# Longitudinal Human Mobility and Real-time Access to a National Density Surface of Retail Outlets

Thomas R. Kirchner[1,2,3] Hong Gao[3], Andrew Anesetti-Rothermel[3,4], Heather Carlos[5], Brian House[6]

[1] Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[2] Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC, USA
[3] The Schroeder Institute at Legacy, Washington, DC, USA
[4] Department of Epidemiology, School of Public Health, West Virginia University, Morgantown, WV, USA
[5] Norris Cotton Cancer Center, Dartmouth College, Hanover, NH, USA
[6] Brown University, Providence, RI, USA

## ABSTRACT

New methods are making it possible to investigate the way points of interest (POIs) within cities affect the health of citizens traveling through their streets. In this paper we present the development of a framework for study of human mobility and real-time access to the landscape of point-of-sale products available across the US, a POI with indisputable implications for a range of health related behaviors. A nationwide density surface of convenience and related retail outlet locations was generated using static bandwidth kernel density estimation (KDE). This resulted in a smooth, continuous density surface where every location in the study area has an assigned density value. This surface was then linked to each real-time mobility coordinate contributed by participants taking part in a real-time geo-location tracking service. Longitudinal human mobility data were collected from a sample of 550 people who continuously recorded their geographic location via their cellular phone every 10-minutes over an average of 8 months. Hourly and daily mobility patterns were characterized by radius of gyration and associated contact with the retail density surface. Analyses explore the way real-time access to retail products varied as a function of both their geographic distribution and the mobility patterns of participants. These data suggest that research on access to retail products must account for the mobility and preferences of individuals as they engage with POI over time. Results demonstrate the kind of insight that can be gained by embracing the dynamic interplay between city-level POIs and human mobility patterns.

## Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

## General Terms

Algorithms, Measurement, Experimentation, Human Factors

## Keywords

human mobility, GIS, urban computing, public health


*UrbComp'14*, August 24, 2014, New York, NY, USA.

## 1. INTRODUCTION

Interest in urban points of interest (POI) and the health of citizens is growing rapidly [1, 8]. Previous work has been founded on the notion that health-related POI can be used to approximate individual-level exposures, because POI are assumed to have a commensurate, stable association with the health of those in near-by residential addresses. This assumption is flawed to the extent that individual POI exposures are determined by non-home mobility. Beyond city-level POI, it is increasingly clear that individual differences in mobility patterns produce a dynamic interaction between citizens and their real-time surroundings [2-4].

Individual mobility patterns and the preferences of citizens actively engaging with their real-time context make exposure and thus the impact of urban POIs on health highly dynamic. The growing empirical literature on human mobility[2, 3] supports the notion that mobility patterns determine POI exposure levels, more than can be approximated by static factors like residential location. Yet methods for real-time quantification of accumulating levels of exposure to urban POI are only now being developed.

In this paper we present the development of a framework for the study of human mobility and real-time access to the landscape of point-of-sale products available across the US, a factor with indisputable implications for a range of health related behaviors [5-7]. This begins with a national probability density surface representing the continuous landscape of convenience, grocery and gas outlets in the US. We then overlay longitudinal human mobility data collected from a sample of 550 people who have voluntarily recorded their real-time geographic location via their cellular phone every 10-minutes (while in motion) over an average of 8 months, resulting in an average of 3,711 observations per user. Analyses examine the extent to which the dynamic nature of participants' mobility patterns interacted with their real-time surroundings and thus determined their day-to-day access to products sold in retail convenience stores. Variations across urban areas and over time provide insight and identify targets of intervention for both urban planners and public health practitioners.

### 1.1 Related work

Urban computing involves the study of urban dynamics using data from sensors, devices, persons, vehicles, buildings, and GIS repositories in urban areas for the benefit of citizens [8]. A rapidly advancing application of urban computing leverages data on the geo-location of citizens within cities to understand socio-behavioral processes and health [2, 4, 9].

The burgeoning literature on human mobility suggests that mobility patterns can reveal important information about the way citizens interact with POI in cities [10-13]. Zheng et al. [8] and Qi

et al. [14] demonstrate that taxi mobility data can be used to identify urban social functions and city planning errors. Complementing this work, Yuan, Zheng et al. [2] were the first to identify functional urban regions by overlaying static POIs of a region with human mobility patterns. Social media platforms provide another example of data on the geographic distribution of citizens within cities [14-16]. Unfortunately, the applied implications of such data for health remains unrealized [17, 18].

Recent advances notwithstanding, the empirical literature on the impact of POIs on health behavior is currently in its infancy. Recent work shows that the geographic density of retail tobacco [19-23] and alcohol [24-28] outlets is associated with aggregated patterns of use and treatment. Geospatial analysis methods have also been applied to the study of physical activity, retail food environments, diet and weight [29-32]. Kirchner et al., [4, 9] was the first to link real-time geographic exposure health-related POIs to addiction-related behavioral outcomes. The present paper advances this literature in a number of ways. Focusing on a reliable, national source of health-related POI data allowed us to identify variation in urban dynamics and behavior across different regions of the US. The use of continuous real-time geo-location tracking provides excellent temporal and spatial resolution, which improved sensitivity to detect dynamic patterns in the data. These methodologic advantages, combined with well-established sources of health-related data, enhance the contribution of this report.

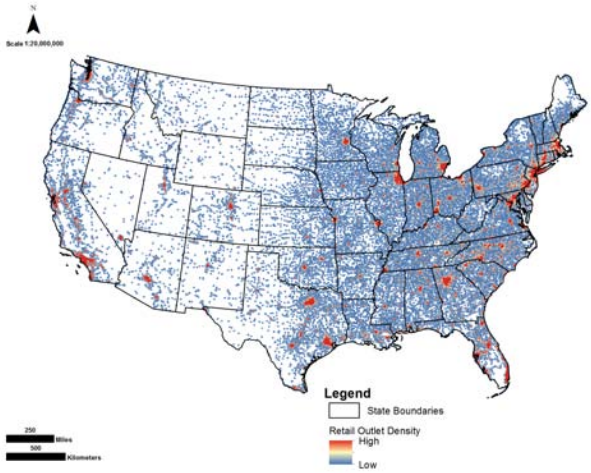
# 2. METHOD

## 2.1 Retail Density Surface

A nationwide density surface of convenience and related retail outlet locations was generated using kernel density estimation (KDE). The empirical basis for this probability density surface was a national dataset of retail outlets, identified by North American Industry Classification Systems (NAICS) codes. Developed by the Office of Management and Budget, NAICS is the standard used by Federal statistical agencies to classify businesses based on their primary activity [33]. In 2012, geocoded data was obtained from OneSource's (now Avention) Global Business Browser.. The following retail categories and corresponding NAIDS codes were included: beer, wine, and liquor stores (NAICS: 445310); supermarkets and other grocery stores (NAICS: 44511); convenience stores (NAICS: 44512); pharmacies and drug stores (NAICS: 446110); gasoline stations with convenience stores (NAICS: 44711); other gasoline stations (NAICS: 44719); department stores (NAICS: 452111); discount department stores (NAICS: 452112); and tobacco stores (NAICS: 453991). The final dataset included N = 269,781 retail outlets (**Figure 1**).

To quantify individuals' real-time access to retail outlets, we used a static bandwidth KDE approach. This non-parametric method extrapolates point data over a study area by calculating the density of the points using a specified bandwidth (e.g., a circle of a given radius centered at the focal location) [34]. This results in a smooth, continuous density surface where every location (i.e., every 250m pixel) in the study area has an assigned density value. The static bandwidth in the KDE analyses was optimized by constructing numerous density surfaces across an array of distance-based bandwidths [1]. Bandwidth selection was based on which density surface optimized the kurtosis of the density surface (e.g., not too peaked or too smoothed), given the underlying point distribution. This bandwidth optimization approach was carried out with the spatial analyst density toolset in ArcGIS v.10.1 software. To produce a final density surface in ArcGIS, a Gaussian kernel with an "optimized" fixed 5-mile bandwidth was used. The resulting density surface had a cell size of 250 meters. We then calculated zonal statistics of the final density surface for zip code tabulation areas (ZCTAs) in the United States. ZCTAs are generalized areal representations of the zip code service areas used by the United States Postal Services [35]. The average retail outlet density per ZCTA was then linked to each real-time mobility coordinate contributed by the participants in our geo-location tracking sample.

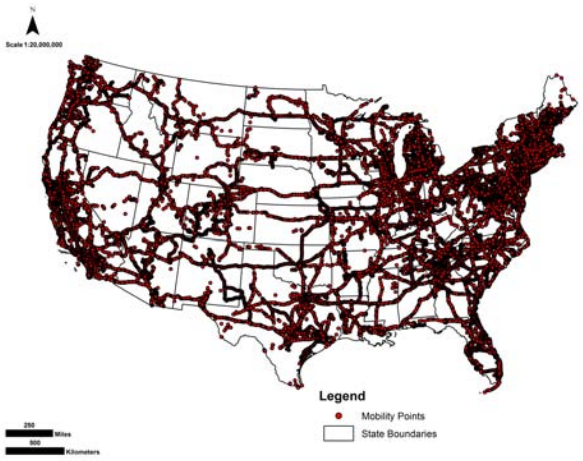
**Figure 1. National Density Surface: US Convenience Retail**




## 2.2 Longitudinal Mobility Data

Geolocation tracking made it possible to physically link each person's real-time location to the probabilistically continuous landscape of convenience stores across the US. Mobility data comes from the OpenPaths (https://openpaths.cc) platform, launched by the New York Times Company Research and Development Lab in May of 2011 [36]. OpenPaths collects GPS location information from volunteers via iOS and Android location tracking applications. These applications utilized multiple approaches to geographic location capture, including both direct satellite GPS coordinates (when available) and wireless network-based "assisted" geo-location estimation via trilateration among cellular towers and wireless data access points.

**Figure 2. Human Mobility Cohort**



Trilateration-based approaches are particularly useful for capture of location coordinates while the user is inside a building or vehicle, where direct satellite coordinates are usually unavailable. The OpenPaths applications wirelessly report the user's location every 10-minutes, unless the user remains stationary (to save battery). Records are updated when a significant location change (more than 5m) is detected, based on cell-tower and Wi-Fi-node trilateration [36]. The longitudinal mobility dataset contained 3,440,821 observations collected from 859 individuals worldwide from 03/01/2012 to 12/31/2013. Each observation includes a unique user ID, date, time, latitude, and longitude. The raw mobility data was clipped to United States using state outline polygons published by United States Census Bureau, yielding a US cohort of 550 individuals with 2,013,042 observations recording during the present observation period. The total number of locations contributed by each individual ranges from 1 to 79,602 with a mean of 3711 (Median=2706) and standard deviation of 2706. The number of days falls between 1 and 993 with a mean of 243.09 days (Median=211.5; SD=183.59).

### 2.2.1 States of residence

Noteworthy, 19.33% of the data (N= 389,058) falls in California, and 14.45% (N= 290,854) in New York. All other states have 66.22% (N= 1,333,130). Given the nice contrast in terms of geo-locations, climate, built-in environment, etc., between New York and California, the two states were entered as covariates in all models.

## 2.3 Radius of Gyration

Radius of gyration ($r_g$) measures the distance a person travels within a certain time period [3]. It defines by the standard deviation between locations and their center of mass:

$$r_g^2 = \frac{1}{N}\sum_{k=1}^{N}(r_k - r_{mean})^2$$

where N is the total number of locations an individual has been to, and $r_{mean}$ is the individual center of mass, or the mean longitude and latitude of all N locations. $(r_k - r_{mean})$ is the great circle distance in kilometers between a specific location and the center of mass calculated using Vincenty's formulae [37]. Hourly $r_g$ was obtained for each individual, resulting in 747,347 observations.

**Table 1. Raw $r_g$ Mean (SD) by State, Weekend, TOD, Season**

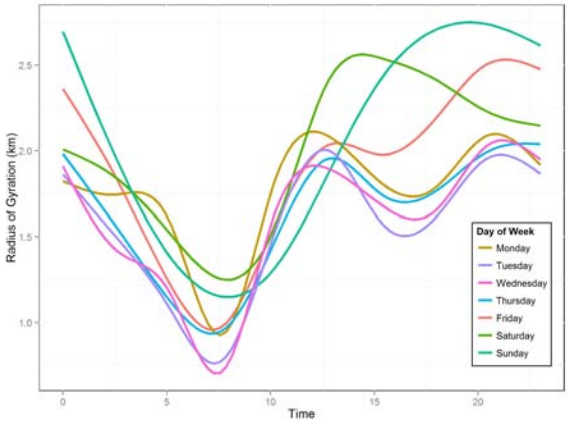
| | | Time of Day | | | |
|---|---|---|---|---|---|
| State | Week End | Early | Day | Evening | Late |
| NY | No | 0.70 | 1.28 | 1.20 | 1.16 |
| | | (2.22) | (3.14) | (3.38) | (3.07) |
| | Yes | 0.77 | 1.41 | 1.43 | 1.38 |
| | | (2.20) | (3.75) | (3.73) | (3.76) |
| CA | No | 1.58 | 1.69 | 1.73 | 1.99 |
| | | (3.76) | (3.36) | (4.05) | (4.10) |
| | Yes | 1.90 | 1.54 | 2.31 | 2.43 |
| | | (4.49) | (3.63) | (4.68) | (5.25) |

| | | Season | | | |
|---|---|---|---|---|---|
| State | Week End | Spring | Summer | Fall | Winter |
| NY | No | 1.25 | 1.19 | 0.96 | 1.22 |
| | | (3.15) | (3.25) | (2.73) | (3.29) |
| | Yes | 1.46 | 1.37 | 1.07 | 1.22 |
| | | (3.77) | (3.65) | (3.11) | (3.42) |
| CA | No | 1.85 | 1.82 | 1.74 | 1.58 |
| | | (3.99) | (4.00) | (3.92) | (3.54) |
| | Yes | 2.31 | 2.27 | 1.99 | 2.01 |
| | | (5.04) | (4.87) | (4.42) | (4.41) |

Because location collectors are motion sensitive, $r_g$ for hours with only one location observation were assumed to be 0. Because this work focuses on day-to-day mobility routines, long-range travel was excluded by dropping observations with $r_g$ larger than 120 kilometers (N=106), the maximum distance a car can travel in one hour (always shorter than that of a plane), yielding 550 individuals and 747,241 observations (99.9% of the original data).

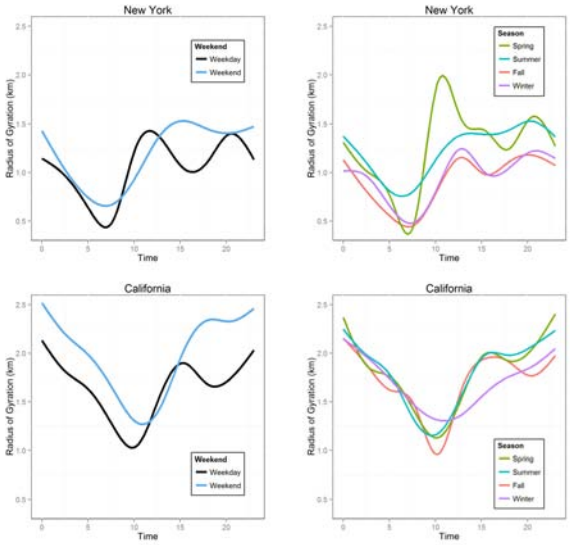
Validity of the data is initially supported by observation of expected patterns of mobility, such as that related to weekdays and weekends. **Figure 3** illustrates the daily drop in mobility across the early morning hours, followed by a steep rise across the middle of the day, and then divergence on Friday, Saturday and Sunday, with late Sunday revealed as the window of greatest mobility, as travelers who departed on either Friday or Saturday return home.

**Figure 3. $r_g$ over 24-hr clock as function of day of week.**




Overall, mean $r_g$ was 1.93 kilometers per hour with a 4.41 kilometers standard deviation. Minimum and maximum were 0 and 119.51 kilometers. Within each day, $r_g$ was the lowest early and higher across the remainder of the day. We see a spike in $r_g$ mid-day on weekdays – and generally more variation on weekdays. Radius of gyration was higher in CA and on weekends than weekdays (**Table 1**). **Figure 4** depicts the markedly different mobility patterns in NY (mid-late peak), and CA (V-shaped).

**Figure 4. $r_g$ over 24-hrs as f() of State, Weekend, Season**



## 2.4 Real-time Retail Access

Real-time access to the retail density surface was defined as the product of each participant's $r_g$ within each hour under observation and their average retail outlet density value for each mobility coordinate recorded within the same hour. Conceptually, this "Access" variable accumulated the number of retail options participants had as they moved, and as expected from a count variable of this kind, the observed distribution was heavily skewed right, and thus not a reasonable fit for the assumptions of the general linear model (see Section 3.1). Access was therefore stratified into deciles (plus an additional level for values of zero), effectively transforming it into a categorical variable with 11 levels from 0 to 10 corresponding to growing access to retail products. Non-parametric categorical data analysis methods were then employed as describe in Section 3.1.

So that we could model socio-temporal patterns of Access, time related variables based on social convention were created for each observation. These were Time of Day, Weekend, and Season. *Time of day* (TOD) was defined by four 6-hr windows: 3:00-9:00 as "early," 9:00-15:00 as "day," 15:00-21:00 as "evening," and 21:00-3:00 as "late," treating observations recorded between midnight and 3AM as part of the preceding day [38]. *Weekend* is binary, indicating whether each observation fell on a weekend, defined as falling after 17:00 Friday through 17:00 Sunday. *Season* was also defined as categorical, with Jan.-Mar. as winter, Apr.-Jun. as spring, Jul.-Sep. as summer, and Oct.-Dec. as fall. "State" captures the US State in which each participant resided on each day of the project, based on his or her mobility pattern (residential addresses were not otherwise available). Because most inter-state travel was excluded during $r_g$ post-processing, 94.95% of observations from each day occurred within the same state.

**Table 2. Raw real-time retail access counts**

| State | Week End | Time | Retail Access 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NY | No | Early | 1,643 | 121 | 48 | 51 | 41 | 59 | 44 | 74 | 105 | 257 | 445 |
| | | Day | 3,251 | 220 | 152 | 105 | 122 | 140 | 117 | 281 | 364 | 607 | 1,991 |
| | | Eve | 3,401 | 155 | 218 | 192 | 159 | 160 | 175 | 401 | 524 | 674 | 1,179 |
| | | Late | 3,721 | 156 | 104 | 136 | 119 | 117 | 145 | 393 | 556 | 754 | 1,895 |
| | Yes | Early | 879 | 93 | 29 | 43 | 18 | 32 | 24 | 71 | 107 | 175 | 327 |
| | | Day | 816 | 137 | 59 | 61 | 55 | 53 | 39 | 69 | 95 | 138 | 287 |
| | | Eve | 1,653 | 144 | 133 | 102 | 101 | 98 | 95 | 252 | 366 | 421 | 691 |
| | | Late | 1,212 | 111 | 51 | 73 | 41 | 43 | 69 | 121 | 212 | 241 | 510 |
| | | | | | | | | | | | | | |
| CA | No | Early | 2,700 | 200 | 277 | 361 | 344 | 301 | 426 | 510 | 441 | 557 | 376 |
| | | Day | 1,700 | 109 | 184 | 276 | 255 | 257 | 313 | 426 | 390 | 428 | 337 |
| | | Eve | 3,660 | 313 | 408 | 580 | 527 | 636 | 695 | 678 | 660 | 627 | 347 |
| | | Late | 3,133 | 270 | 409 | 555 | 578 | 587 | 730 | 843 | 895 | 825 | 484 |
| | Yes | Early | 1,326 | 103 | 169 | 154 | 129 | 131 | 232 | 241 | 233 | 281 | 205 |
| | | Day | 452 | 54 | 75 | 86 | 73 | 47 | 56 | 66 | 70 | 88 | 68 |
| | | Eve | 1,461 | 120 | 260 | 298 | 276 | 272 | 333 | 379 | 413 | 380 | 298 |
| | | Late | 1,295 | 105 | 181 | 259 | 227 | 220 | 296 | 332 | 298 | 336 | 256 |

**Table 2** presents the raw Access counts within New York and California, stratified by Weekend and TOD. Overall, 37.69% of observations were recorded when Access was 0. By TOD, Access was much higher when it was late (21:00- 3:00), and much lower when it was early (3:00- 9:00). Importantly, variation between states was immediately apparent; for example, Access in NY was characterized by spikes on weekdays due to hyper-urban density levels, while Access in CA was more distributed on weekdays. Because these temporal dynamics vary between states in complex ways, formal modeling was used to characterize the joint and conditional associations among the factors at play.

# 3. STATISTICAL ANALYSES

## 3.1 Log-Linear CDA

Retail Access, TOD, Weekend, Season, and State constitute our five variables of interest, and their interactive association can be quantified within a set of multivariate contingency tables that include a total of 704 cells. Given a highly non-linear, count-based polytomous outcome variable (Access), generalized categorical data analysis techniques were utilized [39]. Specifically, we employed exponential, log-linear modeling techniques, developed to analyze multidimensional contingency tables. Log-linear models convert the multiplicative relations among joint and marginal counts in a contingency table to additive, linear associations by transforming the counts to logarithms [39]. Model coefficients and inference are similar to standard Poisson regression for counts. Hierarchically nested model comparison techniques were used to iteratively identify the most parsimonious combination of factors required to explain the observed data (Section 3.2) The saturated model represents the log frequencies for the cell index $(_{h,i,j,k,l})$ of all (non-ordinal) combinations of State (i.e., NY as reference), Weekend, TOD, Season, and Access:

$$log(\mu^{hijkl}) = \lambda + \lambda_h^N + \lambda_i^W + \lambda_j^T + \lambda_k^S + \lambda_l^A + \lambda_{hi}^{NW} + \cdots + \lambda_{hij}^{NWT} + \cdots + \lambda_{hijl}^{NWTA} + \cdots + \lambda_{hijkl}^{NWTSA}$$

## 3.2 Best-in-class Model Selection

**Table 3. Step-down contrasts of "best-in-class" models**

| Model | (Model Terms) | Deviance | df | p | ΔLL*-2 | Δ df | χ2 (Δdf; 0.001) |
|---|---|---|---|---|---|---|---|
| 1 | (N, W, T, S, A) | 1.2E+08 | 5 | 0.00 | -51340765 | 170 | 149.5 |
| 2 | (NWTA) | 1.5E+07 | 175 | 0.00 | -7565142 | 264 | 149.5 |
| **3** | **(NTSA, NWTA)** | **166237** | 439 | **1.00** | **-34674.3** | **45** | **86.66** |
| 4 | (NTSA, NWTA, SWA, SWN, SWT, SW) | 96888.2 | 484 | 1.00 | -11087.5 | 87 | 137.2 |
| 5 | (NWTA, WTSA, NTSA) | 74713.2 | 571 | 1.00 | -13475.4 | 12 | 59.7 |
| 6 | (NWTA, NWST, NTSA, WTSA) | 47762.4 | 604 | 1.00 | -8394.2 | 30 | 27.88 |
| 7 | (NWTA, NWST, NTSA, WTSA, NWSA) | 30973.9 | 613 | 1.00 | -15261.5 | 90 | 137.2 |
| 8 | (NWTSA) | 451.032 | 703 | 1.00 | - | - | - |

N: NY versus CA
W: weekend
T: time of day
S: season
A: access to retail

**Table 3** presents an overview of the best-in-class model selection process that sought to identify the most parsimonious model form, defined as the minimal set of parameters required to provide and adequate fit to the observed data. The initial basis for comparison is Model 8, which is the saturated model that corresponds perfectly to the raw data, having zero degrees of freedom, as the number of parameters is equivalent to the total number of cells generated by all interactive combinations of the 5 factors under study here: retail Access (the conceptual DV), State (i.e., NY as reference), Weekend, TOD, Season. Stepping-down from the saturated model, we evaluated patterns of conditional independence with a model containing every 4-way interaction (Model 7), as well as sets of models with all 4-term 4-way interactions (e.g., Model 6), all 3-term 4-way interactions (e.g., Model 5), all 2-term 4-way interactions (e.g., Model 3), all 1-term 4-way interactions (e.g., Model 2), and the full mutually-exclusive independence model (Model 1). Neither Model 1 nor the Model 2 set (i.e., the five models including a single 4-way interaction term) provided an adequate fit to the observed data. Interestingly, one of the 10 models in the Model 3 set fit the observed data well (see **Table 4**). Model 10 in **Table 4** fit while only modeling 440 of the total 704 cells under study. This is thus the most parsimonious model, effectively isolating an informative pattern in the data that then becomes the basis for inference and conclusions.

**Table 4. Best-in-class determination among all (non-ordinal) double combinations of 4-way interactions**

| Model | D | df | *p* | Term 1 | Term 2 |
|---|---|---|---|---|---|
| 1 | 6928912.0 | 383 | 0.00 | (Access, Weekend, TOD, Season) | (State, TOD, Weekend, Season) |
| 2 | 6335185.0 | 439 | 0.00 | (Access, Weekend, TOD, Season) | (Access, State, Weekend, Season) |
| 3 | 2475403.0 | 223 | 0.00 | (Access, State, Season, Weekend) | (State, TOD, Weekend, Season) |
| 4 | 2195044.0 | 439 | 0.00 | (Access, State, TOD, Season) | (Access, State, Weekend, Season) |
| 5 | 742095.8 | 223 | 0.00 | (Access, State, TOD, Weekend) | (State, TOD, Weekend, Season) |
| 6 | 484712.6 | 439 | 0.00 | (Access, Weekend, TOD, Season) | (Access, State, TOD, Weekend) |
| 7 | 589039.8 | 307 | 0.00 | (Access, State, TOD, Weekend) | (Access, State, Season, Weekend) |
| 8 | 574624.0 | 383 | 0.00 | (Access, State, TOD, Season) | (State, Season, TOD, Weekend) |
| 9 | 383940.0 | 527 | 0.00 | (Access, State, TOD, Season) | (Access, Season, TOD, Weekend) |
| *10 | 188143.5 | 439 | 1.00 | (Access, State, TOD, Season) | (Access, State, TOD, Weekend) |

Model 10 is interesting because it is considerably less complex than the ten other models that include three 4-way interaction terms, as well as the five models that include four 4-way interaction terms, and because it is the only model with only two 4-way interaction terms that provides an adequate fit to the observed data. For these reasons it is important to consider what these two 4-way interaction terms reveal about the pattern of data under observation here. The first of these interactions (Model 10, Term 1, **Table 4**) indicates that Access varied as a function of State, TOD and Season, while the second (Model 10, Term 2, **Table 4**), indicates that Access varied as a function of State, TOD and Weekend. Put another way, the way Access varied within TOD was different across States, and this association varied independently between Weekends versus Seasons.

Inspection of the terms included in Model 10 in **Table 4** reveals that this model provides a specific test of conditional independence between Weekend and Seasonal effects. That this model fits indicates that after accounting for State and TOD, patterns of Access across Weekends and quarterly Season are independent of one another. Nonetheless, returning to Table 3, it is apparent that while it is not necessary to include any interaction between Weekend and Season (Model 3), the addition of 2- and 3-way interaction parameters does incrementally improve model fit (Model 4), which, in turn, is improved following inclusion of each subsequent "best-in-class" set of parameters (Models 5, 6, 7).

**Figure 5** is a model-generated mosaic plot that captures the pattern of results. Mosaic plots present "tiles" weighted by model-predicted log cell frequencies. Access increases down the z-axis. Access in New York versus California is the first column factor, and each is further subdivided into TOD, the second column factor. Colored rows indicate Access on weekdays, while grey rows represent weekends. Rows within each set of weekday and weekend represent effects of Season. Greater Access by State and TOD equates to wider bars, while greater access by Weekend and Season equate to taller bars. Returning to Model 10, immediately striking is the pattern of Access due to State and Weekend. As previously observed, levels of Access are greater in California across the low to moderate range, while Access is greater in New York only near the top of the scale. Access on weekdays is also clearly higher than on weekends, likely due to commuting patterns. Consistent with the model selection process, tile size appears relatively balanced among Seasons within Weekdays across levels of Access.

**Figure 6** highlights the pattern observed for NY, consistent with the **Figure 5** mosaic, making it clear that elevated Access in NY spikes in the middle of the day and again when it is late on weekdays. Access only elevates during the evening on weekends.

**Figure 5. Mosaic plot of retail access by state, time-of-day, weekdays (colored), and season (rows within levels of Access)**

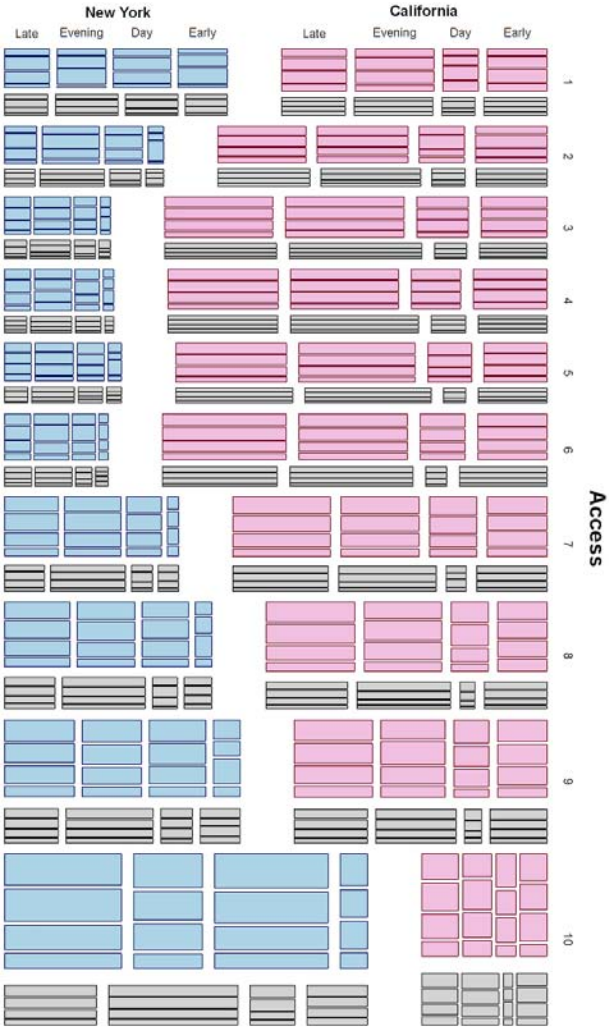



**Figure 6. NY Access as f() of TOD and Weekend**

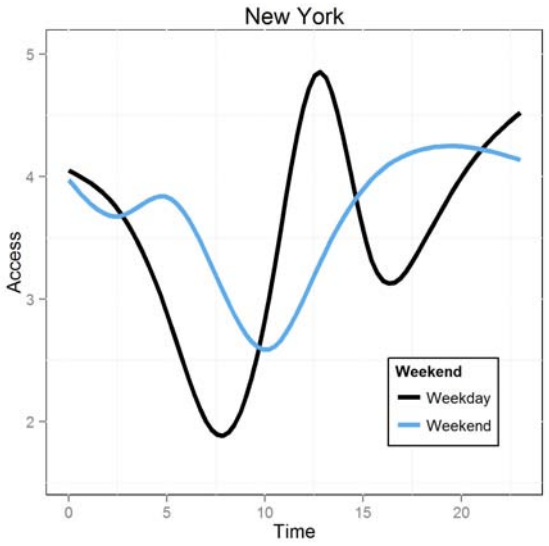
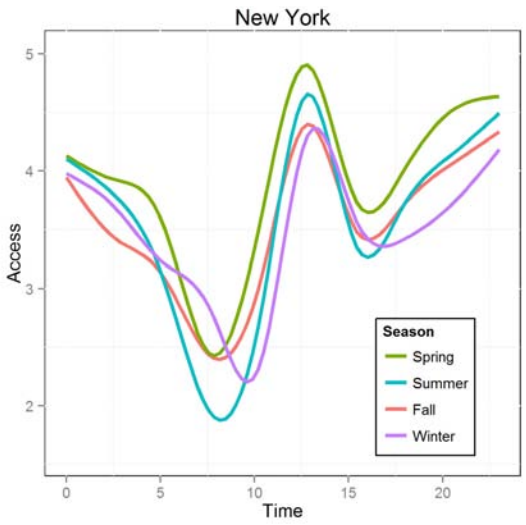
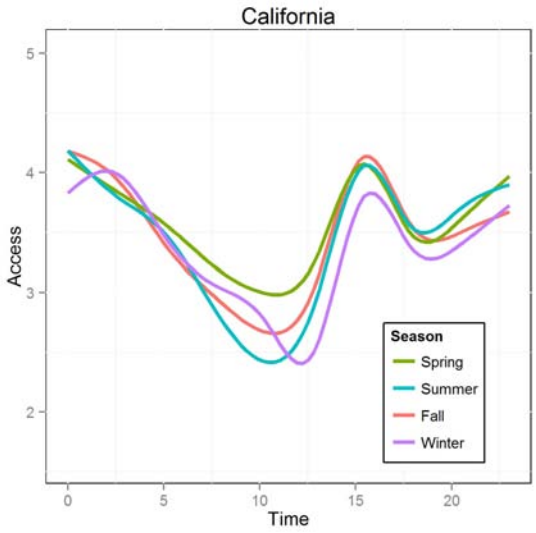


**Figure 7. NY Access as f() of TOD and Season**




Seasonal effects offer another case where Access was not simply a linear product of $r_g$ and density. Overall, $r_g$ in New York varied significantly by Season (**Figure 4**). However, in New York, $r_g$ peaked during mid-day in all seasons. Also noteworthy, mobility in the morning dropped in the summer in both New York and California. One possibility is that Seasons affect commuting and work patterns and thus Access. In New York this could explain especially high mid-day Access in Summer and Spring, which are times that residents are least likely to be on vacation and thus away from the density of New York's urban areas (**Figure 7**). Another possibility is that these patterns correspond to climate-based differences between the Northeast and the Southwest US. In California, summer and winter bring the relative extremes of temperature and therefore may drive down mobility and thus Access levels (**Figure 8**). Current exploratory work is looking into other, more dynamic conceptions of "Season," as we suspect simple partitioning of seasons based on cultural convention may be insufficiently precise when trying to capture the interaction between weather conditions and spatio-temporal mobility.

**Figure 8. CA Access as f() of TOD and Season**




# 4. CONCLUSIONS

Results of this work shed additional light on the nature of real-time retail Access, especially as it compares to the more common use of static POIs, such as aggregated outlet density information. Variations across urban areas and over time provide insight and identify targets of intervention for both urban planners and public health practitioners. The contrast between New York and California is particularly useful on this point. First of all, outlet density peaks much higher in New York than in California, while California and its urban sprawl is characterized by more widely distributed yet moderate to low levels of density. Second, $r_g$ is consistently higher in California than in New York, and exhibits an entirely different cyclic pattern. Thus from hour-to-hour and day-to-day, Access in California is generally higher than in New York. Yet, when individuals move into New York's urban centers they experience a "hyper-density" of outlets and thus accumulate very high Access levels during certain times, usually on weekdays and peaking around mid-day. This pattern of results demonstrates the kind of insight that can be gained by considering both POIs and dynamic mobility patterns. Moving forward it will be useful to ask not just "how much?" but also "where, and when?"

Future work should leverage data of this kind to inform policy, and more generally, to advance our understanding of individual decision-making and behavior change dynamics as a function of POIs. Natural extensions of this work would incorporate other policy- and health-relevant POI like particulate matter, infectious disease spread, and features associated with food and alcohol products. In all cases it will be interesting to examine the extent to which real-time POI exposure can be linked socio-behavioral processes, and ultimately to clinical outcomes.

# 5. ACKNOWLEDGMENTS

The authors are indebted to Michael Tacelosky, Ollie Ganz and Seann Regan for their assistance with this work. Seed funding was provided by the American Legacy Foundation.

## 6. REFERENCES

[1] Cromley, E.K. and Mclafferty, S.L., 2002. *GIS and Public Health*. The Guilford Press.

[2] Yuan, J., Zheng, Y., and Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the ACM KDD,* Beijing, China, ACM Press.

[3] González, M.C., Hidalgo, C.A., and Barabási, A.L., 2008. Understanding individual human mobility patterns. *Nature 453*, 7196 (Jun), 779-782.

[4] Kirchner, T.R., Cantrell, J., Anesetti-Rothermel, A., Pearson, J., Cha, S., Kreslake, J., Ganz, O., Tacelosky, M., Abrams, D., and Vallone, D., 2012. Individual mobility patterns and real-time geo-spatial exposure to point-of-sale tobacco marketing. In *Proceedings of the ACM Wireless Health,* San Diego, CA, ACM Press.

[5] U. S. Department of Health and Human Services, 2012. *Preventing tobacco use among youth and young adults: a report of the Surgeon General.* Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

[6] Institute of Medicine, 2006. *Food Marketing to Children and Youth: Threat or Opportunity?* .

[7] Cairns, G., Angus, K., Hastings, G., and Caraher, M., 2013. Systematic reviews of the evidence on the nature, extent and effects of food marketing to children. A retrospective summary. *Appetite 62*(Mar), 209-215.

[8] Zheng, Y., Capra, L., Wolfson, O., and Yang, H., 2014. Urban computing: concepts, methodologies, and applications. In *Proceedings of the ACM TIST*, ACM Press.

[9] Kirchner, T.R., Cantrell, J., Ansetti-Rothermel, A., Ganz, O., Vallone, D.M., and Abrams, D.B., 2013. Real-time exposure to point-of-sale tobacco marketing predicts daily smoking status and smoking cessation outcomes: A multilevel geo-spatial analysis. *Am J Prev Med 45*, 4, 379-385.

[10] Ge, Y., Liu, C., Xiong, H., and Chen, J., 2011. A taxi business intelligence system. In *Proceedings of the KDD* San Diego, CA, ACM Press, 735-738.

[11] Ge, Y., Xiong, H., Tuzhilin, A., Xiao, K., Gruteser, M., and Pazzani, M.J., 2010. An energy-efficient mobile recommender system. In *Proceedings of the ACM KDD,* Washington, DC, ACM Press, 899-908.

[12] Yuan, J., Zheng, Y., Xie, X., and Sun, G., 2011. Driving with knowledge from the physical world. In *Proceedings of the ACM KDD,* San Diego, CA, ACM Press, 316-324.

[13] Yuan, J., Zheng, Y., Zhang, L., Xie, X., and Sun, G., 2011. Where to find my next passenger? In *Proceedings of the UbiComp,* Beijing, China, ACM Press.

[14] Qi, G., Li, X., Li, S., Pan, G., and Wang, Z., 2011. Measuring social functions of city regions from large-scale taxi behaviors. In *Proceedings of the IEEE PERCOM Workshops,* Seattle, WA, IEEE, 384-388.

[15] Pozdnoukhov, A. and Kaiser, C., 2011. Space-time dynamics of topics in streaming text. In *Proceedings of the ACM GIS,* Chicago, IL, ACM Press, 1-8.

[16] Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T., 2011. Geographical topic discovery and comparison. In *Proceedings of the ACM WWW,* Hyderabad, India, ACM Press, 247-256.

[17] Editorial, 2008. A flood of hard data. *Nature 453*, 7196 (Jun 5), 698.

[18] Lazer, D., Kennedy, R., King, G., and Vespignani, A., 2014. Big data. The parable of Google Flu: traps in big data analysis. *Science 343*, 6176 (Mar 14), 1203-1205.

[19] Andrew Peterson, N., Yu, D., Morton, C., and Reid, R., 2010. Tobacco outlet density and demographics at the tract level of analysis in New Jersey: a statewide analysis. *Drugs: education, prevention and policy 18*, 1, 47-52.

[20] Ogneva-Himmelberger, Y., Ross, L., Burdick, W., and Simpson, S.A., 2010. Using geographic information systems to compare the density of stores selling tobacco and alcohol: youth making an argument for increased regulation of the tobacco permitting process in Worcester, Massachusetts, USA. *Tob Control 19*, 6, 475-480.

[21] Peterson, N.A., Lowe, J.B., and Reid, R.J., 2005. Tobacco outlet density, cigarette smoking prevalence, and demographics at the county level of analysis. *Subst Use Misuse 40*, 11, 1627-1635.

[22] Reid, R.J., Peterson, N.A., Lowe, J.B., and J., H., 2005. Tobacco outlet density and smoking prevalence: Does racial concentration matter? *Drugs: education, prevention and policy 12*, 3, 233–238.

[23] Yu, D., Peterson, N.A., Sheffer, M.A., Reid, R.J., and Schnieder, J.E., 2010. Tobacco outlet density and demographics: analysing the relationships with a spatial regression approach. *Public Health 124*, 7, 412-416.

[24] Matthews, S.A., Mccarthy, J.D., and Rafail, P.S., 2011. Using ZIP code business patterns data to measure alcohol outlet density. *Addict Behav 36*, 7, 777-780.

[25] Kavanagh, A.M., Kelly, M.T., Krnjacki, L., Thornton, L., Jolley, D., Subramanian, S.V., Turrell, G., and Bentley, R.J., 2011. Access to alcohol outlets and harmful alcohol consumption: a multi-level study in Melbourne, Australia. *Addiction 106*, 10, 1772-1779.

[26] Czarnecki, K.D., Goranson, C., Ellis, J.A., Vichinsky, L.E., Coady, M.H., and Perl, S.B., 2010. Using geographic information system analyses to monitor large-scale distribution of nicotine replacement therapy in New York City. *Prev Med 50*, 5-6, 288-296.

[27] Reitzel, L.R., Cromley, E.K., Li, Y., Cao, Y., Dela Mater, R., Mazas, C.A., Cofta-Woerpel, L., Cinciripini, P.M., and Wetter, D.W., 2011. The effect of tobacco outlet density and proximity on smoking cessation. *Am J Public Health 101*, 2, 315-320.

[28] Berke, E.M., Tanski, S.E., Demidenko, E., Alford-Teaster, J., Shi, X., and Sargent, J.D., 2010. Alcohol retail density and demographic predictors of health disparities: a geographic analysis. *Am J Public Health 100*, 10, 1967-1971.

[29] Christian WJ., 2012. Using geospatial technologies to explore activity-based retail food environments. *Spat Spatiotemporal Epidemiol, 3(4)*:287-295.

[30] Hurvitz PM, Moudon AV., 2012 Home versus nonhome neighborhood: quantifying differences in exposure to the built environment. *Am J Prev Med, 42(4)*: 411-417.

[31] Rainham DG, Bates CJ, Blanchard CM, Dummer TJ, Kirk SF, Shearer CL., 2012. Spatial classification of youth physical activity patterns. *Am J Prev Med, 42(5)*: e87-e96.

[32] Krenn PJ, Titze S, Oja P, Jones A, Ogilvie D., 2011. Use of global positioning systems to study physical activity and the environment: a systematic review. *Am J Prev Med, 41(5)*: 508-515.

[33] U.S. Census Bureau, 2014. Introduction to NAICS.

[34] Carlos, H.A., Shi, X., Sargent, J., Tanski, S., and Berke, E.M., 2010. Density estimation and adaptive bandwidths: a primer for public health practitioners. *Int J Health Geogr 9*, 39.

[35] U.S. Census Bureau, 2013. ZIP Code Tabulation Areas (ZCTAs).

[36] House, B., 2013. OpenPaths: empowering personal geographic data. In *Proceedings of the ISEA,* Sydney, Australia.

[37] Vincenty, T., 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Surv Rev 23*, 176, 88-93.

[38] Kirchner, T.R. and Shiffman, S., (2013). *Ecological Momentary Assessment. The Wiley Blackwell Handbook of Addiction Psychopharmacology*. Wiley-Blackwell Publishing Ltd., New York.

[39] Agresti, A., 2012. *Categorical Data Analysis*. Wiley.