is used to measure the effect. Hence, I compute a GPS to restore or at least improve the balancing of covariates and make counties more plausibly comparable to one another.

In the first case, the GPS is takes the form of the bivariate normal density function $\hat{R}(l, e; X) = \phi(L - \hat{L}, E - \hat{E}|X)$. To check the balancing of the covariates, I create two subgroups within each of the four subgroups mentioned above, that is, group 1 (low loan, low ethanol) gets split into group 1 counties with a below-median propensity score and group 1

counties with an above-median propensity score. Then a balancing check is performed across and within subgroups and a weighted average (weighted by the number observations in the subgroups of the initial subgroups) of the t-statistics in the balancing check, and these t-statistics are used to evaluate the GPS-adjusted balancing of the covariates. This adjusted balancing check takes the appearance in Figure 7.

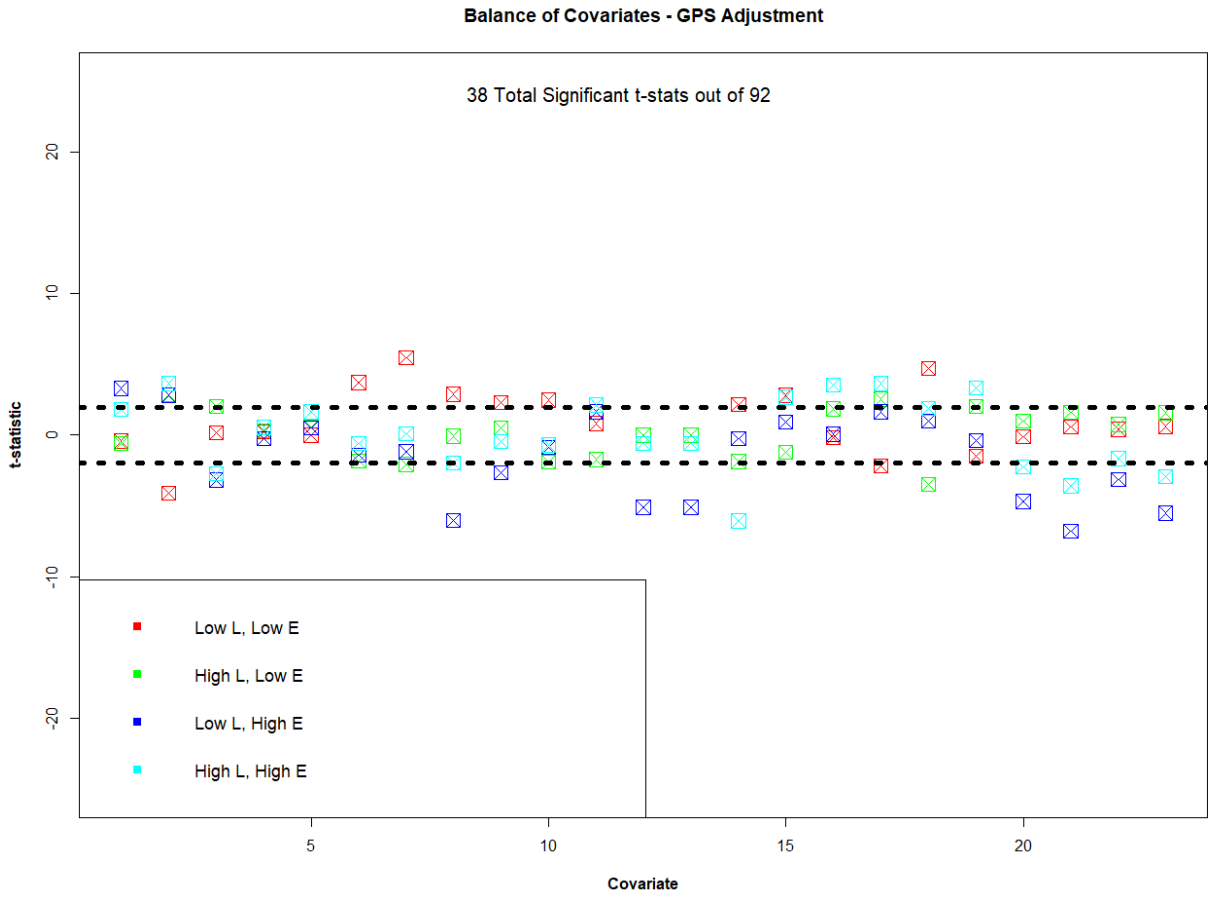


Figure 7: *GPS-adjusted Covariates Balancing (Bivariate Normal)*

Certainly, 38 is far below 87, but I argue that an improved balancing is feasible. Figure 8 maps the value of this propensity function (probability), which only vaguely reflects the geographically clustered ethanol markets appearing in Figure 3.

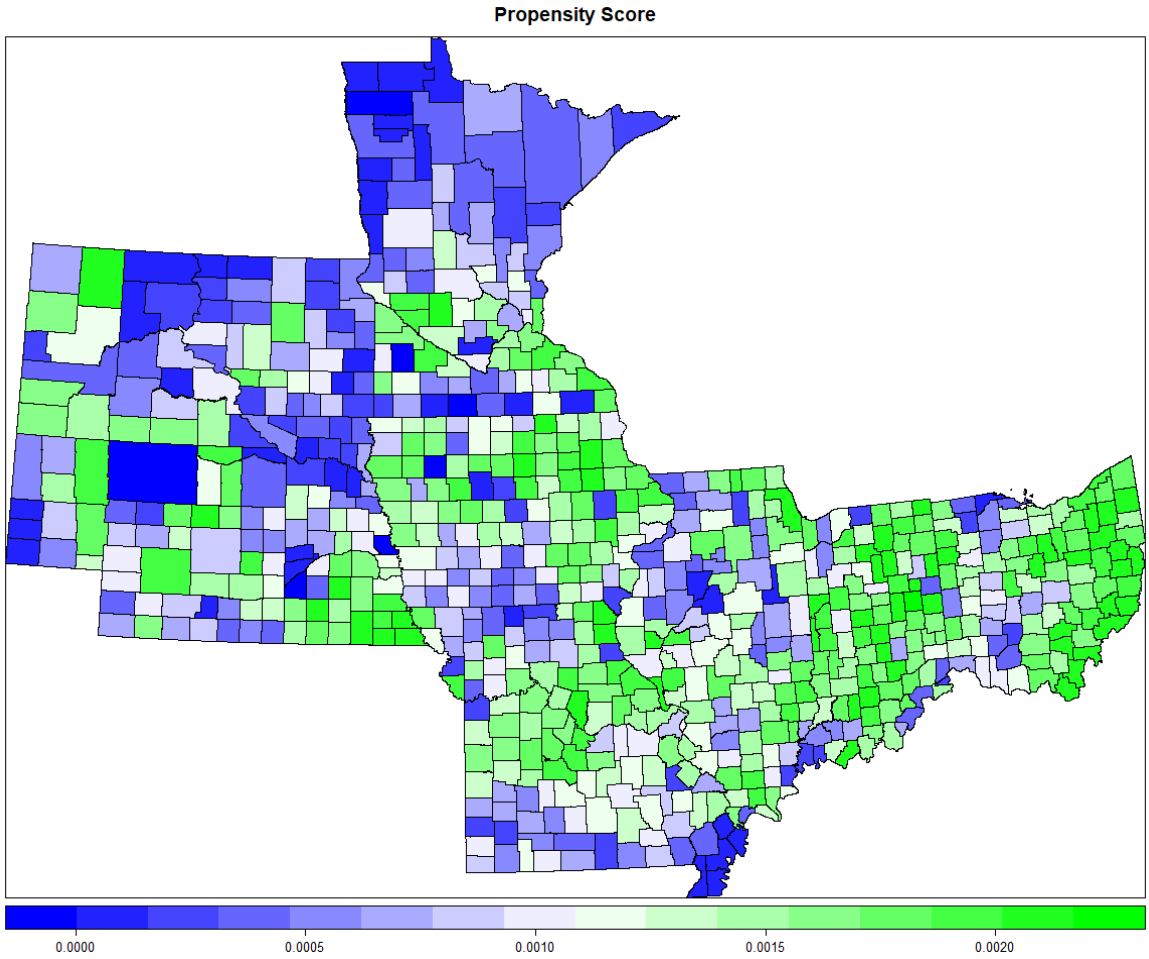


Figure 8: *Map of Bivariate Normal GPS*

Now, rather than forcing the fitted residuals of the loan rate and ethanol selection equations to conform into a bivariate normal density, I propose a new way to estimate a propensity function using the EM Algorithm which I believe improves the balancing of the covariates. I begin by estimating a normal finite mixture model (FMM) on the selection stage residuals from Equations 2 and 3, $\hat{R}(l, e; X) = \boldsymbol{\lambda} * \boldsymbol{\phi}(L - \hat{L}, E - \hat{E}|X)$, where $\boldsymbol{\lambda}$ is the final mixing proportions estimated by the algorithm and $\boldsymbol{\phi}$ is the matrix of estimated posterior probabilities for each of the components in the FMM. Appendix A-3 details the estimated parameters from the EM Algorithm, and Appendix A-4 shows the distribution of the propensity

functions. I argue that this is an improvement over the previous approach to computing the GPS for two reasons. First, it is more flexible than assuming the selection stage residuals are "close enough" to fitting into a bivariate normal density but not knowing for certain, because any FMM is an expectation maximizing approximation to the "true" distribution, which is unknown. Secondly, since the product of the posterior probabilities and mixing proportions is a linear combination, and the expectation is a linear operation, the proofs in Hirano & Imbens (2004) still hold in this case (see Appendix A-2 for a proof sketch for my corollary) and any bias associated with imbalanced covariates is at least reduced.

The primary concern is whether the covariates better balance with my potentially improved approach than with using the traditional approach of using a bivariate normal density function as the propensity score. As Figure 9 shows, this certainly seems to be the case: there are fewer rejections using an estimated propensity score than with the simple bivariate normal density. Therefore, I use the FMM approach proposed above.

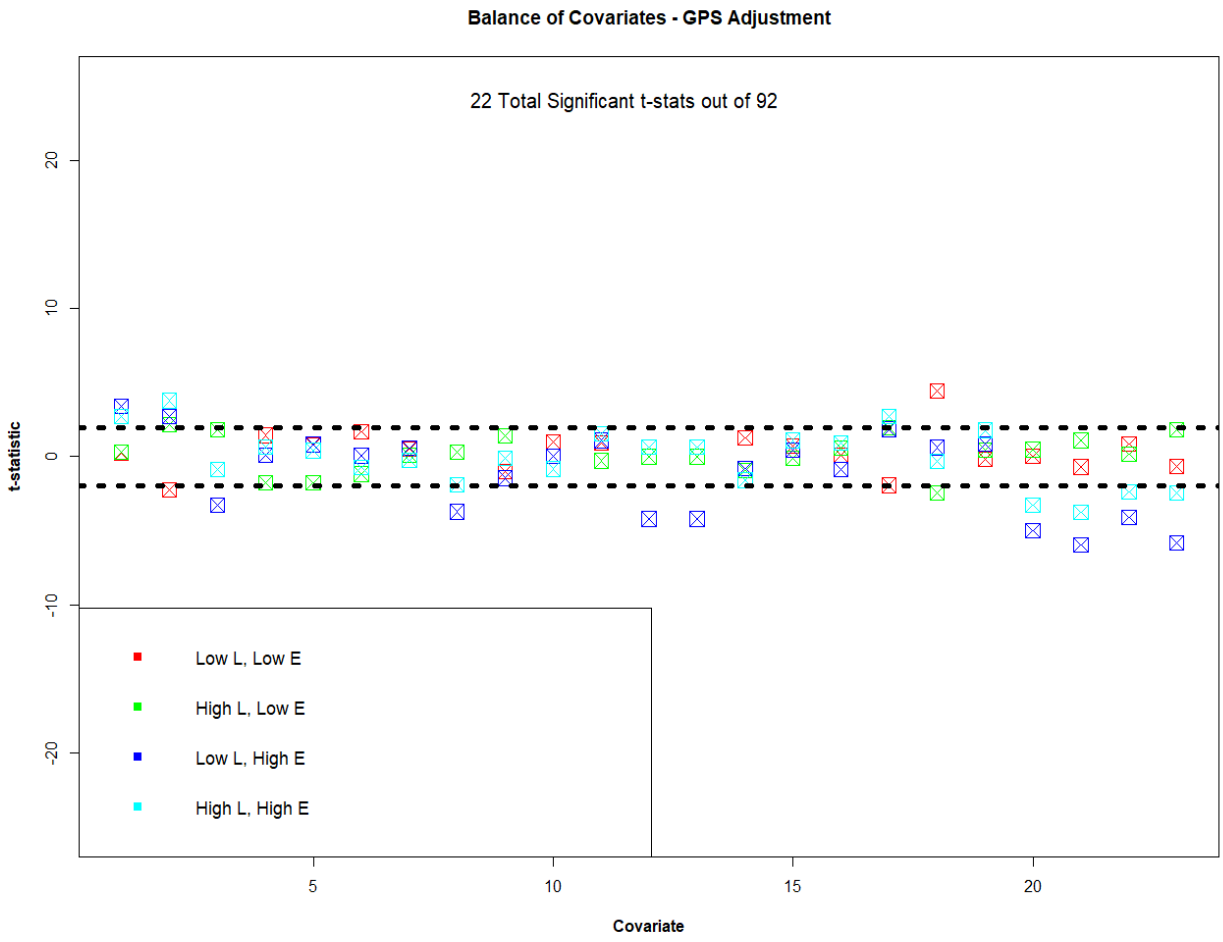


Figure 9: *GPS-adjusted Covariates Balancing (FMM Estimation)*

I choose this GPS moving forward to the analysis since 22 t-test rejections is preferred to 38. Moreover, the map of the score in Figure 10 better detects the regions most associated with ethanol markets (see Figure 3 and the figure in Appendix A-1) in the sample

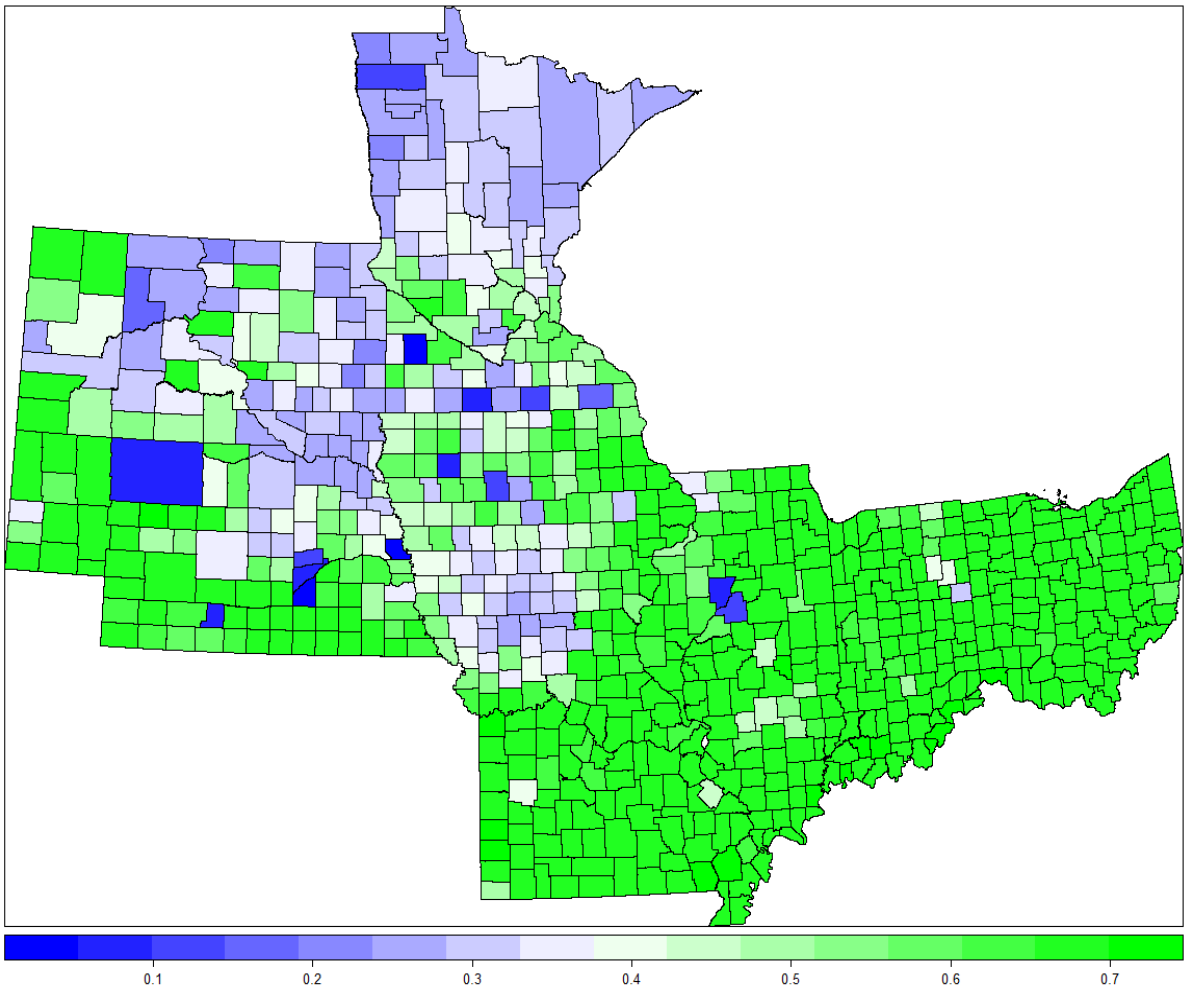


Figure 10: *Map of FMM-Estimated GPS*

4.4 Econometric Estimation (GPS)

The final stage of the GPS procedure begins with estimating Equation (4), a flexible parametric polynomial equation. Controlling for the GPS and its interactions with the treatments, I then estimate the expected conditional outcome at varying levels of ethanol market influence to test whether or not the picture in Figure 4 results.

I estimate Equation 4 using OLS and the results are below in Table 8

Table 8: *Outcome Equation Estimation Results (Equation 4)*

Variable	Estimate	t-statistic
Intercept	-377.30	-2.14
L	2.92	2.00
L^2	0.00	-1.26
E	11.00	2.51
E^2	-0.04	-2.69
$L * E$	-0.06	-2.57
$L^2 * E^2$	0.00	2.78
R	607.10	1.93
R^2	-125.40	-1.13
$R * L$	-3.86	-2.20
$R * E$	-12.23	-1.66
$R^2 * L^2$	0.01	1.76
$R^2 * E^2$	0.05	1.06
$R * L * E$	0.08	1.86
$R^2 * L^2 * E^2$	0.00	-1.45
$Adjusted\ R^2$	0.49	

Consistent with theory, the relationship between land value growth and loan rate is negative, although interpretation from the regression table is not straightforward and would be more easily seen in an illustration. The point estimate on the ethanol quadratic is negative, but the linear term has a substantially larger estimated coefficient, which is consistent with the conclusions in Hofstrand (2009) and Henderson & Gloy (2009). That producer surpluses from being located near an ethanol plant accrue to the landowner and capitalize in higher land values is familiar in economics, a standard result of site rents and dates back to the work of William Petty.

Now that I have the parameter estimates from Equation 4, a useful visualization of the sample and the dynamics between the outcome and the treatments is helpful before testing the hypothesis of this study. Figures 10-13 show the 3-dimensional dose-response surface of land value growth, with representative counties in the sample listed on the surface to illustrate what values of loan rate, ethanol market influence, and land value change look like geographically.

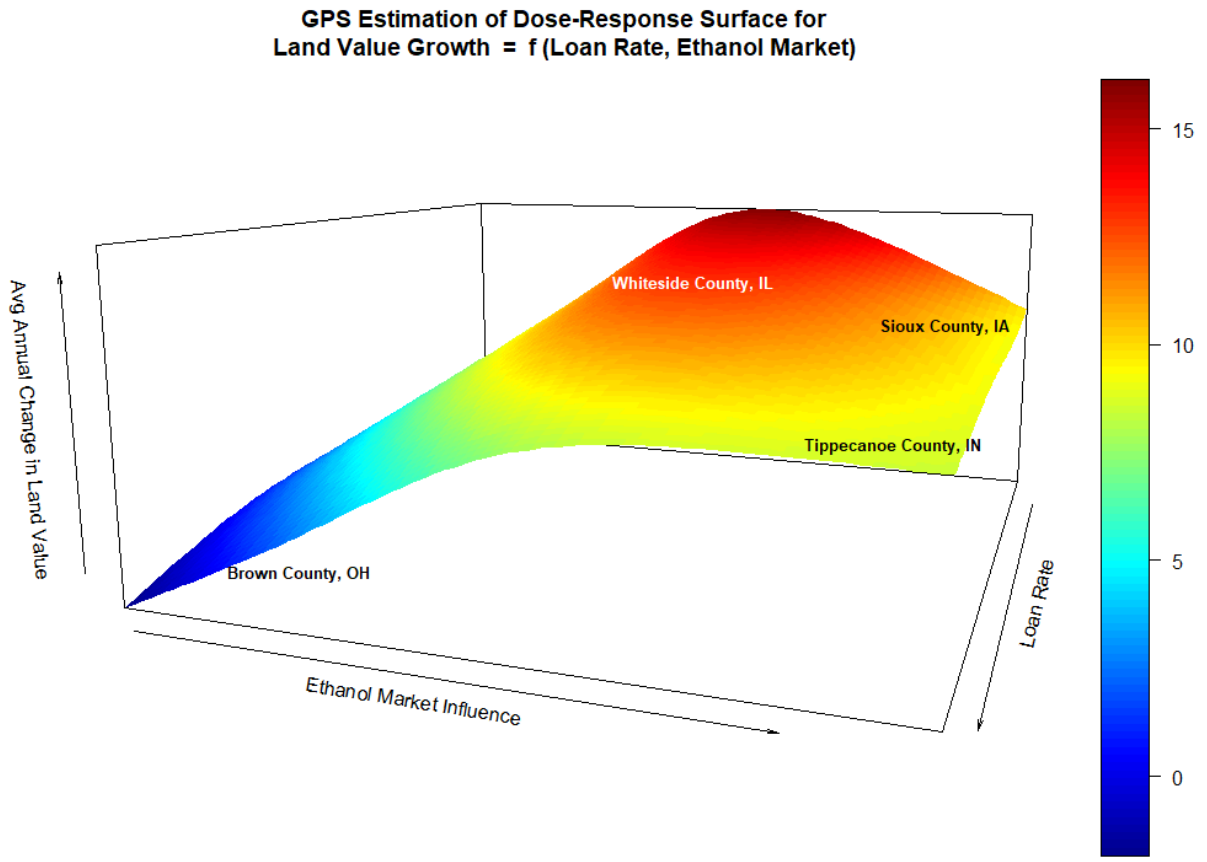


Figure 11: *3D Dose-Response Surface*

Figure 11 shows the response of land value growth to changes in ethanol market influence at the cross section of high values of county loan rate. As the presence of a local ethanol market grows from negligible (e.g. Brown County, Ohio) to highly influential (e.g. Tippecanoe County, Indiana), the estimated annual land value growth rate rapidly increases and tends to remain high or at least drop only very slightly at the end.

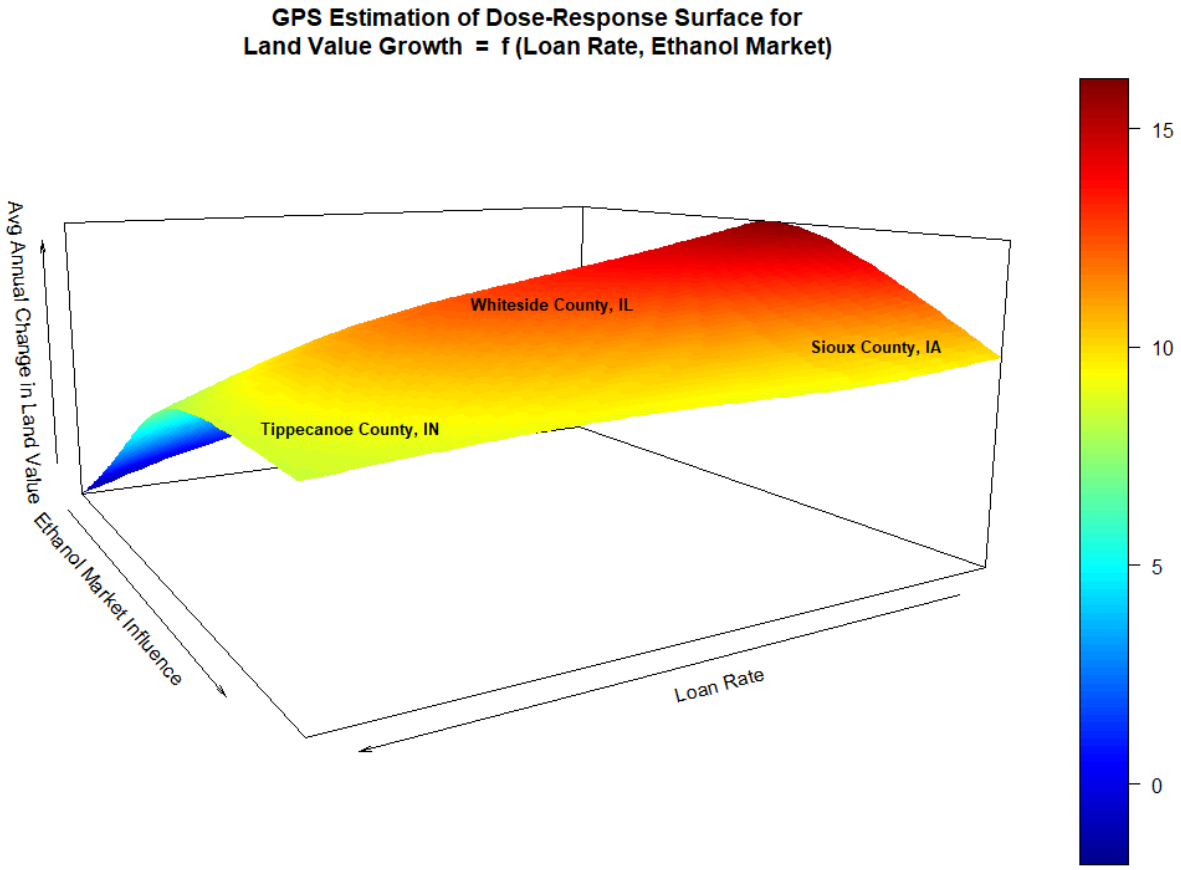


Figure 12: *3D Dose-Response Surface*

Figure 12 shows the response of land value growth to changes in county loan rate for the cross section of counties heavily influenced ethanol market. Under the presence of a strong local ethanol market, a county with a low loan rate (e.g. Sioux County, Iowa) has almost the same annual growth in land values as a county with a higher loan rate (e.g. Tippecanoe County, Indiana). Moving along the loan rate axis to the left, the slope of the cross section of the dose-response surface is only slightly negative, as presented by the dashed line in Figure 4 representing mitigated transport costs (heavy influence from local ethanol market).

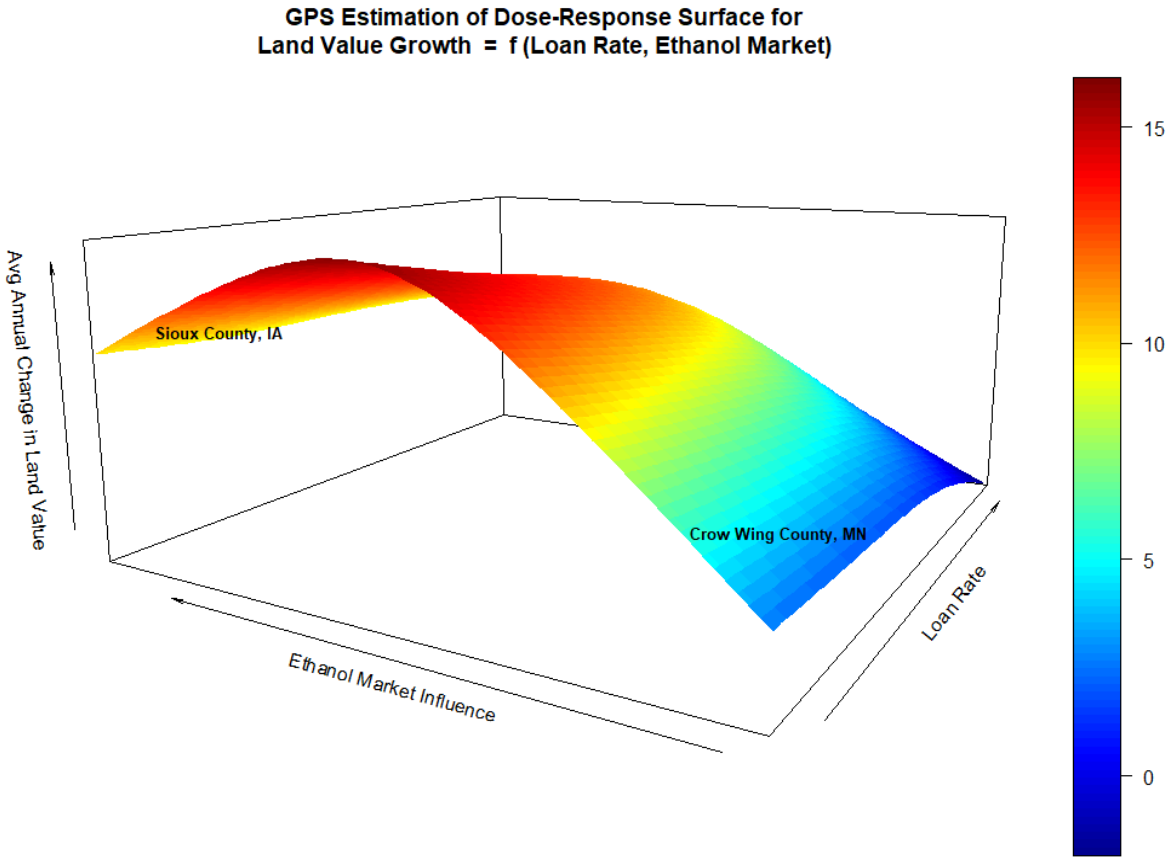


Figure 13: *3D Dose-Response Surface*

In Figure 13, the sample of counties with low values of loan rate, land values appreciate quickly, then level off and even begin to decline slightly. Moving from the environment similar to that of Crow Wing County, Minnesota to a more ethanol-saturated county like Sioux County, Iowa, land values grow quickly but are sluggish to come back down. This pattern is much like that illustrated in Figure 11, moving from a county with limited exposure to a local ethanol market, like Brown County in Ohio, to one more influenced by ethanol, like Tippecanoe County in Indiana. The key difference between these two figures is that the former represents the dose-response of land value growth to ethanol when loan rates are high, and the latter low loan rates.

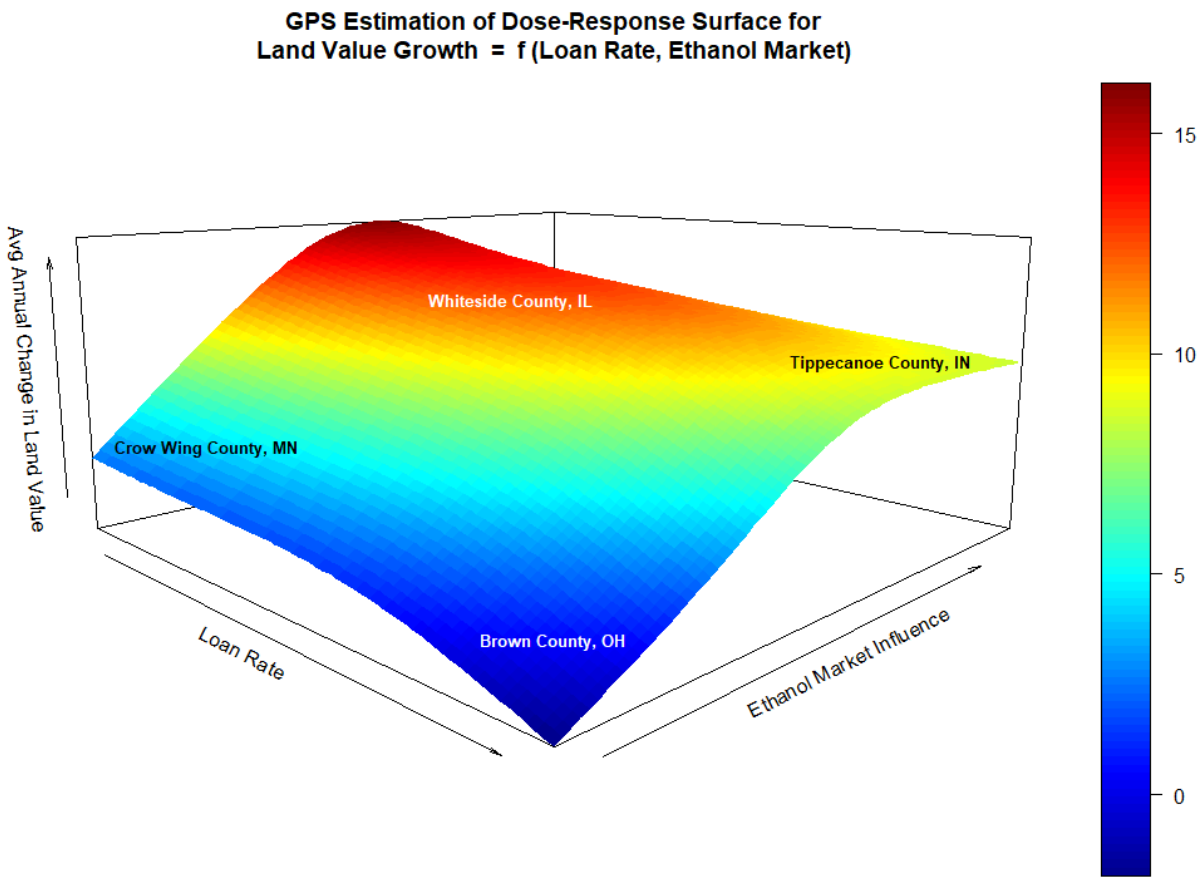


Figure 14: *3D Dose-Response Surface*

Finally, in Figure 14, it is clear that in counties virtually unaffected by ethanol have a dose-response to loan rates much like the original "Benirschka Effect" or the solid line in Figure 4 showing the response under unchanged transport costs. Average land values in Crow Wing County, Minnesota grow more rapidly each year than those in Brown County, Ohio. Moving along the loan rate axis from left to right, the speed of annual land value change declines towards zero.

Figures 11-14 show that, as expected, averaged land values in counties under less influence from a local ethanol market stabilize much more rapidly as the loan rate increases than counties with higher values of ethanol market presence, as was the case in the pre-ethanol world

examined in Benirschka & Binkley (1994). Land values in a county shift upwards and generally remain at their new higher levels as the influence from ethanol grows, as implied from Henderson & Gloy (2009) regarding land parcel distance from an operational ethanol plant and from Hofstrand (2015) regarding a discrete plant/no plant change in ethanol.

To test my hypothesis, I will compute the causal effect of the impact of a local ethanol market on the "Benirschka Effect" by comparing the conditional expected outcome at two different values of the ethanol variable. To visualize, recall Figure 4: the negative, significant relationship between annual land value change and loan rate is less negative in an ethanol-influenced counties than in those under a reduced exposure to the ethanol market. Below in Figure 15 are two level curves or cross-sections of the 3D surface

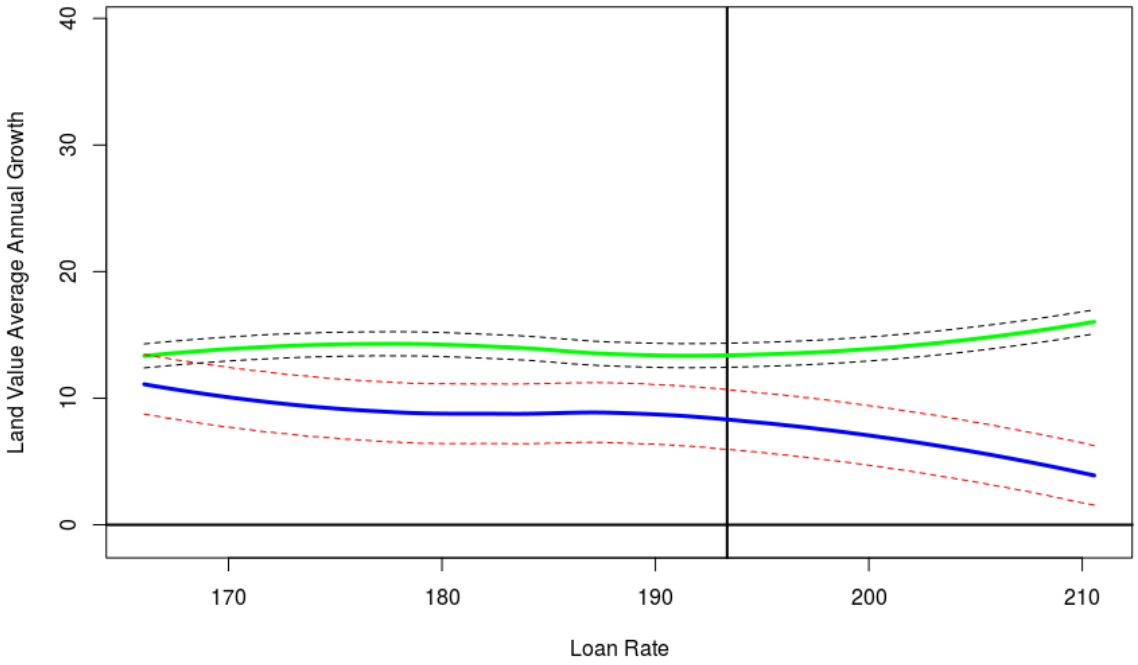


Figure 15: *Cross-Sections of 3D Dose-Response Surface*

The solid blue curve represents the response[7] of annual land value change to changes in the loan rate for a fixed low value of the ethanol variable (1st quartile), and the solid green curve is the response of land value growth to the loan rate for a fixed large value of the ethanol variable (3rd quartile). The dashed red lines are the 95% confidence interval around the blue (low ethanol) curve, and the dashed black lines are the 95% confidence interval around the green (high ethanol) curve. Referring back to Figure 14, the leftmost point of the blue curve has a similar combination of treatments and outcome as Crow Wing County, Minnesota, and the rightmost point of this curve has a similar combination of treatments and outcome as Brown County, Ohio. The relationship between land value growth and loan rate is negative in the absence of a strong ethanol market. Referring back to Figure 13, the leftmost point of the green curve has a similar combination of treatments and outcome as Sioux County, Iowa, and the rightmost point of this curve has a similar combination of treatments and outcome as Tippecanoe County, Indiana. The relationship between land value growth and loan rate is flat in the presence of a strong ethanol market.

Because loan rates represent distance, a county with high loan rate has low transport costs. The blue curve represents the response of annual land value change in a county to changes in loan rate in the "pre-ethanol" world analyzed by Benirschka & Binkley (1994) and the green curve represents the same entity in the "post-ethanol" world. The green curve appears to be flatter than the blue curve, much like the stylized graph in Figure 1. This is consistent with the hypothesis that corn transport costs were affected by ethanol in such a way that market distance is less important in determining land value changes after the Ethanol Boom than it was before.

A critical aspect in Figure 15 is the vertical black line, which represents the maximum level of loan rate in the state of Iowa, $193/bushel, or the "best case" scenario faced by the

[7] The plot is a locally linear nonparametric smoothing of the predicted values of the response variable.

locationally disadvantaged counties in the sample ($n = 323$ counties). The vertical change in the blue curve from the starting point to the reference loan rate value is -1.23, while the same for the green curve is -0.02. The most logical way to test the hypothesis is to compare the slopes of the green & blue response curves in Figure 14 for observations to the left of this line. This will show whether ethanol influence reduces transport costs enough for the "Benirschka Effect" to reflect the reduction in counties with loan rates below the "best case scenario" loan rate. Put differently, this tests for a difference between the conditional expected outcomes' responsiveness to the regressor of interest given two opposing values of ethanol market influence. The slopes are not constant in the estimated outcome, but the average slopes can be estimated and compared to compute the change in average causal effects of loan rates arising from changes in ethanol. Visually, the effect does seem to reduce, and the reduction in predicted land value change from 1.23 to 0.02 indicates that ethanol market influence does stabilize land values, but a statistical test is required to state a conclusion. Below in Tables 9-11 are the test results

Table 9: *Average slope of "blue" outcome function (Low Loan Rate, Weak Ethanol Market)*

Variable	**Estimate**	**p-value**
Intercept	17.19	10.56
L	-0.05	-5.05
$Adjusted\ R^2$	0.07	

Table 10: *Average slope of "green" outcome function (Low Loan Rate, Strong Ethanol Market)*

Variable	**Estimate**	**p-value**
Intercept	14.31	10.00
L	-0.00	-0.11
$Adjusted\ R^2$	-0.00	

Table 11: *Slope of "blue" vs. "green" outcome functions t-test*

Test	**t-statistic**	**p-value**

$\beta_{blue} - \beta_{green}$	-6.58	0.00

The t-test in Table 11 appears to offer strong evidence in favor of the hypothesis: the relationship between land value growth and loan rates or the "Benirschka Effect" reduces significantly in light of increased ethanol market influence. In other words, the average slope of the blue response curve is significantly more negative than that of the green response curve for counties in the western Corn Belt with lower average loan rates. Land values stabilized in the more remote counties of the Corn Belt as a result of influence from ethanol. The standard errors used for the t-test in Table 11 and for the t-statistics in Tables 9 and 10 are residual-bootstrapped with 10,000 iterations.

CHAPTER 5. CONCLUSIONS

Counties far from markets have higher transport costs, which make land prices more volatile, shown in Equation 1. After the Ethanol Boom, ethanol plants in effect moved these counties closer to markets, thus reducing the effect. The goal of this study was to investigate whether this reduction exists and examine its significance.

The study employs two methods: one under the assumption of identified parameter estimates (OLS) and one attempting to correct for a possible lack of identification (GPS). In using both methods and comparing the outcomes, this study effectively establishes an upper and a lower bound on the change in the "Benirschka Effect", although there was a significant change regardless of what one assumes about the data generating process.

The models estimate the effect of the Ethanol Boom on the response of county land value growth to changes in county loan rates (a proxy for transport costs), and computed treatment effects for two subsets of counties: once for counties uninfluenced by a local ethanol market (representative of the "pre-ethanol" world) and once for counties heavily influenced by a local ethanol market (representative of a "post-ethanol" world). As conjectured, land values in remote counties more exposed to a local ethanol market were more stable than land values in remote counties less exposed to a local ethanol market. The conclusion of this investigation is that ethanol markets reduce transport costs faced by corn farmers so that in more remote regions, average county land value growth patterns responded accordingly and were more stable.

There was also a difference in the "Benirschka Effect", albeit insignificant, following a series of events including the Staggers Act of 1980 and the Inland Waterway Trust Act of 1978. The exact magnitude of change resulting from any of these events and policies cannot be identified due to severe data limitations, but an overall effect of directly comparing the slopes on

loan rate for before and after these events has the expected sign. The conclusion of this investigation would be that the infrastructural development and its related policies, at the very least, do not appear to exacerbate the volatility of land values in remotely located counties in the Corn Belt.

The implications of this study are that events thought to mitigate transport costs faced by corn farmers not only improve the local cash basis for corn but also stabilize local land values. This is invaluable in that farmers, ag financiers, researchers, and policy writers know what to watch for when land values are growing (1990s to 2013) and can speak to the severity of an impending downturn (2013 to present in 2018). During the last period of land value growth, multiple such events occurred. The implication for the current downturn is, therefore, that the severity should not be as high as that of the 1980s, and we should observe a lower incidence of farming operations declaring bankruptcy. While having new and powerful insight into the severity of this present downturn in the agricultural economy is valuable, the findings herein cannot speak to the duration of this downturn since it is beyond the scope of this study.

While I am confident in the robustness of my findings, this study is not without its limitations, which are surely addressable with future research. County level data allows primarily for county level conclusions. However, given that measuring the distance to each of the thousands of existing grain markets in North America from each of the millions of land parcels is not feasible or even worth the amount of time, I know of no such data existing, which leaves county loan rates as the next best alternative to measuring market distance. Hence, answering this question below the county level is much too arduous a task to gain precious little precision in the estimates – which may or may not result. Furthermore, county level implications are still relevant in speaking to policies at the county level (e.g. loan rates and CRP rental rates) and can

at least identify patterns at a finer geography than region, state, or district. Another limitation is the lack of availability of the most recent USDA Ag Census data, which would capture the current downturn in land values in the middle of 2012 (the last Ag Census) to 2017. Future research evaluating the relationship between county loan rates and land value growth could not only speak to whether or not the mitigating of the "Benirschka Effect" applies equally to declining land values, but also serves as a robustness check to the present analysis.

With the above findings and implications, this study provides strong and relevant insight into the relationship between land value growth and geographic location, particularly concerning factors that affect this relationship. However long the current decline in land values in the US persists, we can be confident that the severity of financial hardships faced by grain farmers in more remote parts of the US is weakened by events such as the Ethanol Boom. Consequently, as land values continue to find the proverbial floor, there should be fewer instances of bankruptcy as in the previous crisis of the 1980s, which should generally be regarded as a positive, albeit unintended outcome of such an event.

REFERENCES

Alonso, William. 1964. “Location and land use. Toward a general theory of land rent.”

Archer, J. Clark, and Richard E. Lonsdale. 1997. "Geographical aspects of US farmland values and changes during the 1978–1992 period." *Journal of Rural Studies* 13 (4): 399-413.

Benirschka, M., and James K. Binkley. 1994. “Land price volatility in a geographically dispersed market.” *American Journal of Agricultural Economics* 76(2): 185-195.

Bultena, Gordon, Paul Lasley, and Jack Geller. 1986. "The farm crisis: Patterns and impacts of financial distress among Iowa farm families." *Rural Sociology* 51(4): 436.

Caves, Douglas W., Laurits R. Christensen, and Joseph A. Swanson. 2010. “The Staggers act, 30 years later.” *Regulation* 33: 28.

Egger, Peter H., and Maximilian Von Ehrlich. 2013. “Generalized propensity scores for multiple continuous treatment variables.” *Economics Letters* 119(1): 32-34.

Fuller, Stephen, Larry Makus, and Merritt Taylor. 1983. “Effect of railroad deregulation on export-grain rates.” *North Central Journal of Agricultural Economics*: 51-63.

Goodwin, Barry K., and Ashok K. Mishra. 2006. “Are ’decoupled’ farm program payments really decoupled? An empirical evaluation.” *American Journal of Agricultural Economics* 88(1): 73-89.

Henderson, Jason, and Brent A. Gloy. 2009. “The impact of ethanol plants on cropland values in the great plains.” *Agricultural Finance Review* 69(1): 36-48.

Hirano, Keisuke, and Guido W. Imbens. 2004. “The propensity score with continuous treatments." *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164: 73-84.

Hofstrand, Don. 2009. "Corn and Soybean Price Basis." Ag Decision Maker Home Page. *Iowa State University*. http://www.extension.iastate.edu/AGDM/crops/html/a2-40.html

Hofstrand, Don. 2015. "Who profits from the corn ethanol boom?" *Ag Decision Maker Newsletter* 12 (11): 1.

Imai, Kosuke, and David A. Van Dyk. 2004. “Causal inference with general treatment regimes: Generalizing the propensity score." *Journal of the American Statistical Association* 99(467): 854-866.

Koo, Won W., Denver D. Tolliver, and John D. Bitzan. 1993 “Railroad pricing in captive markets: an empirical study of North Dakota grain rates.” *Logistics and Transportation Review* 29(2): 123.

Lambert, Dayton M., Michael Wilcox, Alicia English, and Lance Stewart. 2008. “Ethanol plant location determinants and county comparative advantage.” *Journal of Agricultural and Applied Economics* 40(1): 117-135.

MacDonald, James M. 1989. “Railroad deregulation, innovation, and competition: Effects of the Staggers Act on grain transportation.” *Journal of Law and Economics*: 63-95.

Motamed, Mesbah, Lihong McPhail, and Ryan Williams. 2016. “Corn Area Response to Local Ethanol Markets in the United States: A Grid Cell Level Analysis.” *American Journal of Agricultural Economics*: aav095.

Requena-Silvente, Francisco, Guadalupe Serrano, and Joan Martin-Montaner. 2014. “Industry employment and import competition: a generalised propensity score approach." http://www.etsg.org/ETSG2014/Papers/261.pdf

O’Neill, Stephen, Noémi Kreif, Richard Grieve, Matthew Sutton, and Jasjeet S. Sekhon. 2016. “Estimating causal effects: considering three alternatives to difference-in-differences estimation.” *Health Services and Outcomes Research Methodology* 16 (1-2): 1-21.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1): 41-55.

Rosenbaum, Paul R. 1984. “From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment.” *Journal of the American Statistical Association* 79(385): 41-48.

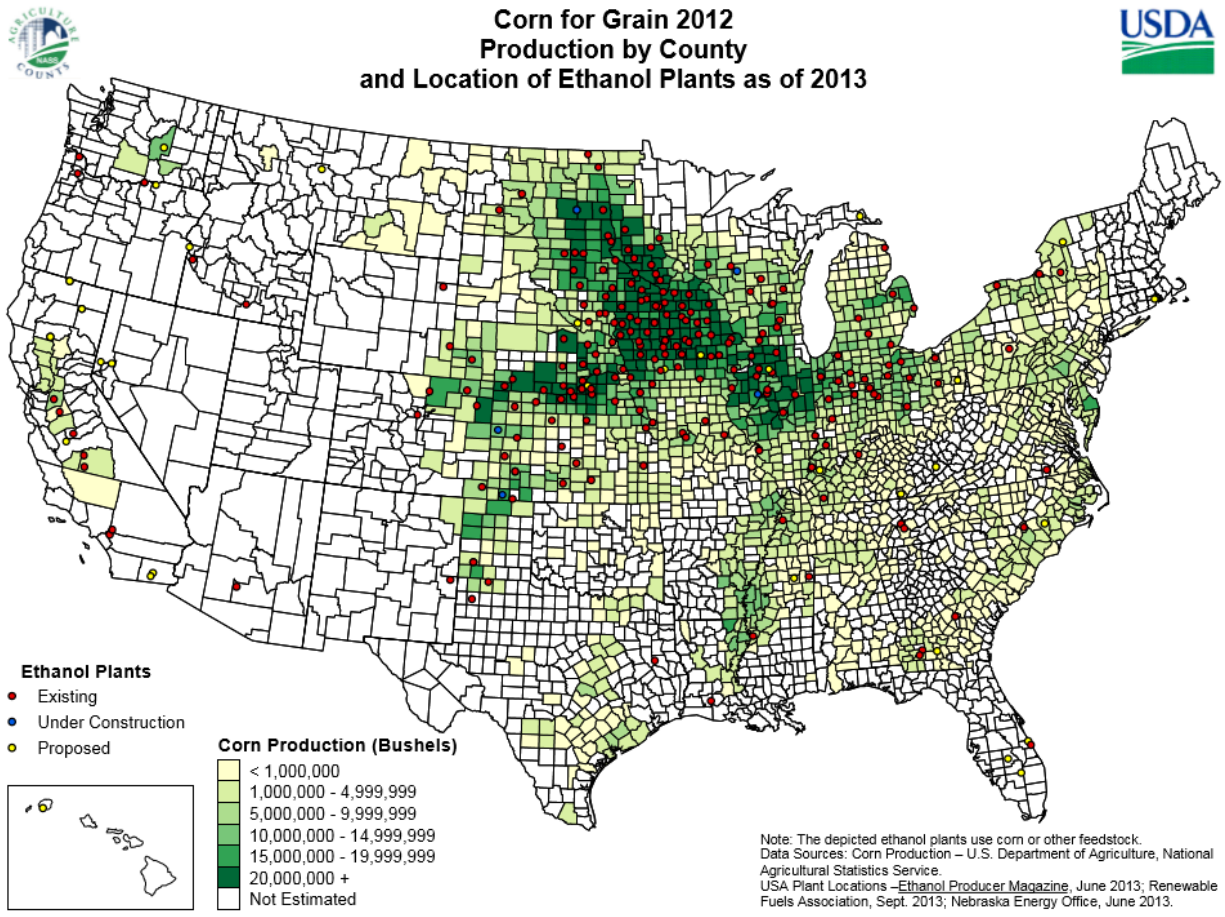
Sarmiento, Camilo, William W. Wilson, and Bruce Dahl. 2012. “Spatial competition and ethanol plant location decisions." *Agribusiness* 28(3): 260-273.

Tigges, Leann M., and Molly Noble. 2012. “Getting to yes or bailing on no: the site selection process of ethanol plants in Wisconsin.” *Rural Sociology* 77(4): 547-568.

Westcott, Paul C., and J. Michael Price. 1999. “Impacts of the US Marketing Loan Program for Soybeans.” *Principal Contributors*: 15.

APPENDIX

A-1.) USDA Ethanol Market Report



A-2.) Proof Outline for Mixture Propensity Score

The definition of the propensity score $\phi(L-l, E-e|X_i)\Pr[L=l \,\&\, E=e|X_i]$ is the conditional probability of observing the combination of loan rate value l and ethanol value e in the i^{th} county in the data given that county's vector of selected covariates X_i. To stay in line with the proofs presented in Hirano & Imbens (2004) and Imai & van Dyk (2004), that bias associated

with imbalanced covariates is reduced or eliminated after controlling for the propensity score, what I construct must measure the same as this.

After estimating a finite mixture model (FMM) with k mixture components, I have estimates for the mean vector μ, the variance-covariance matrix V, the vector of final mixing proportions (how much data comprise each component) λ, and posterior probability matrix $\boldsymbol{\phi}$. By definition, the element of the posterior probability matrix, $\widehat{\phi}_{ik}$, is the posterior probability of observation i belonging to the k^{th} component. From here, I obtain the matrix product of $\boldsymbol{\phi_i \lambda}$ which is the probability of the i^{th} observation belonging to the data. Since the estimated residuals condition on observables, that is, $L - \hat{L}_i = L - X_i\beta = L - E[L_i|X_i]$ and similarly for E, the posterior probability of observing a pair of treatment values $\{l, e\}$ in component k or $\phi_k(L - l, E - e|X_i) = \Pr[L = l \,\&\, E = e,\ in\ component\ k|X_i]$ is also a conditional probability of observing these treatment values in the k^{th} component. Finally, to calculate the conditional probability of observing this combination of loan rate and ethanol in the sample given the covariates is the combination of the posterior probabilities, or $\boldsymbol{\phi\lambda} = \Pr[L = l \,\&\, E = e,$ $in\ component\ 1\ or\ component\ 2\ or\ \ldots component\ k|X_i] = \lambda_1 \Pr[L = l \,\&\, E =$ $e, in\ component\ 1|X_i] + \lambda_2 \Pr[L = l \,\&\, E = e, in\ component\ 2|X_i] + \cdots + \lambda_k \Pr[L = l \,\&\, E =$ $e, in\ component\ k|X_i] =$ $\lambda_1 \phi_1(L - l, E - e|X_i) + \lambda_2 \phi_2(L - l, E - e|X_i) + \cdots + \lambda_k \phi_k(L - l, E - e|X_i) = \Pr[L =$ $l \,\&\, E = e|X_i]$. This interpretation is equivalent to that of the propensity score following Hirano & Imbens (2004) and is suitable to use for my analysis. An attractive feature of estimating the propensity function in this way is that it is much more flexible than assuming a strict, rigid bivariate normal density.

A-3.) Mixture Propensity Score Parameter Estimates

PARAMETER	Loan Rate	Ethanol	
$\boldsymbol{\mu_1}$	-10.78	1.05	
$\boldsymbol{\mu_2}$	3.79	-0.91	
$\boldsymbol{\mu_3}$	-1.14	7.41	
$\boldsymbol{\lambda}$	0.24	0.71	0.05
Log Likelihood	-5212.29		

k=1	VAR[L]	COV[L,E]
COV[E,L]	27.21	6.87
VAR[E]	6.87	68.32
k=2	**VAR[L]**	**COV[L,E]**
COV[E,L]	57.26	-4.45
VAR[E]	-4.45	32.53
k=3	**VAR[L]**	**COV[L,E]**
COV[E,L]	73.99	11.19
VAR[E]	11.19	326.24

A-4.) Propensity Score Densities

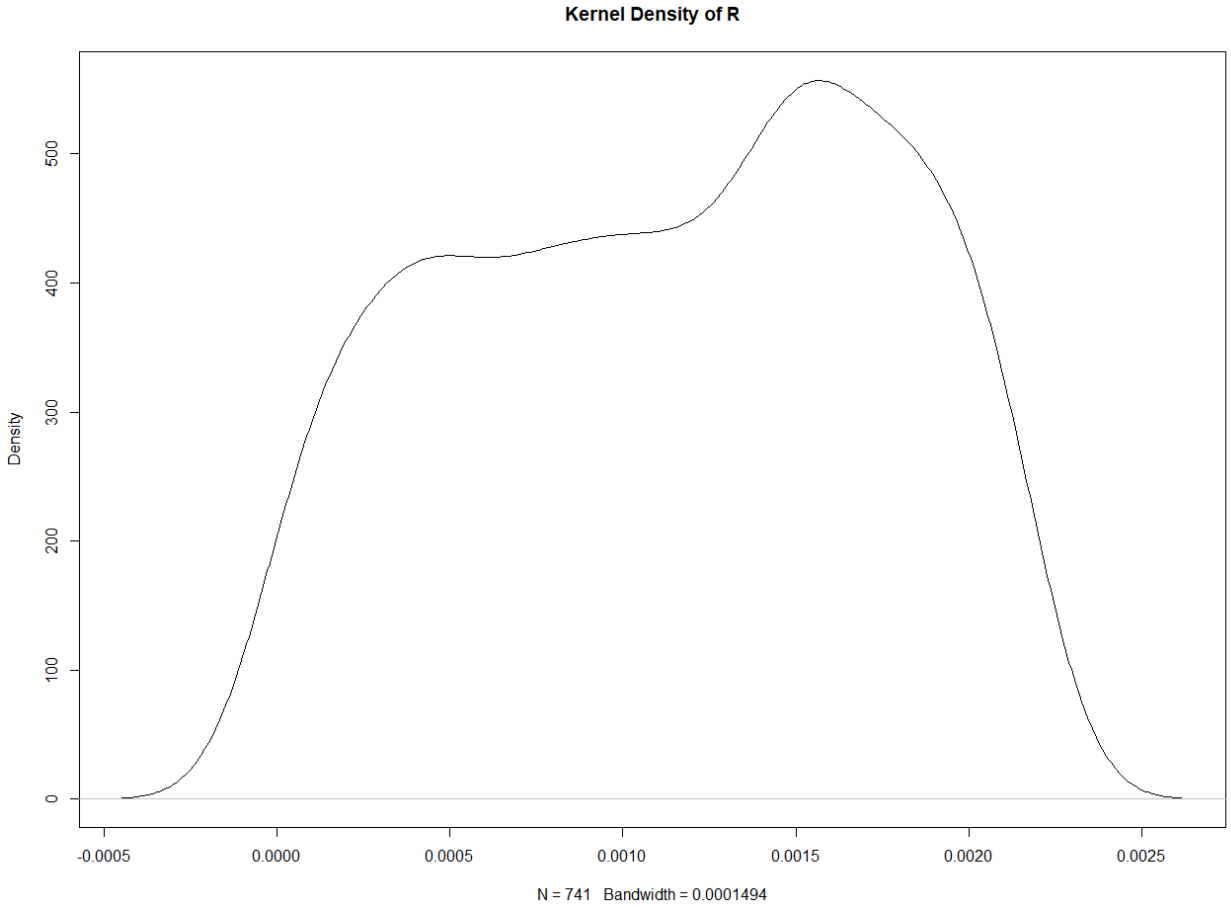


Figure A-4.i: *Bivariate Normal Density GPS*

The range of this score is from 0.000 to 0.002.

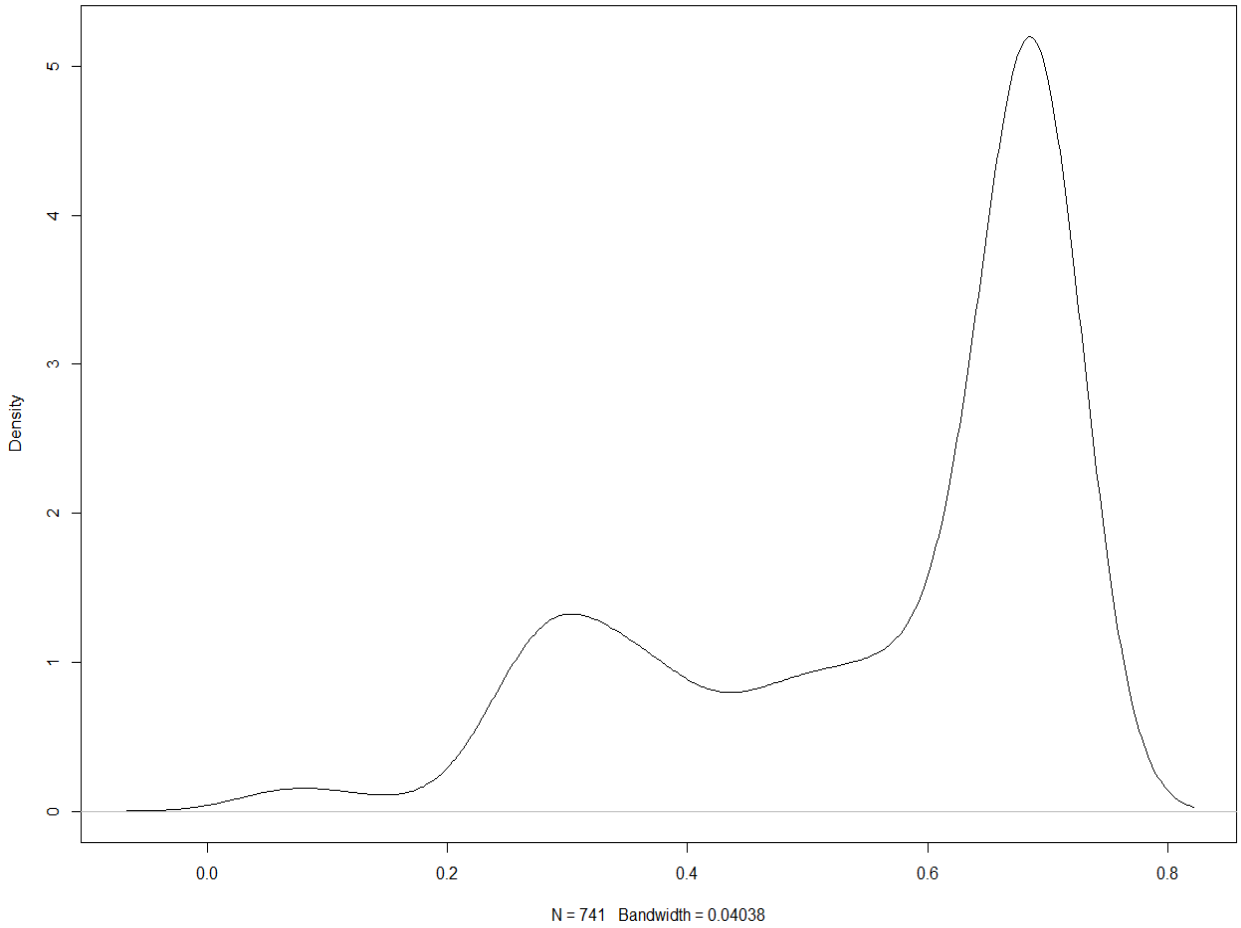


Figure A-4.ii: *FMM-Estimated GPS*

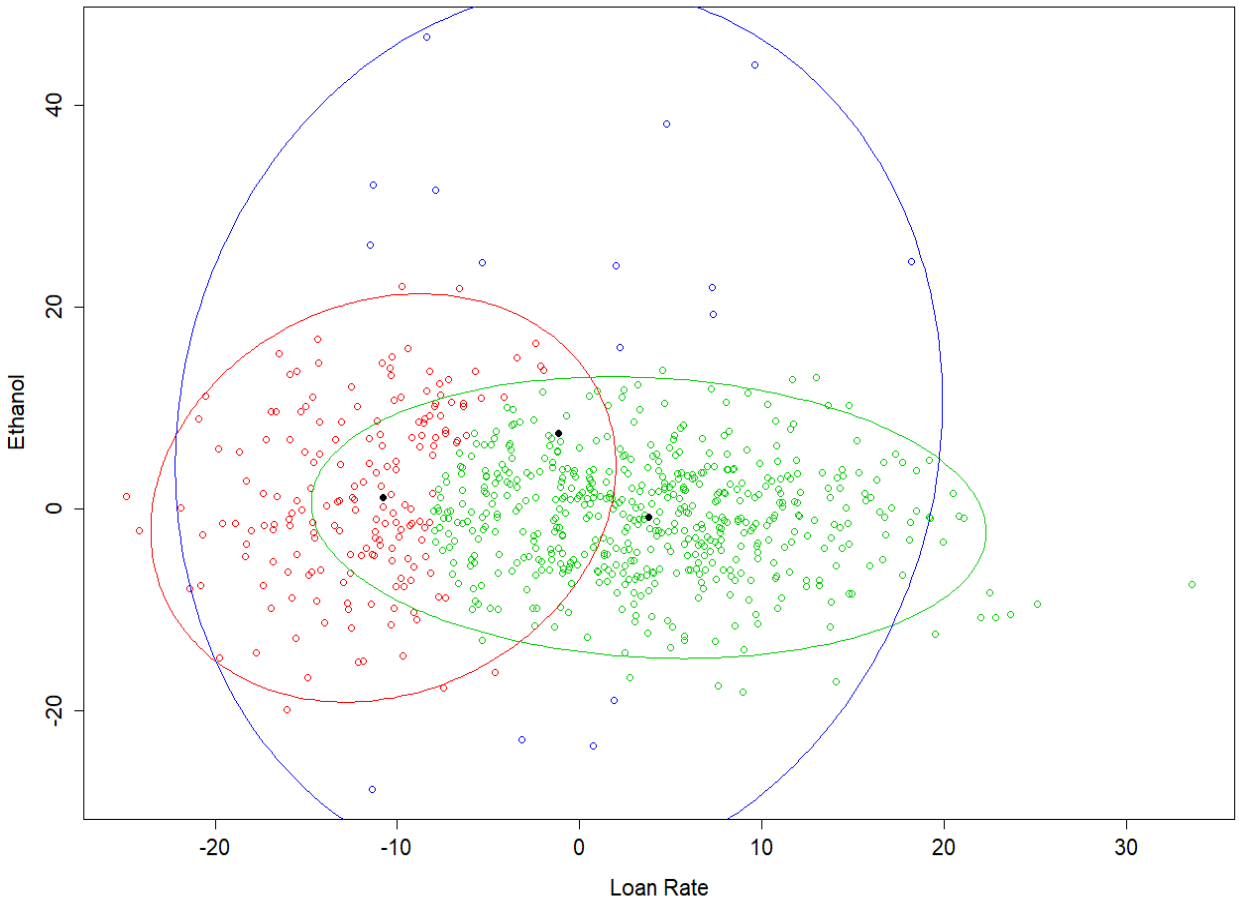


Figure A-4.iii: *FMM-Estimated GPS*

Figure A-4.ii shows the kernel density of the estimated function, which ranges from 0.05 to 0.70 and clearly shows the three-component structure implemented into the estimation, and Figure A-4.iii shows the scatterplot of the selection stage residuals and the estimated 3-component mixture model.

PART II. TRACKING SPATIAL HEALTH PATTERNS WITH GEOGRAPHIC VARIATION IN GROCERY PURCHASING

Some of the data in this work is calculated (or Derived) based on data from The Nielsen Company (US), LLC and marketing databases provided by the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

CHAPTER 6. INTRODUCTION

In this study, regional food marketing data is used to study geographic differences in food purchasing. These differences are then correlated with regional measures of health outcomes. This is largely exploratory in nature and does not lend itself to specific, testable economic hypotheses, although the findings may prove somewhat useful for research related to public health and food policy analysis.

Nearly all current knowledge on the diet-disease relationship in humans includes nutritional epidemiological studies. Most such studies are either cross-country comparisons using national measures of health, mortality, and diets, or longitudinal studies at the individual level, which track individuals over long periods, measuring diets and ultimate health outcomes. Examples of the first are several studies which have led to recommendations regarding the benefits of the "Mediterranean Diet", first identified by ecological studies including cross-county comparisons (Willett, 1995; Kim et al., 2003; Mueller et al., 2006). Examples of the second are many studies using data from the Women's Health Study, which tracks the diet and health behavior of thousands of women across the US for many years, and studies focusing on the Seventh Day Adventists, a group emphasizing vegetarianism (Fraser, 1999; Liu et al., 2000).

An example of the importance of the information arising from nutritional epidemiology is the 2015 World Health Organization recommendations to limit consumption of red meat and processed meat products due to potential carcinogenic effects arising from excessive consumption of bacon, sausage, and similar products. According to the WHO, "the recommendation was based on epidemiological studies suggesting that small increases in the risk of several types of cancers might be associated with high consumption of red meat or processed

meats." (2015) These foods were also associated with increased risk of heart disease, type-2 diabetes, and stroke.

A type of data that has not been used for nutritional epidemiological studies is regional food marketing data. This is surprising, for such data has become widely available, at least on a commercial basis. This data is collected by large tracking firms and sold to food manufacturers, which use it to monitor sales levels in different regions. A possible reason why it has not been recognized as a potential resource for nutritional epidemiology studies is a belief that diets across the US do not greatly vary. However, that is not the case. Purdue has access to an early version of marketing area data in the form of 1990 regional sales indices for 54 cities across the US. This was made available by Selling Area Marketing Inc. (SAMI), a bankrupt tracking firm whose data was used in Larson (1998). This data indicates that there is indeed considerable variation in consumption of specific foods across the US. Consider the case of processed meats: according to the SAMI data, 1990 household bacon consumption ranged from 60 percent of the national average in Cleveland, OH, to 109 percent in Indianapolis, IN, to 186 percent in New Orleans, LA. Similarly, breakfast sausage ranged from a low of 44 percent of the national average in Cleveland, OH, to 149 percent in Indianapolis, IN, to a high of 201 percent in New Orleans, LA. There is also considerable variation in foods as common as canned tuna, with consumption varying between 50 and 200 percent of the national average in Nashville, TN and New York City, respectively.

Furthermore, there is noticeable geographic variation in disease incidence and health outcomes. According to the 2010 State and Metropolitan Area Data Book, the average death rate from heart disease is 2.7 per 1000 persons in Mississippi, the median was Virginia at 1.9 per 1000, compared to Minnesota's 1.3 per 1000 persons, less than half as large as the first. For

stroke incidence, the lowest rate was in Illinois at 0.30 deaths per 1000 persons, Texas was at the median with 0.44 per 1000, and the highest was Utah at 0.59 deaths per 1000 persons. Finally, cancer incidence was highest in Kentucky at 2.1 per 1000, Rhode Island in the middle at 1.8 per 1000, and the lowest was in Utah at 1.4 per 1000. Finally, as reported by the CDC, state rates of adult obesity varied from 21.4 percent in Colorado to 35.5 percent in Mississippi. Certainly obesity is a food-related health problem.

There are three main questions I will attempt to answer in this study. The second is whether or not diet-disease relations currently viewed as holding with a high level of confidence are evident in food marketing data. Three initial cases in focus are red and processed meats, the group of foods involved in a 2016 World Health Organization (WHO) report, nuts and nut products, the heart-health benefits of which have been previously studied and suggested by researchers (e.g. Sabaté, 1999; Hu & Willett, 2002), and fruits and vegetables, which also have been associated with the prevention of obesity as well as various forms of cardiovascular disease (Liu et al., 2000; Hu & Willett, 2002; Bazzano et al., 2003; Woodside et al., 2013). The second question is whether diet-disease relations not previously investigated appear in regional food marketing data. The third and final question is to examine marketing areas with above average and below average health outcomes and determine whether and how the pattern of food sales in these market areas differ. This is particularly pertinent to obesity.

CHAPTER 7. METHODS

7.1 Data

The food purchasing data is from the AC Nielsen Consumer Homescan Panel data, obtained from the Kilts Center for Marketing Data at The University of Chicago Booth School of Business. The datasets cover demographics, geography, and UPC-coded purchases of more than 60,000 households in the US. The data include records from in-home scanners used by the households detailing the source of the purchases (e.g. supermarket chain and city) as well as prices, package sizes, brands, variety descriptions, and UPC codes of food items from specific departments including dry grocery, produce, dairy, deli, and packaged meats. Below the department level, the hierarchy of every product scanned by the household goes down to product group, a Nielsen-assigned numeric code describing the general grouping of products within a given department. The next level lower is product module, which is a Nielsen-assigned numeric code representing the detailed product categories within a group. The final tier is the individual UPC code for every food item, for which there are more than 3.1 million unique values in the data.

One aspect of the Nielsen data is the purchases of non-UPC coded random weight groceries. This is particularly pertinent to produce, much of which is sold in loose form. For most common produce, there are both separate UPC codes as well as aggregate random weight categories. Hence, it is not possible to determine shares of specific from aggregate. A good example is apples. These are sold both in loose form as well as pre-packaged bags. Conversely, bananas are sold almost entirely random weight. Of the 60,648 households recorded in the 2010 Homescan data, not all participated in the random weight portion of data reporting. The number

of households reporting random weight purchases is 27,422 or about 45% of the full sample, varying from 158 in Des Moines, IA to 837 in Los Angeles, CA.

The Nielsen data covers 52 market areas surrounding major metropolitan areas across the continental US, each market consisting of 28 counties on average. Below in Figure 16 is a map with Nielsen market names overlaid

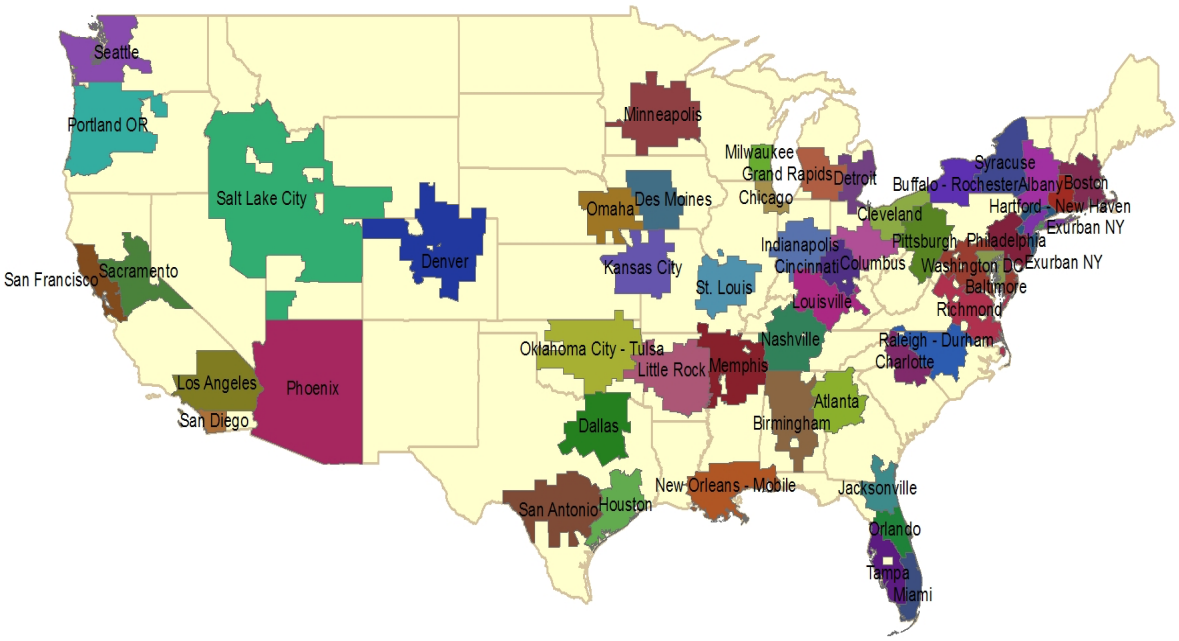


Figure 16: *Layout of the 52 Nielsen Scantrack Markets*

It is reasonable to expect food-purchasing patterns to vary from region to region because of many factors. Indeed, the Nielsen Homescan panel has been shown to exhibit regional patterns by Larson (2004). Since both food purchasing and health outcomes are regional, it is a goal of this study to test for and describe cases in which they similarly regional.

7.1.1 Food Purchasing.

To begin the analysis, I compute grocery purchasing expenditure shares within markets for different foods using the Nielsen data from 2010, selecting specific foods to build off of the

individual-level correlations found in previous work. The foods under investigation include beef and the combined category of bacon and breakfast sausage. According to the WHO report, excessive consumption of these meats correlates to increased risk of colorectal and stomach cancers at an individual level. Additionally, I compute the purchasing measures for nuts, fruit, vegetables, and seafood, all of which are thought to be generally "healthy" and identified as having health benefits, including reduced risk of type-2 diabetes, heart disease, ischemic stroke, and colon cancer (Sabaté, 1999; Hu & Willett, 2002; Liu et al, 2004; Lund, 2013; Wu et al, 2015; Micha et al., 2017). Primarily with regards to diabetes and obesity, sweetened soft drink purchasing shares are examined.[8] While regional variation in purchasing was not the focus of their study, Wang et al. (2016) found evidence that areas with higher rates of obesity purchased more soft drinks regardless of price increases, taxes, or sales. Additional food groupings for which I calculate expenditure shares are salty snacks (e.g. chips, pretzels, and crackers), candies, and baked goods (e.g. pies, cookies, and cakes). Together with sweetened soft drinks, the latter two have been identified as major sources of sugar intake of consumers (Yang et al., 2014). Figure 17 shows the market-level expenditure shares for bacon and breakfast sausage as an example

[8] Calculations from the 2007 National Household and Nutrition Examination Survey (NHANES) indicate that nearly 75% of soft drink purchases are from grocery stores rather than from restaurants.

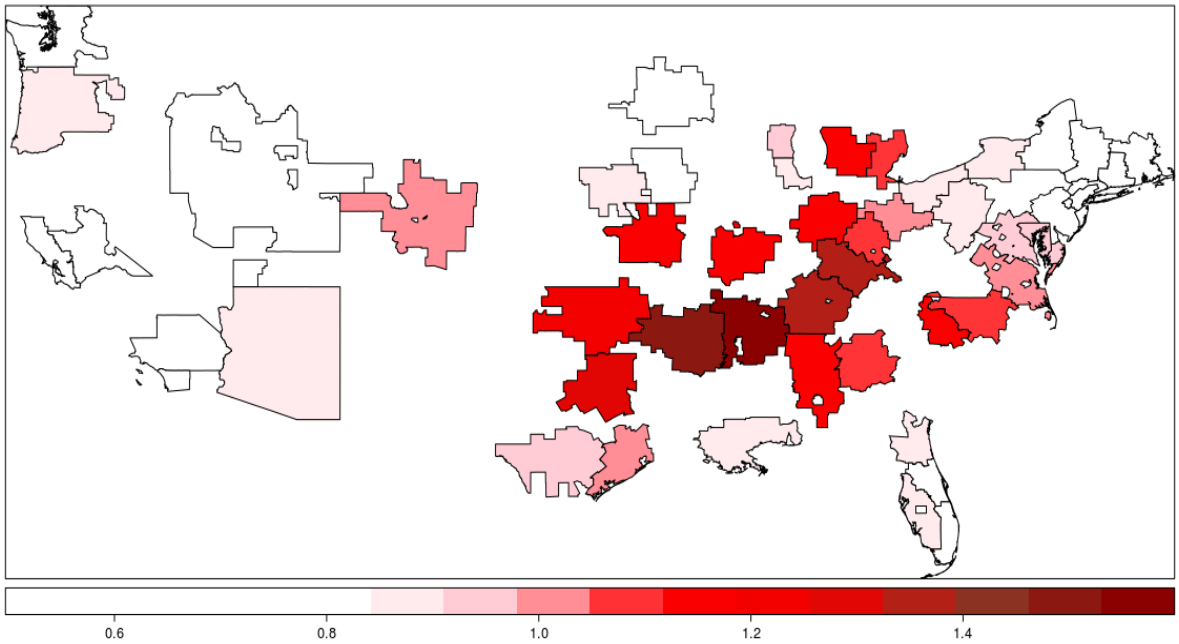


Figure 17: *2010 Grocery Expenditure Share of Bacon & Breakfast Sausage*

The expenditures of the random weight food modules were included with specific UPC-coded, non-random weight modules (e.g. fresh carrots, fresh strawberries) to compute total expenditure of fruit or vegetables. A similar procedure was followed to calculate expenditures for nuts[9], the expenditure shares of which are shown in Figure 18

[9] For 2010 Nielsen data, the category for random weight nuts is combined with random weight candy, while in some of the earlier years, they are categorized separately. Therefore, earlier homescan data from 2006 (which has only 37,786 households in total with 7,526 reporting random weight) is used to estimate the proportion of this combined category belonging to nuts in 2010, since the exact measure of random weight nut expenditure shares is unavailable for 2010.

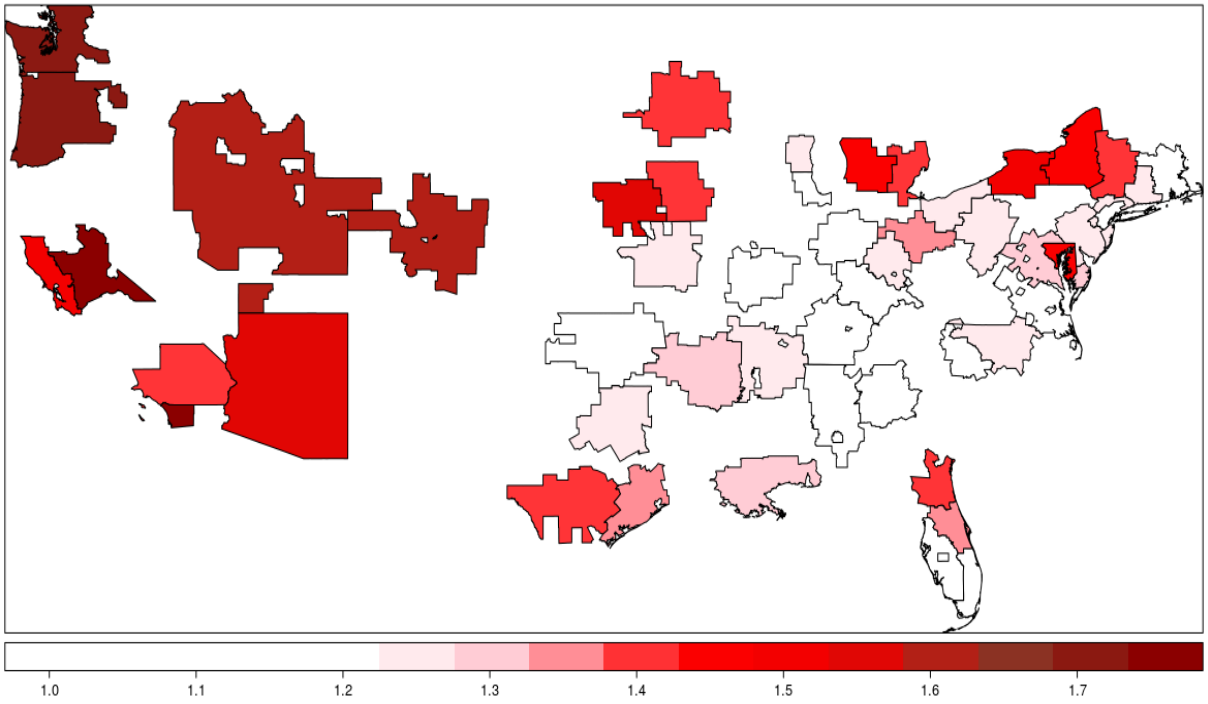


Figure 18: *2010 Grocery Expenditure Share of Nuts*

The summary statistics for all food expenditure shares are displayed in Table 12.

Table 12: *Summary Statistics of Food Expenditure Shares*

	Mean	**Std. Dev.**	**Min**	**Max**	**Median**
Nuts	1.40	0.19	1.10	1.83	1.34
Beef	4.64	0.34	3.89	5.34	4.65
Poultry	2.59	0.37	1.80	4.15	2.61
Sweetened Soft Drinks	*4.35*	*0.84*	*2.96*	*6.50*	*4.21*
Seafood/Fish	*1.90*	*0.51*	*1.09*	*3.30*	*1.89*
Salty Snacks	*6.16*	*0.48*	*5.04*	*7.06*	*6.18*
Baked Goods	*5.50*	*0.41*	*4.26*	*6.68*	*5.49*
Candies	*3.06*	*0.39*	*2.28*	*4.26*	*3.04*
Bacon & Sausage	*1.18*	*0.19*	*0.89*	*1.71*	*1.19*
Fruit	*4.07*	*0.32*	*3.37*	*4.73*	*4.04*
Vegetables	*7.01*	*0.54*	*5.68*	*8.10*	*6.96*

2.1.2 Health.

For health outcomes, there are publicly available data provided by the Center for Disease Prevention and Control (CDC) for select years at the county level. For this study's measures of stomach and colorectal cancer, average incidence calculations by county from 2011-2015 are chosen. For diabetes and obesity, the estimated county prevalence percentages from 2013 are used. For stroke and heart disease, the 2014-2016 average county death rates per 100,000 population for adults aged 35 and up are used. Each of the county level health measures is aggregated up to the Nielsen market level, each county weighted by its share of the total Nielsen households in that market. The map for stroke is shown in Figure 19

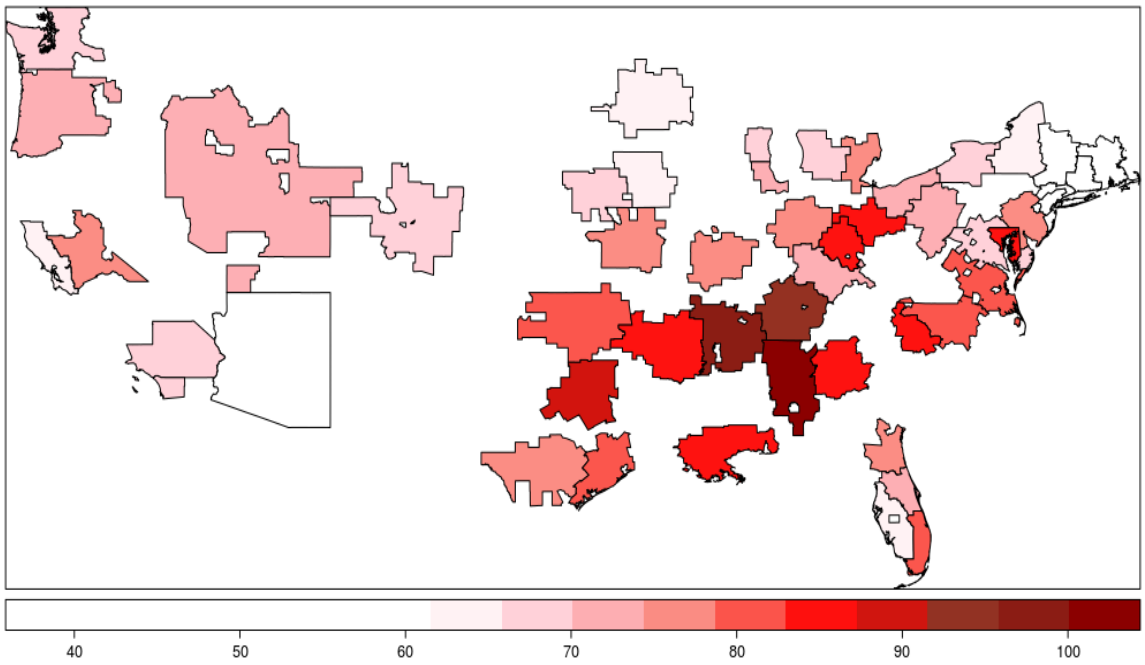


Figure 19: 2014-2016 *Average Stroke Death Rate per 100k population*

The geographic patterns of health outcomes are easily visible. The clustering of adverse health outcomes persists in the southeastern US for all health measures used in this study. Table 13 shows the summary statistics for all health outcome variables at the market level

Table 13: *Market-Level Health Measure Summary Statistics*

	Mean	**Std. Dev.**	**Min**	**Max**	**Median**
Obesity (Prevalence)	28.22	3.55	19.10	34.64	28.41
Diabetes (Prevalence)	10.06	1.48	6.29	13.49	9.79
Colon Cancer (Incidence)	39.90	2.74	33.50	47.90	39.90
Stomach Cancer (Incidence)	6.67	0.91	5.20	10.75	6.65
Stroke (Deaths/100k)	72.89	11.85	40.06	100.16	73.09
Heart Disease (Deaths/100k)	325.83	52.25	210.63	443.73	318.79

7.1.2 Demographics, Behavior, and Environment.

Additional variables that are included in the analysis are the percent adults that are married from the Nielsen Demographics Data (Nielsen); average household size (Nielsen); percent of population graduated from college (Nielsen); average annual household income

(Nielsen); the percent of the population under 18 years old from the US Census Bureau (USCB); the percent of adults without health insurance from the Behavioral Risk Factor Surveillance System (BRFSS); the percent of the adult population who smoke cigarettes daily (BRFSS); the county proportion of male and female adults (USCB); the average age of male and female adults (USCB); the average county proportion that is white, black, Hispanic, and of other minorities (Nielsen); the average annual population growth from 2000 to 2010 to help capture the some of the effect of migration on the longevity of regional food habits (USCB); the proportion of heads of households that are married (Nielsen); and the median Air Quality Index from the Environmental Protection Agency (EPA). Another potentially important[10] variable is some measure of spending on food away from home (FAFH) at fast food and table service restaurants. The USDA Food Atlas provides data for estimated per capita FAFH expenditure by county. This variable is divided by household income, resulting in FAFH expenditure as a share. The resulting measure is the average per capita expenditure on FAFH as a fraction household income, which serves as a point of reference for the magnitude on FAFH spending. Figures 20 depicts this variable after aggregating to the Nielsen market level

[10] Some researchers point to the recent rise of availability of FAFH in the US as a leading cause of the rise in both obesity and its related health problems. This is, in part, because food eaten away from home tends to be more energy-dense than food prepared at home, especially in the case of fast food establishments, which tend to feature tastier energy-dense foods and less fruits, vegetables, and whole grain foods (Jeffery & French, 1998; Stewart et al., 2004; Thompson et al., 2004; Creel et al., 2008). It is also reasonable to expect spending on FAFH to affect the types of food purchased at grocery stores. If that food is associated with a particular health outcome, then per capita expenditures on FAFH may have an impact on the food's effect.

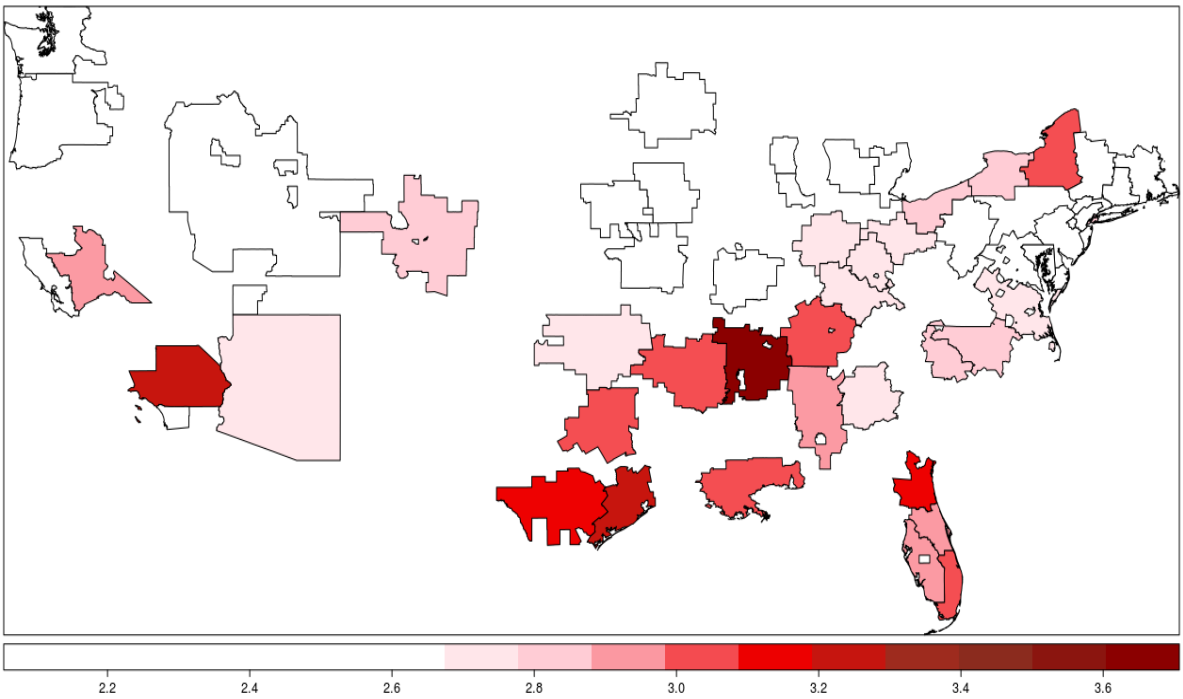


Figure 20: *2012 Average Annual Per Capita Expenditure on FAFH as a Percentage of Household Income*

Table 14 shows the summary statistics of the demographic, behavioral, and environmental variables that may further explain variation in health

Table 14: *Demographics and Environment Summary Statistics*

	Mean	**Std. Dev.**	**Min**	**Max**	**Median**
Physical Inactivity (% Prevalence)	24.05	3.95	15.94	32.41	24.49
Daily Smokers (% of Population)	11.43	2.96	5.22	17.11	11.43
Air Pollution (Air Quality Index)	45.24	7.35	28.17	73.76	44.52
Uninsured Adults (% of Population)	26.69	6.34	13.80	41.68	26.52
FAFH (% of HH Income)	2.72	0.31	2.16	3.61	2.73
Male Adult (% of Population)	49.06	0.53	47.57	50.21	49.05
Female Adult (% of Population)	50.94	0.53	49.79	52.43	50.95
Mean Age (Years)	37.69	2.32	31.39	44.50	37.50

Table 14 continued

Under 18 Years Old (% of Population)	23.84	1.83	20.06	30.17	24.04
2000-2010 Annual Population Change (%)	0.98	0.72	-0.26	2.50	0.88
White (% of Population)	68.51	14.15	33.05	89.39	70.81
Black (% of Population)	12.30	8.03	0.84	33.81	10.61
Hispanic (% of Population)	12.43	10.80	1.20	45.14	8.47
Other Minority (% of Population)	4.65	3.98	1.45	23.18	3.13
HH Size (Persons)	2.39	0.15	2.17	3.11	2.37
HH Income ($1k)	122.71	15.01	89.14	157.34	120.55
Married (% of Population)	64.56	4.93	44.67	78.64	64.36
Graduated College (% of Population)	30.50	5.26	21.02	41.57	30.45

CHAPTER 8. RESULTS

With these data, models are estimated to address the three questions posed at the outset: Do food-health correlations established at the individual level appear at the regional level with food purchases? Do any additional correlations perhaps not as well known appear at the regional level? Are there differences in food purchasing behavior between the extremes of the distribution of a given health measure?

For questions 2 and 3, the method of choice is sorting on each of the health outcomes, followed by a finite mixture model to classify markets by detecting patterns over multiple health dimensions simultaneously. If the diet-disease parallels suggested by individual-level studies hold to a reasonable degree at a more aggregate level, then differences in health outcomes should be accompanied by at least some differences in food purchasing patterns.

For question 1, the primary tool of analysis is linear regression modeling. Fundamentally, these questions ask whether, the food-health associations holding at the individual level are sufficiently in evidence at the level of marketing data. However, it is conceivable that, while there may be correlations of food and health, additional regressors such as demographic, environmental, and behavioral measures are more important in explaining health differences. Therefore, it is desirable to conduct further analysis. I do this by first estimating a series of univariate health-food regressions and examining the correlations between food purchasing and health. These are followed by models with the same food purchasing variables included with all non-food variables, to see whether the correlation seen in the previous iteration of modelling still holds. Finally, a model of each health variable regressed on all food and non-food variables jointly is estimated and compared to the previous two series of models. This not only evaluates

the correlation of food and health at the regional level, but may help to inform the importance of separating the effects of diet and non-food variables in this context.

8.1 Regression Analysis

To address the first research question, a series of linear regression models are estimated, beginning with the univariate cases cited in the nutritional epidemiology literature. For the i^{th} Nielsen market, the model takes the form

$$y_i = \alpha + \boldsymbol{F_i}\beta + \epsilon_i \qquad (1)$$

where y_i is a health measure increasing in disease level, $\boldsymbol{F_i}$ is a vector of food purchasing measures for each of the selected foods. Because many of these variables have different units and points of reference, they are all standardized to have mean zero and unit standard deviation before the models are estimated. The univariate estimation results are displayed in Tables 15-20 below, where each estimate is from a single univariate regression of that particular health outcome on that specific food and nothing else

Table 15: *Univariate OLS Estimation Results*

	Obesity		**Diabetes**		**Colorectal Cancer**		**Stomach Cancer**		**Stroke**		**Heart Disease**	
	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t
Nuts	-0.36	-2.71	-0.54	-4.49	-0.37	-2.85	-0.28	-2.03	-0.17	-1.25	-0.44	-3.44
Beef	-0.19	-1.37	-0.05	-0.36	-0.13	-0.92	-0.15	-1.07	-0.22	-1.63	-0.12	-0.84
Poultry	-0.47	-3.73	-0.09	-0.62	-0.14	-0.98	0.67	6.44	-0.35	-2.60	-0.03	-0.19
Soft Drinks	0.71	7.22	0.48	3.82	0.42	3.30	-0.33	-2.48	0.52	4.30	0.43	3.35
Seafood	-0.65	-6.06	-0.31	-2.28	-0.33	-2.47	0.48	3.89	-0.47	-3.74	-0.35	-2.64
Salty Snacks	0.76	8.35	0.51	4.22	0.43	3.33	-0.28	-2.06	0.56	4.84	0.58	4.97
Baked Goods	0.02	0.16	0.23	1.64	0.17	1.22	0.51	4.23	-0.35	-2.62	0.29	2.12
Candies	0.28	2.09	-0.16	-1.15	0.14	1.01	-0.37	-2.77	0.05	0.34	0.03	0.18
Bacon & Sausage	0.67	6.42	0.60	5.24	0.51	4.20	-0.05	-0.36	0.60	5.36	0.70	6.88
Fruit	-0.66	-6.20	-0.68	-6.51	-0.37	-2.83	0.30	2.19	-0.43	-3.39	-0.58	-5.08
Vegetables	-0.70	-6.95	-0.35	-2.65	-0.42	-3.30	0.27	1.98	-0.37	-2.83	-0.43	-3.35

The above results generally concur with the literature. Foods perceived as "healthy" tend to be negatively associated with disease, while foods perceived as "unhealthy" tend to be positively correlated with disease. Consider the example of produce purchasing. With stomach cancer being the lone exception, higher purchasing of fruit and vegetables is associated with better health at the market level. A standard deviation increase in seafood purchasing correlates with a reduction in stroke, 0.47 standard deviations. Likewise for nut purchasing and heart disease, a standard deviation increase in nut purchasing is linked with a 0.44 standard deviation drop in heart disease mortality. Unsurprisingly, regular soft drink purchasing coincides with higher values of the disease measures, in all cases but stomach cancer. Similarly for bacon and breakfast sausage, increasing purchasing by one standard deviation at the market level associates with a 0.51 standard deviation increase in the colon cancer measure. Increasing purchasing of salty snacks coincides with increased stroke, as well as with increased levels of nearly all disease measures. Beef purchasing does not correlate very strongly with any health measures in this context, and the results for poultry, baked goods, and candy appear to be mixed. Perhaps most surprising is that baked goods are negatively and significantly correlated with stroke and insignificantly correlated with obesity, and similarly for candy being negatively (although insignificantly) correlated with diabetes.

However, it is possible that the effect of food on health may change in the presence of demographics and environmental variables. This could be indirectly through behavioral change or directly. For example, there could be a change directly through biological factors. The next series of models are the same as the above with the inclusion of demographic (excluding as reference groups the male market population proportion and the white proportion of market population), behavioral, and environmental variables. The example cases shown are borrowed

from literature, specifically soft drink purchasing for obesity, vegetables for diabetes, bacon and breakfast sausage for colon cancer, beef for stomach cancer, seafood for stroke, and nuts for heart disease. The estimation results for these models are displayed in Tables 16-21

Table 16: *OLS Estimation Results for Obesity (Univariate Food: Soft Drinks)*

	Obesity	
	$\hat{\beta}$	t
Soft Drinks	0.03	0.33
Female	-0.08	-0.71
Physical Inactivity	0.27	2.85
Daily Smokers	0.10	1.46
Air Pollution	-0.06	-1.00
Uninsured Adults	-0.02	-0.22
FAFH	-0.13	-1.74
Age	0.05	0.32
Under 18	0.26	2.25
2000-2010 Annual Population Change	-0.03	-0.34
Black	0.25	2.53
Hispanic	-0.25	-2.55
Other Minority	0.10	1.33
HH Size	-0.12	-1.19
HH Income	-0.23	-2.09
Married	0.23	1.80
Graduated College	-0.17	-1.59
$\boldsymbol{R^2}$	0.94	

Table 17: *OLS Estimation Results for Diabetes (Univariate Food: Vegetables)*

	Diabetes	
	$\hat{\beta}$	t
Vegetables	0.13	1.54
Female	0.14	1.36
Physical Inactivity	0.39	4.40
Daily Smokers	0.03	0.44
Air Pollution	0.15	2.72
Uninsured Adults	0.18	1.91
FAFH	0.02	0.32
Age	0.07	0.55
Under 18	-0.09	-0.84

Table 17 continued

2000-2010 Annual Population Change	0.14	1.80
Black	0.05	0.52
Hispanic	-0.40	-4.31
Other Minority	0.28	4.16
HH Size	0.03	0.33
HH Income	-0.20	-1.96
Married	0.13	1.28
Graduated College	-0.27	-2.58
$\boldsymbol{R^2}$	0.95	

Table 18: *OLS Estimation Results for Colon Cancer (Univariate Food: Bacon/Sausage)*

	Colon Cancer	
	$\hat{\beta}$	t
Bacon & Breakfast Sausage	0.15	0.73
Female	0.04	0.13
Physical Inactivity	0.31	1.17
Daily Smokers	0.32	1.59
Air Pollution	-0.04	-0.23
Uninsured Adults	0.13	0.44
FAFH	0.16	0.74
Age	-0.03	-0.07
Under 18	0.29	0.90
2000-2010 Annual Population Change	-0.21	-0.93
Black	-0.02	-0.07
Hispanic	-0.22	-0.78
Other Minority	-0.13	-0.68
HH Size	0.11	0.44
HH Income	0.24	0.81
Married	-0.28	-0.88
Graduated College	-0.01	-0.02
$\boldsymbol{R^2}$	0.53	

Table 19: *OLS Estimation Results for Stomach Cancer (Univariate Food: Beef)*

	Stomach Cancer	
	$\hat{\beta}$	t
Beef	-0.06	-0.65
Female	0.42	2.03
Physical Inactivity	0.28	1.61
Daily Smokers	0.03	0.22
Air Pollution	-0.20	-1.82
Uninsured Adults	-0.10	-0.53
FAFH	0.14	1.00
Age	-0.11	-0.42
Under 18	0.31	1.44
2000-2010 Annual Population Change	-0.23	-1.51
Black	-0.07	-0.37
Hispanic	0.31	1.65
Other Minority	0.26	1.95
HH Size	0.34	2.13
HH Income	0.17	0.84
Married	-0.55	-2.66
Graduated College	-0.14	-0.71
$\boldsymbol{R^2}$	0.79	

Table 20: *OLS Estimation Results for Stroke (Univariate Food: Seafood)*

	Stroke	
	$\hat{\beta}$	t
Seafood	-0.38	-1.60
Female	-0.01	-0.05
Physical Inactivity	-0.01	-0.08
Daily Smokers	-0.02	-0.18
Air Pollution	-0.09	-0.70
Uninsured Adults	0.54	2.84
FAFH	0.11	0.81
Age	-0.26	-1.03
Under 18	-0.07	-0.33
2000-2010 Annual Population Change	-0.03	-0.18
Black	0.37	1.86
Hispanic	-0.29	-1.40
Other Minority	0.06	0.45
HH Size	-0.18	-1.02
HH Income	0.40	1.88
Married	0.30	1.33
Graduated College	-0.46	-2.37
$\boldsymbol{R^2}$	0.80	

Table 21: *OLS Estimation Results for Heart Disease (Univariate Food: Nuts)*

	Heart Disease	
	$\hat{\beta}$	t
Nuts	0.25	2.05
Female	-0.01	-0.06
Physical Inactivity	0.85	4.72
Daily Smokers	0.07	0.60
Air Pollution	0.06	0.63
Uninsured Adults	0.03	0.21
FAFH	0.08	0.70
Age	0.24	1.12
Under 18	0.31	1.69
2000-2010 Annual Population Change	-0.21	-1.63
Black	0.09	0.55
Hispanic	-0.02	-0.12
Other Minority	0.13	1.10
HH Size	0.23	1.68
HH Income	-0.19	-1.14
Married	-0.22	-1.27
Graduated College	0.03	0.18
$\boldsymbol{R^2}$	0.81	

The inclusion of demographic and environmental variables either substantially changes or entirely eliminates the associations between food and health in all cases. For example, holding all things constant, the correlation between diabetes prevalence and regular soft drink purchasing not only is no longer significantly different from zero, but also switched signs. Similarly for the link between nuts and heart disease, which is now the opposite sign of previous findings and is statistically significant. The only association maintaining sign and significance from the univariate models was that of seafood and fish with stroke incidence, but even the significance of this association diminishes from the 1% level to the 15% level.

The final set of regression models is all foods, all demographics (again excluding white and male population proportions), all behavioral variables, and all environmental variables predicting health outcome. The estimation results are below in Table 22.

Table 22: *Food + Non-Food OLS Estimation Results*

	Obesity		**Diabetes**		**Colorectal Cancer**		**Stomach Cancer**		**Stroke**		**Heart Disease**	
	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t
Nuts	0.13	0.95	0.14	1.06	-0.42	-1.17	-0.50	-2.30	0.20	0.84	0.12	0.71
Beef	0.07	0.81	-0.03	-0.29	0.14	0.59	-0.20	-1.40	0.11	0.74	-0.10	-0.87
Poultry	-0.20	-1.39	0.11	0.80	0.28	0.74	0.34	1.48	-0.13	-0.52	0.44	2.42
Soft Drinks	-0.08	-0.61	0.18	1.41	0.22	0.64	0.04	0.17	-0.23	-1.02	-0.19	-1.10
Seafood	0.26	1.09	-0.13	-0.58	0.08	0.13	-0.13	-0.35	-0.02	-0.05	-0.23	-0.75
Salty Snacks	0.13	0.94	0.01	0.12	-0.45	-1.30	-0.43	-2.04	0.45	1.98	-0.01	-0.05
Baked Goods	-0.14	-1.25	-0.02	-0.19	-0.16	-0.56	0.21	1.19	-0.33	-1.75	0.07	0.47
Candies	-0.13	-0.78	-0.17	-1.10	0.83	1.96	0.50	1.94	0.02	0.06	0.13	0.62
Bacon & Sausage	0.13	1.18	-0.04	-0.40	0.25	0.87	0.06	0.36	0.11	0.63	0.22	1.58
Fruit	0.08	0.74	-0.03	-0.30	0.09	0.32	-0.14	-0.81	0.10	0.55	-0.12	-0.86
Vegetables	-0.15	-1.15	0.16	1.28	-0.14	-0.40	0.25	1.19	-0.06	-0.29	0.09	0.56
Female	0.30	2.17	0.40	3.01	0.52	1.44	0.13	0.58	0.20	0.85	0.76	4.31
Physical Inactivity	0.12	1.44	0.00	0.04	0.37	1.70	0.09	0.66	0.00	-0.01	0.14	1.34
Daily Smokers	-0.04	-0.46	0.13	1.79	0.10	0.48	-0.13	-1.07	-0.04	-0.28	0.09	0.95
Air Pollution	-0.08	-0.62	0.15	1.27	0.18	0.58	0.02	0.09	0.42	2.04	0.04	0.28
Uninsured Adults	-0.13	-1.22	0.03	0.31	0.26	0.96	0.03	0.21	0.11	0.61	-0.09	-0.71
FAFH	-0.13	-0.82	-0.10	-0.69	-0.01	-0.03	0.02	0.09	-0.24	-0.92	0.39	2.00
Age	0.00	0.02	0.13	0.80	0.34	0.74	-0.07	-0.26	-0.27	-0.91	0.24	1.08
Under 18	0.24	1.76	-0.08	-0.59	0.43	1.20	0.37	1.70	-0.15	-0.64	0.18	1.03
2000-2010 Annual Population Change	-0.01	-0.05	0.15	1.46	-0.05	-0.17	-0.31	-1.84	0.08	0.48	-0.21	-1.57
Black	0.14	1.09	0.02	0.15	0.08	0.22	0.12	0.61	0.24	1.12	0.24	1.47
Hispanic	-0.28	-2.04	-0.42	-3.24	-0.27	-0.76	0.31	1.45	-0.26	-1.16	0.10	0.58
Other Minority	0.14	1.56	0.21	2.36	-0.14	-0.59	0.22	1.55	0.15	0.99	0.09	0.77
HH Size	-0.06	-0.50	0.13	1.12	0.13	0.42	0.14	0.73	-0.06	-0.32	0.16	1.05
HH Income	-0.31	-2.04	-0.12	-0.82	0.51	1.29	0.20	0.85	0.26	1.01	-0.18	-0.96
Married	0.19	1.27	0.01	0.09	-0.46	-1.16	-0.34	-1.39	0.17	0.66	-0.08	-0.43
Graduated College	-0.11	-0.79	-0.31	-2.39	0.12	0.34	-0.23	-1.10	-0.31	-1.37	0.00	-0.02
R^2	0.95		0.96		0.68		0.88		0.87		0.92	

The results are not nearly as clear as the output from the previous section, and are quite ambiguous. There are cases of foods having the opposite sign of the literature cited as well as in the earlier analyses, perhaps the most striking example being the correlation of soft drinks with obesity. However, there are still instances of "correct" sign between food and health – vegetables and obesity, salty snacks and stroke – although the magnitudes of these t-statistics tend to be small.

Including demographic and environmental variables clearly influences the point estimates in the models. Indeed, these variables are correlated with the food expenditure shares. To investigate the degree to which this is the case, I regressed each demographic, behavioral, and environmental variable on all food purchasing variables. The condensed regression output of OLS for each demographic and environmental variable on food purchasing is listed in the table in Appendix A-2. The average R^2 for all demographic variable regressions on food variables is 0.58, which suggests that food is, indeed, correlated with broadly defined consumer groups. This strong link between food and demographics was one of the concluding remarks suggested by Larson (2004) about the regional patterns found using cluster analysis.

I also estimated these equations with different regression techniques, essentially finding the same ambiguities in the results. For example, weighting each observation by the number of Nielsen households did not meaningfully affect the results of any regression model. A spatial error model (a linear regression with spatially correlated residuals) was also estimated to help control for unobservable correlations between neighboring markets.[11] While there is a high degree of spatial autocorrelation in the errors, the point estimates themselves are not

[11] For example, while demographic and environmental variables are included in the model, correlated unobservables such as culture and habits may bias the estimated standard errors. The neighbor-weighting scheme was k-nearest neighbors (k=3 here) since contiguity schemes are not feasible given the layout of the markets.

meaningfully different from those in OLS, therefore offering no clearer implications in the results.

8.2 Elastic Net

It is unclear what the importance of some demographics is in a regression model relative to others, and it is clear that inclusion of certain variables can completely change the sign and significance of the slope on the foods of interest. Some of these variables such as age and race have biological implications for both health (CDC, 2017) and food (Eicher-Miller et al., 2015; Eicher-Miller & Boushey, 2017), and some are tied to the effect of food on health (Saydah et al., 2007; Camhi et al., 2011). This makes building a parsimonious model difficult while avoiding multicollinearity among the predictors. Hence, a data-driven method is preferable to ad hoc variable inclusion and exclusion. To this end, the regression models above are estimated using more advanced techniques from machine learning, specifically variable selection with the elastic net method developed in Zou and Hastie (2005). This is a reasonable next step in model estimating, given no guiding theory from which to build a model, as well as the ambiguity of the results upon the inclusion of other variables in the model. If regional food purchasing patterns are related to spatial variation in health outcomes, then the significance of the point estimates may or may not appear in a simple linear model estimated with OLS.

The algorithm for an elastic net selects variables by penalizing the log-likelihood for any irrelevant regressors included in the model. Mathematically, the objective is

$$\min_{\{\beta_0,\beta\}} \frac{1}{N}\sum (y_i - \beta_0 - \boldsymbol{X_i}\beta)^2 + \lambda \left[\frac{(1-\alpha)\left(||\beta||_{\ell_2}\right)^2}{2} + \alpha||\beta||_{\ell_1} \right],$$

where $||...||_{\ell_2}$ is the ℓ_2 or Euclidean norm and $||...||_{\ell_1}$ is the ℓ_1 norm. There are two special cases of the penalization scheme. First is $\alpha = 0$, known as a ridge regression using only the ℓ_2 norm. This reduces the coefficients of correlated regressors towards one another, but does not lend itself to parsimony (Breiman, 1996). The second special case is when $\alpha = 1$, making the penalization dependent on the ℓ_1 or "Manhattan distance" norm. This is known as a LASSO (Least Absolute Shrinkage and Selection Operator) regression, which chooses one of multiple correlated regressors, discarding the others that are less correlated, regardless of the potential importance of the discarded variables. Thus, one distinct advantage of the elastic net where $\alpha \in (0,1)$ is its robustness to multicollinearity by adopting a mixture of these two variable selection strategies. If there are two variables working in tandem to predict a given health outcome, but whose levels are highly correlated, then the lasso penalization would pick only one, where the elastic net would keep both if needed. Conversely, the ridge regression would keep both even if one did not need to be in the model, but the elastic net would discard the less relevant of the two. The $\alpha -$level is chosen by minimizing the MSE while restricting $\alpha \in (0,1)$ to maintain use of an elastic net penalization scheme, and the value of the complexity parameter, λ, is chosen with a leave-one-out MLE cross validation scheme. The output of the elastic net is the remaining predictors for each health outcome equation, as opposed to a general OLS with all possible regressors on the right-hand side. If the elastic net keeps any food purchasing variables, then we know that there is reason to believe that, at a market level, the correlations are strong enough that the variation in these food purchasing patterns are good predictors of health trends. However, if the elastic net keeps only demographic variables, then the predictive power of those makes them better suited to explain than food purchasing patterns. It is reasonable to expect the algorithm to keep a mixture of both food and demographics in the final output.

Below in Table 23 are the results from the algorithm

Table 23: *Output from Elastic Net Algorithm*

	Obesity	**Diabetes**	**Colorectal Cancer**	**Stomach Cancer**	**Stroke**	**Heart Disease**
	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$
Intercept	0	0	0	0	0	0
Nuts	0	0	0	-0.0721	0	0
Beef	0	0	0	-0.0513	-0.0059	0
Poultry	0	0	0	0.1743	-0.0239	0
Soft Drinks	0.0759	0.0054	0	-0.0022	0.0408	0
Seafood/Fish	0	0	0	0.0437	-0.0375	0
Salty Snacks	0.0957	0	0	-0.0214	0.0835	0
Baked Goods	0	0	0	0.1434	-0.0671	0
Candies	0	-0.0647	0	0	0	0
Bacon & Sausage	0.0317	0	0	0.0083	0.078	0.1655
Fruit	-0.0347	-0.0749	0	0.0359	-0.0373	0
Vegetables	-0.0704	0	0	0	-0.0049	0
Physical Inactivity	0.2013	0.3255	0.1018	0.0248	0.053	0.4972
Daily Smokers	0.0439	0	0	-0.008	0.0025	0
Air Pollution	0	0.0741	0	0	0.0173	0
Uninsured Adults	0	0.1485	0	-0.0069	0.0844	0
FAFH	0	0.0241	0	0.0129	0.0665	0
Male Adult	0	-0.0799	0	-0.0652	0	0
Female	0	0.0068	0	0.0653	0	0
Age	0	0.0603	0	-0.1119	-0.0151	0
Under 18	0	-0.0191	0	0.0622	0.0351	0
2000-2010 Annual Population Change	0	0.036	0	-0.1349	0.0309	-0.0286
White	0	0	0	-0.0967	0	0
Black	0.0577	0.0729	0	0.0043	0.1206	0
Hispanic	-0.1775	-0.2674	0	0.0464	-0.0119	0
Other Minority	0	0.1198	0	0.0999	0	0
HH Size	0	0	0	0.0491	0	0
HH Income	-0.1333	-0.0939	0	0	-0.0322	0
Married	0.0875	0.0516	0	-0.1119	0.0424	0
Graduated College	-0.029	-0.2863	0	0	-0.0442	-0.0822
$\hat{\alpha}$	0.40	0.90	0.95	0.05	0.05	0.80
$\hat{\lambda}$	0.2866	0.0307	0.5328	0.7457	1.6951	0.1970

The output in the table indicates that regional food purchasing variation does offer some predictive power in tracking health outcomes, as do demographic and other variables. The effect

of a standard deviation increase of purchasing shares of soft drinks coincides with a 0.076 standard deviation increase in the prevalence of obesity, a 0.005 standard deviation increase in diabetes, and a stroke rate increase of 0.041 standard deviations. Expenditure shares of fruit are negatively correlated with obesity, diabetes, and stroke, but vegetable shares only with obesity and stroke. This would indicate that at the market level fruit is more correlated with diabetes than vegetables. While the correlation of bacon and breakfast sausage is positive with stomach cancer – consistent with the WHO report – there is zero correlation with colon cancer. In fact, colon cancer is correlated with physical inactivity alone in the output, the reason for which is not immediately clear. As indicated in one of the univariate models, markets with increased seafood purchasing do, indeed, have reduced stroke rates (an estimated reduction of 0.038 standard deviations for every standard deviation increase in seafood purchasing), although stomach cancer mortality tends to be higher. It is also interesting that spending on food away from home (commonly associated with obesity) appears to correlate only with diabetes, stomach cancer, and stroke. An estimated 0.024 standard deviation increase in the market diabetes prevalence coincides with every standard deviation increase in the household income share of food away from home, a change of 0.013 for stomach cancer incidence, and a change of 0.067 in the stroke rate. Physical inactivity by market coincides with increased adverse health outcomes in all cases – most strongly so in the cases of heart disease (estimated effect of 0.497), diabetes (estimated effect 0.326), and obesity (estimated effect 0.201). Markets with higher percent black population tend to have higher prevalence of obesity and diabetes, higher incidence of stomach cancer, and higher stroke mortality, which is consistent with a recent CDC report (2017).

In general, the outcome of the elastic net is more consistent with the literature cited than the larger OLS models, much like the univariate regressions. The colon cancer model was the

only case where $\hat{\alpha}$ is close to 1, a LASSO regression. Stomach cancer and stroke have the smallest $\hat{\alpha}$ values, close to a ridge regression. Thus, the estimations confirm that use of an elastic net is generally preferred to that of LASSO or ridge.

8.3 Discrete Market Groupings

To address the second and third questions, the same general procedure is applied: For a given health outcome variable sorted in descending order, the top 10 markets and the bottom 10 markets are selected[12], and their expenditure shares of each food are compared, once for each health outcome. This is followed by a more data-driven procedure, described below, and the results are compared. Markets that consistently appear in the "unhealthy" category are Memphis and New Orleans-Mobile, also Nashville and Birmingham except in the case of stomach cancer. There exist no markets in the set of "healthy" set every time, although markets scattered through 4 of the 6 healthy categories (e.g. low disease measure in all cases but stomach cancer and diabetes) are Minneapolis, San Diego, Boston, San Francisco, and Denver. In Table 24, I show the mean difference, $\overline{\Delta}$, of shares (purchasing shares in healthiest 10 markets minus those in the unhealthiest 10 markets, for every health measure individually) for each food and the t-test results against the null hypothesis of no difference in market level food expenditure shares in unstandardized percentages

[12] The names of these markets are listed in the appendix, two lists for each health outcome (e.g. the 10 least obese markets' names and the 10 most obese markets' names).

Table 24: *Food Purchasing Differences between the Top and Bottom 10 Markets*

	Obesity		**Diabetes**		**Colorectal Cancer**		**Stomach Cancer**		**Stroke**		**Heart Disease**	
	$\overline{\Delta}$	t	$\overline{\Delta}$	t	$\overline{\Delta}$	t	$\overline{\Delta}$	t	$\overline{\Delta}$	t	$\overline{\Delta}$	t
Nuts	0.11	1.20	0.28	3.48	0.23	3.01	0.16	2.03	0.04	0.52	0.30	3.90
Beef	0.31	2.88	-0.01	-0.05	0.14	0.89	0.10	0.68	0.28	1.82	0.12	0.78
Poultry	0.58	3.67	0.15	1.01	0.21	2.28	-0.52	-2.76	0.34	1.67	0.04	0.26
Soft Drinks	-1.93	-8.50	-1.14	-3.48	-1.10	-3.45	0.63	1.42	-1.36	-4.70	-1.00	-3.14
Seafood	1.05	6.28	0.39	1.92	0.63	3.98	-0.59	-2.49	0.66	3.06	0.56	2.94
Salty Snacks	-1.04	-8.18	-0.63	-3.63	-0.75	-4.08	0.31	1.24	-0.81	-5.93	-0.79	-5.57
Baked Goods	0.16	0.90	-0.32	-1.93	-0.21	-1.49	-0.47	-3.08	0.32	1.89	-0.27	-1.81
Candies	-0.39	-3.17	0.18	1.00	-0.21	-1.66	0.39	2.46	-0.19	-1.26	0.03	0.22
Bacon & Sausage	-0.43	-6.69	-0.29	-3.29	-0.29	-3.94	-0.00	-0.02	-0.33	-4.41	-0.37	-5.39
Fruit	0.61	5.22	0.61	5.39	0.41	2.77	-0.22	-1.41	0.36	3.08	0.55	4.23
Vegetables	1.14	8.93	0.50	2.20	0.75	4.39	-0.47	-1.56	0.63	3.30	0.67	3.54

Many of the differences in food purchasing are consistent with expectations arising from previous work. Beef, poultry, seafood and fish, fruit, and vegetables all have significantly higher food expenditure shares in the least obese 10 markets than in the 10 most obese markets, the largest difference being in purchasing of vegetables. Conversely, the shares for regular soft drinks, salty snacks, candy, and bacon and breakfast sausage are significantly higher in the obese markets, the strongest case being soft drinks. Shares for nuts and baked goods are not significantly different between the two market groups. Similar results appear for four of the five other health measures: soft drinks and salty snacks are generally purchased more in markets with higher values of disease measures, while fruits and vegetables are purchased more in the markets with healthier levels of the health measures. The lone exception for these is stomach cancer, which, as in the regression analyses, tends to deviate from the other health measures in terms of correlation with food purchasing.

The second approach to answering the second and third questions is Bayesian in nature. The clusters of markets are selected not by sorting and extracting one health measure at a time, but with a finite mixture model of the joint distribution of health. I concatenate all six of the

market-level health indicators (obesity, diabetes, heart disease, stroke, colorectal cancer, and stomach cancer) into a 52X6 matrix of health outcomes and run the Expectation Maximization (EM) Algorithm, which is an iterative process that maximizes the expected log-likelihood of the joint distribution of these 6 variables. The 3-component mixture model (log-likelihood of -751) slightly outperforms a 2-component model (log-likelihood of -759). The output of interest is the matrix of estimated posterior probabilities. These are stochastic vectors, each with dimension 52x1 measuring the estimated probability that a given market belongs to one of three mixture components which can be thought of as "healthy markets", "mediocre markets", and "unhealthy markets" with regards to combinations of the levels of each of the 6 health variables. Table 25 shows the correlation of the first and third posterior probability vectors, λ_1 and λ_3, with each of the health outcome variables

Table 25: *Correlating Health and Posterior Probabilities*

	$\lambda_1 = \Pr[Healthy\ Market]$	$\lambda_3 = \Pr[Unhealthy\ Market]$
Obesity	-0.45	0.43
Diabetes	-0.39	0.49
Colorectal Cancer	-0.65	0.67
Stomach Cancer	-0.20	0.02
Stroke	-0.49	0.45
Heart Disease	-0.35	0.27

The first component is the model-estimated cluster of "healthy" markets since the posterior probability is negatively correlated with nearly all disease measures – most strongly with heart disease (ρ =-0.65). There are several markets in the overlap of this clustering and that on the lowest 10 for obesity. The third component appears to be the cluster of "unhealthy" markets since the posterior probability is positively correlated with all disease measures, most strongly with colorectal cancer (ρ =0.67). Again, there is much overlap between this classification method and the top 10 obese markets.

It would appear that the mixture model is effective at sorting the markets according to the joint distribution of the six health measures. The next step is to repeat the previous analysis done for the clustering on obesity by calculating descriptive statistics of food purchasing patterns for markets classified by the EM algorithm. A healthy market is defined as having a posterior probability of the first mixture component, λ_1, greater than 0.50, and an unhealthy market has third posterior probability, λ_3, greater than 0.50. These markets are listed by name in the appendix. Below in Table 26 is the correlation of the first and third posterior probabilities with food shares, followed by the mean differences in food purchasing shares between markets for which $\lambda_1 > 0.50$ and markets for which $\lambda_3 > 0.50$ with their corresponding t-test statistics

Table 26: *Correlations of Posterior Probabilities 1 and 3 with Food Purchasing, Mean Differences, and T-Tests*

	$Corr[Share_i, \lambda_1]$	$Corr[Share_i, \lambda_3]$	$\overline{\Delta}$	t
Nuts	0.46	-0.39	0.23	3.47
Beef	0.22	-0.15	0.14	1.13
Poultry	-0.10	0.13	-0.17	-0.95
Soft Drinks	-0.08	0.07	-0.18	-0.70
Seafood/Fish	0.14	-0.06	0.12	0.82
Salty Snacks	-0.42	0.35	-0.54	-4.11
Baked Goods	-0.43	0.58	-0.55	-5.85
Candies	0.06	0.00	0.06	0.30
Bacon & Sausage	-0.23	0.15	-0.21	-3.92
Fruit	0.28	-0.31	0.25	1.99
Vegetables	0.30	-0.22	0.32	1.86

While the simple correlations are generally intuitive in sign, the results of the t-test for no differences in expenditure shares in healthy versus unhealthy markets are less extreme than in the clustering on any one health outcome individually. For example, while the shares of fruits and vegetables are higher in healthier markets, the level of significance is somewhat diminished. The only other food with significantly higher shares in the healthy markets is nuts. Shares for salty snacks, baked goods, and bacon and breakfast sausage are lower in the healthy markets. Other

foods, however, are not significantly different in purchasing shares between the two classifications of markets.

From these two analyses, I conclude that there is evidence consistent with the hypothesis posed in the third question, that is, there are meaningful differences in food purchases between more healthy and less healthy markets consistent with what has been observed at the individual level. Both segmentation by sorting and a mixture model appear to be effective at classifying markets in terms of broad patterns of health. Moreover, there are several corresponding differences in market-level food purchasing behavior, and these are largely consistent with the epidemiology literature.

Finally, what remains to be addressed is the second question posed at the outset of this chapter, that is, whether there are any food-health associations in the data that may not be as well known as those highlighted in the introduction. The same procedure used to address question 3 will be applied here as well, that is, the sorting on health measures followed by a finite mixture model. The key difference in answering questions 2 and 3 is that here in answering the forming does not use aggregated food purchasing measures, but rather disaggregated product modules as defined by Nielsen in the data. There are 605 such modules for which there is sufficient data to examine purchasing differences across the “healthy” and “unhealthy” markets. After excluding the modules for which the national expenditure falls below 0.025% (a total of 208 modules, with a remaining sample of 397 modules), t-tests are performed for purchasing differences, and the 5 modules with the most extreme positive purchasing differences and most negative purchasing differences are selected from the resulting t-tests. Table 27 shows the top 5 and bottom 5 foods for each of the health outcomes

Table 27: *Product Module Purchasing Differences*

	Obesity	**Diabetes**	**Colorectal Cancer**	**Stomach Cancer**	**Stroke**	**Heart Disease**
5 modules with most negative t-statistic	*-Toaster Pastries, -Breakfast Sausage, -Flavored Milk, -Fresh Buns, -Potato Toppings*	*-Margarine, -Fresh Onions, -Macaroni, -Crackers, -Powdered Creamer*	*-Non-Chocolate Candy, -Bubble Gum, -Carbonated Soft Drinks, -Jelly, -Processed Cheese Loaf*	*-Sweet Rolls, -Fresh Bread, -Grape Juice, -Bottled Water, -Fruit Juice Drinks*	*-Breakfast Sausage, -Canned Green Beans, -Refrigerated Biscuits, -Carbonated Soft Drinks, -Lunch Meat*	*-Chewing Gum, -Toaster Pastries, -Ravioli, -Cake Mixes, -Sugar*
5 modules with most positive t-statistic	*-Frozen Desserts, -Refrigerated Toppings, -Fresh Mushrooms, -Yogurt, -Milk*	*-Granola, -Misc. Fresh Fruit, -Mozzarella Cheese, -Trail Mixes, -Hot Cereal*	*-Mozzarella Cheese, -Milk, -Misc. Fresh Vegetables, -Frozen Desserts, -Fresh Mushrooms*	*-Cottage Cheese, -Canned Pumpkin, -Brown Sugar, -Granola, -Canned Pears*	*-Butter, -Fresh Muffins, -Yogurt -Macaroni, -Refrigerated Toppings*	*- Mozzarella Cheese, -Fresh Herbs, -Crackers, -Herbal Tea, -Granola*

Given the very fine level of disaggregation of the modules, the answer to the second question is not perfectly clear from the results. However, there are some interesting patterns that emerge from the data. For example, mozzarella cheese is purchased more in markets with lower diabetes, colon cancer, and heart disease. Similarly, purchasing shares for granola are significantly higher in markets with lower diabetes, stomach cancer, and heart disease. Carbonated soft drink purchasing is higher in markets with increased colon cancer incidence and stroke mortality, and toaster pastry purchases tend to be higher in markets with more obesity and heart disease. There are also other interesting differences not shown in Table 16. One example is that margarine purchases are significantly higher in the obese markets, whereas butter purchasing is significantly higher in the less obese markets. Another is that bottled water shares are higher in markets with low stroke and diet soft drink shares are higher in high-stroke markets. However, it is likely that some food module purchasing differences are correlated with lifestyle choices rather than indicative of possible diet-disease links. For instance, it is not likely that bubble gum

purchasing is higher in high-colon cancer markets because of previously unknown carcinogenic properties of bubble gum. Another example is refrigerated toppings' purchasing being higher in markets with lower stroke mortality.

The EM Algorithm was run and a finite mixture model was estimated jointly over the six health outcomes, followed by the same t-test on module-level purchasing patterns between markets classified (in the same fashion as done previously using the estimated posterior probabilities) as "healthy" and "unhealthy". The top 5 modules purchased in unhealthy markets are fresh baked cakes, sprayed butter crackers, refrigerated dough, fresh donuts, and dinner rolls. The top 5 modules purchased in healthy markets are trail mixes, coffee, nuts, frozen fruit, and specialty/imported cheese. Similar to the results of the health outcome-specific analysis above, bread and bread-like products tend to be purchased more in less healthy markets, whereas dairy and fiber-rich modules are purchased more in the healthier markets. Comparing these associations along with the remainder of the modules in Nielsen – with attention to modules listed separately but representing similar foods – is a potential topic of future study.

CHAPTER 9. CONCLUSIONS

In this study, the aim was to investigate whether aggregate food marketing data can be related to aggregate health measures, with attention to known relations. The ecological framework is potentially useful for tracking health as well as for potentially discovering new food-health associations not previously investigated. To this end, I addressed three questions related to regional patterns of health: Do food-health correlations established at the individual level appear at the regional level with food purchases? Do any additional correlations perhaps not as well known appear at the regional level? Are there differences in food purchasing behavior between the extremes of the distribution of a given market health measure? The aim was to exploit geographic differences in food purchasing in relation to geographic variation in health outcomes. Both are known to exhibit regionality, but the overarching goal was to detect instances where they are similarly regional and compare the resulting associations to those previously established by other means.

There were two possible outcomes to arise from this work: first, that upon using longitudinal and other individual-level nutritional epidemiology studies as a baseline for any particularly strong diet-disease associations, the same sign and possibly significance of these associations appear also at an aggregated regional level, at least in a large number of cases. The second possible outcome was that the market-level associations were either inconsistent, weak, have signs different from those in the individual-level studies, or some combination of these three shortcomings. The first case, at least to a reasonable degree, appears to be the outcome of this study. It is therefore reasonable to say that, according to the findings in this study, regional food marketing data has the potential to track health at a broad level.

The methods used in this study are likely of some benefit to future research as well. The sorting over the distribution of each health outcome helped identify many of the same markets chosen by a finite mixture model estimated with the EM Algorithm. While the results agreed in many of the cases, it appears that the mixture model is more conservative in its classification of market health status. It is also interesting to see how a mixture model performs at identifying regions with higher incidence, risk, mortality, or prevalence of multiple diseases jointly. By estimating the posterior probabilities of each market to belong to a particular class of markets, this provides a continuum on which to base analyses of these markets and allows for a more flexible taxonomy. Employing a machine-learning algorithm such as an elastic net performed well at answering the question of which variables mattered and to what degree. The results from OLS were not perfectly clear in their implications, but some of the effects of food purchasing on health were consistent with prior expectations. This was particularly evident in regressing health outcomes on individual foods in the univariate models.

There are limitations to this study as conducted. The selected Nielsen data cover a short timespan, which could be expanded by more years in future work. While the 11 foods in this study were selected by grouping combinations of like foods among 163 product modules out of more than 600 possible modules, there are likely other foods whose relation to health may be worth investigating. For example, the earlier years of the Nielsen Homescan data disaggregate the random weight foods. This disaggregation would permit, for example, the calculation of random weight apple sales and hence total apple sales, thus permitting the examination of apples' association with health. Furthermore, despite the array of demographic, behavioral, and environmental variables considered herein, there may be other confounding factors that, while not measured, may be important.

REFERENCES

Anding, Jenna D., Richard R. Suminski, and Linda Boss. 2001. "Dietary intake, body mass index, exercise, and alcohol: are college women following the dietary guidelines for Americans?" *Journal of American College Health* 49(4): 167-171.

Bazzano, Lydia A., Mary K. Serdula, and Simin Liu. 2003. "Dietary intake of fruits and vegetables and risk of cardiovascular disease." *Current atherosclerosis reports* 5(6): 492-499.

Breiman, Leo. 1996. "Stacked regressions." *Machine learning* 24(1): 49-64.

Camhi, Sarah M., George A. Bray, Claude Bouchard, Frank L. Greenway, William D. Johnson, Robert L. Newton, Eric Ravussin, Donna H. Ryan, Steven R. Smith, and Peter T. Katzmarzyk. 2011. "The relationship of waist circumference and BMI to visceral, subcutaneous, and total body fat: sex and race differences." *Obesity* 19(2): 402-408.

Center for Disease Control and Prevention. 2017. “Stroke Facts”. https://www.cdc.gov/stroke/facts.htm

DiNicolantonio, James J., Sean C. Lucan, and James H. O’Keefe. 2016. "The evidence for saturated fat and for sugar related to coronary heart disease." *Progress in cardiovascular diseases* 58(5): 464-472.

Eicher-Miller, Heather A., Victor L. Fulgoni III, and Debra R. Keast. 2015. "Energy and nutrient intakes from processed foods differ by sex, income status, and race/ethnicity of US adults." *Journal of the Academy of Nutrition and Dietetics* 115(6): 907-918.

Eicher-Miller, Heather A., and Carol J. Boushey. 2017. "How often and how much? Differences in dietary intake by frequency and energy contribution vary among US adults in NHANES 2007–2012." *Nutrients* 9(1): 86.

He, Ke, F. B. Hu, G. A. Colditz, J. E. Manson, W. C. Willett, and S. Liu. 2004. "Changes in intake of fruits and vegetables in relation to risk of obesity and weight gain among middle-aged women." *International journal of obesity* 28 (12): 1569-1574.

Imamura, Fumiaki, Laura O’Connor, Zheng Ye, Jaakko Mursu, Yasuaki Hayashino, Shilpa N. Bhupathiraju, and Nita G. Forouhi. 2015. "Consumption of sugar sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: systematic review, meta-analysis, and estimation of population attributable fraction." *Bmj* 351: h3576.

Hu, Frank B., and Walter C. Willett. 2002. "Optimal diets for prevention of coronary heart disease." *Jama* 288(20): 2569-2578.

Kim, Soowon, Pamela S. Haines, Anna Maria Siega-Riz, and Barry M. Popkin. 2003. "The Diet Quality Index-International (DQI-I) provides an effective tool for cross-national comparison of diet quality as illustrated by China and the United States." *The Journal of nutrition* 133(11): 3476-3484.

Liu, Simin, Mary Serdula, Sok-Ja Janket, Nancy R. Cook, Howard D. Sesso, Walter C. Willett, JoAnn E. Manson, and Julie E. Buring. 2004. "A prospective study of fruit and vegetable intake and the risk of type 2 diabetes in women." *Diabetes care* 27(12): 2993-2996.

Lund, Elizabeth K. 2013. "Health benefits of seafood; is it just the fatty acids?." *Food chemistry* 140(3): 413-420.

Mueller, S., Saunier, K., Hanisch, C., Norin, E., Alm, L., Midtvedt, T., and Clavel, T. 2006. "Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study." *Applied and environmental microbiology* 72(2): 1027-1033.

Must, Aviva, Jennifer Spadano, Eugenie H. Coakley, Alison E. Field, Graham Colditz, and William H. Dietz. 1999. "The disease burden associated with overweight and obesity." *Jama* 282 (16): 1523-1529.

Saydah, Sharon, Catherine Cowie, Mark S. Eberhardt, Nathalie De Rekeneire, and KMVenkat Narayan. 2007. "Race and ethnic differences in glycemic control among adults with diagnosed diabetes in the United States." *Ethnicity & disease* 17(3): 529-535.

Slavin, Joanne L., and Beate Lloyd. 2012. "Health benefits of fruits and vegetables." *Advances in Nutrition: An International Review Journal* 3(4): 506-516.

Willett, Walter C., Frank Sacks, Antonia Trichopoulou, Greg Drescher, Anna Ferro-Luzzi, Elisabet Helsing, and Dimitros Trichopoulos. 1995. "Mediterranean diet pyramid: a cultural model for healthy eating." *The American journal of clinical nutrition* 61(6): 1402S-1406S.

Wu, Lang, Zhen Wang, Jingjing Zhu, Angela L. Murad, Larry J. Prokop, and Mohammad H. Murad. 2015. "Nut consumption and risk of cancer and type 2 diabetes: a systematic review and meta-analysis." *Nutrition reviews* 73(7): 409-425.

Yang, Quanhe, Zefeng Zhang, Edward W. Gregg, W. Dana Flanders, Robert Merritt, and Frank B. Hu. 2014. "Added sugar intake and cardiovascular diseases mortality among US adults." *JAMA internal medicine* 174(4): 516-524.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2): 301-320.

APPENDIX

A-1.) Regressing Demographic and Environmental Variables on Food Purchasing

	y = Physical Inactivity t-statistic	**y = Air Pollution t-statistic**	**y = Daily Smokers t-statistic**
Intercept	0	0	0
Raw Nuts	-1.99	1.47	-0.22
Beef	-0.27	-0.19	-0.80
Poultry	0.39	1.14	-0.73
Soft Drinks	0.69	-0.01	0.62
Seafood/Fish	0.00	-2.00	-1.09
Salty Snacks	0.11	0.65	-0.37
Baked Goods	1.58	0.44	1.46
Candies	-0.19	-2.55	0.33
Bacon & Sausage	1.67	-0.53	-1.36
Fruit	-3.48	0.98	-3.91
Vegetables	-0.86	-0.56	0.39
$\boldsymbol{R^2}$	0.81	0.25	0.62

	y = No Health Insurance t-statistic	**y = FAFH Percent t-statistic**	**y = Proportion Male t-statistic**
Intercept	0	0	0
Raw Nuts	0.55	1.64	2.03
Beef	-0.50	0.66	1.35
Poultry	-0.61	-0.05	-2.87
Soft Drinks	-0.97	-0.73	-0.86
Seafood/Fish	0.08	-1.32	0.13
Salty Snacks	-0.51	-1.47	-0.66
Baked Goods	-4.03	-0.04	-2.52
Candies	-3.18	-2.88	-0.49
Bacon & Sausage	1.60	3.77	0.97
Fruit	0.09	0.17	2.31
Vegetables	-1.24	-0.80	-1.49
$\boldsymbol{R^2}$	0.63	0.51	0.74

	y = Female t-statistic	**y = Household Income t-statistic**	**y = Household Size t-statistic**
Intercept	0	0	0
Raw Nuts	-2.03	-0.54	-2.25

Beef	-1.35	-0.29	-0.27
Poultry	2.87	-1.17	1.17
Soft Drinks	0.86	-0.19	-1.51
Seafood/Fish	-0.13	2.58	0.31
Salty Snacks	0.66	2.03	1.10
Baked Goods	2.52	-1.03	-2.21
Candies	0.49	0.98	2.25
Bacon & Sausage	-0.97	-1.93	-0.89
Fruit	-2.31	2.60	0.66
Vegetables	1.49	1.55	-0.63
$\boldsymbol{R^2}$	0.74	0.71	0.35

	y = Proportion White t-statistic	**y = Black t-statistic**	**y = Hispanic t-statistic**
Intercept	0	0	0
Raw Nuts	-2.05	-0.30	1.27
Beef	0.39	-0.44	0.30
Poultry	-2.17	1.00	0.81
Soft Drinks	-0.89	0.83	-0.57
Seafood/Fish	-0.35	1.49	-0.60
Salty Snacks	-0.07	1.36	-1.05
Baked Goods	2.05	-0.58	-2.23
Candies	3.68	-0.91	-2.55
Bacon & Sausage	-0.36	1.15	0.21
Fruit	-1.89	-0.39	1.90
Vegetables	1.04	-0.25	-1.16
$\boldsymbol{R^2}$	0.72	0.49	0.61

	y = Other Minority t-statistic	**y = College Graduate t-statistic**	**y = Under 18 t-statistic**
Intercept	0	0	0
Raw Nuts	2.59	-0.63	-0.21
Beef	-0.61	-0.37	-0.61
Poultry	1.80	-0.54	1.57
Soft Drinks	2.04	0.13	-1.25
Seafood/Fish	-0.47	0.71	-1.07
Salty Snacks	-0.10	0.17	0.51
Baked Goods	0.86	-1.73	-3.46
Candies	-1.77	0.68	-0.25
Bacon & Sausage	-1.58	-1.11	0.63
Fruit	1.59	1.77	1.51
Vegetables	0.31	1.33	-1.43
$\boldsymbol{R^2}$	0.54	0.51	0.51

	y = Population Change t-statistic	**y = Age t-statistic**	**y = Married t-statistic**
Intercept	0	0	0
Raw Nuts	0.51	-0.90	-2.03
Beef	-2.03	1.17	-0.05
Poultry	-1.20	-2.44	-0.78
Soft Drinks	-1.22	-0.13	-0.60
Seafood/Fish	-0.30	0.90	-0.18
Salty Snacks	-1.90	0.08	0.52
Baked Goods	-3.35	2.58	-2.47
Candies	-1.97	0.64	2.06
Bacon & Sausage	0.80	-0.11	-0.19
Fruit	-1.54	-1.51	-1.03
Vegetables	-0.64	1.72	0.49
R^2	0.57	0.50	0.57

A-2.) Market Segmentation: Sorting & Discrete Selection

	Obesity	**Diabetes**	**Colorectal Cancer**	**Stomach Cancer**	**Stroke**	**Heart Disease**
Top 10 Markets (Highest Ranking)	*Memphis, Little Rock, New Orleans-Mobile, Birmingham, Oklahoma City-Tulsa, Nashville, Louisville, Detroit, Grand Rapids, Indianapolis*	*Memphis, Birmingham, Little Rock, New Orleans-Mobile, Nashville, Louisville, Cincinnati, Cleveland, Orlando, Tampa*	*Louisville, Memphis, New Orleans-Mobile, Pittsburgh, Chicago, St. Louis, Birmingham, Nashville, Cincinnati, Cleveland*	*Urban NY, Suburban NY, Los Angeles, Chicago, Exurban NY, Hartford-New Haven, San Francisco, Memphis, New Orleans-Mobile, Louisville*	*-Birmingham, Memphis, Nashville, Dallas, New Orleans-Mobile, Cincinnati, Little Rock, Baltimore, Charlotte, Columbus*	*Memphis, Birmingham, Oklahoma City-Tulsa, Little Rock, Detroit, New Orleans-Mobile, Nashville, Pittsburgh, St. Louis, Cleveland*
Bottom 10 Markets (Lowest Ranking)	*San Diego, Denver, San Francisco, Los Angeles, Urban NY, Miami, Suburban NY, Exurban NY, Boston, Hartford-New Haven*	*Denver, San Diego, Minneapolis, Salt Lake City, Seattle, San Francisco, San Antonio, Suburban NY, Chicago, Sacramento*	*Grand Rapids, Phoenix, San Diego, San Francisco, Raleigh-Durham, Tampa, Seattle, Boston, Miami, Los Angeles*	*Tampa, Phoenix, Des Moines, Omaha, Portland OR, Denver, Indianapolis, Grand Rapids, St. Louis, Orlando*	*Urban NY, Suburban NY, Boston, Exurban NY, Albany, Hartford-New Haven, Phoenix, Tampa, Syracuse, Minneapolis*	*Minneapolis, San Francisco, Denver, San Diego, Portland OR, Seattle, Phoenix, Boston, Miami, Omaha*

A-3.) Market Segmentation: Mixture Model Posterior Probabilities

"Unhealthy" Markets $(\boldsymbol{\lambda_3 > 0.5, n = 26})$	*Albany, Baltimore, Birmingham, Boston, Buffalo-Rochester, Cincinnati, Cleveland, Columbus, Detroit, Exurban NY, Grand Rapids, Hartford-New Haven, Indianapolis, Little Rock, Louisville, Memphis, Nashville, New Orleans- Mobile, Oklahoma City-Tulsa, Philadelphia, Pittsburgh, St. Louis, Suburban NY, Syracuse, Urban NY, Washington DC*
"Healthy" Markets $(\boldsymbol{\lambda_1 > 0.5, n = 11})$	*Denver, Miami, Minneapolis, Omaha, Phoenix, Portland OR, Raleigh-Durham, San Diego, San Francisco, Seattle, Tampa*

PART III. DEFICIENT DIETARY BEHAVIOR IN LOW-INCOME AMERICANS: ASSESSING THE ROLE OF DIET COSTS

Some of the data in this work is calculated (or Derived) based on data from The Nielsen Company (US), LLC and marketing databases provided by the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

CHAPTER 10. INTRODUCTION

In 1977, the US Senate Select Committee on Nutrition and Human Needs began the development of the Dietary Goals for the American People. These goals made both nutrient-based and food-based recommendations for the American people (USDA, 2015). This led to the development of the better-known Dietary Guidelines for Americans (DGA), which first appeared in 1980 as a collaborative effort of the US Department of Agriculture and the US Department of Health & Human Services. The purpose of the DGA was to promote improved diets in order to prevent chronic disease in light of emerging research at the time suggesting a link between diet and health outcomes. Since its initiation, the Guidelines undergo thorough revisions every five years. These revisions are based on developments in nutrition research and current health trends in the US.

One major element highlighted in the revisions of the DGA since the 1970s[13] is combatting the increasingly serious epidemic of obesity. A staggering 70% of adults were classified as overweight and nearly 37% as obese in 2017 (CDC, 2017). This is problematic because research shows that obesity is related to various adverse health outcomes and chronic disease and contributes to as many as an estimated 112,000 preventable deaths in the US (Flegal et al., 2007).

Certainly, obesity is a complex issue with multiple contributing factors, but the greatest contributor is a lack of balance between a person's energy intake and energy expenditure (NIH, 2017). Thus, it is reasonable to say that an individual's diet is a principal link to whether or not that person is overweight or obese. In addition to obesity, there is considerable evidence that

[13] In 1980, the prevalence of obesity in US adults was 13.4%, jumping to 37% in 2017 (Office of the Surgeon General, 2010).

poor diets contribute to increased risk of coronary heart disease and type-2 diabetes (Must et al., 1999; Anding et al., 2001; He et al., 2004; Slavin & Lloyd, 2012; Imamura et al., 2015; DiNicolantonio et al., 2016).

The average American falls short of consuming a diet adhering to the DGA[14], which from its beginning has emphasized moderation of consuming fats, sugars, and sodium. In fact, the 2010 report of the Surgeon General's office outlines multiple strategies for preventing obesity that concern these nutrients. Such strategies include individuals reducing "consumption of energy dense foods that primarily contain added sugars or solid fats" as well as "sodas and juices with added sugars", and choosing to consume more "low-fat or non-fat dairy products" and "fruits, vegetables, whole grains, and lean proteins" (2010). Furthermore, the Department of Health & Human Services Healthy People 2020 initiative promotes an objective specifically targeting overweight and obesity, including a section which aims at reducing Americans' consumption of fat and sugar (HP2020, 2017).

The problem of obesity is not equally distributed across socioeconomic groups. For instance, obesity is disproportionately prevalent in minorities and those in lower income categories (Drewnowski & Specter, 2004; Guo et al., 2004). Studies on dietary behavior show that, on average, Americans with lower income have more calorie-dense, high-fat, and high-sugar diets than their high-income counterparts, and tend to consume insufficient fruits and vegetables (Drewnowski & Specter, 2004; Guo et al., 2004; Golan et al., 2008). Naturally, much research and speculation about the potential causes of this widespread pattern of nutritional deficiency, particularly the role of income, has developed. An obvious potential culprit is the

[14] A diet that is compliant with the Guidelines will be informally referred to henceforth as a "healthy" diet or something similar.

price of food, bringing into question whether or not low-income consumers can afford a healthy diet. Therefore, I propose testing the following null hypothesis:

- ***Hypothesis***: There is a positive relationship between the healthfulness of a diet and its cost.

CHAPTER 11. LITERATURE REVIEW

11.1 Literature for/against the Affordability Axiom

There have been several studies contending that "healthy" foods are more expensive than "unhealthy" foods (Mooney, 1990; Jetter & Cassady, 2006; Maillot et al., 2007; Drewnowski, 2010; Monsivais et al., 2011; Rao et al., 2013). Their general conclusion is that low-income households can only easily afford energy-dense foods, which are said to be low cost. The affordability axiom (a healthy diet costs more money than an unhealthy diet) is a recurring outcome in these studies. However, as discussed below, the methods and reasoning used in these studies are suspect. If the null hypothesis stated above does not hold and these studies' conclusions are not valid, which is what I argue, then there are other forces giving way to low-income Americans' lack of compliance with the DGA.

External influences such as the popular press or TV news may also contribute to consumers' belief in the alleged high cost of healthier eating. A study by Pettigrew (2016) speculates that these sources may be influencing consumers by suggesting ideas such as, "bad food is tasty" and "healthy food is boring and expensive". As a result, consumers cultivate the belief that healthy food is more expensive than unhealthy food and is less tasty, thus excusing their bad dietary habits. The author also cites evidence that healthy food campaigns and social marketing programs have proved relatively ineffectual in shifting consumption. Similarly, Hill et al. (2016) find that an individual's belief that healthier food is more expensive is not only easily influenced by outside forces, but also is manipulated by the individual's ultimate food intake goals. Participants on a special diet almost uniformly rejected the belief that healthy foods cost more than unhealthy foods, and the same was true for participants who self-identified as restrained eaters. The opposite was true for people who neither were on a diet nor had any

intentions of dieting and did not self-identify as restrained eaters. Although their study is not without its limitations (small sample size being chief of these), it still offers evidence that the affordability axiom is largely a psychological construct rather than an economic phenomenon.

A major flaw in many of the of the studies purporting the higher cost of healthy eating is that they often measure food cost as dollars per calorie, or "food energy cost", which makes very little economic sense. Paraphrasing Chen et al. (2012), energy density has a negative shadow value in a world of excess calories, and Darmon et al. (2005) even acknowledge that fresh fruits and vegetables have a very high nutrient-to-price ratio. Moreover, consumers do not tend to make their grocery purchasing decisions based on the caloric or nutrient content of foods, but rather quantity-based criteria such as edible weight and serving sizes (Rolls et al., 2002; Krukowski et al., 2006). In a USDA study, Carlson & Frazão (2012) show that the method used to measure the cost can determine the outcome of the question of whether or not healthier foods cost more. They demonstrate this by calculating the cost of food in three ways, (i) dollars per calorie, (ii) dollars per edible gram, and (iii) dollars per serving. Only in the first case was healthier food found to be more expensive. A more recent study by Davis & Carlson (2015) showed that the correlation between energy cost and energy density is spurious and energy density does not reduce cost of food. Put differently, the negative relation between energy cost and energy density is *necessarily* negative – calories in the denominator on one side of the equation and calories in the numerator on the other side of the equation is a mathematical artifact leading to an inverse association. Later work by Drewnowski (2013; 2015) appeared to retreat from measuring food cost in dollars per calorie. His 2013 study acknowledges that one's choice of cost metric alone can determine conclusions, and his 2015 study uses a new "nutrient density per dollar" measure, which can only increase when nutrient density increases (the numerator of

this cost metric is "good" nutrients per 100 kcal minus "bad" nutrients per 100 kcal, and the denominator is dollars). Unsurprisingly, empirical evidence for this positive relationship between his newly derived cost measure and nutritional value of foods was presented as new evidence of the role of the alleged costliness of healthy food in the prevalence of obesity of low-income Americans.

11.2 Validation of Dissenting Literature

Even simple correlations fail to support the affordability axiom. Figure 1 shows the scatterplot of dollars per 100g on the vertical axis and the estimated "healthiness" of food items as measured by a nutrient profiling index developed by Arsenault et al. (2012), and a line illustrating a univariate OLS regression using data from NHANES 2003[15]

[15] The measures of nutritional quality and food prices and the data from which they are calculated are described below. I use later years of these same data and the same formula developed by Arsenault et al. (2012) in my modeling to test the null hypothesis.

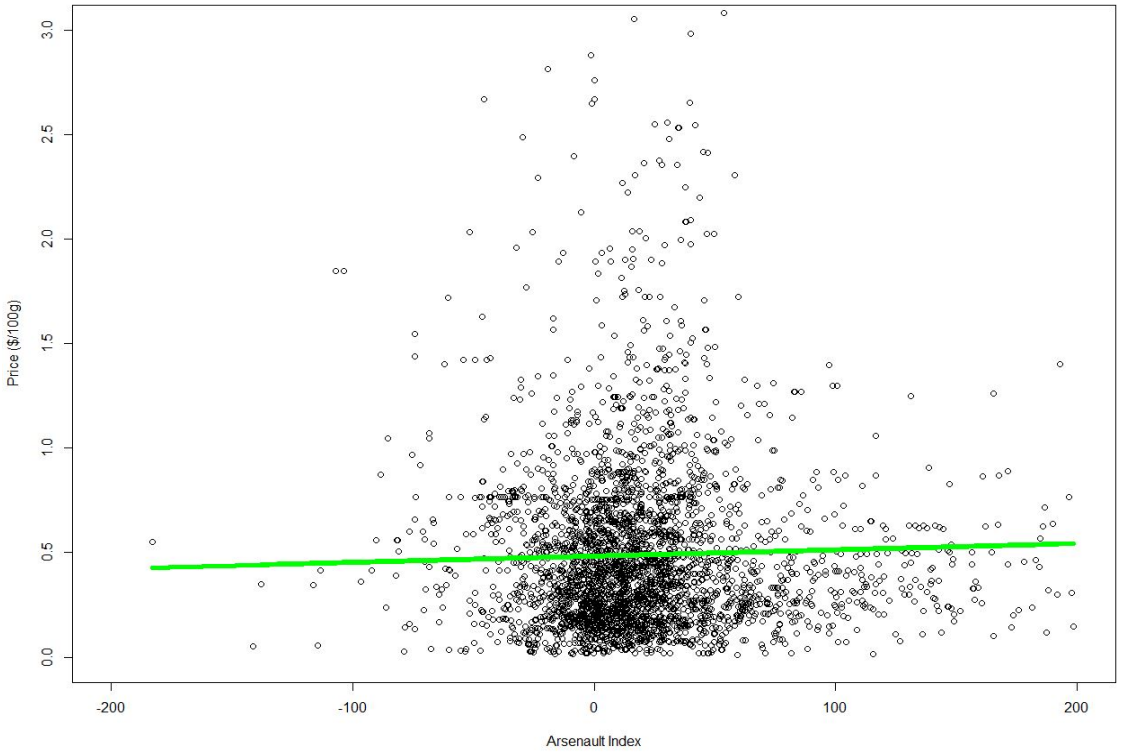


Figure 21: *Scatterplot of Food Price and Nutritional Value*

There is no apparent positive relationship between the two variables, as the null hypothesis would suggest. In fact, Table 28 shows that the univariate regression of food prices on weighted nutrient density measured by the index developed by Arsenault et al. (2012) does not support the claim that healthy foods are costly.

Table 28: *Price Regressed on Weighted Nutrient Density (OLS)*

VARIABLE	Parameter Estimate	t-statistic
Intercept	0.48	62.90
Nutrition Index	$4.1(10^{-5})$	1.69
$\boldsymbol{Adjusted\ R^2}$	0.00	

While the point estimate on the nutrition index is positive and significant at the 10% level, it is of a trivial magnitude: a unit change in the Arsenault nutrient profiling index (approximately 5% of

the mean of the index) accompanies a change in the price of food smaller than one one-hundredth of a penny per 100g (less than 0.001% of the mean of the price). A standardized regression estimates that less than a 0.05 standard deviation change in price follows a 1 standard deviation change in the nutrition index, also quite small. If healthier foods cost more, then a nontrivial correlation would be expected, which is not what the data show.

The same is not true of the relationship between calorie density and price, which is contrary to the main argument of the literature claiming calorie-dense foods are so cheap (e.g. Drewnowski & Specter, 2004; Guo et al., 2004). Figure 22 depicts the scatterplot of food price and energy density (kcal per 100g). There is a clear positive relationship

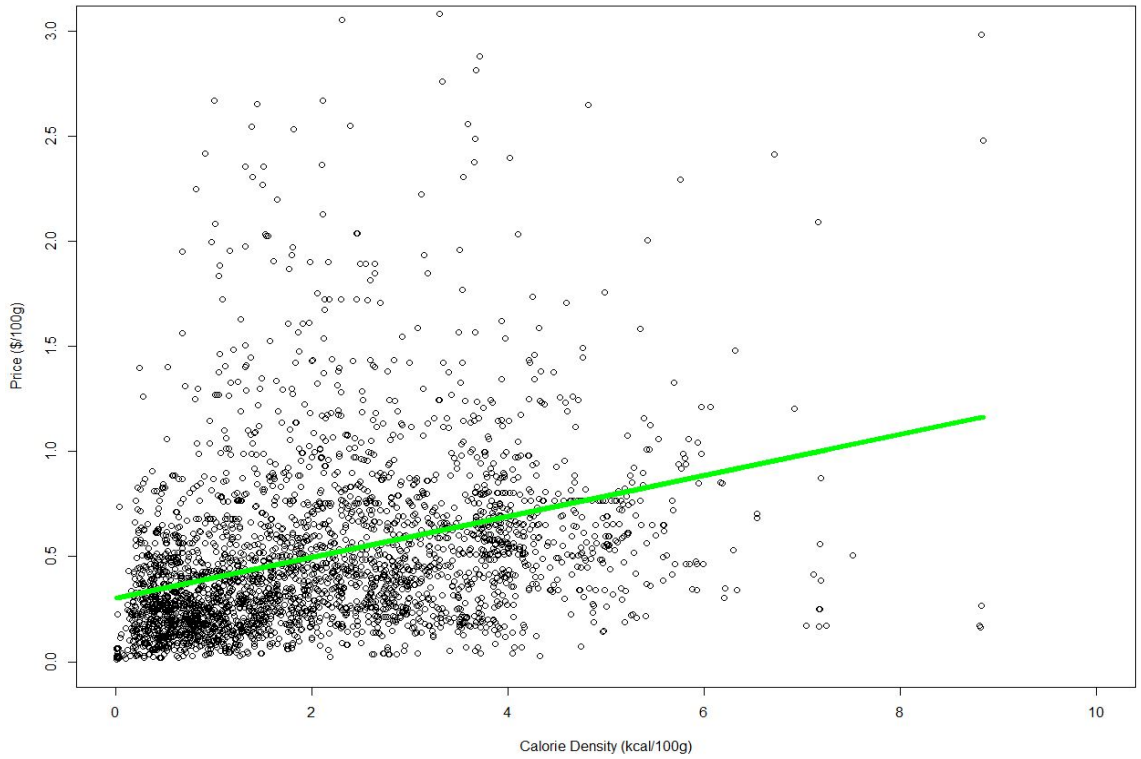


Figure 22: *Scatterplot of Food Price and Calorie Density*

The green line is a univariate OLS regression of food price on calorie density, the results of which are in Table 29

Table 29: *Price Regressed on Calorie Density*

VARIABLE	Parameter Estimate	t-statistic
Intercept	0.30	27.32
Calorie Density	0.10	20.89
$\boldsymbol{Adjusted\ R^2}$	0.09	

A unit change in calorie density (approximately 52% of the mean calorie density) coincides with a change of 9.7 cents per 100g (just less than 20% of the mean price). A standardized beta in this regression is 0.294, six times larger than that of the healthiness regression. Also notice that the adjusted R^2 is 0.09, which is much higher than that reported in Table 28.

4.3 Alternatives

The above correlations suggest that it is not healthy foods that tend to be priced more, hence a healthy diet is likely achievable at low cost. Indeed, the USDA has a developed portfolio of foods and food groupings that are both healthy and low in cost, known as the Thrifty Food Plan. One serious problem with the Thrifty Food Plan is that it ignores palatability. Drewnowski & Eichelsdoerfer (2010) point out that many of the foods in the TFP are, indeed, cheap and healthy, but may not be very tasty. This is not a new idea. Stigler (1945) developed an optimal diet for an average American male, minimizing the total expenditure on food subject to his minimum nutritional requirements. This diet included dried navy beans, raw cabbage, raw spinach, wheat flour, and evaporated milk. Silberberg (1980) observed this and proposed an Engle Curve type of explanation. First treating healthiness and tastiness of a diet as substitutable "inputs", he showed using survey data that as the ratio of actual cost of food to minimal cost of food increases with income, nutrition's share of expenditure falls with income. His theory

proposes that, once the individual's nutritional requirements are satisfied, any additional money spent on food improves palatability.

The conclusions from Silberberg's model estimated on survey data appear in recent studies as well. For example, Irz et al. (2015) developed a mathematical optimization program of dietary simulations. Their model simulations showed that reducing expenditure on a diet of fixed nutritional quality significantly reduces the palatability of that diet. A paper by Binkley & Golub (2011) previously argued in favor of this taste-nutrition tradeoff and provided empirical evidence in consumers' grocery purchasing. They found a significantly positive relationship between income and purchases of low fat, reduced sugar, and high fiber groceries, after controlling for demographics. That is, consumers with lower income tend to choose the options with higher fat, more sugar, and less fiber, even though the less healthy options of the foods examined cost either the same or more than the healthier options. Moreover, the commodities chosen were so simply differentiated (e.g. sugar or calorie content in diet versus regular soft drinks) that lack of knowledge about nutrition was unlikely to be an obstacle to choosing the healthier of two options, either. A similar conclusion followed in the work by Chen et al. (2012).

Prioritizing the palatability or taste of foods over their healthiness is well known. A recent choice experiment conducted by Malone & Lusk (2017) found that American consumers are willing to pay more money to have tastier meat than they are for healthier or even safer meat. This is consistent with a study by Glanz et al. (1999) that found US adults tend to place their highest value on the taste of food, followed by cost. In this same vein, Blaylock et al. (1999) write,

> In the long run, taste considerations may simply prevail: habits and other forces may be too difficult to overcome. Similarly, the uncertain future benefits of better nutrition—you have to die of something—may outweigh the perceived potential benefits of healthy eating. Put differently, for many people healthy eating is

> just not worth the effort and sacrifice…. Convincing people of the long-run benefits of good nutrition is clearly made more difficult if immediate gratification is given a higher priority.

This certainly appears to be consistent with what we observe: for many American adults, there is considerable evidence of a tradeoff between the nutritional quality of a diet and its palatability. If pleasure trumps health, then a consumer is unwilling to sacrifice taste for improved health – if, indeed, such a sacrifice becomes necessary – and it is, therefore, no surprise that the tendency of American adults is to value palatability over diet quality.

If, in fact, healthier eating does cost more, then the prevalence of the dietary behavior of low-income Americans comes as no surprise; in fact, one would conclude such an outcome naturally. However, if no such positive "healthiness-costliness" relation exists, and the dietary deficient behavior of low-income consumers goes beyond cost, then the problem requires further investigation, which is the purpose of this study. While the dietary deficient behavior of consumers with lower income is not due to the higher cost of healthier diets, it is still inextricably linked with their limited income.

CHAPTER 12. THEORY

12.1 Income and Constrained Utility

The first theory is a stylized, conceptual model of utility maximization. Suppose that there exist two consumers, identical in all regards except food budgets. Thus, both have the same preferences and their utility is derived from the foods they consume. Further assume that, neither consumer is initially aware of any value of nutrition, making their utility a function of tastiness only. That is, the two consumers have the same utility function

$$U = u(taste)$$

but face different food budget constraints. Because neither consumer is aware of nutritional value of the foods they eat, both will choose diets maximizing taste subject to their food budget constraints.

Now suppose that both consumers are made aware of which dietary choices constitute healthy eating and are faced with imposing an additional constraint on U to meet minimum dietary quality requirements. There are two possible outcomes, either that utility or taste will not change because the foods already chosen happen to satisfy the nutrition constraint, or that taste must be sacrificed to accommodate nutrition. In the latter case, the only ways to maintain the current level of taste are to either increase the portion of income allocated to food or ignore the nutrition constraint. Increasing the food budget is less of a problem for the wealthier consumer. The pressure to ignore the nutrition constraint is thus greater for the consumer with limited income, and as a result, those with less income will be more likely to consume less healthy foods.

The implications are straightforward: when the food budget expands for a consumer, more options are considered – options that were not previously considered due to their high cost (the initial set of choices will tend to emphasize low cost foods). The foods that are in the

expanded set but not in the initial set are considered for the reason that they will preserve or improve the initial level of taste as well as satisfy the nutrition constraint. Because these new foods appear only in the expanded set, cost will not decrease. Hence I anticipate finding that diets with simultaneously high levels of tastiness and healthiness tend to cost more, as do diets with more variety.

12.2 Expected Utility and Health as a Means to Longevity

The second theory that can help explain the link between poverty and dietary deficiencies is not new to economists. In short, consumers with less income place a lower value on longevity and hence the means to attain it, which in the context of this essay is a healthy diet. Because income produces utility, low-income consumers are giving up less expected utility by reducing expected lifespan. Hence they have less incentive to adopt healthy behavior, especially when they must forego pleasure to do so. Becker & Murphy (1988) showed that an increase in expected earnings raises the cost of consuming unhealthy and addictive goods because associated negative effects on productivity caused by health problems or death can incur greater losses in those increased expected earnings. This same argument goes as far back as the work presented by Grossman (1972) who modeled the demand for health as a form of human capital, particularly with regards to years invested in education and the associated value of loss incurred from early death. Thus, low-income consumers have lower expected utility, and a resulting behavior is that low-income Americans will tend to have less healthy diets.

More recently, Binkley (2010) presented a similar argument with regards to smoking cessation, finding that low-income Americans are significantly less likely to quit smoking than those with higher income, which is consistent with the theory that consumers with lower income tend to place a lower value on health as a means to longevity.

This same theory can be tested concerning food-purchasing behavior across different income levels. Indeed, subsequent work by Binkley & Golub (2011) and Chen et al. (2012) found further evidence consistent with this theory. Their results show that given the choice of more healthy and less healthy versions of the same foods at the same price, low-income consumers are more likely than high-income consumers to choose the less healthy version. In other words, affordability is not the limiting factor. However, neither of these studies controlled for access of the healthier versions of foods. A contribution that I can offer is to improve on these models by controlling for access.

CHAPTER 13. THEORY 1 METHODS

13.1 Data

The first task of this study is to show empirical evidence for the tradeoff of a diet's healthiness and palatability working in conjunction with its associated cost. In order to do so, I compute measures of healthiness, tastiness, and variety and estimate a model of diet cost. The measure of nutritional quality of diets is described below, but the procedure used to obtain it also aids me in my measure of diet palatability. What is critical about these two variables is their interaction, specifically its effect on diet cost, which was discussed in the presentation of the theory. Diet variety, as it pertains to diversity of foods consumed, can be measured in multiple ways and will serve primarily as a control in my model.

Calculating each of these measures requires data on individual food intake, food preferences, and food prices. I use data from the National Health And Nutrition Examination Survey (NHANES) over 2007-2010. This survey, administered by the CDC, collects dietary intake through 24-hour food recalls and nutrient intake data for individuals and includes sampling weights, stratification variables, and sampling units to ensure a representative sample of the US population. A sub-sample of individuals has two days of dietary intake data, to which I restrict my sample[16], along with excluding individuals under the age of 18, giving a remaining sample size of 10,133 adults (4,868 in 2007-2008, 5,265 in 2009-2010). The first day's observation of food intake (more importantly, detailed nutrient intake) is recorded in an in-person interview, and the second via a follow-up telephone interview 4 to 11 days after the in-person interview.

[16] Since the majority of individuals are observed for both days, an average pattern of dietary choice can better be represented than those with only one day worth of data. While two observations per person is small, I know of no other data with such information as NHANES that would allow me to conduct this study in this way.

13.1.1 Diet Cost.

Since my analysis focuses on diet cost, a metric of food prices is required. For this, I choose the USDA's Center for Nutrition Policy & Promotion's (CNPP) national average price per 100g in dollars as the outcome variable because it makes the most economic sense as a measure for the average cost of a diet. These prices were calculated by CNPP researchers who used price data from the AC Nielsen Consumer Homescan Panel to calculate the nationally representative average prices[17] for the foods found in NHANES 2003-2004 (CNPP, 2008). Assuming that relative food prices did not meaningfully change from 2003 to 2010, I use these prices for foods consumed in the NHANES 2007-2010 data and update by filling in any prices missing from these years, which occurs whenever a food appeared in the 2007-2010 data but not 2003-2004. To do this, I exploit the USDA food coding system.[18] Taking the prices of individual foods and quantities consumed of those foods, I computed the average diet cost weighted by each food item's share of the individual's total grams as my outcome. The summary statistics of this variable are below in Table 30.

[17] This ignores branded vs. private label, organic vs. conventional and other secondary distinguishing factors of foods that may reflect different pricing mechanisms. The CNPP calculated these prices with the intent of describing what price a consumer in the US would face on average per 100g of that food item in 2003 and 2004 in those dollars. Full description of calculations can be found at https://www.cnpp.usda.gov/sites/default/files/usda_food_plans_cost_of_food/PricesDatabaseReport.pdf

[18] I first sort the price data by the USDA's 8-digit food code, and then for any foods with missing prices, I take the average of the immediately previous and next non-missing prices. This is because very similar foods are ordered sequentially. The first digit of the food code corresponds to which of the 9 major food categories the food belongs, the second and third digits and some fourth digits represent the increasingly granular subcategories, and the last four digits describe specific foods in numerical sequence. Further details and examples can be found at https://www.cdc.gov/nchs/tutorials/Dietary/SurveyOrientation/ResourceDietaryAnalysis/Info2.htm

Table 30: *Average Diet Cost ($/100g)*

Min	Max	Mean	StdDev	Median
0.04	0.82	0.19	0.07	0.18

13.1.2 Diet Quality.

A commonly used measure of an individual's compliance with the DGA is the Healthy Eating Index (HEI). The HEI is a density based metric used to score an individual's consumption of select food categories to emphasize (e.g. leafy green vegetables, dairy) and to limit (e.g. solid fats, added sugars), compared with total energy, with a higher overall score indicating a closer alignment to the Guidelines. The HEI is revised every 5 years to better measure an individuals' dietary quality (Guenther et al., 2008).

By using the HEI, my analysis would be confined to groupings of foods, and in the present context, the goal is to evaluate an overall basket of foods purchased, not the food groups from which they came. For this reason, a measure designed to evaluate the nutritional value of foods is necessary. Therefore, I turn to a nutrient profiling model developed by Arsenault et al. (2012), which was displayed on the horizontal axis of Figure 1. Their model computes a weighted sum of nutrients to calculate a nutrient density per 100 grams, given by 1.4*(protein) + 3.3*(fiber) + 1.0*(calcium) + 2.51*(unsaturated fat) + 0.37*(vitamin c) – 2.95*(saturated fat) – 1.34*(sodium) – 0.52*(added sugars). The coefficient for each nutrient was estimated as a linear regression using OLS. The dependent variable was the 2005 HEI, and each predictor was the intake value of each of the chosen nutrients consumed by an individual for whom the HEI was calculated. Their initially chosen set of nutrients was selected from those in the 2005 and 2010 Dietary Guidelines for Americans. The algorithm estimated the model for various combinations of nutrients until finding a model with the smallest number of nutrients for which variation in the

adjusted R^2 "reached a plateau". The adjusted R^2 for the final model was 0.65, meaning that variation in the final set of nutrients explains 65% of the variation in the HEI scores calculated for NHANES 2005-2008. The dot product of the estimated coefficient vector from the final model and the vector of nutrient levels for a given food computes a scalar value of the weighted nutrient density for that food item. Higher values of the computed index signal higher nutritional quality, thereby reflecting the "healthfulness" of any given food item in the data. The average of this index across the foods eaten over two days weighted by each food's share is an indicator of the "health" of dietary intake.

13.1.3 Palatability.

There is a gap in the food policy and consumer science research concerning diet palatability. I know of no research that has developed a plausible, non-subjective measure for the palatability of a diet. Drewnowski & Eichelsdoerfer (2010) cite a need for such work, because little or no research addresses whether or not Americans will actually eat the foods recommended in the Guidelines.

It is clear that a measure for diet palatability is desirable. I argue that creating such a measure is feasible when given the right data. Questionnaires in the 2007-2008 and 2009-2010 NHANES data include the following series of questions: "When you buy food from a grocery store or supermarket, how important is…?" (a) taste, and (b) nutrition. I make use of these to develop a measure of taste. Each of these is scored on a Likert-type scale[19] ranging from "not at all important" to "very important". While the nominal values of these variables are not useful ("very important" may not mean the same thing for two different respondents, since each choice is subject to perception), the difference of taste versus nutrition would speak to the degree of an

[19] The original scale of these variables is reversed to make higher scores indicate more importance.

individual's relative value placed on food taste: the higher the value of this difference, the greater the prioritization of taste over nutrition. The distribution of possible values of this variable is consistent with literature. That is, the average of the difference of the importance of taste and the importance of nutrition is significantly greater than 0 at a 1% significance level.[20] One such study finding evidence of this is Glanz et al. (1998) who found that Americans value food taste significantly more than their nutrition. A similar conclusion appears in Malone & Lusk (2017). Furthermore, a two-sample t-test between individuals classified as obese have slightly greater values of this variable on average (albeit insignificant) – echoing the findings of Hill et al. (2016). Figure 23 shows the distribution of the possible values of this variable.

[20] This is true for both 2007-2008 and 2009-2010 NHANES samples and for the two samples pooled. Of course, there is little reason to expect preferences to have changed much.

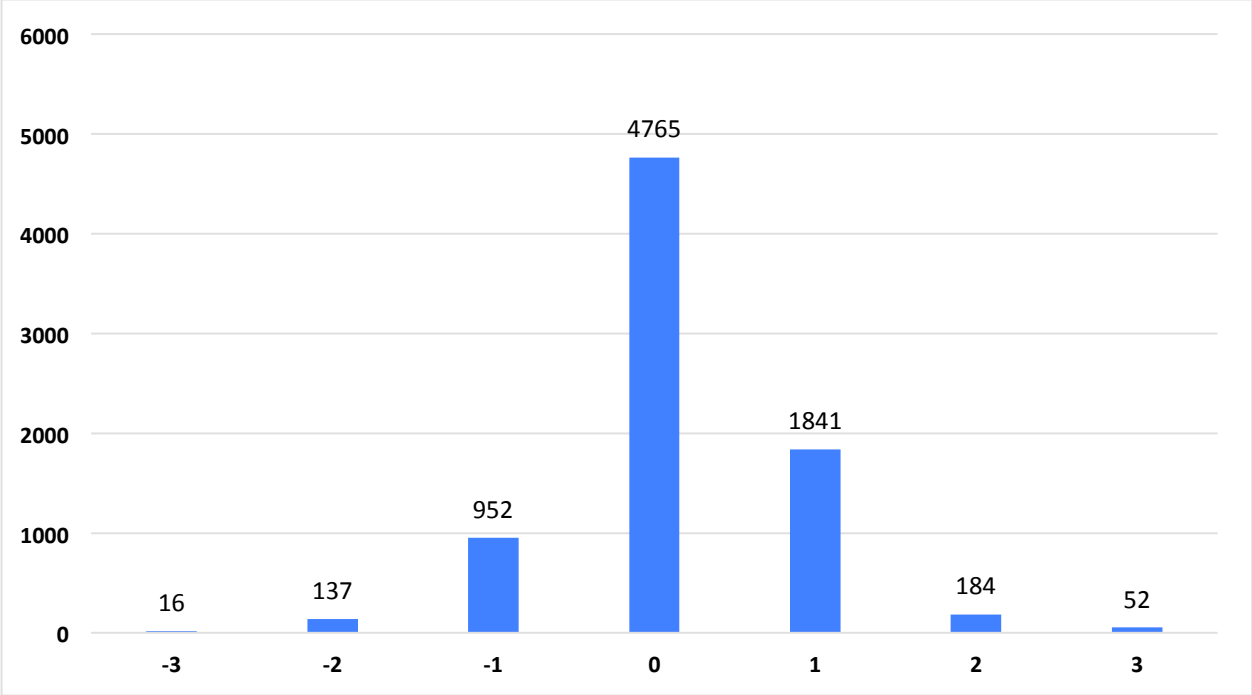


Figure 23: *Distribution of Importance of Taste relative to Importance of Nutrition*

It is reasonable to expect there to be some correlation between this variable and individuals' diet compositions within the NHANES data. There are obvious factors such as sugar, sodium, and saturated fat content, which can help to signal palatability (Mattes 1997, 2006; Glanz et al., 1998; Yeomans, 1998). Similarly, some literature implies that a reasonable proxy for palatability is calorie density (e.g. Drewnowski & Specter, 2004; Darmon et al., 2005; Drewnowski, 2010; Binkley & Golub, 2011). Conversely, foods high in vitamins, minerals, and fiber tend to be considered less tasty – typically grains and whole grain products as well as fresh fruits and vegetables (Drewnowski & Rock, 1995; Drewnowski, 1997; Glanz et al., 1998). One study by Adelaja et al. (1997) found that individuals who indicated that they were more concerned with how their food tastes tended to consume significantly more saturated fat in their diets. More recently, a survey by Morning Consult revealed that among the top 10 words that US consumers claimed to make groceries less appealing, the words "sugar free" and "fat free" made

the top 5 (2018), which implies that because the absence of these nutrients is perceived as less tasty, their inclusion is perceived as more tasty.

Borrowing from the procedure in Arsenault et al. (2012), I estimate a model of this measurement of an individual's stated preference for taste relative to nutrition. The initial candidate nutrients (selected from the above citations) are added sugars, sodium, saturated fat, dietary fiber, vitamin C, vitamin K, vitamin A, calcium, protein, caffeine, carbohydrates, poly- and monounsaturated fats, moisture, potassium, cholesterol, and natural sugars. The model for the i^{th} individual takes the form

$$(T - N)_i = \alpha_i + \boldsymbol{X}\beta_i + \epsilon_i \qquad (1)$$

where $(T - N)_i$ is the i^{th} individual's stated importance of tastiness minus their stated importance of healthiness, and $\boldsymbol{X}_i$ is a matrix of the i^{th} individual's intake of the selected nutrients across two days[21]. The interpretation of β_{ij} is the change in individual i's prioritization of palatability over nutrition given an increase in their intake of nutrient j per 100g of intake. Since some of the pre-selected candidate nutrients may matter more than others, and with the possibility that some excluded nutrients may be important predictors, the final combination of nutrients is chosen with stepwise regression, minimizing the AIC. This is similar to the method used by Arsenault et al. (2012) who maximized the adjusted R^2 to the point of minimal variation. The results of the procedure are listed below in Table 31, displaying the final set of nutrients, the estimated slopes, the t-statistic for each slope, the R^2, and the number of steps it took for the algorithm to minimize the AIC

[21] The initial matrix of nutrients, $\boldsymbol{X}$, contains the nutrients for which there are priors on their sign.

Table 31: *Nutrient Intakes as Predictors of Importance of Tastiness Relative to Healthiness*

Nutrient	**Estimate**	**t-statistic**
Intercept	8.4330	4.17
Added Sugars	2.5009	2.78
Sodium	0.0895	2.30
Saturated Fat	8.8021	1.91
Moisture	-8.3194	-4.10
Fiber	-14.7172	-2.64
Vitamin C	-1.2532	-2.57
Vitamin K	-0.2749	-1.68
Monounsaturated Fat	-9.3464	-1.64
Protein	-12.2842	-4.96
Caffeine	0.9745	4.16
Carbohydrates	-8.1290	-3.73
Magnesium	-2.0561	-4.10
$\boldsymbol{R^2}$	0.06	
Number of Steps	7	

Unsurprisingly, nutrients generally associated with taste-enhancement (sodium, saturated fat, and added sugars) are positively and significantly correlated with an individual increasingly valuing taste over nutrition. Moreover, the nutrients associated with grains and vegetables (fiber, vitamin C, vitamin K, monounsaturated fat) are negatively and significantly correlated with the dependent variable. While the R^2 is only 0.06, the variation in the dependent variable is also small, taking on only 7 possible values as indicated in Figure 3.

In the same fashion as Arsenault et al. (2012), the dot product of the estimated $\hat{\beta}$ shown in Table 4 and the matrix of each food item's content on the nutrients listed in Table 4 computes the weighted nutrient density of food items. The key difference between this calculation and that in Arsenault et al. (2012) is that these estimated nutrient weights reflect consumer's stated preferences of taste versus nutrition, and the calculated index is a measure of the tastiness of food items. This product henceforth serves as a measure of palatability for each food in

NHANES 2007-2010. Table 32 shows some example foods at different magnitudes of the tastiness index, and similarly for the weighted nutrient density score from Arsenault et al. (2012)

Table 32: *Examples of Palatability & Nutritional Quality (NHANES 2007-2010)*

Quantile	**Examples of Nutritional Quality**	**Examples of Palatability**
2 Std. Dev. Below Mean	*--Coffee Mocha w/Whole Milk,* *--Soy Sauce,* *--Energy Drink*	*--Raisin Bran Cereal,* *--Dried Shrimp,* *--Papaya*
1 Std. Dev. Below Mean	*--Smoked Sausage,* *-- Fruit Juice Drink,* *--Low-Fat Soft-Serve Ice Cream*	*--Dry Cowpeas,* *--Cooked Spinach,* *--Bell Pepper*
Mean	*--Lasagna w/Spinach,* *--Turkey & Vegetables w/Cheese,* *--Mashed Potatoes*	*--Fried Catfish,* *--Marinara Sauce,* *--Energy Drink*
1 Std. Dev. Above Mean	*--Raw Tomato,* *--Turnip Greens,* *--Kiwifruit*	*--Smoked Sausage,* *--Ham & Biscuit,* *--Devil's Food Cake*
2 Std. Dev. Above Mean	*--Watercress,* *--All-Bran w/Extra Fiber,* *--Bell Pepper*	*--Bacon,* *--Milk Chocolate Candy,* *--Whipped Cream*

It is worth noting that some foods appearing in the "low-taste" space appear in the "high-health" space, and vice versa (e.g. bell peppers versus smoked sausage). The values of the indices have no meaningful units and are not displayed, although higher values of either index signal tastier or healthier foods. The set of foods high in both nutritional quality and palatability is much smaller than any of the sets described in Table 5. Some examples include cooked

mushrooms, milk, okra, and salsa. Foods measured as neither tasty nor healthy include seaweed in soy sauce, frozen unsweetened rhubarb, and candied sweet potatoes. In the same fashion as before when computing the nutritional quality of diets, I compute the weighted average of the tastiness index for each individual's diets to measure diet palatability.

13.1.4 Diet Variety.

Earlier I inferred from my theory that as a consumer's food budget expands, so does the set of feasible options for consumption. The expanded set would then include foods not previously considered for consumption due to their higher cost, since the initial set of choices would tend to emphasize lower cost foods. Hence, I include variety, which I expect to have a positive effect on diet cost.

There are multiple measures of diet variety: the simplest measure would be the number of foods eaten by an individual, but this would not speak to the shape of the distribution of that individual's consumption. Another measure is Simpson's index. This is computed as the sum of the squared diet shares of each food item (the Herfindahl Index of concentration) subtracted from 1. Alternatively, a different measure of variety is entropy. This is computed by the negative sum of the logged shares across all foods consumed, and varies between 0 and the natural logarithm of the number of foods consumed. This would place a higher weighting on foods that contribute relatively little to the overall diet because of the mathematical properties of the natural logarithm. I use an alternative measure of variety proposed in Jekanowski & Binkley (2000). This measure is computed as the number of foods necessary for a consumer to achieve 75% of their total grams of intake, which can be done by computing the diet share of each food for a consumer in descending order and taking the cumulative sum until it reaches 0.75. The ordinal food for which that cumulative total hits 75% is the measure of variety for that person. Measuring this variable

in this fashion assigns variety as a feature of a basket of food items in which there is a relatively small number of dominant members, and places less emphasis on the foods contributing less to the cumulative diet shares. The summary statistics of this variable are in Table 33.

Table 33: *Diet Variety*

Min	Max	Mean	StdDev	Median
1.00	49.00	12.94	5.57	12.00

The average adult in NHANES 2007-2010 ate about 13 unique USDA-coded foods to achieve 75% of their total intake over the two days for which I have data on them. In my econometric model, this dietary diversity measure will not be interacted with nutritional quality or palatability since there is no theoretical justification for doing so, although final robustness checks of the model can verify the relative importance of such interactions.

13.2 Models: Choosing Healthiness, Tastiness, and Costliness of Diets

With measures of diet cost, tastiness, variety, and healthiness chosen, I now proceed to estimate an econometric model in search of evidence either in favor of or contrary to the first theory. That is, a low-income consumer is more likely to forego adhering to the Guidelines than a high-income consumer because a healthy and tasty diet is costly, while a healthy diet of lesser palatability or variety is not.

The econometric model estimates the average cost of a diet per 100g in dollars as a function of that diet's healthiness, its palatability, and its variety. For the i^{th} individual, the model takes the form

$$Cost_i = \beta_0 + \beta_1 H_i + \beta_2 T_i + \beta_3 V_i + \beta_4 H_i T_i + \beta_5 D_{2007-2008} + \epsilon_i. \qquad (3)$$

H_i is Arsenault index for individual i, T_i is their tastiness index, V_i is diet variety, and $D_{2007-2008}$ is a year fixed effect. The variables H, T, and V are normalized to have mean zero and unit standard deviation, but the outcome variable is left unstandardized because its units are more meaningful.

I would reject the assertion that healthier eating is more expensive, if the result shows $\beta_1 \leq 0$, since a significantly positive effect of nutritional quality on diet cost would support this claim. Otherwise, I reject the null hypothesis. Furthermore, if the model supports the utility maximization theory developed above, then the coefficient on the interaction $H * T$ will be positive and significant. This is because the new choices being considered were not previously considered when the food budget expands because of their high cost. As indicated above, the coefficient on variety, β_3 should be non-negative. While the inclusions of T and $D_{2007-2008}$ are necessary for the model, any priors on the signs of β_2 or β_5 are not immediately clear.

13.3 Results

I estimate the model using OLS with the NHANES survey design[22]. The estimation results are displayed below in Table 34.

Table 34: *OLS estimation of Equation (3)*

Variable	Estimate	t-statistic
Intercept (β_0)	0.1962	73.14
H (β_1)	-0.0323	-36.97
T (β_2)	-0.0085	-5.37
V (β_3)	0.0110	9.51
H*T (β_4)	0.0060	5.27
$D_{2007-2008}$ (β_5)	-0.0065	-1.97
$\boldsymbol{R^2}$	0.21	

Overall, the model supports the theory. The univariate effect of nutritional quality, β_1, is not positive, and the slope of the interaction H*T is positive and significant. Specifically, a standard deviation change in H is accompanied by a -0.03 standard deviation change in diet cost, while a standard deviation change in the interaction of H and T coincides with a 0.006 standard deviation change in cost. As expected, the sign on variety is also positive. While no priors were given for β_2 or β_5, both are significantly negative. The marginal effect of healthiness on cost, $\beta_1 + \beta_4 T$ is of particular interest. Figure 24 shows this marginal as a linear function of tastiness with 95%

[22] This includes sampling weights for each individual in NHANES as well as the surveying hierarchy with clustering and stratification variables. All are included in the NHANES demographics data. A thorough description of the NHANES sampling methods and survey design can be found at https://www.cdc.gov/nchs/data/series/sr_02/sr02_160.pdf.

confidence intervals, and similarly Figure 25 depicts the marginal effect of tastiness as a linear function of healthiness with 95% confidence intervals

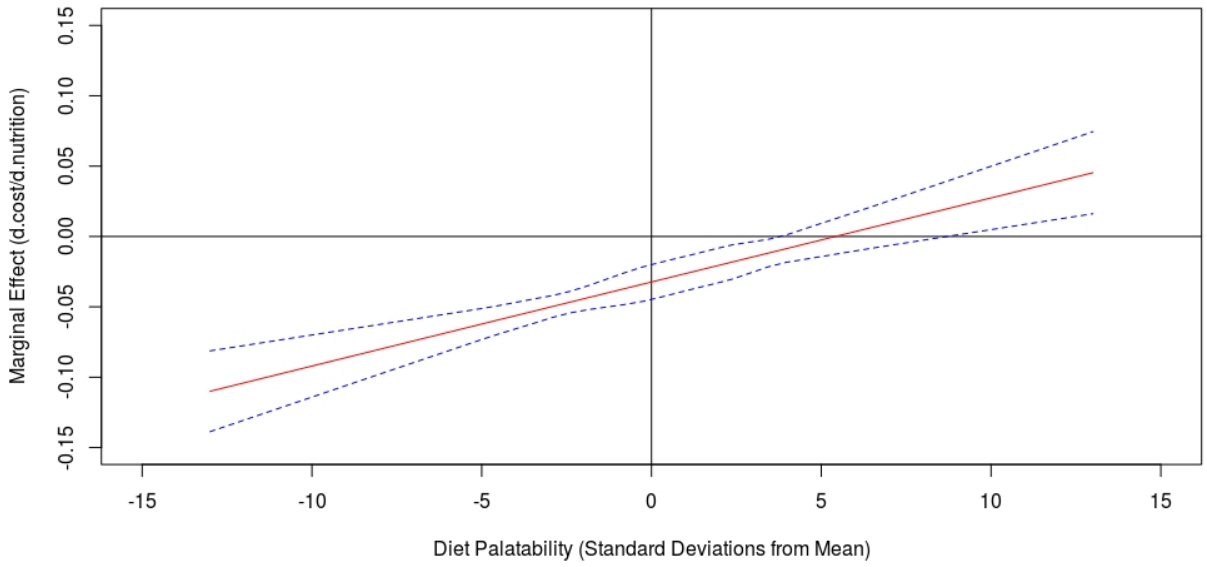


Figure 24: *Marginal Effect of Healthfulness on Cost*

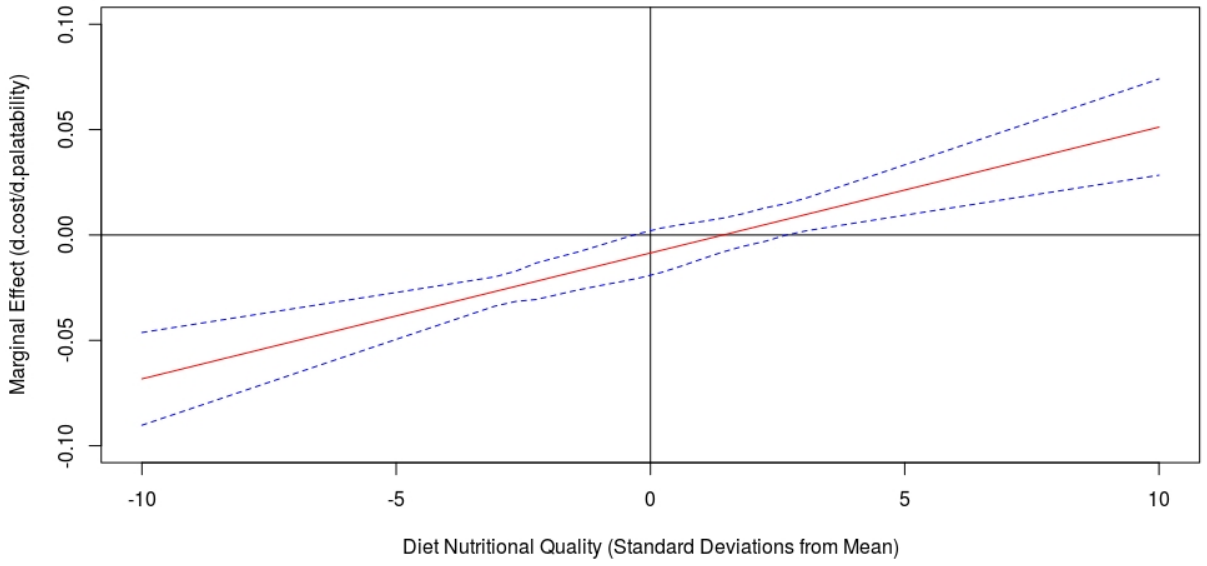


Figure 25: *Marginal Effect of Tastiness on Cost*

The nutritional quality marginal does not cross the horizontal axis until tastiness is 6 standard deviations above the mean, and is statistically significantly greater than zero when tastiness is 11 standard deviations above the mean. The palatability marginal crosses the horizontal axis and becomes positive when healthiness is 2 standard deviations above the mean, and is significantly greater than zero when healthiness is 4 standard deviations above the mean. These findings offer strong evidence to reject the null hypothesis that healthier eating comes at a higher cost. Conversely, these findings support the theory: when an individual imposes a constraint on their diet's nutritional quality, their set of foods either remains unchanged or expands to consider other options in order to satisfy the new constraint while maintaining a sufficiently high level of palatability. If the set expands to include other foods, then the food budget must also increase, since the new food choices were not previously considered due to their high cost. Therefore, because this is less of an obstacle to a high-income consumer, those with less income tend have more of an incentive not to adopt a healthy diet.

I have four final comments about the findings in this section. First, I also estimated a model similar to Equation (3) using Generalized Propensity Score modeling following Hirano & Imbens (2004), Imai & van Dyk (2004) and Eggers & Von Ehrlich (2013). The point estimates differ little from those in Table 7, and the estimated outcomes do not differ qualitatively. I still reject the null hypothesis that healthier diets tend to cost more, and I still find that simultaneously increased levels of healthiness and tastiness tend to come at a higher cost. Second, I included interaction terms of variety and healthiness as well as variety and tastiness, and the results did not differ. The interaction of variety and healthiness has a negative and significant estimate, while variety and tastiness is negative and insignificant. Third, I estimated the same model in Equation (3), but with household income included. The point estimates did not change

much and the statistical significance of each estimate increased slightly, as did the R^2 of the model. Fourth, I changed the measure of healthiness from the weighted average of the Arsenault et al. (2012) index to each person's HEI. While the resulting estimates were less clear[23], the conclusion partially remained intact that a healthier diet does not come at a significantly higher cost.

4.4 Descriptive Statistics: Subclasses of Diets

Below in Figure 26, I show the variation in diet cost between individuals at the top and bottom quartiles of the distribution of nutritional quality.

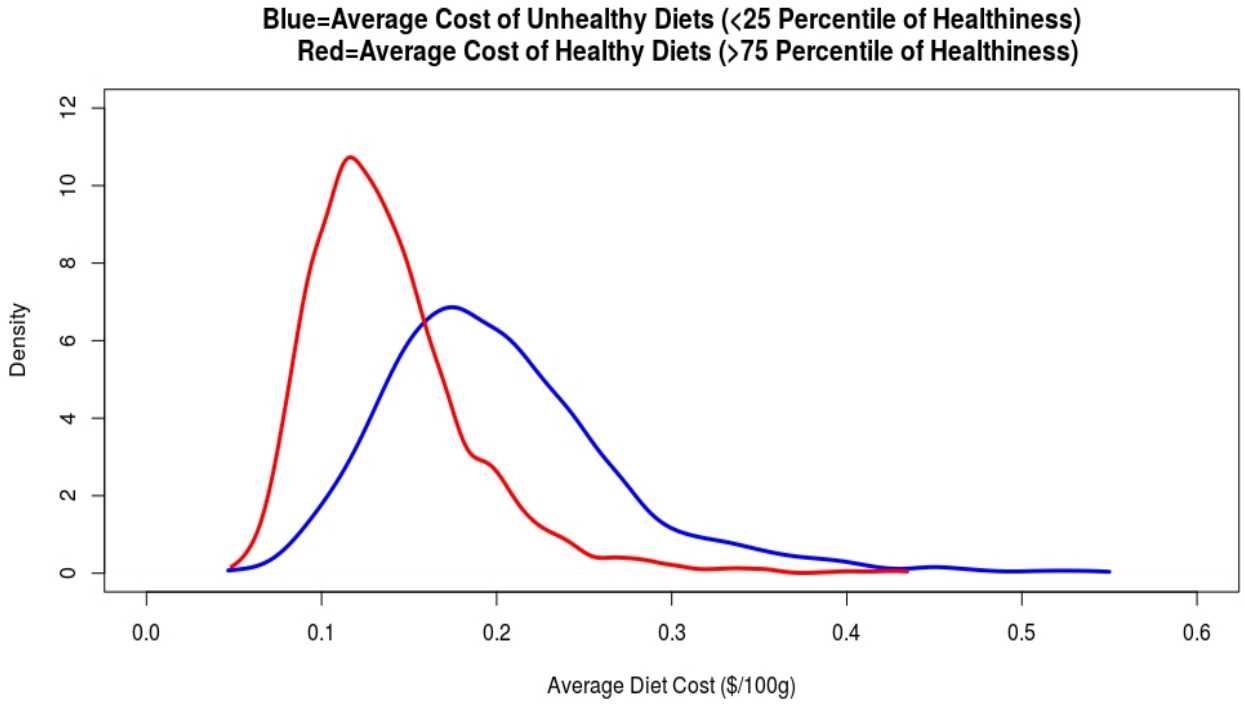


Figure 26: *Distributions of Diet Costs for Healthiest, Unhealthiest Quartiles*

[23] For example, the sign on HEI interacted with diet palatability is the opposite of that on the interaction with the Arsenault et al. (2012) index for diets, and the R^2 of the model was only slightly above zero.

The red curve is the distribution of the average diet cost of adults in NHANES 2007-2010 whose diet healthiness is at or above the 75th percentile of the distribution of diet quality, and the blue curve is the same for those whose diet healthiness is at or below the 25th percentile. Both subsamples consist of 2,531 individuals. There are several things worth noting in the figure. First, the overlap of the two distributions is far greater than would be expected if healthier eating is more expensive: just over 91% of the observations in the healthy sample have lower costs than the average cost of the unhealthy sample. Second, the average cost of the healthy sample is \$0.14/100g, and the average cost in the unhealthy sample is \$0.20/100g. A t-test for a difference in mean diet cost reject with a t-statistic of -37.75 and an associated probability value of 0. In light of these findings, it is very difficult to argue that high cost is the principal barrier to maintaining a healthy diet.

To illustrate further the output of the model results in Table 5, I explore the tail behavior of the red and blue curves by calculating and comparing descriptive statistics of the lowest quartile and highest quartile of each curve, that is, in each case the cheapest 25% of diets and costliest 25%. Each of the subgroups contains 633 individuals. The descriptive statistics and their concomitant t-tests are presented below in Table 35

Table 35: *Descriptive Statistics Underlying Figure 4*

BLUE (unhealthy)	**Mean of Bottom Quartile**	**Mean of Top Quartile**	**t-statistic**		**RED (healthy)**	**Mean of Bottom Quartile**	**Mean of Top Quartile**	**t-statistic**
Cost	0.12	0.29	66.32		*Cost*	0.09	0.20	61.01
Tastiness	0.07	-0.05	-2.03		*Tastiness*	-0.11	0.06	2.86
Number of Foods	26.70	30.37	6.57		*Number of Foods*	28.74	33.58	8.69
Variety	9.36	12.59	12.16		*Variety*	10.02	13.98	13.38

For the subsample of individuals with healthier diets, it is clear that the diets of higher cost have significantly higher palatability and more foods. On average, a costly healthy diet

contains about 5 more foods overall and requires 4 more foods to achieve 75% of total intake. Furthermore, the foods contained in these diets differ from those contained in the low-cost healthy diets, as well. Some examples of foods for which the expensive healthy diets contain significantly more grams are fish, pasta, chicken, pork, and eggs. Conversely, the cheap healthy diets contain significantly more grams of beans, cabbage, rice, carrots, and potatoes. For the subsample of less healthy diets, similar patterns emerge. Like the regression model, this exercise does not support the affordability axiom, and shows that a healthier diet, if anything, costs less on average.

CHAPTER 14. THEORY 2 METHODS

14.1 Foundations

I now turn to the second possible explanation for the dietary deficiencies in low-income Americans. That is, the tendency of these consumers to place a lower value on longevity, and thus health as a means to obtaining it. Their lower expected future income lowers expected future utility, making them less willing to sacrifice present pleasure for the chance of longer life. The explanation is that presented in Binkley (2010), and later validated empirically in Binkley & Golub (2011) and Chen et al. (2012). The first study dealt with smoking, the second two with food consumption.

The evidence presented in these studies supports the explanation of lower expected utility driving low-income consumers' less healthy dietary choices. However, a gap remains in that access is not controlled in the two food-focused studies. If consumers with lower income have reduced access to healthier types of food, then this could inflate the correlation between income and healthy food consumption, thus compromising these studies. Addressing this gap is important and is a more rigorous way of testing for evidence supporting or refuting the second theory.

There is much research attributing the link between income and diet quality to the problem of access. The term "food desert" has gained popularity as a description of regions or neighborhoods without sources of fresh produce, and the presence of fast food restaurants is widespread. These areas also tend to have a higher concentration of convenience stores, which tend to feature limited, higher-priced selections of healthy foods (Moore & Diez Roux, 2006). Studies contend that the lack of access to and availability of fresh produce and low-calorie, low-fat varieties of foods force the poor who live in food deserts to buy only what is available, and,

hence, their dietary quality suffers as a result (Zenk et al., 2005; Moore & Diez Roux, 2006; Hilmers et al., 2012). Similar studies argue that the prevalence of outlets for food away from home in these limited access areas entice the residents of these areas to eat out more, which tends to increase a consumer's caloric intake since food eaten away from home tends to be more energy-dense than food prepared at home (Jeffery & French, 1998; Stewart et al., 2004; Thompson et al., 2004; Creel et al., 2008; Hilmers et al., 2012; Laska et al., 2015).

As stated above, the study reported here builds on Binkley & Golub (2011) and Chen et al. (2012). In light of the above discussion, I extend their analyses by controlling for access. Chen et al. (2012) could not control for access since their analysis used data from NHANES, which has no information on buying options available to respondents. Binkley & Gollub (2011) used aggregate market level data. The commodity studied is nonorganic fluid milk, the same as that in Chen et al. (2012) and one of the commodities included in Binkley & Gollub (2011). Milk is an ideal commodity to study. It is a homogeneous commodity that is widely available with fixed container sizes. It is also differentiated primarily by fat content, which makes it very easy to determine the "healthiness" of the various choices. For all choices, there is no associated time cost of preparation or other differences in convenience. Further, milk of different fat content is typically sold for the same price, and when this is not the case, price tends to increase with fat content.[24] Also, milk is often featured in discussions of access. There is much research examining the access and availability of low-fat milk, some finding evidence of limitations (e.g. Cheadle et al., 1990; 1991; Glanz et al., 2007) and some finding no issues with access (e.g. Hosler et al., 2006; Liese et al., 2007).

[24] I verify this with simple statistics in Table 10 on the next page.

14.2 Data

The data I use for this model is the AC Nielsen Consumer Homescan Panel, obtained from the University of Chicago Booth School of Business' Kilts Marketing Center. The data cover the grocery purchases of more than 40,000 households in the US. It includes prices and purchasing sources of more than three million UPC-coded foods as well as household demographic information. Nielsen markets – of which there are 52 primary – consist of a cluster of counties in and around a major metropolitan area. The samples in the markets are representative of the region, and the data includes a projection factor to make any subsamples representative of the US population.

14.2.1 Low-Fat Milk Purchasing.

The outcome variable is the quantity share of low fat milk (skim and 1%) of a household's total gallons of milk purchased during the year 2010. A key difference between low fat and high fat milk is taste, since the primary distinction between nutrient content is saturated fat. Therefore, buying low fat milk involves sacrificing taste in order to gain health by avoiding excess saturated fat. The data allow me to model of this variable as a linear function of household income and demographic variables. Further, because these data record the purchasing sources of milk, imposing restrictions on which data are used allows for controlling access. What is required is that the restricted data pertain to purchases made under conditions where the household could have bought milk of any fat content. This is done by limiting which stores appear as the source in the set of milk purchases made by households – specifically, store chains that offer all types of milk in a given market area. To accomplish this, only stores for which the average price per gallon for all four fat content levels was not missing are included. This was

further restricted for a store chain to have sold at least 100 gallons during the year in a given market.

14.2.2 Household Income.

The variable of interest is annual household income. Nielsen measures annual household income by category, in $10,000 to $25,000 increments. This variable was rescaled to the midpoint of the income bracket, making the interpretation of a unit increase in household income to be a $1000/year change rather than a discrete, categorical change. The models from Binkley & Golub (2011) and Chen et al. (2012) that I follow tested the effect of income on the purchasing share of low fat milk. I estimate a slight variant of this model using data constructed in such a way to control for access.

Income correlates positively with the dependent variable, the summary statistics for which are shown in Table 36.

Table 36: *Low Fat Milk's Average Share of Total Milk Purchases by Income Range*

Income Bracket	Mean	StdDev
<$20k	0.31	0.43
>$20k to <$50k	0.36	0.45
>$50k to <$100k	0.44	0.46
>$100k	0.53	0.46

The average share of low fat milk of total household milk purchasing is 31% for households whose annual income is under $20,000. For households with annual income over $100,000, 53% of their total milk purchasing is low fat milk.

14.2.3 Milk Prices.

In order to control for cost, the average prices for 2010 in dollars per fluid ounce for gallons of skim, whole, 1%, and 2% milk across each store chain for every market was calculated. Doing so allowed the subsetting of the data to include only the store chains for which the price of each type of milk was not missing. Hence, if a consumer purchased milk from that store and that purchase was recorded in the data, then access to both high fat and low fat milk was not an issue. With the share of low-fat milk as the outcome variable, the coefficient on the price of low-fat milk should be negative and that on high-fat milk should be positive (since these calculations are for shelf prices, not prices paid by consumers including coupons, sales, etc.). I also restrict the data to be purchases of gallons only, which is the most commonly purchased size of milk. In Table 37, it is clear that the average price per gallon of milk increases by fat content.

Table 37: *Milk Prices by Fat Content*

Fat Content	Min	Max	Mean	StdDev	Median
0% (Skim)	$0.42	$5.01	$2.08	$0.41	$2.04
1%	$0.75	$8.04	$2.14	$0.51	$2.11
2%	$0.87	$6.04	$2.14	$0.42	$2.11
3.5% (Whole)	$0.82	$4.92	$2.21	$0.44	$2.19

14.2.4 Demographics.

I also include demographic variables that may be correlated with income as well as a household's share of low-fat milk purchasing. Race indicators are in the model to help capture differences in milk purchasing by cultural norms and habits, and also because of their correlation with household income. Age and presence of children indicators for infants, youth, and teens are included because consumption of high fat milk tends to be higher for young children. For similar reasons, household size is also included. Educational attainment indicators for highschool and college graduate are included because of the implied increase in knowledge about health and

nutrition. Marital status and age bracket (young adult, middle-aged, senior) dummy variables of the head of household are also included, since these tend to correlate with dietary quality (Ervin, 2011). Indicator variables for previous or current WIC participation are in the model as well. Aside from their obvious correlation with income, and in addition to milk being in the food package subsidized by WIC, one reason to include the variables is that WIC participants are required to attend nutrition education courses in order to continue receiving benefits. This could skew the household's purchasing towards low fat milk, at least in the short run. Total ounces of breakfast cereal purchased in 2010 is also included because the taste of milk is less likely to be important when combined with cereal than it is when drinking milk by itself. Finally, market fixed effects are in the model to help capture geographic differences in milk purchasing.

14.3 Model

To increase sample validity, I exclude any households who bought less than 10 gallons in 2010, making the final data a sample of 27,437 households in 52 market areas across the US. I estimate the model using sampling weights provided by Nielsen for each household. These sampling weights ensure a representative sample of the US population, since the sample in each Nielsen market is not random.

For household i, the model takes the form

$$Y_i = \beta_0 + \beta_1 Inc_i + \boldsymbol{X}_i\delta + \epsilon_i. \qquad (4)$$

Y_i is household i's share of low-fat milk, Inc_i is their annual household income, and $\boldsymbol{X}_i$ is a vector of the other variables. Since, on average, low fat milk is the cheaper option, one could argue that this estimate should be negative: low fat milk is at worst more affordable, and thus should be more attractive to low income consumers.

5.4 Results

In Table 38 are the results of estimating Equation (4) with OLS

Table 38: *Estimation Results*

Variable	**Estimate**	**t-statistic**
Intercept	0.3191	12.88
Household Income	0.0011	16.28
College Graduate	0.1596	21.79
Highschool Graduate, No College	0.0379	6.45
Married	0.0417	6.69
Young Adult	-0.0192	-2.85
Middle-Aged	-0.0128	-1.84
Child Age < 2 present	-0.0212	-2.55
Child Age 3 to 12 present	0.0421	6.91
Child Age 13 to 17 present	0.0223	2.98
Black	-0.1615	-15.63
Hispanic	-0.0599	-7.28
East US	0.0160	0.93
South US	-0.1375	-11.92
West US	-0.1196	-8.21
Price of Low-Fat Milk	-0.0673	-4.17
Price of High-Fat Milk	0.0977	5.85
Ounces of Cereal Purchased	0.0001	13.61
Previously Enrolled in WIC	-0.0455	-5.95
Currently Enrolled in WIC	0.0067	0.38
$\mathbf{R^2}$	0.12	

From this estimation, it is clear that even when controlling for access, the effect of income is positive and very significant. A $1,000 per year increase in household income coincides with a 0.11% increase in the share of low-fat milk purchased by a household. In Table 10, a discrete jump in household income from <$20,000 to between $20,000 and $50,000 coincided with a 5% increase in the share, which is comparable. The signs on the prices correctly reflect the law of demand: when low fat milk becomes more expensive, holding the price of high fat milk constant, consumers buy relatively less low fat milk. The opposite is true for the price of high fat milk.

Households with middle aged and young heads purchase less low fat milk than the omitted group, which is senior-headed households, which is consistent with previous findings that dietary behavior is generally better among the elderly (Hann et al., 2001; Ervin, 2011). The estimate on cereal indicates that for each additional ounce of breakfast cereal purchased in 2010 by a household, the share of low fat milk purchasing increases by 0.0001. Thus, for each additional 15-ounce box of cereal purchased per year by a household, the milk-purchasing share of low fat milk increases by 0.15%, which is consistent with the assertion that milk taste is less important when milk is consumed with cereal than when consumed by itself. The effects of the education variables are consistent with prior expectations. The negative slope on the race indicator for a black household may reflect lower expected future prospects and reduced longevity. Therefore, like in the main argument with regards to the effect of income by itself, less value is placed on health as a means to longevity. Current WIC participation coincides with non-decreased purchasing of low fat milk, while previous enrollment is negatively correlated with low fat milk purchasing. One possible explanation for this is that, while enrolled in WIC, some households did shift their milk purchasing towards low fat, but after enrollment ceased and benefits discontinued, once-enrolled households reverted back to purchasing milk with higher fat content.

I have five comments with regards to this analysis. First, I estimated with only married households and only single households. The results did not differ, particularly with regards to the role of income. Second, the data is extensive enough to run the model using only data from one chain operating in many markets. There are several such chains. This restriction should control access even more. When this was done, the results differ little from those in Table 11. In particular, the coefficient on income is always positive and significant. Third, Equation (4) was estimated using a censored regression. This is because the dependent variable is left-censored at

0 (no low fat milk purchasing) and right-censored at 1 (purchasing is exclusively low fat). The marginal effect of income at the mean was slightly larger than that estimated with OLS. Fourth, the dependent variable was transformed into a categorical variable where 0 indicates no low fat milk purchasing, 1 indicates a share between 0% and 100%, and 2 indicates exclusive purchasing of low fat milk. Then, Equation (4) was estimated using an ordered logit regression. The results are consistent with everything previously presented: a change in income produces a significant increase in the odds ratio of increasing categories of low fat milk purchasing. The fifth and final comment is that the results of this model hold even when the restriction criterion for stores selling all types of milk changes. In the present analysis, each store's prices for all four types had to be non-missing. Another way to do this is to choose which type had the fewest sales of the four and restrict the number of sales to be greater than some arbitrary threshold. The results for the estimate on income are robust to this alternative as well.

The results in Table 11 strongly support the second theory, as do the follow-up analyses. It is clear that in the case of one commodity for which there are no meaningful differences in cost or access but there are differences in taste and nutritional quality, low-income consumers tend to purchase more of the tastier, less healthy version. This outcome is robust to estimation procedure as well as any subsampling or restrictions on the data.

CHAPTER 15. CONCLUSIONS

The purpose of this essay has been to investigate the role of food cost in the prevalence of the dietary deficiencies of low-income Americans. There is a body of literature claiming that healthier diets are more expensive, and that is why we observe this problem. However, there is also a growing body of literature showing that this healthiness-costliness link may be spurious.

In this study, I offered two alternative explanations for why low income can lead to less healthy diets. The first is that lower income limits the choices an individual faces. For a given food budget and level of diet palatability, increasing the nutritional quality of the same diet without decreasing its palatability requires increasing the food budget, since new candidate foods were not previously considered for consumption because of their high cost. This is less of an issue for a wealthier consumer who can more easily afford to buy such foods. For the individual with lower income, increasing the food budget is much more difficult. Thus, the incentive to compromise nutritional quality in order to maintain sufficient palatability and low cost is stronger. The second explanation follows from the point that increasing palatability brings immediate utility and improving nutrition does not (utility of healthy eating is realized in the future, primarily through longer expected life). But the utility of longevity increases with expected income, and is therefore less for individuals with lower income. As a result, they are less willing to sacrifice the immediate pleasure of tasty eating for the future payoff of healthy eating.

At the outset, this study showed that there is no meaningful association between healthiness and costliness in the case of individual foods. This holds for the case of all options being available, and there may be a small number of special cases with additional obstacles or constraints for which consuming healthier foods may be difficult. Second, this study offered

empirical evidence rejecting the null hypothesis that healthier diets cost more. While the evidence was sufficient to reject the null, it was also strongly supportive of the first alternative explanation. Simultaneously increasing palatability and nutritional quality of a diet tends to come at significantly higher cost. This study also validated the second alternative explanation by showing that increased income is linked to a significant reduction of high fat milk purchasing, even though it is more costly than and equally accessible as the healthier alternative.

The role of cost in the dietary deficiencies of low-income Americans is negligible in the sense of healthier diets costing more. After showing this to be the case, I proposed and validated two alternative explanations for this widespread problem. Since cost is not the main barrier to healthy eating, perhaps emphasis should be placed on policies and programs whose aims are to improve nutritional knowledge and awareness. The food palatability measure developed in this study is the first of its kind, and undoubtedly leaves room for improvement, but it does at least appear to successfully capture forms of palatability in a general sense (see Drewnowski, 1997; Mattes 1997, 2006; Yeomans, 1998). There is potential future work in possibly improving this measure and applying it to more questions related to consumer behavior and diet choice.

This study is not without its limitations, however. In addressing the first theory, the use of data from 24-hour food recalls implicitly assumes that measurement error due to misreporting[25] is negligible, and that two days' worth of dietary intake are representative of an individual's food habits. While these were not tested assumptions in the context of this study, NHANES is used because of its ease of use and public availability, as well as its large sample size and detailed dietary intake information facilitating the analysis herein. The choice of measure for dietary quality (Arsenault et al., 2012) is useful in detailing the "health" of individual food items, unlike

[25] For instance, misreporting could include underreporting less healthy foods, overreporting more healthy foods as detailed in Pryer et al. (1997) and Lafay et al. (2000)

the HEI. When aggregated up to the diet level, the result is highly correlated with the HEI. A potential gap in using this measure to describe diet health is that, with a small number of nutrients considered in its calculation, it is conceivable that a diet consisting of foods with top individual ratings could have a lower HEI than a diet consisting of a balanced intake of all critical vitamins and minerals in a 24-hour period. The palatability index developed in this study likely has room to improve, having little external validation and no methodological precedent other than the method set forth by Arsenault et al. (2012) to calculate a similar measure. The milk model at the end of the study should be applied to other products whose varieties have clear health implications. These further applications would lend further validity to the theory being tested in the model, in addition to the single commodity case given in this study.

REFERENCES

Adelaja, Adesoji O., Rodolfo M. Nayga, and Tara C. Lauderbach. 1997. "Income and racial differentials in selected nutrient intakes." *American Journal of Agricultural Economics* 79(5): 1452-1460.

Anding, Jenna D., Richard R. Suminski, and Linda Boss. 2001. "Dietary intake, body mass index, exercise, and alcohol: are college women following the dietary guidelines for Americans?" *Journal of American College Health* 49(4): 167-171.

Arsenault, J. E., Fulgoni III, V. L., Hersey, J. C., & Muth, M. K. 2012. "A novel approach to selecting and weighting nutrients for nutrient profiling of foods and diets." *Journal of the Academy of Nutrition and Dietetics* 112(12): 1968-1975.

Becker, Gary S., and Kevin M. Murphy. 1988. "A theory of rational addiction." *Journal of political Economy* 96(4): 675-700.

Binkley, James K., Jim Eales, and Mark Jekanowski. 2000. "The relation between dietary change and rising US obesity." *International journal of obesity* 24(8): 1032.

Binkley, James. 2010. "Low income and poor health choices: the example of smoking." *American Journal of Agricultural Economics* 92(4): 972-984.

Binkley, James K., and Alla Golub. 2011. "Consumer demand for nutrition versus taste in four major food categories." *Agricultural Economics* 42(1): 65-74.

Blaylock, James, David Smallwood, Kathleen Kassel, Jay Variyam, and Lorna Aldrich. 1999. "Economics, food choices, and nutrition." *Food Policy* 24(2): 269-286.

Carlson, Andrea, and Elizabeth Frazão. 2012. "Are healthy foods really more expensive? It depends on how you measure the price."

Center for Disease Prevention and Control. 2017. "Overweight and Obesity" https://www.cdc.gov/obesity/

Center for Disease Prevention and Control. 2018. "National Health and Nutrition Examination Survey" https://wwwn.cdc.gov/nchs/nhanes/default.aspx

Cheadle, A., Psaty, B., Wagner, E., Diehr, P., Koepsell, T., Curry, S. & Von Korff, M. 1990. Evaluating community-based nutrition programs: assessing the reliability of a survey of grocery store product displays. *American Journal of Public Health* 80(6): 709-711.

Cheadle, A., Psaty, B. M., Curry, S., Wagner, E., Diehr, P., Koepsell, T., & Kristal, A. 1991. Community-level comparisons between the grocery store environment and individual dietary practices. *Preventive medicine* 20(2): 250-261.

Darmon, Nicole, Michel Darmon, Matthieu Maillot, and Adam Drewnowski. 2005. "A nutrient density standard for vegetables and fruits: nutrients per calorie and nutrients per unit cost." *Journal of the American Dietetic Association* 105(12): 1881-1887.

Davis, George C., and Andrea Carlson. 2015. "The inverse relationship between food price and energy density: is it spurious?" *Public health nutrition* 18(6): 1091-1097.

DiNicolantonio, James J., Sean C. Lucan, and James H. O'Keefe. 2016. "The evidence for saturated fat and for sugar related to coronary heart disease." *Progress in cardiovascular diseases* 58(5): 464-472.

Drewnowski, Adam, and Cheryl L. Rock. 1995. "The influence of genetic taste markers on food acceptance." 506-511.

Drewnowski, Adam. 1997. "Taste preferences and food intake." *Annual review of nutrition* 17(1): 237-253.

Drewnowski, Adam, and S. E. Specter. 2004. "Poverty and obesity: the role of energy density and energy costs." *The American journal of clinical nutrition* 79(1): 6-16.

Drewnowski, Adam, and Petra Eichelsdoerfer. 2010. "Can low-income Americans afford a healthy diet?" *Nutrition today* 44 (6): 246.

Drewnowski, Adam. 2013. "New metrics of affordable nutrition: which vegetables provide most nutrients for least cost?" *Journal of the Academy of Nutrition and Dietetics* 113(9): 1182-1187.

Drewnowski, Adam. 2015. "The carbohydrate-fat problem: can we construct a healthy diet based on dietary guidelines?" *Advances in Nutrition: An International Review Journal* 6(3): 318S-325S.

Egger, Peter H., and Maximilian Von Ehrlich. 2013. "Generalized propensity scores for multiple continuous treatment variables." *Economics Letters* 119(1): 32-34.

Ervin, R. Bethene. 2011. "Healthy Eating Index–2005 total and component scores for adults aged 20 and over: National Health and Nutrition Examination Survey, 2003–2004." *National Health Statistics Report* 44: 1-9.

Flegal, Katherine M., Barry I. Graubard, David F. Williamson, and Mitchell H. Gail. 2005. "Excess deaths associated with underweight, overweight, and obesity." *Jama* 293(15): 1861-1867.

Glanz, Karen, Michael Basil, Edward Maibach, Jeanne Goldberg, and D. A. N. Snyder. 1998. "Why Americans eat what they do: taste, nutrition, cost, convenience, and weight control concerns as influences on food consumption." *Journal of the American Dietetic Association* 98(10): 1118-1126.

Glanz, K., Sallis, J. F., Saelens, B. E., & Frank, L. D. 2007. Nutrition Environment Measures Survey in stores (NEMS-S): development and evaluation. *American journal of preventive medicine* 32(4): 282-289.

Golan, Elise, Hayden Stewart, Fred Kuchler, Diansheng Dong, and John A. Kirlin. 2008. "Can low-income Americans afford a healthy diet?" *Amber Waves* 6 (5): 26.

Grossman, Michael. 1972. "On the concept of health capital and the demand for health." *Journal of Political economy* 80(2): 223-255.

Guenther, Patricia M., Jill Reedy, and Susan M. Krebs-Smith. 2008. "Development of the healthy eating index-2005." *Journal of the American Dietetic Association* 108(11): 1896-1901.

Guo, X., B. A. Warden, S. Paeratakul, and G. A. Bray. 2004. "Healthy eating index and obesity." *European journal of clinical nutrition* 58 (12): 1580-1586.

Guthman, Julie. 2013. "Too much food and too little sidewalk? Problematizing the obesogenic environment thesis." *Environment and Planning A* 45(1): 142-158.

Hann, Clayton S., Cheryl L. Rock, Irena King, and Adam Drewnowski. 2001. "Validation of the Healthy Eating Index with use of plasma biomarkers in a clinical sample of women." *The American journal of clinical nutrition* 74(4): 479-486.

He, Ke, F. B. Hu, G. A. Colditz, J. E. Manson, W. C. Willett, and S. Liu. 2004. "Changes in intake of fruits and vegetables in relation to risk of obesity and weight gain among middle-aged women." *International journal of obesity* 28 (12): 1569-1574.

Healthy People 2020. 2017. “2020 Topics & Objectives: Nutrition and Weight Status” https://www.healthypeople.gov/2020/topics-objectives/topic/nutrition-and-weight-status

Hill, Sarah E., Kaily Baskett, Hannah K. Bradshaw, Marjorie L. Prokosch, Danielle J. DelPriore, and Christopher D. Rodeheffer. 2016. "Tempting foods and the affordability axiom: Food cues change beliefs about the costs of healthy eating." *Appetite* 107: 274-279.

Hilmers, Angela, David C. Hilmers, and Jayna Dave. 2012. "Neighborhood disparities in access to healthy foods and their effects on environmental justice." *American Journal of Public Health* 102(9): 1644-1654.

Hirano, Keisuke, and Guido W. Imbens. 2004. “The propensity score with continuous treatments." Applied Bayesian modeling and causal inference from incomplete-data perspectives 226164: 73-84.

Hosler, A. S., Varadarajulu, D., Ronsani, A. E., Fredrick, B. L., & Fisher, B. D. 2006. Low-fat milk and high-fiber bread availability in food stores in urban and rural communities. *Journal of Public Health Management and Practice*, 12(6): 556-562.

Imai, Kosuke, and David A. Van Dyk. 2004. “Causal inference with general treatment regimes: Generalizing the propensity score." Journal of the American Statistical Association 99(467): 854-866.

Imamura, Fumiaki, Laura O’Connor, Zheng Ye, Jaakko Mursu, Yasuaki Hayashino, Shilpa N. Bhupathiraju, and Nita G. Forouhi. 2015. "Consumption of sugar sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: systematic review, meta-analysis, and estimation of population attributable fraction." *Bmj* 351: h3576.

Jeffery, Robert W., and Simone A. French. 1998. "Epidemic obesity in the United States: are fast foods and television viewing contributing?" *American journal of public health* 88(2): 277-280.

Jekanowski, M. D., & Binkley, J. K. 2000. “Food purchase diversity across US markets.” *Agribusiness: An International Journal*, 16(4): 417-433.

Kennedy, Eileen T., James Ohls, Steven Carlson, and Kathryn Fleming. 1995. "The healthy eating index: design and applications." *Journal of the American Dietetic Association* 95(10): 1103-1108.

Krukowski, Rebecca A., Jean Harvey-Berino, Jane Kolodinsky, Rashmi T. Narsana, and Thomas P. DeSisto. 2006. "Consumers may not use or understand calorie labeling in restaurants." *Journal of the American Dietetic Association* 106(6): 917-920.

Lafay, L., L. Mennen, A. Basdevant, M. A. Charles, J. M. Borys, E. Eschwege, and M. Romon. 2000. "Does energy intake underreporting involve all kinds of food or only specific food items? Results from the Fleurbaix Laventie Ville Sante (FLVS) study." *International journal of obesity* 24(11): 1500.

Lantz, Paula M., James S. House, James M. Lepkowski, David R. Williams, Richard P. Mero, and Jieming Chen. 1998. "Socioeconomic factors, health behaviors, and mortality: results from a nationally representative prospective study of US adults." *Jama* 279(21): 1703-1708.

Larson, Nicole I., Mary T. Story, and Melissa C. Nelson. 2009. "Neighborhood environments: disparities in access to healthy foods in the US." *American journal of preventive medicine* 36(1): 74-81.

Laska, Melissa N., Caitlin E. Caspi, Jennifer E. Pelletier, Robin Friebur, and Lisa J. Harnack. 2015. "Peer Reviewed: Lack of Healthy Food in Small-Size to Mid-Size Retailers Participating in the Supplemental Nutrition Assistance Program, Minneapolis–St. Paul, Minnesota, 2014." *Preventing chronic disease* 12.

Liese, A. D., Weis, K. E., Pluto, D., Smith, E., & Lawson, A. 2007. Food store types, availability, and cost of foods in a rural environment. *Journal of the American Dietetic Association*, 107(11): 1916-1923.

Malone, Trey, and Jayson L. Lusk. 2017. "Taste Trumps Health And Safety: Incorporating Consumer Perceptions Into A Discrete Choice Experiment For Meat." *Journal of Agricultural and Applied Economics* 49(1): 139-157.

Mattes, Richard D. 1997. "The taste for salt in humans." *The American Journal of Clinical Nutrition* 65(2): 692S-697S.

Mattes, Richard D. 2006. "Orosensory considerations." *Obesity* 14(S7): 164S-167S.

Moore, Latetia V., Ana V. Diez Roux, Jennifer A. Nettleton, David R. Jacobs, and Manuel Franco. 2009. "Fast-food consumption, diet quality, and neighborhood exposure to fast food: the multi-ethnic study of atherosclerosis." *American journal of epidemiology* 170(1): 29-36.

Morning Consult. 2018. "Consumer Trends in the Food and Beverage Industry". https://morningconsult.com/wp-content/uploads/2018/05/Morning-Consult-Consumer-Trends-In-The-Food-and-Beverage-Industry.pdf

Must, Aviva, Jennifer Spadano, Eugenie H. Coakley, Alison E. Field, Graham Colditz, and William H. Dietz. 1999. "The disease burden associated with overweight and obesity." *Jama* 282 (16): 1523-1529.

National Institutes of Health. 2017. "What causes obesity & overweight?" https://www.nichd.nih.gov/health/topics/obesity/conditioninfo/Pages/cause.aspx

Pryer, Jane A., Martine Vrijheid, Robert Nichols, Matthew Kiggins, and Paul Elliott. 1997. "Who are the 'low energy reporters' in the dietary and nutritional survey of British adults?." *International journal of epidemiology* 26(1): 146-154.

Rolls, Barbara J., Erin L. Morris, and Liane S. Roe. 2002. "Portion size of food affects energy intake in normal-weight and overweight men and women." *The American journal of clinical nutrition* 76(6): 1207-1213.

Slavin, Joanne L., and Beate Lloyd. 2012. "Health benefits of fruits and vegetables." *Advances in Nutrition: An International Review Journal* 3(4): 506-516.

Short, Anne, Julie Guthman, and Samuel Raskin. 2007. "Food deserts, oases, or mirages? Small markets and community food security in the San Francisco Bay Area." *Journal of Planning Education and Research* 26(3): 352-364.

Stewart, Hayden, Noel Blisard, Sanjib Bhuyan, and Rodolfo M. Nayga Jr. 2004. "The demand for food away from home." *US Department of Agriculture-Economic Research Service Agricultural Economic Report* 829.

Stewart, Hayden, and J. Michael Harris. 2005. "Obstacles to overcome in promoting dietary variety: the case of vegetables." *Review of Agricultural Economics 27*(1): 21-36.

Stewart, Hayden. 2011. *How much do fruits and vegetables cost?* No. 71. DIANE Publishing.

Thompson, Olivia M., C. Ballew, K. Resnicow, A. Must, L. G. Bandini, H. D. W. H. Cyr, and W. H. Dietz. 2004. "Food purchased away from home as a predictor of change in BMI z-score among girls." *International journal of obesity* 28(2): 282-289.

United States Department of Agriculture. 2015. “History of Dietary Guidance Development in the United States and the Dietary Guidelines for Americans.” *Federal Presentations*. https://health.gov/dietaryguidelines/2015-BINDER/meeting1/historyCurrentUse.aspx

United States Department of Agriculture Center for Nutrition Policy & Promotion. 2008. “Development of the CNPP Prices Database.” https://www.cnpp.usda.gov/sites/default/files/usda_food_plans_cost_of_food/PricesDatabaseReport.pdf

United States Department of Agriculture Center for Nutrition Policy & Promotion. 1995. “The Healthy Eating Index.” https://www.cnpp.usda.gov/sites/default/files/healthy_eating_index/HEI89-90report.pdf

Ver Ploeg, Michele, and Ilya Rahkovsky. 2016. "Recent Evidence on the Effects of Food Store Access on Food Choice and Diet Quality." *Amber Waves*: 1C.

Yeomans, Martin R. 1998. “Taste, palatability and the control of appetite.” *Proceedings of the Nutrition Society* 57(4): 609-615.

Zenk, Shannon N., Amy J. Schulz, Teretha Hollis-Neely, Richard T. Campbell, Nellie Holmes, Gloria Watkins, Robin Nwankwo, and Angela Odoms-Young. 2005. "Fruit and vegetable intake in African Americans: income and store characteristics." *American journal of preventive medicine* 29(1): 1-9.