Demand forecasting in airline industry

Shadi Sharif Azadeh
Alireza Nikseresht

Erasmus University Rotterdam

22 June 2018

Outline

Introduction

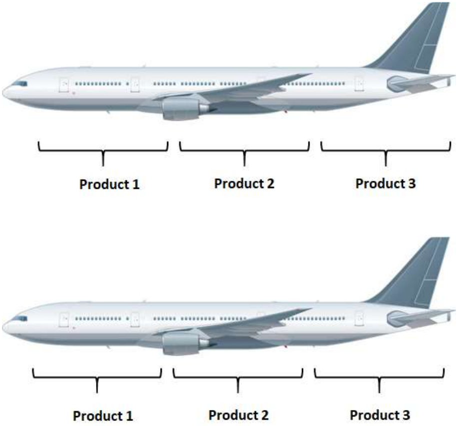
Revenue management

The application of disciplined tactics that predict consumer behaviour at the micro-market level and optimize product availability and price to maximize revenue growth (Robert Cross, 1998)

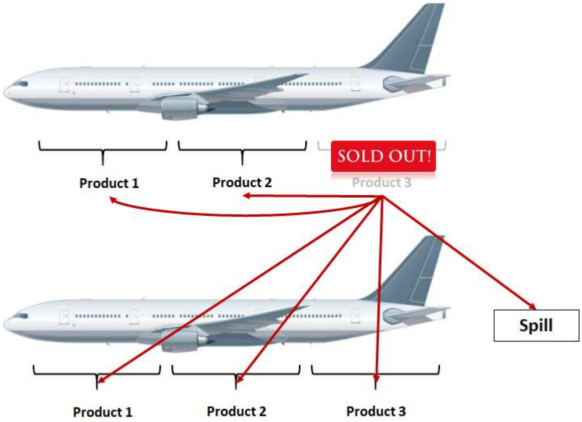
Resources

- Transactional sales.
- No information about different customer segments' preferences.

Spill and recapture



Spill and recapture

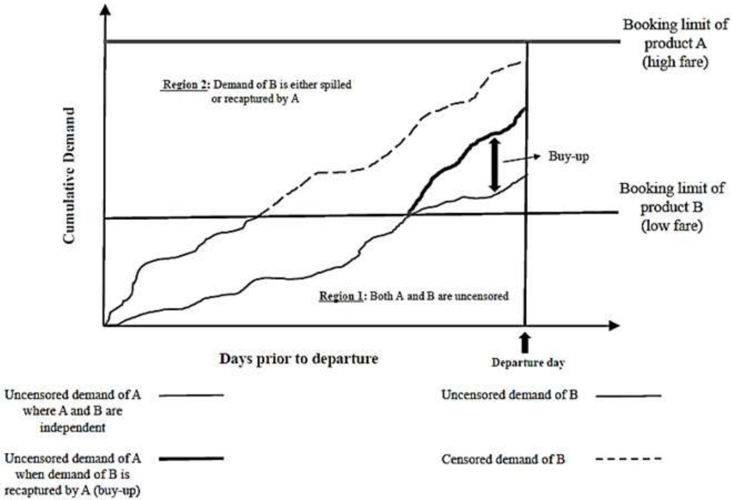


Example of Underestimating Demand

The company said it sold 13.1 million iPads in 2017, down 22 percent from the 16.1 million it sold during the holiday season in 2016. iPad revenue was $5.5 billion, down 19 percent from $7.1 billion a year before.

CEO Tim Cook insisted the iPad still had a bright future, and blamed this in part due to underestimating of demand.

Motivation


Booking limit of product A (high fare)
Region 2: Demand of B is either spilled or recaptured by A
Cumulative Demand
Buy-up
Booking limit of product B (low fare)
Region 1: Both A and B are uncensored
Days prior to departure
Departure day
Uncensored demand of A where A and B are independent
Uncensored demand of B
Uncensored demand of A when demand of B is recaptured by A (buy-up)
Censored demand of B

Motivation

Product	A	B	C	D
General information				
Availability	1	1	0	1
Booking limit	8	6	0	23
Observed demand	8	0	0	22
Actual demand	15	**0**	**9**	22

Motivation

Product	A	B	C	D
General information				
Availability	1	1	0	1
Booking limit	8	6	0	23
Observed demand	8	0	0	22
Actual demand	15	**0**	**9**	22

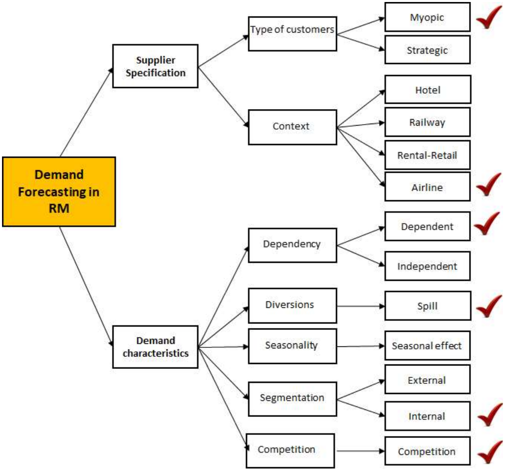
- on high-demand flights, underestimating the demands about 12.5% to 25% results in the revenue around 3% (Weatherford and Belobaba, 2002; Guo, Xiao et al. 2012; Sharif Azadeh et al. 2014, 2015).

Motivation

Product	A	B	C	D
General information				
Availability	1	1	0	1
Booking limit	8	6	0	23
Observed demand	8	0	0	22
Actual demand	15	**0**	**9**	22

- on high-demand flights, underestimating the demands about 12.5% to 25% results in the revenue around 3% (Weatherford and Belobaba, 2002; Guo, Xiao et al. 2012; Sharif Azadeh et al. 2014, 2015).
- censored data used as true demands causes a spiral down effect over time (Cooper et al. 2006; Little et al. 2002).

Framework and assumptions



Framework and assumptions

Classes	Penalty due to change	Cancellation	Price$
c_1	No	No	200
c_2	No	No	240
c_3	Charge 40%	Charge 50%	300
c_4	Charge 10%	Charge 20%	400
c_5	No charge	No charge	600

Framework and assumptions

Classes	Penalty due to change	Cancellation	Price$
c_1	No	No	200
c_2	No	No	240
c_3	Charge 40%	Charge 50%	300
c_4	Charge 10%	Charge 20%	400
c_5	No charge	No charge	600

- every customer has a preference defined based on class attributes.
- customer preferences are estimated for each class.
- using the estimated preference vector, true, spilled and recaptured demands are calculated.

Notations

- single flight and single leg network.
- n classes: $c_1, c_2, ..., c_n$.

Notations

- single flight and single leg network.
- n classes: $c_1, c_2, ..., c_n$.
- booking horizon: start of booking (1), departure date (T); $t \in \{1, ..., T\}$ non-uniformly discretized.
- $C_t = \{c_{1t}, c_{2t}, ..., c_{nt}\}$ subset of classes offered at time t.

Notations

- single flight and single leg network.
- n classes: $c_1, c_2, ..., c_n$.
- booking horizon: start of booking (1), departure date (T); $t \in \{1, ..., T\}$ non-uniformly discretized.
- $C_t = \{c_{1t}, c_{2t}, ..., c_{nt}\}$ subset of classes offered at time t.
- number of observed sales for class j in interval t: $z_{jt} \sim F_z$.
- true demand for class j in interval t: $d_{jt} \sim F_d$. Definition: the actual demand if all classes were available at all times.

Notations

- single flight and single leg network.
- n classes: $c_1, c_2, ..., c_n$.
- booking horizon: start of booking (1), departure date (T); $t \in \{1, ..., T\}$ non-uniformly discretized.
- $C_t = \{c_{1t}, c_{2t}, ..., c_{nt}\}$ subset of classes offered at time t.
- number of observed sales for class j in interval t: $z_{jt} \sim F_z$.
- true demand for class j in interval t: $d_{jt} \sim F_d$. Definition: the actual demand if all classes were available at all times.
- class availabilities defined exogenously.

- $\hat{z_{jt}}$: estimated number of observed sales for class j at time t.
- Z_t: $\{z_{1t}, z_{2t}, ..., z_{nt}\}$ set of observed sales for offered classes at time t.
- M_t: aggregated number of observed sales at time t.
- $\hat{M_t}$: estimated total number of observed sales at time t.
- v_{jt}: customer preference weight towards class j in period t.
- v_0: no purchase.
- V_t: preference weight vector in period t.
- p_{jt}: Probability of choosing class j in time frame t.
- $\hat{z}_{0t}$: Denote the estimated number of no purchase in time frame t.
- A_t: estimated number of customer arrival in time frame t.
- r_{jt}: recaptured to class j in time frame t.
- s_{jt}: spilled or turned away demand from class j in time frame t.
- S_t: aggregated spilled demand in time frame t.
- e_{jt}: difference between observed and estimated sales of class j in time frame t.

Assumptions

- MNL choice model.
- information at hand: aggregated sales over each time interval t, availability of classes .
- homogeneous customers in terms of their preferences.

Probability function

$$p_{jt} = \frac{v_{jt}}{\sum_{i \in C_t} v_{it} + v_0} \tag{1}$$

If class j is not available in period t, then $p_{jt} = 0$.

$$p_{0t} = \frac{v_0}{\sum_{i \in C_t} v_{it} + v_0} \tag{2}$$

$$\sum_{j=0}^{n} p_{jt} = 1 \tag{3}$$

aggregated number of customers arriving in period t is A_t. The estimated number of observed sales of class j in period t is:

$$\hat{z_{jt}} = p_{jt} A_t \tag{4}$$

Estimated demand

The estimated total number of sales in the period t is achieved through a summation over all available classes' estimated sales.

$$\hat{M}_t = \sum_{j \in C_t} \hat{z}_{jt} = \frac{\sum_{j \in C_t} v_{jt}}{\sum_{j \in C_t} v_{jt} + v_0} A_t \tag{5}$$

Using total observed sales in each period t (M_t) and estimated total sales in that period ($\hat{M}_t$), we can estimate the number of arrivals:

$$A_t \approx \left(\sum_{j \in C_t} z_{jt}\right) \frac{\sum_{j \in C_t} v_{jt} + v_0}{\sum_{j \in C_t} v_{jt}} \tag{6}$$

Estimated demand

The estimated number of no purchases is equal to, the number of arrival minus the number of estimated total sales in period t,

$$\hat{z_{0t}} = A_t - \hat{M}_t = A_t - \frac{\sum_{j \in C_t} v_{jt}}{\sum_{j \in C_t} v_{jt} + v_0} A_t = \frac{v_0}{\sum_{j \in C_t} v_{jt} + v_0} A_t = p_{0t} A_t \quad (7)$$

Estimated demand

The estimated number of no purchases is equal to, the number of arrival minus the number of estimated total sales in period t,

$$\hat{z_{0t}} = A_t - \hat{M}_t = A_t - \frac{\sum_{j \in C_t} v_{jt}}{\sum_{j \in C_t} v_{jt} + v_0} A_t = \frac{v_0}{\sum_{j \in C_t} v_{jt} + v_0} A_t = p_{0t} A_t \quad (7)$$

Finally, true demand is the demand that would have been seen, if all n classes had been available, calculated as follows:

$$d_{jt} = p_{jt} A_t = \frac{v_j t}{\sum_{i=1}^{n} v_{it} + v_0} A_t \quad (8)$$

True, recaptured and spilled demand

When class j is available in period t, the estimated number of observed sales of j in that period ($\hat{z}_{jt}$) consists of:

- the number of customers interested in class j as their first choice (true demand or d_{jt}),

True, recaptured and spilled demand

When class j is available in period t, the estimated number of observed sales of j in that period ($\hat{z}_{jt}$) consists of:

- the number of customers interested in class j as their first choice (true demand or d_{jt}),
- the number of customers buying j because their first choice was not available (recapture or r_{jt}),

True, recaptured and spilled demand

When class j is available in period t, the estimated number of observed sales of j in that period ($\hat{z_{jt}}$) consists of:

- the number of customers interested in class j as their first choice (true demand or d_{jt}),
- the number of customers buying j because their first choice was not available (recapture or r_{jt}),
- number of customers unable to buy class j as their preferred class (spill or s_{jt}).

$$\hat{z_{jt}} = d_{jt} + r_{jt} - s_{jt} \tag{9}$$

Algorithm

1) estimating preference weights V_t using the observed transactions data and availability set using Mean Square Error.

Algorithm

1) estimating preference weights V_t using the observed transactions data and availability set using Mean Square Error.
2) number of arrivals at time t: A_t,

Algorithm

1) estimating preference weights V_t using the observed transactions data and availability set using Mean Square Error.

2) number of arrivals at time t: A_t,

3) true demand of each class j at time t: d_{jt},

Algorithm

1) estimating preference weights V_t using the observed transactions data and availability set using Mean Square Error.
2) number of arrivals at time t: A_t,
3) true demand of each class j at time t: d_{jt},
4) number of spilled demand at time t: S_t,

Algorithm

1) estimating preference weights V_t using the observed transactions data and availability set using Mean Square Error.

2) number of arrivals at time t: A_t,

3) true demand of each class j at time t: d_{jt},

4) number of spilled demand at time t: S_t,

5) number of recaptured demand in class j at time t: r_{jt}.

Algorithm

1) estimating preference weights V_t using the observed transactions data and availability set using Mean Square Error.

2) number of arrivals at time t: A_t,

3) true demand of each class j at time t: d_{jt},

4) number of spilled demand at time t: S_t,

5) number of recaptured demand in class j at time t: r_{jt}.

Initialization

- preferences initialization:

$$v_{j0} = \frac{\sum_{t=1}^{T} z_{jt}}{\sum_{t=1}^{T} \sum_{i=1}^{n} z_{it}} \tag{10}$$

- if class j at time t not available: $\hat{z_{jt}} = r_{jt} = 0$ and $S_t = \sum_{j \notin C_t} d_{jt}$.

Recapture

We assume that all offered classes in period t (C_t) are arranged based on their preference weight in a descending order. The first class becoming unavailable in each period are those with higher preference weight. Suppose that classes 1 to j are not available;

- their associated requests shift to other available classes.
- they leave the system without purchase.

Recapture

We assume that all offered classes in period t (C_t) are arranged based on their preference weight in a descending order. The first class becoming unavailable in each period are those with higher preference weight. Suppose that classes 1 to j are not available;

- their associated requests shift to other available classes.
- they leave the system without purchase.

Recaptured demand by class k at time t, could be estimated by the product customers who could not buy their first choice and the probability of choosing class k:

$$r_{kt} = S_t \frac{v_{kt}}{\sum_{j \in C_t} v_{jt} + v_0} \quad k \in C_t \tag{11}$$

Summary of algorithm

After initializing V, and using the observed transactions, we estimate number of arrivals at each time interval, A_t using:

$$A_t \approx (\sum_{j \in C_t} z_{jt}) \frac{\sum_{j \in C_t} v_{jt} + v_0}{\sum_{j \in C_t} v_{jt}}$$

Summary of algorithm

After initializing V, and using the observed transactions, we estimate number of arrivals at each time interval, A_t using:

$$A_t \approx (\sum_{j \in C_t} z_{jt}) \frac{\sum_{j \in C_t} v_{jt} + v_0}{\sum_{j \in C_t} v_{jt}}$$

Then, the true demand is calculated as:

$$d_{jt} = p_{jt} A_t = \frac{v_j t}{\sum_{j=1}^{n} v_{jt} + v_0} A_t$$

Summary of algorithm

After initializing V, and using the observed transactions, we estimate number of arrivals at each time interval, A_t using:

$$A_t \approx (\sum_{j\in C_t} z_{jt})\frac{\sum_{j\in C_t} v_{jt} + v_0}{\sum_{j\in C_t} v_{jt}}$$

Then, the true demand is calculated as:

$$d_{jt} = p_{jt}A_t = \frac{v_j t}{\sum_{j=1}^{n} v_{jt} + v_0}A_t$$

Finally, spill and recapture are calculated as mentioned previously:

$$S_t = \sum_{j\notin C_t} d_{jt}$$

Summary of algorithm

After initializing V, and using the observed transactions, we estimate number of arrivals at each time interval, A_t using:

$$A_t \approx (\sum_{j \in C_t} z_{jt}) \frac{\sum_{j \in C_t} v_{jt} + v_0}{\sum_{j \in C_t} v_{jt}}$$

Then, the true demand is calculated as:

$$d_{jt} = p_{jt} A_t = \frac{v_j t}{\sum_{j=1}^{n} v_{jt} + v_0} A_t$$

Finally, spill and recapture are calculated as mentioned previously:

$$S_t = \sum_{j \notin C_t} d_{jt}$$

and,

$$r_{kt} = S_t \frac{v_{kt}}{\sum_{j \in C_t} v_{jt} + v_0} \quad k \in C_t$$

Objective function

At the end of each iteration, the difference between observed and estimated sales is reduced (error):

$$e_{jt} = z_{jt} - \hat{z}_{jt} = z_{jt} - (d_{jt} + r_{jt} - s_{jt}) \tag{12}$$

Objective function

At the end of each iteration, the difference between observed and estimated sales is reduced (error):

$$e_{jt} = z_{jt} - \hat{z_{jt}} = z_{jt} - (d_{jt} + r_{jt} - s_{jt}) \tag{12}$$

Assumption: when a class is available there is no spilled demand.

$$\min_{V} E_t(V_t) = \sum_{j=1}^{n} e_{jt}^2 \tag{13}$$

Line search method is used to find the local minimum with adaptive step size. The objective function is non-convex.

Simulated data generation

Simulated data is used to evaluate the goodness of the model then the model is tested using real data. Two types of datasets:

Simulated data generation

Simulated data is used to evaluate the goodness of the model then the model is tested using real data. Two types of datasets:

uncensored demand generation

- preference vectors for each dataset, randomly generated from a Gumbel distribution,
- arriving customers at each time period are generated via a Non-homogeneous Poisson Process,
- all classes are available in all periods,
- MNL probability of choosing class j at time t.

censored demand generation

- exogenously generated indicator matrix to show the class availability at each time period.
- demand, spill and recapture are calculated.

Simulated data

A sample of simulated data for a specific flight, z_{jt}:

↓ Observed Data ↓															
	Booking Periods														
Classes	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
c1	20	17	11	14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
c2	9	13	8	6	13	20	NA	NA	NA	NA	NA	NA	NA	NA	NA
c3	5	14	4	6	4	7	9	12	8	NA	NA	NA	NA	NA	NA
c4	6	3	2	3	4	4	6	5	5	8	12	7	NA	NA	NA
c5	0	1	1	1	2	1	1	1	1	2	3	3	2	1	4
↓ Hidden Data ↓															
No Purchases	18	16	14	18	21	16	30	34	20	51	56	30	44	43	47
#Arrivals	58	64	40	48	44	48	46	52	34	61	71	40	46	44	51

Estimated preference vector

Classes	Preference weights	True values (V_t)	Estimated values ($\hat{V}_t$)
c_1	v_1	1	1.02
c_2	v_2	0.7	0.64
c_3	v_3	0.4	0.35
c_4	v_4	0.2	0.18
c_5	v_5	0.5	0.6

Unconstrained demand using estimated preference vector

Estimated true demand using the proposed algorithm, d_{jt}:

	Booking Periods														
Classes	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
c1	20	17	11	14	13.14	18.28	13.62	15.32	11.91	16.41	24.62	16.41	12.79	6.39	25.57
c2	9	13	8	6	9.22	14.74	8.53	9.59	7.46	10.28	15.42	10.28	8.01	4	16.01
c3	5	14	4	6	1.94	4.13	4.12	6.51	3.73	5.6	8.41	5.6	4.37	2.18	8.73
c4	6	3	2	3	2.91	2.48	3.41	2.09	2.73	3.16	4.73	2.16	2.32	1.16	4.63
c5	0	1	1	1	1.69	0.58	0.28	0.19	0.37	0.65	0.97	1.65	0.65	0.32	1.29
No Purchases	17.75	21.3	11.54	13.31	12.81	17.83	13.28	14.95	11.64	16.01	24.02	16.01	12.46	6.25	24.97
#Arrivals	57.75	69.3	37.54	43.31	41.71	58.04	43.24	48.65	37.84	52.11	78.17	52.11	40.6	20.3	81.2

Recaptured and spilled demand using estimated preference vector

Recaptured and spilled demand, r_{jt} and s_{jt}:

	Booking Periods														
Classes	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
c1	0	0	0	0	-13.14	-18.28	-13.62	-15.32	-11.91	-16.41	-24.62	-16.41	-12.79	-6.39	-25.57
c2	0	0	0	0	3.78	5.26	-8.53	-9.59	-7.46	-10.28	-15.42	-10.28	-8.01	-4	-16.01
c3	0	0	0	0	2.06	2.87	4.88	5.49	4.27	-5.6	-8.41	-5.6	-4.37	-2.18	-8.73
c4	0	0	0	0	1.09	1.52	2.59	2.91	2.27	4.84	7.27	4.84	-2.32	-1.16	-4.63
c5	0	0	0	0	0.31	0.42	0.72	0.81	0.63	1.35	2.03	1.35	1.35	0.68	2.71

Predicated sales

Predicted number of transactions (sales) using the estimated preference vector, $\hat{z_{jt}}$:

Classes	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
c1	18.19	21.82	11.82	13.64	0	0	0	0	0	0	0	0	0	0	0
c2	11.39	13.67	7.4	8.54	12.01	16.71	0	0	0	0	0	0	0	0	0
c3	6.21	7.45	4.04	4.66	6.55	9.11	9.53	10.72	8.34	0	0	0	0	0	0
c4	3.3	3.95	2.14	2.47	3.47	4.83	5.06	5.69	4.42	7.82	11.73	7.82	0	0	0
c5	0.92	1.1	0.6	0.69	0.97	1.35	1.41	1.59	1.23	2.18	3.27	2.18	2	1	4
No Purchases	17.75	21.3	11.54	13.31	18.71	26.04	27.24	30.65	23.84	42.11	63.17	42.11	38.6	19.3	77.2
#Arrivals	57.75	69.3	37.54	43.31	41.71	58.04	43.24	48.65	37.84	52.11	78.17	52.11	40.6	20.3	81.2

Comparison between estimated and original data

Mean, standard deviation, and mean square error of the difference of estimated and simulated true demand, d_{jt} and $\hat{d}_{jt}$:

Classes	Mean	SD	MSE
c1	0.476	5.551	31.02
c2	0.248	3.868	15.01
c3	0.128	2.236	5.01
c4	0.069	1.287	1.66
c5	0.085	3.092	1.10

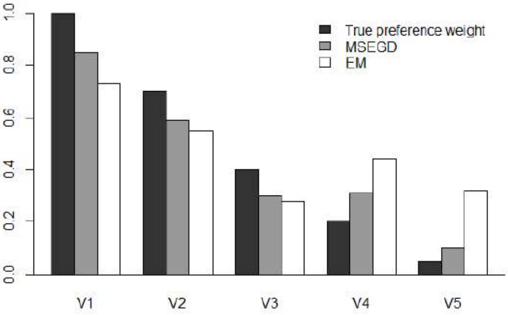
Computational time

Mean execution time of the proposed algorithm on different size datasets:

Dataset	Mean Convergence Time(sec) Adaptive step size	Mean Convergence Time(sec) Fixed step size (l=0.0001)	Mean iteration count Fixed step size
5X15	0.021	0.2656	69.58
10X100	0.187	10.4099	137.41
15X150	0.412	32.4776	207.87
20X300	0.858	61.626	150.61

Comparison with EM Method

The result of the comparison between true preference vector and estimated values, using EM method and our proposed algorithm (MSEGD).



Real data case study

class	Mean Fare	Mean Seat Available in periods before flight	Mean reservation in each flight	class	Mean Fare	Mean Seat Available in periods before flight	Mean reservation in each flight
c1	83	0.59	7.94	c7	579	45.76	12.83
c2	148	2.22	13	c8	686	55.6	9.4
c3	202	4.55	15.42	c9	848	62.16	5.73
c4	256	12.02	25.46	c10	1009	70.56	5.91
c5	363	22.66	18.52	c11	1279	73.59	2.14
c6	471	35.7	18.91	c12	1386	76.49	1.62

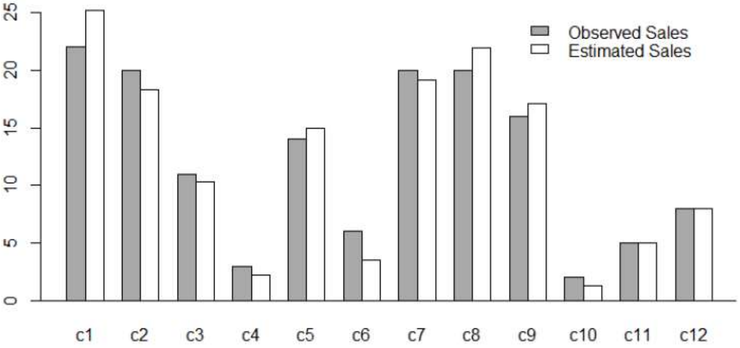
Estimated preference

We use our proposed algorithm to estimate the preference vector using real data case study.

Estimated V for each classes

Classes	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12
V values	0.94	1.07	0.60	0.23	0.68	0.16	0.95	0.39	0.20	0.15	0.07	0.05

Estimated versus observed



Conclusion and remarks

- a non-parametric algorithm has been introduced to estimate preference weights, true demand, spill and recapture in the presence of only registered booking data.
- it has been tested for both simulated and real data.