Estimating demand in online search markets, with application to hotel bookings

Online Appendix

Sergei Koulayev[*]

May 18, 2014

Contents

1 Identifying variation in the data **1**

2 Derivation of individual likelihoods **7**

2.0.1 k=1,t=1 10

2.0.2 k=1, t>1 10

2.0.3 k>1, k<=t 11

2.1 Unconditional likelihoods 12

2.2 Useful properties of EV Type 1 distribution 13

3 Map of Chicago hotels **15**

1 Identifying variation in the data

The identification strategy outlined above rests on two assumptions: that hotel's price, as observed by a consumer, is uncorrelated with consumer's idiosyncratic tastes and that hotel's

[*] www.sergeikoulayev.com.

first-page membership is also uncorrelated with tastes. In our case, none of these assumptions holds, which introduces bias into the estimation results. However, we find it worthwhile to proceed with the estimation, for two reasons. First, some of our results are derived by comparing the predictions of various discrete choice models, and, if their estimates are biased in a similar way, our conclusions should hold at least qualitatively. Second, we attempt to alleviate the concerns about endogeneity.

Price variation

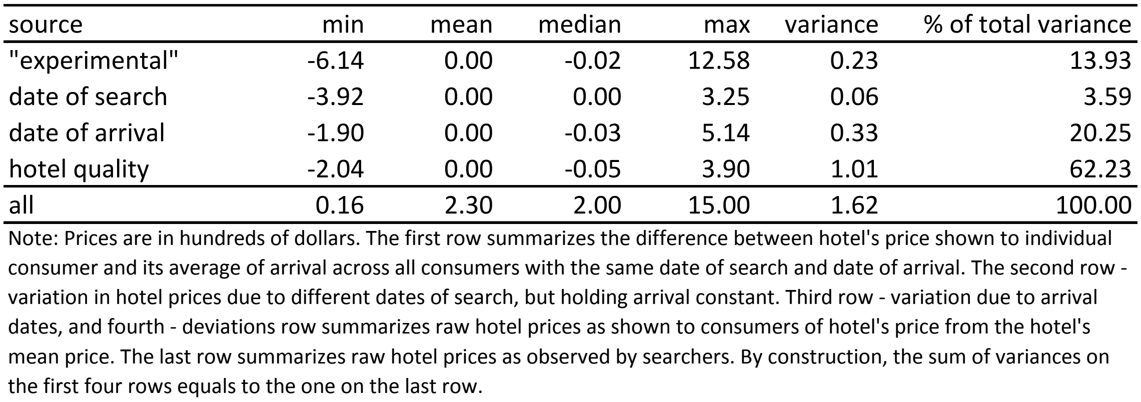
Hotel prices observed by website visitors are a product of hotel's revenue management systems as well as markups imposed by online distribution channels. Although these prices were not set in a direct response to individual users' tastes, it is possible that hotel's price can be correlated with the error term the utility equation. This is due to permanent or temporary shocks to hotel's quality that shift preferences of multiple consumers in a similar way. For example, a hotel may be located on a noisy street – a factor that may be known to travelers, but not to the econometrician, and that permanently reduces demand for that hotel. As an example of a temporal shock, a US Open tournament may increase demand and prices for all hotels in the city, and hotels closer to the stadium will receive larger premiums.

A common solution to this problem is instrumental variables approach. However, it is very difficult to conceive a valid instrument for the hotel market. Given that temporary shocks to hotel prices are typically correlated across time and geography, one cannot use past hotel prices (or prices of hotels in other locations) as an instrument. Exogenous changes in marginal costs are a valid, but weak instrument, because marginal costs are a small component of price.

Instead, we including various controls in the utility specification: dummies for hotel's neighborhood and brand as well as weekend and month dummies, to account for seasonality.

We also find a substantial component of within-hotel price variation that cannot be explained by these economic factors: searchers with the same or very similar combinations of date of search and date of arrival are shown different prices for the same hotel. This is surprising because such searchers look identical from the hotel's perspective. Our hypothesis is

Table 1: Variance decomposition of hotel prices



source	min	mean	median	max	variance	% of total variance
"experimental"	-6.14	0.00	-0.02	12.58	0.23	13.93
date of search	-3.92	0.00	0.00	3.25	0.06	3.59
date of arrival	-1.90	0.00	-0.03	5.14	0.33	20.25
hotel quality	-2.04	0.00	-0.05	3.90	1.01	62.23
all	0.16	2.30	2.00	15.00	1.62	100.00

Note: Prices are in hundreds of dollars. The first row summarizes the difference between hotel's price shown to individual consumer and its average of arrival across all consumers with the same date of search and date of arrival. The second row - variation in hotel prices due to different dates of search, but holding arrival constant. Third row - variation due to arrival dates, and fourth - deviations row summarizes raw hotel prices as shown to consumers of hotel's price from the hotel's mean price. The last row summarizes raw hotel prices as observed by searchers. By construction, the sum of variances on the first four rows equals to the one on the last row.

that hotels or OTA's are engaged in a sort of "experimental" pricing, where they randomly change prices in order to capture some of the high-value consumers[1].

To document this phenomenon, we use all 23,959 unique search sessions by consumers who visited the website during May 2007. From their observation histories, we obtain 721,848 price observations. Such wealth of data allows us to look into very narrow consumer segments, to eliminate almost all observable heterogeneity. We define segments using 3-day windows around the date of search and the date of arrival, which results in 220 consumer "types" per hotel. Matching these types to hotels, we obtain 28,219 of unique hotel-date of search-date of arrival combinations.

Table 1 presents the results of variance decomposition of hotel prices. Let p_{hasi} - price of hotel h shown to consumer i with request parameters (s,a). The following equality holds:

$$V(p_{hasi}) = V(p_{hasi} - \bar{p}_{has}) + V(\bar{p}_{has} - \bar{p}_{ha}) + V(\bar{p}_{ha} - \bar{p}_{h}) + V(\bar{p}_{h} - \bar{p})$$

Rows 1-4 in present summary statistics of each of the four variables on the right side of the equation: $(p_{hasi} - \bar{p}_{has})$ are "experimental" price deviations observed for hotel h by consumers within the same (s,a) cell; $(\bar{p}_{has} - \bar{p}_{ha})$ are price deviations due to different dates of search; $(\bar{p}_{ha} - \bar{p}_{h})$ - price deviations due to different dates of arrival; $(\bar{p}_{h} - \bar{p})$ - price differences due to

[1] To be sure, the search aggregator does not set prices. It retrieves prices from other websites, such as Expedia or hotel's own site, which themselves could be engaged in a dynamic price discrimination.

varying hotel quality. The final row summarizes variation in raw prices, p_{hasi}. As expected, differences in hotel qualities contribute a large part of the observed variation, about 62%. Variation in the dates of arrival contributes another 20%. The added contribution of changing inventory is small, only 3%. In contrast, 14% of price variation has "experimental" nature, as it is not explained by hotel identities and parameters of request. This "experimental" price variation reduces the potential correlation between price and error term in the utility model.

First page variation

The contents of the first page of results, that is observed by all users, is chosen by the website's recommendation system. From conversations with website managers we found that hotels are ranked according to their past click through rates. Therefore, the contents of the first page may be correlated with unobserved demand shocks.

Figure 1 plots the frequencies of appearance of individual hotels on the first page (the data is truncated to a set of 46 hotels with at least a 5% rate). The top 15 hotels appear on 40-60% of first pages observed by the users; there is a hotel that appears on 82% of pages. Thus, there is a certain structural persistence in the composition of the first page. This does not mean, however, that users observe the same page; in fact, among 23,959 search histories in our dataset, a total of 12,455 of unique first pages were displayed. One reason for this diversity is that the first page fits many hotel options – this fact aids the identification of the model in an important way[2].

The advantages of our data, which includes a complete set of search activity over a month, allow us to evaluate a potential correlation between hotel's past attractiveness and its prominence on a given search session. Although we can only obtain an approximation (we do not know the exact formula for the default ranking), we should be able to detect a strong correlation, if it exists.

For each search session i, we observe dates of search and arrival, (s_i, a_i). The outcome variable is y_{hi} - an indicator whether a given hotel was located on the first page ($y_{hi} = 1$) or not ($y_{hi} = 0$). Explanatory variables are: I_h - hotel fixed effect, which controls for a time-

[2] Thanks to the anonymous referee for this observation.

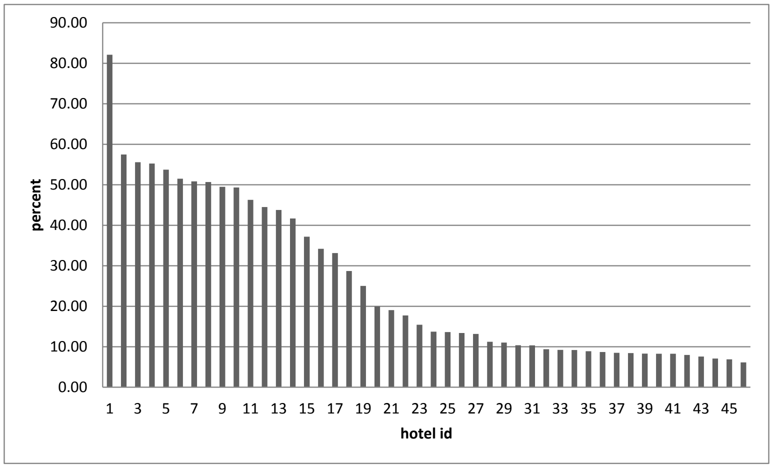


Figure 1: Appearances of individual hotels on the first page

Table 2: Logistic regression of hotel's appearance on first page

Var	Coef	sd	ME	sd	Coef	sd	ME	sd
CTR hotel	0.041	(0.008)	0.005	(0.001)				
CTR_hotel_search	-0.121	(0.007)	-0.015	(0.001)	-0.091	(0.013)	-0.006	(0.001)
CTR_hotel_search_arrive	0.033	(0.001)	0.004	(0.000)	-0.021	(0.002)	-0.001	(0.000)
log(Price)	0.344	(0.004)	0.043	(0.000)	-0.312	(0.010)	-0.021	(0.001)
Hotel FE	NO				YES			
N obs	1,822,557				1,822,557			

Note: Results of a logistic regression, outcome variable is the appearance of a hotel on the first page in a given session. Regressors include: hotel click rate among all searchers; click rate among previous searchers; click rate among previous searchers with the same date of arrival as a given searcher; and a set of hotel fixed effects. Marginal effects are computed. Click rates are measured in percents. Standard errors are in parentheses.

invariant hotel quality; x_{hs} - click rate on hotel h among all searches made prior to s_i, which reflects varying hotel popularity over time; x_{hsa} - click rate on hotel h among searches made prior to s_i, with arrival date a_i, which controls for the effects of future shocks for hotel quality on its current popularity.

Table 2 presents the results from a logistic regression, together with marginal effects. We obtain that popularity effects are small and often have incorrect signs: for example, a 1% increase in the past click rate decreases the chances of prominence by 0.006 percentage points. We interpret this as the evidence of only a weak correlation between first page participation and our measures of temporal shocks to hotel quality.

One reason behind this result is the following. Website managers indicated that a certain amount of "random reshuffling" is introduced into the ranking, to increase the variety of hotels that may appear on the first page. Indeed, we found through simulations that if a pre-determined past click-based formula is used, the first page quickly becomes stationary: limited search leads to a feedback property, where hotels that were prominent yesterday receive a lion's share of clicks today and continue to occupy the first page tomorrow.

2 Derivation of individual likelihoods

In this section, we derive closed form expressions for individual likelihoods of joint searching and clicking decisions. The associated inequalities were derived in the paper. Taking these inequalities as constraints on the unobserved product-specific utility shocks, we analytically integrate out these shocks. The resulting likelihoods will remain conditional on consumers-specific unobservables: tastes for product characteristics and search costs. Derivations produced in this section dramatically reduce the burden of numerical integration in a sequential search model, because the number of product-specific shocks typically vastly exceeds the number of consumer-specific unobservables.

For convenience, we reproduce here notation adopted earlier. Let k - index of the clicked page (where outside option is also part of the first page), t - total number of pages observed. For brevity, we suppress all consumer specific indices in this section. Further, u_g is the maximal utility on page $g = 1..t$, x_k - utility of the clicked hotel (so that $x_k = u_k$) and y_k - maximal utility of remaining hotels on the clicked page. Depending on the combination (k, t), the joint inequalities are the following:

clicked page	observed pages	search	click
$k = 1$	$t = 1$	$x_k > r_t$	$x_k > y_k$
$k = 1$	$t > 1$	$x_k < \min\{r_k, .., r_{t-1}\}$ $x_k > r_t$	$x_k > y_k$ $x_k > u_g, g = k+1..t$
$k > 1$	$t > 1$	$x_k < \min\{r_k, .., r_{t-1}\}, k < t$ $x_k > r_t$ $u_g < \min\{r_g, .., r_{k-1}\},\ g = 1..k-1$	$x_k > y_k$ $x_k > u_g, g = 1..k-1$ $x_k > u_g, g = k+1..t$

(1)

We start with an assumption on the structure of utility, also commonly made in discrete choice demand estimation, including this study. Utility of a product j for consumer i is:

$$u_{ij} = \mu_{ij} + \varepsilon_{ij}$$

- where μ_{ij} is the mean utility function that depends on consumer tastes and product characteristics, and ε_{ij} is an EV Type 1 error term, i.i.d across products and consumers. Is it turns out, the extreme value distribution possesses some extremely valuable properties. In the Section (2.2), we derive various results concerning the extreme value distribution that will be referred to during derivations.

Given a vector of mean utilities μ_{ij} for all hotels observed by consumer i, the following quantities are computed:

(1) μ_g^r - mean utility of a hotel located on the position r on page g;

(2) M_g - mean utility of the best product on page g. Using Claim (1),

$$M_g = \log(\exp(\mu_g^1) + .. + \exp(\mu_g^{15})) \tag{2}$$

(3) Similarly we can define

$$M_{g_1:g_2} = \log(\sum_{g=g_1}^{g_2} \exp(M_g))$$

- mean utility of the best product on pages $g_1, .., g_2$ combined;

(4) μ_x - mean utility of the clicked product, so that $x_k = \mu_x + \varepsilon_x$;

(5) M_k^y - mean utility of the best among non-clicked products on page k, so that $y_k = M_k^y + \varepsilon_y$;

Using the set of reservation utilities, we define the following statistic:

$$\rho_k^t = \begin{cases} \min\{r_k, .., r_{t-1}\} & 1 \leq k < t \\ +\inf & k = t \end{cases} \tag{3}$$

We proceed in two steps. First, we integrate out all product utilities other than the utility of the clicked product, x_k. Second, we integrate out x_k to obtain likelihoods as functions of reservation $r_1, .., r_{t-1}$ and mean utilities of the observed products. In what follows, $F(x)$ is the CDF of standard EV distribution of type II.

Conditional likelihoods - I

Utilities that were observed after the preferred product, y_k and $u_g,\ g = k+1..t$, are not involved in search decisions. They are only involved in the click-related events (see (1)):

$$\begin{aligned} x_k &> y_k \\ x_k &> u_g, g = k+1...t, k < t \end{aligned}$$

whose probabilities conditional on x_k are:

$$\begin{aligned} P(x_k > y_k | x_k = x) &= F(x - M_k^y) \qquad (4) \\ P(x_k > u_{k+1}, .., x_k > u_t | x_k = x) &= F(x - M_{k+1})...F(x - M_t) \\ &= F(x - M_{k+1:t}),\ k < t \qquad (5) \end{aligned}$$

For observations with $k > 1$, utilities $u_1, .., u_{k-1}$ are subject to conditions:

$$\begin{aligned} u_g &< \rho_g^k \equiv \min\{r_g, .., r_{k-1}\},\ g = 1..k-1, k > 1 \\ u_g &< x_k,\ g = 1..k-1, k > 1 \end{aligned}$$

The probability that both inequalities hold is,

$$P(u_g < \min\{\rho_g^k, x_k\} | x_k = x) = F(\min(x, \rho_g^k) - M_g), k > 1 \qquad (6)$$

Putting all this together, obtain the likelihood of joint searching and clicking decision, conditional on the utility of the clicked product:

$$\begin{aligned} L(k,t|x) &= \prod_{g=1}^{k-1} F(\min(x,\rho_g^k) - M_g),\ k>1 \\ &\times \ F(x - M_k^y) \\ &\times \ F(x - M_{k+1:t}),\ k<t \\ &\times \ I(x < \rho_k^t),\ k<t \\ &\times \ I(x > r_t) \end{aligned} \tag{7}$$

On the second step, we integrate this expression with respect to the utility of the clicked product, x. Throughout, we will assume that the rationality constraint $r_t < \rho_k^t$ is satisfied.

Conditional likelihoods - II

2.0.1 k=1,t=1

In this case, the conditional likelihood reduces to:

$$L(k,t|x) = F(x - M_k^y) I(x > r_t)$$

Using Claim (2), we immediately obtain:

$$L(k,t) = \frac{\exp(\mu_x)}{\exp(M_1^y)} \left[1 - F(r_1 - M_1)\right] \tag{8}$$

2.0.2 k=1, t>1

The conditional likelihood is:

$$\begin{aligned}L(k,t|x) &= F(x-M_k^y) \\ &\times F(x-M_{k+1:t}),\ k<t \\ &\times I(x<\rho_k^t),\ k<t \\ &\times I(x>r_t)\end{aligned}$$

Using Claim (2), we obtain:

$$L(k,t) = \frac{\exp(\mu_x)}{\exp(M_{1:t})}\left[F(\rho_k^t - M_{1:t}) - F(r_t - M_{1:t})\right] \tag{9}$$

2.0.3 k>1, k<=t

A separate case $k=t$ can be avoided by defining $\rho_k^t = +\inf$. Consider the first two lines in (7), which involve expressions $\min(x,\rho_g^k)$, $g=1,..,k-1$, provided $k>1$. Because $\rho_g^k \leq \rho_{g+1}^k$, there exists a sequence of indices $g_1,..,g_2$, such that $x<\rho_{g_1}^k,..,\rho_{g_2}^k<\rho_k^t$. This sequence could be empty, but it cannot be disjoint. Thus we obtain thresholds of integration for x:

$$S_x = \{r_t, \rho_{g_1}^k,..,\rho_{g_2}^k,\rho_k^t\} \tag{10}$$

The integral of (7) over x can be represented as a sum of $\#S_x - 1$ elements, by the number of integration intervals in S_x:

$$\begin{aligned}L(k,t) &= \int_{r_t}^{\rho_k^t} L(k,t|x)\exp(-e^{-(x-\mu_x)})e^{-(x-\mu_x)}dx \qquad (11) \\ &= L_1(k,t)+L_2(k,t)+..+L_{(\#S-1)}(k,t)\end{aligned}$$

The individual elements are:

$$L_n(k,t) = \int_{S_x(n)}^{S_x(n+1)} L(k,t|x)\exp(-e^{-(x-\mu_x)})e^{-(x-\mu_x)}dx \tag{12}$$

Consider $S_x(n) < x < S_x(n+1)$. When the utility of the clicked product belongs to that interval, the conditional probability (7) simplifies to:

$$\begin{aligned} L(k,t|x,x &\in (S_x(n),S_x(n+1))) = \prod_{g:\rho_g^k \leq S_x(n), g<k} F(\rho_g^k - M_g) \prod_{g:\rho_g^k \geq S_x(n+1), g<k} F(x - M_g) \qquad (13) \\ &\times \ F(x - M_k^y) \\ &\times \ F(x - M_{k+1:t}), \ k < t \end{aligned}$$

Integrating over $x \in (S_x(n), S_x(n+1))$, we obtain:

$$\begin{aligned} L_n(k,t) &= \prod_{g:\rho_g^k \leq S_x(n), g<k} F(\rho_g^k - M_g) && (14) \\ &\quad \times \frac{\exp(\mu_x)}{S_n} \left[F(S_x(n+1) - \log(S_n)) - F(S_x(n) - \log(S_n))\right] && (15) \\ S_n &= \sum_{g:\rho_g^k \geq S_x(n+1), g<k} \exp(M_g) + \sum_{g=k}^{t} \exp(M_g) \end{aligned}$$

2.1 Unconditional likelihoods

So far we have derived likelihoods of observed clicking and searching decisions, conditional on the choice of search strategy (as well as a vector of reservation utilities and other consumer-specific traits). For consumer i, denote this conditional likelihood as $L_i(k_i, t_i | r_{1i}^{s_i}, .., r_{t_i i}^{s_i}, \theta_i\dot{)}$, where s_i is the index of the chosen strategy. (Of course, s_i is observed only for those who searched; for non-searchers, it has to be integrated out). Assuming that the choice of search strategy is multinomial logit, the unconditional likelihoods are obtained as:

$$\begin{aligned} L_i(k_i, t_i, s_i | \theta_i\dot{)} &= L_i(k_i, t_i | r_{1i}^{s_i}, .., r_{t_i i}^{s_i}, \theta_i\dot{)} \frac{\exp(r_{1i}^{s_i})}{\sum_{s=1}^{S} \exp(r_{1i}^{s})}, \ t_i > 1 && (16) \\ L_i(k_i, t_i | \theta_i\dot{)} &= \sum_{s_i=1}^{S} L_i(k_i, t_i | r_{1i}^{s_i}, \theta_i\dot{)} \frac{\exp(r_{1i}^{s_i})}{\sum_{s=1}^{S} \exp(r_{1i}^{s})}, \ t_i = 1 && (17) \end{aligned}$$

These likelihoods are then integrated over unobserved component in θ_i - vector of consumer-specific traits. In practice, this is done by taking 200 Halton draws from the appropriate search cost distribution and the distribution of consumer tastes.

2.2 Useful properties of EV Type 1 distribution

Suppose x is EV Type 1 random variable with location parameters a and a unit scale. Its CDF and PDF are:

$$\begin{aligned} F_x(x) &= \exp(-e^{-(x-a)}) \\ f_x(x) &= \exp(-e^{-(x-a)})e^{-(x-a)} \end{aligned}$$

If $F(x)$ is a CDF of a standard EV Type 1 (with location zero and scale one), then $F(x-a) = F_x(x)$.

Claim 1 *The distribution of a maximum of n independent EV Type 1 random variables with location parameters $a_1, .., a_n$ and unit scale, is also EV Type 1 with location parameter given by $M(a_1, .., a_n) = \ln(\exp(a_1) + .. + \exp(a_n))$.*

Proof. The CDF of the maximum is: $P(\max(x_1, .., x_n) < x) = F(x - a_1)...F(_n - a_n)$. The product of CDF's can be written as:

$$\begin{aligned} F(x-a_1)..F(x-a_n) &= \exp\left(-e^{-(x-a_1)}.. - e^{-(x-a_n)}\right) \\ &= \exp(-e^{-x}e^{a_1}.. - e^{-x}e^{a_n}) \\ &= \exp(-e^{-x}(e^{a_1} + .. + e^{a_n})) \\ &= \exp(-e^{-(x-M(a_1,..,a_n))}) \\ &= F(x - M(a_1, .., a_n)) \end{aligned}$$

■

Claim 2 *Let x, y - independent EV Type 1 random variables with location parameters μ_x and a, respectively. Let - constants. The probability of an event: $x > y, x_L < x < x_H$, where*

$x_L < x_H$ *are constants, is given by:*

$$
\begin{aligned}
P(x &> y, x_L < x < x_H) = \int_{x_L}^{x_H} F_y(x) f_x(x) dx \\
&= \frac{\exp(\mu_x)}{\exp(M(a,\mu_x))} \left(F(x_H - M(a,\mu_x)) - F(x_L - M(a,\mu_x))\right)
\end{aligned}
$$

Proof. First, we substitute the definition of CDF and PDF of extreme value distribution and make some simplifications:

$$
\begin{aligned}
\int_{x_L}^{x_H} F_y(x) f_x(x) dx &= \int_{x_L}^{x_H} F(x-a) f_x(x) dx \\
&= \int_{x_L}^{x_H} \exp(-e^{-(x-a)}) \exp(-e^{-(x-\mu_x)}) e^{-(x-\mu_x)} dx \\
&= \int_{x_L}^{x_H} \exp(-e^{-x} e^{a} - e^{-x} e^{\mu_x}) e^{-x} e^{\mu_x} dx \\
&= \int_{x_L}^{x_H} \exp(-e^{-x}(e^{a} + e^{\mu_x})) e^{-x} e^{\mu_x} dx
\end{aligned}
$$

Now we can make a substitution: $t = e^{-x}$, $dt = -e^{-x} dx$.

$$
\begin{aligned}
\int_{x_L}^{x_H} \exp(-e^{-x}(e^{a} + e^{\mu_x})) e^{-x} e^{\mu_x} dx &= \int_{\exp(-x_H)}^{\exp(-x_L)} \exp(-t(e^{a} + e^{\mu_x})) e^{\mu_x} dt \\
&= -\frac{e^{\mu_x}}{(e^{a} + e^{\mu_x})} \exp(-t(e^{a} + e^{\mu_x}))|_{\exp(-x_H)}^{\exp(-x_L)} \\
&= \frac{e^{\mu_x}}{(e^{a} + e^{\mu_x})} \left(F(x_H - a) F(x_H - \mu_x) - F(x_L - a) F(x_L - \mu_x)\right) \\
&= \frac{\exp(\mu_x)}{\exp(M(a,\mu_x))} \left(F(x_H - M(a,\mu_x)) - F(x_L - M(a,\mu_x))\right)
\end{aligned}
$$

■

3 Map of Chicago hotels

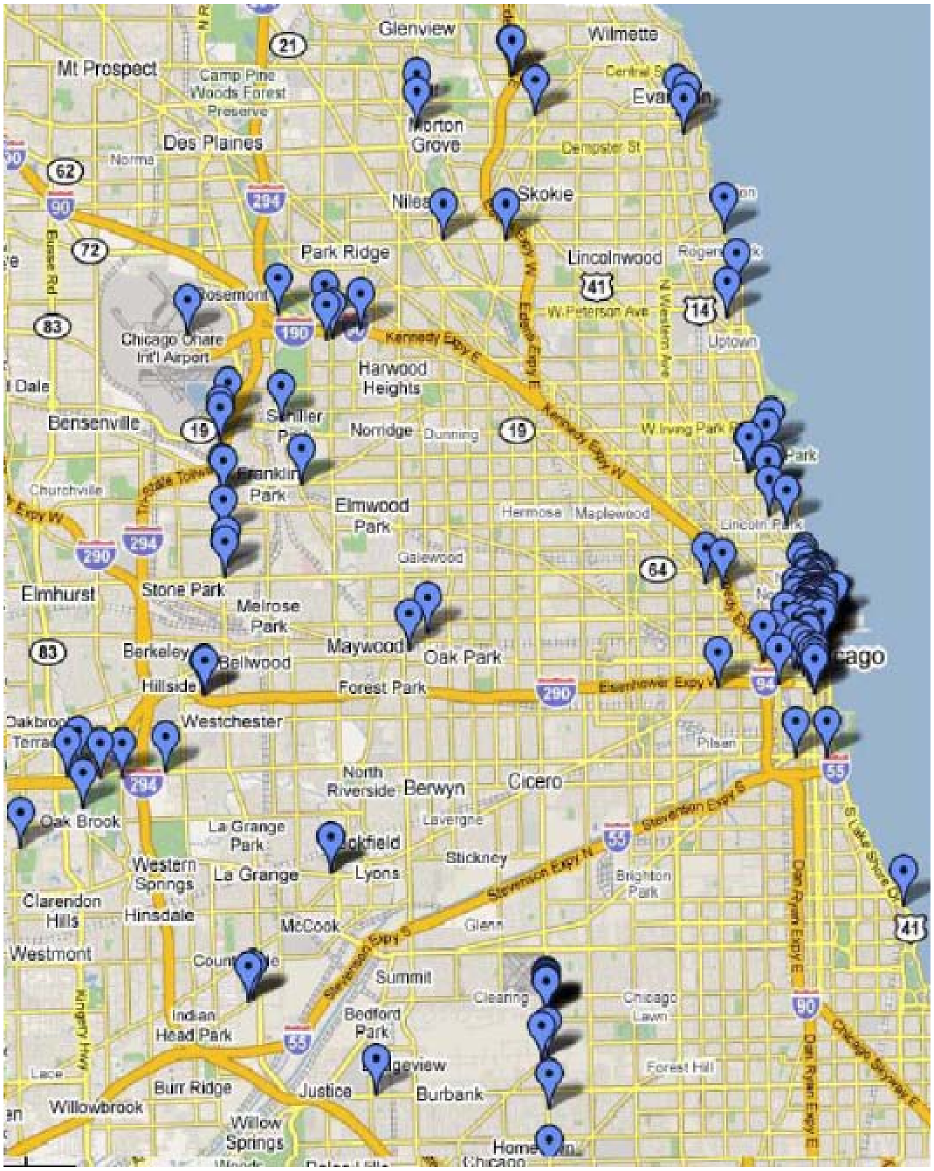


Figure 2: Geographical dispersion of Chicago hotels