Utilizing Big Data from Online Reviews to Understand Local Tourist Travel

Andrei Kirilenko
Juan Hernández
Svetlana Stepchenkova
Lijuan Su
Florida tourism

116.5 million visitors traveled to Florida in 2018 (Visit Florida)

• 111.8 million domestic visitors (+7%)
• 3.5 million Canadian visitors
• Other countries: 10.8 million
• cf: Australia 9.2 million (Tourism Australia 2019)
Study purpose

• Research goal: understand utilization of Florida recreational resources by Floridians

• Research questions:
  1. Can we use the attraction review data for market segmentation?
     • Identify groups of attractions that tourists tend to visit together
     • Identify the groups of tourists that tend to visit similar set of attractions
  2. Which social demographic and economic factors affect utilization of Florida recreational resources by Floridians?

• Data: social media (TripAdvisor reviews)

• Instruments: network analysis; spatial analysis
Data

• **TripAdvisor reviews**, Question 1:
  • 76,110 reviews of 8,049 *attractions* from 522 locations in Florida published by 14,088 reviewers from Florida

• **TripAdvisor reviews**, Question 2:
  • 431,829 reviews of Florida hotels (18%), restaurants (65%), and attractions (17%) from 591 locations in Florida published by 18784 reviewers from Florida
  • 60,294 reviews of hotels, restaurants, and attractions outside of Florida

• **Public profile information of users:**
  • self-identified place of living (city, state, and country)
  • declared travel interests (e.g., “Nature lover”)
  • Age and gender

• **Latitude and longitude** of the attractions and self-identified *place of living* of the reviewers were obtained from TripAdvisor search and from geonames.org ([Kirilenko and Stepchenkova, 2014](#)), respectively.

• Data was carefully cleaned for erroneous entries, such as attractions located outside Florida.

• **2017 American Household Survey**
Data description: Top destinations

Top 5 destinations
Orlando: 14.1%
St. Augustine: 6.8%
Key West: 5.4%
Tampa: 4.8%
Miami 3.3%
Research question 1:
Can we use the attraction review data for market segmentation?

• Market segmentation [. . .] consists of viewing a heterogeneous market (one characterized by divergent demand) as a number of smaller homogeneous markets (Smith, 1956)

• Tourism market segmentation: Selecting homogeneous groups in heterogeneous tourist market

• Goals: better tailored services resulting in customized tourist experience, higher satisfaction, repeat visitation, more revenues for business

• Approaches:
  • \textit{a priori} (common-sense): factors identifying traveler groups selected prior to the study. E.g., geographic segmentation; demographic segmentation; psychographic segmentation
  • \textit{a posteriori} (data-driven): identification of traveler groups result from the analysis of observed travel patterns
Visitation pattern as a network

• Attraction visitations can be presented as a bipartite network
  • Two types of nodes: attractions, interests
  • Edges: a tourist with interest $A$ visiting attraction $\alpha$.
  • Edge weight: number of tourists with same interest $A$ to same attraction

Unipartite

Attractions

Visits

Tourists
Visitation pattern as a network

• Attraction visitations can be presented as a bipartite network
  • Two types of nodes: attractions, interests
  • Edges: a tourist with interest A visiting attraction α.
  • Edge weight: number of tourists with same interest same attraction
Network community structure

- Community: semi-autonomous parts of the network.

- A network has a community structure if the network nodes can be grouped in such a way that the nodes in each group are connected more densely inside as compared with outside connections.

- Community detection: finding the best partition of nodes in groups or clusters in such a way that the density of links among nodes inside every group is higher than the density of edges among nodes belonging to different groups.


- Based on the maximization of the modularity function.

- The modularity measures the difference between the connectivity in every cluster and the connectivity in a null model.
# Interests– Attractions clusters

<table>
<thead>
<tr>
<th>Excitement</th>
<th>Nature</th>
<th>Heritage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interests</strong></td>
<td><strong>Attractions</strong></td>
<td><strong>Interests</strong></td>
</tr>
<tr>
<td>Family Vacationer</td>
<td>Orlando: Magic Kingdom Park</td>
<td>Nature Lover</td>
</tr>
<tr>
<td>Foodie</td>
<td>Orlando: Epcot</td>
<td>Peace and Quiet Seeker</td>
</tr>
<tr>
<td>Thrill Seeker</td>
<td>Orlando: SeaWorld</td>
<td>Beach Goer</td>
</tr>
</tbody>
</table>

**Finding 1:** Three attractions clusters
- **Excitement** – visiting primarily amusement parks
  - family vacationers, thrill seekers, and shopping fanatics:
- **Nature** - visiting beach and other nature-based locations:
  - beachgoers, nature lovers, and peace and quiet seekers
- **Heritage** - concentrating on cultural attractions: history, art, and architecture lovers who are also interested in local life

*Note: similar interests were found for Florida visitors, however the set of correspondent attractions was different*
Research question 2: Which socio-demographic and economic factors affect utilization of recreation?

• Unequal utilization of recreational resources:
  ▪ Affected by geographic, demographic, and socio-economic factors (e.g., Lindsey, Marai, & Kuan, 2001; Stodolska, 1998)
  ▪ Availability: reduced in minority and low income areas (Moore et al., 2008)
  ▪ Differences in income, age, and race between users and non-user of public recreation services (Howard & Crompton, 1984)

• Dependent variable: number of trips per person
  • County granularity

• Independent variables: socio-economic profile of the county

• Geographic position of the county
Social-demographic variables

Population

Hispanic Ratio

Black Ratio

Median Age

Median Income

Average Family Size
Distribution of the mean travel distance

Mean trip distance is correlated with:
- Within-state travel: distance to the center of Florida; longitude
- Out-of-state: latitude
## Regression Analyses

<table>
<thead>
<tr>
<th>Variables</th>
<th>Std. B</th>
<th>t</th>
<th>p-value</th>
<th>Std. B</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>-3.942</td>
<td>0.000</td>
<td></td>
<td>-0.906</td>
<td>0.368</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.110</td>
<td>1.010</td>
<td>0.317</td>
<td>-0.369</td>
<td>-3.487</td>
<td>0.001</td>
</tr>
<tr>
<td>Age</td>
<td>0.346</td>
<td>3.169</td>
<td>0.002</td>
<td>-0.030</td>
<td>-0.284</td>
<td>0.778</td>
</tr>
<tr>
<td>Income</td>
<td>0.466</td>
<td>4.414</td>
<td>0.000</td>
<td>0.506</td>
<td>4.915</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Lat</strong></td>
<td>-0.397</td>
<td>-2.234</td>
<td>0.029</td>
<td>-0.101</td>
<td>-0.571</td>
<td>0.570</td>
</tr>
<tr>
<td><strong>Long</strong></td>
<td>0.178</td>
<td>1.306</td>
<td>0.197</td>
<td>-0.266</td>
<td>-1.974</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Note: Union, Dixie, Franklin, Lafayette, Taylor counties have no with-in-state reviews, excluded.
Geographically weighted regression

- Linear regression model hints that the noticed relationships vary spatially
- GWR: extension of linear regression to spatial data
- The First Law of Geography: "everything is related to everything else, but near things are more related than distant things.“ (Tobler, 1970)
- Regression models ignore variation of a phenomena over space
- GWR (Fotheringham et al. 2002) explores whether locally weighted regression coefficients move away from their global values.
## Within-state travel: OLS vs. GWR

<table>
<thead>
<tr>
<th>Variables</th>
<th>Beta</th>
<th>OLS Coefficients</th>
<th>GWR Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Std. B</td>
<td>t</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.02</td>
<td>0.11</td>
<td>1.01</td>
</tr>
<tr>
<td>Age</td>
<td>0.15</td>
<td>0.35</td>
<td>3.17</td>
</tr>
<tr>
<td>Income</td>
<td>0.00</td>
<td>0.47</td>
<td>4.41</td>
</tr>
<tr>
<td>R2</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Finding 2: Two socio-demographic variables explain 49% of in-state travel variability

- **Age** – aged population take more travels
- **Income** – higher income population take more travels

GWR leads to a significant improvement in explanatory power of the model
Contribution and Implications

• The study demonstrated usability of the social network data and network analysis for the analysis of travel patterns and segmentation of different tourist types
  • A novel way to tourist segmentation via attractions that tourists visited at a destination
  • Analysis from the tourists point of view as opposed to a theoretical perspective
  • Tourist typology based on tourists’ interests, motivations, and activities at a destination, which is “custom-tailored” for a specific destination and firmly grounded in empirical data.

• Study 1: three attraction-tourist type clusters found
  • Attraction clusters for family vacationers and nature lovers, older urban travelers seeking peace and quiet, and those looking for thrills, good food and drinks as well as an attractive nightlife scene are clearly separated
  • Support in the tourist literature, as similar clusters of tourists interests (nature, heritage, entertainment) have also been identified in other studies.

• Study 2: utilization of attractions are strongly affected by socio-demographic profile of the county
  • GWR proved superior to regular linear regression method
  • Surprisingly, age and income were the only statistically significant variables (49% variance explained) despite significant heterogeneity in racial composition, unemployment rate, family size, etc.
  • Research in progress
Application examples

- Destination professionals would appreciate the proposed way to detect inconsistencies in how different origin markets view destination attractions and, thus, the ability to identify clusters of under-performing attractions, both image-wise and geographically.
- Which attractions can be co-promoted by DMOs?
- How the attractions are (re)viewed by different tourist types?
- Which attractions does a particular group of hotels serve?
- Which restaurants are preferred by visitors of particular hotels?
- What are the travel patterns and attraction utilization of different socio-economic population groups living in different parts of the state?
- What is an impact of tourist traffic on road system
Limitations

• Representativeness: Certain population groups are underrepresented
  • Example: Younger ages

• Model-related uncertainly: how the change in methodology will change results?
Questions?

Publications