Cluster Analysis in market segmentation research

Đỗ Huệ Hương, PhD
Outline

- Cluster Analysis: principles
- Cluster Analysis: practices
- Example: profiling visitors to Cat Tien National Park
Cluster Analysis

- Exploratory multivariate data analysis technique
- Grouping customers
- Different classes of data: psychographics, behaviour, product ratings, usage information, perceived needs or benefits
- CA always forms clusters whether ‘natural’ clusters exist or not
- Solid conceptual support from the literature (Hair et al., 2010)
Market segmentation

“a process of dividing customers whose valuations of a product or service vary greatly into groups or segments containing customers whose valuations vary very little within the group by vary greatly among groups”

(Lilien & Rangaswamy, 1998, p. 56)
Market segmentation

- Better understand customers → target marketing efforts to the right segments
- Homogeneity
- Parsimony
- Accessibility

(Lilien & Rangaswamy, 1998)
What to look at when clustering?

- Separation
- How many observations per cluster?
- Cluster profiles
- Validation

(STAT, 2010)
CA decision process

- Objectives of CA
- Research design
- Assumptions in CA
- Deriving clusters and assessing overall fit
- Interpretation of the clusters
- Validation and profiling of the clusters
Example: Profiling visitors to Cat Tien National Park
Overall study
Stage 1 - Objectives of CA

- Research questions
- Selection of clustering variables
Research background

• Ecotourism as a major component of global tourism (Weaver, 2008) with an estimation of contribution up to 20 per cent by UNWTO (Wight, 2001)

• Dominance of the Western conventional market (Eagles & Higgins, 1998)

• Increasing evidence of potential sizable markets in Asia (Cochrane, 2006; Weaver, 2002)
Research background (cont.)

- Two mandates of protected areas: to protect nature and to accommodate visitors

- New prosperous middle class in Viet Nam

- Increasing trend of visitation in protected areas (Cochrane, 2007)

- Potential incompatibility of the two mandates → ecotourism

→ This requires a better understanding of ecotourist component of protected area visitation
Literature review

Ecotourism

Ecotourist

Hard-soft dimension

Demographics

Psychographics

Behaviours
Definition of ecotourism

Three core criteria of ecotourism (Blamey, 1997)

- Nature-based attractions
- Learning/education motives
- Environmental, socio/cultural and economic sustainability - Triple bottom line (TBL)
Ecotourists

Definition:

- **Supply side:** people who visit a relatively wild and undisturbed areas (Ceballos Lascuráin, 1986)

- **Demand side:** people have at least an ecotourism experience (Blamey, 1995)
# Ecotourist typologies

<table>
<thead>
<tr>
<th>Source</th>
<th>Hard</th>
<th>Soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual</td>
<td>Deep ecology</td>
<td>Shallow ecology</td>
</tr>
<tr>
<td>Naess (1973)</td>
<td>Hard</td>
<td>Soft</td>
</tr>
<tr>
<td>Acott et al. (1989)</td>
<td>Hard-core</td>
<td>Dedicated</td>
</tr>
<tr>
<td>Lindberg (1991)</td>
<td>Self-reliant ecotourism</td>
<td>Small group ecotourism</td>
</tr>
<tr>
<td></td>
<td>[based on a sample of 545 general travellers passing through Cardwell, northern Qld]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24%</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>[based on a sample of 507 users of NSW state forests during time of participation]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18%</td>
<td>22%</td>
</tr>
<tr>
<td></td>
<td>[based on a sample of 207 travellers through Belize’s international airport]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3%</td>
<td>45%</td>
</tr>
<tr>
<td></td>
<td>[based on a sample of 245 visitors who have recent nature-based travel experience in North Carolina and donate money or belong to nature/environmental organisations]</td>
<td></td>
</tr>
<tr>
<td>Diamantinis (1999)</td>
<td>Frequent</td>
<td>Occasional</td>
</tr>
<tr>
<td></td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>[based on a sample of 1760 UK residents from databases of ecotourism-related tour operators and organisations]</td>
<td></td>
</tr>
<tr>
<td>Weaver &amp; Lawton (2002)</td>
<td>Harder</td>
<td>Structured</td>
</tr>
<tr>
<td></td>
<td>34%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>[based on a sample of 1180 overnight patrons of two ecolodges in Lamington National Park, southeastern Qld]</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Weaver (2008)
Hard-Soft ecotourists
Motivational/behavioural segmentation

A research in Lamington N.P., Australia. Adapted from Weaver and Lawton (2002)
Hard-Soft ecotourists (cont.)

Variables for segmentation & comparison

Demographics
- Age
- Gender
- Income
- Education

Psychographics
- Motivations
- Values
- Attitudes
- Self definition

Behaviour
- Length of stay
- Group size
- Trip planning
- Preferred activities
Research problem

- Lack of research on core criteria and hard-soft typology of Asian and Vietnamese ecotourists

- Lack of research on comparison between Western and Vietnamese ecotourists in the same site

→ To what degree do Vietnamese & Western ecotourists in Viet Nam differ with regard to their ecotourism affiliation?
## CA variables

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature-based, Harder</td>
<td>3</td>
</tr>
<tr>
<td>Nature-based, Softer</td>
<td>3</td>
</tr>
<tr>
<td>Learning, Harder</td>
<td>3</td>
</tr>
<tr>
<td>Learning, Softer</td>
<td>3</td>
</tr>
<tr>
<td>Sustainability, Harder</td>
<td>3</td>
</tr>
<tr>
<td>Sustainability, Softer</td>
<td>3</td>
</tr>
<tr>
<td>Asian distinctiveness</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>24</strong></td>
</tr>
</tbody>
</table>
Stage 2 - Research design in CA

- Sample size
- Detecting outliers
- Measuring similarity/distance
- Standardizing data
Research design

- Sample size: 500 VN - 500 Western visitors
- Stratified sampling, exit survey
- 1082 usable questionnaires (500 VN - 582 Western) after excluding outliers
- Cronbach alpha: 0.662
- Squared Euclidean distance measure
- No need for standardization because all clustering variables are measured on the same scale (1-5)

\[ \text{Distance} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \]
Stage 3 – Asumptions in CA

- CA is not a statistical inference technique but a method for quantifying the structural characteristics of a set of observations.
- Strong mathematical properties but not statistical foundations.
- No requirement on normality, linearity, homoscedasticity.
- Two critical issues:
  - Representative of the sample
  - Impact of multicollinearity
Multicollinearity

- Impact of multicollinearity to clustering
- Is there any underlying factors among 22 variables?
- Factor analysis as a frequent used pre-processing technique prior to clustering (Dolnicar, 2002)
- Principal Component Analysis
Principal Component Analysis

- Each item correlates at least .30 with at least one other item
- Barlett’s test of sphericity is significant (<.05)
- Kaiser-Meyer-Olkin (KMO) ≥ .60
- Measure of sampling adequacy in anti-image correlation matrix ≥ .50
- Total variance explained ≥ .60%

(Hair et al., 2010)
Principal Component Analysis

- All Eigenvalues ≥ 1
- All communalities for each item ≥ .40
- Fator loading for each item > .50
  (Hair et al., 2010)

- Cronbach’s alpha at least .60 in exploratory research
  (Nunnally, 1978)
Stage 4 – Deriving clusters and assessing overall fit

- Select the partitioning procedure used for forming clusters
- Make the decision on the number of clusters to be formed
Types of CA

- **Hierarchical clustering**
  - No particular number of clusters identified in advance

- **Non-hierarchical (K-mean) clustering**
  - Number of clusters is specified
  - Assign cases into clusters
Hierarchical clustering

- Agglomerative methods
- Divisive methods

(Sayad, 2012)
Agglomerative methods

- Single-linkage
- Complete-linkage
- Average linkage
- Centroid method
- Ward’s method

(Hair et al., 2010)
Deriving clusters

- Run the C.A in SPSS
- Examine each solution
Stage 5 – Interpretation of the clusters

- Run ANOVA to compare differences among groups
- Finalise cluster solution
- Name each cluster
6 cluster solution

- Sociable wildlife engagers (cluster 1, n=208 or 19%)
- Unenthusiastic visitors (cluster 2, n=239, or 22.2%)
- Typical visitors (cluster 3, n=283, or 26%)
- Service shunners (cluster 4, n=89, or 8%)
- Service seekers (cluster 5, n=92, or 8%)
- Classic Western ecotourists (cluster 6, n=171, or 16%)
Stage 6 – Validation and Profiling of the clusters

**Validation**
- Run C.A in a separate sample
- Divide the sample in two half
- Cross-tabulation
- Examine motivation, attitude items that have relationship with behaviour items, but not included in C.A

**Profiling of clusters**
- Discriminant Analysis
- Qualitative interviews
Conclusions

- C.A is the "art" of finding groups in data
- Theoretical background is important
- Random sampling
- Factorial Cluster Analysis
- Validation
References