# Prediction of Shopping Behavior Using a Huff Model Within a GIS Framework

James D. Hibbert[1], Sarah E. Battersby[2], Angela D. Liese[1]

1. Department of Epidemiology and Biostatistics, and
Center for Research in Nutrition and Health Disparities, Arnold School of Public Health, University of South Carolina Columbia, South Carolina

2. Department of Geography, University of South Carolina
Columbia, South Carolina

# Background

- A common retail marketing question is “How can I estimate consumer demand for individual retail locations?”
- Reilly (1931) applied the gravity concept to retail trade analysis:
  - Determine the retailing pulling power of two competing cities on an intervening area
  - Drawing power of a facility is directly proportional to the size of the facility and inversely proportional to distance
  - Focused on stores rather than consumers

# Huff Model

- The Huff Model (1963) is a probabilistic retail gravity model used to predict consumer behavior among competing retail shopping locations.

$$P_{ij} = \frac{\frac{S_j}{T_{ij}^{\lambda}}}{\sum_j^n \frac{S_j}{T_{ij}^{\lambda}}}$$

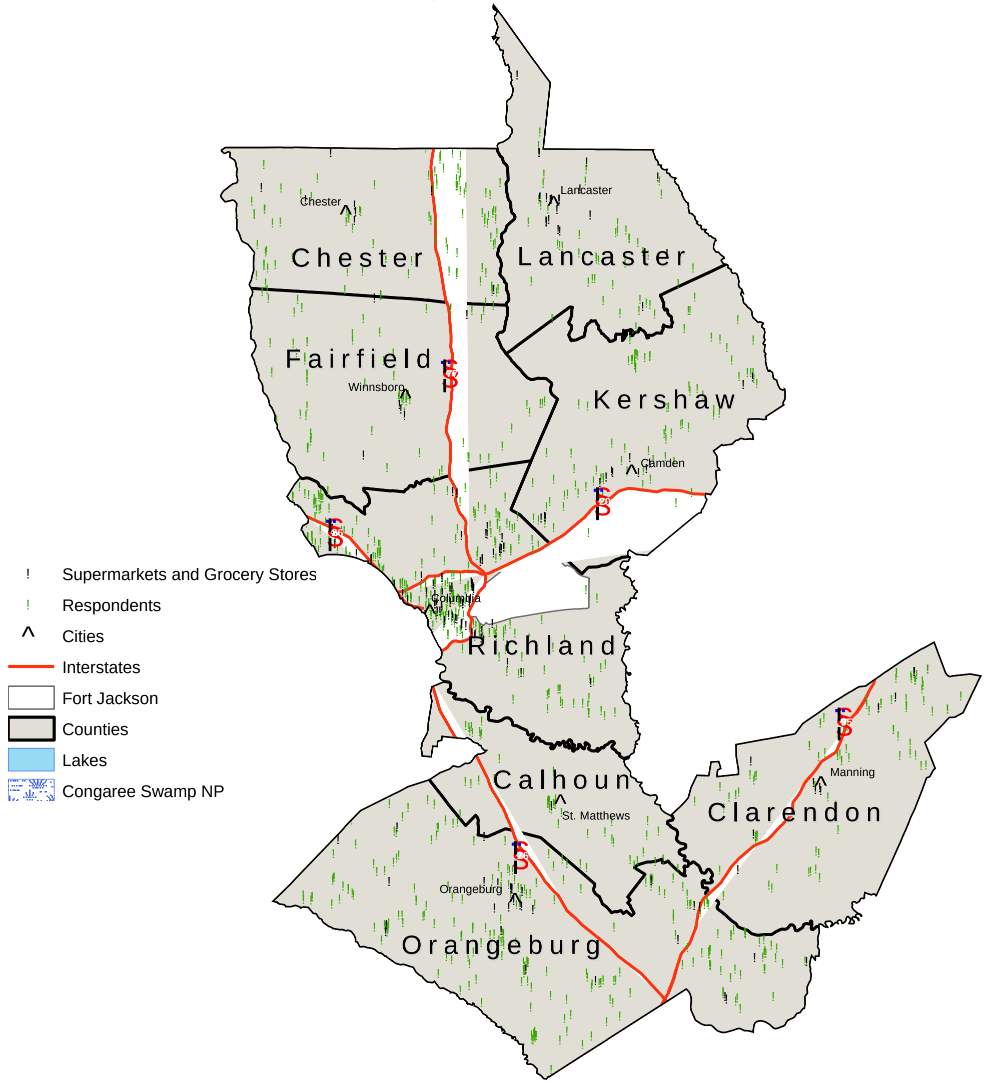
| | |
|---|---|
| $P_{ij}$ | The probability of a consumer at point *i* travelling to retail location *j* |
| $S_j$ | Size of retail location *j* |
| $T_{ij}$ | Travel time (or distance) from consumer at point *i* to travel to location *j* |
| $\lambda$ | Estimated parameter to reflect effect of travel on shopping trips (1-2) |

# Study Region and Respondents

- For this study, a dataset of 968 respondents were geocoded and queried for their primary grocery shopping location
- These locations were then matched to a GPS-verified dataset of supermarkets, supercenters, warehouse clubs, and grocery stores in a 8 county region of South Carolina
- 18 could not be geocoded,  146 shopped at a location not accounted for in the verified data – 804 in the Huff model analysis

# Study Region and Respondents

- Respondents sampled by ZIP code to be in study area (~15 per ZIP code)
- 166 total supermarkets, large grocery stores, supercenters, and warehouse clubs identified in area
- All 166 served in model as potential destinations – 91 were actually utilized by respondents



# Methods

- Traditionally the model used planar distances to make it less computationally intensive to solve, however GIS allows for a more realistic approach using street distances and travel time

- Origin-Destination cost matrix
  - Calculates the shortest street distance from each respondent to each store.

- Employee number comes from InfoUSA retail outlet data
  - This was the most complete descriptive measure, and served as a proxy for attractiveness.

- Model accuracy was determined by percent agreement of store with largest probability to each respondents' self-reported primary food shopping outlet

# Methods

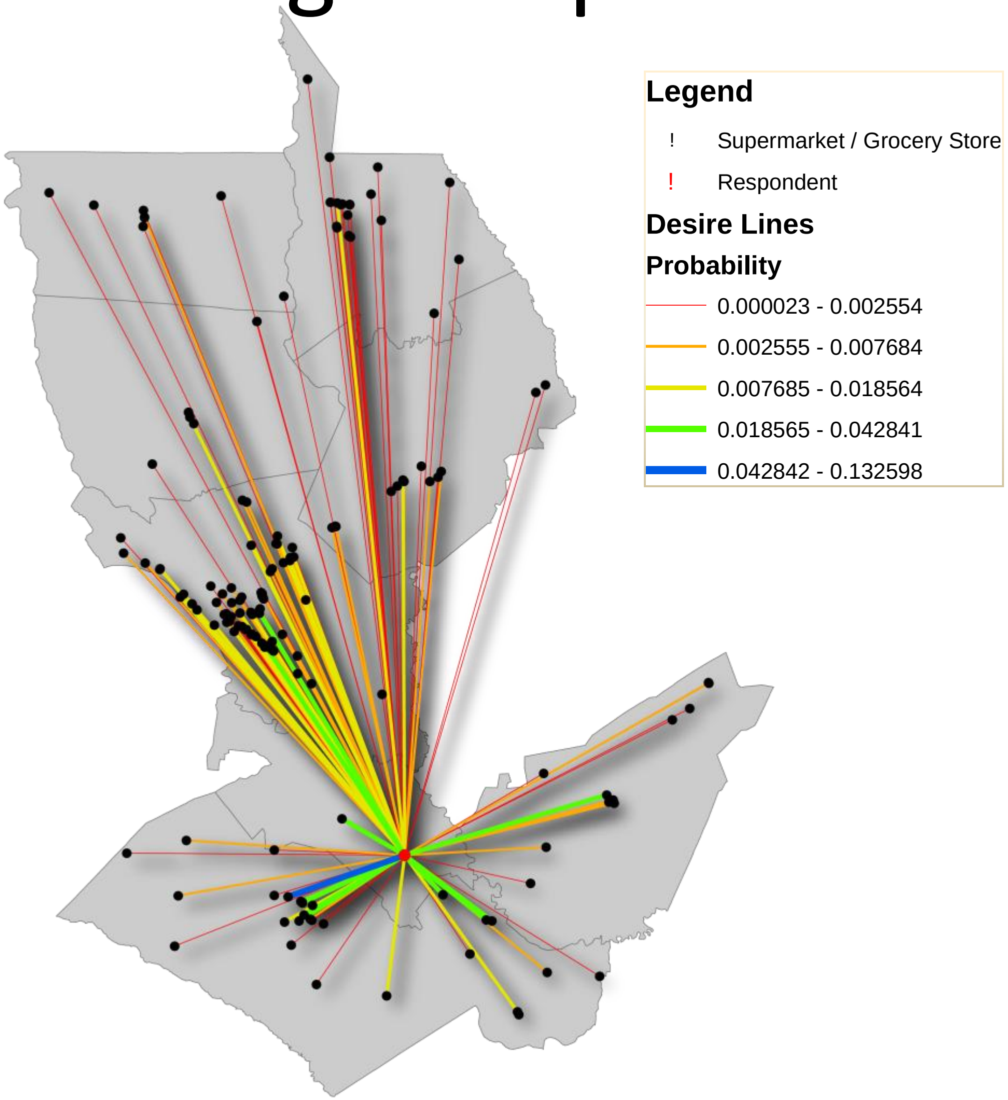
- Maximum probability as determined by the model is selected for each respondent

| | Store 1 | Store 2 | Store 3 | Store 4 | Store 5 | Store Utilized | Store Predicted |
|---|---|---|---|---|---|---|---|
| Respondent 1 | .225 | .055 | .466 | .150 | .104 | Store 3 | Store 3 |
| Respondent 2 | .890 | .005 | .083 | .007 | .015 | Store 5 | Store 1 |
| Respondent 3 | .001 | .075 | .750 | .060 | .114 | Store 3 | Store 3 |

Output table: List of store preferences (percent of time expected to go to each store)

# Example output of a single respondent

- 160 probabilities calculated for each respondent
- Represented as "desire lines" – connecting respondents to their potential stores.
- Red represents stores of lower probability of use, blue and green of higher probability of use



# Results (Distance based)

| County | % Agreement | Population Density (persons / mile°) |
|---|---|---|
| Calhoun | 31.4 | 36.4 |
| Chester | 34.5 | 58.2 |
| Clarendon | 36.4 | 44.2 |
| Fairfield | 20.3 | 32.0 |
| Kershaw | 24.1 | 67.5 |
| Lancaster | 29.5 | 104.4 |
| Orangeburg | 39.7 | 81.9 |
| Richland | 35.2 | 409.6 |
| **Overall** | **33.5** | |

# Results (Time based)

| County | % Agreement | Population Density (persons / mile°) |
|---|---|---|
| Calhoun | 28.6 | 36.4 |
| Chester | 33.3 | 58.2 |
| Clarendon | 36.4 | 44.2 |
| Fairfield | 21.9 | 32.0 |
| Kershaw | 26.5 | 67.5 |
| Lancaster | 27.3 | 104.4 |
| Orangeburg | 38.7 | 81.9 |
| Richland | 33.8 | 409.6 |
| **Overall** | **33.0** | |

# Results Summary (Probability)

| | Rural (Distance) | Rural (Time) | Urban (Distance) | Urban (Time) |
|---|---|---|---|---|
| **Mean** | 0.22 | 0.19 | 0.24 | 0.22 |
| **Median** | 0.17 | 0.14 | 0.19 | 0.17 |
| **Mode** | 0.49 | 0.38 | 0.20 | 0.19 |
| **Standard Deviation** | 0.16 | 0.14 | 0.16 | 0.16 |
| **Sample Variance** | 0.03 | 0.02 | 0.03 | 0.02 |
| **Range** | 0.91 | 0.91 | 0.94 | 0.94 |
| **Minimum** | 0.05 | 0.05 | 0.06 | 0.06 |
| **Maximum** | 0.97 | 0.96 | 1.00 | 1.00 |
| **Sum** | 141.28 | 122.94 | 38.80 | 35.31 |
| **Count** | 641 | 641 | 163 | 163 |

# Discussion

- Overall, model predicted correctly for about 33% of respondents

- Results for distance and time as accessibility measures were similar (T and D highly correlated)

# Discussion

- Better accessibility data (Modifies ***T***)
  - Incorporate into the distance/time calculations enhanced travel times based on more realistic conditions (cost, and barriers)

$$P_{ij} = \frac{\frac{S_j}{T_{ij}^{\lambda}}}{\sum_j^n \frac{S_j}{T_{ij}^{\lambda}}}$$

- This study considered all stores in the study region – can be limited by neighborhood size (Modifies ***n***)

# Discussion

- Employee number was used as an "attractiveness" parameter since it was most complete – better parameters can be used such as sales volume, square footage, or consumer survey data (Modifies ***S***)

- Modify λ to reflect the sensitivity of the model with respect to distance

$$P_{ij} = \frac{\frac{S_j}{T_{ij}^{\lambda}}}{\sum_j^n \frac{S_j}{T_{ij}^{\lambda}}}$$

# Next Steps

- Parameterize the current Huff Model
- Actual market basket values used as attractiveness factor (attractiveness varies inverse to food costs)
- Exploration of other types of spatial interaction models

# Questions / Comments?