Data-Driven Customer Segmentation and Personalized Information Provision in Public Transit

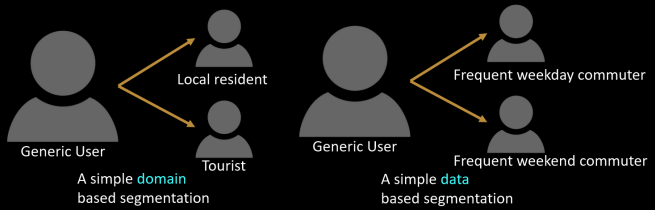
TransitData 2019 Workshop and Symposium
Paris, France
July 7-11, 2019

Abhishek Basu [1] Jinhua Zhao [2] Haris Koutsopoulos [3] Rabi Mishalani [4]

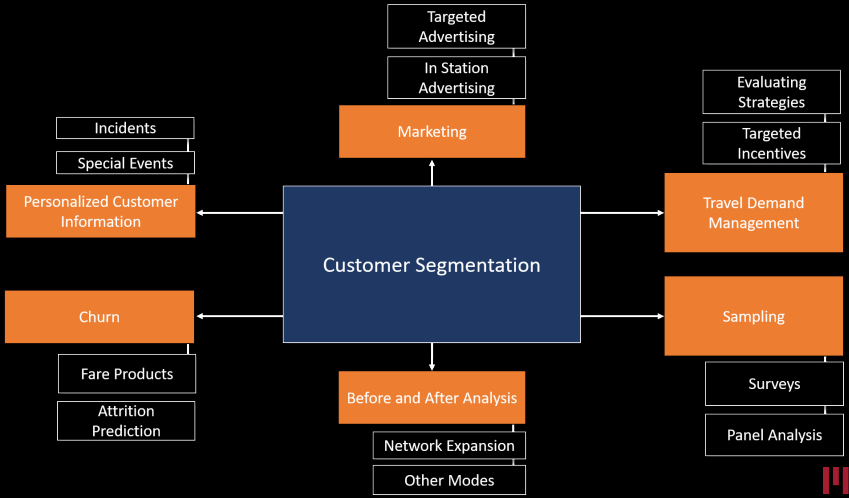
[1]Massachusetts Institute of Technology
[2]Associate Professor, Massachusetts Institute of Technology
[3]Professor, Northeastern University
[4]Visiting Professor, MIT; Professor, The Ohio State University

Customer Segmentation

- Pervasive strategy in general marketing. Product differentiation and Segmentation are two sides of a coin.
- Segmentation: Viewing a heterogeneous market as a number of smaller homogeneous markets in response to differing product preferences among important market segments.
- Domain knowledge/Rule based versus data based segmentation. Historically domain knowledge based, but pervasiveness of ADCS can change that.



Applications of segmentation



Overall structure of segmentation

- Short-term segments
  - Aim: Track finer changes in travel patterns.
  - Criterion: Users who have traveled at least once in the previous month.
- Long-term segments
  - Aim: Track long term trends in travel patterns.
  - Criterion: Users who travel at least half of the weekdays each month in the past year.

Short-term segments

Considers the spatial and temporal aspects of the user's travel pattern.

Features:

- SP: spatial probability or the number of trips that begin or end at a certain station
- TP: temporal probability or the number of times a trip terminates in a given hour

Methodology

- Use unsupervised clustering using k-means++.
- Select number of clusters based on DB index.
- Utilize May 2016 as the base data-set (100,000 users randomly sampled).
- We obtain 24 Spatial and 15 Temporal short-term segments.

Examples of short-term segments

Example of Spatial short term segment

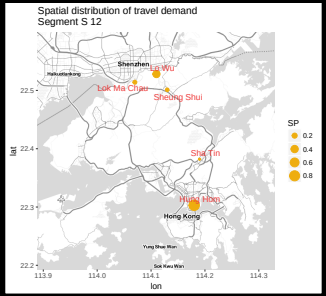


Figure: Long Cross Border users

Example of Temporal short term segment

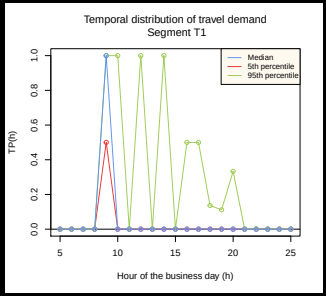
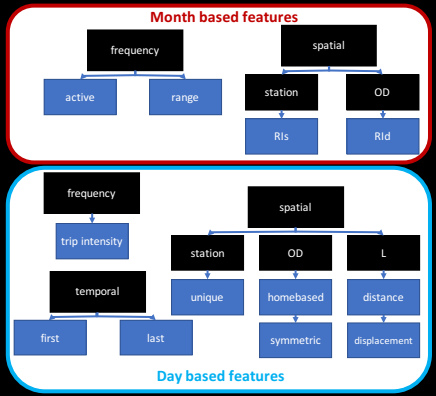


Figure: Work-based *less-variable*

Long-term segments

- A total of 26 features are created and utilized for clustering.
- Random sample of 100,000 long term users from January 2015 dataset.
- Eight long-term segments are obtained.
- Temporal stability of segments:
  - May 2015: 96.7%
    June 2015: 96.9%
  - Segmentation is found to be robust throughout the year 2016 as well (97% for eight segment solution).

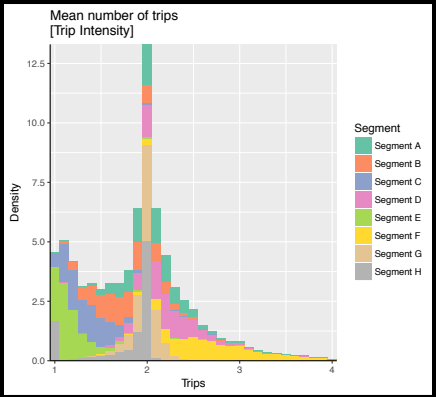


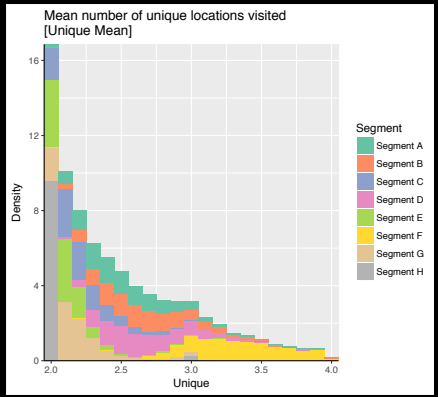
Long-term segments: results

- Group A: Long distance travelers with high deviation in *symmetric* and *displacement*.
- Group B: Travel on the least number of days with highest deviation in *first*.
- Group C: Single trip that begins/ends at home or return early.
- Group D: Travel on most days and make two or more trips on any given day.
- Group E: Single trip everyday in the evening (single OD) of a moderate distance.
- Group F: Highest *trip intensity*, cover long distances and visit most number of stations.
- Group G: Travel early in the morning, with highest *symmetric* and lowest *displacement*.
- Group H: All trips are HB, and single (O,D). Fixed time of *first* and *last* trips.

Long-term segments: results

Analyze the features to describe the corresponding segments.





Application

- The segments and the features used to create those could be used for many useful applications, an example of which is information provision.
- MTR, like many other transit agencies across the globe, has a dedicated smartphone app that can be used by a customer as a means to obtain information pertaining to the operation of the network among other things.
- The usual way for a customer to obtain relevant information is to actively choose (self-select) the types of information they need (e.g., the lines/stations/times), or in a broadcast mode where every customer receives this information. Here we focus on provision of this information through a passive mode (i.e. anticipating the need for information based on past travel behavior).
- The above is necessitated by the trend in information technology to personalize a user's experience, which involves providing relevant information by sensing the user's information needs (not broadcasting and not solely relying on user to self select).
- Here we demonstrate the applicability of the segmentation framework for provision of personalized information in the event of incidents/disruptions.

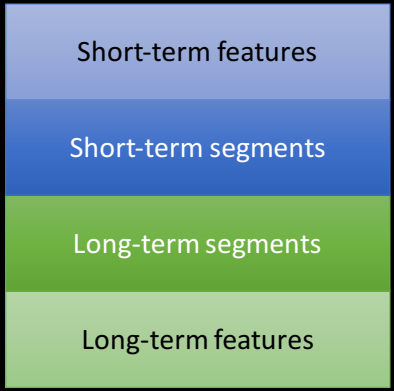
Personalized information provision

- In this study:
  - By providing information to users based on their past travel behavior, we do not imply that users cannot self-select. They may still choose to override this passive mechanism.
  - The mechanism we employ provides the agency the capability to cast a wider or smaller net (i.e. send information to more users or less) depending on factors such as severity of incident.
  - The users who are identified through this mechanism to be possible beneficiaries of this information are deemed to be informed on priority.
- The data used for this demonstration consists of 1.2 million users in the year 2016.
- The approach involves the following:
  - The short-term segments are used to identify and examine the affected users.
  - Then, the long-term segments are used to examine the nature/characteristics of the affected users.

Procedure

- *Aim:* Provide information pertaining to a given incident on a priority basis as inferred from their past spatio-temporal travel behavior.
- Sort segments based on station and time of incident.

  [Table of 24x90S and 15x21T]
- Pick top $\{n1, n2\}$ $\{spatial, temporal\}$ segments
- Provide personalized information to users that belong to these segments.

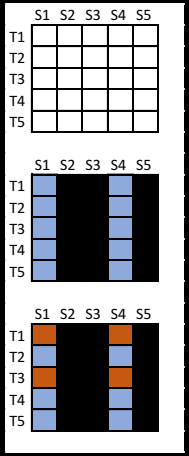


Select relevant short-term segments by sorting using features

Filtering:

- Incident occurs at Station x and at hour h.
- Sort segments in descending order based on SP(z(x)) and TP(h), for station and time of incident.
- Pick top $\{n1, n2\}$ $\{spatial, temporal\}$ segments.
  In this example: $\{2, 2\}$
- Send information to filtered list of users.

These short-term segments also yield information on what other stations (& times) these users would access.

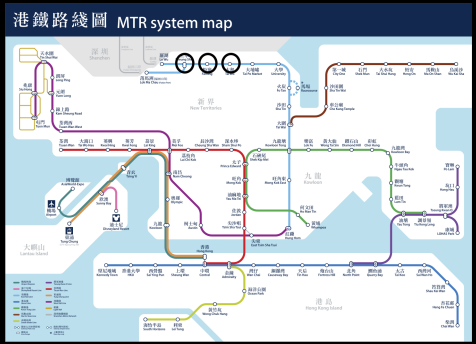


Case study

Incident:

- Date: 08/25/2016
- Line: East rail line (EAL)
- Stations: Sheung Shui, Fanling, Tai Wo
- Start time: 10:51:00
- End time: 15:39:00

Utilize July 2016 clusters as base.
Scope: Inform users who may access the three affected stations.



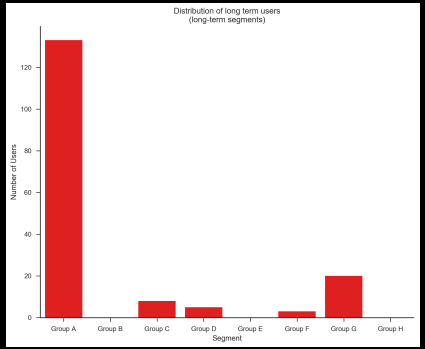
Analyze efficacy of the personalized information provision system

- A maximum of 5,381 affected passengers may be informed.
- Some users who were affected and did belong to the informed segments in August 2016, but not in July 2016.
- There are 446 (201S+245T) unfavourable shifts.
- We reach 4,027 (effectively 81%) of the users affected in the incident.
- Reduction compared to broadcast mode 1,218,876 $\rightarrow$ 300,131
- We see a 75% effective reduction in spam.

Impact on long-term users

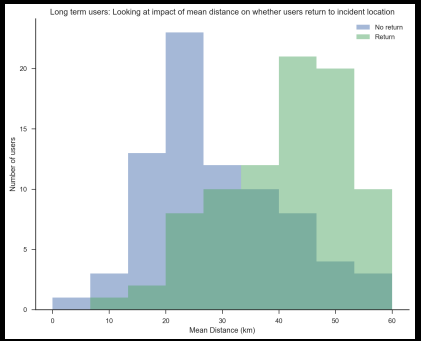
The agency may examine long-term segments (for the long-term users) to further understand their characteristics.

- Predominantly, a single long-term segment was effected by the incident - Group A.
- Travel the longest distances: Considerable number of cross-border travelers.
- High deviation in mean symmetric: important to provide information to ensure return trip.
- High distance and displacement deviations: Make one-way trips or they actively switch modes - Group G.



Impact on long-term users (cont'd.)

- Utilizing the features themselves.
- We recognize that mean distance traveled (& deviation) is a distinguishing feature of Group A.
- Consider mean distance traveled among those who made the return trip v/s those that did not.
- Clearly users that did return back to the affected stations later in the day are those who travel longer distances on average.
- This knowledge can help set further filters on the type of information to be provided to users in the future.



Summary

- A two-tiered data-driven segmentation was created in this study. In the first tier, short-term travel features were explored where the specific times of travel and stations visited by a customer were used to identify segments.
- In the second tier, long-term features such as distance of travel, number of trips from inferred home location, etc. were explored.
- The knowledge of segments obtained from the above methodology could be used to identify potential customers who may benefit from certain types of information.
- In this study, information provision during service disruptions was analyzed. It was found that by applying the developed segmentation framework, the transit agency could provide targeted information to customers who would be affected due to a disruption on a priority basis.
- With the availability of new sources of data (e.g., agency's smartphone app) in the near future, the framework presented in this study could be extended potentially to provide more targeted and accurate information.