Appendix A: Summary of Variables

Table 1 Summary of Variables.

Data Source	Variable	Definition
Retailer's sales and operations data	Sales	Total dollar sales from location-day opeartion
	Coupons	Total dollar amount of coupons redeemed as part of sales
	Location tenure	Number of months that the location has been operating
	Fulfillment center	Fulfillment center for the pick-up location
	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 0.5, 1, 3, 5, 7 and 10 miles radii and in 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	
	Tenure of nearby pick-up locations	Maximum, average and minimum tenure of nearby pick-up locations within 0.5, 1, 3, 5, 7 and 10 miles radii and in 0.5 to 1, 1 to 3, 3 to 5, 5 to 7 and 7 to 10 miles circular rings of the location
	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	Number of business/organizations within 0.3, 0.5, 1, 3, 5, 7, 10 and 15 miles radii and in 0.3 to 0.5, 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	

Appendix B: Data Description and Limitation

Each observation in our setting is a pick-up location operation that is tied to a geographic location and a date. Therefore, all the variables from the retailers data, including operating frequency, nearby locations and home delivery availability, are time-variant. Unfortunately, the census and the Open Street Map data are time-invariant due to a data limitation: the census data is released with a two year lag, and we use the

Table 2 Variables with Top 10% Varible Importance .

Region 1 (All Data)	
Feature	Variable importance
Coupon	100.0
Day of week	37.2
Median age of population in 3 to 5 miles	33.7
Location tenure	20.7
Week in year	15.9
No. grocery stores in 10 miles	14.4
Max nearby location tenure within 10 miles	13.9
Month	13.9
No. church in 7 to 10 miles	13.2
Number of households in 3 miles	11.1

Region 1 (Ex-ante Data Only)	
Feature	Variable importance
Week in year	100.0
Month	82.6
Year	80.5
Day of week	53.5
Median age of population in 3 to 5 miles	40.0
No. office delivery in 10 to 15 miles	32.1
Home delivery availability	28.4
Number of households in 1 to 3 miles	25.7
Median housing value in 3 to 5 miles	24.6
No. schools in 7 to 10 miles	23.5

Region 2 (All Data)	
Feature	Variable importance
Coupon	100.0
Day of week	45.5
Week in year	36.8
No. church in 5 to 7 miles	26.6
Month	26.2
Avg nearby location tenure in 3 miles	21.9
Location tenure	21.5
No. schools in 0.3 to 0.5 miles	17.2
Population in the block group	15.0
Household mean income in 3 to 5 miles	12.7

Region 2 (Ex-ante Data Only)	
Feature	Variable importance
Year	100.0
Week in year	93.0
Month	64.9
Day of week	54.6
No. church in 5 to 7 miles	30.2
Home delivery availability	23.4
No. office delivery in 3 miles	23.0
Median housing value in 3 miles	17.0
No. Starbucks in 5 miles	16.9
Population in the block group	15.3

most up-to-date census data for 2014 which only covers the first year in our dataset. We assume that the demographic data did not change drastically in the next two years in our data. Similarly, Open Street Map provides a snapshot of most current location data and as a result, we use a snapshot of 2016 data. Once again, we assume that the location data did not change radically. This is a data limitation in our setting. However, if more complete data is available, our prediction model can be updated such that these variables are also time-varying. Finally, for each pick-up location, the demographic data and the location data from Open Street Map are uniquely defined, although they are not time-varying.

Appendix C: Discussion of the Random Forests Model.

In Table 2, we report 10 of the top 25 important variables for two regions. Tables in the first column report results when using all variables and tables in the second column report results using only the ex-ante variables (i.e. without the coupon and tenure variables). The importance measure is calculated based on how much the variable affects the prediction accuracy when the variable is permuted (i.e. randomly shuffled) in the out of sample data (Friedman at al. 2008). If a subset of variables is highly correlated, then these variables have similar importance. As a result, we find that there are many highly correlated variables in the top 10% of importance (e.g., median age in the population within 3 miles and median age in the population within 5 miles). Since we are most interested in finding out which characteristics of a location are important predictors

of high sales, we excluded those variables that have repeating characteristic from the top 10% of the most important variables (i.e. only one median age variable is included even if another median age variable was in the top 10% of important variables).

We find that the coupon and tenure variables are always selected in the models that include all variables. In all models, the time trends variables, which capture seasonality and growth trends, are important (e.g., week-in-year, year, day of week). In each model, there is a good mix of demographic variables and location variables in the top 10% of important variables, suggesting that combing these two data sources is beneficial.

Appendix D: Potential Selection Bias in Revenue Estimation.

In this section, we describe how we rule out potential selection bias in our revenue estimation. First, note that we design our prediction model based on the factors managers consider in the location selection problem. In addition, we restrict the set of potential locations based on the distance from the actual locations to reduce erroneous extrapolation. Nonetheless, one may be concerned that our data suffers from selection bias because the retailer carefully chooses the locations, rather than randomly. Because the managers select locations that they believe will generate high sales, our prediction model might over-predict sales for those locations that were not selected, to the extent that our prediction model excludes factors the managers believe lead to higher sales. Therefore, we conduct two tests, one using an instrumental variable approach and another using a Heckman Selection model, to examine whether we suffer from selection bias.

We first formally present the potential selection bias. Suppose we could observe selection likelihood for locations, then we could model sales as:

$$Sales_{ld} = \gamma_0 + \gamma_1 Sales\ Prediction\ by\ Researcher_{ld} + \gamma_2 Selection\ Likelihood\ by\ Manager_{ld} + e_{ld}. \quad (1)$$

The coefficient of *Selection Likelihood by Manager* would then represent the effect of the manager's assessment of a location's attractiveness not captured by *Sales Prediction by Researcher* on *Sales*. Therefore, in our current model shown in equation (15) where we do not include the manager's assessment of the location, the potential selection bias can be described as an omitted variable problem.

$$Sales_{ld} = \beta_0 + \beta_1 Sales\ Prediction\ by\ Researcher_{ld} + u_{ld}. \quad (2)$$

If $\gamma_2 \neq 0$ in the true model (equation (14)), then in our model (15), β_1 wll be inconsistent. Therefore, we use an instrumental variable method to estimate a consistent β_1.

Table 3 Selection Bias Test using an Instrument.

	(1)	(2)	(3)
Variables	First Stage: Predicted Sales	Second Stage: Actual Sales (IV)	Actual Sales (OLS)
Distance from Fulfillment Center	4.907***		
	(1.913)		
Fulfillment Center Control	Y		
Fitted Sales Prediction from First Stage		0.926***	
		(0.028)	
Sales Prediction			0.945***
			(0.015)
Constant	913.24	155.75	134.12
No. of observations	17130	17130	17130
R-squared	0.210	0.786	0.786

Robust standard error clustered by location in parentheses
*** p<0.01, ** p<0.05, * p<0.10

The instrumental variable approach requires a variable that is correlated with *Selection Likelihood by Manager*. In addition, the variable needs to be uncorrelated with e_{ld} in equation (14). In other words, a valid instrument should not directly affect *Sales*, except through *Selection Likelihood by Manager*. A variable that satisfies these conditions is *Distance from Fulfillment Center*. First, *Distance from Fulfillment Center* is a valid instrument because it does not affect sales outside of its effect on opening decisions. Customers are unlikely to consider distance to fulfillment center when making purchasing decisions. Even if they did, customers do not have access to the distance information because fulfillment center addresses are not public. Second, distance is a valid instrument because it is correlated with managers' opening decisions. Managers, who are located at the fulfillment center, require a higher search cost to open distant locations because these locations require higher travel costs. In addition, locations farther away have higher fulfillment costs and therefore, the likelihood of selection will be lower for locations farther away. Distance is used as an instrument in various settings in empirical literature (Card (1999), Zheng (2016)).

Our resulting two stage least squares regressions is as follows:

$$Sales\ Prediction\ by\ Researcher_{ld} = \delta_0 + \delta_1 Distance\ From\ FC_{ld} + r_{ld}. \tag{3}$$

$$Sales_{ld} = \beta_0 + \beta_1 \widehat{Sales\ Prediction\ by\ Researcher}_{ld} + \epsilon_{ld}. \tag{4}$$

In Table 3, we report estimates from the first stage regression described by equation (16) in column 1 and the second stage regression described by equation (17) in column 2. We also report the baseline OLS

regression where we do not use the instrumental variable in column 3. In the first stage regression, as expected, we find evidence that *Distance from FC* has a non-zero and statistically significant coefficient and therefore likely meets the relevancy condition for valid instruments. In particular, we reject the null hypothesis that the *Distance from FC* variable has zero coefficient, based on the robust F-test statistic of 13.388.

We do not find that the coefficient on predicted sales in the second stage (column 2) differs significantly from the coefficient on our sales prediction variable (column 3). Specifically, we do not reject the hypothesis that predicted sales is exogenous based on the Wooldridges score test (Wooldridge (1995)) (p-value = 0.372). This is analogous to the Durbin-Wu-Hausman test for endogeneity (Durbin (1954), Wu (1973), and Hausman (1978)). Since the standard errors are clustered by location, we use the Wooldridge's score test which is based on the robust standard errors instead of Durbin-Wu-Hausman test. Based on this analysis, we conclude that our results do not suffer from selection bias.

The selection bias problem can also be addressed using the Heckman correction method, also called Heckit (Heckman (1976)). In order to apply the Heckman correction, we need to observe the full population. In our setting, this is equivalent to observing both locations that were selected and not selected. Since we are only able to observe locations that were selected, we randomly sample the non-selected locations from the potential location set to create a non-selected sample and apply the Heckman correction model. The Heckman correction method follows a two-step procedure where in the first step, we estimate a likelihood of a location being selected to open based on the location attributes in a probit model. In the second step, we estimate sales of all location attributes including those that can be only obtained if the observations are selected and the estimated inverse Mills ratio from the first stage. Essentially, the estimated inverse Mills ratio can be thought of as *Selection Likelihood by Manager* in equation (14). For more details, refer to Chapter 19 of Wooldridge (2010).

To avoid multi-collinearity, we include location attribute variables within the 5 miles radius of the location, and we regress *Sale* on these location attributes. In column (1) of Table 4, we report the regression output from the selection bias corrected model. Column (2) contain the probit regression estimates for the selection model based on the location attributes. Finally, column (3) contains the regression output of the selected sample (i.e. actual locations that operated). As we can see from the output, the Heckit and OLS estimates are very close to each other, and the coefficient on the inverse Mills ratio term is statistically insignificant. Based on the likelihood-ratio test, we conclude that the residuals of the first stage and the second stage are

Table 4 Selection Bias Test based on Heckman Selection Model.

VARIABLES	(1) f(Sale): Heckman	(2) select: Heckman	(3) f(Sale): OLS
Coupon	2.273***		2.272***
	(0.119)		(0.120)
ln(Location tenure)	206.603***		203.971***
	(22.277)		(22.026)
No. Schools within 5 miles	1.194*	-0.004	1.472**
	(0.654)	(0.002)	(0.687)
No. Universities within 5 miles	-12.039	0.037**	-12.293
	(8.305)	(0.017)	(8.320)
No. Churches within 5 miles	1.034*	0.010***	0.955*
	(0.562)	(0.002)	(0.557)
No. Supermarkets within 5 miles	-4.039	0.093***	-4.717*
	(2.582)	(0.012)	(2.483)
No. Starbucks within 5 miles	3.384	0.198***	3.314
	(3.252)	(0.036)	(3.265)
Median Income in 5 miles	0.000	0.000*	0.000
	(0.002)	(0.000)	(0.002)
Population in 5 miles	0.014	0.000*	0.013
	(0.010)	(0.000)	(0.010)
Median Age in 5 miles	2.786	-0.043*	5.171
	(13.816)	(0.026)	(12.832)
rho		0.199	
		(0.147)	
sigma		421.929***	
		(20.497)	
lambda		83.835	
		(63.550)	
Quarter-Year Control	Y	Y	Y
Constant	334.572	0.637	292.151
	(498.027)	(0.932)	(482.648)
Observations	24,496	24,496	12,947

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

uncorrelated (χ^2=1.74, p-value = 0.1873). As a robustness check, we also conduct the same analysis using location variables within 3 miles and 1 miles, and find consistent evidence that our main results do not suffer from selection bias.

Based on these two analyses, we believe that we can rule out selection bias in our estimation. This is unsurprising because that the retailer found the location selection problem to be challenging as evidenced by the large variation in sales for locations in our data, and because our choice of variables was motivated by discussions with managers about the factors they consider when opening locations.

Appendix E: Proof of Theorem 1

Proof. We prove the theorem by showing how to convert a given non-integral optimal solution to QP into integral values without decreasing the objective function. Let us combine the location and day indices and write $i=(l,d)$, $j=(m,e)$, $\mathbb{U}=\mathbb{L}\times\mathbb{D}$, and $A_{ij}=S(l,d;m,e)+T(l,d;m,e)$. Then, the objective function

in (QP) can be expressed more compactly as $\sum_{i\in\mathbb{U}} x_i \cdot \{R_i - \sum_{j\in\mathbb{U}} x_j A_{ij}\}$. Note that A_{ij} is symmetric (i.e. $A_{ij} = A_{ji}$).

Let x^* be an optimal solution for (QP). Let $\mathbb{K} = \mathbb{U} \setminus \{i, j\}$. We can order the indices such that

$$R_i - 2\sum_{k\in\mathbb{K}} x_k^* A_{ik} - 2x_j^* A_{ij} \geq R_j - 2\sum_{k\in\mathbb{K}} x_k^* A_{jk} - 2x_i^* A_{ji}.$$

Let $y_k = x_k^*$ for $k \in \mathbb{K}$, $y_i = x_i^* + \epsilon$, and $y_j = x_j^* - \epsilon$ for some $\epsilon > 0$. We will show that y does not decrease the objective function over x^* while remaining feasible, as follows.

$$\begin{aligned}
&\sum_{k\in\mathbb{U}} x_k^* R_k - \sum_{k\in\mathbb{U}}\sum_{h\in\mathbb{U}} x_h^* x_k^* A_{hk} \\
&= \sum_{k\in\mathbb{K}} x_k^* R_k - \sum_{k\in\mathbb{K}}\sum_{h\in\mathbb{K}} x_h^* x_k^* A_{hk} + x_i^* R_i + x_j^* R_j - 2\sum_{k\in\mathbb{K}} x_i^* x_k^* A_{ik} - 2\sum_{k\in\mathbb{K}} x_j^* x_k^* A_{jk} - 2x_i^* x_j^* A_{ij} \\
&\leq \sum_{k\in\mathbb{K}} x_k^* R_k - \sum_{k\in\mathbb{K}}\sum_{h\in\mathbb{K}} x_h^* x_k^* A_{hk} + x_i^* R_i + x_j^* R_j - 2\sum_{k\in\mathbb{K}} x_i^* x_k^* A_{ik} - 2\sum_{k\in\mathbb{K}} x_j^* x_k^* A_{jk} - 2(x_i^* + \epsilon)(x_j^* - \epsilon)A_{ij} - 2\epsilon(x_i^* - x_j^*)A_{ij} \\
&\leq \sum_{k\in\mathbb{K}} x_k^* R_k - \sum_{k\in\mathbb{K}}\sum_{h\in\mathbb{K}} x_h^* x_k^* A_{hk} + (x_i^* + \epsilon)(R_i - 2\sum_{k\in\mathbb{K}} x_k^* A_{ik}) + (x_j^* - \epsilon)(R_j - 2\sum_{k\in\mathbb{K}} x_k^* A_{jk}) - 2(x_i^* + \epsilon)(x_j^* - 2\epsilon)A_{ij} \\
&= \sum_{k\in\mathbb{K}} y_k R_k - \sum_{k\in\mathbb{K}}\sum_{h\in\mathbb{K}} y_h y_k A_{hk} + y_i R_i + y_j R_j - 2\sum_{k\in\mathbb{K}} y_i y_k A_{ik} - 2\sum_{k\in\mathbb{K}} y_j y_k A_{jk} - 2y_i y_j A_{ij}
\end{aligned}$$

To convert a fractional variable of the optimal solution to be integral, choose ϵ such that either y_i or y_j is integral, i.e. $\epsilon = \min(x_j, 1 - x_i)$. This reduces the number of non-integral solution variables without decreasing the objective function. Repeating this procedure until we eliminate all non-integral solution variables will yield an integral optimal solution for (QP).

References

Card D (1999) The causal effect of education on earnings. *Handbook of labor economics*, volume 3, 1801–1863 (Elsevier).

Durbin J (1954) Errors in variables. *Revue de l'institut International de Statistique* 23–32.

Hausman JA (1978) Specification tests in econometrics. *Econometrica: Journal of the econometric society* 1251–1271.

Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement, Volume 5, number 4*, 475–492 (NBER).

Wooldridge JM (1995) Selection corrections for panel data models under conditional mean independence assumptions. *Journal of econometrics* 68(1):115–132.

Wooldridge JM (2010) *Econometric analysis of cross section and panel data* (MIT press).

Wu DM (1973) Alternative tests of independence between stochastic regressors and disturbances. *Econometrica: journal of the Econometric Society* 733–750.

Zheng F (2016) Spatial competition and preemptive entry in the discount retail industry .