Optimal Retail Location: Empirical Methodology and Application to Practice

Marshall Fisher, Chloe Kim Glaeser, Xuanming Su
The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104

September 21, 2016

Abstract

We empirically study the spatio-temporal location problem motivated by an online retailer that uses the Buy-Online-Pick-Up-In-Store fulfillment method. Customers pick up their orders from trucks parked at specific locations during specific times, and the retailer's problem is to determine where and when these pick-ups occur. We combine demographic and economic data, business location data, and the retailer's historical sales and operations data to predict demand at potential locations. We introduce a novel two step procedure that combines machine learning and econometric techniques. First, we use a random forests algorithm to predict demand when a particular location operated in isolation. Then, we use a fixed effects regression to estimate spatial and temporal cannibalization effects that cannot be captured in the first step. Based on the predicted demand, we develop heuristics to improve the pick-up location configuration and schedule. We estimate a 36% increase in revenue from the improved location configuration and schedule.

1 Introduction

The advancement of technology and the e-commerce industry has introduced many new forms of business operations. Many traditional brick-and-mortar retailers have converted into omnichannel retailers and are attempting to capture more sales by providing more options to customers. The Buy-Online-Pick-Up-In-Store (BOPS) format is one aspect of omnichannel retail that allows customers to shop online and obtain the goods offline. This format reduces the customer wait time for goods relative to the wait time if the goods were shipped.

We partner with an online retailer that was founded on the BOPS fulfillment idea: instead of having offline stores, the retailer uses delivery trucks parked at various convenient pick-up locations, which include schools, businesses, gyms, and parking lots to deliver goods to customers. Upon placing an order, customers choose the date and location at which to pick-up their order from a list of pre-specified options. While this business format may be less convenient to the customers than the traditional brick-and-mortar format due to limited availability, the retailer can use its delivery trucks to capture a greater share of consumer demand by accessing multiple markets in a single week with a limited capital investment. These additional markets can include small towns with few traditional brick-and-mortar retailers, and in these towns this new format enhances customer convenience. As the retailer has expanded, it has created new pick-up locations and closed underperforming locations. Some of the locations are open once a week whereas others open multiple times a week. When adding a new location, the retailer considers other pick-up locations nearby, its competitors' locations, the income level of the area provided by the US census, and other relevant information based on a site visit. However, despite the use of this information, there are still large differences in the performance of locations. In fact, optimizing the operations in terms of both location and time is a low cost, yet difficult problem for many reasons: 1. the retailer needs to be able to accurately estimate potential profitability in each market; 2. the retailer needs to identify a specific location that can capture as much demand as possible within a market; 3. if the market demand is big enough for one location operating once a week, the retailer needs to determine whether it should expand in the market by operating another day of week at the same location, or operating two separate locations. Additional operations may cannibalize the sales from other locations, and therefore, the optimal course of action may not be clear. The company believes that fine tuning its pick-up locations and schedule using analytics will improve its revenue, most of which will be converted into profit. This unique setting provides high-frequency data on the performance of different locations and creates an opportunity to study the location problem in both the time and space dimensions, which we refer to as the spatio-temporal location problem throughout the paper.

We formalize the spatio-temporal location problem as an integer program where the retailer determines when and where it should operate its locations to maximize revenue. In addition, we account for two types of cannibalization effects. The first type of cannibalization happens among locations, and we refer to it as spatial cannibalization. It represents the decrease in daily revenue when there is another nearby pick-up location. Another type of cannibalization happens within

a location, and we refer to it as temporal cannibalization. It represents the decrease in a daily revenue when a location operates more than once a week. Both cannibalization effects happen due to customers who prefer the other pick-up location or time, yet would have used this location if the other option did not exist. In summary, we consider the space and time dimensions of the location problem, while accounting for cannibalization effects in each dimension.

In the operations research literature, the space dimension of location problems has been studied extensively. The prior literature includes the facility location problem (also known as the plant, warehouses, or distribution center location problem), which was studied since the early 1960's (see an early review by Krarup and Pruzan (1983) or a recent textbook by Daskin (2013)). These models assume a deterministic set of demand points and require choosing K facility locations to minimize the sum of distances (or cost to travel) from facilities to all demand locations. Although the demand for each facility is determined by its location, the total demand that will be captured is fixed. On the other hand, our model focuses on the demand (revenue) impact of a chosen set of locations, accounting for cannibalization across locations. In this spirit, out model resembles the competitive facility location problem. (see Berman et al. (2009) for a review of the competitive facility location problem literature). Another demand-focused location is the classic Hotelling model, which splits demand among locations by assuming customers care only about proximity and choose the closest location; see Hotelling (1990). The gravity model, as reviewed in Ghosh et al. (1995), assumes that each location has an attractive factor in addition to a distance decay function that determines the demand. Contrary to these previously studied spatial location problems, where demand is given and cannibalization follows from the imposed structure of the researcher's model, we estimate demand and cannibalization empirically.

The time dimension of the location problem has also been studied by researchers who have built on the traditional facility location problem model (see Owen and Daskin (1998), Snyder (2006), Melo et al. (2009) for detailed reviews). There are two main types of location problems that consider the time dimension. In dynamic location problems, the objective is to locate facilities during specified time periods within which a currently attractive location can become unattractive later. In stochastic location problems, some parameters may fluctuate from time to time; these models contain either a probability distribution of uncertain parameters or different outcomes for a set of scenarios that result from the uncertain parameters. In contrast to these other models, our model considers demand cannibalization effects, which we estimate empirically. Furthermore, we consider seasonality in various time scales (e.g. day of week, month, year) in our model.

To estimate demand for each location, we combine demographic and economics data, business location data and the retailer's historical sales and operations data. Using these the varied sources of data, we try to capture the micro factors (e.g. competitors, traffic generating businesses, etc.) and macro factors (e.g. population, income level, etc.) that affect demand at each potential location. We end up with more than 200 attributes in our dataset. With this massive data, we use random forests, a machine learning algorithm, to predict demand; see Hastie et al. (2009) for a book on statistical learning. Ferreira et al. (2015) also uses a machine learning technique, which is very similar to ours, to predict demand for an online retailer. They compare the prediction performance against five other regression models and find that over multiple measures, regression trees with bagging performs the best.

Because the location and operating schedule are carefully chosen by the retailer, we observe in our data that locations that operate twice a week or have a large number of nearby pick-up locations have historically higher sales than those locations that are operating once a week or operating in an isolated area. Therefore, both frequent operations and more nearby locations are associated with greater consumer demand. As a result, the random forests algorithm would incorrectly conclude that nearby locations and more frequent operations will lead to higher sales. However, this is the result of unobservable time-invariant attributes of the location (e.g. high latent consumer demand). To address this issue, an econometric specification with fixed effects can be used to remove these unobservable differences across locations. Therefore, we use a two-step procedure. First, we estimate demand using the random forests algorithm while excluding endogenous variables (e.g. operating frequency) chosen by the retailer. Second, we use fixed effects regression to accurately estimate the spatial and temporal cannibalization effects. This two-step procedure yields our demand prediction model.

In retail operations management, demand estimation using empirical methods has been an active research area. Fisher and Raman (1996) Fisher and Raman (1996) utilize dispersion in expert opinion and historical data to estimate demand probability distributions for fashion products. Caro and Gallien (2012) use a multiplicative regression model to predict demands for clearance prices in a fashion retail setting. Fisher and Vaidyanathan (2014) estimate demand for attribute levels and substitution probabilities of products using a maximum likelihood estimation in a retailing setting. Ferreira et al. (2015) use regression trees with a bagging method for their demand estimation. We add to this literature by presenting a combined method that utilizes both machine learning method as well as an econometrics model, to overcome limitations of machine learning methods.

Importantly, we find that our combined method outperforms both random forests and least squares regression in out of sample prediction accuracy.

Finally, we present a simple heuristic that can be easily implemented to improve retailer performance when the retailer is faced with the spatio-temporal location problem. In our setting, we estimate that our heuristic can result in a 6% to 18% increase in profit when 3 locations in each market are optimized.

We believe our method can provide a road-map for companies to improve their location selection and scheduling. One direct application would be to optimize the location selection for the growing number of Healthcare-on-wheels operations, such as ambulance location problems (e.g.Brotcorne et al. (2003)). These operations face the spatio-temporal problem and can similarly use our empirical strategy to ensure they are at the right place at the right time. More generally, our methodology can also be applied to more traditional retailers, as they must also optimize the location and operating time of their stores. The spatio-temporal problem is especially relevant for traditional retailers with a large number of stores. One example is Starbucks. In many urban areas, there are more than one Starbucks location within a 1 mile radius. Our results suggest that Starbucks and similar companies should not only consider the environmental attractiveness of potential locations, but also consider the potential cannibalization from nearby locations.

Our methodology can also be applied to retailers that only consider the spatial problem, without the temporal component (e.g., a retailer that operates 10-6 without variation, or 24-7). For example, Amazon.com offers customers the option to pick-up orders at Amazon lockers. Amazon needs to determine where and how many lockers to add to each market, but does not need to consider the operating hours of its lockers for most locations. Car-sharing businesses, such ZipCar, also need to determine their car pick-up locations. Undoubtedly, Amazon and ZipCar's problems are complicated by the availability of lockers and cars, respectively. However, their flexibility to move locations provides them with varied and high-frequency data that can be readily paired with our methodology to determine areas of high demand. Ultimately, any company that physically interacts with customers needs to determine where the point-of-contact will be and can potentially benefit from adopting our methodology.

The remainder of the paper is organized as follows. We formulate the spatio-temporal problem in a general setting and present special cases in §2. We present our prediction method in §3 and describe two algorithms to improve location-scheduling solution in §4. In §5, we discuss results and we conclude in §6.

2 The Spatio-Temporal Location Problem

2.1 Problem Formulation

A firm's spatio-temporal problem is to determine where to locate its retail locations and when to operate each location to maximize its revenue for a given week. Let $\mathbb{L}$ denote the set of all potential locations and $\mathbb{T}$ denote the set of all possible days of operation.

For each potential location $i \in \mathbb{L}$ and day (or time period) $t \in \mathbb{T}$, the firm decides whether to operate. Let $x_{it} = 1$ if firm operates location i on day t, and otherwise $x_{it} = 0$.

We formulate the spatio-temporal problem as an integer program:

$$max \sum_{i \in \mathbb{L}} \sum_{t \in \mathbb{T}} x_{it}(R_{it} - \sum_{i \neq j} S_{ij} \cdot \max_{t \in \mathbb{T}}\{x_{it}\} \cdot \max_{t \in \mathbb{T}}\{x_{jt}\} - \sum_{s \neq t} T_{ts} \cdot x_{is}) \tag{1}$$

$$s.t. \sum_{i} x_{it} \leq K, \tag{2}$$

$$x_{i,j} \in \{0, 1\} \forall i, j \in \mathbb{L}. \tag{3}$$

Next, we define the variables used in the integer program:

- R_{it}**:** The firm earns revenue of R_{it} when location i operates on day t if it is the only operating time and location.

- S_{ij}**:** If there is another location j within D miles, regardless of location j's operating day, location i's revenue on day t will be cannibalized and decreased by S_{ij}. The parameter D will be determined in our empirical analysis. This *spatial cannibalization* captures the loss of demand from customers who prefer location j over location i, yet without the presence of location j would have used location i.

- T_{ts}**:** If location i is also open on day s, location i's revenue on day t will be cannibalized and decreased by T_{ts}. This *temporal cannibalization* captures the loss of demand from customers who prefer day s over day t, yet if location i only operated on day t would have come on day t.

- K**:** The firm has limited capacity and can operate maximum of K times in a given day. If there is no resource constraint, $K = n(\mathbb{L}) * n(\mathbb{T})$.

Note that when the subscript is a location-time pairing it indicates a location-day operation.

When the subscript is a location-location or time-time pairing, it indicates cannibalization of the first subscript event by the second subscript event.

The objective function is the sum of revenue earned at each operation (i.e. location-day pairing) the firm decides to open. The second term in (1) is the sum of spatial cannibalization effects on location i, and the max terms captures the intuition that regardless of operating day of other locations, a spatial cannibalization occurs as long as two locations are open. The third term in (1) captures the temporal cannibalization that occurs when the location is open more than once a week.

2.2 Special Cases of Spatio-Temporal Location Problem

If we impose further restrictions, our spatio-temporal problem reduces to a known class of problems. In this sub-section, we present three special cases.

Case 1 (No Cannibalization Effects): Suppose $S = T = 0$.

When there is no spatial or temporal cannibalization effect, the optimal solution to the spatio-temporal problem is selecting the K operations with largest R_{it}. In this setting, a greedy algorithm finds the optimal solution. In a more general facility location problem setting, prior work finds that the greedy algorithm performs sufficiently well. (Cornuejols et al. (1977))

Case 2 (No Spatial Cannibalization Effects): Suppose $S = 0$ and $R_{it} = R_i$.

When there is no spatial cannibalization effect, the problem reduces to finding the optimal number of times to operate for each location. Due to the temporal cannibalization effect, additional operating days have diminishing marginal returns. In our formulation, the optimal solution is achieved when $R_i = T \cdot \sum_{t \in \mathbb{T}} x_{it}$ (i.e. marginal return = marginal cost) at each location.

Case 3 (No Temporal Cannibalization Effects): Suppose $T = 0$ and $R_{it} = R$.

When all potential locations provide the same profit and there is no temporal cannibalization effect, our problem reduces to a version of densest k-subgraph problem. We describe the densest k-subgraph problem and articulate how this special case of our optimization problem can be characterized as a densest k-subgraph problem. (Feige, Peleg, and Kortsarz, 2001)

Let $G(V, E)$ be an undirected graph with $|V|$ vertices and $|E|$ edges. Suppose $U \subseteq V$ and let $E(U)$ be edges in U. The density $d(U)$ is defined as $|E(U)|/|U|$. The densest k-subgraph problem is to find U^*, a subgraph of G with k vertices such that G^* is of maximum density denoted as $d^*(G, k)$.

Now suppose each vertex represents a potential location. Since there is no temporal cannibalization effect, we ignore the temporal part of the integer programming. Because all edge weights are non-negative in the dense subgraph problem, we assign an arbitrarily large weight, W, to each edge at start. As one location (i.e. vertex) is chosen to be open, the nearby locations within D miles will suffer from the spatial cannibalization if open. For the vertex representing these locations, we subtract S from the edge weight. Ultimately, the problem reduces to finding the densest K-subgraph after accounting for the spatial cannibalization by changing the edge-weights based on the vertex selection.

These special cases highlight some of the unique characteristics of our formulation, as well as the complexity of our problem setting.

3 Empirical Method

In this section, we describe our data and how we use our data to estimate revenue, spatial cannibalization, and temporal cannibalization for potential locations. We were asked to ensure confidentiality of the retailer's data, and to avoid disclosure we present the retailer specific statistics and results scaled by a factor. Specifically, any results regarding a sales figure is linearly transformed.

3.1 Data

We combine three different sources of data. The first source is our partner retailer's sales transaction data from January 2014 to December 2015. For each pick-up location that operated, we have the following variables: location ID, location name, location type (e.g. business, school, etc.), total sales, date, day of week, latitude and longitude. From this data, we can also track for each location-day operation the number of other pick-up locations operated in the same week within 0.3, 0.5, 1, 3, 5, 7 and 10 miles radii and also for various sized circular rings (i.e., annuli). Note that only a few observations operate on the same day and in close proximity to one another. More commonly, nearby locations will operate on different days of the week (e.g. Monday and Thursday). Therefore, in our empirical model, we allow nearby locations that operate in the same week to have a cannibalization effect on one another. This reflects the intuition that a location operating on Thursday will cannibalize sales from a nearby location that operates on Monday.

The retailer also offers home delivery and office delivery services. Home delivery is for one order and office delivery is for multiple orders. To capture a potential cannibalization and/or word-of-

mouth effect from these operations, we also track whether or not home delivery was available in each ZIP code for each week, and the number of office delivery locations operating on the same week within various radii and various sized circular rings.

The second source is the U.S. Census Bureau's demographic and economic data which provides macro environment factors. These data are available in two publicly accessible databases. The first is the American Community Survey (ACS) 5-year Estimates in block group geo-database. Block groups are the most granular data that ACS is provided and they generally contain between 600 and 3000 people. (Bureau of the Census (1994))

The ACS contains over 19000 statistics on average age, sex, race, place of birth, education attainment, marital status, family status, income level, housing value and more. Of these, we use the following variables: *total population, female population, male population, population with post-secondary degree, number of households, number of households with minor, number of households with income greater than $75,000, median age, median housing value, median income and mean income*. We choose to focus on these variables after talking to our partner retailer on their target market and the characteristics it looks for when selecting locations.

The other U.S. Census Bureau database we use is the County Business Patterns (CBP) in ZIP code level database for the last three years. In this data set, ZIP code is the smallest geographic entity for which business pattern data are available. The database contains the employment size for each ZIP code. For the ZIP codes where this value is suppressed for confidentiality and is provided as a range of the employment size, we use the median of the range. For each ZIP code, we then compute the labor density in square miles based on this data.

The third source we use is OpenStreetMap which provides data on business locations. OpenStreetMap offers user-contributed business location data for free. More details can be found on `http://www.openstreetmap.org/`. We decide which business locations to use in our prediction model based largely on which micro factors the retailer indicated they are most affected by. Based on these conversation, we use 7 different types of locations: school, university, kindergarten, church, Starbucks, our retailer's competitors (e.g. department stores), a specific direct competitor (e.g. Nordstrom). We decide to focus on these locations for three reasons. First, these data were more accurate than some other location types, such as gym or yoga studio, when cross-validated with other map data. Second, many of these location types are used as a pick-up location because consumers visit them on a regular basis. Third, we want to capture the effect of location competition on location performance.

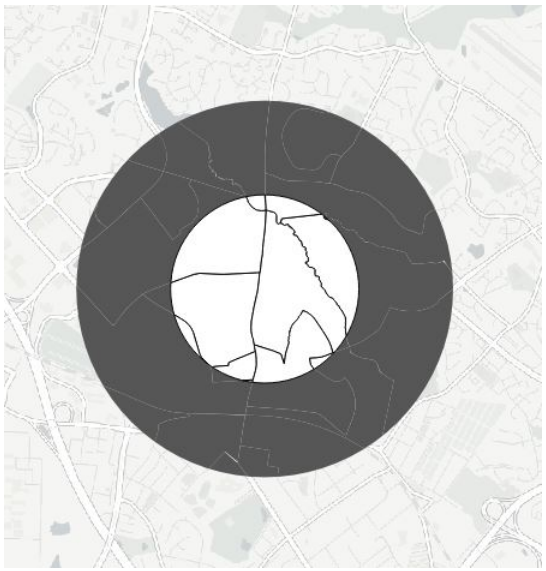

Figure 1: Surrounding Area Partitioned by Block Groups For A Location. The smaller circle in white is a 0.5 miles radius and the outer ring in grey represents an area from 0.5 miles to 1 mile radius from a location.

We then combine these data to create a panel data set. For demographic data and nearby business location data, we determine how each attribute will be measured. For demographic information, we compute the area-weighted mean of each variable within 0.5, 1, 3, and 5 miles radii and various sized circular rings for each location. Similarly, for business location data, we compute the number of each type of business locations within various radii and in various sized circular rings for each pick-up location. These data were mainly processed using a open-source geographical information system application called QGIS.

To depict this process, we include a map of a location in Figure 1 with two defined areas: a 0.5 miles radius circle and a 0.5 miles to 1 mile radius circular ring. The areas are partitioned by the block group sectioning and confined by the defined radii. Each block group includes its demographic and economic attributes that we described earlier. These attributes are weighted by the area of block group. By summing these weighted attribute for all block groups within the defined area, we calculate surrounding area attributes for each location. For example, to calculate attributes for the 0.5 miles radius circle specific to this location, we process 12 data points. Note that this location is in a suburban area and some 0.5 miles radius circle areas for locations in an urban area can consist of over 30 blockgroups and obviously a much greater number of data points for larger surrounding areas.

In total, for each location-day operation, we use 205 explanatory variables to predict our de-

Data Source	Variable	Definition
Retailer's sales and operations data	Sales	Total dollar sales from location-day opeartion
	Location type	Type of pick-up location (e.g. Business, School, etc.)
	Day of week	Day of week
	Month	Month
	Year	Year
	Home delivery availability	1, if home delivery service was available in the zip code of the location this week; 0, otherwise
	Office delivery	# of office delivery/pick-up locations within 0.3, 0.5, 1, 3, 5, 7 and 10 miles radii and in 0.3 to 0.5, 0.5 to 1, 1 to 3, 3 to 5, 5 to 7 and 7 to 10 miles circular rings of the location
	Nearby pick-up locations	
	Operating Frequency	Number of operating days of this location this week
US Census Bureau - American Community Survey 5 Year Estimates	Total population	Area weighted average of each variable within 0.5, 1, 3, 5 miles radii and in 0.5 to 1, 1 to 3, 3 to 5 miles circular rings of the location
	Female population	
	Male population	
	Population with post-secondary degree	
	Households	
	Households with minor	
	Households with income greater than $75000	
	Median age	
	Median housing value	
	Median income	
	Mean income	
US Census Bureau - County Business Patterns	Labor density	Number of employees per square miles in zipcode of the location
	Population density	Number of residents per square miles in zipcode of the location
	Business district	1, if labor density is greater than population density; 0, otherwise
Open Street Map	School	# of office delivery/pick-up locations operated within0.5, 1, 3, 5, 7, 10 and 15 miles radii and in 0.5 to 1, 1 to 3, 3 to 5, 5 to 7, 7 to 10 and 10 to 15 miles circular rings of the location
	University	
	Kindergarten	
	Church	
	Starbucks	
	Competitors	
	Direct Competitor	

Figure 2: Summary of Variables

pendent variable, sales. The variables are summarized in Figure 2.

3.2 Revenue Prediction Using Machine Learning: Estimating R

To predict location revenue, we use random forests which is a machine learning algorithm developed by Breiman (2001). In this subsection, we briefly describe the algorithm and discuss the features of the algorithm relevant to our setting.

From the training data, the random forests algorithm draws a default of 500 bootstrap samples (this default number can be changed depending on the data; on our setting, 500 samples did just as well as 1000 samples, and hence, we use 500 samples.) For each sample, the algorithm grows

a regression tree. Regression trees recursively partition the data based on the attributes. The mean response in each partitioned group becomes the prediction value for the regression tree. The prediction value from the random forests is an average prediction from all 500 regression trees. Further information can be found on Breiman (2001) and Liaw and Wiener (2002).

We choose to use random forests because they are known to be more accurate than other machine learning algorithms or regression models. Ferreira et al. (2015) find that the regression tree with bagging method, on which random forests is based, predicts sales more accurately when compared to five other regression methods. We also find that random forests predict better than least square regressions (we share this result in §3.4).

Furthermore, random forests can run on large databases efficiently, and it can also incorporate a large number of explanatory variables without variable deletion. This latter feature is important in our setting, as many of our variables are linearly dependent: the number of nearby locations in a circular ring is defined as a linear combination of the number of nearby locations in the larger radius and the number in a smaller radius. As a result, in a linear regression, these variables will be multi-collinear and one will be excluded from the empirical model. However, in reality, these variables may be important determinants of a location's success.

Last but not least, random forests is a non-parametric method. As mentioned in Ferreira et al. (2015) , this improves prediction accuracy when there are nonmonotonic relationships between the response variable and explanatory variables. This is precisely the case for many of our independent variables. For example, if a location is in a low demand area where there are no other competitors, the sales performance may be worse than when having one or two competitors in a high demand area. However, if the market is overcrowded by too many competitors, the sales performance would suffer. Similar arguments can apply for demographic variables. Since the company is an online retailer, a location in an area with many retirees (i.e. high median age) may not perform as well. However, since the retailer sells luxury goods, a location in a college town (i.e. low median age) may not perform well either. The ideal target market is then perhaps somewhere in between these groups.

One commonly criticized feature of random forests is the lack of interpretability given that it uses a large number of explanatory variables and the prediction values are based on a large number of different regression trees. Despite this downside, we still prefer using random forests for its superior prediction accuracy.

A perhaps less discussed feature of machine learning, which is very important in our setting,

is the inability to address endogeneity. In our setting, two variables suffer from omitted variable bias: *Operating frequency* and *Number of nearby pick-up locations*. Both these terms are positively correlated with market demand, and market demand is also positively correlated with sales. In fact, when we look at the raw data, we see that locations that operate more than once and locations that have other nearby locations have higher sale than those locations that operate only once and have no other nearby locations. As a result, the random forests associate an increase in *Operating frequency* and *Number of nearby pick-up locations* with an increase in sales. However, this is clearly not true as sales would decrease due to cannibalization effects. Therefore, we remove operating frequency and number of nearby pick-up locations variables from the random forests model and proceed with the prediction using 190 explanatory variables.

3.3 Cannibalization Effects Estimation Using Fixed Effects Regression: Estimating S & T

Because the revenue prediction for potential locations does not account for the cannibalization effects of *Operating frequency* and *Number of nearby pick-up locations*, we use a fixed effects regression to estimate these cannibalization effects. Our strategy builds on prior work that has use fixed or random effects to address endogenous effects. For example, Cachon and Olivares (2010) use fixed effect estimation to address endogeneity in their automobile data. Similarly, Rajagopalan (2013) uses random effects estimation to address omitted variable bias in his data. The author states that he could not use fixed effects estimation due to the short time span of his data. Because the time span of our data is longer, we are able to use fixed effects estimation to address bias in our setting.

The idea is simple. The intrinsic market demand for a location does not change after controlling for time-varying location features (e.g. time trends and seasonality). Therefore, by controlling for location with fixed effects, we can effectively capture the cannibalization effects. In other words, if a location changed operating frequency or if the number of nearby locations changed throughout its lifetime, we can capture the average effect of those changes for all locations after controlling for seasonality and day of week.

Because it is practically impossible to estimate spatial and temporal cannibalization for every potential location, we assume that the spatial cannibalization and temporal cannibalization parameters are the same across all locations and time. Note that varying cannibalization effects can be addressed by solving the spatio-temporal problem separately on the level at which the problem

varies. For example, if we found evidence that cannibalization effects vary across markets, we could address this issue by solving the spatio-temporal location problem separately in each market. In our setting, the data at each market level is too sparse to estimate the cannibalization effects robustly.

We assume that the spatial cannibalization effect decreases as the distance between two locations increases. This assumption is similar to that of the gravity model. A notable difference is that we assume a constant cannibalization effect within a range of distances. More specifically, for each specified distance range, there is a constant spatial cannibalization effect. Furthermore, there is a threshold distance beyond which the spatial cannibalization no longer occurs. We empirically determine the thresholds and estimate the corresponding spatial cannibalization effects.

Similarly, we postulate that the temporal cannibalization effect varies based on how recent the location was open. This is especially applicable in our setting because the retailer sells consumable goods and the customers may want to shop more than once a week. The retailer does not change the location configuration or operation on a weekly basis, and therefore, we assume that the location operating days are the same in the previous week for modeling purposes. For example, suppose a location is open twice a week, on Monday and Thursday. For the Monday location, the last open day is Thursday the week before, and it would be 4 days since the location was open. For the Thursday location, the last open day is Monday, and it would be 3 days since the location was open. Going forward, we will call this duration as recency duration. We speculate that the temporal cannibalization effect increases as the recency duration decreases and show that indeed our hypothesis is supported in the data.

Now we present our model specification to estimate the cannibalization effects:

$$Sale_{ijts} = \alpha_0 + \sum_{\delta} \beta_\delta NearbyLocations_{it}^{\delta} + \sum_{\tau} \gamma_\tau i.RecencyDuration_{it}^{\tau} + \sum_{i} \upsilon_i Location_i + \sum_{j} \omega_j DayOfWeek_j + \sum_{s} \psi_s YearMonth_s + \epsilon_{ijts}. \quad (4)$$

The subscript i indicates *Location*, j indicates *Day of Week*, t indicates *Date* and s indicates *YearMonth*. The variable, *Nearby Locations*, is continuous and allow us to estimate the spatial cannibalization in each distance range. We use 0.5, 1, 3, 5, and 7 miles as a threshold for each range (i.e. within 0.5 miles, from 0.5 miles to 1 mile, etc.). The *Recency Duration* variables are binary variables for τ from 1 to 7 and indicate how many days ago the location was open most recently. For example, if the location was open 3 days ago, $i.RecencyDuration^3 = 1$ and all other

$i.RecencyDuration^{\tau} = 0$. Since we are interested in the temporal cannibalization within a week, we group any observations with recency duration greater than 7 days (e.g. a location opens once a week, but did not open on a particular week due to a public holiday) with the observations with 7 days recency duration. We use this group as the base line for the regression results. Therefore, the *Recency Duration* variables capture the change in sales from once a week operation to when the most recent operation of the location was less than 7 days ago. The three control variables, *Location*, *Day Of Week* and *YearMonth* are all binary variables and control for location-specific effects, day-of-week effects, and seasonality and sales growth effects that vary by month and year, respectively. While not explicit in the model specification, we omit one binary variable for each control to avoid perfect multicollinearity.

The regression result for the specification (6) is summarized in (a) of Table 1. We use *Total Sales* as a dependent variable. First, the estimates for *Nearby Locations* variables suggest that the spatial cannibalization effect is highest within 0.5 miles and the estimated effect for having one additional nearby location within 0.5 miles is −$108.68. The spatial cannibalization effect from 1 mile to 3 miles is −$21.40 which is approximately one-fifth of the the effect for within 0.5 miles. The other nearby variables are statistically insignificant, suggesting that the spatial cannibalization effect disappears beyond a 3 miles radius.

Turning to the *Recency Duration* variables, we find that the temporal cannibalization effect is much greater when a location was open 2 days ago, −$549.95, compared to the effect for longer recency durations (estimates ranging from −$188.82 to −$376.02). Based on a F-test of the difference of coefficients, none of the coefficients of the *Recency Duration* variables for 4 to 6 days are statistically different from one another, nor are the 3 and 6 days recency duration variables statistically different from one another. Therefore, we conclude that the temporal cannibalization effect drops significantly when the recency duration is 3 days or longer. Note that there was only one observation that was open on consecutive days and as a result, we are not able to effectively estimate the effect of temporal cannibalization for a one day recency duration. However, this is because such durations are never used by the retailer, and not because of a deficiency in our modeling strategy.

In order to capture the cannibalization effects parsimoniously, we run another specification with

Table 1: Spatial and Temporal Cannibalization Effect.

Model	(a)	(b)
Nearby locations within 0.5 miles	-108.68***	-99.53***
	(34.150)	(34.340)
Nearby locations from 0.5 miles to 1 mile	-10.29	
	(17.366)	
Nearby locations from 1 mile to 3 miles	-21.40***	
	(4.652)	
Nearby locations from 3 miles to 5 miles	-1.24	
	(6.075)	
Nearby locations from 5 miles to 7 miles	3.14	
	(5.545)	
Nearby locations from 0.5 miles to 3 miles		-20.40***
		(4.151)
i.Recency duration: 1 day	-37.49	
	(40.885)	
i.Recency duration: 2 days	-549.95***	
	(66.438)	
i.Recency duration: 3 days	-188.82***	
	(51.649)	
i.Recency duration: 4 days	-376.02***	
	(50.385)	
i.Recency duration: 5 days	-321.11***	
	(76.840)	
i.Recency duration: 6 days	-288.29***	
	(103.091)	
i.Recency duration < 3 days (if operating twice a week)		-589.09***
		(66.487)
i.Recency duration ≥ 3 days (if operating twice a week)		-244.46***
		(40.414)
Location Control	Y	Y
MonthYear Control	Y	Y
Day of Week Control	Y	Y
No. of observations	12,803	12,803
R-squared	0.650	0.647

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

new variables:

$$Sale_{ijts} = \alpha_0 + \sum_{\delta} \beta_\delta NearbyLocations_{it}^{\delta} + \Gamma_1 i.RecencyDurationLessThan3 + \Gamma_2 i.RecencyDurationGreaterThan2 + \sum_i \upsilon_i Location_i + \sum_j \omega_j DayOfWeek_j + \sum_s \psi_s YearMonth_s + \epsilon_{ijts}. \quad (5)$$

In this specification, we include continuous variables capturing the number of nearby locations within 0.5 miles and from 0.5 miles to 3 miles. We do not include farther distances based on the evidence in column (a) that locations which are more than 3 miles away do not have a cannibalization effect. We introduce two new indicator variables to capture the temporal cannibalization effect for locations that are open twice a week. This is based on the evidence in column (a) that there are two levels of cannibalization effects. The *Recency Duration Less Than 3* variable indicates whether the location's recency duration is 1 or 2 days and the *Recency Duration Greater Than 2* variable indicates whether the location's recency duration is 3 days or longer.

The regression result of our final model is summarized in Table 1 column (b) . $TotalSales$ is the dependent variable. Our estimates suggest that having one additional location within 0.5 miles radius results in a sales decrease of \$99.53. When this range changes to 0.5 miles to 3 miles, the spatial cannibalization effect for an additional location becomes −\$20.40. In addition, there is a notable difference between the temporal cannibalization when the recency duration is less than 3 days and is greater than equal to 3 days. These estimates suggest that when the location is open twice a week and the recency duration is less than 3 days, sales decrease by \$589.09. Similarly, when the location is open twice a week and the recency duration is greater than equal to 3 days, sales decrease by \$244.46.

As a robustness check, we estimate the temporal cannibalization effect of operating twice a week using the random forests method. We use only the observations that operated once a week in the training set. Using this random forests model, we estimate sales for observations that operated twice a week. The model estimates sales for these observations as if they operated once a week, and therefore when we compare the prediction result against the actual sales of these observations, the average over-prediction can be interpreted as an estimate for the temporal cannibalization effect. The average over-prediction is \$111.72, which is smaller than what we obtain in the fixed effects regression.

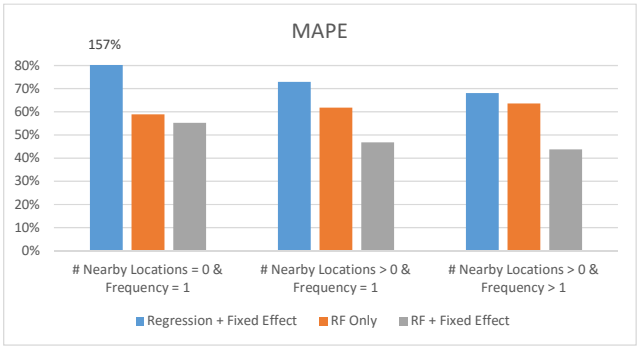

Figure 3: Accuracy Comparisons of Three Models.

3.4 Accuracy Comparisons

In this subsection, we present the out-of-sample prediction accuracy of our estimation method and compare the performance against two other models: random forests model with all explanatory variables including *Operating frequency* and *Number of nearby pick-up locations*, and least square regression model without *Operating frequency* and *Number of nearby pick-up locations*, and later adjusted based on the fixed effects regression.

We use the first 18 months in our data as a training set and the last 6 months as a validation set. We use the cannibalization estimates from the training set using the fixed effects specification (7) in §3.3. In Figure 3, we present the accuracy comparisons using Mean Absolute Percent Error (MAPE). When we use other measurements such as Median Absolute Percent Error (MdAPE), Mean Absolute Error (MAE), and Median Absolute Error (MdAE), the results are similar. We group the observations in the validation set into three categories: locations that operate once a week and have no nearby locations, locations that operate once a week and have nearby locations, and locations that operate twice a week and have nearby locations (there was no observation where it operated twice a week and had no nearby locations). Note that the axis of the graph is manually adjusted for ease of comparison, and the MAPE for no nearby locations and operating once a week observations (the first blue bar in the graph) is 157%.

First, we note that both random forest methods outperform least square regressions significantly in all three categories, but especially in the first category to which most observations belong. In all three categories, the combined method model outperform both the regression model and the random forests only model.

Finally, we point out that this accuracy comparison is a conservative measure on the benefit that the fixed effect regression provides since the validation set is just as endogenous as the training

set (i.e. the retailer strategically chose to add additional operations in a high demand area). As a result, random forests with the endogenous variables do not suffer from the incorrect association between frequency or nearby locations and sales as much as randomly chosen locations. In other words, predicting high demand when more operations in an area is observed would yield a rather accurate forecast. However, as we stressed, this will not result in an accurate forecast when we apply it to randomly selected potential locations.

4 Potential Location Set Generation and Heuristic Improvement

Given the set of potential locations and the operation performance predictions from our empirical model, we present a greedy algorithm and an interchange algorithm to improve the retailers pick-up location operation in both the time and space dimensions. The most valuable feature of our prediction method is that it can be applied to any potential location that needs to be evaluated as long as we can generate the attributes used in the prediction model. In order to estimate the profitability of potential locations, we create a grid of points in 0.4 mile increments across the states that the retailer is operating in. Of these locations, we filter out the ones that do not have target market characteristics based on density and median income level. In addition, to prevent inappropriate extrapolation, we only consider those locations that are within 10 miles of historical locations. This generates 56,443 potential locations.

For each potential location, we create six feasible day events and predict revenue based on the location's attributes. Assuming that the retailers do not have any locations operating (i.e. all potential location will operate once a week and there are no other locations nearby), we predict each locations performance using a random forests model. This yields a revenue prediction map which can guide the retailer when evaluating potential new markets to enter. Figure 4 shows a snapshot of an area with potential locations in circles and actual locations in stars. The color map is based on the percentile of the average actual sales for a particular month. Despite the fact that the potential locations are out-of-sample predictions for that month and are not accounting for operating frequency and nearby locations, we can see that most predictions are aligned with the actual performance observed.

Given these predictions and cannibalization estimations, we use a greedy algorithm to identify an improved location configuration and operating schedule for N number of operations.

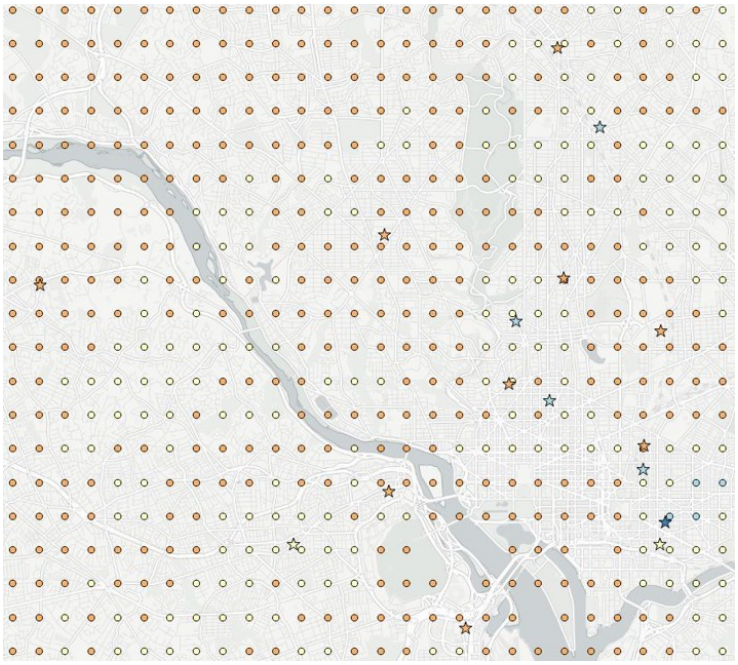

Figure 4: Initial Prediction Against Actual Sale.

Greedy Algorithm:

1. Predict each potential locations daily revenue from Monday to Saturday using random forests.

3. Subtract the cannibalization effects from the applicable predictions.

4. Determine the new location operations with the highest predicted revenue.

5. If this is N^{th} iteration, terminate. Otherwise, go back to step 2.

This algorithm can be applied when the retailer is looking to enter a brand new market. For the existing markets, our partner retailer wanted to limit the scope of potential changes. Therefore, we also develop an interchange algorithm to close N currently operating locations and open N new locations. These additional operations could be a new physical location or an additional operating day in the week of a currently operating location.

Interchange Algorithm:

1. Update the potential locations to reflect current operation based on distance.

2. Predict each location's daily revenue from Monday to Saturday using random forests.

3. Subtract the cannibalization effects from the applicable predictions.

4. Determine the new location operation with highest predicted revenue.

5. Compare 4 with the lowest prediction for operating locations.

6. If 5 < 4, close the operating location and open the new location. Otherwise, terminate the

algorithm.

7. If this is N^{th} iteration, terminate. Otherwise, go back to step 3.

Note that our setting is characterized by many of the features that Zanakis and Evans (1981) suggest lend themselves to the use of heuristics instead of solving for an exact solution. Most notably, we are optimizing over predictions and oftentimes, the prediction errors will outweigh any improvement in the algorithm despite the use of a highly accurate method. In addition, our heuristic still yields a significant improvement in predicted sales over the current configuration, and it is easy to understand for anyone such that it can provide confidence for our partner retailer.

5 Results

5.1 Algorithm Performance and Choices

To estimate the retailer's expected benefits of implementing our heuristics, we obtain the recent location configuration and schedule for a given week, as well as the number of trucks available to each fulfillment center. We find the improved configurations using the greedy algorithm and the interchange algorithm, keeping the number of operations in each market and the number of trucks available to each fulfillment center the same as actual. The number of trucks add a constraint on the maximum number of locations that can operate each day in the areas serviced by the corresponding fulfillment center. We then compare the predicted revenue from the improved configurations against the actual sales earned in that given week in Figure 5.

The greedy algorithm results in a 34.8% estimated increase in a weekly revenue. We also ran the interchange algorithm beginning with the optimal configuration based on the greedy algorithm, but did not find any changes that result in an increase in revenue.

The interchange algorithm starting from the current configuration results in a 35.6% increase in weekly revenue after 35 interchanges. Additional iterations beyond the 35th do not yield any improvement. Note that one round of interchange closes and opens one operation in each area serviced by a fulfillment center. The starting point of the interchange algorithm is lower due to the prediction error.

The expected improvements from the greedy algorithm and the interchange algorithm have less than a percent difference and we believe that they are close to the true optimal. Of the selected locations, the two algorithms share 73%. Finally, we emphasize that any realized benefit from implementation will directly transfer to the bottom line since this is a very low-cost fix.

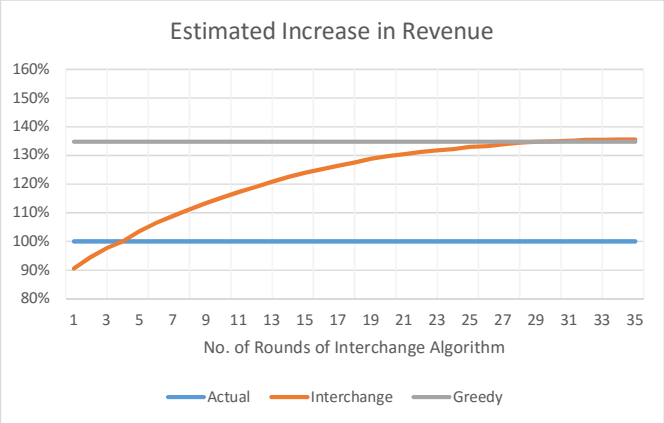

Figure 5: Algorithm Performance.

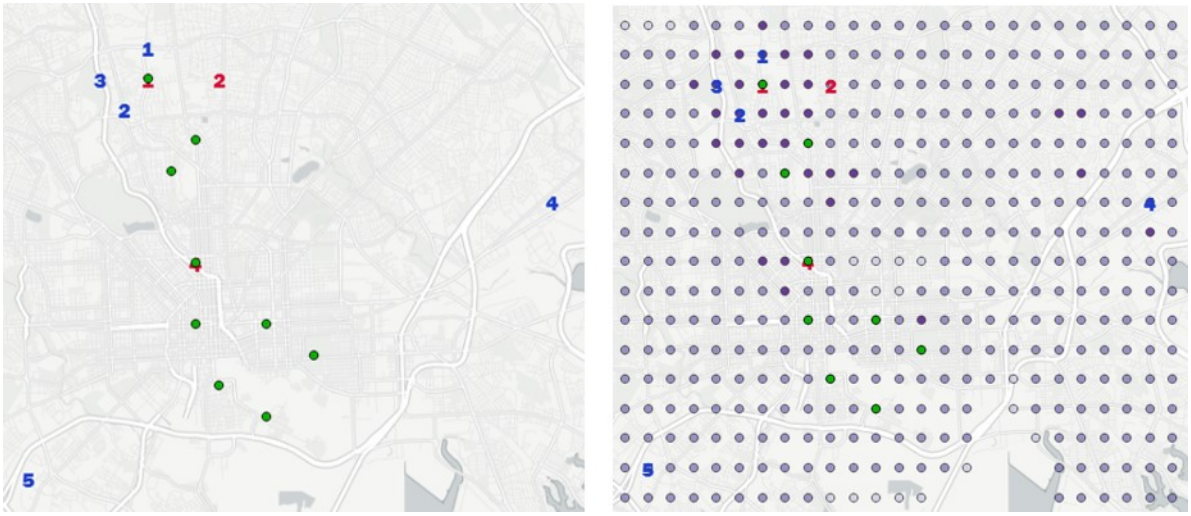

Figure 6: Interchange Algorithm for 5 Rounds: Market 1 (Partial View).

Next, we examine the location changes suggested by the interchange algorithms. Figure 6 and 7 show the interchange algorithm suggestions in two different markets. The green dots represent the currently operating locations. The red and blue numbers represent the locations selected to be closed and opened, respectively, in the corresponding round of the interchange algorithm. We overlay a heat map representing the predicted sales for each location (without accounting for the cannibalization effects) on the right hand side maps. The locations with higher predicted revenue are darker in the heat map.

Figure 6 provides a partial snapshot of Market 1. There are a number of locations that operate twice a week. We see that in the first round a location in the top left corner is chosen to be closed . This location was previously open twice a week. After the closure, the location now operates

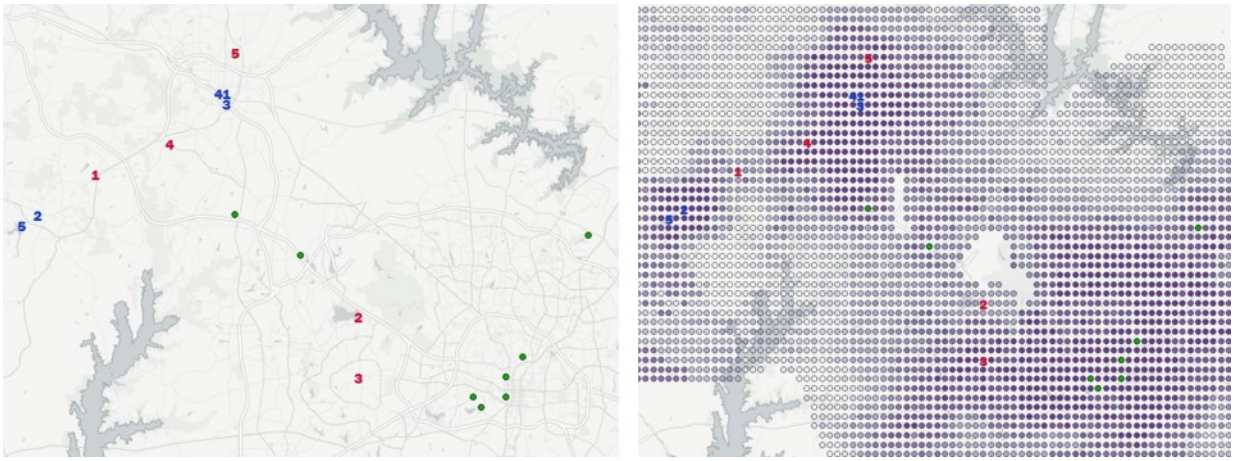

Figure 7: Interchange Algorithm Results for 5 Rounds: Market 2.

once a week. In such a situation, we see that there is a green dot overlaying on a red number. A location right above is chosen to be open instead in round 1. This is not surprising as our estimates for temporal cannibalization effects are much greater than those for spatial cannibalization effects. Additional locations are selected to open in the top left corner because predicted revenues are high. These new locations are selected to be locations more than 0.5 miles away from each other in round 2 and 3. The other closed locations either operate in a low predicted sales area or operate twice a week.

We see a similar pattern in Figure 7. The locations in the low predicted sales areas are moved to the areas with high predicted sales. This suggest that the retailer can improve profits by focusing on high performing areas rather than by exploring new areas. Consistent with this interpretation, the company's management team agrees that the algorithm's suggestions are in line with their location selection strategy going forward.

5.2 Varying The Cannibalization Estimates

In § 3.3, we stated that some of the spatial and temporal cannibalization effects (e.g. *1 Day Recency Duration*) could not be estimated due to lack of data, and we had to extrapolate the estimation by grouping the 1 and 2 days recency durations together. Like any other regression estimates, these cannibalization effect estimates are not definitive. In this subsection, we illustrate how the estimates affect the location configuration.

Figure 8 on the left contains a spatio-temporal configuration for a selected market when we define the cannibalization effects to be the upper bound of the 95% confidence interval. The figure

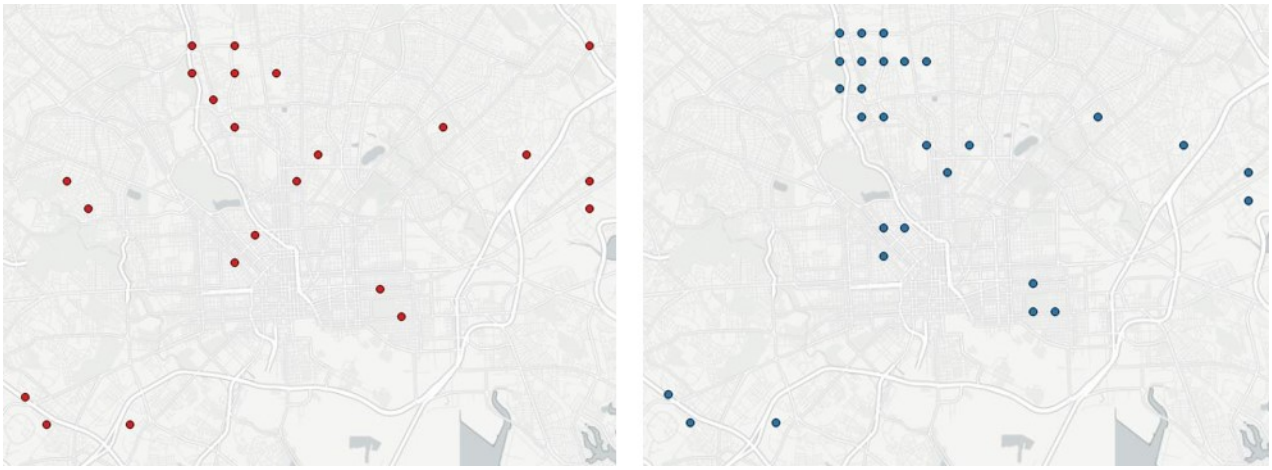

Figure 8: Spatio-Temporal Configuration Based on a Greedy Algorithm. High Cannibalization vs. Low Cannibalization.

on the right contains a spatio-temporal configuration when we define the cannibalization effects to be the lower bounds of the 95% confidence intervals. These configurations are constructed using a greedy algorithm. We see that when spatial and temporal cannibalization effects are large (i.e. in the left hand graph), the locations are less concentrated and operate less frequently. In contrast, when the spatial and temporal cannibalization effects are small (i.e. in the right hand graph), the locations are more concentrated and operate more frequently in the high demand area.

5.3 Profit Maximization

So far, we have focused on maximizing the retailer's revenue since there is no location-specific variable costs associated with operating pick-up locations. One relevant issue that the retailer wishes to address is identifying the optimal number of locations for each market. In this subsection, we examine how the operating cost factors into the profit maximization. The fixed cost associated with purchasing a delivery truck is not accounted for in this analysis.

We apply the greedy algorithm to determine the location configuration and schedule for each county, for up to 102 operations. We present results for three counties serviced by a same fulfillment center. On the left hand side of Figure 9, we present how predicted profit changes as we add operations. On the right hand side of Figure 9, we present how the heuristic optimal number of operations (i.e. the y axis) changes as we change variable costs from \$400 to \$900 (i.e. the x-axis).

First, we see that within an area (i.e. county), adding operations increases profits initially, but due to the temporal and spatial cannibalization effects, adding operations eventually decreases

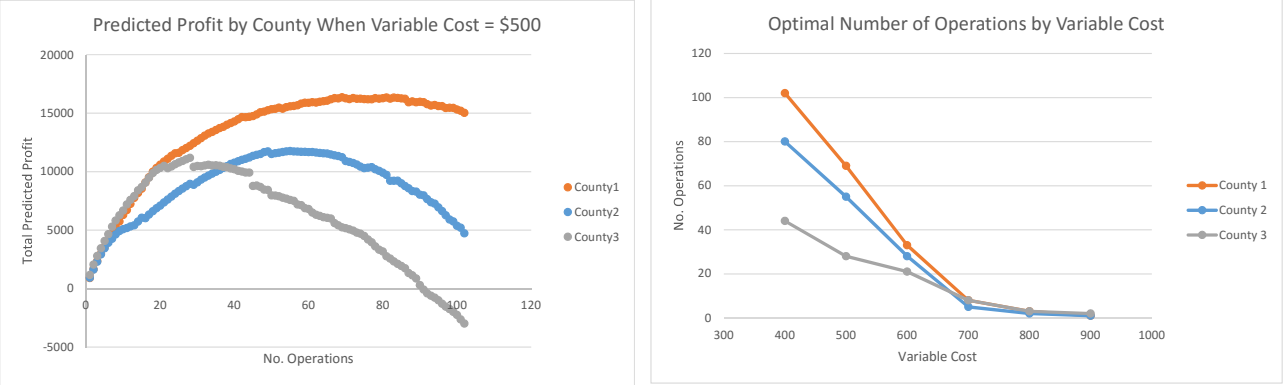

Figure 9: Profit Maximization.

profit. The optimal number of operations is determine by the average predicted revenue for all potential locations in the area and the size of the area. As the average revenue predictions and the size of the county increase, the optimal number of operations increases (given a constant variable cost). Both County 1 and County 2 have greater average revenue predictions than County 3 (106.5% and 105.9% of the average revenue in County 3 respectively). In addition, they are also larger counties than County 3. County 1 is 2.9 times the size of County 3 and County 2 is 1.3 times the size of County 3. Not surprisingly, the optimal number of operations decreases as variable costs increase. As the variable costs approaches the average revenue of the selected locations, the optimal number of operations approaches 0.

6 Conclusion

In this paper, we present our collaborated work with an online retailer on improving spatio-temporal location configuration. We demonstrate how retailers can utilize big data by collecting external data that are freely accessible and combining them with the retailer's sales and operations data to predict the performance of potential locations. We use random forests as our main prediction model for its superior prediction accuracy. However, random forests cannot address endogeniety issues present in our data. In order to circumvent this issue, we use a novel two-step procedure: first, we predict sales using exogenous variables and second, we estimate the effect of endogenous variables using a fixed effects regression and combine two results. We present a greedy algorithm and an interchange algorithm to improve location structure and operating frequency, that predicts a revenue increase of up to 36%. Since these location changes are not costly, most increases in the revenue will directly result in increases in profits.

Based on the results of our heuristics, we generate a concrete set of recommendations to our partner retailer's operation. In addition, we present a numerical analysis of profit maximization for a range of operating costs for each location. We are continuing our collaboration with the retailer to implement and validate our results. This is a lengthy process since the company has to find appropriate locations near the algorithm selected locations. When the actual location search ends, we can first predict the weekly revenue based on the changes in the configuration and validate against the total actual sales after controlling for the sales growth rate.

We believe our work can be used as an empirical blueprint for many retailers to evaluate new markets and potential locations. Furthermore, our work highlights the abundance of accessible external data and geographic information systems that allows us to study location problems in new ways.

References

Berman O, Drezner T, Drezner Z, Krass D (2009) Modeling competitive facility location problems: New approaches and results. *TutORials in Operations Research. INFORMS Annual Meeting: San Diego CA*, 156–181.

Breiman L (2001) Random forests. *Machine learning* 45(1):5–32.

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. *European journal of operational research* 147(3):451–463.

Bureau of the Census (1994) Geographic areas reference manual. `http://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf`.

Cachon GP, Olivares M (2010) Drivers of finished-goods inventory in the us automobile industry. *Management Science* 56(1):202–216.

Caro F, Gallien J (2012) Clearance pricing optimization for a fast-fashion retailer. *Operations Research* 60(6):1404–1422.

Cornuejols G, Fisher ML, Nemhauser GL (1977) Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management science* 23(8):789–810.

Daskin MS (2013) *Network and Discrete Location: Models, Algorithms, and Applications* (John Wiley & Sons).

Ferreira KJ, Lee BHA, Simchi-Levi D (2015) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* 18(1):69–88.

Fisher M, Raman A (1996) Reducing the cost of demand uncertainty through accurate response to early sales. *Operations research* 44(1):87–99.

Fisher M, Vaidyanathan R (2014) A demand estimation procedure for retail assortment optimization with results from implementations. *Management Science* 60(10):2401–2415.

Ghosh A, McLafferty S, Craig CS (1995) Multifacility retail networks. *Facility Location–A Survey of Applications and Methods. Berlin ua O* 301–330.

Hastie T, Tibshirani R, Friedman J (2009) Unsupervised learning. *The elements of statistical learning*, 485–585 (Springer).

Hotelling H (1990) Stability in competition. *The Collected Economics Articles of Harold Hotelling*, 50–63 (Springer).

Krarup J, Pruzan PM (1983) The simple plant location problem: survey and synthesis. *European Journal of Operational Research* 12(1):36–81.

Liaw A, Wiener M (2002) Classification and regression by randomforest. *R news* 2(3):18–22.

Melo MT, Nickel S, Saldanha-da Gama F (2009) Facility location and supply chain management–a review. *European journal of operational research* 196(2):401–412.

Owen SH, Daskin MS (1998) Strategic facility location: A review. *European Journal of Operational Research* 111(3):423–447.

Rajagopalan S (2013) Impact of variety and distribution system characteristics on inventory levels at us retailers. *Manufacturing & Service Operations Management* 15(2):191–204.

Snyder LV (2006) Facility location under uncertainty: a review. *IIE Transactions* 38(7):547–564.

Zanakis SH, Evans JR (1981) Heuristic optimization: Why, when, and how to use it. *Interfaces* 11(5):84–91.