# Utilizing Big Data from Online Reviews to Understand Local Tourist Travel

# Utilizing Big Data from Online Reviews to Understand Local Tourist Travel

## Introduction

Importance of leisure time for human wellbeing, and especially usage of recreational resources during one's free time, has been established in several fields of study. Emotional stress is a major contributing factor to the six leading causes of death in the United States (e.g., cancer, coronary heart disease, accidental injuries) (www.urmc.rochester.edu/encyclopedia), and recreational activities help manage stress. However, utilization of recreational resources by various population groups is often unequal, as it is affected by a number of geographic, demographic, and socio-economic factors (e.g., Lindsey, Marai, & Kuan, 2001; Stodolska, 1998). For example, different availability of recreational resources in minority and low income areas has been registered (Moore et al., 2008), and differences in income, age, and race between users and non-user of public recreation services have been recorded as well (Howard & Crompton, 1984). Policymaking in the sphere of recreational resource management, resource accessibility and distribution requires regular monitoring of the state of affairs in this area; the data, however, is limited. This study, therefore, is making a case for employing online content as a source of data to investigate how residents of the state of Florida are using the state attractions and other recreational resources.

Florida has abundant and diverse recreation and tourism resources, including beaches, natural preserves, theme parks, and historic towns and settlements, and we are interested in how Floridians utilize recreation resources, specifically how they visit attractions within the state. Travel distance is an important measure of resource utilization, but not many studies are focused on the factors affect travel distance. This study investigates how the social demographic and economic factors affect Floridians' travel distance and the spatial variations of travel distance at the county level. We (Kirilenko et al., 2019) used network analysis of the TripAdvisor data to investigate utilization of Florida attractions by in-state, out-of-state, and international travelers and found significant differences between these segments, especially with regard to the natural resources in Florida. This paper continues this line of research, focusing on the local tourists and concentrating on the distanced traveled as opposed to attraction visitation. The overall research inspiration is to aid the state in incorporation of tourist travel demand into the transportation modeling.

## Data

We created a database of the online reviews of Florida (further FL) hotels, attractions, and restaurants posted by Florida residents on the TripAdvisor website (for methodology see Kirilenko et al., 2019). To better understand the travel pattern of the counties located in Northern Florida, especially the Westernmost Panhandle region, we supplemented the abovementioned database with reviews posted by Floridians travelling to the neighboring states (Georgia, Alabama, Louisiana, Mississippi, and South Carolina; further collectively referred as NS). Note that only the data from the frequent reviewers was included as a simple measure to reduce the number of fake reviewers in the database whose aim is to generate positive image for their properties and/or negative image for competitors (Trend, 2013) and who tend to publish very few reviews (Feng et al., 2012; Mayzlin et al., 2014). Specifically, (1) We identified all tourists staying in FL and NS hotels who posted hotel reviews in 2016 – 2018 by performing a TripAdvisor search; (2) We filtered out infrequent reviewers (those who left lesser than 10 reviews overall) and non-Florida residents; (3) For these reviewers, we collected all reviews of FL and NS+ hotels, attractions, and restaurants; (4) We identified the locations of these properties; (5) We identified the geographical

latitude and longitude the reviewers' self-identified place of living using geotagging service geonames.org (see Kirilenko and Stepchenkova, 2014 for details); (6) We obtained geographical latitude and longitude of the reviewed properties from TripAdvisor search; and (7) we cleaned the data. The final database comprised 492,123 reviews.

The vectors $V_r^p$=(reviewer *r* place of residence, reviewed property *p*) were used to estimate travel distance for the individual travels. First, the Euclidian distance $|V_r^p|$ was computed from the geographical latitude and longitude of the reviewer place of residence and reviewed property, respectively, using the haversine formula. Next, the reviews published within one trip were combined and the maximum distance was used to characterize a trip *t*: $|V_r^t|$= max($|V_{r\in t}^p|$). We define reviews as belonging to one trip ($r \in t$) as those reviews published consequently without breaks (the maximum break for any review sequence was set at 5 days, approximating a holiday break trip length). In that way we limited the effect of multiple reviews published within a single trip. To concentrate on the long-distance travel, only the trips of at least 100 km were left in the database. Thus, we created a database of Florida resident in-state (FL) and out-of-state (to NS) travels with individual travel distances. Finally, the data was combined for each of the 68 counties of Florida and complemented with socio-demographic data from the US Census (Figure 1).

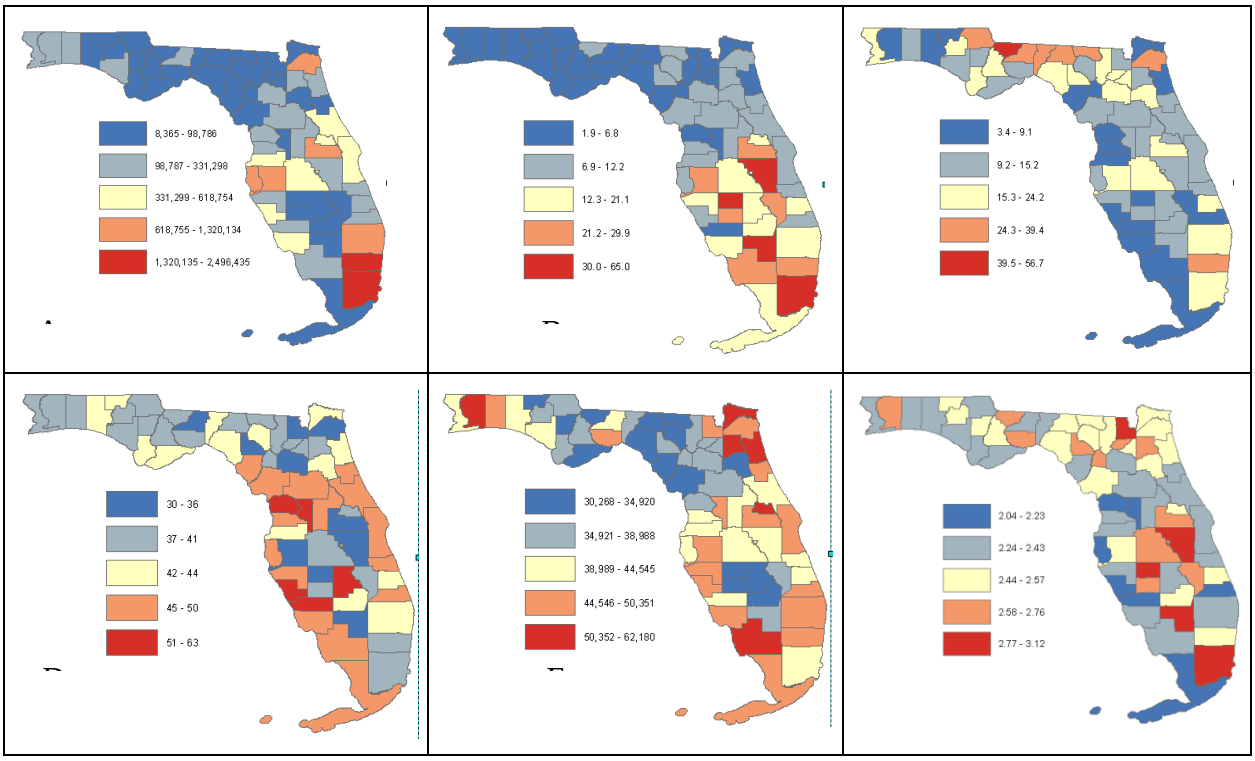


**Figure 1.** Socio-demographic information for the area of interest. A: Population; B: Percent Hispanics; C: Percent Blacks; D: Median age, E: Median household income (USD), F: Average household size.

## Results

### *Geographical Analysis*

Both the number of trips and the average trip distance show high variability throughout the state. The number of in-state trips per county decreases from east to west (Figure 2A), as the distance to the main Florida attractions increases. At the same time, the number of out-of-state-trips increases (Figure 2B). Interestingly, the number of the out-of-state trips also increases, but moderately, for the southern tip of Florida. Overall, the number of trips is significantly fewer in the remote western region of Panhandle as compared to the main Florida peninsular (Figure 2C). For the travel distance, the geographical location of the place of residence becomes important. The main travel destinations for Florida residents are Orlando (amusement parks, 14.1%), St. Augustine (the oldest city in the continental US, 6.8%), Key West (the southernmost continental US point, 5.4%), followed by Tampa (4.8%) and Miami (3.3%) – see Figure 3B. Accordingly, the mean distance for the in-state trips is larger for the westernmost part of Florida (Figure 2D; note an outlier in the center of the state – this is a sparsely populated county poorly represented in the database) due to large distances to the main state attractions. Similarly, the in-state travel distance increases for the southern tip of Florida (Figure 2D), which is also remotely located from the main attractions.

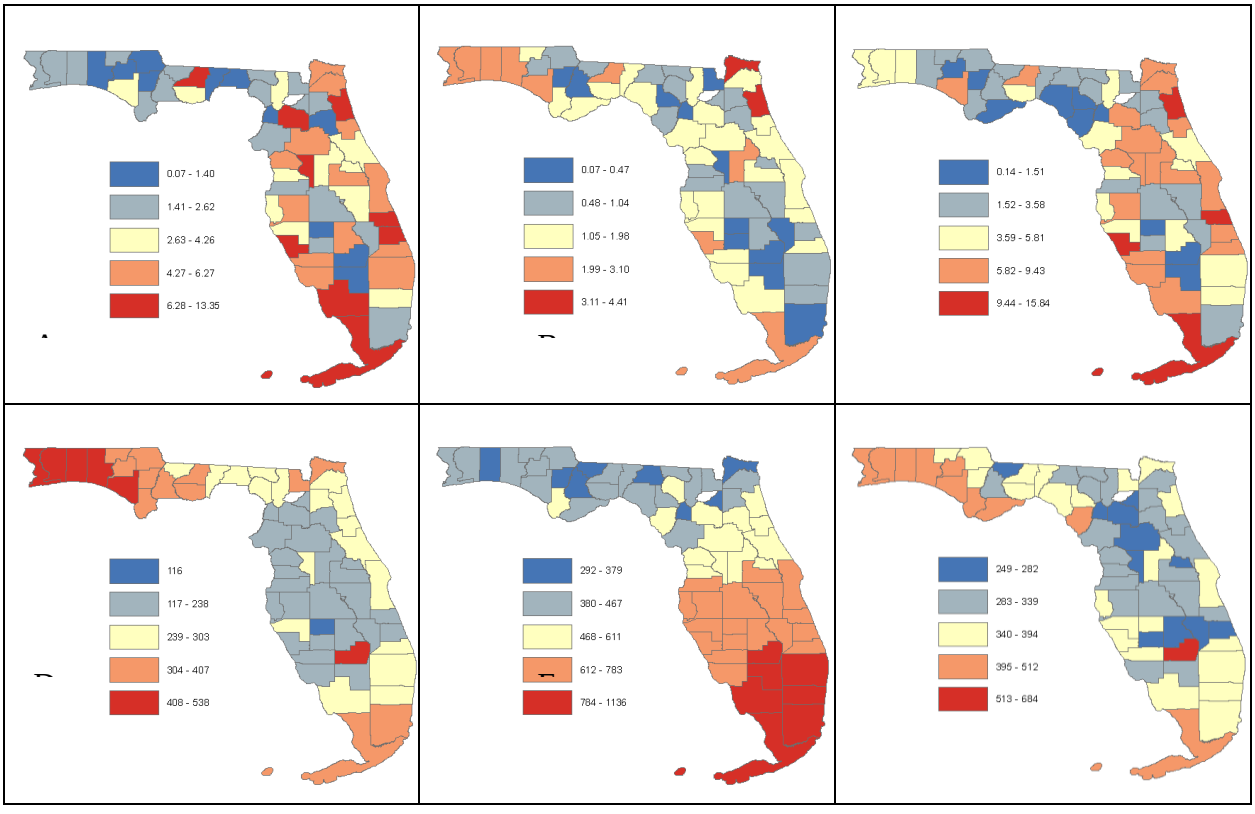


**Figure 2.** Florida travel data. A, B, C: relative number of trips for the in-state, out-of-state, and combined travel, respectively. D, E, F: mean trip distance for the in-state, out-of-state, and combined travel, respectively.

For the out-of-state travel, the latitudinal position of the county becomes the most important, as the distance to other states increases. Combined, both the number of trips and the trip distance

seem to be higher for the coastal counties of the Florida peninsular; it also decreases in the more rural counties located in the middle part of Florida Panhandle. In terms of the travel pattern, trip origins coincide with population centers (Figure 3A), while trip destinations concentrate around the main Florida attractions: amusement parks, historical sites, natural areas, and coastal zone (Figure 3B).

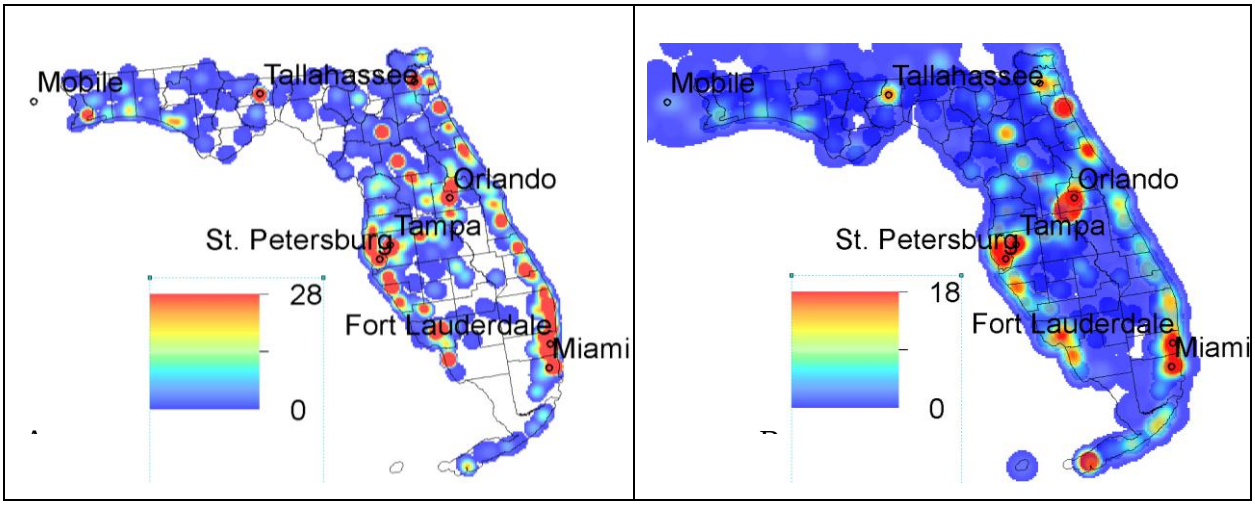


**Figure 3.** Travel pattern expressed as the number of reviews per km$^2$. A: travel origins; B: travel destinations.

### *Regression Analysis*

The study modeled travel by Floridians using Hierarchical Linear Regression (HLR) with 67 Florida counties as the units of analysis. The number of trips in each state was expressed as the number of trips per 1000 people (Trips1000) to account for differences in population sizes among the counties. Trips1000 was the dependent variable and the independent variables included in the model are listed below:

- Latitude and longitude of the county seat. This variable was included to account for the geographical position of the residents and, consequently, for their proximity to tourism resources of Florida and the NS states.
- Median household income (US 2010 Census). This is the main economic variable representing the discretionary income of residents that could be applied toward travel.
- Average household size (US 2010 Census). This variable reflects the life stage of the respondents.
- Median age of residents (US 2010 Census). This variable was included to reflect Florida's role as a retirement place; accordingly, some counties have a significant senior population with the county's medium household age above 60 years.
- Percentage of Black and Hispanic population minorities in a county (US 2010 Census). This variable reflects the differences in the ethnic composition of the counties.

The latitude and longitude were entered in the model at Step 1. At Step 2, all other variables were entered. The model with seven variables explained 63% of variance in the dependent variable Trips1000. The geographical position of a county (latitude and longitude combined) accounted for

6% of the variance in the model, while the socio-economic variables explained 57%. The model exhibited slight multicollinearity, but all VIFs were smaller than 4.6, and the distribution of model residuals was acceptable. There was one outlying observation, Monroe County (located on Florida Keys), which had a disproportionately large number of Trips1000; removal of this county, however, did not affect the results.

**Table 1.** The effect of the geographical location and socio-economic characteristics of Florida counties on residents' travel (100 km and more): Regression Analysis.

| | Dependent Variable: Trips1000 | | |
|---|---|---|---|
| Independent Variables* | Std. B | *t* | *p*-value |
| *Block 1: $R^2$ = .095; Adj.$R^2$ = .067; F(2, 63) = 3.316; p = 0.043* | | | |
| Latitude | -0.17 | -1.05 | ns |
| Longitude | 0.17 | 1.03 | ns |
| *Block 2: $R^2$ = .666; Adj.$R^2$ = .626; $\Delta R^2$ = .571; F(7, 58) = 16.551; p < 0.001. D.-W. = 1.777* | | | |
| Latitude | 0.06 | 0.37 | ns |
| Longitude | 0.21 | 1.87 | 0.067 |
| Median Income | 0.50 | 6.10 | <0.001 |
| Average household size | -0.84 | -5.43 | <0.001 |
| Median age | -0.30 | -2.28 | 0.026 |
| Black population (%) | -0.05 | -0.50 | ns |
| Latino population (%) | 0.27 | 1.69 | 0.097 |

ns - not significant at 0.10 level.

* Union county does not have reviews in the database.

## Conclusion and Discussion

In this extended abstract, we would like to focus on three takeaways from this study. First, in the absence of data about residents' movements across the state for tourism purposes, the online reviews data seem to provide a good approximation of travel patterns of residents. The results of the regression analysis are feasible from the standpoint of what we know about individual's travel behavior. The larger the median income in a county, the more discretionary income its residents can put toward travel. Both average household size and median age reflect the composition of the state population from a life stage perspective: whether younger single residents, families with kids, or empty-nesters and seniors prevail. The larger the average household size, the less trips county residents conduct. So, large families with many children travel less than singles, young couples without kids, empty-nesters, or seniors. The younger the county population, the more trips its residents make, so the younger people will travel more than seniors.

The second takeaway is that the geographical position of a county, that is, its proximity to main attractions and natural resources of the Florida state, is not overwhelmingly important and it does not preclude people from travel. With all variables in the model, the only counties that "feel" their

locational disadvantage are the westernmost counties on the Florida Panhandle (the smaller the longitude, the fewer trips). These counties, however, are not very prosperous, either, which is a stronger factor in the model affecting FL and NS travel. Finally, the racial composition of the county population has little effect on the number of trips. The counties with a large Hispanic population seem to be at a slight disadvantage; however, these counties are primarily located in the southern part of Florida, which is rich in its own attractions but the travel distance of 100 km required for a trip to be counted precluded us from counting their travel to a full extent.

Some limitations of the study need to be mentioned. The most significant limitation is the relationship between the number of TripAdvisor reviews and the actual travel intensity. It should be kept in mind that TripAdvisor data reflect the behavior of those travelers who post their reviews, which is not necessarily generalizable to all travelers. The main social networks such as Twitter and Facebook are biased towards the younger generations, leaving the older people under-represented. In our dataset, the genders are presented equally (48% male, 52% female), similar to Facebook (52% female, 48% male), but not to Twitter (34% female, 66% male). As opposed to both Facebook and Twitter, the younger generation is under-represented in our study: compare 13% of 18-34 years old domestic outside of Florida reviewers in our dataset with 23% of Florida visitors in the same age category according to Visit Florida (2018).

At the conference, we will additionally present the results (1) of the short-distance models that would reflect how residents utilize recreational resources close to their home and (2) of the models that distinguish between three types of Florida recreational resources: nature parks and beaches, entertainment attractions (e.g., Orlando parks), and historic and cultural places (e.g., St. Augustine or Cape Canaveral).

**References**


Feng, S., Xing, L., Gogar, A., Choi, Y., 2012. Distributional Footprints of Deceptive Product Reviews. ICWSM 12, 98–105.

Kirilenko, A.P., Stepchenkova, S.O., 2014. Public microblogging on climate change: One year of Twitter worldwide. Global Environmental Change 26. https://doi.org/10.1016/j.gloenvcha.2014.02.008

Kirilenko, A. P., Stepchenkova, S., & Hernández, J. M. (2019). Clustering destination attractions by visitations and tourist interests with network and spatial analyses of online reviews. Tourism Management (in print).

Howard, D. R., & Crompton, J. L. (1984). Who are the consumers of public park and recreation services? An analysis of the users and non-users of three municipal leisure service organizations. Journal of Park and Recreation Administration, 2(3), 33-48.

Lindsey, G., Maraj, M., & Kuan, S. (2001). Access, equity, and urban greenways: An exploratory investigation. The Professional Geographer, 53(3), 332-346.

Mayzlin, D., Dover, Y., Chevalier, J., 2014. Promotional reviews: An empirical investigation of online review manipulation. American Economic Review 104, 2421–55.

Moore, L. V., Roux, A. V. D., Evenson, K. R., McGinn, A. P., & Brines, S. J. (2008). Availability of recreational resources in minority and low socioeconomic status areas. American journal of preventive medicine, 34(1), 16-22.

Stodolska, M. (1998). Assimilation and leisure constraints: Dynamics of constraints on leisure in immigrant populations. Journal of leisure research, 30(4), 521-551.

Trend, N., 2013. TripAdvisor and the issue of trust. October 23, 2013. The Telegraph.

Visit Florida, 2018. Profile of domestic visitors to Florida. Research department, Visit Florida, Tallahassee, FL.