# Choice based demand forecasting in RM

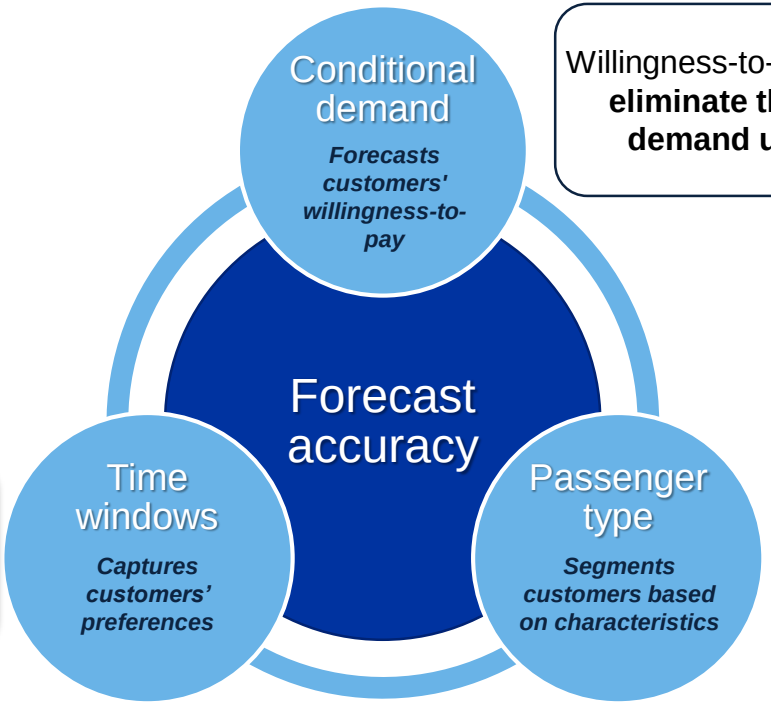
Gizem Keysan and Nick Paavola

May 9th, 2019

# Did you hear we have a new livery?

# Last year, we talked about three design choices that support the measurement and improvement of forecast accuracy

Willingness-to-pay-based forecasts **eliminate the guesswork of demand un-constraining**




Forecasting each **OD/time window** instead of each possible itinerary leverages the statistical magic of aggregation

Forecasting demand for **distinct customer profiles (business, leisure, award, etc)** allows for more granular segmentation

# As RM systems evolve, they relax constraints on customer choice

## Fixed itineraries

- Customer preference is static
- Only the price of the single itinerary matters
- Other itineraries don't influence decisions

## Discrete time preferences

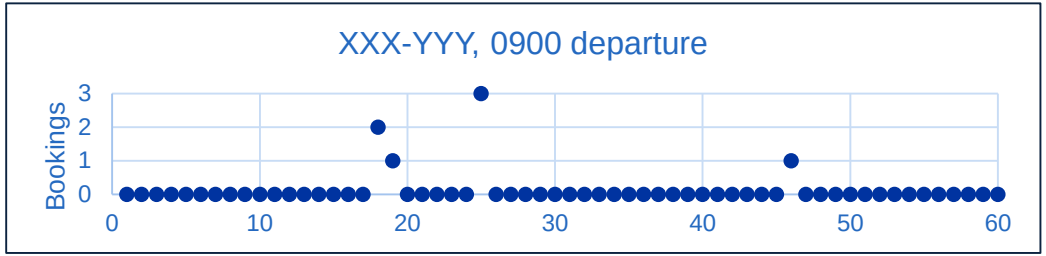
- Customer preference is sticky within discrete bands
- Neighboring time bands might influence decisions

## Continuous time preferences

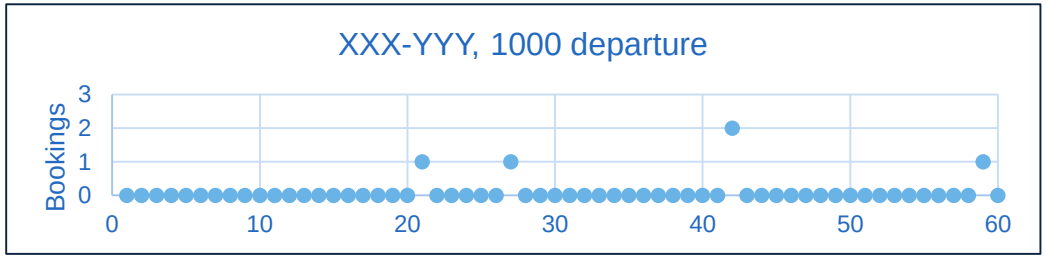
- Customer preference spans across hours and maybe days
- Characteristics of a broad set of options might influence decisions

Wittman, Fiig, Adelving, Belobaba
AGIFORS 2017

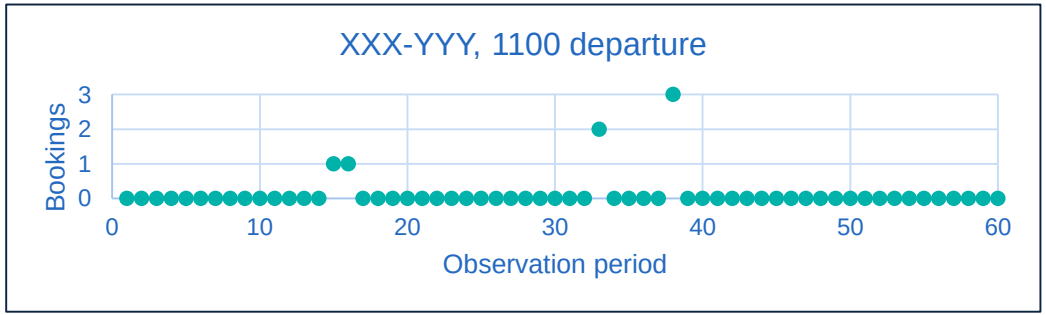
# Forecasting fixed itineraries assumes independence across choices and exacerbates the “small numbers” problem



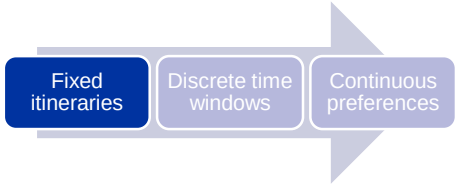

$$CV = \frac{\sigma}{\mu} = 4.2$$



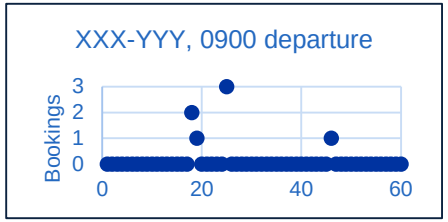

$$CV = \frac{\sigma}{\mu} = 4.0$$



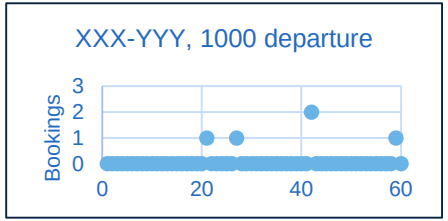

$$CV = \frac{\sigma}{\mu} = 4.2$$



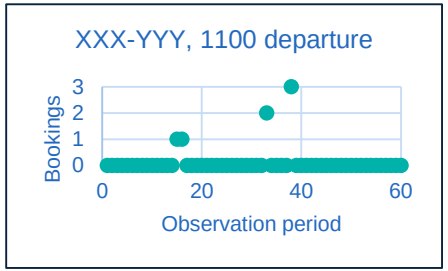
# Time windows assume discrete time preferences and are used to aggregate demand across substitutable itineraries



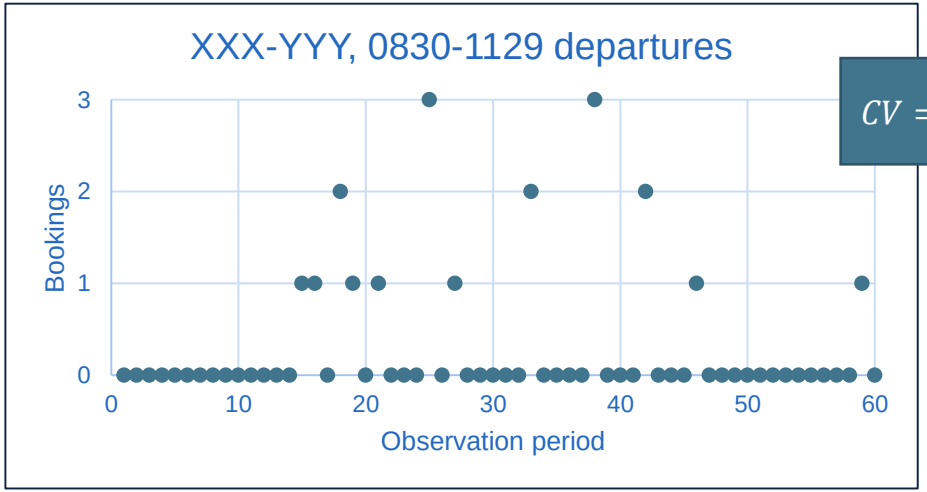

$CV = 4.2$



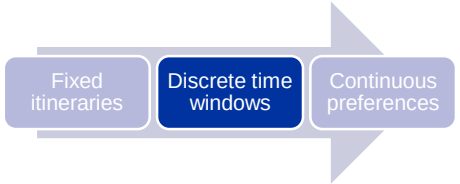

$CV = 4.0$



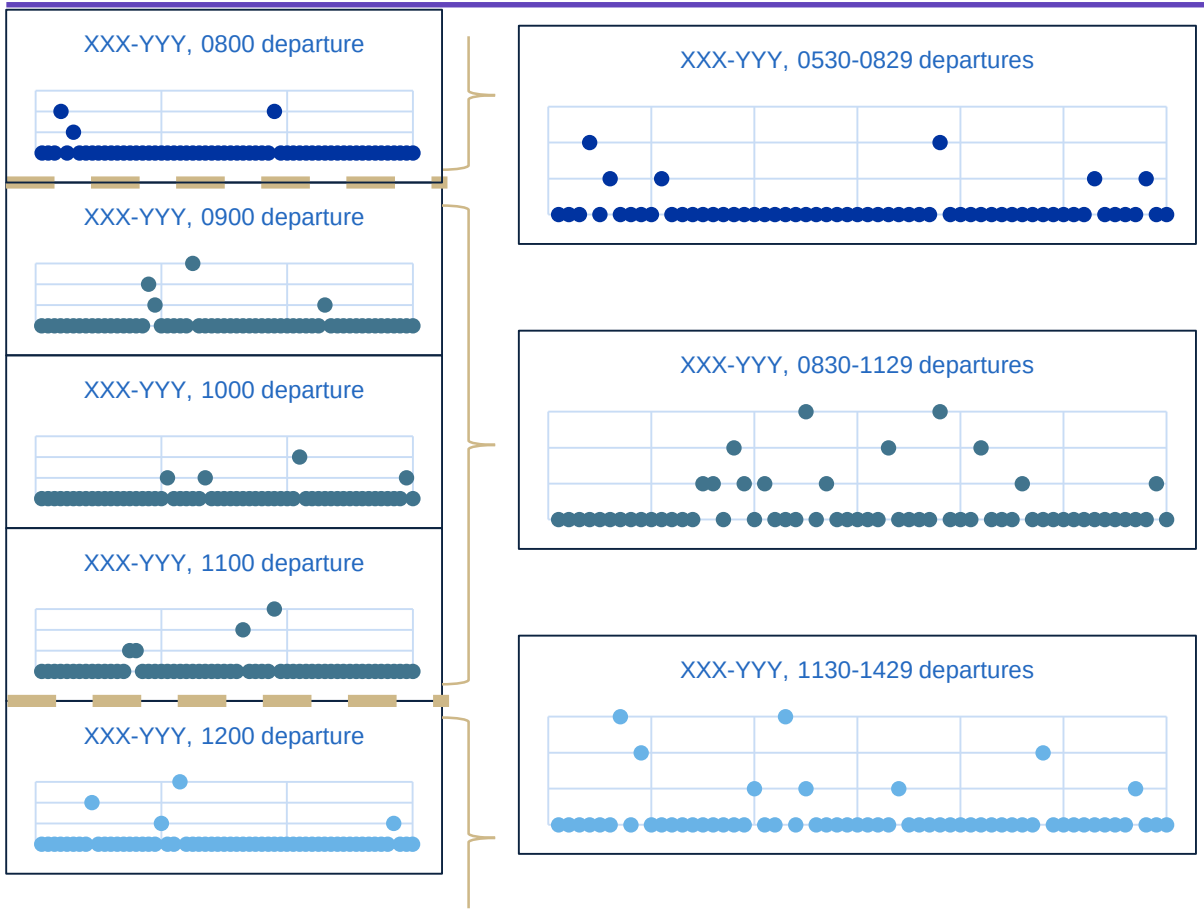

$CV = 4.2$

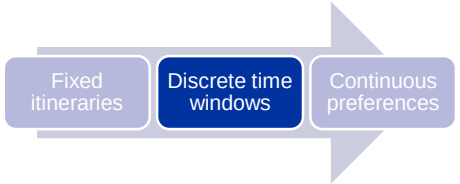



$CV = 2.3$



# Discrete time windows imply independence across time windows and are subject to other calibration decisions

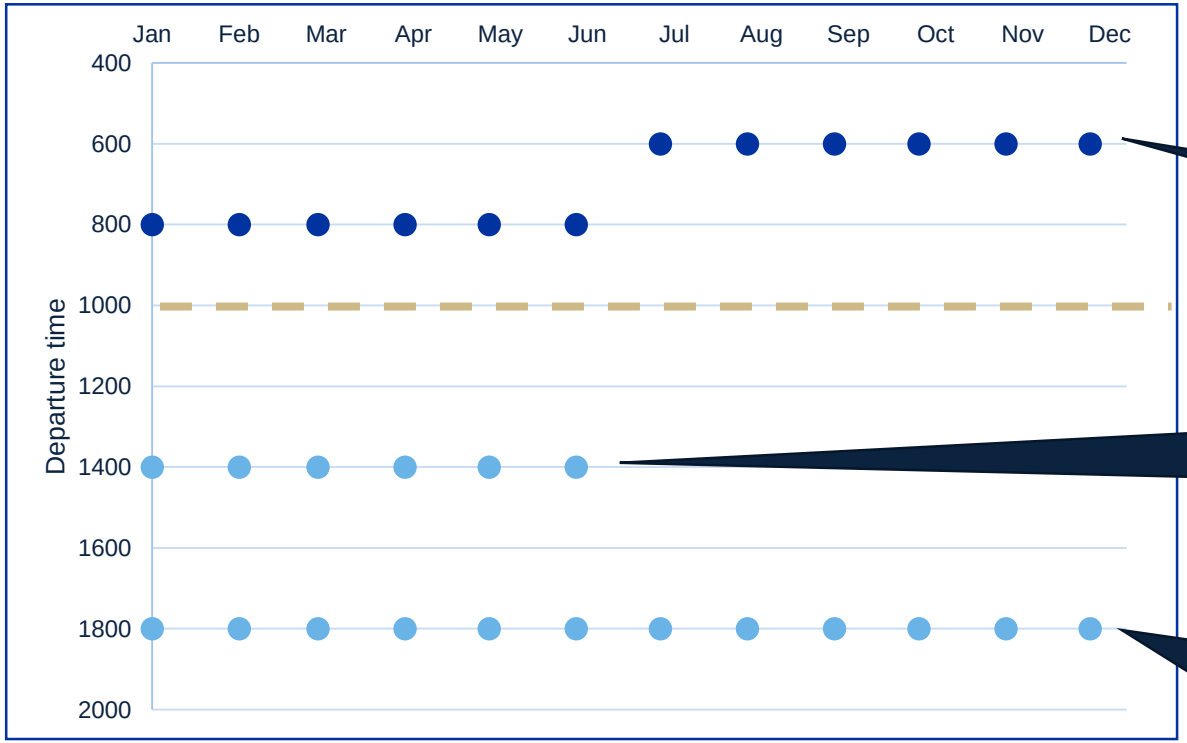


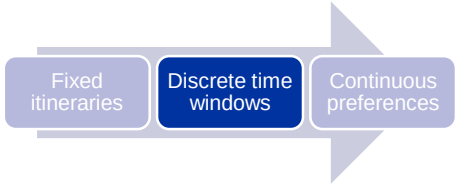




But why these exact choices?

Clustering algorithms can be sensitive to parameter choices and can lead to second-guessing

# Discrete time windows also don’t easily account for schedule changes that could affect preferences



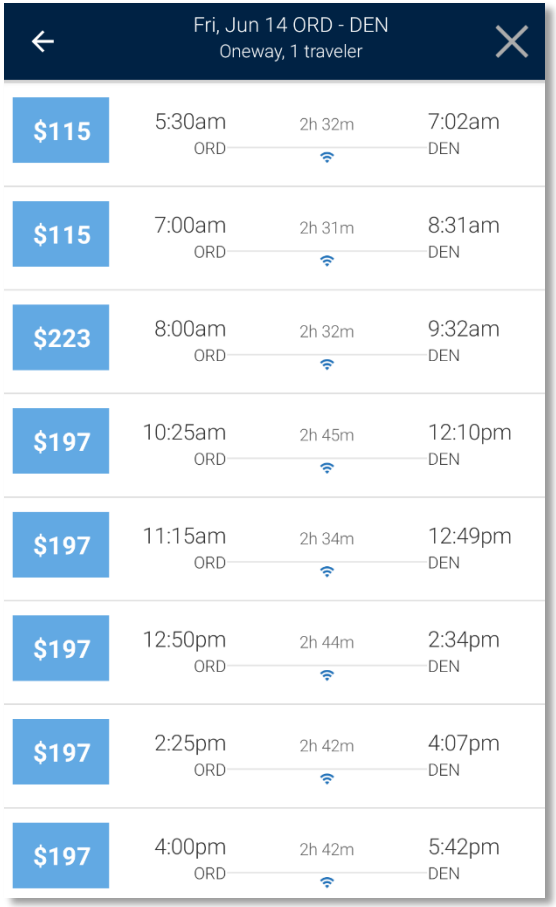



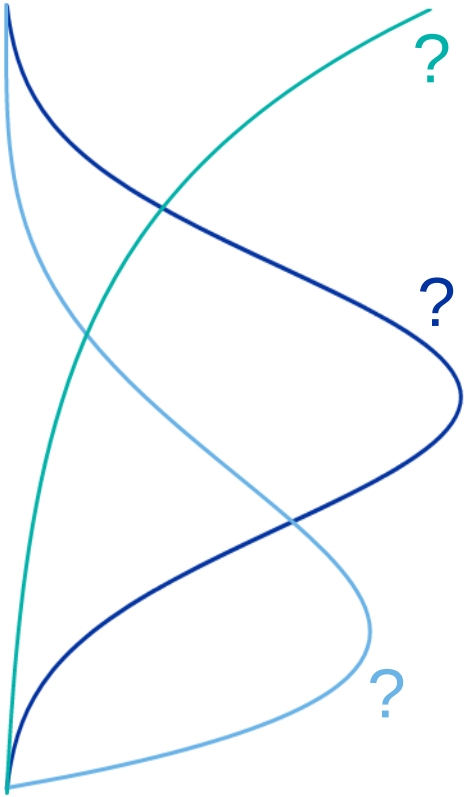

This flight might now be less preferable, but discrete time windows would not capture the reduction in demand

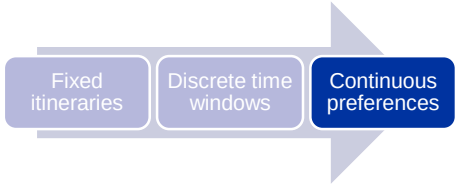
Customers with strong mid-day preferences might choose to take connecting options instead

Some demand will spill onto this 1800 flight (and maybe even some to 0600), but precisely how much will depend on preferences

# True preferences are probably continuous; can we extend the discrete model?

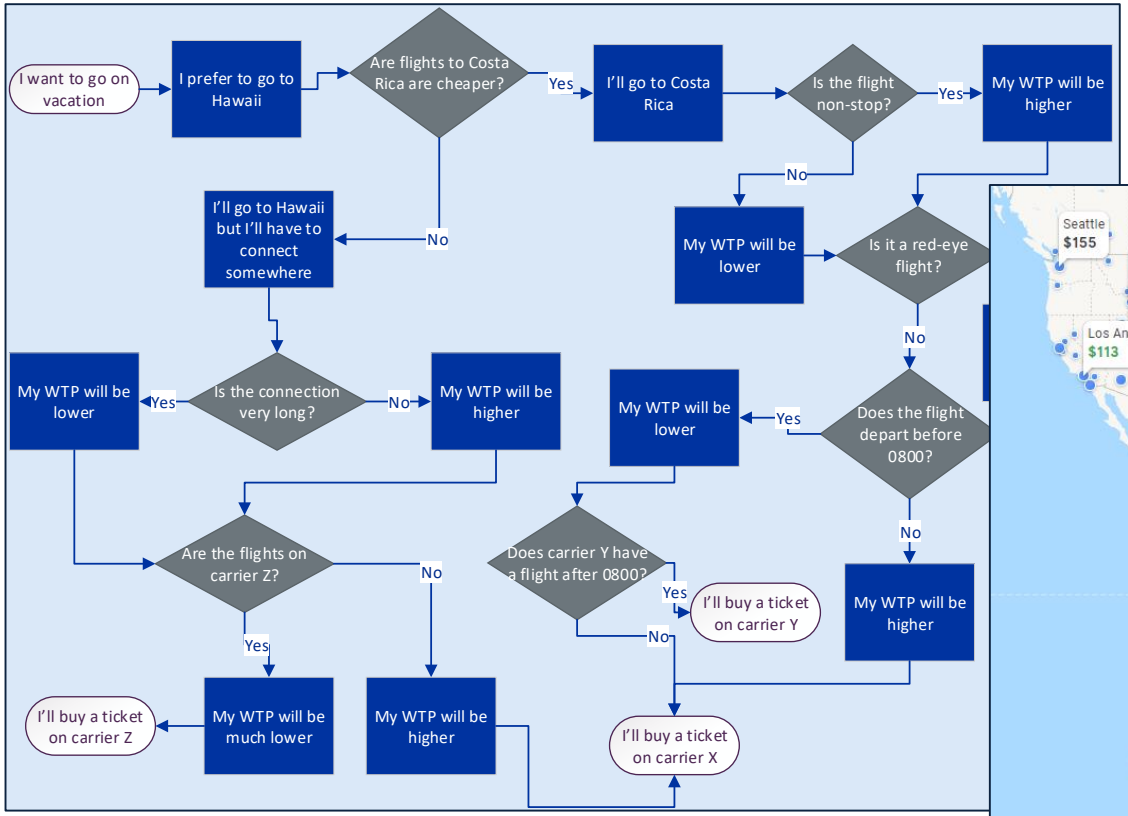


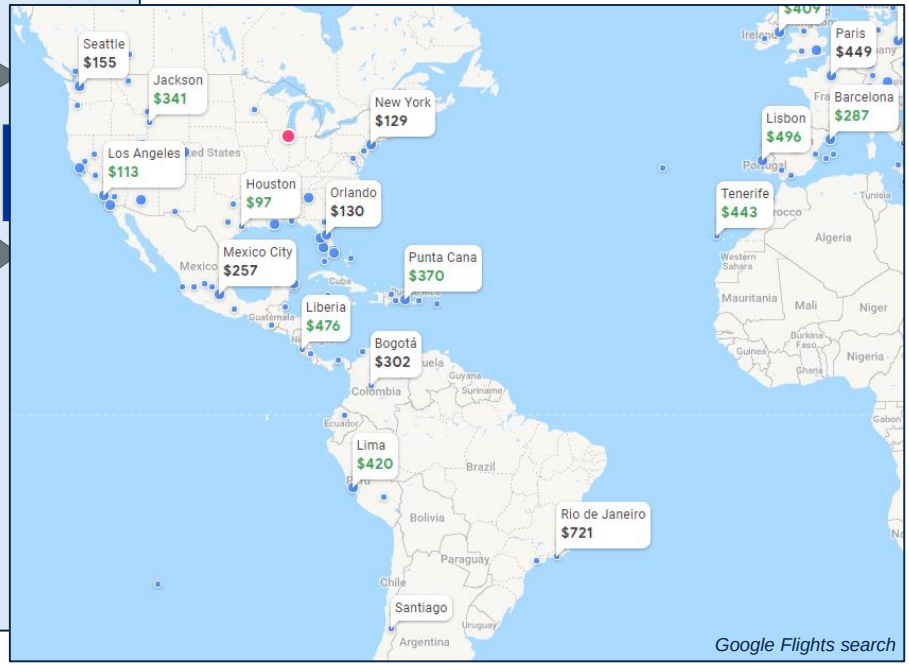




A simple extension to a continuous model might not easily capture interactions across itineraries

# In reality, customers consider an incredibly large number of factors




*Google Flights search*

# A passenger choice model that understands the tradeoffs between itinerary attributes can help adapt the forecaster to network changes

**Passenger Choice Model**

Market A-C
Departure: 6/20/2019

- Itinerary 1
  A-C at 7 am, 3 hours,
  A320, $200: **40%**
- Itinerary 2
  A-B-C at 2 pm, 5 hours,
  E175-A319, $100: **18%**
- Itinerary 3
  A-C at 11 am, 3.5 hours,
  A319, $150: **12%**
- Itinerary 4
  A-C at 5 pm, 3 hours,
  A320, $150: **30%**

Total daily demand forecast in a market → Itinerary share from passenger choice model → Itinerary level demand

# A lot of work has been done to understand the formulation of passenger choice modelling for airline industry

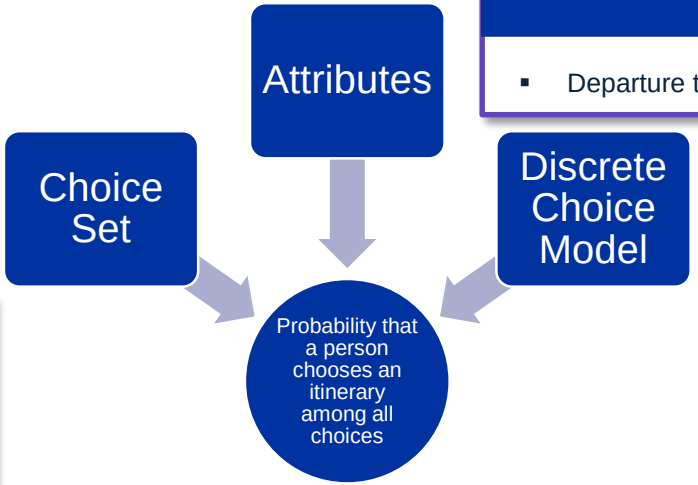
**Network Planning**

- Conditional logit models
- Time of day preference usually not handled explicitly
- Fare sensitivity not handled, uses average fares for each itinerary in a market

**Revenue Management and Pricing**

- Virginie Lurkin, Laurie Garrow, (2017) : Estimates an itinerary choice model with time-of-day preferences and controls for price endogeneity using 2-stage control function, and compares the results of baseline MNL model to more advanced GEV models
- C. Angelo Guevara(2016) : Assesses five methods (proxies, 2-stage control function, maximum likelihood, multiple indicator solution, latent variables) to correct for price endogeneity in discrete choice models
- Laurie A. Garrow (2013) : Formulates choice-based RM with a 2-step approach allowing estimation of all parameters including price, and estimates nested logit models accounting for censoring by applying the GEV choice-based sampling probabilities
- Sabre presentations on Dynamic Pricing and Joint Forecasting for Airline Pricing and Revenue Management (2018) use discrete choice models

# Literature search reveals the most commonly used choice model to analyze customer travel preferences is conditional logit




**Attributes**

- Departure time, equipment, elapsed time, fare, etc.

**Choice Set**

- For a given market, departure date and days from departure, list of all itineraries including non-stop and connection

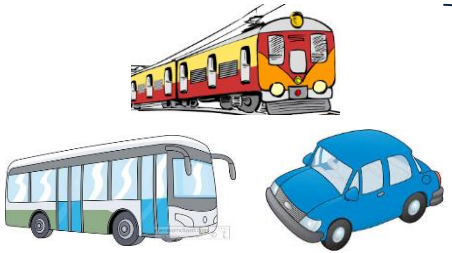
**Discrete Choice Model**

- Conditional Logit Model

**Probability**

- $P_i = \frac{\exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}{\sum_{j=1}^{J} \exp(\boldsymbol{x}_j^t \boldsymbol{\beta})}$, $x_i$: value of predictors, $\boldsymbol{\beta}$: coefficients
- Coefficients obtained by Maximum Likelihood Estimation (MLE)

# Traditional applications of conditional models have low computing complexity



**Choice attributes ($X_i$)**

- Travel Time
- Travel Cost
- Mode preference

**Passenger attributes ($Z_{ij}$)**

- Income
- Location
- Distance to Station

**Passenger Utility ($U_{ij}$)**

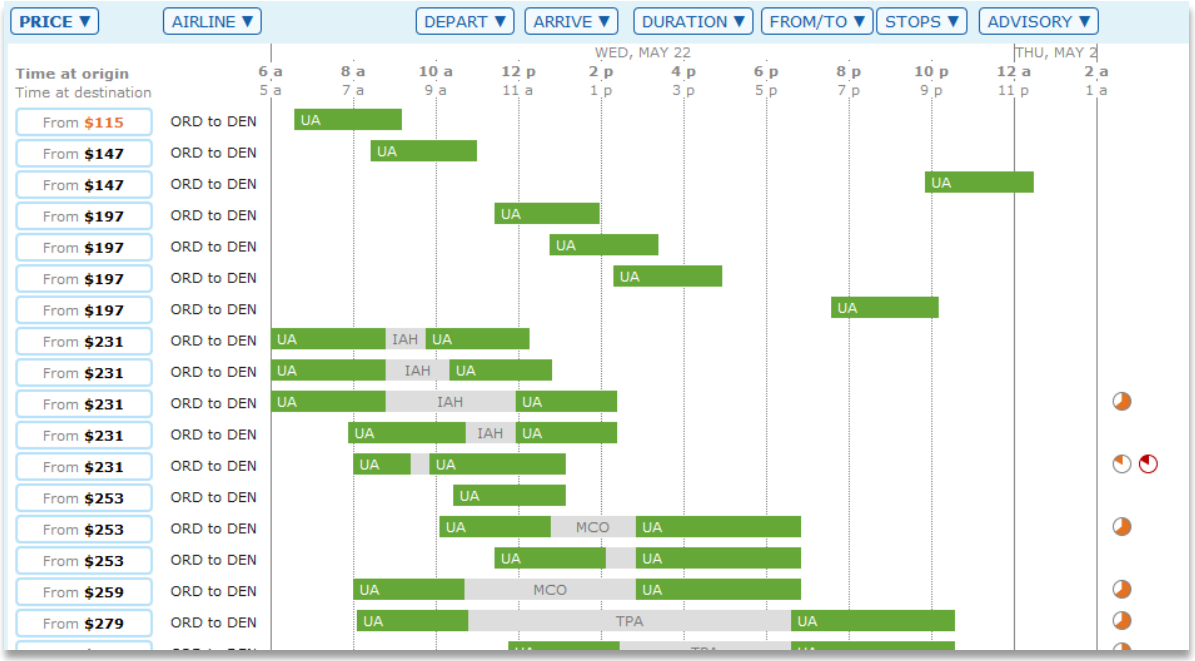
$$X'_i \beta + Z'_{ij} \gamma$$

***Choice Probability***

$$P_{ij} = \frac{\exp(\boldsymbol{U}_{ij})}{\sum_{i=1}^{3} \exp(\boldsymbol{U}_{ij})}$$

- Individual level records
- Less number of alternatives
- Choice set remain similar for individuals
- Multiple packages available to solve efficiently
  - SAS - MDC
  - R - Mlogit

# High number of alternatives for airline travel creates computation challenges




## Choice attributes ($X_i$)

- DOW - Time of day
- Non-Stop/ Connecting
- Elapsed time
- Price
- Tight connection
- Equipment

## Passenger attributes ($Z_{ij}$)

- Business/Leisure
- Market

*Too many alternatives -> No longer can use individual level records*
*Choice set changes with schedule changes and market*

Source: https://matrix.itasoftware.com/

# Limited options exist to solve the problem efficiently

- "Larch" in Python was recently developed to improve the efficiency of the estimation

**Computation times[1]**

| | Larch | Biogeme | Stata |
|---|---|---|---|
| Data loading | 6.63 secs | 9 mins 10 secs | 4.87 secs |
| MNL | 0.72 secs | 3 mins 9 secs | 15.34 secs |
| 2-level NL | 2.06 secs | 286 mins 34 secs | 7 mins 37 secs |
| 3-level NL | 2.30 secs | NA | 21 mins 41 secs |
| Constrained OGEV with $T = 1$ | 2.14 secs | 3 days 23 hours | NA |
| Constrained OGEV with $T = 2$ | 2.50 secs | 4 days 23 hours | NA |

- Any statistical software using Maximum likelihood function
  - Maximize the likelihood function

$\hat{\beta}=\text{argmax}_{\beta}[\ln(L(\beta))]=\text{argmax}_{\beta}[\sum_i lnP_i = \sum_i \sum_j y_{ij} lnP_{ij}]$ *where $y_{ij}$: 1 if individual i made choice j and 0 otherwise*

1: LARCH: A package for estimating multinomial, nested, and cross-nested logit models that account for semi-aggregate data, AGIFORS 2016

# Initial approach was to use discrete time intervals to model time of day preference

## Discrete Time of Day Approach

**Conditional Logit Model**

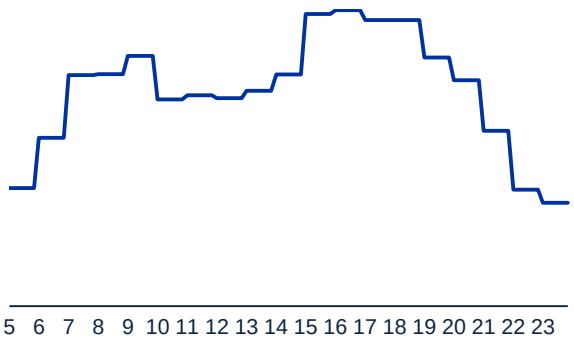
- Utility function = $\beta_1 \text{TOD_5}_i + \ldots + \beta_{19}\text{TOD_23}_i + \beta_{20}\text{Average_Fare}_i + \beta_{21}\text{Narrow_Body}_i + \beta_{22}\text{Wide_Body}_i + \ldots$

**Time of day preference**

- $\hat{\pi}_i = \frac{\exp(\hat{\beta}_i)}{\sum_{j=1}^{19}\exp(\hat{\beta}_j)}$ for $i$=1,…,19

**Controlling for other factors, their coefficients cancel out in $\hat{\pi}_i$**

## Discrete Time of Day Curve




**Cons**

- Same time of day preference within the discrete interval
- Boundary issues

# Moved to continuous modeling with Fourier Series to address the issues with discrete time of day modeling

## Continuous Time of Day Approach

**Conditional Logit Model**

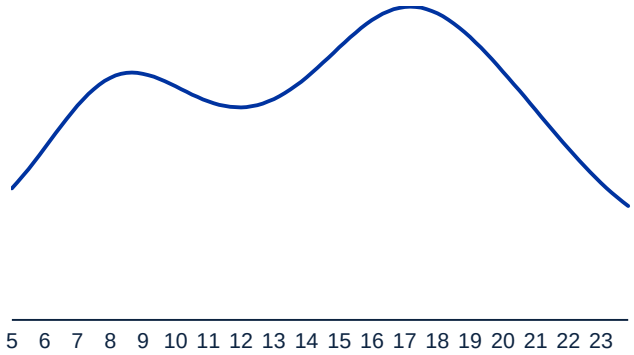
- Utility function = $\sum_{p=1}^{P}[\beta_{s,p} \sin\left(2\pi p \frac{DepTm}{1440}\right) +$
- $\beta_{c,p} \cos\left(2\pi p \frac{DepTm}{1440}\right)] +$
- $\beta_1$Carrier$_i$+ $\beta_2$Narrow_Body$_i$ + $\beta_3$Wide_Body$_i$ + $\beta_4$Average_Fare$_i$+…

**Time of day preference**

- $\hat{\pi}_i = 2\pi \frac{DepTm}{1440}$ for *each minute in the day*

**Controlling for other factors, their coefficients cancel out in $\hat{\pi}_i$**

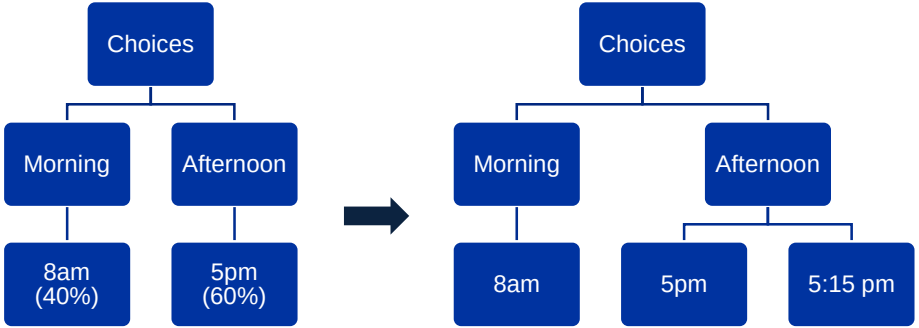
## Continuous Time of Day Curve




**Pros**

- More granular preference throughout the day
- Smooth transitions

# Nested structures can be explored to improve the model accuracy



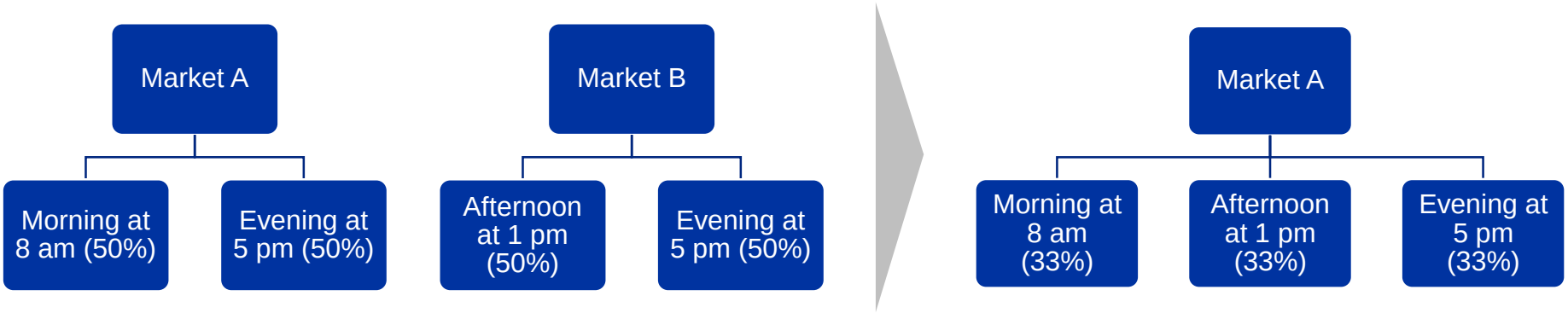

| Share | Continuous time of day without nest | Continuous time of day with nest | Independent time windows |
|---|---|---|---|
| Itinerary 1 at 8 am | 25% | 35% | 40% |
| Itinerary 2 at 5 pm | 37.5% | 32.5% | 30% |
| Itinerary 3 at 5:15 pm | 37.5% | 32.5% | 30% |

## Proportional substitution

Introduction of a new mode or improvements to any existing mode will reduce the probability of existing modes in proportion to their probabilities before the change ( implies equal competition between all pairs of alternatives )

$$\frac{P_{ni}}{P_{nk}} = \frac{e^{V_{ni}} / \sum_j e^{V_{nj}}}{e^{V_{nk}} / \sum_j e^{V_{nj}}} = \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni} - V_{nk}}$$

# Model segmentation is crucial for capturing time of day preference accurately




**Time of day preference of a new time channel in a market can be measured more accurately by segmenting similar markets together**

# Next steps include building a nested model structure and model segmentation

## Nested Model

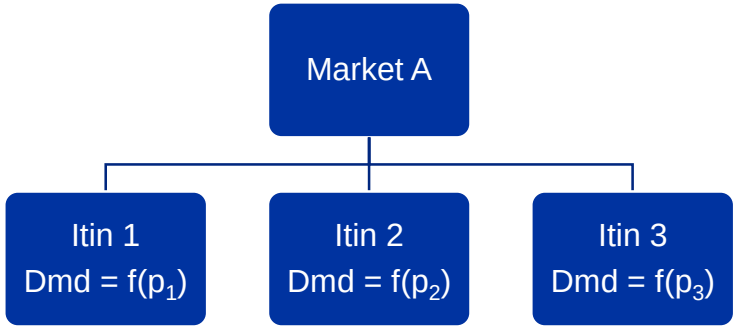
- Number of levels, cross nests, etc.
- Alternative hierarchy of variables

## Model Segmentation

- Market entities
- Passenger type, cabin, point of commencement

# In the future, optimization could also be improved to capture the interplay between price and itinerary quality

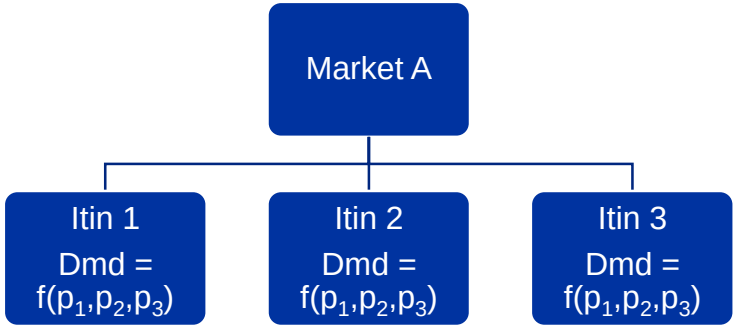
## Traditional Optimization




Max Revenue = Dmd(Itin 1)*$p_1$ + Dmd(Itin 2)*$p_2$+ Dmd(Itin 3)*$p_3$

s.t. capacity constraints

## Choice Based Optimization




Max Revenue = Dmd(Itin 1)*$p_1$ + Dmd(Itin 2)*$p_2$+ Dmd(Itin 3)*$p_3$

s.t. capacity constraints