Tourism Big Data Research

Bing Pan, Ph.D.
Department of Recreation, Park, and Tourism Management
College of Health and Human Development
Penn State University

SODA 501 Guest Lecture
February 1, 2018
TODAY’S TOPICS

What is tourism?
A Behavioral Model of Tourism Big Data
Relevancy and Bias of Search Data
...Website Traffic Data
...Social Media Data
...Mobile Data
The Stage and Future of Tourism Big Data
WHAT IS TOURISM?

Tourism: the sum of the phenomena and relationships arising from the interaction in generating and host regions, of tourists, business suppliers, governments, communities, and environments.

----- Tribe (1997)
A Behavioral Model of Forecasting Tourist Behavior with Big Data
THE STARTING OF FORECASTING WITH SEARCH DATA - 2009

United States Flu Activity

Influenza estimate

Google Flu Trends estimate • United States data

United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.
Search Data for Tourism Forecasting

Past forecasting methods are based on historic performance or macro-economic indicators, which rely on a consistent pattern of tourist activities and a stable economic structure.
Advantages of Using Search Data

Online data are real-time, high-frequency (daily and weekly instead of quarterly or annually), and sensitive to user behaviour's small changes.

Can we Forecast Weekly Hotel Occupancy data with search engine volume data? (Pan, Wu, & Song, 2012)
MORE ON TOURISM SEARCH DATA

• The additional search engine volume data of 5 keywords can increase the forecasting accuracy by at least 2-3 percent for Charleston, South Carolina (Pan, Wu, & Song, 2012)

• Baidu keyword search volume can improve Summer Palace visitor volume forecast by 15% - Huang etc. (2015)

• Baidu keyword search volume can help improve Hainan, China monthly visitor volume forecasts accuracy; Baidu performed better than Google (Li, Pan, Law, 2017)

• Forecasting in Cyprus with search data needs to correct for language and platform bias (Dergiades, Mavragani, Pan, 2018)
Figure 2. Market shares in tourists’ arrivals for Cyprus per country.
Thank you for stopping by.

Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and
THE INACCURACY IN GOOGLE FLU FORECAST (2014)

Google estimates more than double CDC estimates

Google starts estimating high 100 out of 108 weeks
Limitations of Search Queries

Changing Google interface;
Changing user behavior;
Old models do not work well...

In Tourism - do managers need 100% of forecasting accuracy? Good enough forecasting is good-enough.
FORECASTING WITH WEBSITE DATA

Website traffic of local convention and visitors bureau can increase hotel demand forecast by 6-10%, according to different measurements (Yang, Pan, & Song, 2015)

Using website data to identify potential markets with a gravity model (Park & Pan, 2017)
A gravity model of origin cities to Charleston, SC

\[ V_{ic} = \alpha_0 \text{Pop}_i^{\alpha_1} \text{Pop}_c^{\alpha_2} \text{GDP}_i^{\alpha_3} \text{GDP}_c^{\alpha_4} D_{ic}^{\alpha_5} \text{Price}_{ic}^{\alpha_6} D_{i}^{\alpha_7} D_{c}^{\alpha_8} \]

- Direct flight is expected to increase flight passengers by almost 200%, mobile devices by 14%, and hotel guests by 57%.
- Flight time, distance, and relative cost of flying show negative correlations.
- Income and population show positive correlation but the degree is not very great.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Flight passengers</th>
<th>Mobile devices</th>
<th>Hotel guests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
</tr>
<tr>
<td>Income</td>
<td>0.0139***</td>
<td>0.0223***</td>
<td>0.0149***</td>
</tr>
<tr>
<td>Population</td>
<td>0.0010***</td>
<td>0.0013***</td>
<td>0.0010***</td>
</tr>
<tr>
<td>Distance</td>
<td>0.0889***</td>
<td>-0.5255***</td>
<td>-0.1642***</td>
</tr>
<tr>
<td>Flight Time</td>
<td>-0.3908***</td>
<td>-0.6168***</td>
<td>-0.7282***</td>
</tr>
<tr>
<td>Distance*Flight Time</td>
<td>0.0044***</td>
<td>0.0492***</td>
<td>0.0249***</td>
</tr>
<tr>
<td>Relative cost</td>
<td>-1.5554***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance* Relative cost</td>
<td>-0.4885***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Flight</td>
<td>1.0843***</td>
<td>0.1355***</td>
<td>0.4615***</td>
</tr>
<tr>
<td>Constant</td>
<td>8.9778 ***</td>
<td>9.5030***</td>
<td>9.8146***</td>
</tr>
</tbody>
</table>
Website Data can Identify potential markets

1. Find the outliers
2. Find those that have the greatest difference
3. Compare web traffic to the potential
4. Red dot: potential market
Results: Potential markets

<table>
<thead>
<tr>
<th>MSA</th>
<th>Population</th>
<th>Income</th>
<th>Driving Time</th>
<th>Current Flight time</th>
<th>Expected Flight time</th>
<th>Web traffic</th>
<th>Market potential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portland-South Portland, ME</td>
<td>526,295</td>
<td>73,200</td>
<td>16h48</td>
<td>4h09</td>
<td>2h29</td>
<td>8282</td>
<td>241</td>
</tr>
<tr>
<td>Minneapolis-St. Paul-Bloomington, MN-WI</td>
<td>3,524,583</td>
<td>85,700</td>
<td>19h52</td>
<td>4h02</td>
<td>2h41</td>
<td>45,117</td>
<td>207</td>
</tr>
<tr>
<td>Jacksonville, FL</td>
<td>1,449,481</td>
<td>63,300</td>
<td>3h42</td>
<td>3h50</td>
<td>0h43</td>
<td>46,417</td>
<td>68</td>
</tr>
<tr>
<td>Allentown-Bethlehem-Easton, PA-NJ</td>
<td>821,173</td>
<td>71,200</td>
<td>10h45</td>
<td>3h40</td>
<td>1h36</td>
<td>12,230</td>
<td>64</td>
</tr>
<tr>
<td>Kansas City, MO-KS</td>
<td>2,087,471</td>
<td>74,700</td>
<td>15h59</td>
<td>4h05</td>
<td>2h18</td>
<td>28,905</td>
<td>44</td>
</tr>
<tr>
<td>Language</td>
<td>Sessions</td>
<td>% Sessions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en-us</td>
<td>456</td>
<td>44.93%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(not set)</td>
<td>192</td>
<td>18.92%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secret.google.com</td>
<td>134</td>
<td>13.20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zh-cn</td>
<td>52</td>
<td>5.12%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en-gb</td>
<td>40</td>
<td>3.94%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pt-br</td>
<td>24</td>
<td>2.36%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitaly rules google</td>
<td>17</td>
<td>1.67%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ko</td>
<td>10</td>
<td>0.99%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en-au</td>
<td>8</td>
<td>0.79%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ko-kr</td>
<td>6</td>
<td>0.59%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BIASES IN WEBSITE TRAFFIC DATA

More searches from locals, school students, crawlers, and spammers;

More big data sources (search, website, flight passenger traffic, weather, events, etc.) do not mean more accurate forecasting results since they are highly correlated (Pan & Yang, 2017)

We need more diverse data sources.
RELEVANCY IN SOCIAL MEDIA DATA

Online reviews are used for hotel decisions; age and gender differences in using online reviews (Gretzel & Yoo, 2008)

Text analysis can tell hotel experience and satisfaction from online reviews (Xiang, Schwartz, Gerdes, & Uysal, 2015)
Dianping is a Chinese service review site similar to yelp.com
29 restaurants in Haikou, Hainan Province, China
1,646 intercept surveys obtained from 29 restaurants (2015 to 2016);
39,803 online reviews crawled for 30 restaurants in the fall of 2016 – most were posted in the last 3-4 years.
Both online reviews and surveys contains ratings on overall, taste, service, and atmosphere.
COMPARISON OF DIANPING AND SURVEY RESULTS

Comparison of Overall Ratings

Comparison of Taste Ratings

Recreation, Park, and Tourism Management
COMPARISON OF DIANPING AND SURVEY RESULTS

Comparison of Atmosphere Ratings

Comparison of Service Ratings

Online Reviews vs. Survey Results

\[ r = 0.422 \]

\[ \text{sig.} = 0.023 \]

\[ r = 0.413 \]

\[ \text{sig.} = 0.026 \]
Online reviews are only marginally useful for evaluating atmosphere and service of restaurants in China; Online reviewers are younger, more educated, and higher income earners; Taste is very subjective, especially for Chinese dishes?
FLICKR DATA AND PARK VISITS

MOBILE DATA VERSUS INTERCEPT SURVEY

WIRELESS NETWORK OPERATORS

Signaling Data → ANONYMIZING → Anonymous & Aggregate Data

AirSage Wireless Signal Extraction (WISE) Technology Platform

AirSage WiSE Analytical Platform

LOCATIONS • PATTERNS • INTELLIGENCE

Target Marketing
Out-of-Home Media
Tourism Studies

Equity Research
Insurance Analytics
Real Estate Site Selection
Transportation Planning
Emergency Management

AirSage
VISITOR ORIGIN STATES

Mobile Data

Survey Results

Human Development

Recreation, Park, and Tourism Management
MOBILE DATA

Mobile data can provide county-level visitor origin.
## The Origin Areas of Charleston Tourists

<table>
<thead>
<tr>
<th>Region</th>
<th>AirSage Percentage</th>
<th>AirSage Rank</th>
<th>Survey Percentage</th>
<th>Survey Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southeast</td>
<td>74.63%</td>
<td>1</td>
<td>44.28%</td>
<td>1</td>
</tr>
<tr>
<td>Mideast</td>
<td>10.93%</td>
<td>2</td>
<td>19.79%</td>
<td>2</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>6.15%</td>
<td>3</td>
<td>13.11%</td>
<td>3</td>
</tr>
<tr>
<td>New England</td>
<td>2.98%</td>
<td>4</td>
<td>6.40%</td>
<td>4</td>
</tr>
<tr>
<td>Far West</td>
<td>1.82%</td>
<td>5</td>
<td>4.13%</td>
<td>7</td>
</tr>
<tr>
<td>Southwest</td>
<td>1.56%</td>
<td>6</td>
<td>4.65%</td>
<td>6</td>
</tr>
<tr>
<td>Plains</td>
<td>1.09%</td>
<td>7</td>
<td>6.21%</td>
<td>5</td>
</tr>
<tr>
<td>Rocky Mts.</td>
<td>0.84%</td>
<td>8</td>
<td>1.42%</td>
<td>8</td>
</tr>
</tbody>
</table>
Five Survey Locations in downtown Charleston
DIFFERENT DEFINITIONS OF TOURISTS

The cost is based on the size of the data; we can only afford four month’s of data. If a device appeared more than half of time at nights outside of the Charleston area within a month, we consider it as a visitor device.

The official WTO definition is overnight stay or 50 miles or more one way within a year (World Tourism Organization).
<table>
<thead>
<tr>
<th></th>
<th>AirSage Report</th>
<th>Intercept Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated Total Cost</td>
<td>$15,000</td>
<td>$14,000</td>
</tr>
<tr>
<td>Total Visitors</td>
<td>474,500</td>
<td>723</td>
</tr>
<tr>
<td>Per Respondent Cost</td>
<td>$0.03</td>
<td>$19.36</td>
</tr>
<tr>
<td>Data Points per Response</td>
<td>9</td>
<td>124</td>
</tr>
<tr>
<td>Cost per Data Point</td>
<td>$0.004</td>
<td>$0.16</td>
</tr>
<tr>
<td>Number of Questions</td>
<td>9</td>
<td>34</td>
</tr>
<tr>
<td>Cost per Question</td>
<td>$1,667</td>
<td>$412</td>
</tr>
<tr>
<td>Margin of Error (for ratios)</td>
<td>±1%</td>
<td>±4%</td>
</tr>
</tbody>
</table>
What is the stage and future of tourism big data?
Gartner Hyper Cycle, 2012

The Stage of Big Data
Gartner Hyper Cycle, 2013

The Stage of Big Data
Gartner Hyper Cycle, 2015

Stage of Big Data
THE FUTURE OF TOURISM BIG DATA

The “bankruptcy” of AirSage *(to be confirmed)*

We need to move from correlation to precision

Need the sharing and a trading platform of big data sources

Need transparency in data generation and calculation method and replicability in studies
THE FUTURE OF TOURISM BIG DATA

Need to correct big data biases and connect with official tourism statistics;
The merging of several big data sources to produce a more accurate picture – ongoing studies
Tourism big data needs to land in tourism bureaus’ and hotel managers’ desktops – accurate, graphic, easy-to-understand, and real-time.
Tourism Big Data
Questions and Discussions?
Email: bingpan@psu.edu