Identification of Tourist Flows in Florida to Support Development of Tourist Travel Module for FDOT Florida Transportation Model

Master agreement BDV31-977-118

Final report

Date:	August 23, 2020
To:	Thomas Hill Transportation Data and Analysis Office Email address: Thomas.Hill@dot.state.fl.us
From:	Dr. Andrei Kirilenko (Principal Investigator) University of Florida UF Department of Tourism, Hospitality, and Event Management Email address: andrei.kirilenko@ufl.edu Phone number: (352) 294-1648
Task 8:	Final Report. Upon Department approval of the draft final report, the university will submit the Final Report in PDF and Word formats electronically to the Research Center at research.center@dot.state.fl .us The Final Report is due by the end date of the task work order.
Deliverable:	Final report

Disclaimer

The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the State of Florida Department of Transportation. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability of the contents or use thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Identification of Tourist Flows in Florida to Support Development of Tourist Travel Module for FDOT Florida Transportation Model		5. Report Date June 2020
		6. Performing Organization Code
7. Author(s) Andrei Kirilenko, Jin Won Kim, Shihan Ma, Eunjung Yang		8. Performing Organization Report No.
9. Performing Organization Name and Address University of Florida Department of Tourism, Hospitality and Event Management 240B Florida Gym, P.O. Box 118208 Gainesville, FL 32611-8208		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. BDV31-977-118
12. Sponsoring Agency Name and Address Florida Department of Transportation 605 Suwannee Street, MS 30 Tallahassee, FL 32399		13. Type of Report and Period Covered Draft final, August 2019 – May 2020
		14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

Continued use of Florida Statewide Model (FLSWM) poses a number of important limitations in terms of tourist-oriented traffic flow, accidents, potential revenue and maintenance considerations. The overall goal of the project was to enable Florida Department of Transportation (FDOT) to better monitor and forecast traffic flow by separate consideration of tourist flows. This report provides (1) analytical review of the up-to-date approaches and methods of accounting for tourist flow impact on transportation system; (2) mapping of tourism regions in Florida based on spatial distribution of tourism objects; (3) identification of tourism supply components in Florida using social media data; (4) estimating tourism flows using the tourism supply components alone; (5) estimating tourism flow in Florida using social media and cell phone data and cross-validation; and (6) implications for practical modeling.

17. Key Word	18. Distribution Statement
FLSWM; tourism flow; social media; cell phone tracking; fine resolution tourism data	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages 256	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Executive summary

Florida is the largest tourism destination worldwide, receiving over 100 million visitors annually and contributing over $89 billion to the state's economy; importantly, Florida tourism occurs throughout the year. As such, tourism travel to and within Florida has substantial impact on the state's highway system. Previous iterations of the Florida Statewide Model (FLSWM) included tourism trip generation; however, the current transportation model does not. The goal of the project was to develop more advanced tools using primary and secondary data needed to accurately forecast visitor trips to regions within Florida and to estimate their impact on the Florida transportation system.

The main obstacles to including tourism in transportation models is a lack of data. While there are multiple datasets estimating the flows for a particular place or attraction, the data are patchy, and there is no dataset covering the entire state. This project developed methodology and estimated fine-granularity tourism flows in Florida. Specifically, the project demonstrated how the fusion of multiple innovative data sources, including traditional data, social media, and cell phone tracking data, help in estimation of the tourism flow with a granularity of a census tract or a county. The project pays special attention to data validation process by cross-referencing different sources of data. Finally, the project results suggest a methodology to include the new data into transportation model.

The results of this project will enable FDOT to better monitor and control traffic flow, evaluate alternative strategies for improving highway accessibility to potential and existing tourism products, evaluate potential strategies for improving visitor experiences as they travel through the state, and forecast problems or conflicts created by tourists and Florida residents. Continued use of a FLSWM which does not include tourism-related travel poses a number of important limitations in terms of traffic flow, accidents, potential revenue and maintenance, and other issues of highway use by travelers. The results of this project will enable FDOT to (1) better monitor and control traffic flow; (2) evaluate alternative strategies for improving highway accessibility to potential and existing tourism products; (3) evaluate potential strategies for improving visitor experiences as they travel through the state; and, (4) forecast problems or conflicts created by tourists and Florida residents.

Table of Contents

List of tables

List of acronyms

ADT: Average Daily Traffic

ANN: Artificial Neural Network

AON: All or Nothing

AR: Autoregressive

ARMA: Autoregressive Moving Average

ARIMA: Autoregressive Integrated Moving Average

ATS: American Travel Survey

DOT: Department of transportation

EVA: Production – Distribution – Mode Choice (from German terms Erzeugung – Verteilung - Aufteilung)

FAST: Fixing America's Surface Transportation [Act]

GA: Genetic Algorithm

GPS: Global Positioning System

GRNN: Generalized Regression Neural Network

HBO: Home-based Office

HBW: Home-based Work

HPMS: Highway Performance Monitoring System

HSR: High-speed Rail

KySTM: Kentucky Statewide Travel Model

LD: Long-distance

LDA: Latent Dirichlet Allocation.

MA: Moving Average

MDCEV: Multiple Discrete-Continuous Extreme Value [model]

MLP: Multi-layer Perceptron

MPO: Metropolitan Planning Organization

MTC: Metropolitan Transportation Commission

MTP: Metropolitan Transportation Planning

NCHRP: National Cooperative Highway Research Program

NHB: Non-home-based [work]

NHPN: National Highway Planning Network

NHTS: National Household Travel Survey

SACOG: Sacramento Area Council of Governments

SARIMA: seasonal ARIMA

SCAG: Southern California Association of Governments

SD: System Dynamics

SVM: Support Vector Machine

TAZ: Transportation Analysis Zone

TCI: Tourism Climate Index

VFR: Visiting friends and relatives

VMT: Vehicle Miles Traveled

List of figures

Chapter 1. Literature review on tourism flow and analytical methodology in transportation planning

Task description

Provide a literature review on research analyzing tourist flow impact on transportation system. The intent of the literature review is to identify current tourism modeling efforts in published scientific papers and reports to clearly define an implementable methodology for the development of the tourism flow model. This effort will also develop an analytical review of the up-to-date approaches and methods with links to published reports as well as define issues and concepts not clearly reported in the review.

Deliverable 1: Upon completion of Task 1, the University shall submit to the Research Center at research.center@dot.state.fl.us a case study of the impact of tourists on transportation systems. An online presentation will be provided to FDOT staff that will summarize the findings of the case study and the recommended methodology.

1.1 Introduction

Tourism and transportation are intrinsically related as the concept of tourism is the geographical phenomenon, involving the movement of tourists from one place to one or more destinations via a complex multimodal transportation network (Lundgren, 1984). Tourism generates travel demand to and between major tourist attractions, which not only increases transportation needs but is also distinctly different from commuter travel and commercial transportation. Meanwhile, the goal of effective transportation planning is balancing transportation needs of different traveler groups, which requires coordination between transportation and tourism agencies. This literature review focuses on academic research and current practices related to estimation of tourist flow in transportation systems. Chapter 1.2 of this review identifies transportation models and practices that include tourism-related travel data that are primarily in use in different states. Chapter 1.3 outlines classic theories of tourist movement and mobility found in academic tourism research. Chapter 1.4 and Chapter 1.5 review the methodology and data used in current tourism travel demand model researches. Finally, Chapters 1.6 and 1.7 summarize and envision the potential usage of new data sources in developing tourism module for Florida transportation model.

1.2 Transportation models and practices

1.2.1 Travel demand models in transportation planning

A key process in transportation planning is estimation of current and forecast future travel in the transportation system, including highway, transit, non-motorized, and freight modes. These travel forecasts are generally accomplished through a computer simulation modeling of traffic network, known as travel demand models (Transportation Research Board, 2005).The basic modeling approach is a sequential four-step process including (1) Trip generation: estimate total travel demand, (2) Trip distribution: distribute traffic volumes among the origin and destination zones, (3) Modal split: divide traffic according to the mode of travel, and (4) Traffic assignment: assign travel flows to appropriate highway and transit networks. Note that in areas where the travel mode is homogeneous, the mode of travel step (3) is omitted, resulting in a three-step process (Transportation Research Board, 2007).

Hence, travel demand models require data collection regarding three aspects: (1) roadway and transit system, (2) the sociodemographic and economic attributes of travelers such as income, education, residential location, job location and vehicle ownership, and (3) day-to-day travel behavior patterns such as trip purpose, departure time, mode of transport, activity duration, activity location, travel route, party composition, and traffic condition. The roadway and transit system data have been traditionally collected using roadside, global positioning system (GPS), on-board, and smart card techniques. The sociodemographic and economic attributes of travelers can be extracted from numerous survey datasets including the U.S. census data. Meanwhile, the data on day-to-day travel behavior are more complex and include face-to-face, telephone, mail-out-mail-back, Web-based, and on-board (on transit for example) surveying methods (Rashidi et al., 2017).

Travel demand models are used in both metropolitan planning (urban and regional models) and statewide planning (statewide model). The statewide models are frequently built upon practices originating in urban modelling; however, they differ from urban models with special consideration of long-distance intercity

and interstate trips. Long-distance trips include multiple trip purposes, including not only business related trips (Transportation Research Board, 2012); leisure travel such as visiting friends and family, shopping, relaxation, sight-seeing, outdoor recreation, and entertainment is a significant segment of long-distance travels (Outwater et al., 2015).

1.2.2 Long-distance models with tourism/leisure purposes

In the past decade, multiple federal and state agencies have developed a stronger interest in modeling long-distance passenger movements as part of their highway infrastructure planning (Outwater et al., 2015). Meanwhile, several studies from the National Cooperative Highway Research Program (NCHRP 329, 419, 735) has urged inclusion of tourism data into current transportation models (Transportation Research Board, 1998, 2004, 2012). In particular, the NCHRP Synthesis 329 reports on 11 state DOTs that have already account for tourism data in their transportation planning (Transportation Research Board, 2004) and a number of states adopted the long-distance/tourism/leisure travel components into their statewide models (Table 1.1), estimating transportation-related features of long-distance intercity and interstate tourists/visitors such traffic volume, speed, destination choice, and others. Case 1 provides details on one of such models. Mathematically, these models can be based on: a negative binomial regression where the dependent variable is the number of leisure travels for each purpose, season, and accompaniment combination; a linear regression where the dependent variable is the time budget of the household decides to participate in leisure travel; a Multiple Discrete-Continuous Extreme Value (MDCEV) (probabilistic travel selection model that allocates the annual time budget to combinations of trips of various purposes in different seasons; Poisson regression, which normally is used for road crush modeling, but also can be used to estimate the number of tours for each purpose-season-accompaniment type combination, end others.

Table 1.1 Selective long-distance models including tourism/leisure purposes

State	Context	Distance	Modelling Structure	Travel Purpose	Data sources
California Statewide High-Speed Rail Ridership(California High-Speed Rail Authority, 2016)	HSR ridership and revenue for proposed line	Over 100 miles	Interregional model: frequency, destination choice and mode choice; assignment of all trips including urban and external	Business, Commute, ***Recreation and Other***;	stated preference surveys, Household Travel Surveys, SCAG, MTC, SACOG, network data and Census data
Integrated Florida Statewide Model (McCullough, 1998)	Model system including passenger and freight components	Over 40 miles	4-step model ("mode" is just auto occupancy; assignment is joint with freight)	business and four types of visitor	NHTS, ATS, and Florida Visitors Survey
Kentucky Statewide Travel Model (Bostrom, 2006)	Model of long- distance trips within KY and roughly halfway into neighboring states	Over 100 miles	US Macro model (3-step, no mode choice), combined with micro model of Kentucky (denser network)	Business, ***tourism***, or other	ATS, National Highway Planning Network, HPMS
Michigan Statewide Travel Demand Model (Nellett et al., 1999)	4-step person-trip model for all motorized ground transportation	trip between urban areas	4-step model	Home-based trips (work, ***vacation*** and other), and non- home-based trips	Census, state employment agencies, roadway inventory, traffic counts
Wisconsin Multimodal Intercity Passenger Demand Model (Proussaloglou & Popuri, 2004)	Interurban model of all roads, including HSR	Over 50-miles between states, counties, and urban areas	Cross classification trip generation mode, destination and mode-choice models that are run simultaneously	Business, personal business, and ***pleasure-related travel***	2001 NHTS add-on

Note. SACOG: Sacramento Area Council of Governments; MTC: Metropolitan Transportation Commission; SACOG: Sacramento Area Council of Governments; NHTS: National Household Travel Survey; ATS: American Travel Survey; HPMS: Highway Performance Monitoring System; NHTS: National Household Travel Survey

Case 1. Kentucky statewide model

The Kentucky statewide model (KySTM), in addition to the traditional home-based work (HBW), home-based office (HBO), and non-home-based (NHB) trips, also forecasts long distance interstate and intrastate trips including tourist travel. Although the number of these long-distance (LD) trips is relatively small, the vehicle miles traveled (VMT) impact on the statewide arterial system is significant. State to state data from the 1995 American travel survey (ATS) was used to develop traffic analysis zone (TAZ)-level long-distance trip tables. Some of the details of the statewide model is as follow:

- Number of zones = 1,530 (includes 823 Kentucky zones)
- Number of links = 28,282
- Trip purposes: home-based work (HBW), truck, tourist, external, other person
- Software: 4-stage transportation planning software MinUTP
- Current year: 1999; future year: 2030
- Assignment methodology: All or Nothing (AON)
- Network development: existing 1991 Kentucky network plus National Highway Planning Network (NHPN) outside Kentucky

Long distance travel models use multiple sources of data, including travel surveys, householder characteristics (i.e., age, race, employment status and ethnicity), economic characteristics (i.e., household size, income, and vehicles owned), and residence characteristics (i.e., tenure, housing type, location and family structure). In addition, even the definition of a long-distance travel varies: some models use a specific distance such as 100 miles or more while others define long distance travel as a trip between urban centers. The influence of some variables on the number of leisure travels is shown in Table 1.2.

Table 1.2 Variable affecting selected leisure travel types

State	Visit	Relaxation	Sightsee	Outdoor Recreation	Entertainment
Personal Characteristics					
Age	+	+/-	+	+/-	-
Ethnicity other than Caucasian		+/-	-	+/-	
Working Position (full and part-time)	+	+	+/-	+/-	+/-
Household Characteristics					
Household Size	+	+			+
Household Income	-	+	+	+	
Number of Personal Vehicles	-		-	+	+
Household Structure	-	+/-	+/-	+/-	+
Residence Characteristics					
Tenure (own, rent)		+	+	+	+
Housing Type (house, apartment)		+			+
Household Location (9 Regions)	-	+/-	+/-	-	-

1.2.3 Trends in transportation modelling

The first trend in transportation modelling is that advanced four-step and tour-/activity-based microsimulation models have been developed in a few states to address several known limitations of the traditional four-step modeling approach.

The advanced four-step model incorporated advanced features that significantly improve the model capability and reliability of basic four-step model, yet still loosely follow the four-step modeling procedure. In comparison, the trip generation and distribution steps in traditional models are considered at the aggregate TAZ level, which could be improved with disaggregate individual choice models in the advanced four-step models; Moreover, the advanced four-step models incorporate tour-based methods, recognizing a complete tour is formed by multiple trips (e.g. home to work to shopping and back to home) and individual trips are interdependent due to scheduling, mode choice, travel companion and other constraints, significantly different from the trip-based assumption in traditional four-step models.

The most advanced models are those depart from four-step models with integrated land use-economic-transport analysis and tour-based/activity-based microsimulation. These models are referred as "OModels" since Oregon and Ohio were the first to adopt such models in practice (Xiong & Zhang, 2013). In these models, travel is regarded as a consequence of diverse human and economic activities and

therefore travel demand should be modeled at a behavioral level using tour-based or activity-based microsimulation. They hence require for large amounts of purposely collected data, long development time and high cost (Zhang et al., 2011). For instance, the updated 2007 Ohio statewide model has been replaced the four-step approach by a microsimulation of household activity-travel decisions in two steps: household synthesis and personal travel tours. The household synthesis is created by using a TAZ-level Monte Carlo simulation to demonstrate household characteristics; and the personal travel tours model consists of components dealing with short-distance home-based, long-distance home-based, commercial work-based, and visitor tours separately (Costinett & Stryker, 2007). Apart from the traditional socioeconomic data, traffic counts and travel survey data, Ohio model needs an extensive list of demand-side data as follows:

- Socioeconomic data (U.S. Census, County Business Patterns, ES-202, and BEA Regional Economic Information System Program);
- Land use data from Department of Natural Resources and County Auditors;
- Land value data from county assessor;
- IMPLAN I/O data used for the aggregate demographic modeling;
- A traditional one-day household survey, a small subset of which were GPS-based, covering a total of 25,000 households;
- A two-week long distance (over 50 miles) travel survey, covering 2,000 households;
- TRANSEARCH data;
- A business establishment survey of about 800 establishments, supplementing the
- TRANSEARCH data;
- CTPP outside Ohio;
- Roadside surveys taken at approximately 700 locations; and
- Other data sources, including traffic counts, travel time studies, etc.

Another trend is led by the Fixing America's Surface Transportation (FAST) Act passed in 2015, which added new Metropolitan Transportation Planning (MTP) requirements for MPOs. Under the FAST Act, the MPOs planning should take into consideration activities such as "tourism" that are affected by transportation. In addition, MPOs should consider enhancing travel and tourism as a new planning factor when developing MTPs, which requires mode detailed data as compared to long-distance travel. In such context, FDOT has made their initiative to develop a Statewide Tourism Travel Demand Model in 2016 (Pourabdollahi et al., 2017), an advanced behavior-based tourism modeling which incorporates important behavioral elements of tourism travels and captures the specific characteristics of tourism trips such as seasonality in Florida.

In summary, while initially tourism was primarily considered in long distance transportation planning and only in a few models, the current trend is towards adoption of tourism related traffic in a wider spectrum of models and at finer scales.

1.3 Tourist movement theories

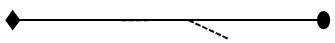
Tourist movement is a significant element in academic tourism research, representing how tourists travel from a place of origin to a chosen destination (Interdestination), and also move within a destination (Intradestination). The following movement patterns are conceptual models with idealized assumptions. Although they have not been fully integrated in current travel demand models, these theoretical tourist mobility patterns are valuable references for calibration and validation of travel pattern in building a tourism-related travel demand models, where the interdestination movement patterns are suitable for nationwide and statewide modeling and intradestination movement patterns for urban and local modelling. In addition, the tourist mobility patterns are also likely to be feasible for clustering different tourist segments based on their spatial movement.

1.3.1 Interdestination and multi-destination movements

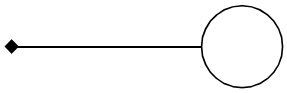
Travel itinerary or travel route is a tourist travel path from an origin to a destination, including the routes they travel along, the stopping points, and activity patterns (Lew & McKercher, 2006; McKercher & Lau, 2008). The itinerary provides a way of abstracting the movement of tourists from origin to destination, as well as across destinations.

There are four basic types of interdestination movements (Lue et al., 1993; McKercher & Lew, 2004):

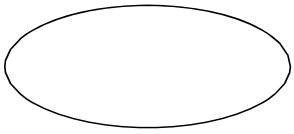
(1) Single destination, with or without side trips (a diamond indicates home location and a circle indicates a destination)



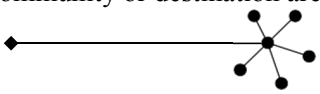
(2) Transit leg and circle tour at a destination (transport modes may vary)



(3) Circle tour with or without multiple access, egress points; different itinerary styles possible at different destination areas (transport mode may vary)



(4) Hub-and-spoke style (from home community or destination area)



There are various factors influencing tourist's decision of taking a trip and of the multi-destination trip type. These intervening factors include but not are limited to sociodemographic, psychological, physical and economic considerations (Table 1.3).

Table 1.3 Factors influencing interdestination movement

Theme	Factor	Detail and remarks
physical factors	1. attractions set	+ cumulative destinations attractions (Lue et al., 1996)
		+ distance proximity (Mckercher, 1998)
	2. travel distance	
human factors	1. tourist motivation	+ leisure – VFR **, business (McKercher & Wong, 2004)
	2. travel party	+ friend family - single (Fesenmaier & Lieber, 1988)
	3. modes of transport	+ car (Tideswell & Faulkner, 1999)
	4. previous visits experience *	+ first time visit: Hwang et al. (2006); Tideswell & Faulkner, (2009), + repeat visit (Lau & McKercher, 2006)
		+ cultural distance: McKercher & Chow So-Ming (2001); Pizam & Sussmann (1995)
	5. sociocultural differences	
time and budget	1. length of stay	Chavas et al. (1989)
	2. budget	McKean et al. (1995)
other	1. seasonality	De Oliveira Santos et al. (2012)

Notes: + indicates tourists more likely to travel in multiple destination, - indicates the opposite.

* There are contradict findings regards previous visit experience to influence interdestination movement

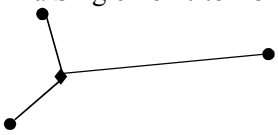
**VFR: Visiting friends and relatives;

1.3.2 Intradestination movements

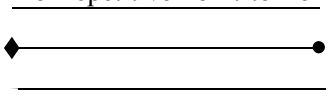
Tourist movement within a destination is the trajectory from accommodation location (hotel or home) to various attraction site or stops. Researches have classified three types of patterns based on movement linear paths as shown below. In the illustration, a diamond the origin point (location of accommodation) and a circle indicates attraction sites or stops.

- The first type is ***Point-to-Point Patterns***, with three sub-categories as:

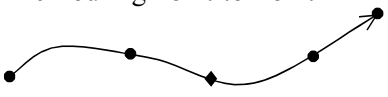
P1a Single Point-to-Point



P1b Repetitive Point-to-Point

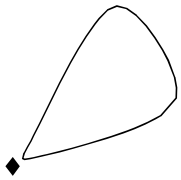


P1c Touring Point-to-Point

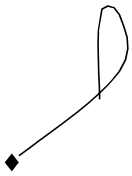


- The second type is ***Circular Patterns***, with two sub-categories as:

P2a Circular Loop

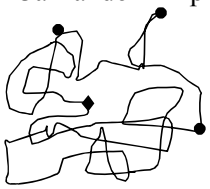


P2b Stem and Petal

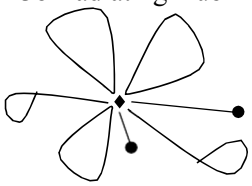


- The third type is ***Complex Patterns***, with two sub-categories as:

P3a Random Exploratory



P3b Radiating Hub



There are various destination and tourist features that might impact the intradestination movement. Table 1.4 contains a summary of these features.

Table 1.4 Factors influencing intradestination movement (Lew & McKercher, 2006)

	Themes	Impacts
Destination Characteristics	1. Trip Origins/Accommodation Locations A. Clustered or Dispersed B. Type: Hotel, Resort, Home, Other C. Clientele / Market Segments	Diversity and complexity of itineraries Identification of geographic market segments Customization of services and products
	2. Trip Destinations/Attraction Locations A. Number, Diversity/ Types, Hierarchy B. Clustering or Isolated C. Intervening or Substitutable Attractions	Diversity and complexity of itineraries (including organized vs independent travel) Identification of thematic districts Customization of services and products Importance of relative location to Accommodations
	3. Transportation Accessibility A. Traffic Network: Dense / Concentrated or Linear Topography / Site Characteristics B. Transportation Modes: Public, Tour Company, Self-drive vehicle, Walking C. Quality, Ease, Congestion, Cost and Affordability, Information, Signage D. Limitations/Barriers, Distance Decay	Degree of freedom or restriction of movement Number of preferred and alternative linear paths Perceived ease of travel and willingness to wander or explore Value of locations for development Variable access to attractions Transport mode options or restrictions
Tourist Characteristics	1. Time Budgets A. Trip Length, Visit Length B. Time Value C. C. Outcome or Process Oriented	Number of activities or attractions that can be visited Depth of participation in an activity Perception of acceptable itinerary distances Tolerance of transportation experience
	2. Motivations, Interests and Composition A. Allocentric or Psychocentric B. Special Interests or Generalists C. Recreation or Education Oriented D. Age and Physical Disabilities E. Travel Group Dynamics	Selection set of acceptable attractions, including substitutable attractions Perception of acceptable linear paths, distances and content Freedom or restriction of movement Decision making process
	3. Destination Knowledge and Emotional Value A. Information Sources, Gatekeepers B. First Time or Repeat Visit C. Primary or Secondary Destination	Emotional attachment to destination or attraction Relative appeal of attractions Perception of acceptable itinerary distances Selection set of acceptable attractions, including substitutable attractions

1.4 Methodology in tourism travel demand models

This section presents the review of the state of the art in tourism travel demand models, especially the methodology that was applied in each model.

The methodologies and approaches used in current models for tourism estimation and forecasting have been much complicate and frequently with mixture of multiple techniques. Here we briefly classify them into three categories based on their origins and evolvement overtime (Ghalehkhondabi et al., 2019; Goh & Law, 2011; Song & Li, 2008).

1.4.1 Time series methods

A time series is an ordered sequence of values of a variable on constant time intervals. Time series models are applied on time series data to estimate and predict future trends based on previously observed values. Time series models have been the most common methods in forecasting tourist arrivals when historical data on visitor arrival patterns are available.

Some traditional, commonly used univariate time series models continue to appear in studies, including the no change (Naïve I) and constant growth rate (Naïve II) models, while different exponential smoothing (ES) models and more sophisticated time series methods have emerged in recent studies. Following are several terminologies and brief introduction of these methods (Table 1.5):

- Autoregressive (AR): an AR process or AR model assume the current value of a dependent variable depends only on its values in previous periods, plus a random error.
- Moving average (MA): The dependence of a variable could also be specified on its own past using a MA process. It is one of the most common smoothing methods.
- Autoregressive moving average (ARMA): The ARMA models are a type of stationary stochastic models that consist of two models of autoregressive and moving average models.
- Autoregressive integrated moving average (ARIMA): in ARIMA, a series is transformed into a stationary covariance condition, and is then identified, estimated, diagnosed, and forecasted.
- SARIMA**:** ARIMA is extended to a seasonal ARIMA (SARIMA) when the time series is seasonally non-stationary.

* ARMA, ARIMA and SARIMA were first introduced and applied by Box and Jenkins (1976), and hence these models are commonly referred to as Box-Jenkins models.

Table 1.5 Selected research highlights with time series models

Authors	Purpose	Demand type and period	Determinants	Modeling
Oh and Morzuch (2005)	Comparing some time series forecasting models	Monthly travelers' arrivals	Historic data of arrivals	Naïve I/Naïve II/Linear regression/Winters's model/ Autoregressive Integrated Moving Average (ARIMA)/ Sine-wave regression
Andreoni and Postorino (2006)	Predicting the air transport demand	Annual airport passengers	Historic annual airport; passenger demand	ARIMA/ARIMAX
Chu (2008)	Comparing ARMA-based methods to forecast the tourism demand	Monthly tourist arrivals	Historic data of international tourist arrivals	Naïve/Linear regression/Cubic polynomial regression/Sine wave nonlinear; regression/ARIMA/ARFIMA/SARIMA
Ibrahim et al. (2010)	Predicting the tourism demand	Seasonal international tourist arrivals	Historic seasonal data of international tourist arrivals	Box-Jenkins SARIMA
Lin et al. (2011)	Comparing different methods to forecast the tourism demand in Taiwan	Monthly tourism demand	Historic data of monthly visitors to Taiwan	ARIMA/ANNs/Multivariate adaptive regression splines
Nguyen et al. (2013)	Predicting the inbound tourism demand	Inbound arrivals per month	Historic monthly arrivals of International tourists	ARIMA/Grey forecasting/Fourier series
Claveria and Torra (2014)	Comparing the tourism demand forecast models of time series and ANNs Predicting the inbound tourism demand	Monthly tourism demand	Monthly data of tourist arrivals and overnight stays	ARIMA/Self-exciting threshold autoregressions/ANNs
Tang et al. (2015)	Predicting the inbound tourism demand	Monthly tourist arrivals	Monthly data of international tourist arrivals	Time series/Likelihood-based belief function Multiple
Cankurt and Subaşi (2016)	Forecast the multivariate tourism demand in Turkey	Monthly tourist arrivals	Financial and demographic information	Multiple linear regression/Artificial neural network/ Support vector regression
Anvari et al. (2016)	Predicting the urban rail passenger demand	Periodic number of passengers	Historic periodic passenger's data	Box–Jenkins ARIMA

1.4.2 Econometric methods

An econometric model determines the causal statistical relationship between a particular economic phenomenon called dependent variable and various economic quantities called explanatory variables. The econometric methods used in tourism-related travel demand models are to explain and forecast the future tourism demand based on the quantitative relationship between quantity demanded and its determinants, using a multivariate mathematical function.

The early adoption of econometric methods mainly utilized multiple regression models (ordinary least squares (OLS) estimation method, SEM, etc.), assuming the forecasting demand to be a static value. Later, dynamic econometric models have been developed to forecast time-dependent variables, incorporating basic and advanced time series models into econometric methods, such as:

- Structural time series models (STSM): STSM is mainly to deal with the dynamics of time-indexed variables. STSM decomposes the time series into four components, which can be further linked with ARIMA or regression methodologies to formulate multivariate structural time series models (M-STSM).
- Cointegration and error correction model (ECM): if Yt and Xt are cointegrated CI (1, 1), then there must exist an error correction model (ECM), and vice versa. An ECM contains both short-run and long-run information, with the deviation from long-run equilibrium corrected gradually through partial short-run adjustments (Engle & Granger, 1987).
- Vector autoregression (VAR): VAR bypasses the need for structural modeling and is advocated both for forecasting systems of interrelated time series without imposing a priori restrictions and for analyzing the dynamic impact of innovations on the system of variables(Sims, 1980). All variables are treated as endogenous in the VAR model and there is no a priori endo-exogenous division of variables as in the SEM.
- Pooled data and Panel Data Approaches: in both approaches, the information on cross-sectional units is observed over time. A pooled data approach is obtained by sampling randomly from a large population at different points in time; whereas in a panel data approach the same samples are followed across time.

Gravity-based models are mathematical models based on Newton's gravitational law. These models have been used in social sciences to account for aggregate human behaviors related to spatial interaction, fitting well to explain international flows of trade, migration and foreign direct investment. Gravity model has been used widely to model tourism demand, especially when spatial and structural factors are analyzed. However, gravity model assumes independence among tourist flows once the effect of distance is controlled, and the spatial spillover effect is beyond the consideration of a gravity model. An alternative approach, spatial econometric modeling, which takes origin-destination dependence into account and is able to capture the spatial interaction in the modeling process, has received growing interests in the recent literature.

Selective econometric models are presented in Table 1.6.

Table 1.6 Selected research highlights with econometric methods

Authors	Purpose	Demand type	Determinants	Modeling
Varagouli et al. (2005)	Forecasting the travel demand	Number of passengers traveling by car	GDP of both origin and destination zones/Population of the both origin and destination zones/Number of cars per thousand inhabitants of the origin zone/Trip time by car or trip length/Trip price by car	Multiple linear regression
Grosche et al. (2007)	Estimation air passenger between city-pairs	Annually total passenger volume between cities	Population/Buying power index/Gross domestic product/Geographical distance/Average travel time/ Number of competing airports/ Average distance to competing airports	Gravity model logarithmic regression model
Wu et al. (2012)	Forecasting the tourism demand based on the external effective factors	Monthly international tourist arrivals	Travel demand by each origin country/Income of origins/ Prices in destination/Transportation costs/Foreign exchange rate/Population of the origin country	Sparse Gaussian process regression/ARIMA/v-SVM/ g-SVM
Sivrikaya & Tunç (2013)	Predicting the domestic air transport demand	Number of passengers carried	Urban population/Bedding capacity/Distance/ Transit/Price/Airline count/Travel match/Schedule consistency/Travel time	Semi-logarithmic regression model
Marrocu and Paci (2013)	Estimate domestic tourism flows among provinces	Number of tourists stay in province	Origin variables (GDP/Density) /Origin-destination variables(distance/price) / destination variables (GDP/density/Accessibility/tourism resources)	Panel data Gravity model spatial autoregressive model
Chu (2014)	Predicting the tourism demand	Monthly tourist arrivals	Historic monthly tourist arrival data	Logistic growth regression/ SARIMA/Naïve 1
Semeida (2014)	Predicting the taxi passenger demand	Number of trips per person per year	Distance/Population/Area/Income/Travel time/Travel cost/Trip frequency	Multiple linear Regression/ Generalized linear modeling
Chinnakum & Boonyasana (2017)	Modeling the tourism demand for Thailand	Annual tourist arrivals	Gross domestic product per capita/Relative price of tourism in Thailand/Exchange rate/Population	Panel data regression models

1.4.3 Artificial Intelligence (AI) Methods

The adoption of AI-based methods in tourism-related travel demand studies begun relatively recently. These methods are based on computational methods used to forecast human behaviors. We observe a significant growing of such method applications in the past decades, as well as the rapid development of hybrid methods integrating AI methods with time series methods and econometric methods. This section introduces some of the most popular AI methods used in tourism demand estimation.

1.4.3.1 Support vector machines (SVM)

Support vector machine (SVM) was developed by Boser et al. (1992) as a machine learning for classification and regression. SVM is based on the idea that non-linear trends in input space can be mapped to linear trends in a higher-dimensional feature space and recognizes the subtle patterns in complex data sets by using a learning algorithm (Vapnik, 2013). SVM contains two main categories: support vector classification and support vector regression (SVR). SVR/SVM has been applied to tourism and hotel forecasting by several studies (Chen & Wang, 2007; Hong et al., 2011; Xu et al., 2016) – see Table 1.7.

Table 1.7 Selected research highlights with SVR/SVM

Authors	Purpose	Demand type and period	Determinants	Modeling
Pai, Hong, & Lin (2005)	Predicting the tourism demand	Annual number of visitors	Service price/Foreign exchange rate/ Population/Market expenses/Gross domestic expenditure/Average hotel rate	Back-propagation neural networks/Multifactor support vector machine model
Samsudin et al. (2010)	Forecasting the tourism demand using a hybrid algorithm	Number of visitors per month	Historic monthly tourist arrivals data	Group method of data handling/Least squares support vector machine
Hong et al. (2011)	Forecasting the tourism demand using a hybrid algorithm	Annual tourist arrivals	Historic annual tourist arrivals data Capacity of international flights/GDP/CPI/ Foreign exchange rate	Support vector regression/ Chaotic genetic algorithm
Lin and Lee (2013)	Forecasting the tourism demand using a hybrid algorithm	Monthly tourist arrivals	Average hotel price/Number of hotel rooms/	Multivariate adaptive regression splines/ANN/Support vector regression
Rafidah et al. (2017)	Forecasting the tourist arrivals to Malaysia from Singapore	Monthly tourist arrivals	Historic monthly tourist arrivals data	Support vector machine model

1.4.3.2 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN), a nonparametric and data-driven technique, has its capability of mapping linear or nonlinear function without any assumption imposed by the modeling process. ANN simulates biological neural systems, especially human brains, by including input, hidden and output layers; each layer containing one or more neurons. Neurons of the input layer represent the input variables, such as economic and demographic data. Hidden layers are used for the network's internal understanding of the nonlinear data trend, and the output layer represents the solution to the problem. Different ANN models have been applied to tourism and hotel forecasting practice, including multi-layer perceptron (MLP), radial basis function (RBF), generalized regression neural network (GRNN) and Elman neural network (Elman NN) – see Table 1.8.

Table 1.8 Selected research highlights with ANN

Authors	Purpose	Demand type and period	Determinants	Modeling
Celebi et al. (2009)	Predicting light rail passenger demand Passenger	Passenger demand per 15 min	Historic daily passenger data	ANN/ARIMA
Chen et al., (2012a)	Forecasting tourism demand by decomposing data into a finite set of intrinsic mode functions	Monthly tourism demand	Historic tourist arrivals series	ARIMA/Back propagation neural network/Empirical mode decomposition
Chen et al., (2012b)	Predicting the air passenger and cargo demand	Annually air passenger and cargo demand	Population/GDP/GNP/ CPI/Economic growth rate/Hotel rate	Back-propagation neural network
Cuhadar et al. (2014)	Predicting the cruise tourism demand	Monthly passenger demand	Monthly foreign tourist arrivals by cruise	Radial basis function ANN/Multi-layer perceptron ANN/Generalized regression ANN
Claveria & Torra (2014)	Predicting the tourism demand	Monthly tourist arrivals from different countries	Monthly data of tourist arrivals	Multi-layer perceptron ANN/Radial basis function ANN/Elman recurrent neural networks
Noersasongko et al. (2016)	Forecasting tourist arrivals in Indonesia	Monthly foreign tourist arrivals	Historic tourist arrivals to three cities in central Java	Genetic algorithm based neural network

1.4.3.3 Other methods

There are some other computational advanced algorithm or methods have been applied in travel demand model in recent researches, yet not as prominent as SVM or ANN, such as:

- Grey system models: grey systems are those dealing with both known and unknown information. More precisely, as white systems have completely known information, and black systems have completely unknown information, grey systems are defined as the systems with partially known and partially unknown information.
- System dynamics (SD): SD is a computer-oriented approach that uses the inter-relation of variables in a complex setting. The main characteristics of SD include the existence of a complex system, time-to-time variations regarding the system behavior, and the existence of the feedback in a closed loop.
- Fuzzy logic models: fuzzy forecasting methods apply fuzzy numbers to consider uncertainties in the input data. A fuzzy system maps the nonlinearity of an input vector to a scalar output and is able to deal with both numerical values and linguistic variables.
- Genetic algorithm (GA): GA is a metaheuristic that mimics the natural selection process as described by Darwin.

AI-based methods are frequently combined with other AI-based techniques, and the researches are often present their methods as hybrid mothed. For example, Genetic algorithms have also been applied to an SVR model (Chen & Wang, 2007; Hong et al., 2011). Pai et al. (2014) further incorporated the fuzzy system, SVR technique and genetic algorithms into a new model which has demonstrated superior forecasting performance over a number of other models. Chen et al. (2010) applied the adaptive network-based fuzzy inference system model to forecast tourist arrivals to Taiwan and demonstrated its superior forecasting performance over the fuzzy time series model, grey system model and Markov residual modified model.

1.5 Data in tourism travel demand models

1.5.1 Current data sources in tourism mobility and demand forecast

Various data have been utilized to analyze tourist mobility, ranging from the traditional survey, interview and travel journals, to the new technological tools such as mobile devices, GPS trace and user-generated-contents published on social media and searching engines. Currently, there are several popular data collection methods and data sources for analysis and the comparison of these data sources and tools is listed in Table 1.9:

- GPS trace approach allows real-time high-precision (within a few meter resolution) locational tracing of a mobile device which has a relevant application installed.
- Somewhat similar to the GPS trace, mobile phone tracking potentially allows locational tracing of a single mobile device with moderate spatial resolution (~100 m and worse) through the process of multilateration of radio signals from cell towers. Since the mobile phone tracking is highly privacy invasive, only generalized data are available from the third parties.
- Geotagged social media utilizes publicly available digital footprint of the travelers and available either for fee or freely from numerous from various social media platforms. It is one of the most popular data in latest researches, yet there is an ongoing discussion on ethical ways of using

social media use in business and research, with higher concerns expressed towards using the photography.

- Last but not least, Web search data effectively captures public attention toward a certain tourism product or destination, thus, can be used as powerful predictors for tourism demand. However, Web search data is merely a general reflection of travel demand, and the estimation bias cannot be ignored.

Table 1.9 Comparison of data source and tools for tourist movement

data source	advantages	disadvantages	examples
Information collected before trip			
Web search	• no/low privacy concerns • low costs	• estimation biases • no access to user profile • generalized trends (not necessarily actual tourists)	(Gunter & Önder, 2016; Pan et al., 2011; Peng et al., 2017)
Information collected on site			
GPS trace	• accurate data	• recruited in advance • no access to user profile	(Ahas et al., 2007; Kidd et al., 2018; Yun et al., 2018)
mobile networks	• large volume	• Invasion of privacy • no access to user profile	(Bastianoni et al., 2008; Nishad & Abraham, 2017)
Information collected after trip			
geotagged social media content	• no/low privacy concerns • low costs • large volume	• bias in the sample • hard to create a tourist profile	(Cai et al., 2016; Chareyron et al., 2013; Chua et al., 2016)
travel journals	• less privacy issues	• non-representative sample • fluctuation in data quality	(Chung et al., 2017; Tussyadiah & Fesenmaier, 2007)
surveys	• level of detail • user profile	• time-consuming • small sample	(Luo et al., 2017; Smallwood et al., 2012)

1.5.2 New data to represent tourism components in transportation model

There are also emerging studies to integrate tourism component with transportation model or develop tourism-specific travel demand models through innovation in data collection (Table 1.10). The following are three examples have their respective novelties in employment of new tourist data or variables to transportation models:

- Hofer et al. (2016) integrated tourists in the existing macroscopic transport model of Salzburg (VerMoSa). An activity-based EVA approach was used to calculate tourist demand. In contrast to the traditional 4-step-model, EVA allows a disaggregated analysis of different trip purposes and

tries to find the equilibrium of transport supply and demand. The number of tourists at a vacation spot is used for the structural property at the origin and a newly defined tourism attraction potential is used for the structural property at the destination. A survey in the planning area was conducted to obtain mobility behavior data of tourists at the vacation spot. The data for model parametrization and calibration comes from the surveys. The advantage of this model is that is builds traffic network entirely based on tourism activities generating three types of tourist flows: Hotel – Attraction, Attraction – Attraction, and Attraction – Hotel. Note that the data collected for the model were mostly survey-based similar to the traditional approach.

- Yue and Ksaibati (2018) followed the traditional four-step travel demand model to predict traffic volumes on low-volume roads Yellowstone and Grand Teton National Parks. This study used gravity model to distribute trips among destinations. Tourism-related parameters, including traffic volumes at park entrances, park area, and number of campsites in park were collected and input into the travel demand model for estimating and mapping the average daily traffic (ADT). The advantage of this study is that the data was mostly secondary and could be obtained automatically. Note that this approach seems mostly limited to national parks and similar areas where the traffic is dominated by tourism activities and is relatively homogenous. To apply the model in an urban area with highly heterogeneous travel flows seems challenging.
- Davis et al.(2018) pioneered in social media data used in a wide-area transportation modeling. Their model was built upon the California statewide long-distance model. The model used travel log data from the 8-week California Household Travel Survey, which includes summaries of daily travel diary, household sociodemographic information, and details on the place of residence. Variables used for analysis includes three censored variables (tour miles by air, miles driving, and miles by public transportation) and two reported variables (main trip tour purpose and number of overnight stays). In addition to employing the new travel survey data, the most innovative of this study is to incorporate the social network data (Foursquare), to describe destinations and their attractiveness.

Table 1.10 Selected examples of tourism travel demand modelling research.

Place, reference	Determinant/Data input	Data output	Basic model	Highlights
Salzburg, Austria (Hofer et al., 2016)	Structural data: # of tourists, attraction potential; Survey data on tourists' behaviors: mobility rates, production rate, occupancy rate, modal split; socio-demographics	Public and private traffic hotel-attraction, attraction-attraction, attraction-hotel: travel distance, time, and speed	Activity-based transportation model	Integration of tourists in an existing macroscopic transport model; Survey data is specifically tourist-oriented
North Wyoming (Yue & Ksaibati, 2018)	Traffic volumes at park entrances; park area; number of campsites	ADT in TAZ level assignment (transportation network, road volumes, directions)	Four-step Transportation Model; gravity model	tourists flow and traffic within destinations; not applicable to Interstates/statewide models
California (Davis et al., 2018)	Household Travel Survey travel log: single-day diary household sociodemographic; place of residence characteristics; check-in social media data	Mode of transport (miles of air, bus, auto) Stay night Trip purpose	Path analysis; Activity-based transportation model; Multilevel regression	supporting the statewide model in California; used social media for cross-validation of location

1.6 Summary

1.6.1 Trends in tourism demand modeling

After reviewing the methodologies and new data regarding tourism travel demand models, we observe the following three trends in current and future studies:

Prominent AI methods with big data analytics

There is a significant increase of soft computing and artificial intelligence methods such as ANN and SVM, especially applied in the development of hybrid methods with other AI-based or traditional methods. Combination of forecasting methods can integrate the advantages of various methods and provide useful tools to deal with non-linear patterns in data or intermittent and lumpy demands.

Another reason for the growing dominance of AI methods lies in its outperformance in model accuracy and efficiency, which is largely attributable to the advance of big data analytics and applications. AI-based demand forecasting methods using big data are likely to be the most interesting topic for the future studies in this field.

Lack of high-resolution data

Although the model accuracy and performance has been improved with the development of AI modelling algorithms, the determinant variables in multiple models still largely rely on historical records with low spatial and temporal resolution. It is thus suggested to bring near-real-time data into the forecasting methods to further increase the accuracy of forecasts and search for additional data with finer scales in spatial and temporal resolution.

New explanatory variables

The selection of tourism demand's determinants is far more diverse than its measurement. Apart from the classic economic factors, transportation cost, population density, and other social, cultural, geographic and political factors, new explanatory variables have appeared in recent empirical studies, and some are particularly strong in explaining tourism and hotel demand trends and changes (Wu et al., 2017). One of most significant categories of variables is the tourist online behavior variables. As user-generated data and online consumer behavior data have become increasingly available to researchers, these data have been widely used in conjunction with traditional economic data to improve forecasting performance (Yang et al., 2014). The early researches mainly used search query data and Web traffic data for forecasting the general trend of travel demand (Pan et al., 2012; Yang et al., 2015), while the more recent studies are leveraging new data such as tourist social media to identify detailed tourist behavioral pattern for travel demand modelling.

1.6.2 Social media and location intelligence data

Advantages

New data sources such as social media and locational intelligence (for example, generalized cell phone signal tracks) data have significant potential to be accommodated in tourism travel demand models and transportation models alike, with their advantages in the following aspects:

(1) Data instantaneousness

Despite the advantage of survey data in level of details, they provide historical records. The GPS trace, mobile tracking solutions, and location-based social media allow for real-time extraction of data on tourist mobility.

(2) Data resolution

Tourism-related transportation model requires methodologies that allow to obtain and analyze data with a higher spatial and temporal resolution and applicable to smaller scale such as attraction site or within a destination. Compared to travel survey, geotagged tweets, activity logs on mobile networks and geotagged photos are more promising in provide accurate origin, destination, and trajectory data.

(3) Data quality and richness

Social media data frequently contain fine-resolution data on the temporal and geographical footprints of the tourists, which is beneficial for generating tourist-related movement. Tourist's attitudes, consumption behaviors extracted from user-generated data, as well as the additional demographic information from user profiles are all the possible variables to improve multiple methods and approaches in travel demand modelling.

Feasibility

It is also feasible to retrieve determinant variables for a transportation model such as destination choice, mode choice, time allocation and expenditure from aforementioned new data with data mining techniques (Rashidi et al., 2017).

- Trip purpose: the purpose of travel activity can be extracted from the online textual data including photography captions and hashtags with theme modeling approaches such as Latent Dirichlet Allocation (LDA).
- Departure time and activity location: the timestamp and geocode associated with the social media make it easier to detect the time and location attributes.
- Travel route, activity duration and traffic condition: with multiple social media posts, the trajectory of data is likely to provide the information about travel route, direction and duration, although there is a high requirement of data density.
- Mode of transportation and party composition: this information can potentially be estimated with machine learning using textual data.
- Socio-demographic attributes: it is possible to identify certain attributes from user profiles or through data mining techniques.
- In addition, content analysis including sentiment mining can provide auxiliary data on travel and traveller satisfaction, which is largely absent in traditional survey data.

Perspective new methodologies

The perspective to integrate the new data into travel demand models is summarized in the following table 1.11:

Table 1.11 New methodological approaches in travel demand models

Model	Method/approach	New data and tools application potential	
		Data	Analysis
Tourism Travel Demand Model			
Time series modelling	univariate time series	• alternative historic record • high-resolution temporal data (more cross-section and time intervals)	
	multivariate time series	• alternative historic record • high-resolution temporal data (more cross-section and time intervals)	
Econometric model	regression method	• additional determinants	• cross-validation
	dynamic econometric methods	• panel/ pooled data • high-resolution temporal data • additional determinants	• cross-validation
	spatial econometric methods (gravity model)	• easy availability • accuracy • high-resolution spatial data • additional determinants	• cross-validation • additional spatial analysis • network analysis
AI-based model	SVM, ANN, etc.	• large data volume • structured data • richness in meta data	• data training and test • big data analytics feasibility
Transportation Travel Demand Model			
Destination choices	regression analysis	• high-resolution spatial data	• cross-validation
	econometric approach	• additional determinants • high-resolution spatial data	• cross-validation
Mode choice	discrete choice analysis	• additional determinants	• cross-validation
	econometric approach	• additional determinants	• cross-validation
Time allocation	regression analysis	• instantaneousness • accuracy • high-resolution spatial data	• cross-validation • additional temporal analysis

1.7 Methodology based on the literature review

1.7.1 Overview

This project is suggesting the methodology as described below.

1. A spatial socio-econometric analysis based on tourism supply components developed in Task 2 & 3 can define general tourism resources and destination attractiveness in Florida on TAZ scale.

2. Fine spatial resolution social media data from TripAdvisor reviews of Florida attractions posted by tourists countrywide are likely to be used to generate, validate and spatially downscale the model to a level of individual attractions with high tourism popularity;

3. Locational intelligence data are promising to validate the model and to enhance it to differentiate between trips on different days of the week and different months.

1.7.2 Framework

The suggested framework of tourism travel demand modeling for the State of Florida is presented in three modelling layers:

1. Tourist Segmentation: macro-level decisions
The origins of tourists can be retrieved from online social media profiles, partially validated by generalized cellphone data. Together with the travel purposes and tourist group (based on self-description of tourists in the social media), tourists can be segmented into multiple featured groups (Figure 1.1).

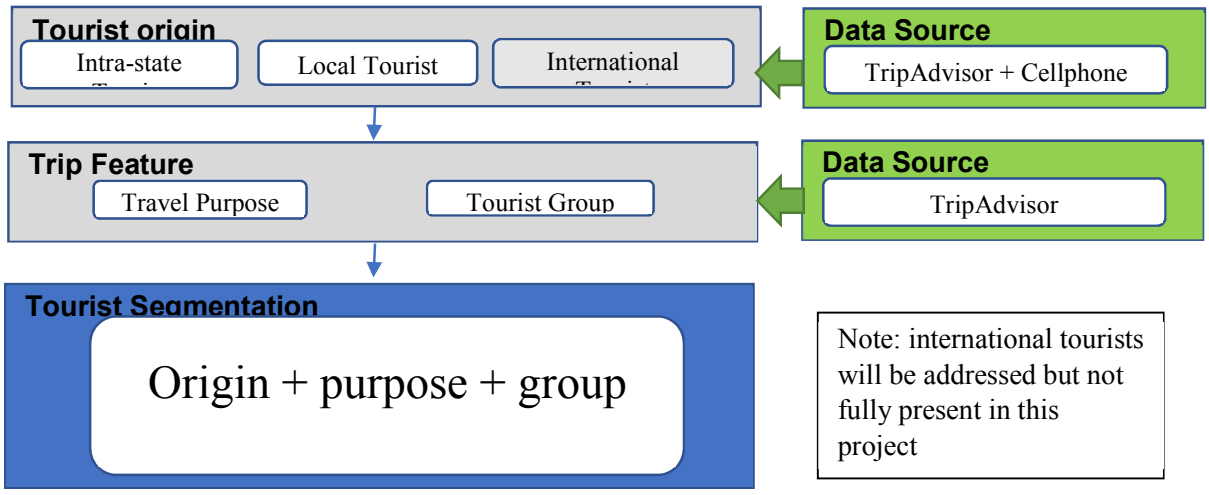


Figure 1.1 Tourist segmentation: macro-level decisions

2. Tour Generation: meso-level travel plans

Tours generated between TAZs are available based on locational intelligence (for example, generalized cell phone signal tracks) data, particularly destination choices. Likewise, individual trips are detectable from tourist travel footprints on TripAdvisor and the crowdsourced travel recorded can be aggregated to TAZ scale with comparable structured tour generation dataset (Figure 1.2).

Figure 1.2 Tour generation: meso-level travel plans

3. Within-destination Trip Simulation: Micro-level activity-travel choices

TripAdvisor provides fine spatial resolution travel records to a level of individual attractions/facilities with high tourism popularity, thus it is possible to identify relative preferences for travel within

destinations, and further distinguish different types of tourist tours, such as hotel-attraction, attraction-service, hotel-hotel, etc. (Figure 1.3).

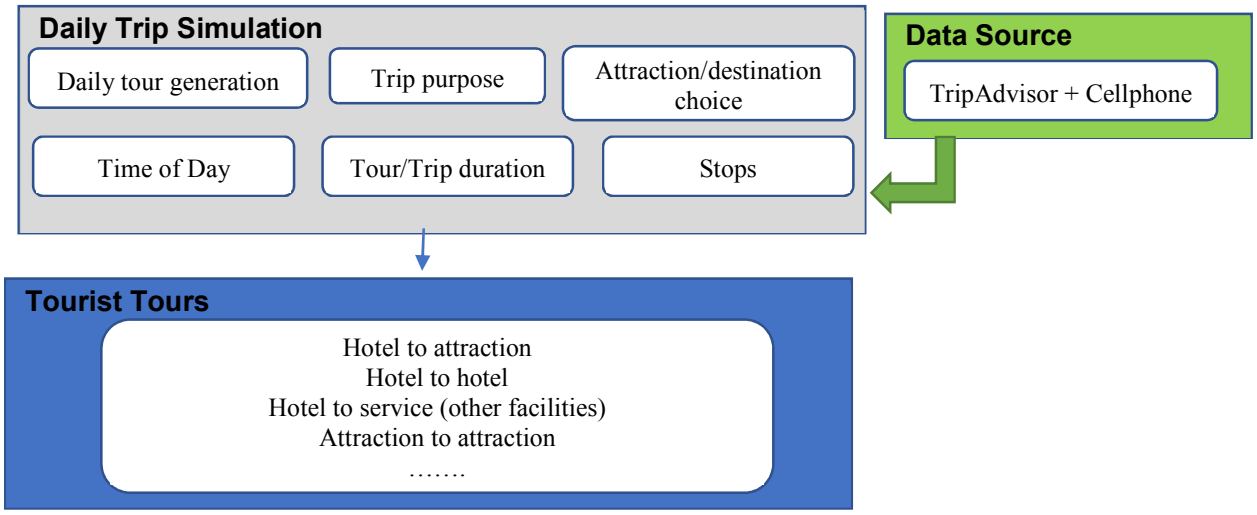


Figure 1.3 Within-destination trip simulation: micro-level activity-travel choices

1.7.3 Methodology based on current FLSWM

1.7.3.1 Trip generation

The trip generation stage within the passenger model determines the number of trips that originate or terminate within each TAZ. Current FLSWM model only considers the first notion, namely the production of trips originated from each TAZ. With the employment of social media data and locational intelligence data, we propose two options for trip generation depending on the data structure and data quality after collecting.

The traditional trip generation model simulates trip production based on assumed variables and true traffic counts, and further forecasts trip generation for each TAZ, especially for those lacking observational data. For tourist trip generation, social media provides data on tourist trip generation which incorporate practically the entire assembly of data from tourists who reported their travel experiences on TripAdvisor. All the observed data are also eligible to be segmented based on tourist groups (intra-state, domestic and international). This data however may not be representative for smaller TAZs. Hence, the use the standard trip generation module as described in the current FLSWM will be required for smaller TAZs. Similar to the current FLSWM, trip generation of each TAZ should be based on their respective socioeconomic features such as Commercial Employment, Population, Household size, auto ownership etc. Instead of using traffic counts, the observed tourist trip data could be used as the input of dependent variables for linear regression.

1.7.3.2 Trip distribution

The trip distribution stage of the passenger model determines how the trips created in the generation stage of the model get allocated. The output of trip distribution is a trip table in which the origins and destinations of individual trips are identified. The current FLSWM utilizes both gravity model and multinomial logit destination choice model for trip distribution. Specifically, gravity model is used for TT (Truck Taxi), SDEI (Short Distance External/Internal), and Long-Distance Business (LDB) trip purposes,

while destination choice model is used for internal person trip purposes (HBW, HBSH, HBO, HBSR and NHB).

The three main tourist segmentations (intra-state tourist, domestic tourist and international tourists) have shared characteristics with internal person trip purposes as well as SDEI and LDB. Because of that, is too early to decide which approach is more suitable to adopt tourist flow into FLSWM, and for the purposes of this task both approaches are explained.

Gravity model

The gravity model for trip distribution in current FLSWM is as follows:

$$V_{ij} = \frac{O_i D_j F_{ij}}{\sum_{j=1}^{n} D_j F_{ij}} \quad (1.1)$$

Here,

V_{ij} = Trips (volume) originating at analysis area *i* and destined to analysis area *j*;
O_i = Total trips originating at *i*;
D_j = Total trips destined at *j*;
F_{ij} = Friction factor for trip interchange *ij*,
i = Origin analysis area number, $i = 1, 2, 3, \ldots n$;
j = Destination analysis area number, $j = 1, 2, 3, \ldots n$; and
n = Number of analysis areas.

The friction factors (F_{ij}) can be based on travel time or distance between analysis areas, which are measures of trip impedance. The FLSWM highway network includes link distances and travel times that can be used to compute the required travel times between TAZs.

We suggest two possible adjustments to current gravity model to employ tourism flows into current FLSWM:

(1) Tourism trips, either originating or destined, are observed based on tourist behaviors, namely, data collected from tourist review records on TripAdvisor; the volume of the trips can later be cross validated by traffic counts and cellphone network data;
(2) New parameter W_j adds to the equation, adjusting the formula as:

$$V_{ij} = \frac{O_i W_j D_j F_{ij}}{\sum_{j=1}^{n} W_j D_j F_{ij}} \quad (1.2),$$

where: W_j = weighted index indicating tourism supply resources in the destination area, the evaluation of which is based on result of Task 2 and 3;

Destination choice model

In current FLSWM, the destination choice model is specified as a typical multinomial logit model. For each possible destination zone, a systematic utility function is introduced as:

$$V_i = \sum_k \beta_k X_{ik} + \log\left(\sum_j \exp(\gamma_j) Z_{ij}\right) \quad (1.3)$$

The first term represents the qualitative part of the utility function which describes the relative attractiveness of any particular destination in a zone, where β is a vector of parameters and X is a vector of qualitative. The second term represents the quantitative part of the utility function which describes the

total number of individual possible destinations (i.e., jobs, shopping or activity opportunities, etc.) in a zone, where γ is vectors of parameters and Z is a vector of quantitative variables representing the attributes of alternative i.

To complete the multinomial logit model, an error term is added to the systematic utility, with an extreme value distribution that is independent across alternatives and across observations, so that the probability of choosing any particular destination zone i is given by:

$$P(i) = \frac{\exp(V_i)}{\sum_j \exp(V_j)} \quad (1.4)$$

The data used in FLSWM destination choice model estimation includes three primary data sources: trip observations from the 2009 NHTS Florida Add-on survey data, socio-demographic data (population and employment) by travel analysis zone, and congested travel conditions that give the zone-to-zone highway travel distance for all zones in Florida.

A similar approach is suggested for tourist travel destination choice. Specifically, the utility function representing the attractiveness of the destinations can be formulated in a similar way for each of the destination zones based on the available tourism resources (produced in Tasks 2 and 3). The trip observation data can be retrieved from tourist footprints on TripAdvisor, instead of the survey data. The model can be validated using the mobile phone data. Further, for the most important tourist destinations the probability of choosing any particular destination zone i can be estimated directly from the social media data as the relative number of review for this TAZ:

$$P(i) = \frac{Obs(V_i)}{\sum_j \mathrm{Obs}(V_j)} \quad (1.5)$$

where: Obs (Vi) = observed trip to destination zone *i.* In such case, the volume of the trips to a particular destination zone is no longer based on the estimation of utility function (namely, based on destinations attributes), but replied upon the observed social media data of true tourist travel records.

Chapter 2. Analysis of regional patterns of tourism resources in Florida

Task description

Develop and apply a methodology for the analysis of regional patterns in tourism resources in Florida. Using newly developed FDOT geocoded tourism inventory data (a total of 78 types of geographic dataset: 8 tourist origins and 70 tourist destinations), this task identifies spatial patterns in Florida's tourism resource base, connects tourism inventory with highway accessibility/availability (e.g., miles of highway system for each areal unit), and then delineates the state's tourism regions. The research team will accomplish the following subtasks for Task 2:

2.1 Use the FDOT tourism inventory data to create a series of indices summarizing the spatial patterns of Florida's tourism resources and define specific types of tourism regions. Factor analysis will be used to define specific types of tourism regions, e.g., urban tourism region, beach tourism region, theme park tourism region, cultural tourism region, park and outdoor recreation region. These objectively defined tourism regions will be then mapped with a granularity of a county or census tract where possible;

2.2 Develop a tourism region map layer complemented with the measures of highway accessibility/availability (e.g., miles of highway system for areal unit) at a census tract granularity;

2.3 Employ geographically weighted regression (GWR) techniques to explore spatial variations in the relationships between the spatial patterns of Florida's tourism resources and highway accessibility/availability. The generated GWR model will be used to complement objective data in locations where the data is missing. The results of this subtask will enable FDOT to better establish strategies for improving highway accessibility to potential and existing tourism regions by identifying the spatial mismatches between tourism resources and the Florida highway system.

Deliverable 2: Upon completion of Task 2, the University shall submit to the Research Center@dot.state.fl.us the results and findings of this task. The submission will be in a form of (1) a written report and (2) GIS layers identifying: (1) tourism regions; (2) spatial patterns of tourism regions; (3) highway accessibility/availability; and (4) spatial association between tourism resource indices and highway accessibility/availability with census tract granularity. All ArcGIS layers will be provided electronically via the FDOT FTA protocol and supplemented with metadata describing each layer. An additional geodatabase copy will be provided on a flash drive to the FDOT within a 15-day time window.

2.1 Introduction

Florida is the largest tourism destination worldwide, receiving over 100 million visitors annually and contributing over $89 billion to the state's economy (Lee et al., 2019); importantly, Florida tourism occurs throughout the year. As such, tourism travel to and within Florida has substantial impact on the state's highway system. Tourism, as a socioeconomic activity, does not occur randomly. Some regions, destinations, or sites appear to be more successful than others in offering tourism activities and in attracting travelers (Gunn, 1972). The identification and analysis of existing patterns of tourism resources are critical steps in assessing the potential for attracting tourists to a given area (Formica & Uysal, 2006). The findings from analyzing regional patterns in tourism resources will enable FDOT to forecast visitor trips to regions within Florida and to estimate their impact on the Florida transportation system.

The purpose of Task 2 is to develop and apply a methodology for the analysis of regional patterns in tourism resources in Florida. Using newly developed FDOT geo-coded tourism inventory data (a total of 92 types of geographic dataset: 8 types of tourist origins and 84 types of tourist destinations), this task identifies spatial patterns in Florida's tourism resource base, connects tourism inventory with highway accessibility/availability (e.g., miles of highway system for each areal unit), and then delineates the state's tourism regions.

To achieve the task purpose, three objectives are identified.

1. Create a series of tourism resource indices and define specific categories of tourism regions;
2. Develop tourism region maps complemented with a highway accessibility/availability measurement;
3. Explore spatially varying relationships between Florida's tourism resources and highway accessibility/availability.

The findings from these objectives will enable FDOT to better establish strategies for improving highway accessibility to potential and existing tourism regions by identifying the spatial mismatches between tourism resources and highway systems in Florida.

2.2 Methodology

The overall process of analyzing regional patterns of tourism resources in Florida is a complex process that involves several steps. It begins with review resources and ends with exploring the spatial relationships between tourism resources and highway accessibility/availability. Figure 2.1 presents an overall flowchart for task 2.

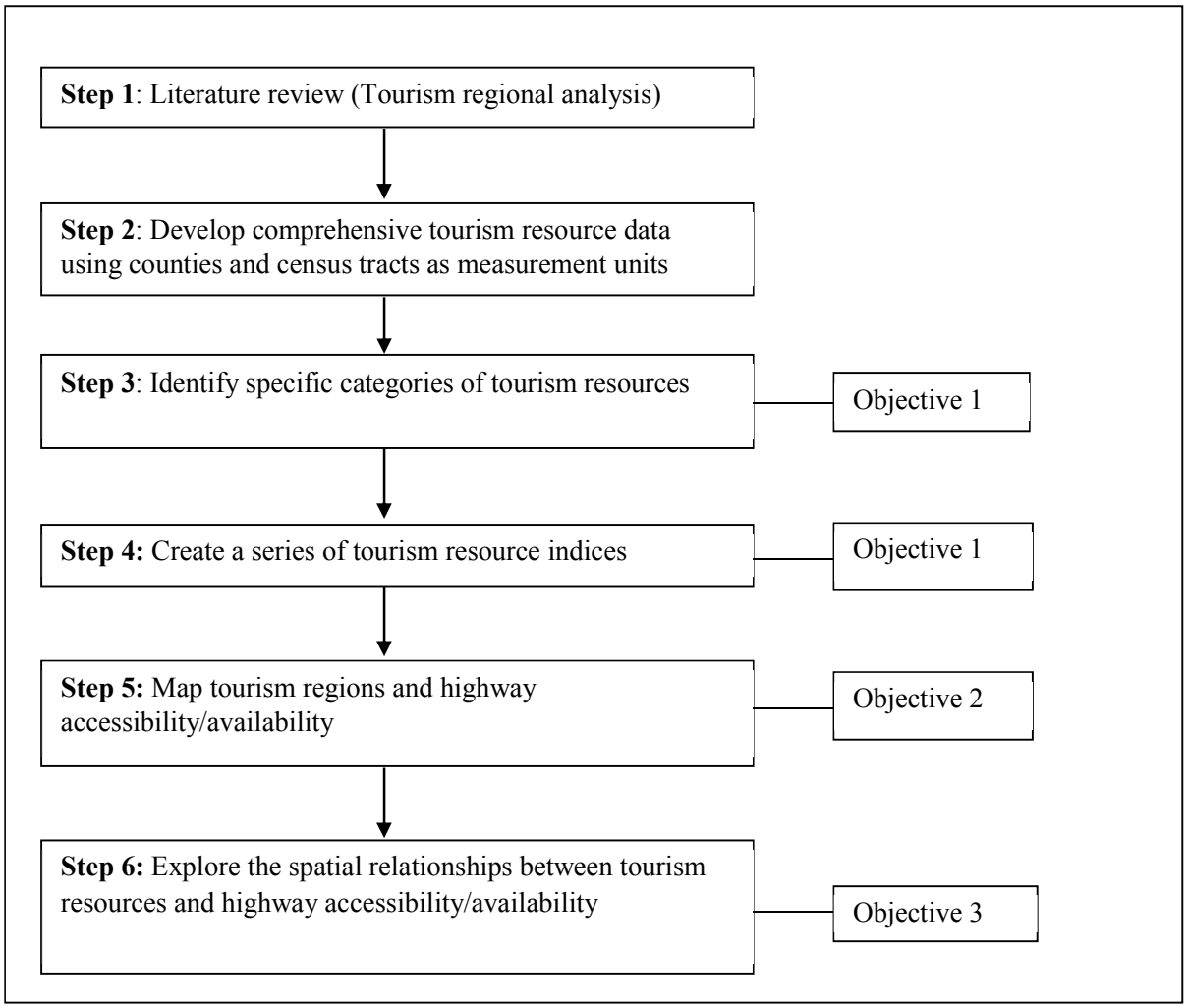


Figure 2.1 Flowchart for Task 2

2.3 Step 1: Review of literature on tourism regional analysis

Smith (1987) was the first who applied regional analysis to the tourism field in an analysis of travel and tourism resources. He used factor analysis and GIS mapping to identify the spatial patterns in Ontario's travel and tourism resources as "urban tourism," "outdoor recreation," "cottaging and boating," and "urban fringe tourism." Lovingood and Mitchell (1989) applied Smith's (1987) regional analysis approach and obtained similar results via a case study of South Carolina. The spatial patterns in South Carolina's travel and tourism resources were identified as "urban recreation – amenities rich," "urban recreation – tourism," "boating (fishing) and camping," and "outdoor recreation – nature oriented." Spotts (1997) used same methodology to explore the spatial patterns of Michigan's tourism resources and defined six tourism resource factors: urban tourism resources, general wildland tourism resources, general coastal tourism resources, parkland tourism resources, Lake Michigan coastal tourism resources, and canoeing/ORV riding tourism resources. Lastly, Formica and Uysal (2006) used similar methodological approach to explore the spatial patterns of tourism resources in Virginia and identified four key factors of tourism resources: tourism services/facilities, cultural/historical, rural lodging, and outdoor recreation. Table 2.1 summarizes the identification of tourism resources in previous tourism regional analysis studies.

Table 2.1 Summary of previous tourism regional analysis studies

Author	Study Area	Method	Unit (N)	Factors
Smith (1987)	Ontario, Canada	Factor analysis, GIS mapping	County (47)	1. Urban tourism 2. Outdoor recreation 3. Cottaging and boating 4. Urban fringe tourism
Lovingood & Mitchell (1989)	South Carolina, USA	Factor analysis, GIS mapping	County (46)	1. Urban recreation -amenity rich 2. Urban recreation - tourism 3. Boating (fishing) and camping 4. Outdoor recreation – nature oriented
Spotts (1997)	Michigan, USA	Factor analysis, GIS mapping	County (83)	1. Urban tourism 2. General wildland tourism 3. General coastal tourism 4. Lake Michigan coastal tourism 5. Canoeing/ORV riding tourism
Formica & Uysal (2006)	Virginia, USA	Factor analysis, GIS mapping	County (95)	1. Tourism services/facilities 2. Cultural/historical 3. Rural lodging 4. Outdoor recreation

2.4 Step 2: Developing comprehensive tourism resource data using counties and census tracts as units

Selection of tourism resources in Florida was based upon a review of previous journal articles, tourist websites (e.g., visitflorida.com), and the geospatial database recently collected by FDOT which describes the location and capacity of all major tourism resources (including tourism attractions, accommodations

and others). In particular, the FDOT includes 62 categories of tourism resources, representing tourist origins (8 categories) and tourist destinations (54 categories), which were from a variety of agencies, organizations, companies, or educational institutions.

The choice of areal unit is critical in any spatial analysis (Kim et al., 2018). Previous studies have typically used the county as a measurement unit (Formica & Uysal, 2006; Lovingood & Mitchell, 1989; Smith, 1987; Spotts, 1997). However, the use of a county cannot explore the local patterns of tourism resources. So, the census tract was also employed as a unit of analysis. Finally, all geographic data, describing existing Florida tourism resources (lodging, beach, golf course, shopping center, restaurant, scenic drive, cultural center, civic related, and park) were collected for each census tract and county in the state of Florida. Table 2.2 shows a comprehensive list of tourism resources, including time, type, source, and date.

Table 2.2 Comprehensive list of tourism resources in Florida.

Item	Number of Features	Type of data	Source	Date
Tourist Origin Data				
Lodging				
Hotel	7,840	Address, Point	STR, UFGC, WS	2014
Motel	2,798	Address, Point	STR, UFGC	2014
Bed and Breakfast	258	Address, Point	UFGC, SSL	2011
Airbnb	112,743	Address	AirDNA	2016
Campground	12	Latitude/Longitude	UFGC	2016
Mobile home park	2369	Point	UFGC	2009
Resort condominium	128	Address	USCB	2012
Timeshare	364	Address	AIF	2015
Tourist Destination Data				
Beach				
Beach access point	2,184	Latitude/Longitude	FDEP	TBD
Beach area	302	Polygon	FFWCC	2002
Golf Course	1,124	Point & Polygon	UFGC	2015
Historic related				
Historic structure	163,623	Point	BAR	2016
Historic bridge	1,254	Line	BAR	2016
Historic building	930	Polygon	BAR	2016
Historic church	1	Polygon	BAR	2016
Shopping Center	707	Point	ESRI	2014
Restaurant				
Restaurants	49,659	Point	GRI, FRLA	2016
Bars (Drinking Places)	4,441	Point	GRI, FRLA	2016
Cafeteria/Buffet	463	Point	GRI, FRLA	2016
Scenic Drive				
Scenic highway	18	Line	FDOT	2012
Scenic byway	23	Line	FDOT	2012
Cultural Center				
Aquarium/Zoological	47	Point & Polygon	UFGC	2015
Arboreta/Botanic garden	22	Point & Polygon	UFGC	2015
Art council	12	Point & Polygon	UFGC	2015
Motion picture theater	279	Point & Polygon	UFGC	2015
Museum/Art gallery	566	Point & Polygon	UFGC	2015
Planetarium	6	Point	UFGC	2015
Theater/Performing art center	106	Point & Polygon	UFGC	2015

Table 2.2 (continue)

Tourist Destination Data				
Civic related				
Amphitheater/Outdoor venue	17	Point	UFGC	2012
Banquet hall/Facility	828	Point	UFGC	2012
Bingo hall	122	Point	UFGC	2012
Bowling alley	301	Point	UFGC	2012
Casino	38	Point	UFGC	2012
Carnival facility	14	Point	UFGC	2012
Civic center	49	Point	UFGC	2012
Conference center	118	Point	UFGC	2012
Convention center	235	Point	UFGC	2012
Fairground	19	Point	UFGC	2012
Greyhound/Horse tract	22	Point	UFGC	2012
Hall auditorium/Ballroom	197	Point	UFGC	2012
Skating rink	99	Point	UFGC	2012
Speedway	54	Point	UFGC	2012
Sportclub organization/facility	91	Point	UFGC	2012
Stadium/Arena	105	Point	UFGC	2012
Trade show exposition	47	Point	UFGC	2012
Aquatic center	34	Point	UFGC	2016
Parks				
National park	3	Polygon	ESRI	2015
National forest	2	Polygon	NPS/ESRI	2015
National preserve	1	Polygon	UFGC, ESRI	2015
National seashore	2	Polygon	NPS, ESRI	2015
National monument	3	Polygon	NPS, ESRI	2015
National memorial	2	Polygon	NPS, ESRI	2015
National wildlife refuge	2	Polygon	NPS, ESRI	2015
State park	175	Polygon	ESRI	2015
Local park	1431	Polygon	ESRI	2015
Marine protected area site	391	Polygon	UFGC	2016
National marine sanctuary	1	Polygon	UFGC	2016
Amusement park	34	Point	UFGC	2016
Trail	1635	Line	UFGC	2016
Water access	774	Point	UFGC	2016
Community garden	6	Point	UFGC	2016
Historic park	37	Point	UFGC	2016

Note. AIF: ARDA International Foundation; BAR: Bureau of Archaeological Research; ESRI: Environmental Systems Research Institute; FAROC: Florida Association of RV Parks and Campgrounds; FDOT: Florida Department of Transportation; FFS: Florida Forestry Service; FFWC: Florida Fish and Wildlife Conservation Commission; FDEP: Florida Department of Environmental Protection; FRLA: Florida Restaurant and Lodging Association; NOAA: National Oceanic and Atmospheric Administration;

NPS: National Park Service; SSL: Superior Small Lodging; STR: Smith Travel Research; UFGC: University of Florida GeoPlan Center; USCB: U.S. Census Bureau

2.5 Step 3: Identification of specific categories of tourism resources

The identification (or naming) of tourism resources is one of the most basic and widespread practices in tourism marketing (Spotts, 1997). Factor analysis was used to identify specific categories of tourism resources based on previous studies (Formica & Uysal, 2006; Lovingood & Mitchell, 1989; Smith, 1987; Spotts, 1997). Factor analysis is a statistical technique that is used to reduce a large number of variables into a fewer number of factors. Based on previous studies, we excluded tourism resources that are missing or nearly missing in Florida (such as skating rinks) and combined similar resources (such as state and and county parks). Table 2.3 summarizes the selected variables included in factor analysis.

Suitability of data for factor analysis was evaluated using Bartlett's test of sphericity and the Kaiser-Meyer-Olkin Measure (KMO) of sampling adequacy (Table2. 4) and was judged satisfactory: Bartlett's test of sphericity $p<0.001$; KMO>0.5. As a result, 12 tourism resource factors (**Park tourism, theme park tourism, urban tourism, recreational boating/outdoor recreation, beach tourism, camping tourism, event tourism, aquarium/zoo tourism, sports tourism, cultural/heritage tourism, amusement park/casino tourism, and garden tourism**) were identified with eigenvalues greater than 1.0. These twelve factors explained 60% of variance in the data (Table 2.5). The resulting factors rotated with varimax are shown in Table 2.6.

Table 2.3 Variables included in factor analysis

Variable	Operational Definition	Date	Source(s)
Park Tourism			
Water resource	Acres of water resources, including rivers, springs, and lakes	2016	UFGC
Marine protected area	Acres of marine protected areas	2016	UFGC
National park	Acres of national parks	2016	ESRI
Park area	Acres of state, county, regional, and local parks	2015	ESRI
Theme Park Tourism			
Tourist attraction	Number of tourist attractions	2018	TripAdvisor
Hotel	Number of hotel rooms	2014	STR, UFGC, WS
Theme park	Number of theme parks and fairgrounds	2012	UFGC
Golf course	Acres of golf courses	2015	UFGC
Urban Tourism			
Drinking place	Number of drinking places, including bar and café	2016	GRI, FRLA
Restaurant	Number of full-service restaurants, cafés, and grills	2016	GRI, FRLA
Museum	Number of museums	2015	UFGC
Art gallery	Number of art galleries	2015	UFGC
Recreational Boating/ Outdoor Recreation			
Boat ramp	Number of boat ramps	2016	UFGC
Trail	Total miles of trails	2016	UFGC
Water access	Number of water access points (Lakes, rivers, and springs)	2016	UFGC
Scenic drive	Total miles of scenic highways and byways	2012	FDOT
Beach Tourism			
Beach area	Acres of beach areas	2002	FFWCC
Beach access	Number of beach access points	TBD	FDEP
Marina/Pier	Number of marinas and piers	2016	UFGC
Camping Tourism			
Mobile home park	Number of mobile home parks	2009	UFGC, FAROC
Campground	Number of campgrounds	2016	UFGC, FAROC

Table 2.3 (continue)

Variable	Operational Definition	Date	Source(s)
Event Tourism			
Horse tract	Number of horse tracts	2012	UFGC
Race tract	Number of race tracts	2012	U.S. Census Bureau
Aquarium/Zoo Tourism			
Aquarium	Acres of aquarium areas	2015	UFGC
Zoo	Number of zoos	2015	UFGC
Sports Tourism			
Airbnb	Number of Airbnb rooms	2016	AirDNA
Stadium arena	Acres of stadium arenas	2012	UFGC
Cultural/Heritage Tourism			
Historic building	Acres of historic building areas	2016	BAR
Historic site	Number of historic sites	2016	BAR
Amusement Park/Casino Tourism			
Amusement park	Acres of amusement parks	2016	UFGC
Casino	Number of casinos	2012	UFGC
Garden Tourism			
Botanic garden	Acres of botanic gardens	2015	UFGC

Note: AirDNA: AIRDNA Inc.; BAR: Bureau of Archaeological Research; ESRI: Environmental Systems Research Institute; FAROC: Florida Association of RV Parks and Campgrounds; FDOT: Florida Department of Transportation; FFWCC: Florida Fish and Wildlife Conservation Commission; FDEP: Florida Department of Environmental Protection; FRLA: Florida Restaurant and Lodging Association; GRI: Geographic Research Inc.; NPS: National Park Service; STR: Smith Travel Research; UFGC: University of Florida GeoPlan Center; USCB: U.S. Census Bureau; WS: Web-scrapping program.

Table 2.4 Results of KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.652
Bartlett's Test of Sphericity	Approx. Chi-Square	28293.042
	df	0.496
	Sig.	0.000

Table 2.5 Eigenvalues for and percentage of variance explained by the twelve-factor models

	Factor												
	1	2	3	4	5	6	7	8	9	10	11	12	Total
Eigenvalues	3.201	2.951	1.962	1.708	1.489	1.261	1.212	1.166	1.130	1.080	1.032	1.028	
Percentage of total variance explained	10.004	9.222	6.133	5.336	4.654	3.939	3.788	3.645	3.531	3.375	3.226	2.212	60.065
Percentage of common variance explained	16.655	15.353	10.211	8.884	7.748	6.558	6.307	6.068	5.879	5.619	5.371	3.683	100

Table 2.6 Factor loadings (loadings of 0.40 or greater are shown)

Interpretation	Variables	Factors											
		1	2	3	4	5	6	7	8	9	10	11	12
Parks	Water resource	0.95											
	Marine prot. area	0.86											
	National park	0.82											
	Park area	0.53											
Theme parks	Tourist attraction		0.86										
	Hotel		0.86										
	Theme park		0.73										
	Golf course		0.44										
Urban	Drinking place			0.76									
	Restaurant			0.72									
	Museum			0.64									
	Art gallery			0.63									
Recreational boating/outdoor	Boat ramp				0.79								
	Trail				0.66								
	Water access				0.52								
	Scenic drive				0.48								
Beach	Beach area					0.82							
	Beach access					0.81							
	Marina/Pier					0.41							
Camping	Mobile home park						0.81						
	Campground						0.71						
Events	Horse tract							0.79					
	Race tract							0.77					
Aquarium and zoo	Aquarium area								0.78				
	Zoo								0.75				
Sports	Airbnb									0.73			
	Stadium Arena									0.65			
Cultural and heritage	Historic building										0.73		
	Historic site										0.62		
Amusement	Amusement park											0.70	
	Casino											0.69	
Gardens	Botanic garden												0.88

The factors were interpreted and labeled as follows:

- Factor 1 was labeled "Park Tourism." The variables that loaded highly on this factor were Water resource (acres of water resources, including rivers, springs, and lakes), Marine protected area (acres of marine protected areas), National park (acres of national parks), and Park area (acres of state, county, regional, and local parks).
- Factor 2 was labeled "Theme Park Tourism." The variables that loaded highly on this factor were Tourist attraction (number of tourist attractions), Hotel (number of hotel rooms), Theme park (number of theme parks and fairgrounds), and Golf course (acres of golf courses).
- Factor 3 was labeled "Urban Tourism" and included the following variables: Drinking place (number of drinking places, including bar and café), Restaurant (number of full-service restaurants, cafés, and grills), Museum (number of museums), Art gallery (number of art galleries).
- Factor 4 was labeled "Recreational Boating/Outdoor Recreation" and contained Boat ramp (number of boat ramps), Trail (total miles of trails), Water access (number of water access points), and Scenic drive (total miles of scenic highways and byways).
- Factor 5 was labeled "Beach Tourism." The variables that loaded highly on this factor were Beach area (acres of beach areas), Beach access (number of beach access points), and Marina/Pier (number of marinas and piers).
- Factor 6 was labeled "Camping Tourism" and contained variables Mobile home park (number of mobile home parks) and Campground (number of campgrounds).
- Factor 7 was labeled "Event Tourism." The variables that loaded highly on this factor were Horse tract (number of horse tracts) and Race tract (number of race tracks).
- Factor 8 was labeled "Aquarium/Zoo Tourism." The variables that loaded highly on this factor were Aquarium (acres of aquarium areas) and Zoo (number of zoos).
- Factor 9 was labeled "Sports Tourism." The variables that loaded highly on this factor were Airbnb (number of Airbnb rooms) and Zoo (acres of stadium arenas).
- Factor 10 was labeled "Cultural/Heritage Tourism." The variables that loaded highly on this factor were Historic building (acres of historic building areas) and historic site (number of historic sites).
- Factor 11 was labeled "Amusement Park/Casino Tourism." The variables that loaded highly on this factor were Amusement park (acres of amusement parks) and Casino (number of casinos).
- Factor 12 was labeled "Garden Tourism." The variables that loaded highly on this factor were Botanic garden (acres of botanic gardens).

2.6 Step 4: Development of tourism resource indices

The purpose of an index is to combine a number of related measures into a single measure. Based on Smith's (1987) regional analysis approach, we created a series of tourism resource indices by producing factor scores. These scores were obtained by multiplying the component loading for a particular variable on a particular factor by the county's (or census tract) original

score for that variable. This process was repeated for every variable on a factor. The products were then summed. This was repeated for every other factor for that county (or census tract), and ultimately for all counties (67) and census tracts (4245). Finally, the factor scores were standardized with a mean of 0.0 and a standard deviation of 1.0. These scores were used to visualize the county (or census tract)-level variations of identified tourism resources.

2.7 Step 5: Mapping tourism regions and highway accessibility and availability

2.7.1 Spatial distribution of tourism resource factors

To illustrate how the twelve tourism resource factors were distributed over space, standardized scores for each factor were computed and mapped (Figures 2.2-2.27). The scores were aggregated into five ranges: greatly below average (scores below -1.9 standard deviations); below average (-1.0 to – 1.9 standard deviations); near average (-0.9 to 0.9 standard deviations); above average (1.0 to 1.9 standard deviations); and greatly above average (scores above 1.9 standard deviations). This classification is justified by adopting Spotts' (1997) and Smith's (1987) approach.

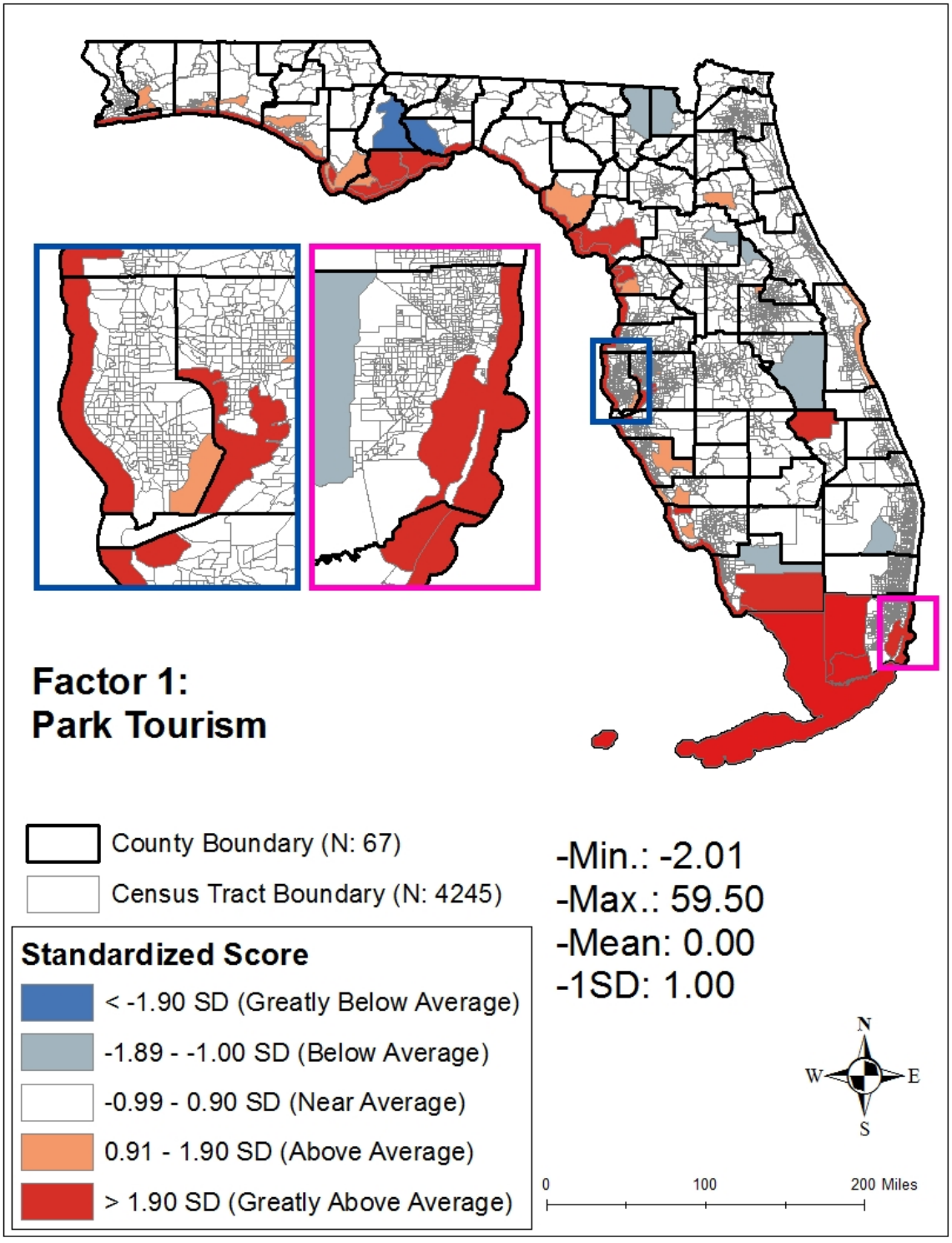


Figure 2.2 Standardized scores for factor 1: Park Tourism (census tract)

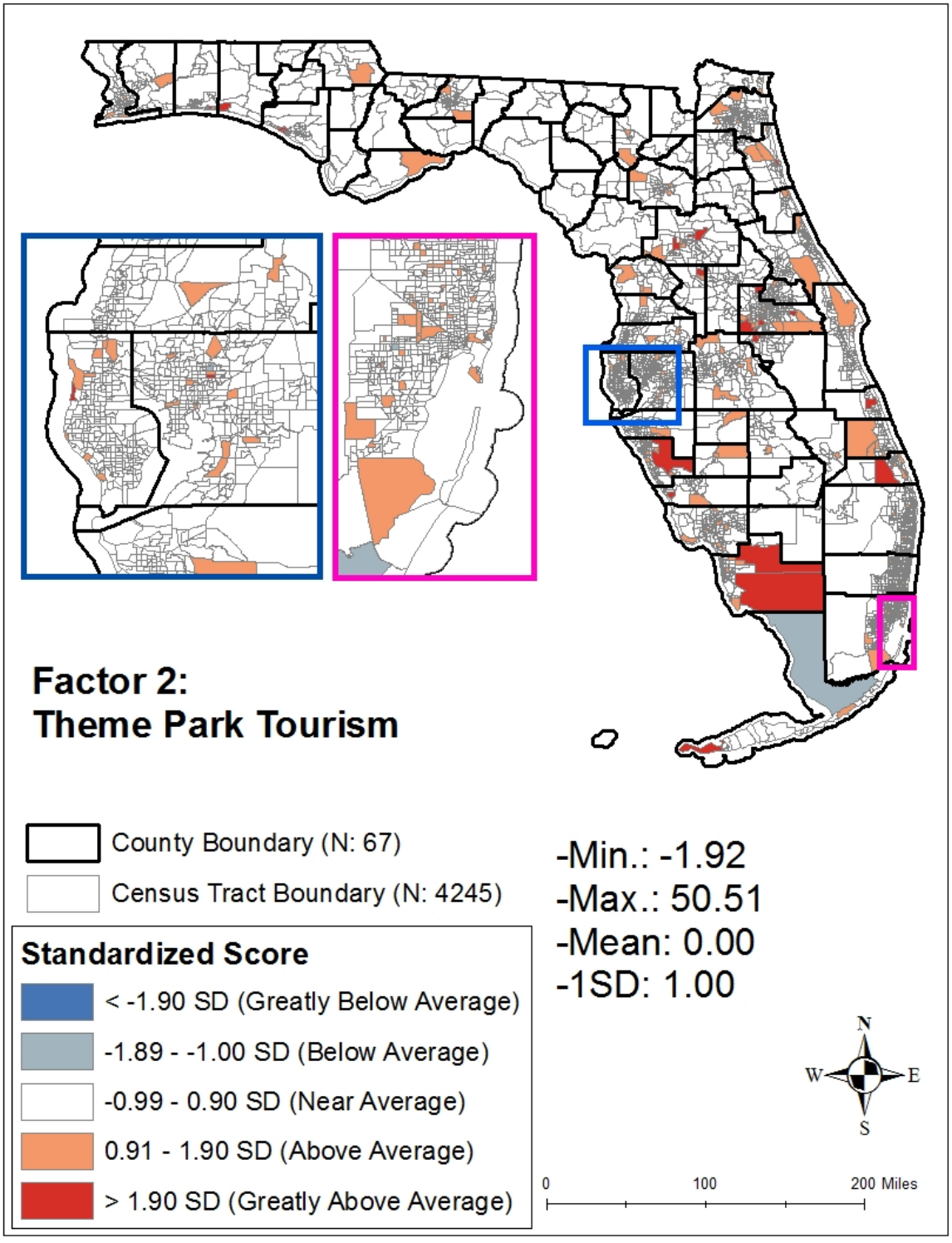


Figure 2.3 Standardized scores for factor 2: Theme Park Tourism (census tract)

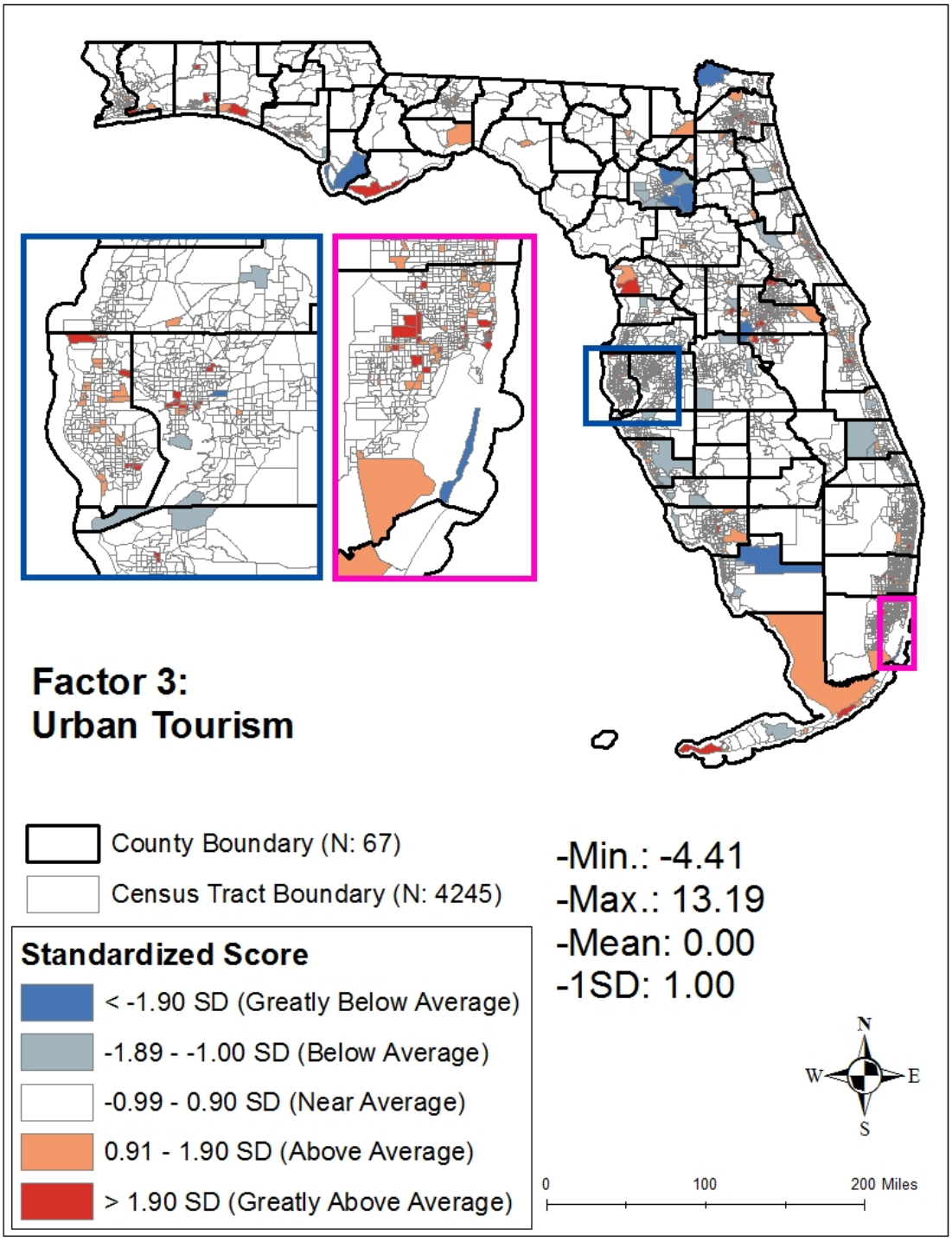


Figure 2.4 Standardized scores for factor 3: Urban Tourism (census tract)

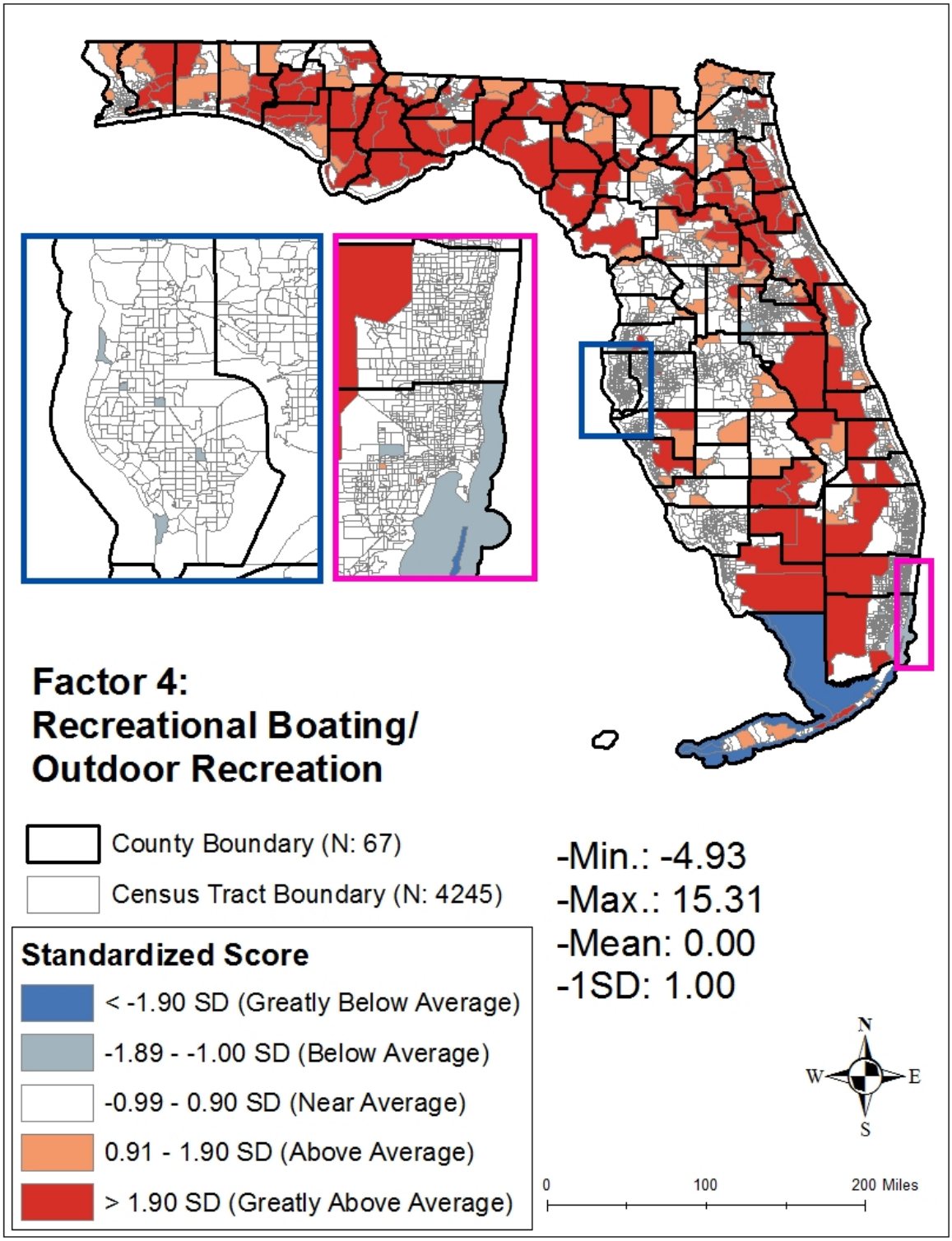


Figure 2.5 Standardized scores for factor 4: Boating/Outdoor Recreation (census tract)

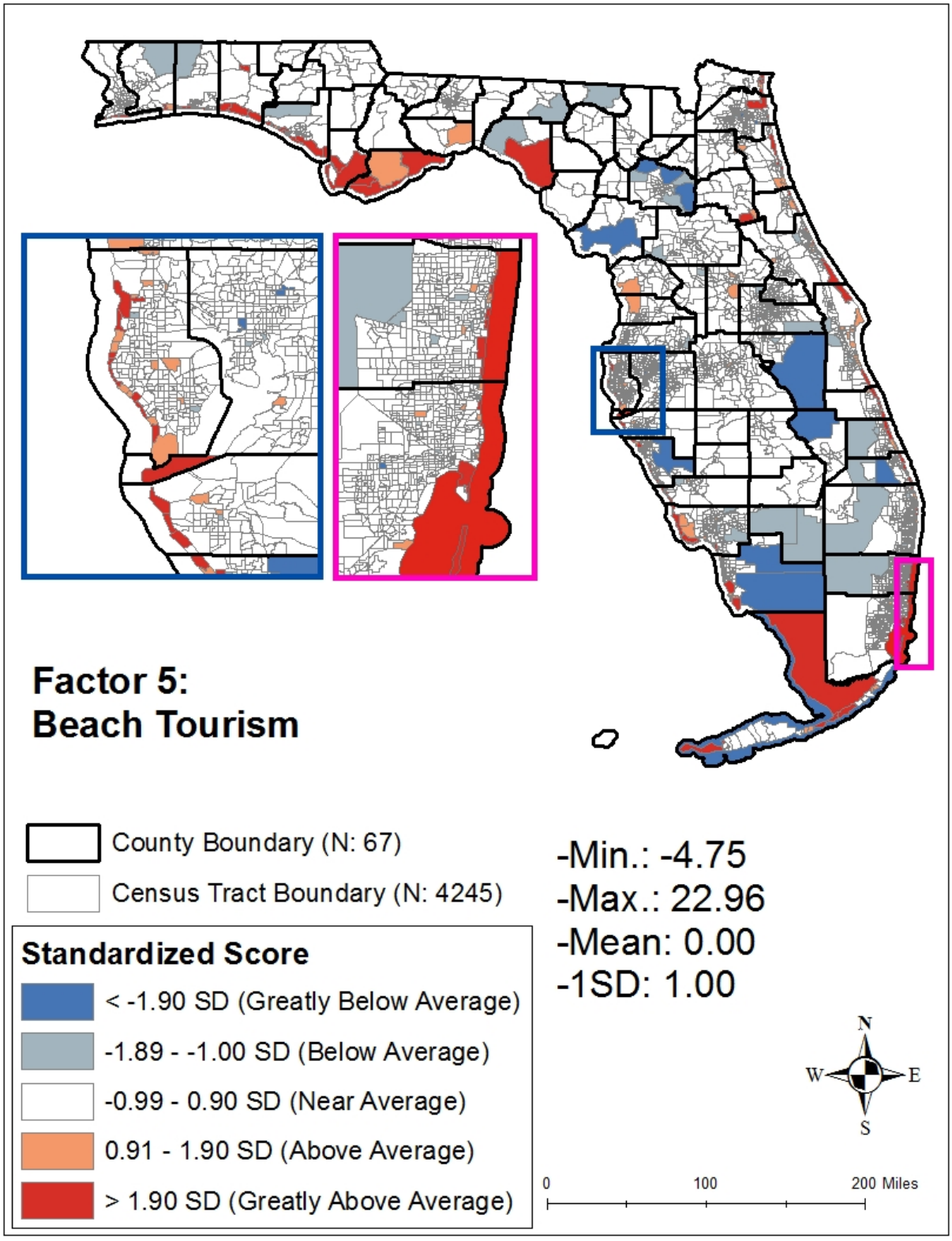


Figure 2.6 Standardized scores for factor 5: Beach Tourism (census tract)

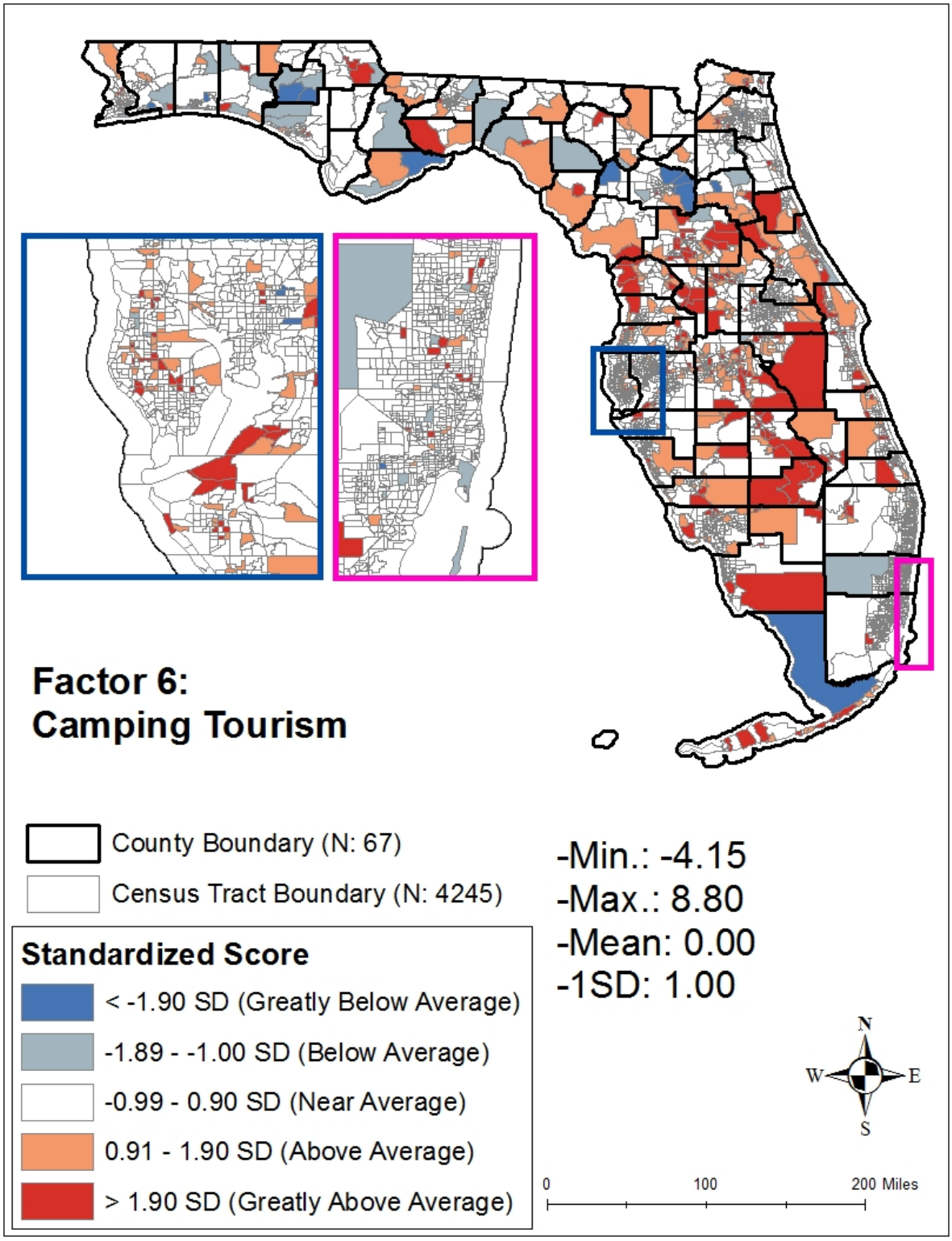


Figure 2.7 Standardized scores for factor 6: Camping Tourism (census tract)

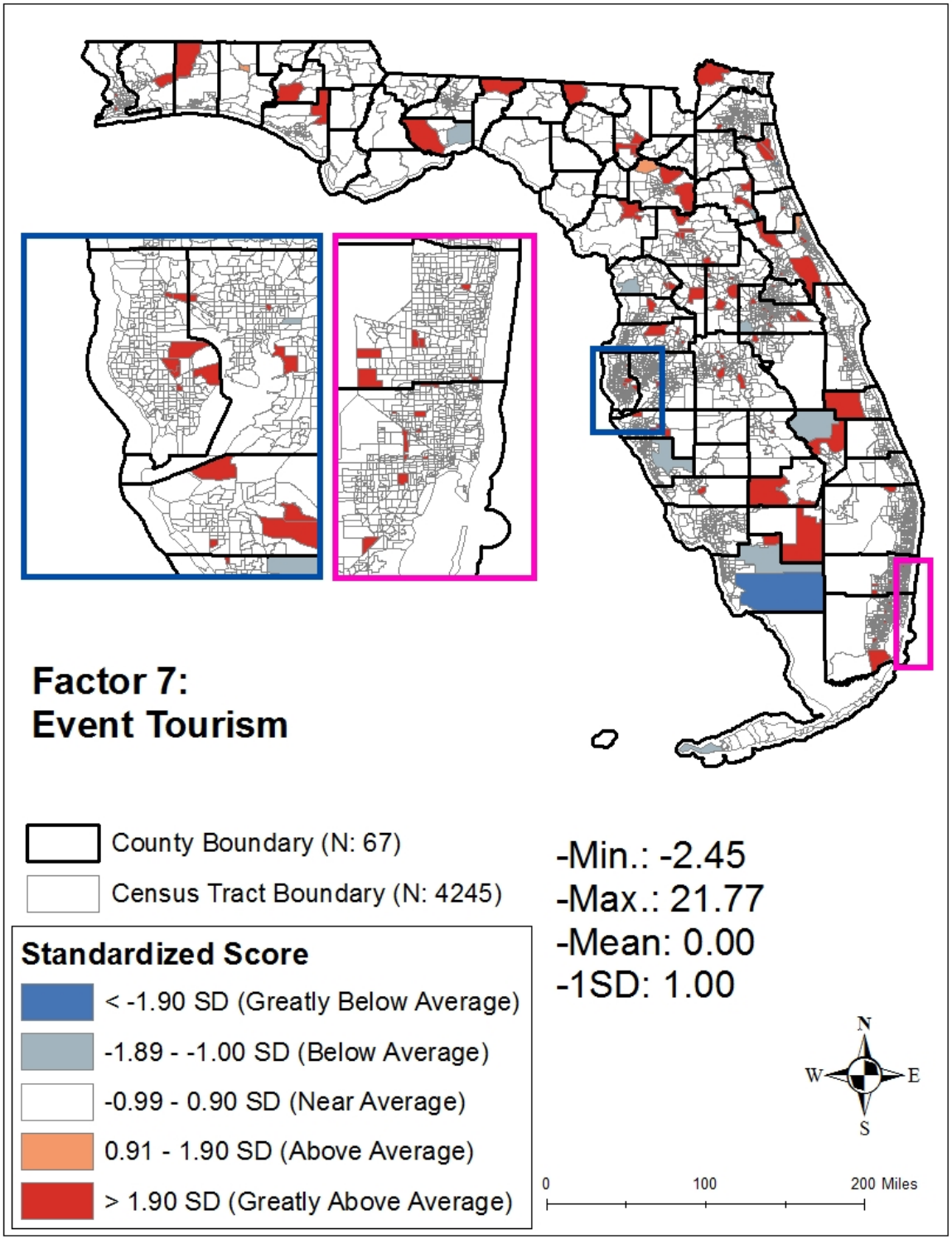


Figure 2.8 Standardized scores for factor 7: Event Tourism (census tract)

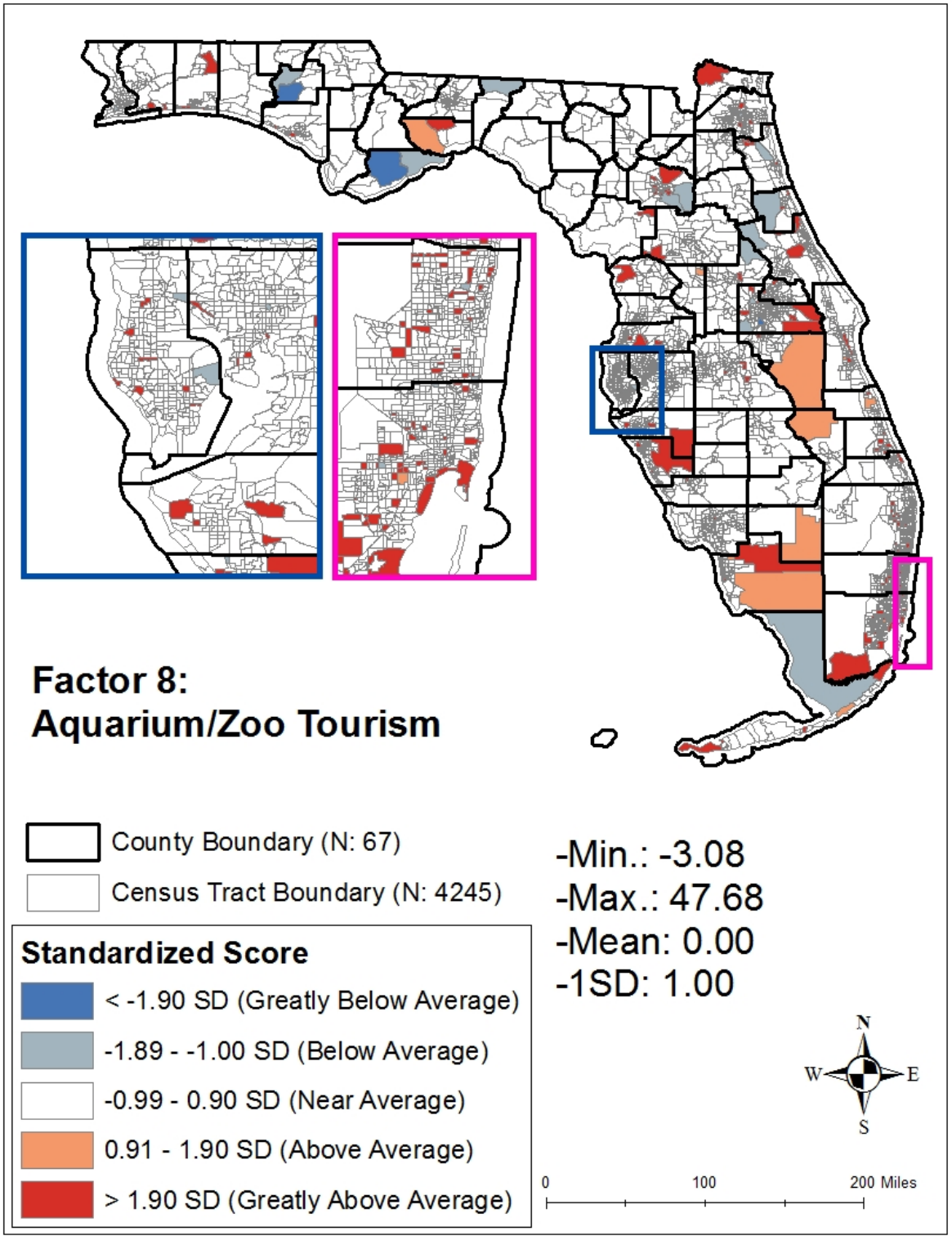


Figure 2.9 Standardized scores for factor 8: Aquarium/Zoo Tourism (census tract)

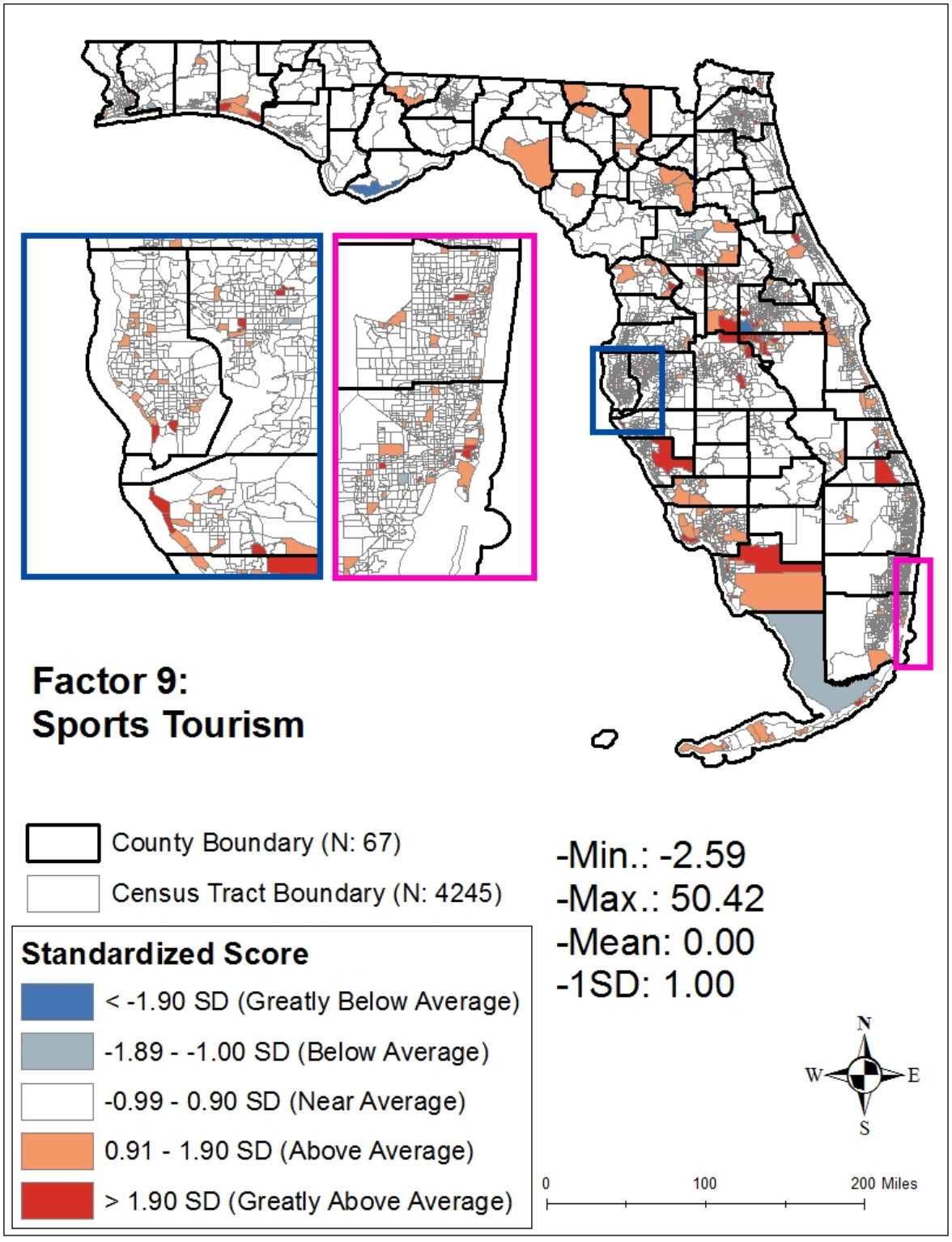


Figure 2.10 Standardized scores for factor 9: Sports Tourism (census tract)

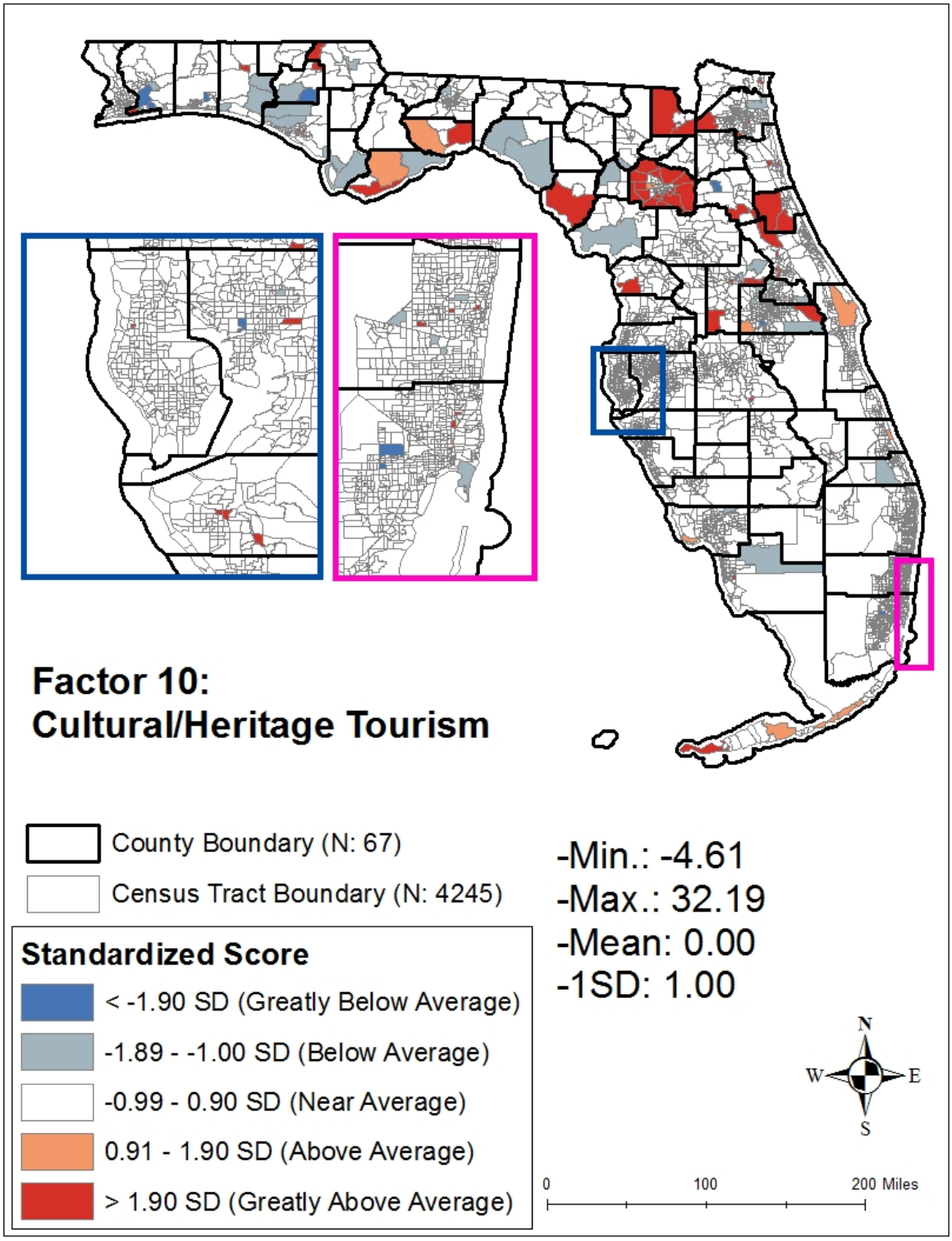


Figure 2.11 Standardized scores for factor 10: Cultural/Heritage Tourism (census tract)

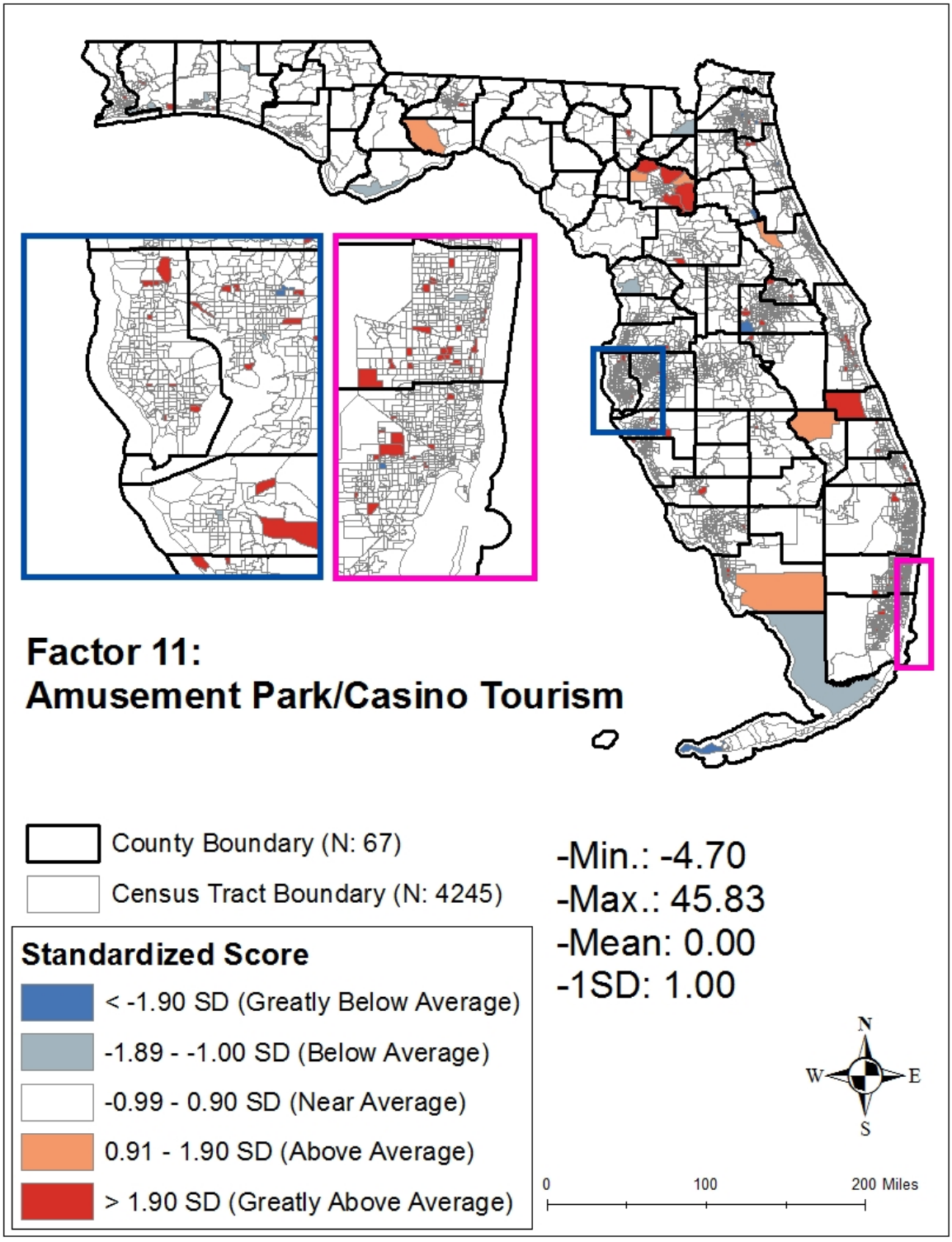


Figure 2.12 Standardized scores for factor 11: Amusement Park/Casino Tourism (census tract)

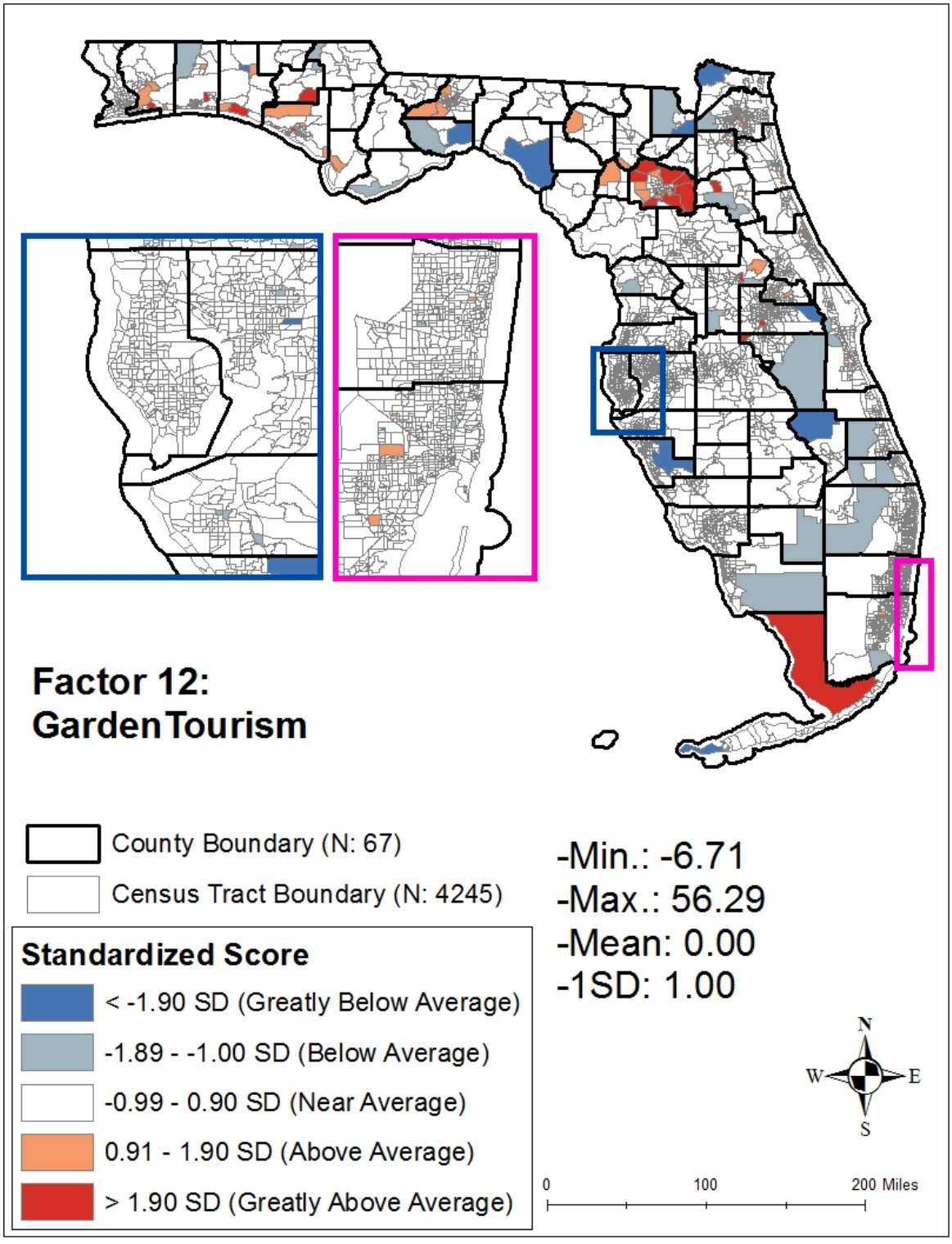


Figure 2.13 Standardized scores for factor 12: Garden Tourism (census tract)

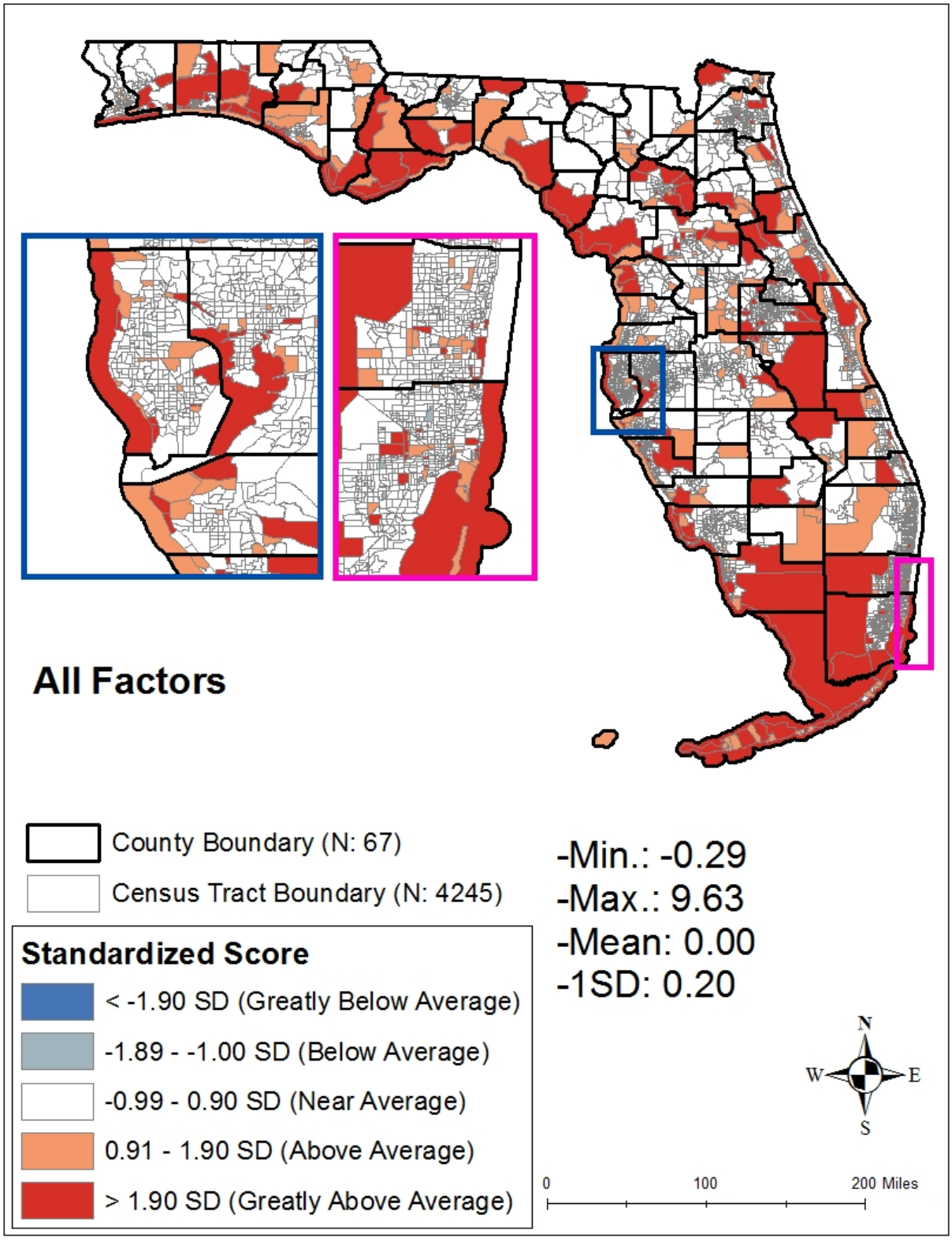


Figure 2.14 Standardized scores for all combined factors (census tract)

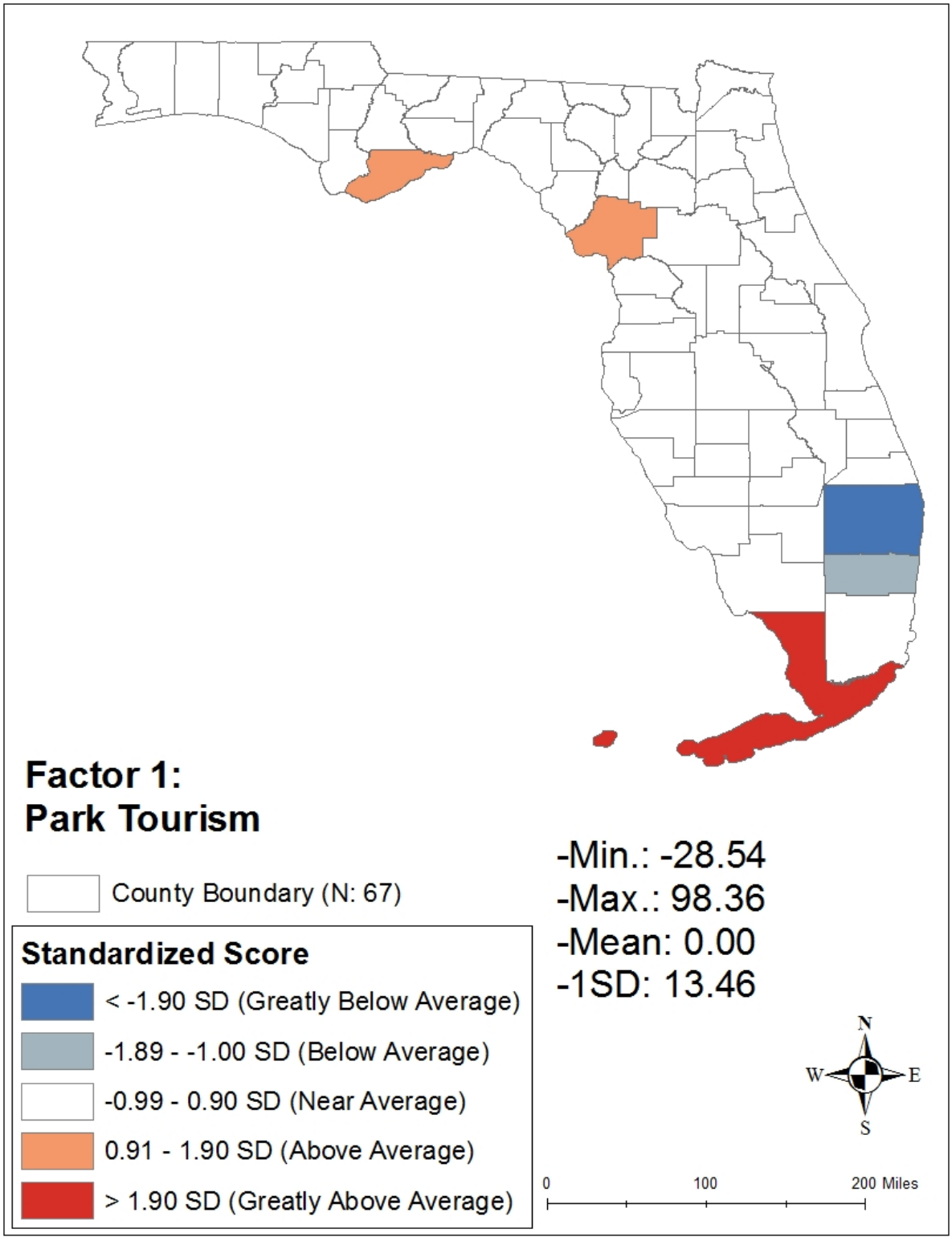


Figure 2.15 Standardized scores for factor 1: Park Tourism (county)

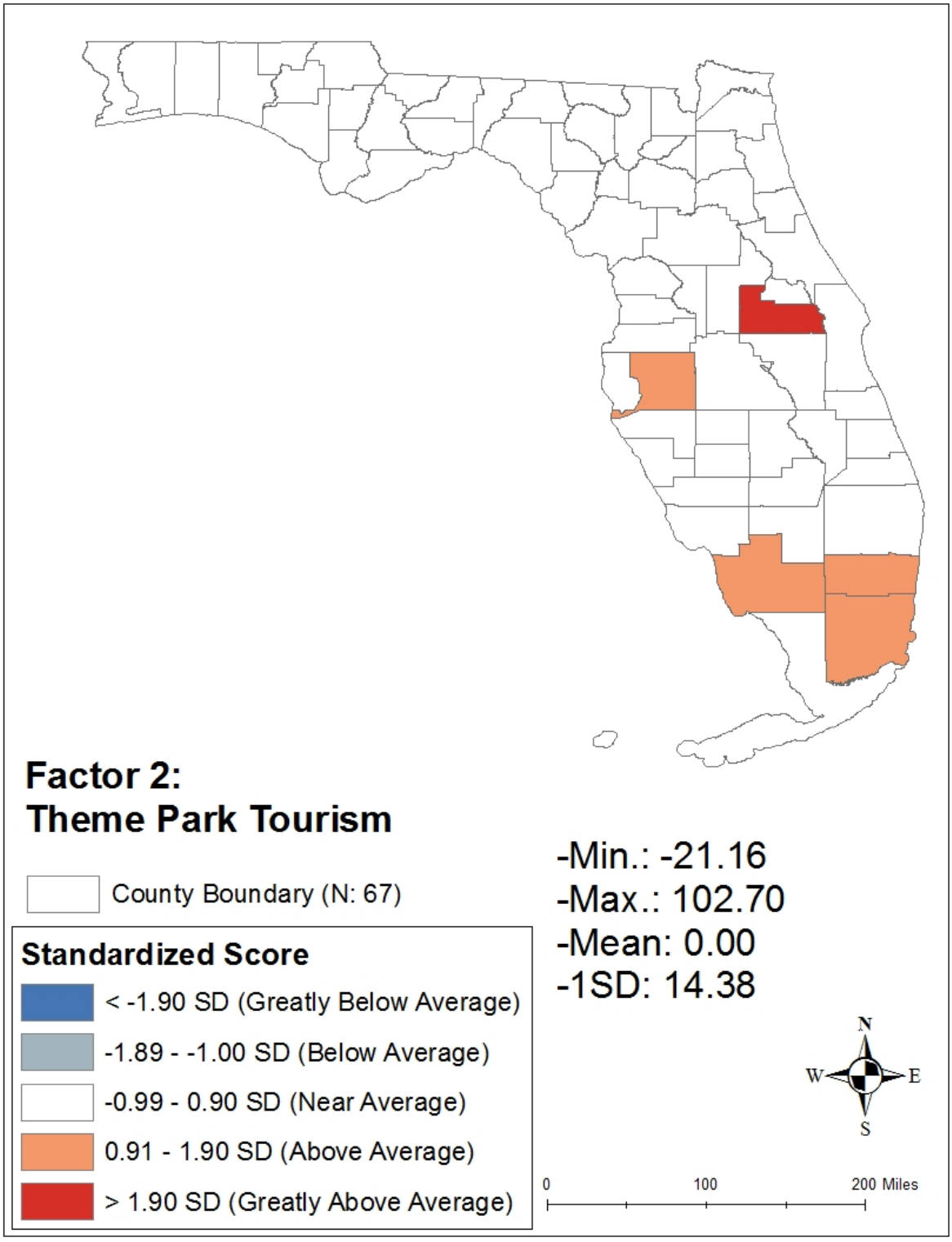


Figure 2.16 Standardized scores for factor 2: Theme Park Tourism (county)

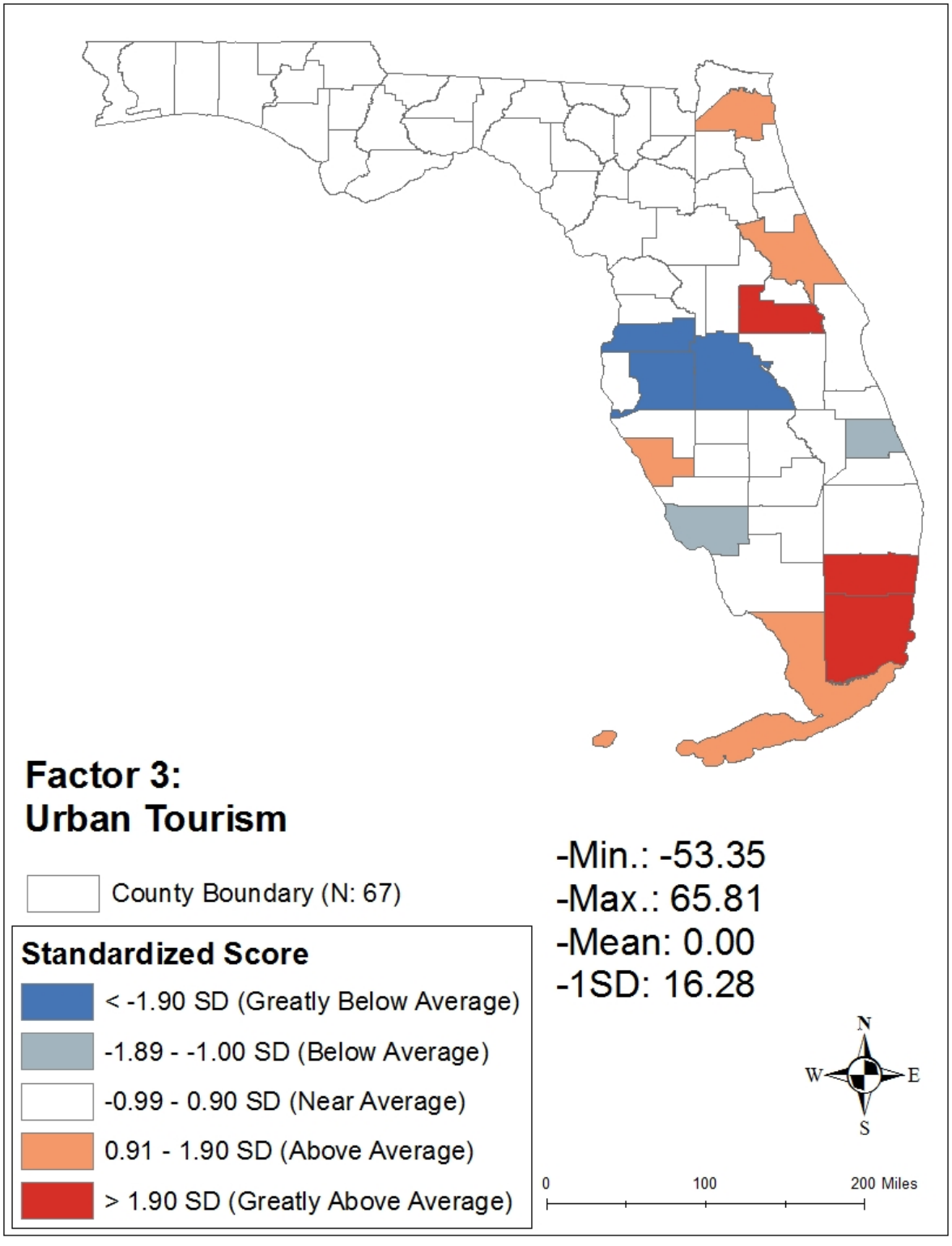


Figure 2.17 Standardized scores for factor 3: Urban Tourism (county)

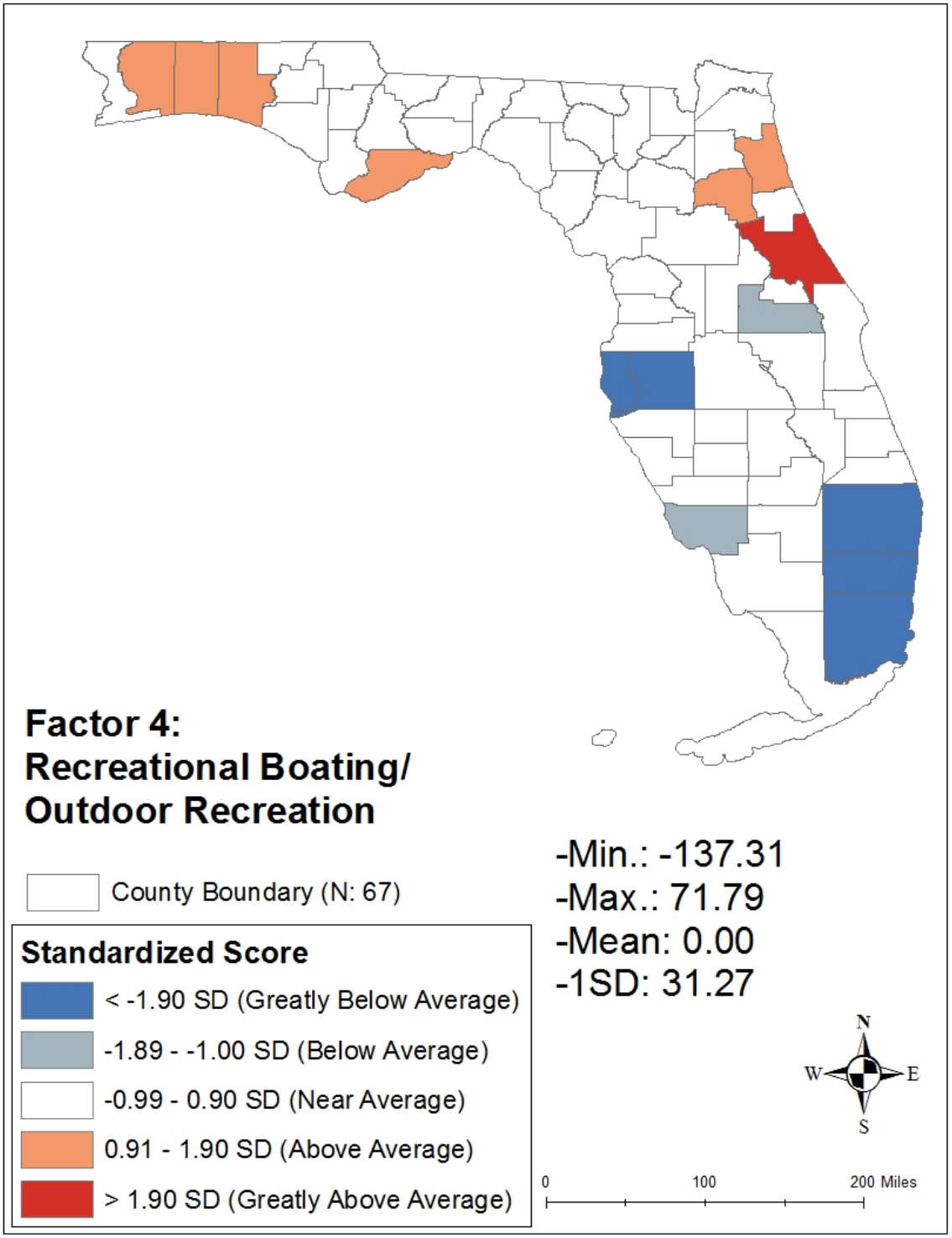


Figure 2.18 Standardized scores for factor 4: Recreational Boating/Outdoor Recreation (county)

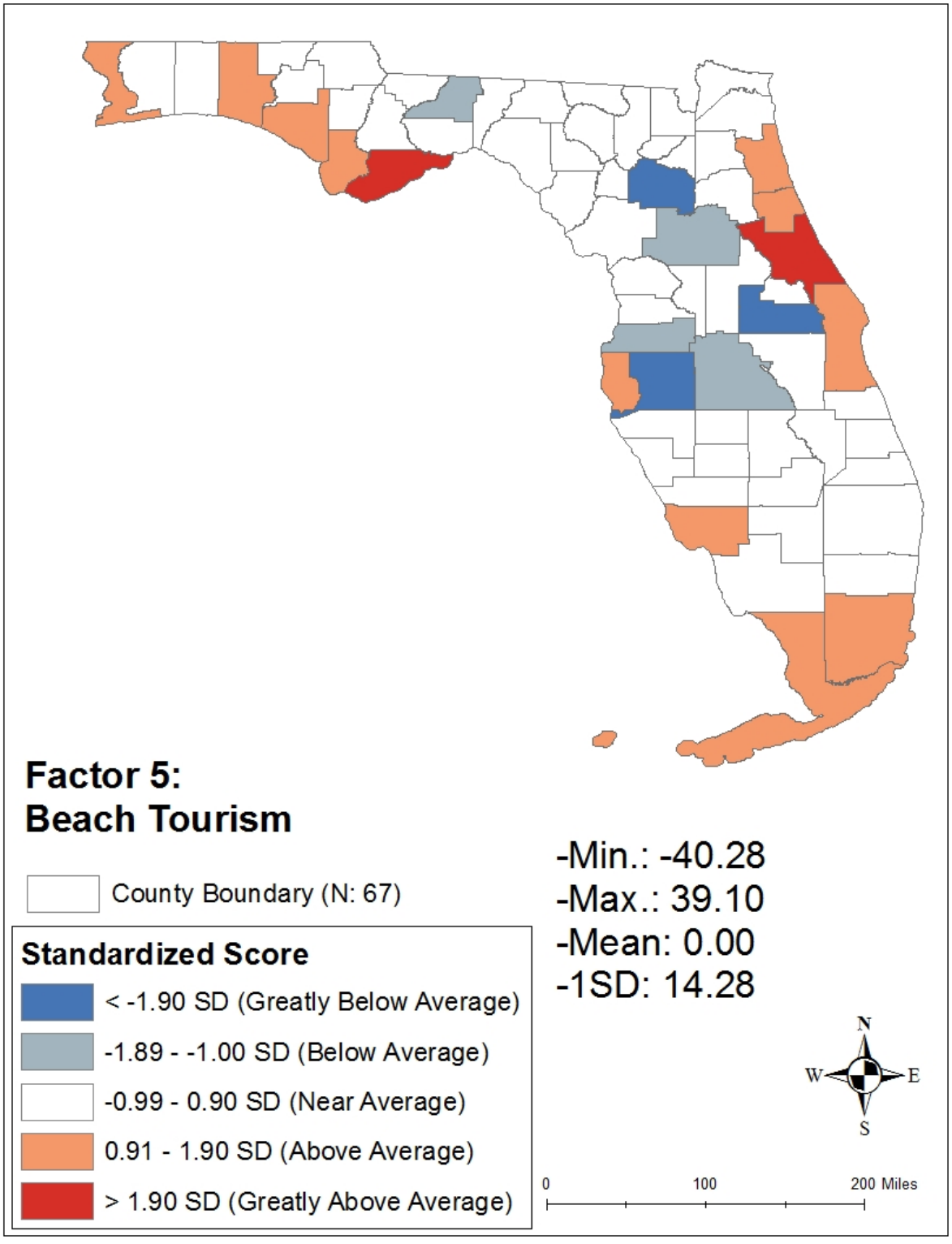


Figure 2.19 Standardized scores for factor 5: Beach Tourism (county)

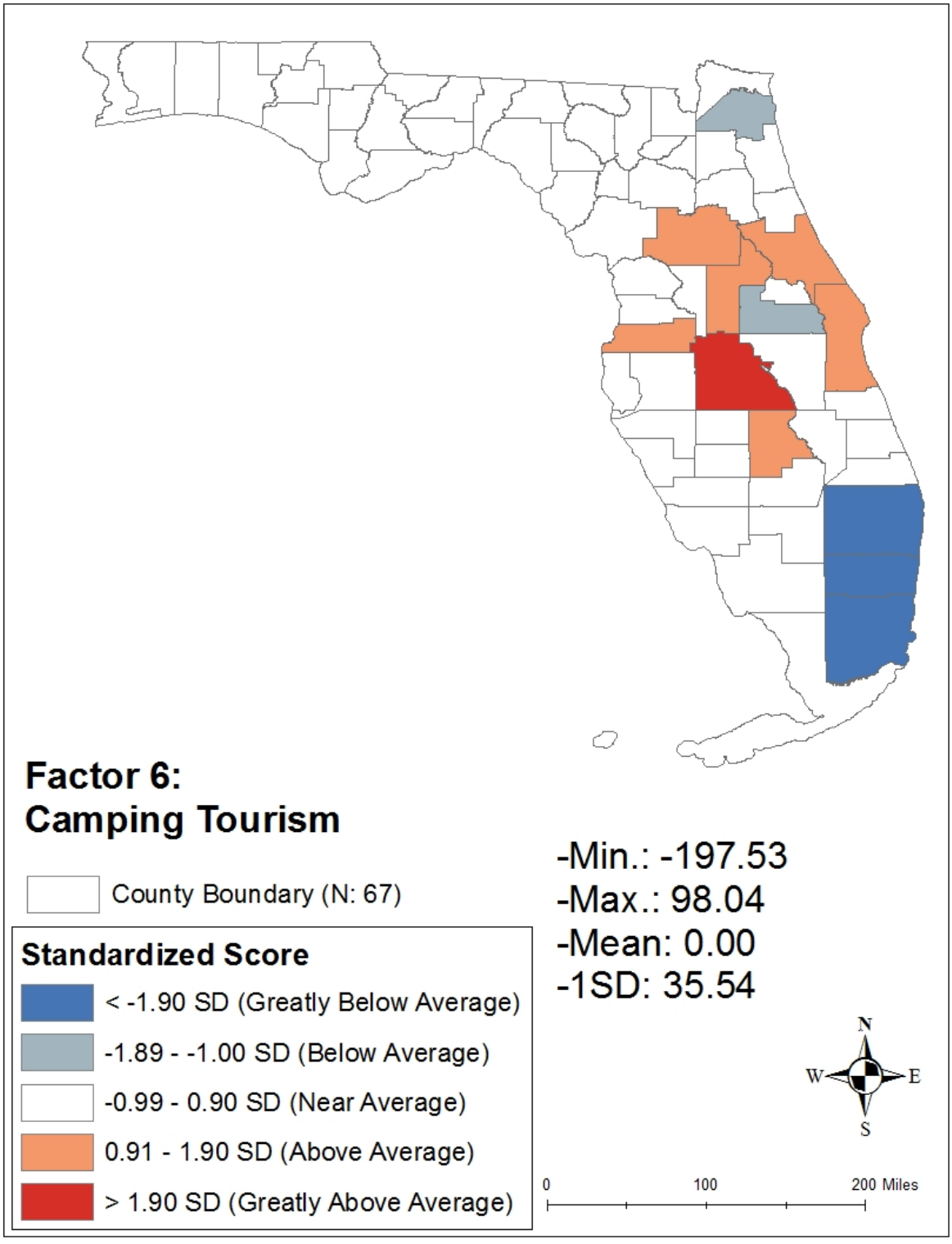


Figure 2.20 Standardized scores for factor 6: Camping Tourism (county)

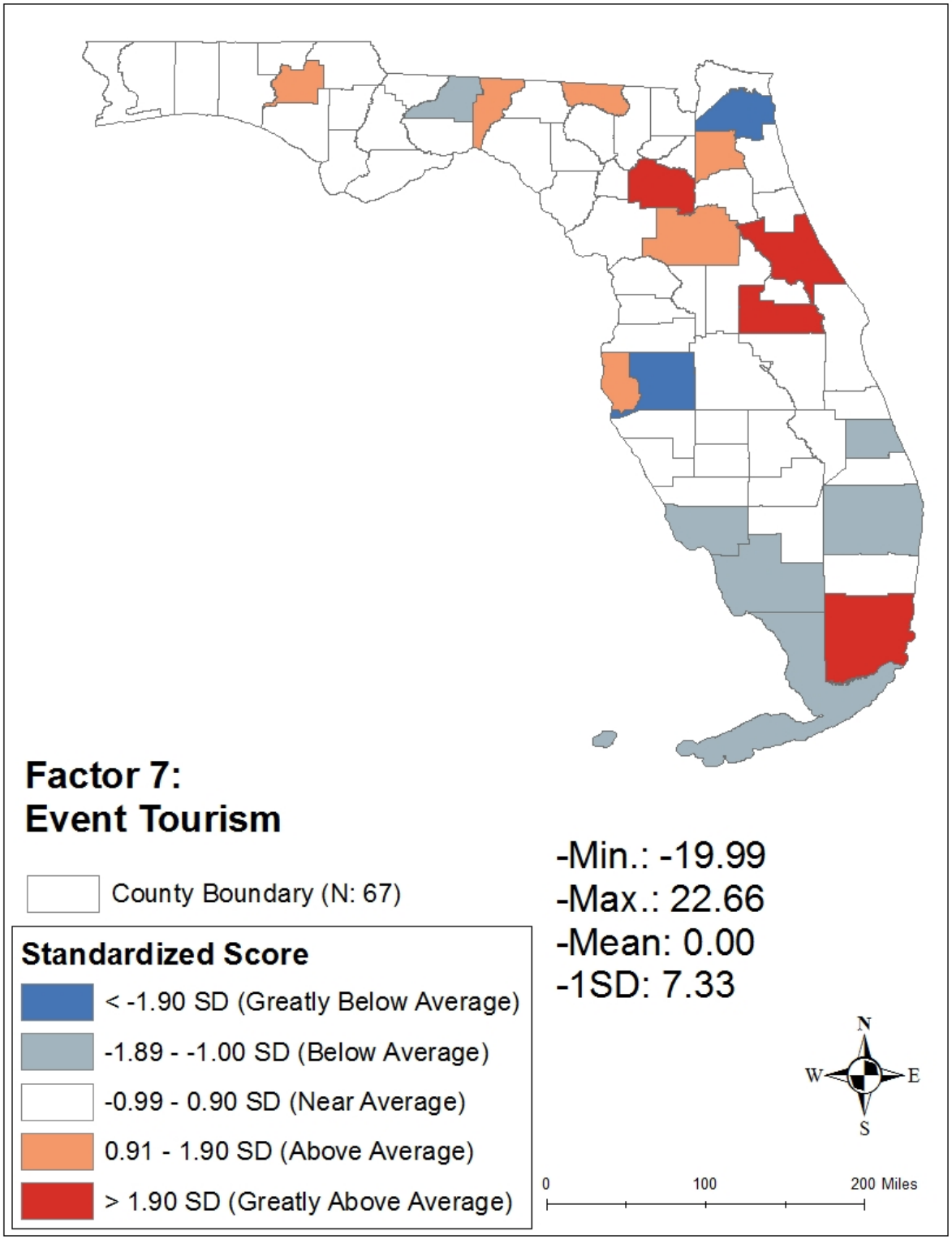


Figure 2.21 Standardized scores for factor 7: Event Tourism (county)

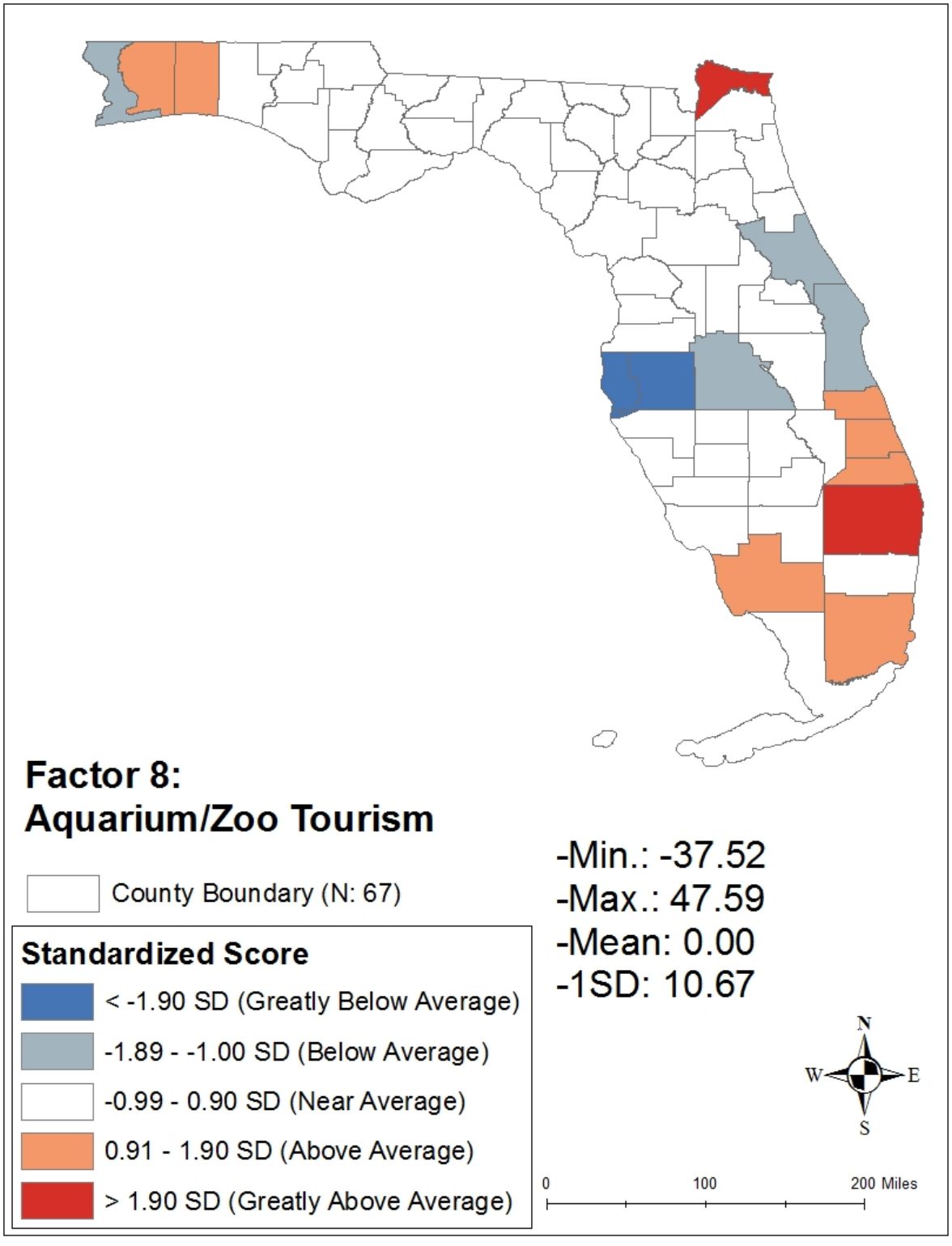


Figure 2.22 Standardized scores for factor 8: Aquarium/Zoo Tourism (county)

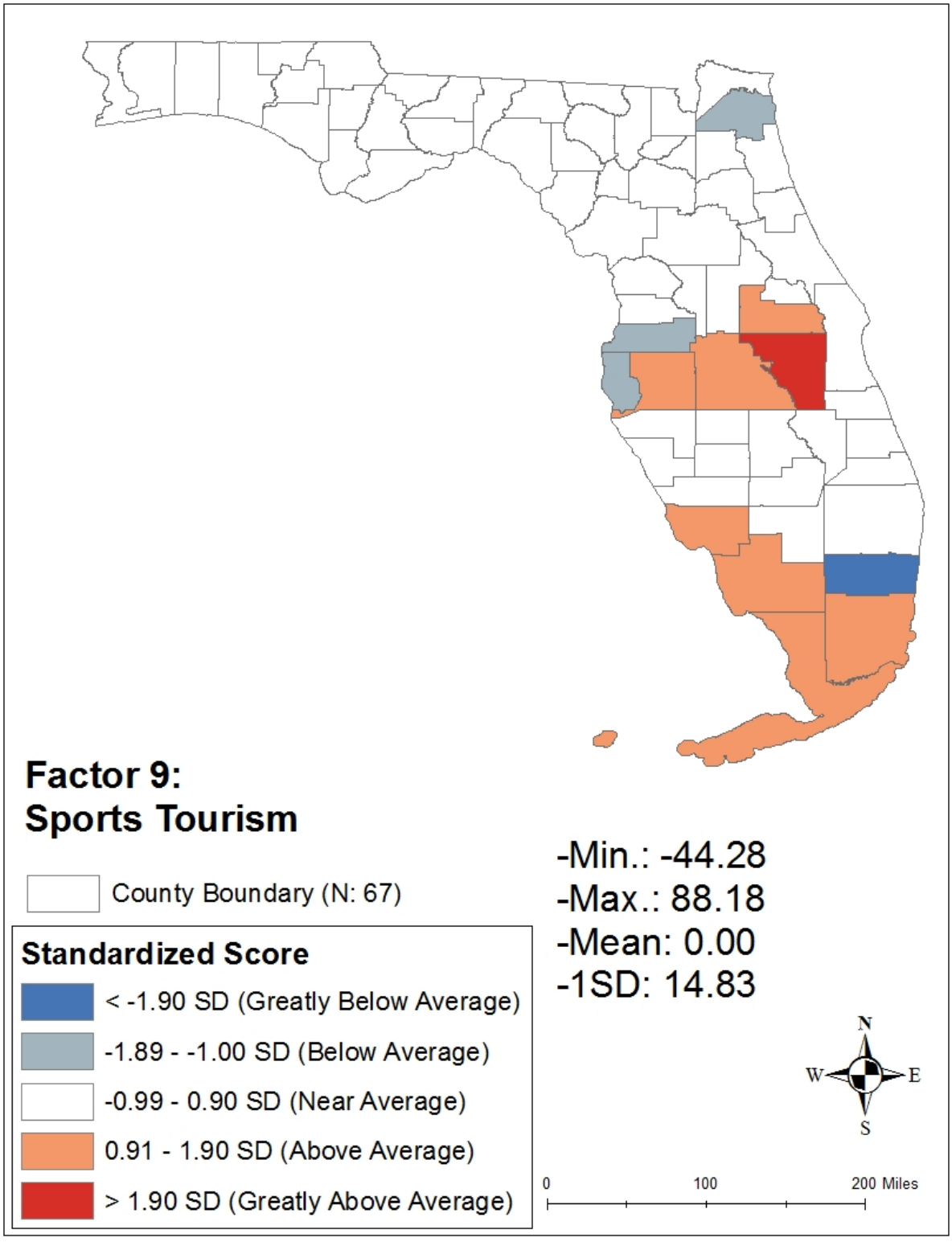


Figure 2.23 Standardized scores for factor 9: Sports Tourism (county)

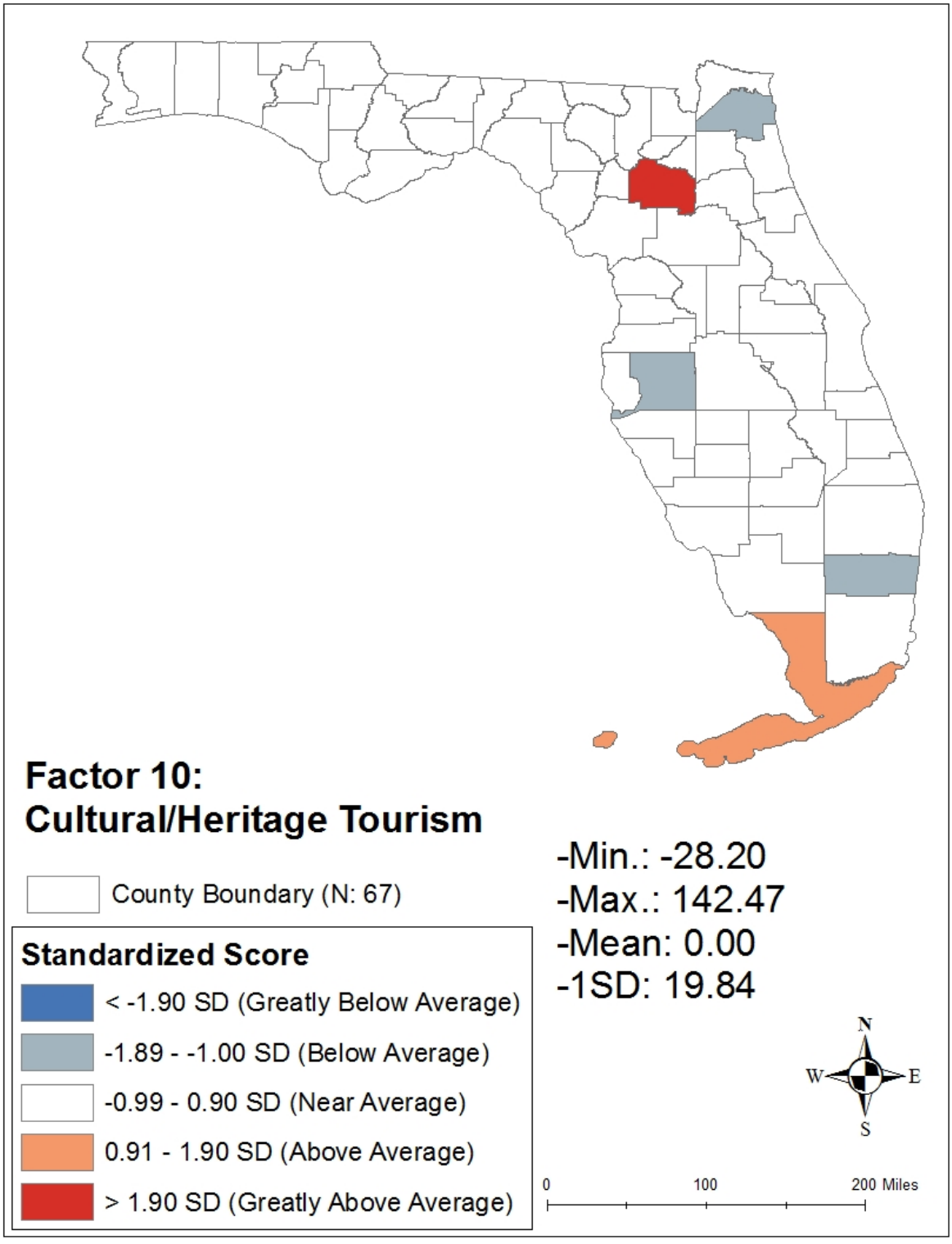


Figure 2.24 Standardized scores for factor 10: Cultural/Heritage Tourism (county)

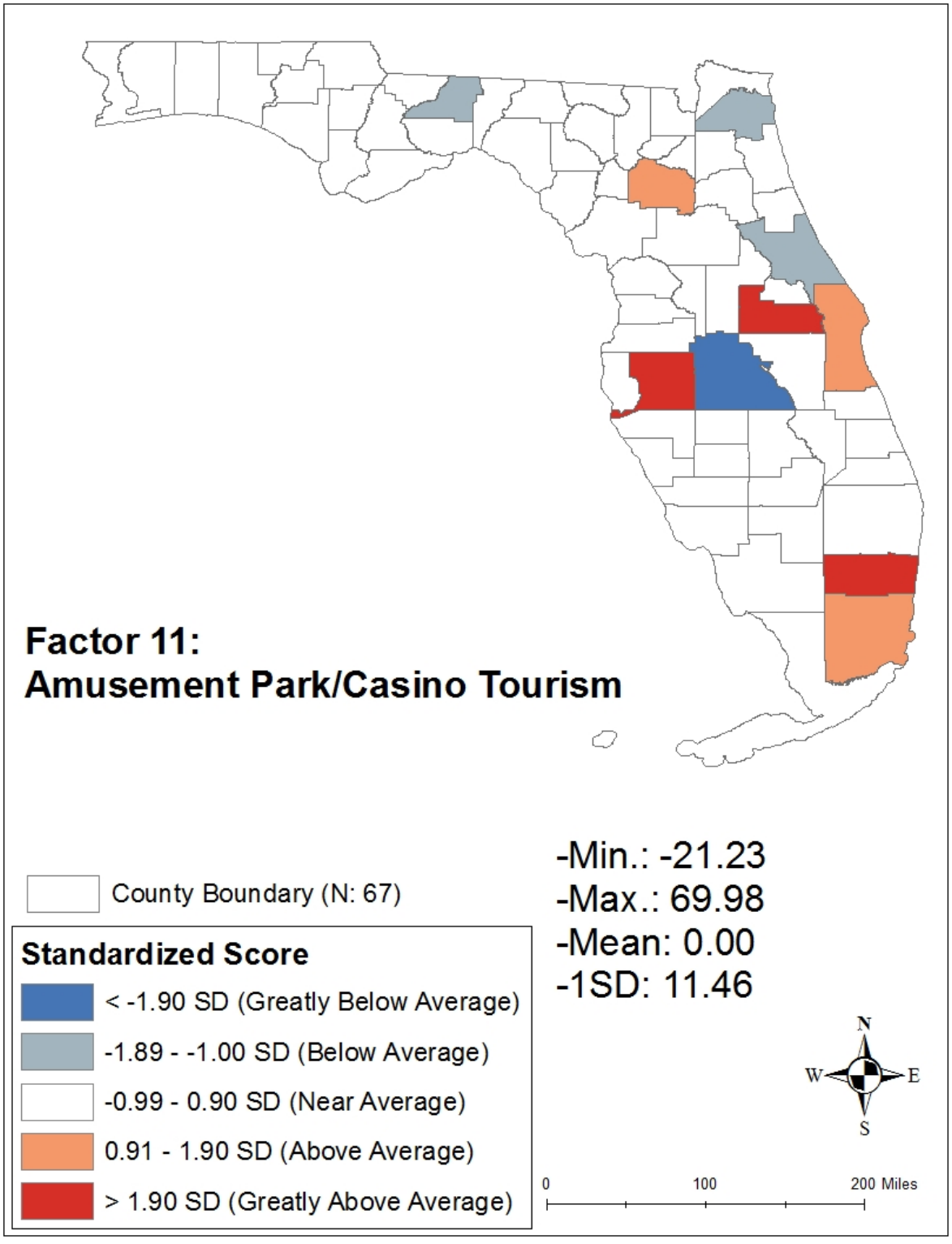


Figure 2.25 Standardized scores for factor 11: Amusement Park/Casino Tourism (county)

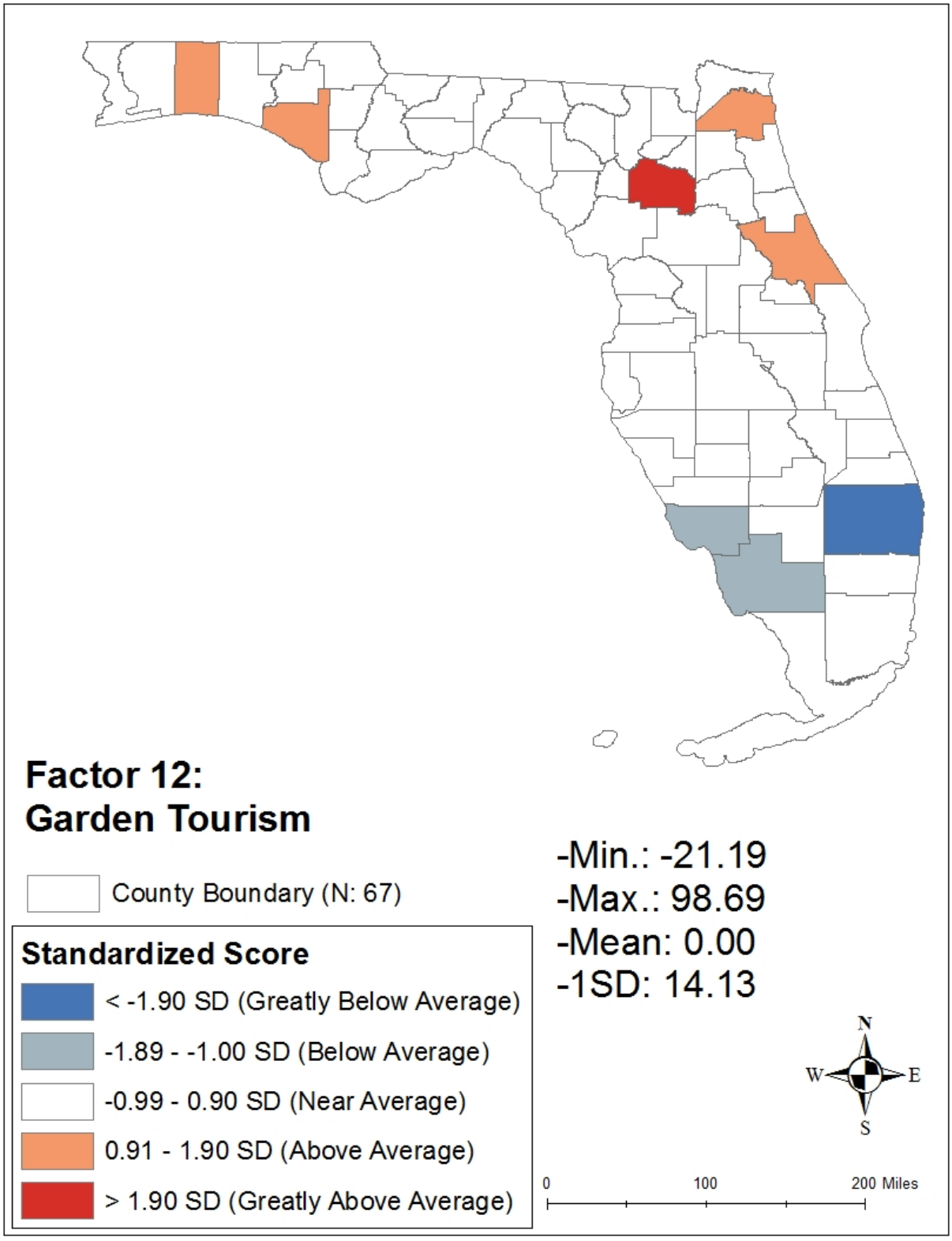


Figure 2.26 Standardized scores for factor 12: Garden Tourism (county)

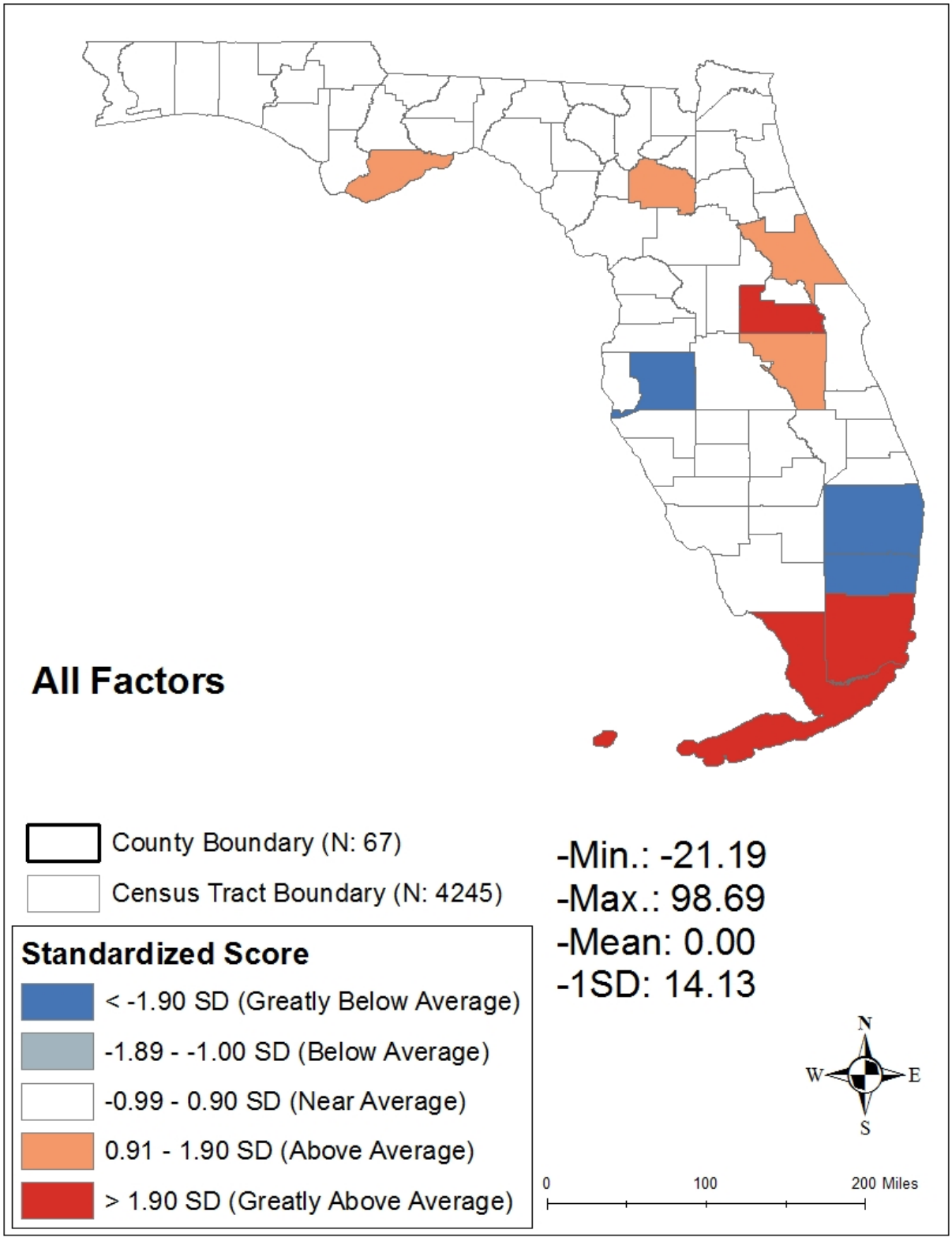


Figure 2.27 Standardized scores for all combined factors (county)

2.7.2 Exploring the spatial patterns of tourism resource factors

While mapping with tourism resource factors using GIS can show the spatial patterns of the tourism resource factors, general mapping technique cannot reference statistical significance of the spatial patterns. To reveal the significant association between the factor scores, local indicators of spatial association (LISA) analysis was employed. As one of the exploratory spatial data analysis techniques, LISA can identify the location and type of spatial cluster in a data set based on the concept of spatial dependence (Kang et al., 2014). In this task, LISA can be calculated as follows:

$$I_i = \frac{(x_i - \mu)}{m_2} \sum_i w_{ij}(x_j - \mu),\ m_2 = \sum_i (x_i - \mu)^2 / N \quad (2.1)$$

where I_i is the local Moran's I statistic at county (or census tract) i; w_{ij} is the matrix of weights such that $w_{ij} = 1$ if county (or census tract) i and county (or census tract) j are adjacent; otherwise, $w_{ij} = 0$, x_i is the attribute value of a specific factor at county (or census tract) i, x_j is the attribute value of a specific factor at county (or census tract) j, μ is the average attribute value of a specific factor, and N is the total number of counties (n: 67) and census tracts (n: 4245).

Results of LISA analysis can be presented in five categories: (1) High-High (HH): spatial clusters of counties (or census tracts) with high factor scores, indicating hot spots; (2) High-Low (HL): counties (or census tracts) with high factor score adjacent to counties (or census tracts) with low factor scores, indicating spatial outliers; (3) Low-High (LH): counties (or census tracts) with low factor scores adjacent to counties (or census tracts) with high factor scores; (4) Low-Low (LL): spatial clusters of counties (or census tracts) with low factor scores, indicating cold spots; and (5) Not Significant: no clustering between counties (or census tracts) (Jang et al., 2017; Jang & Kim, 2018). ArcGIS 10.4.1 was also applied for the LISA analyses. Figures 2.28 -2.53 show the hot spots and cold spots of tourism resource factors.

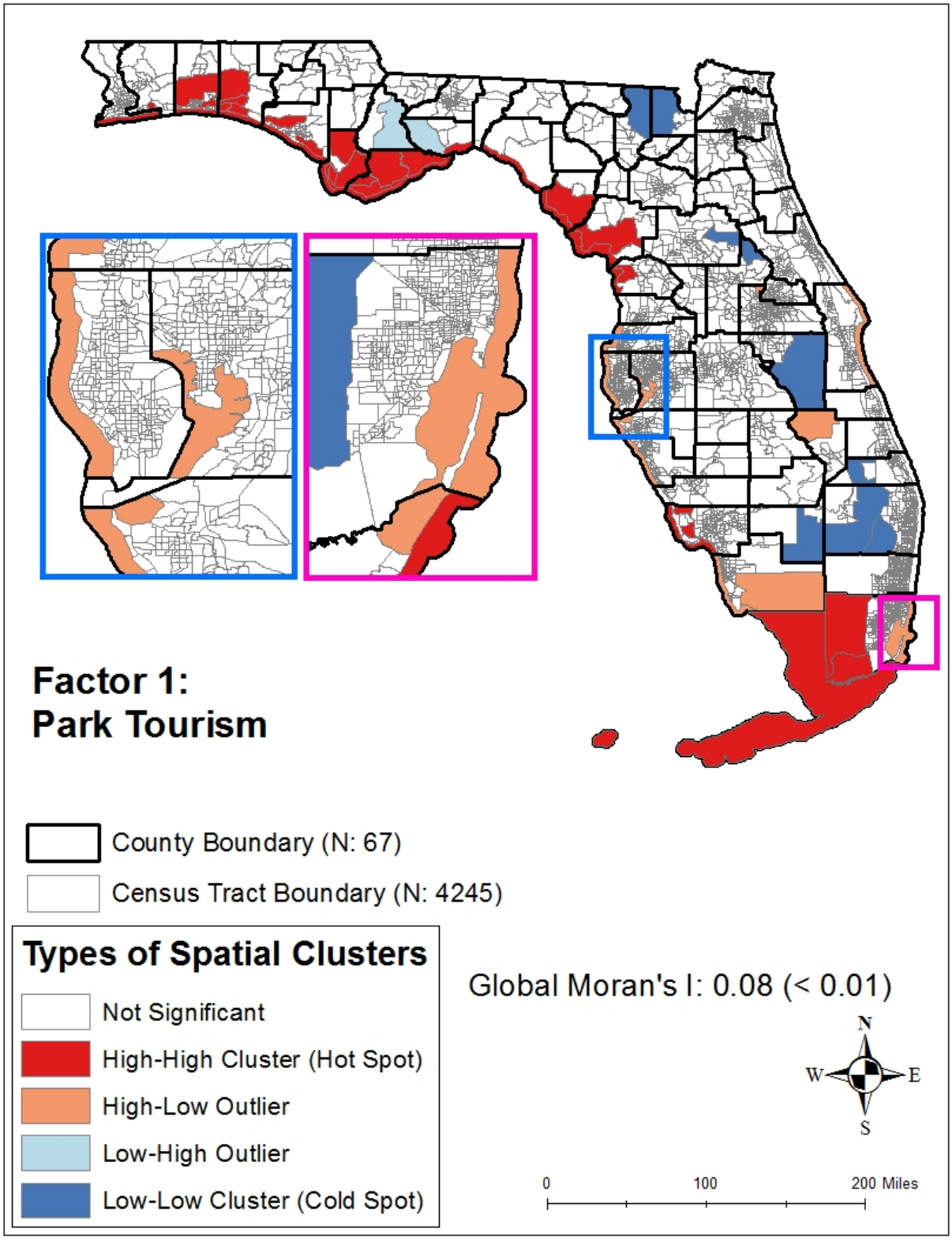


Figure 2.28 Spatial clustering of factor 1: Park Tourism (census tract)

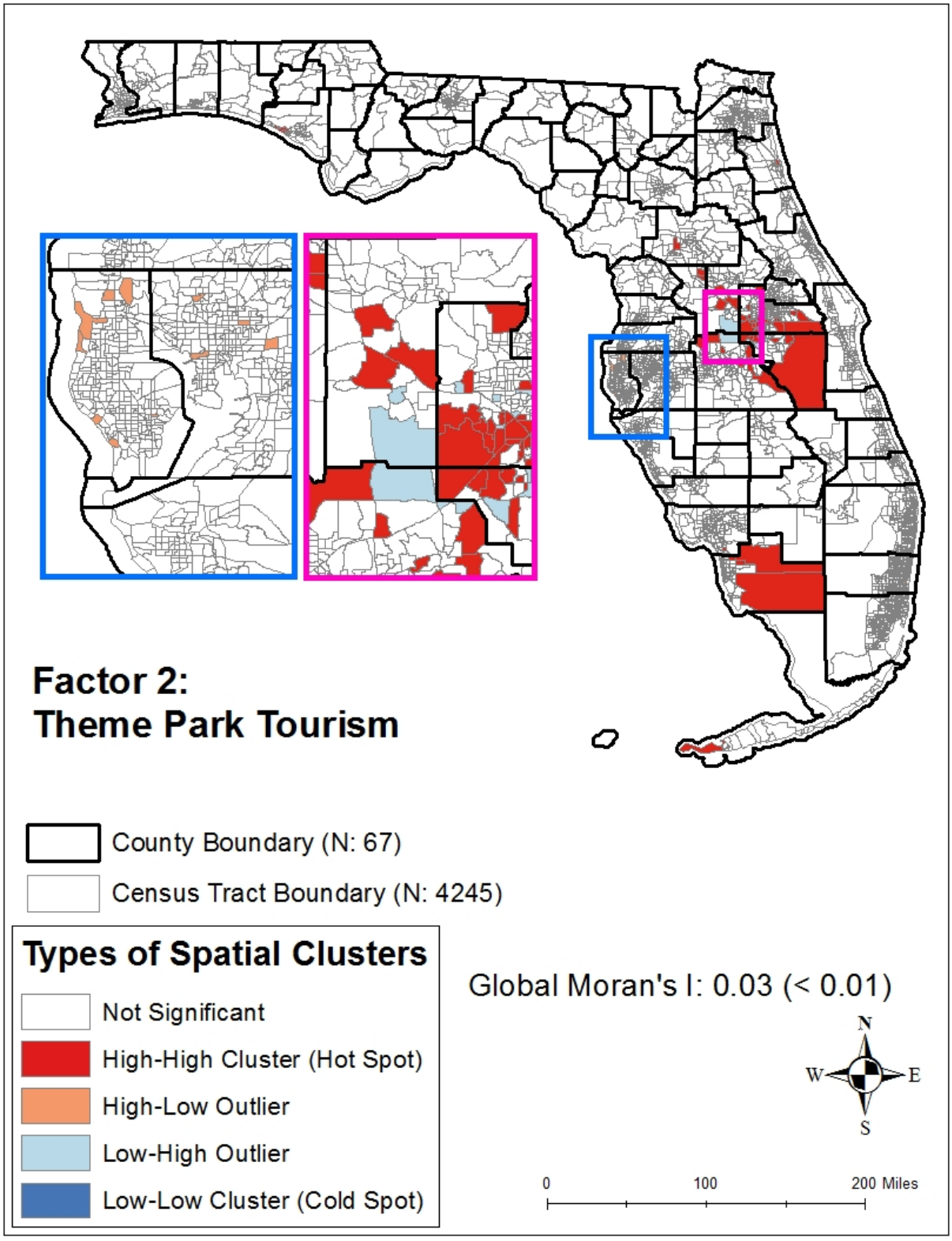


Figure 2.29 Spatial clustering of factor 2: Theme Park Tourism (census tract)

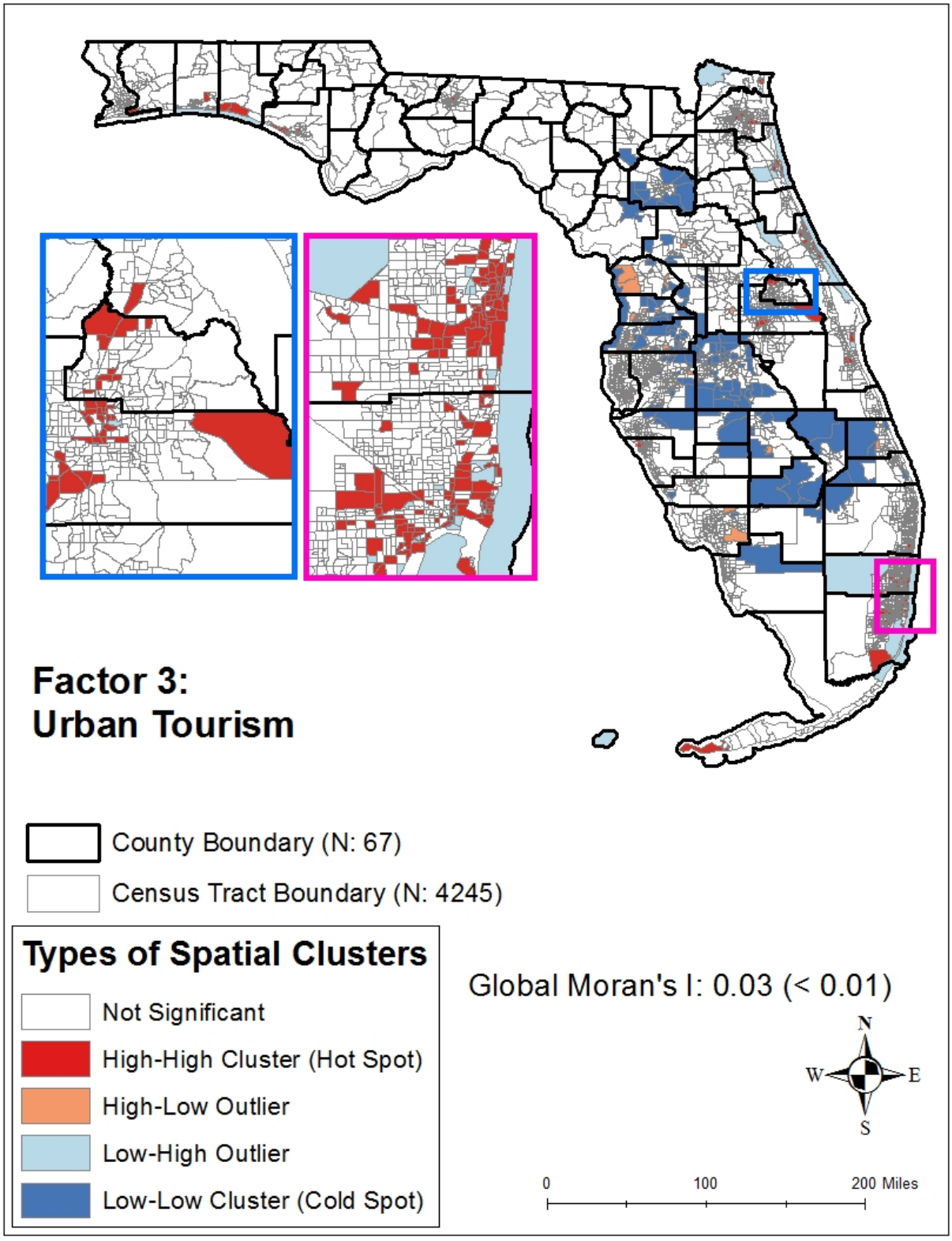


Figure 2.30 Spatial clustering of factor 3: Urban Tourism (census tract)

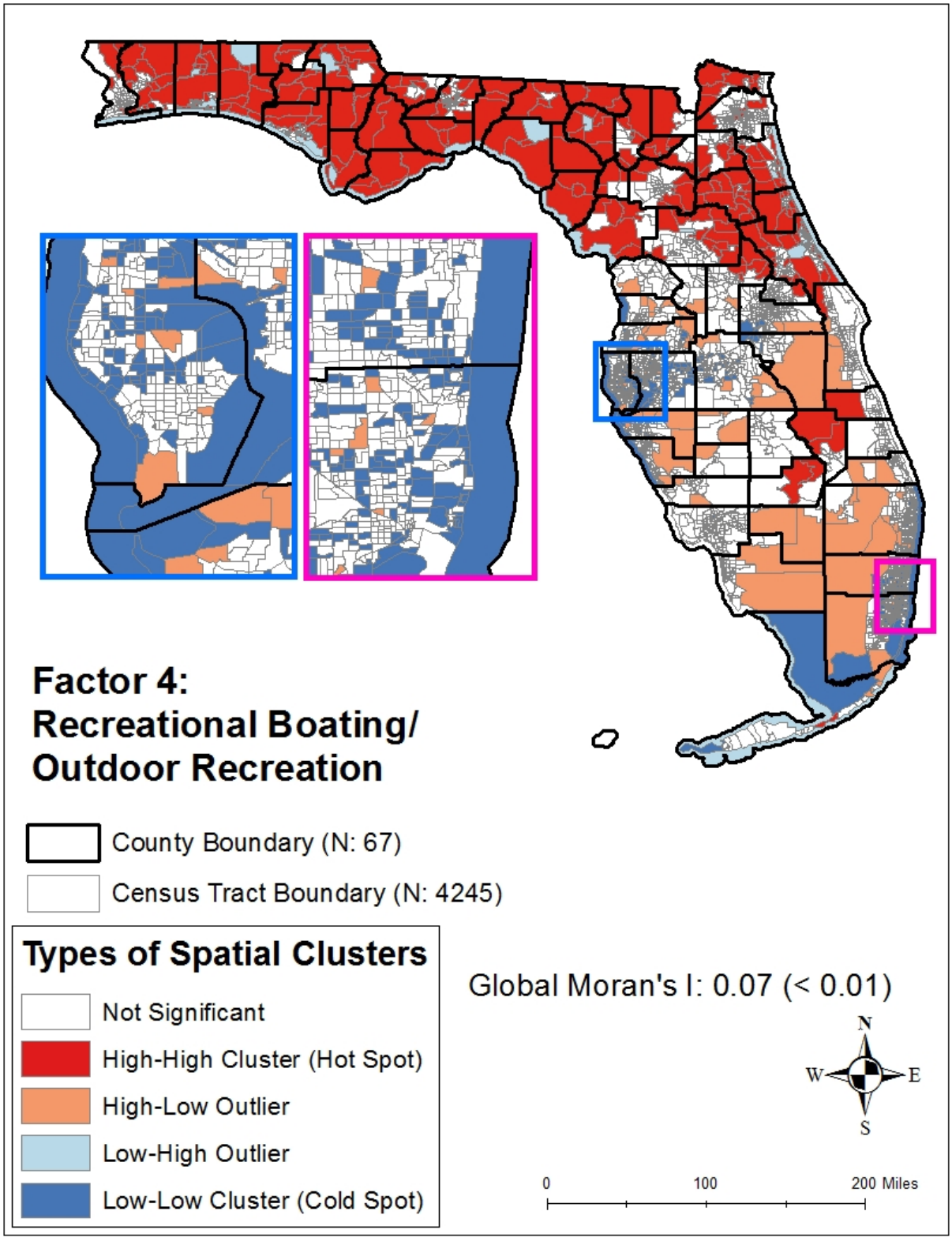


Figure 2.31 Spatial clustering of factor 4: Recreational Boating/Outdoor Recreation (census tract)

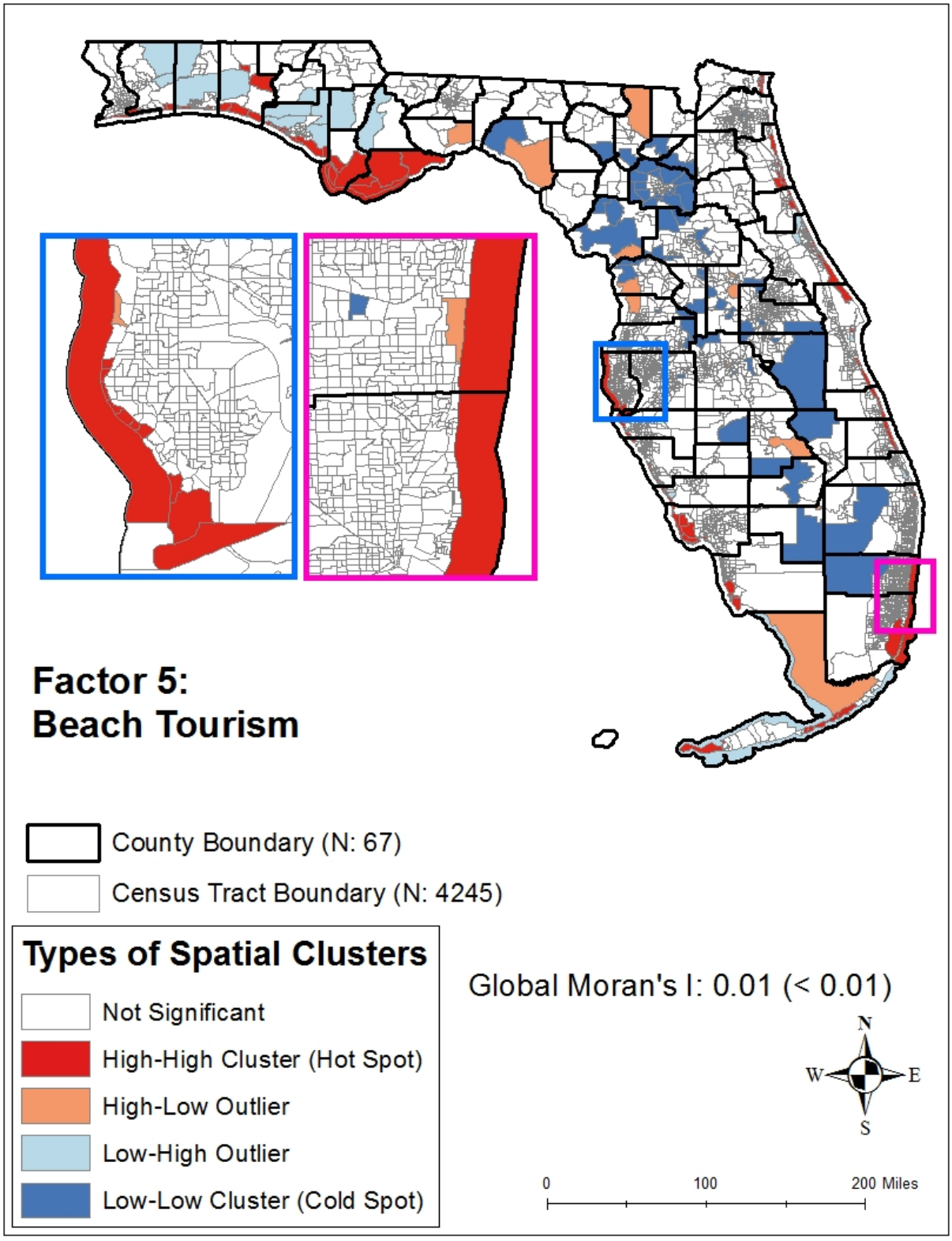


Figure 2.32 Spatial clustering of factor 5: Beach Tourism (census tract)

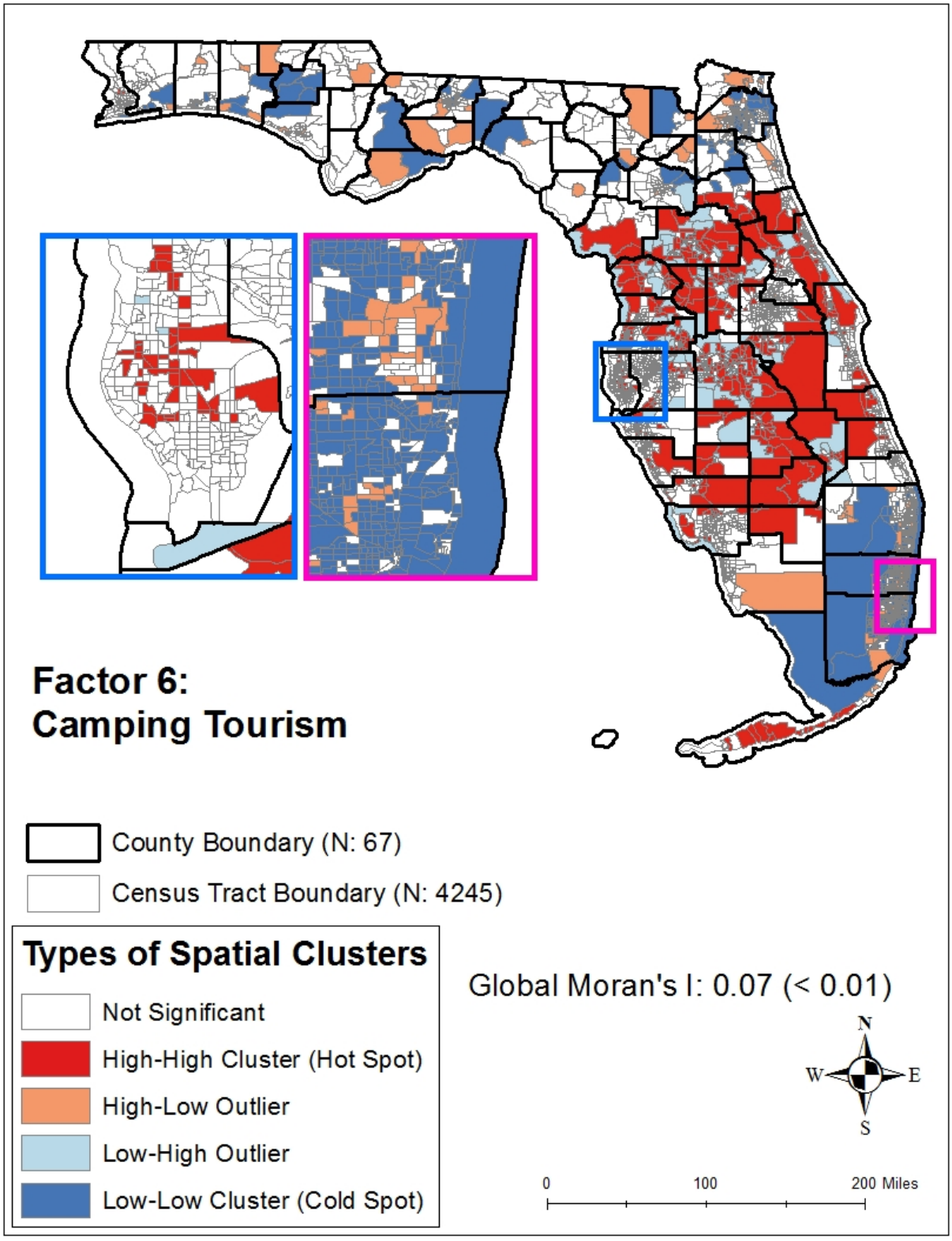


Figure 2.33 Spatial clustering of factor 6: Camping Tourism (census tract)

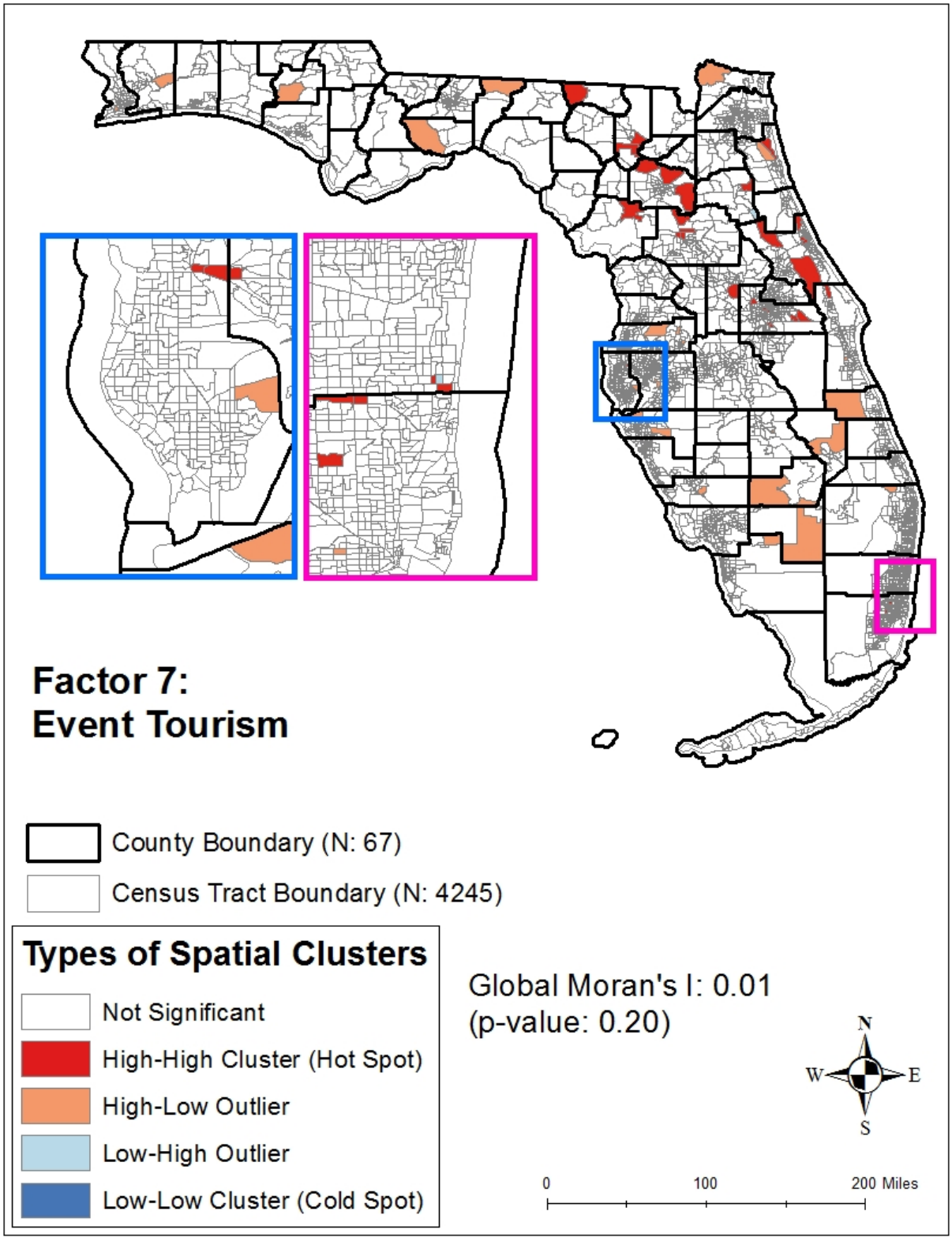


Figure 2.34 Spatial clustering of factor 7: Event Tourism (census tract)

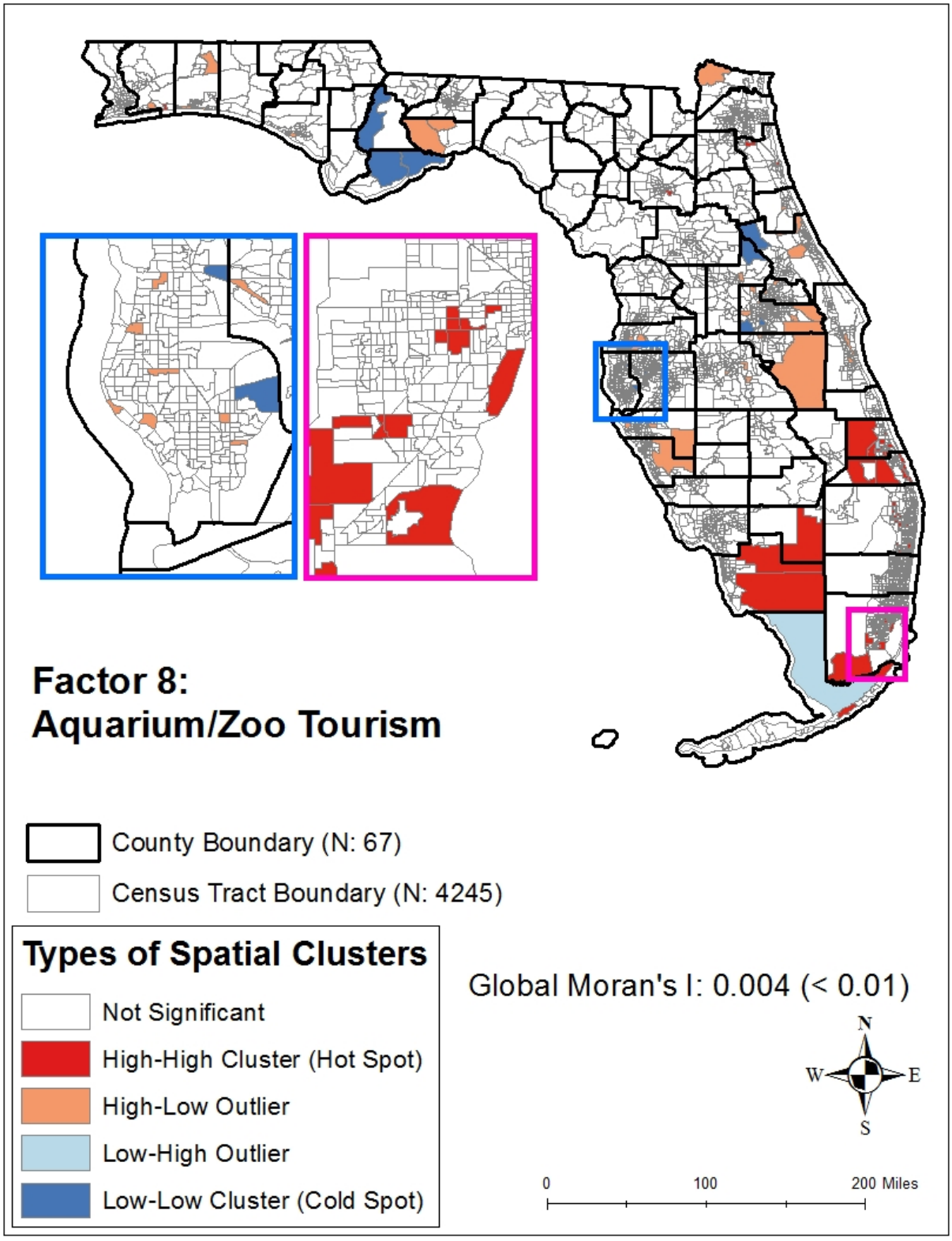


Figure 2.35 Spatial clustering of factor 8: Aquarium/Zoo Tourism (census tract)

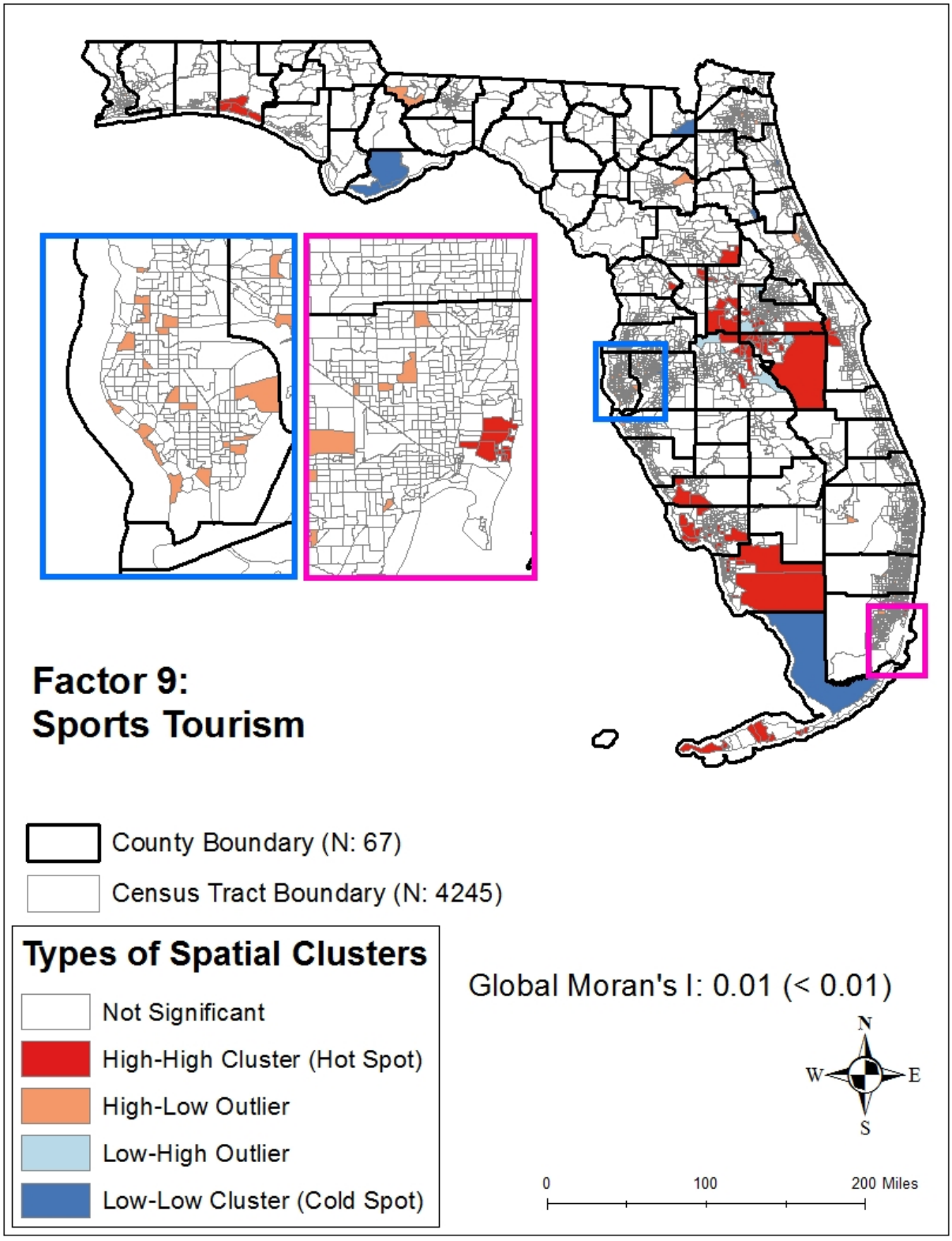


Figure 2.36 Spatial clustering of factor 9: Sports Tourism (census tract)

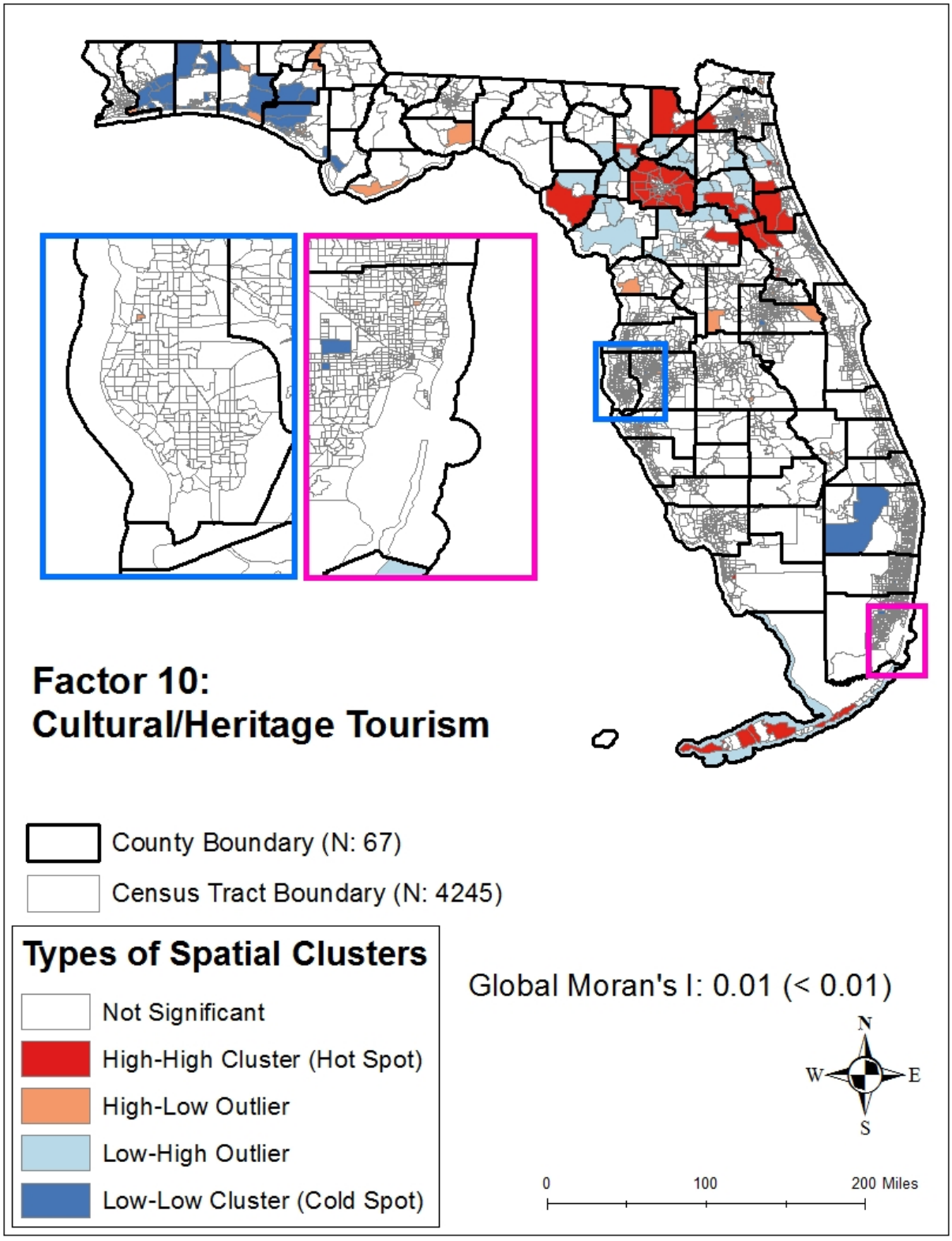


Figure 2.37 Spatial clustering of factor 10: Cultural/Heritage Tourism (census tract)

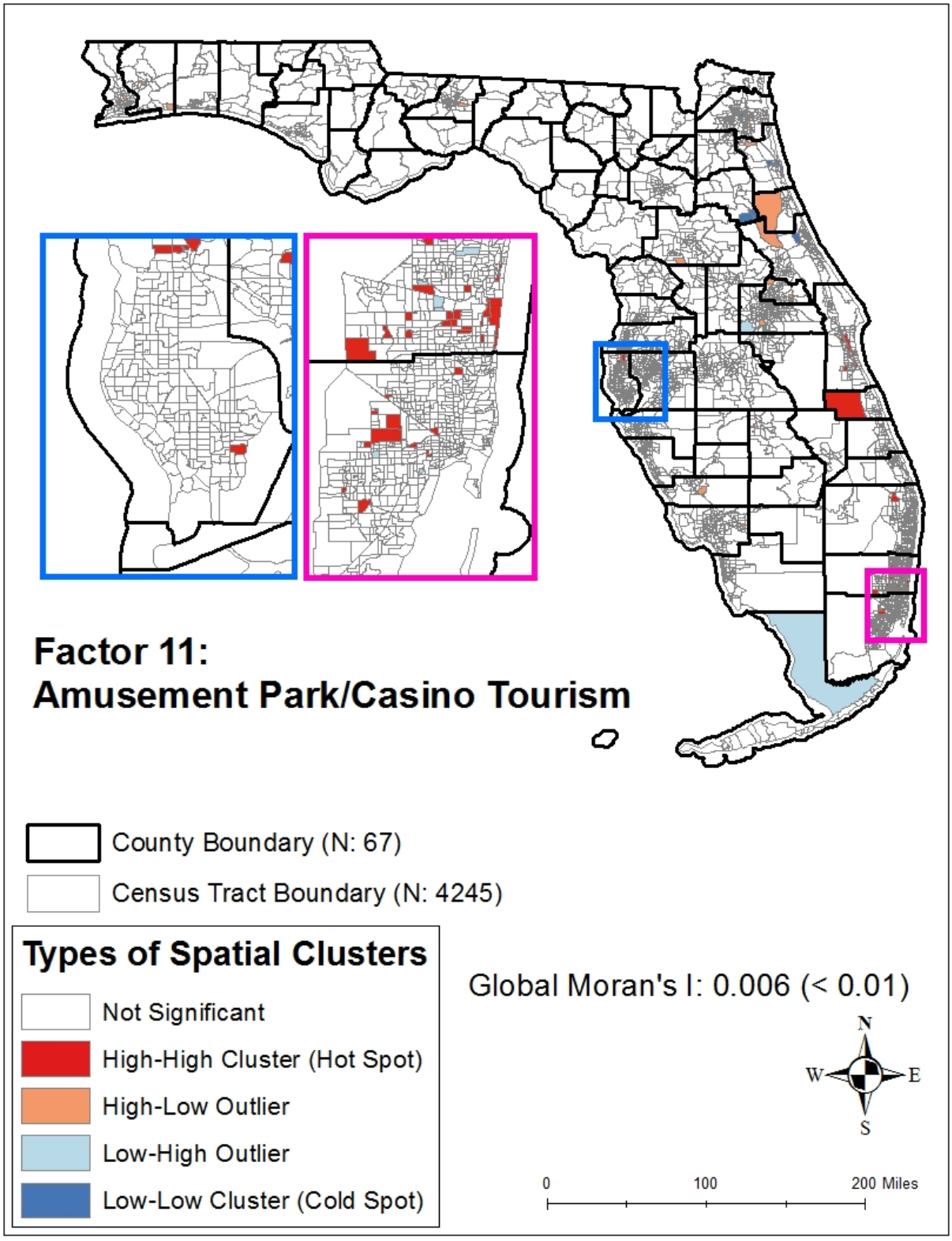


Figure 2.38 Spatial clustering of factor 11: Amusement Park/Casino Tourism (census tract)

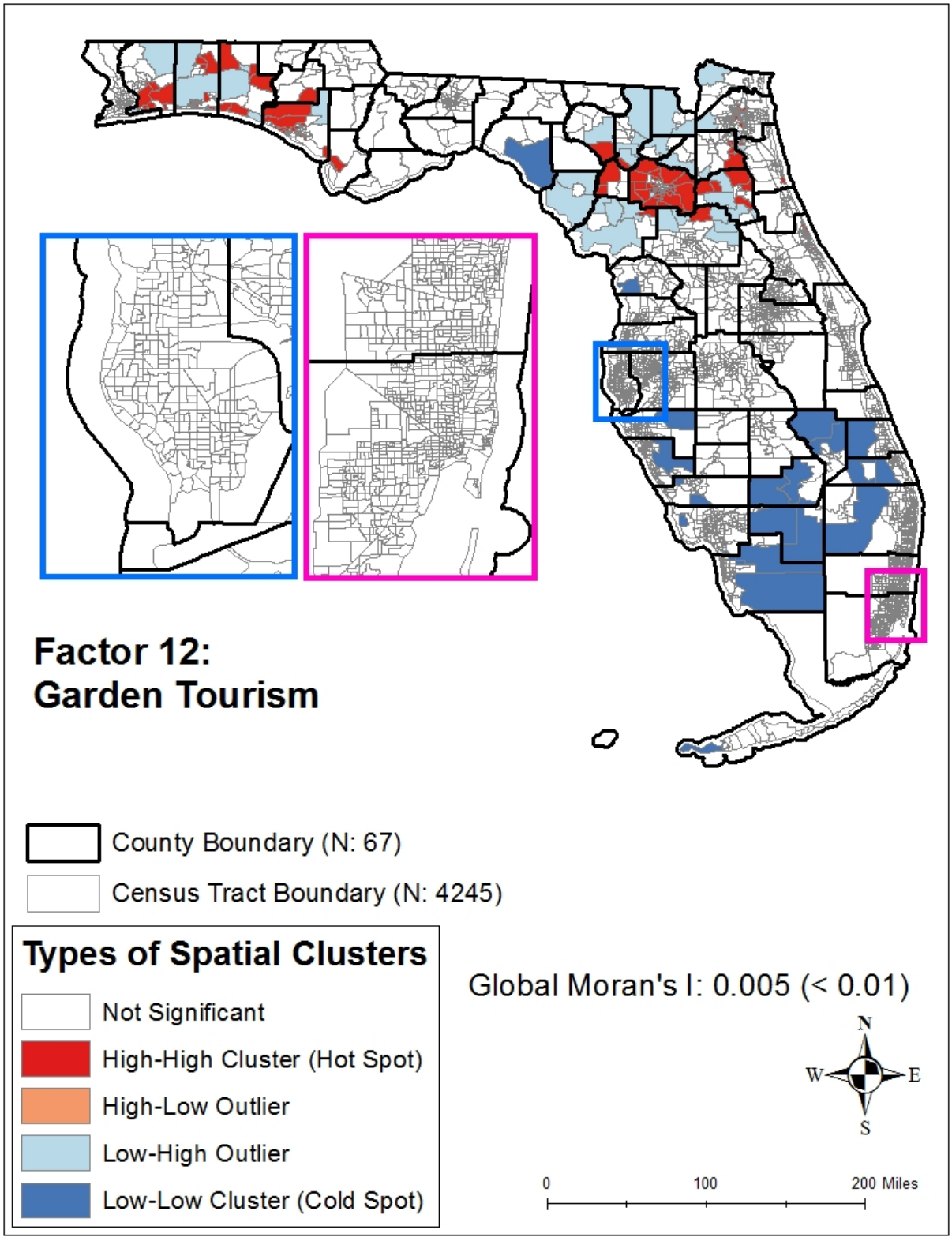


Figure 2.39 Spatial clustering of factor 12: Garden Tourism (census tract)

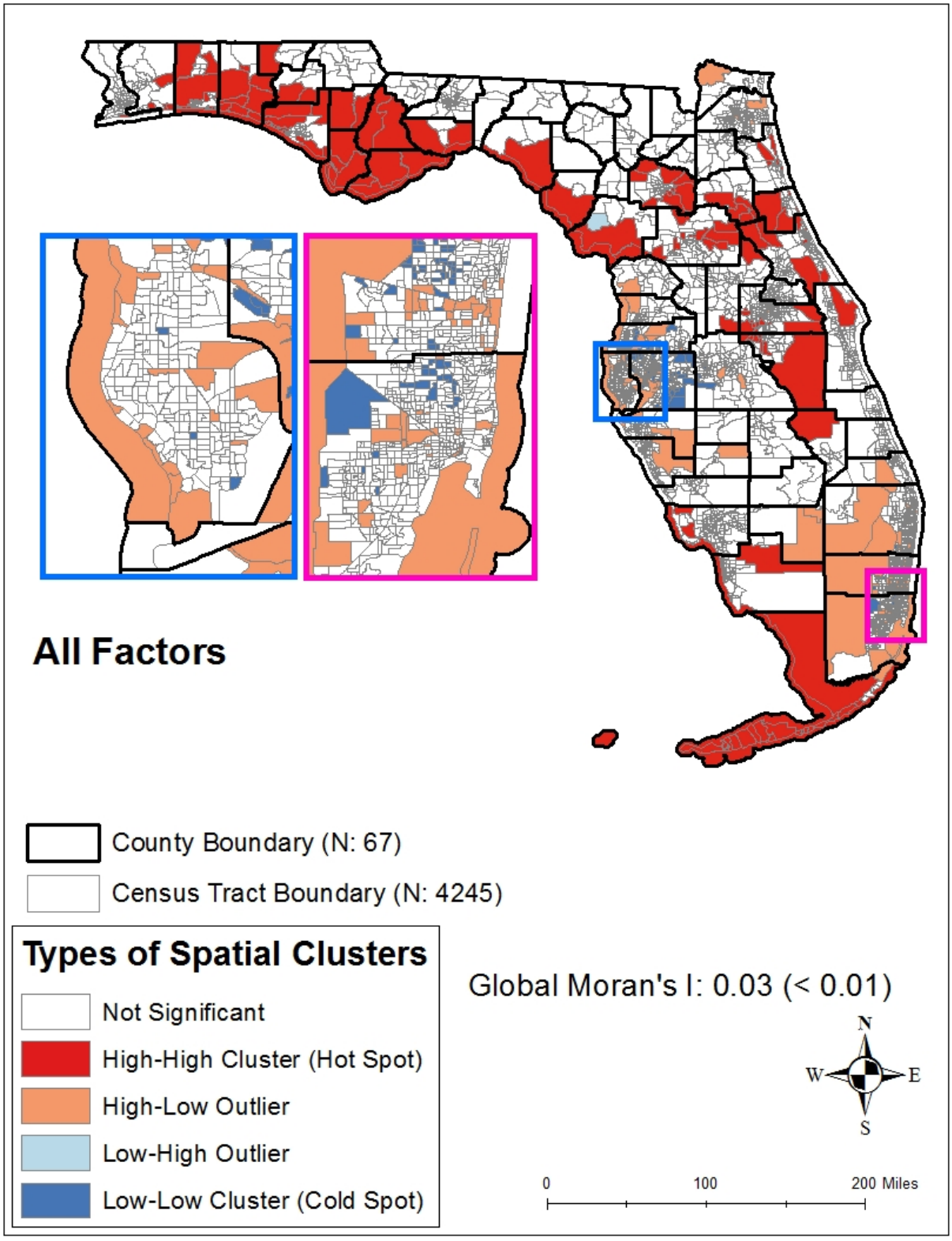


Figure 2.40 Spatial clustering of all combined factors (census tract)

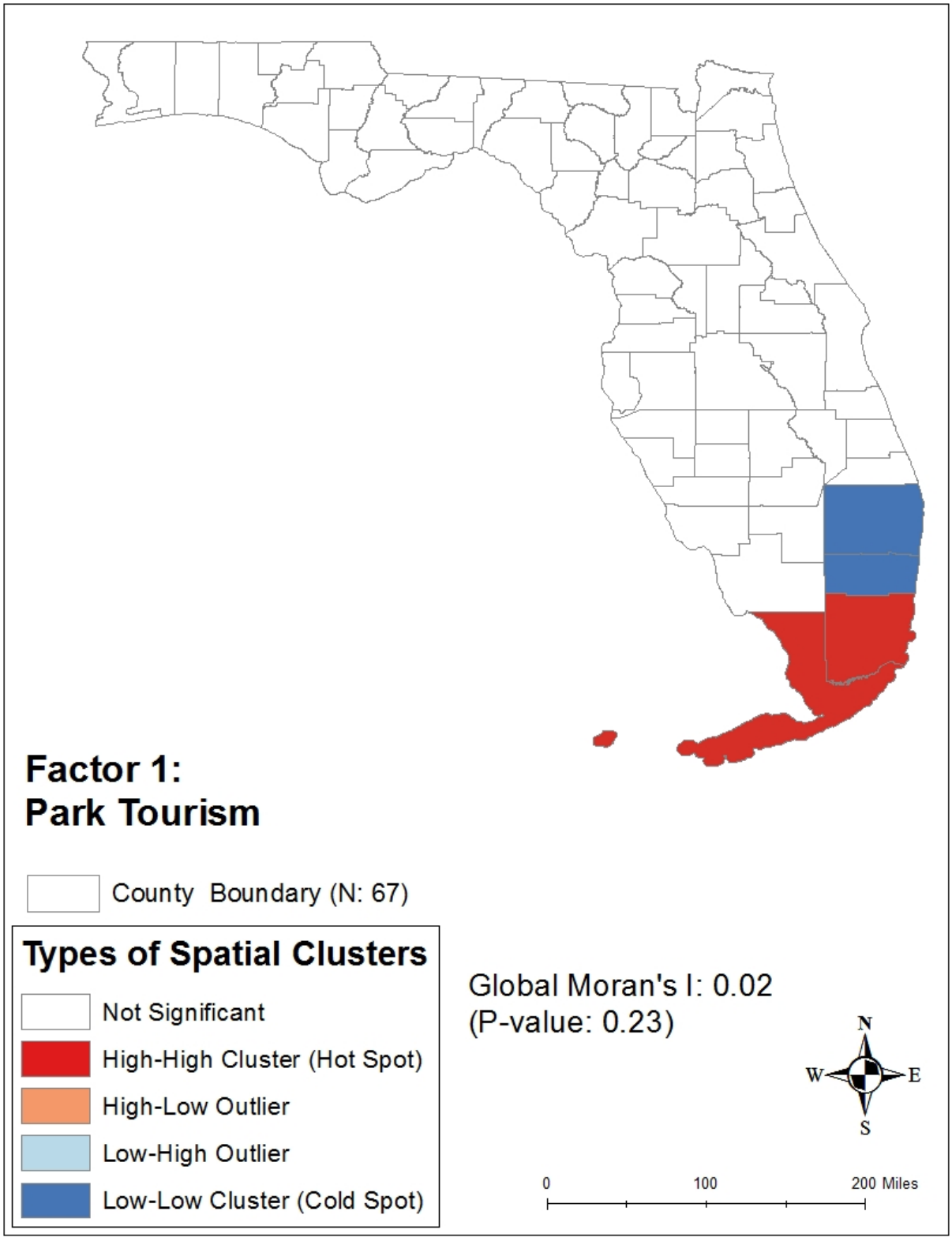


Figure 2.41 Spatial clustering of factor 1: Park Tourism (county)

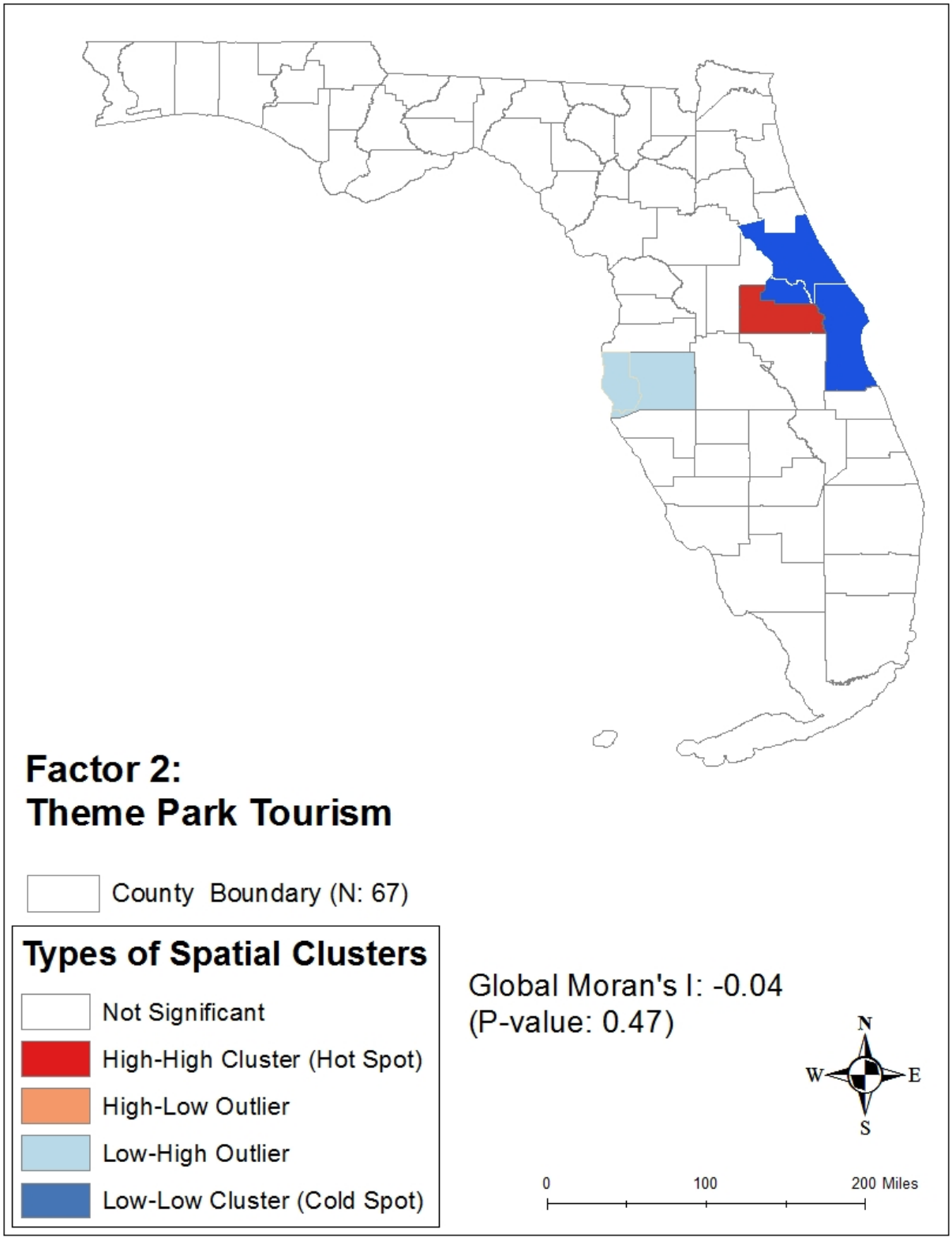


Figure 2.42 Spatial clustering of factor 2: Theme Park Tourism (county)

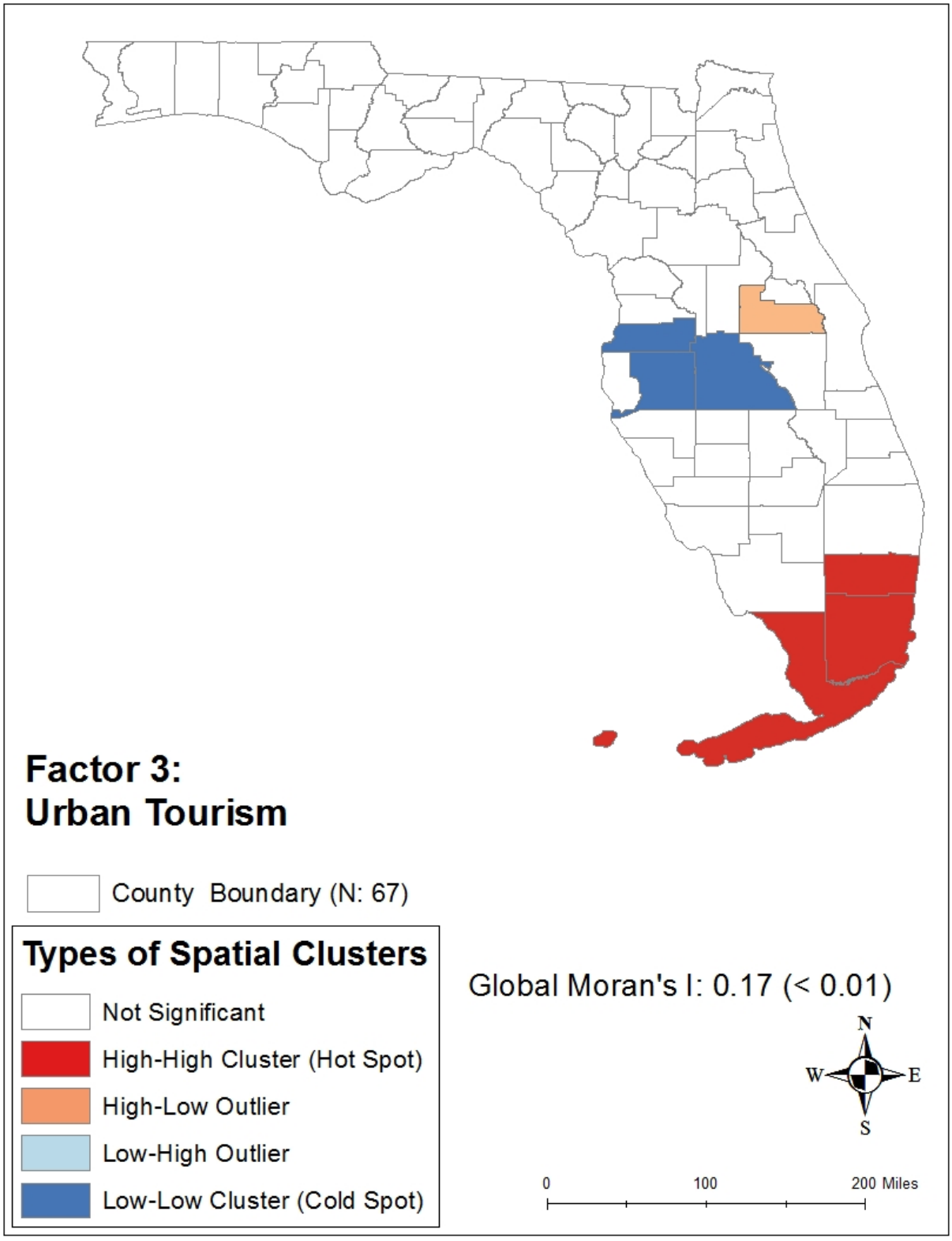


Figure 2.43 Spatial clustering of factor 3: Urban Tourism (county)

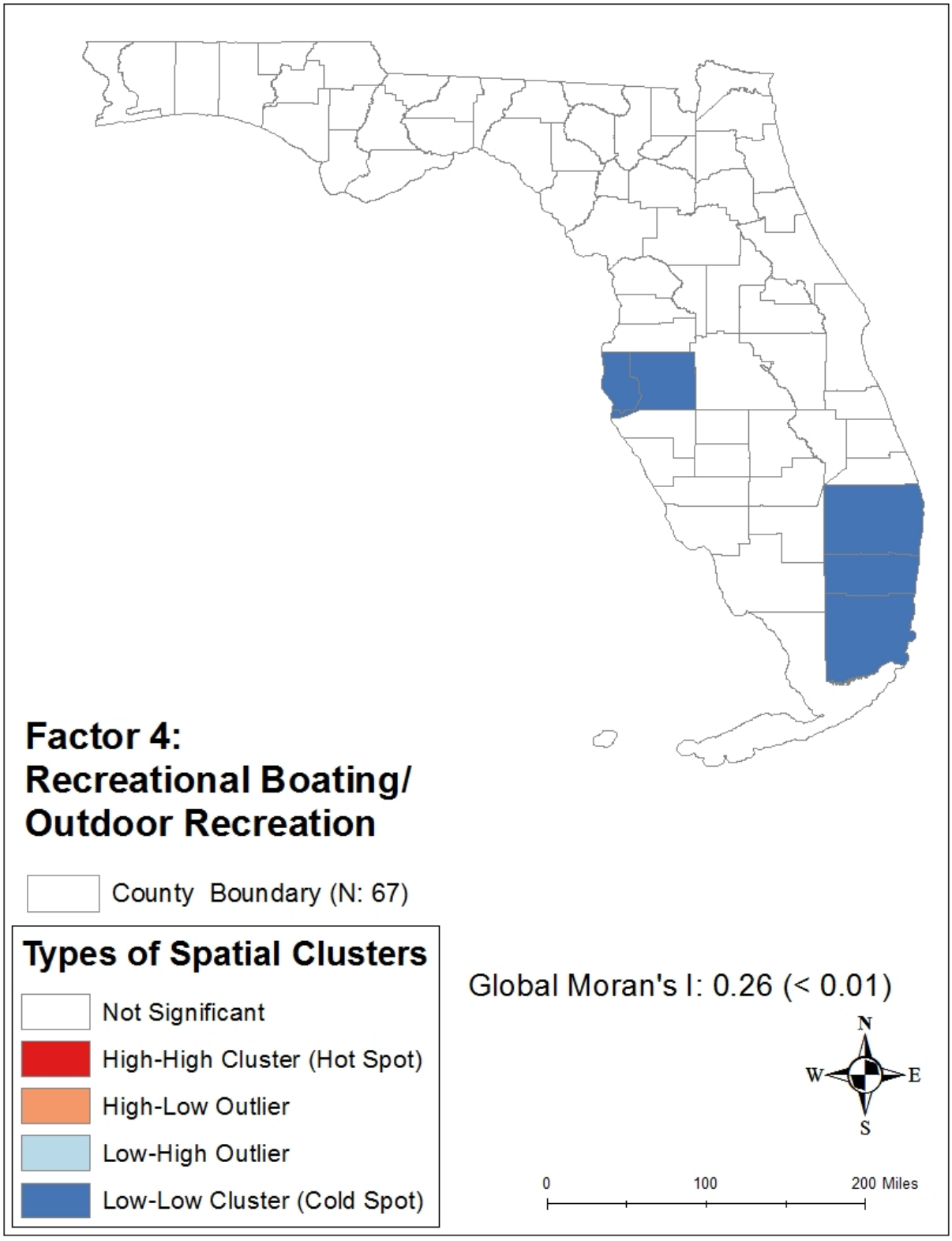


Figure 2.44 Spatial clustering of factor 4: Recreational Boating/Outdoor Recreation (county)

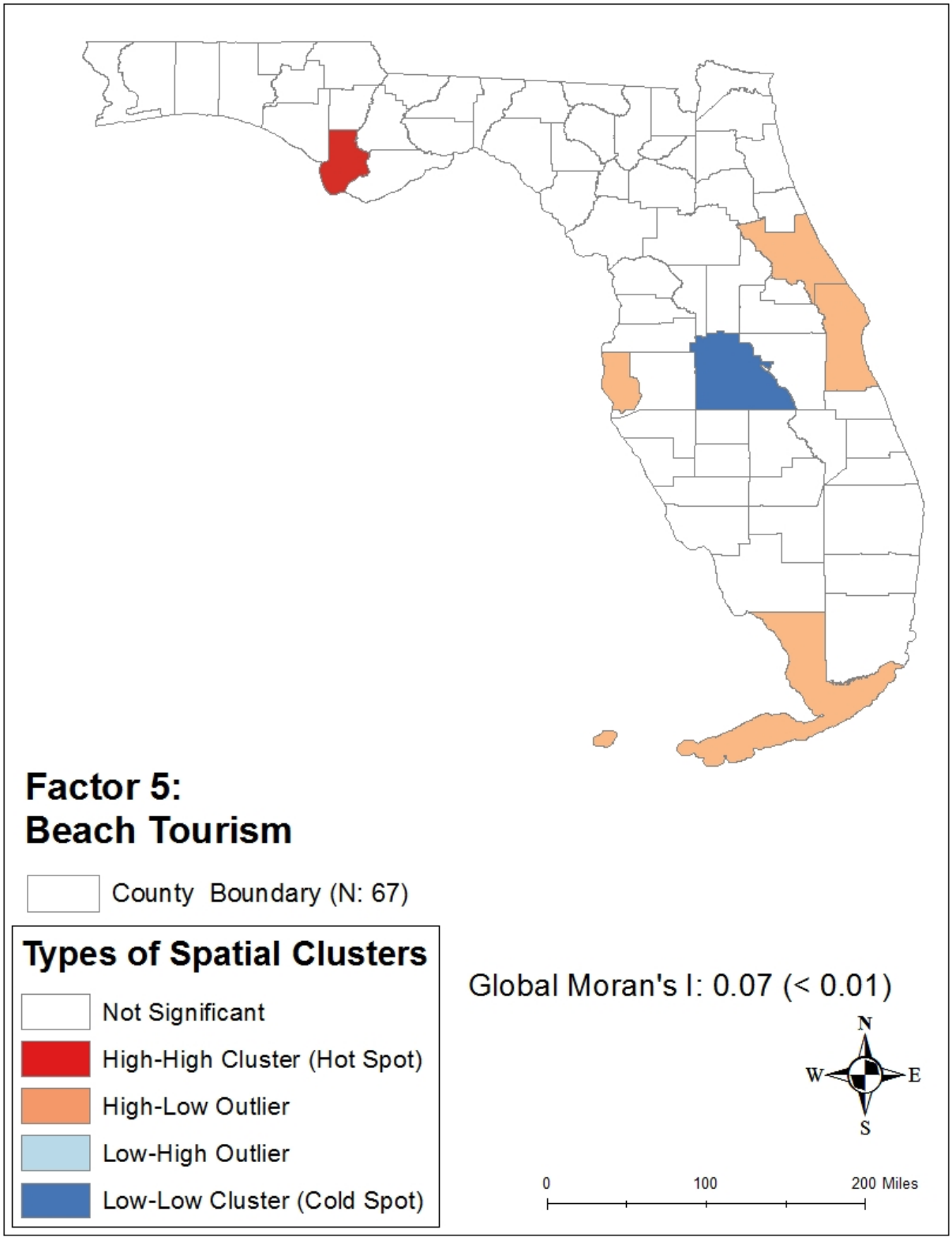


Figure 2.45 Spatial clustering of factor 5: Beach Tourism (county)

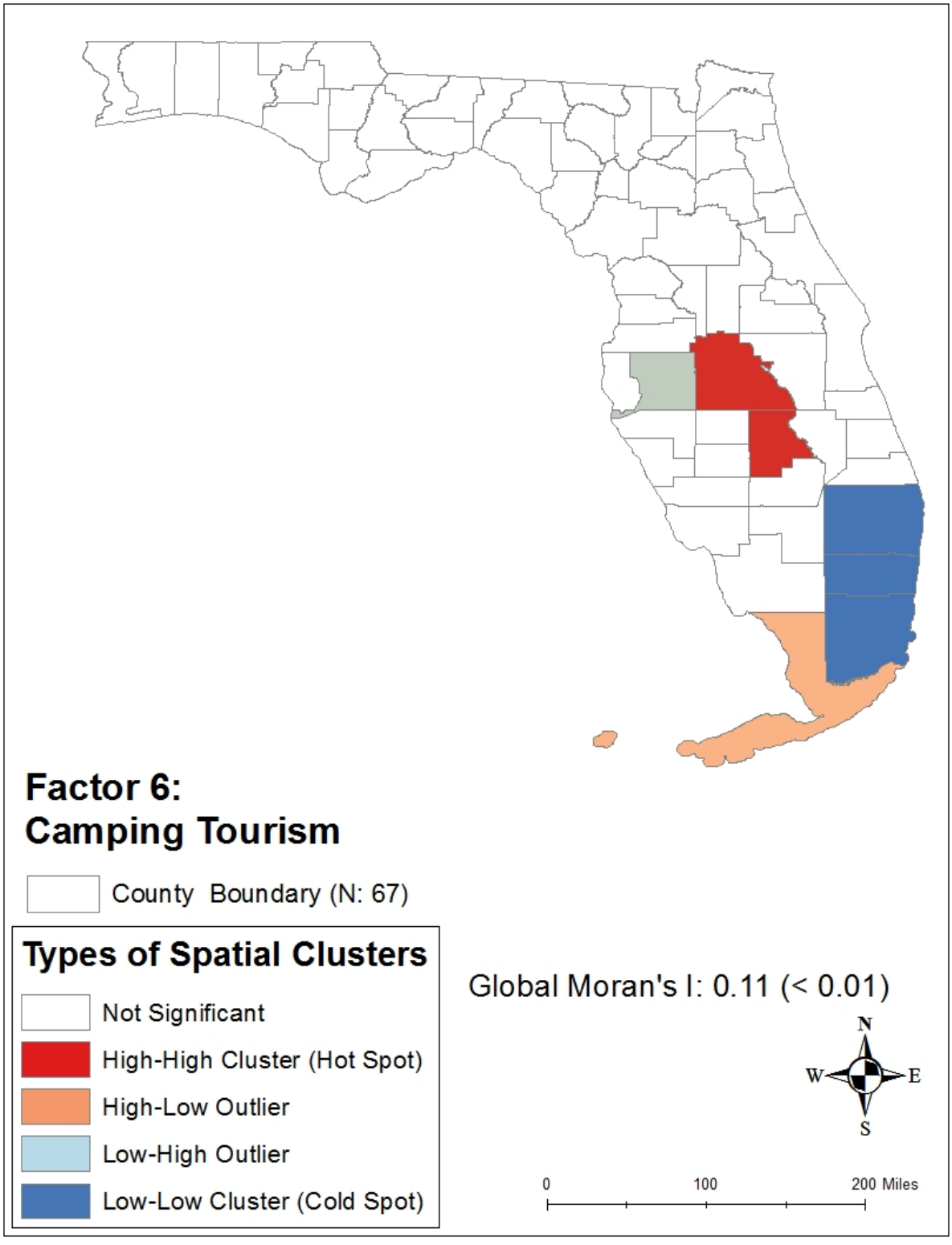


Figure 2.46 Spatial clustering of factor 6: Camping Tourism (county)

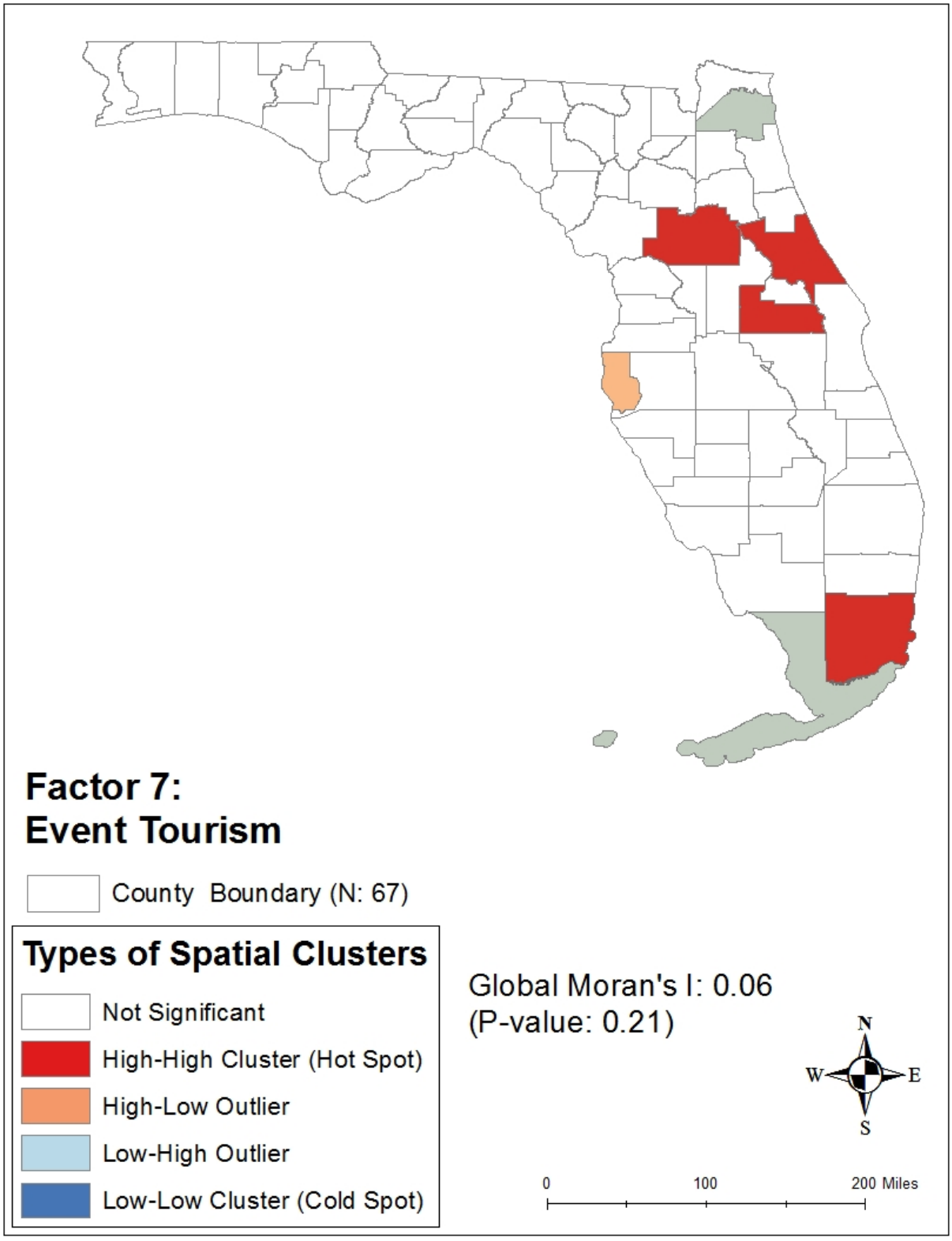


Figure 2.47 Spatial clustering of factor 7: Event Tourism (county)

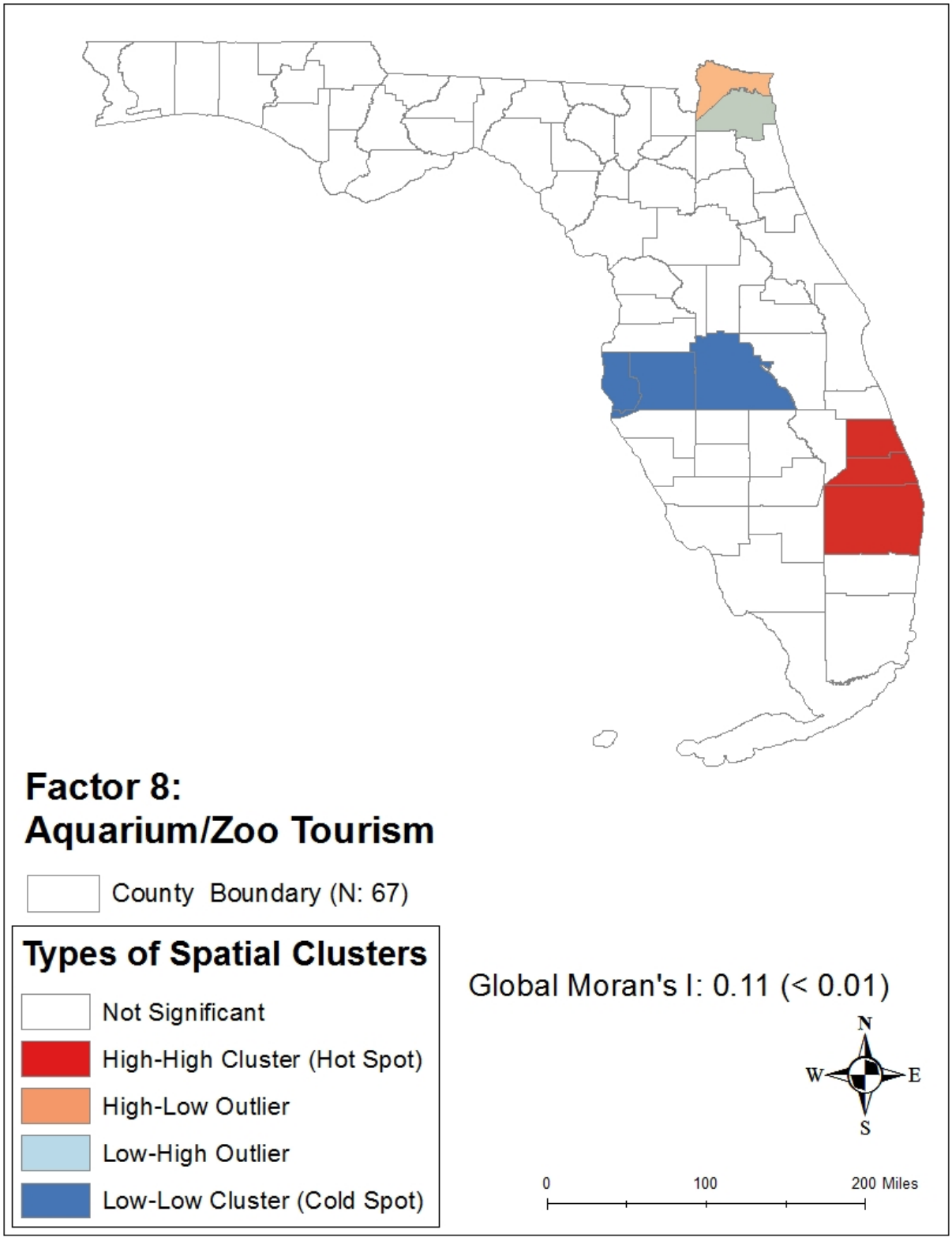


Figure 2.48 Spatial clustering of factor 8: Aquarium/Zoo Tourism (county)

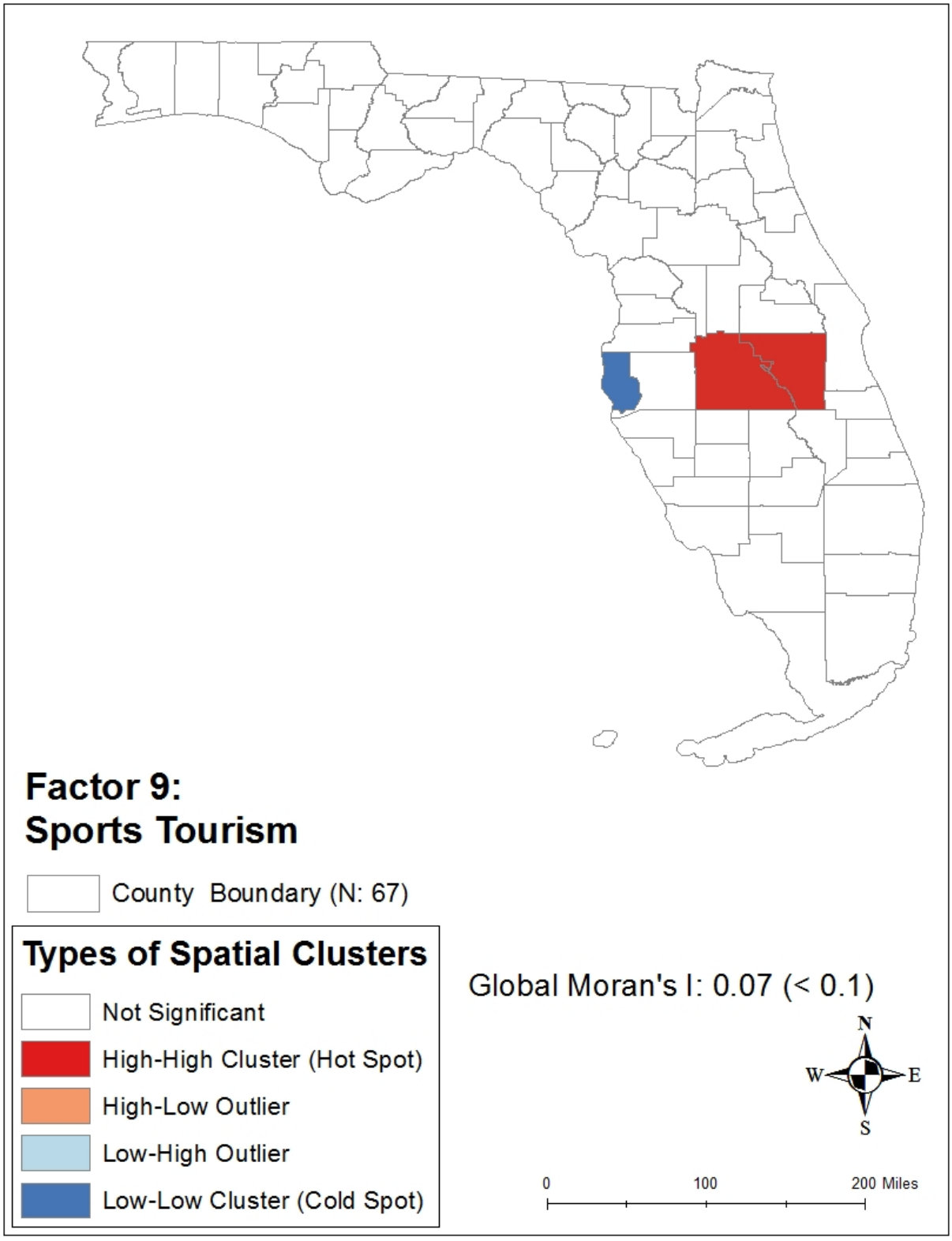


Figure 2.49 Spatial clustering of factor 9: Sports Tourism (county)

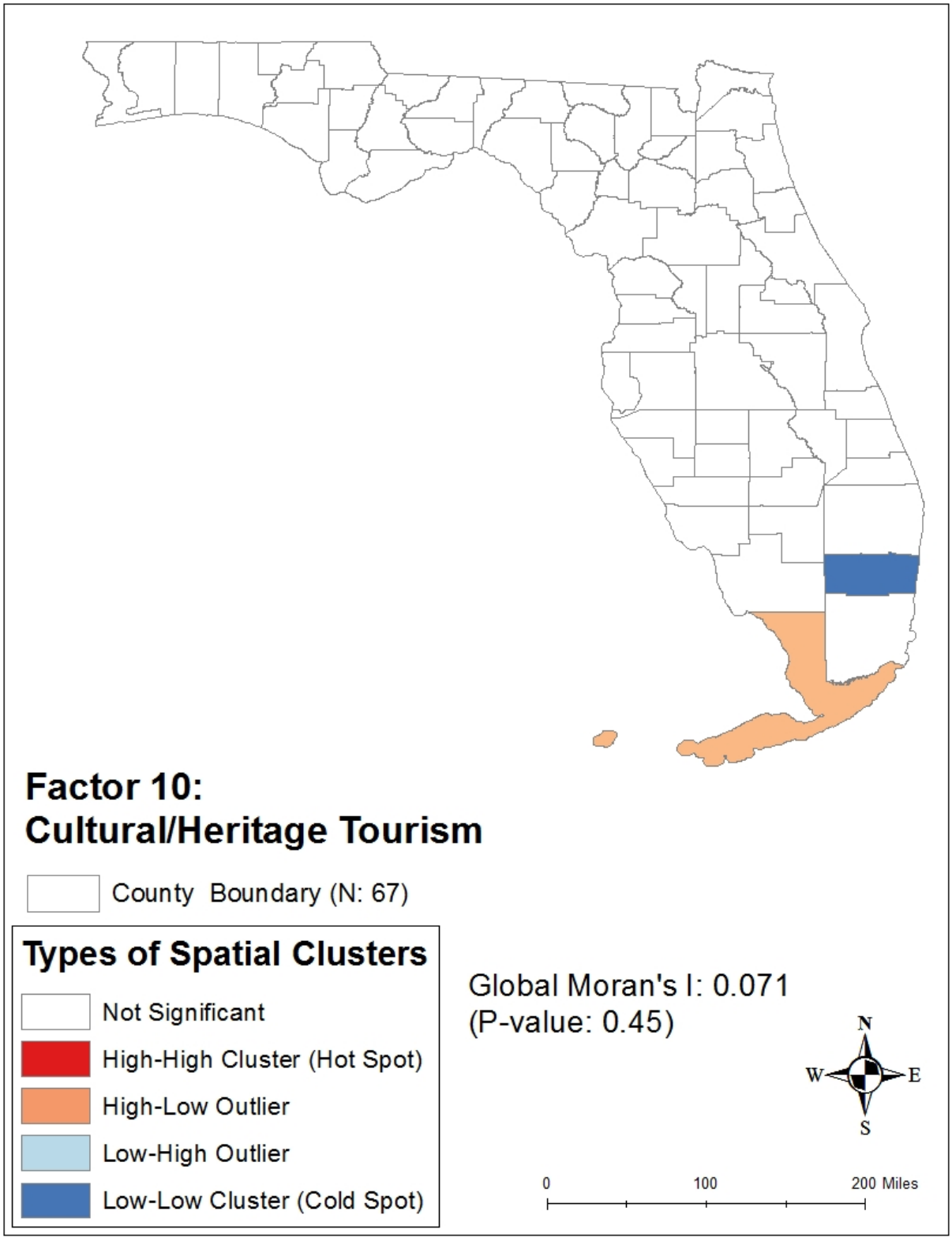


Figure 2.50 Spatial clustering of factor 10: Event Tourism (county)

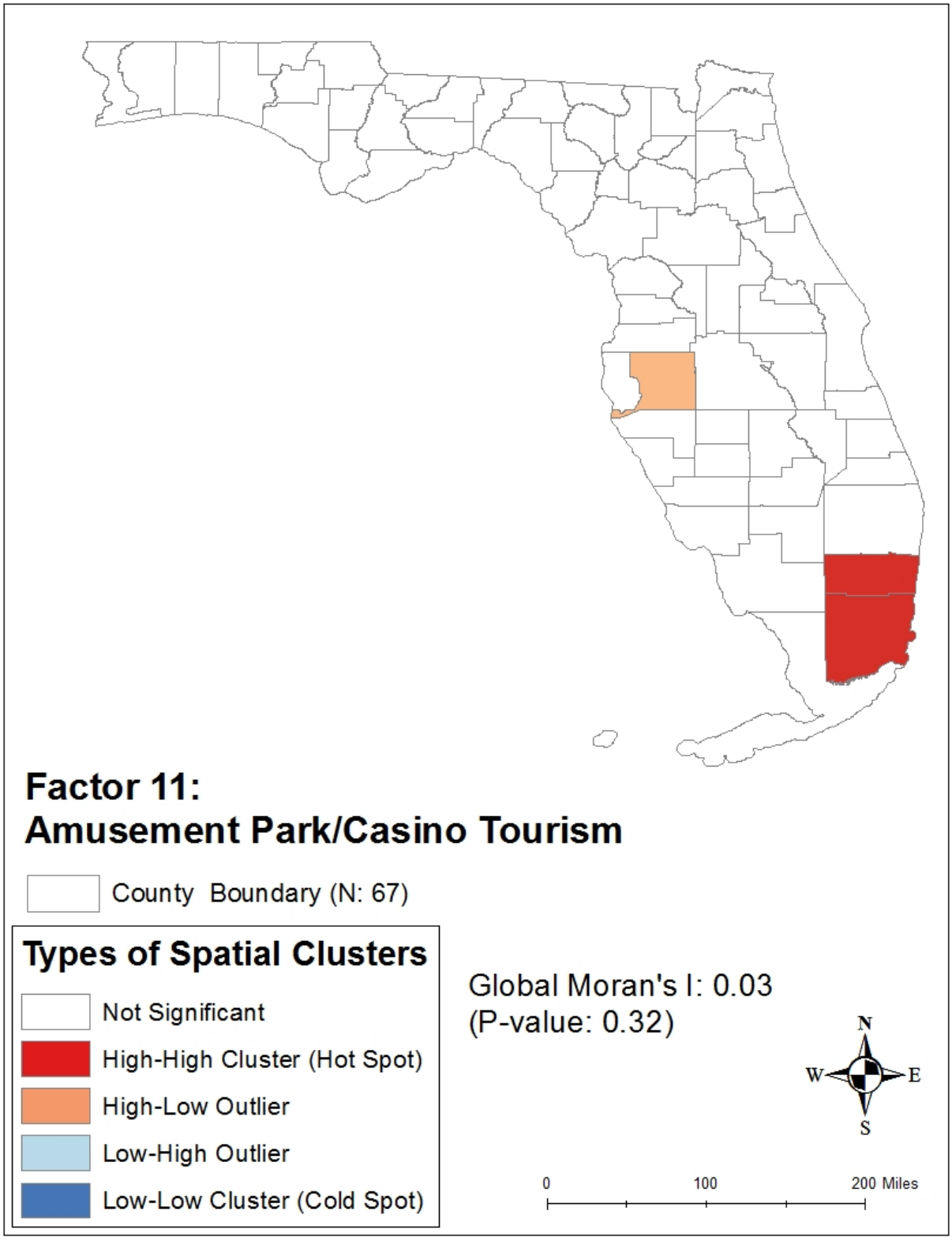


Figure 2.51 Spatial clustering of factor 11: Amusement Park/Casino Tourism (county)

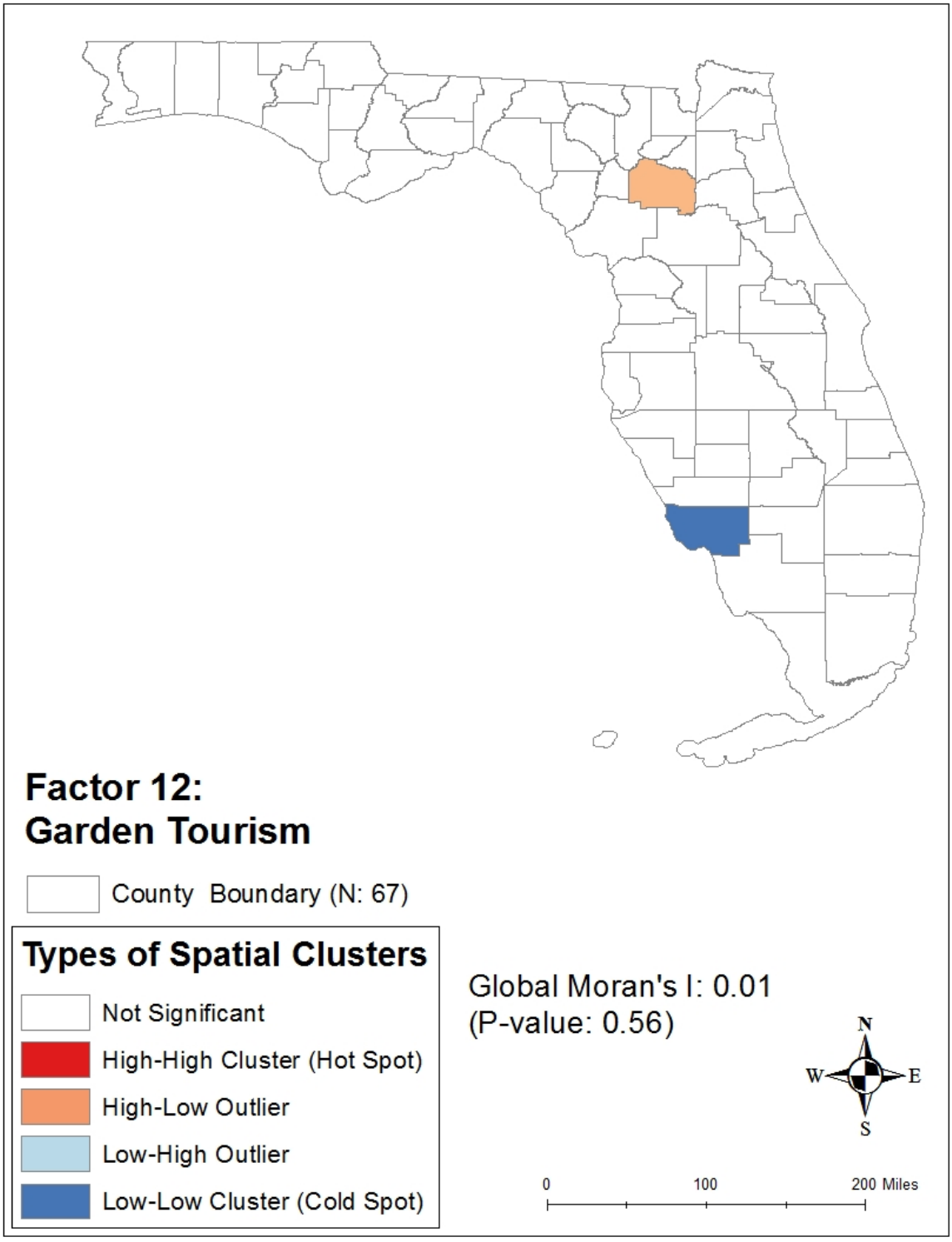


Figure 2.52 Spatial clustering of factor 12: Garden Tourism (county)

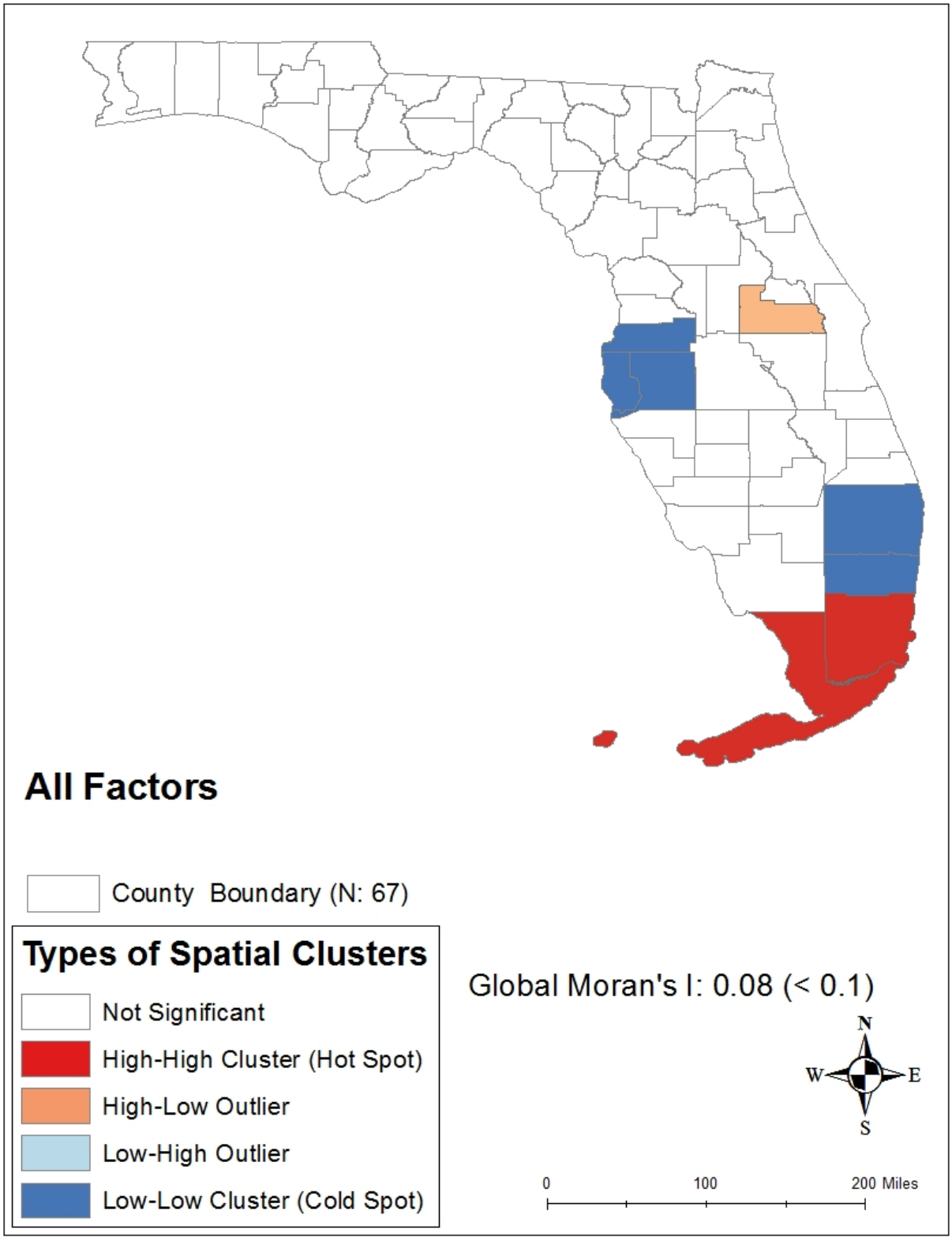


Figure 2.53 Spatial clustering of all combined factors (county)

2.7.3 Mapping highway accessibility/availability

Accessibility is generally referred to as the ease with which activities or services can be reached or obtained (Johnson et al., 2000; Morris et al., 1979; Nicholls, 2001). Highway accessibility/availability is a key driver for tourist movements. In this task, highway accessibility/availability was operationally measured as the total miles of primary and secondary roads. This access measure is justified based on the container approach explained by Kim and Nicholls (2016).

The container approach defines access according to the presence of resources/facilities within a geographic unit, such as a census tract, zip code, or local neighborhood unit (e.g., total miles of road networks within the geographic unit) (Zhang et al., 2011). A container index Z_i^C is calculated as follows:

$$Z_i^C = \sum_j S_J, \in I \quad (2.2)$$

where Z_i^C is a container index for residential neighborhood i, and the number or aggregate size, S_J, is summed for those facilities located within the boundaries I of i. The higher the number or total miles of highway systems within each unit of analysis, the higher the level of access to highway systems enjoyed by residents of that unit. The container approach has been employed extensively in political science and urban planning research due to its simplicity (Talen & Anselin, 1998; Lindsey et al., 2001).

- Primary roads are generally divided, limited-access highways within the interstate highway system or under State management, and are distinguished by the presence of interchanges. These highways are accessible by ramps and may include some toll highways.
- Secondary roads are main arteries, usually in the U.S. highway, state highway, and/or county highway system. These roads have one or more lanes of traffic in each direction, may or may not be divided, and usually have at-grade intersections with many other roads and driveways.

Figures 2.54-2.55 show the distribution of total miles of highway systems for each census tract and county.

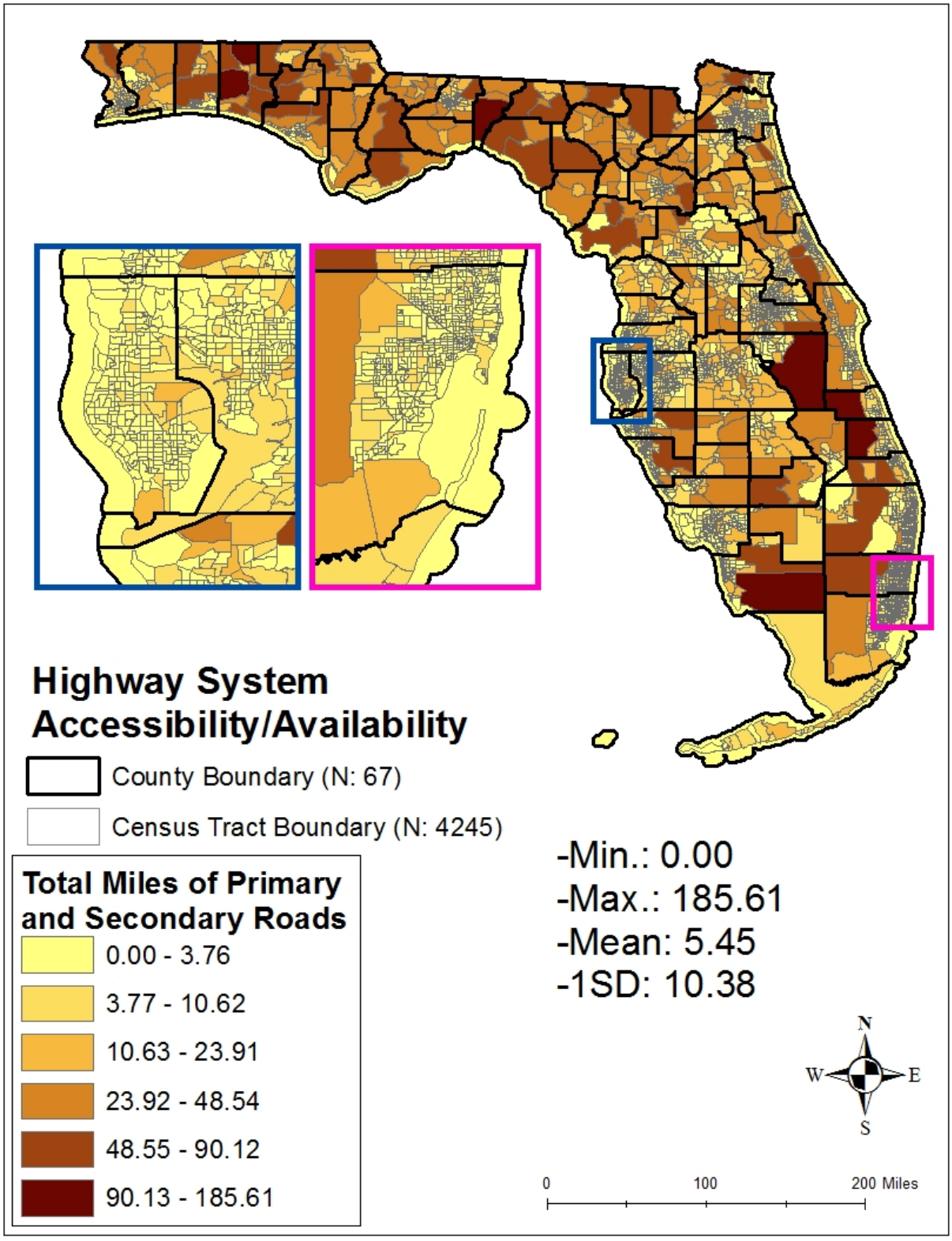


Figure 2.54 Spatial distribution of highway accessibility/availability (census tract)

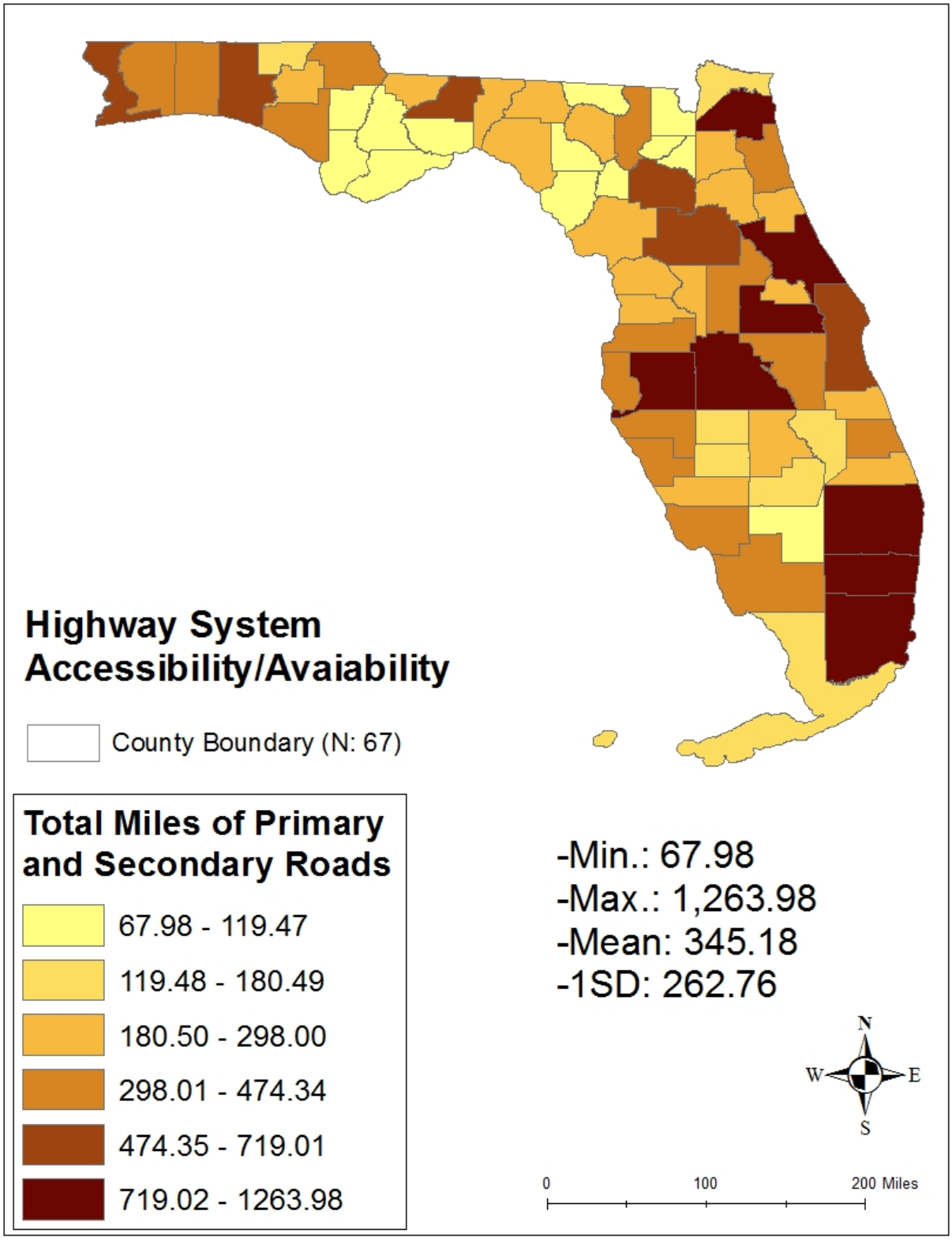


Figure 2.55 Spatial distribution of highway accessibility/availability (county)

2.8 Step 6: Explore the spatial relationships between tourism resources and highway accessibility/availability

We employed geographically weighted regression (GWR) technique to explore spatial variations in the relationships between the spatial patterns of Florida's tourism resources and highway accessibility/availability. GWR is a spatial regression technique, which can model local spatial heterogeneity between variables. GWR assumes that relationships between variables may differ from location to location (Fotheringham et al., 2002). In other words, GWR generates a set of local regression coefficients for each county (or census tract) in Florida. Figures 2.56-2.57 show the distribution of local correlation between tourism resources and highway accessibility/availability and their spatial clustering patterns. These results will enable FDOT to better establish strategies for improving highway accessibility to potential and existing tourism regions by identifying the spatial mismatches between tourism resources and the Florida highway system.

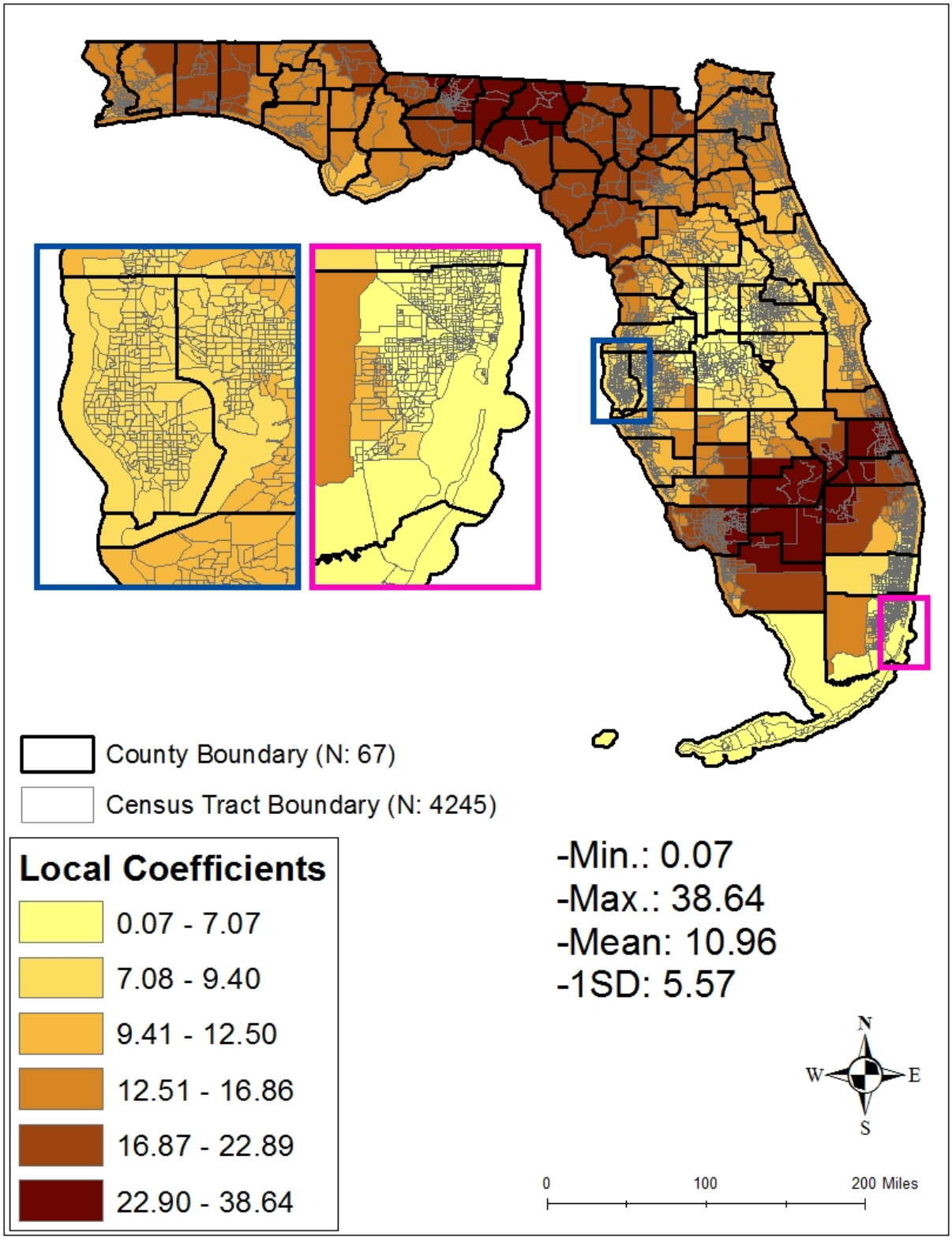


Figure 2.56 Local correlation between tourism resources and highway systems (census tract)

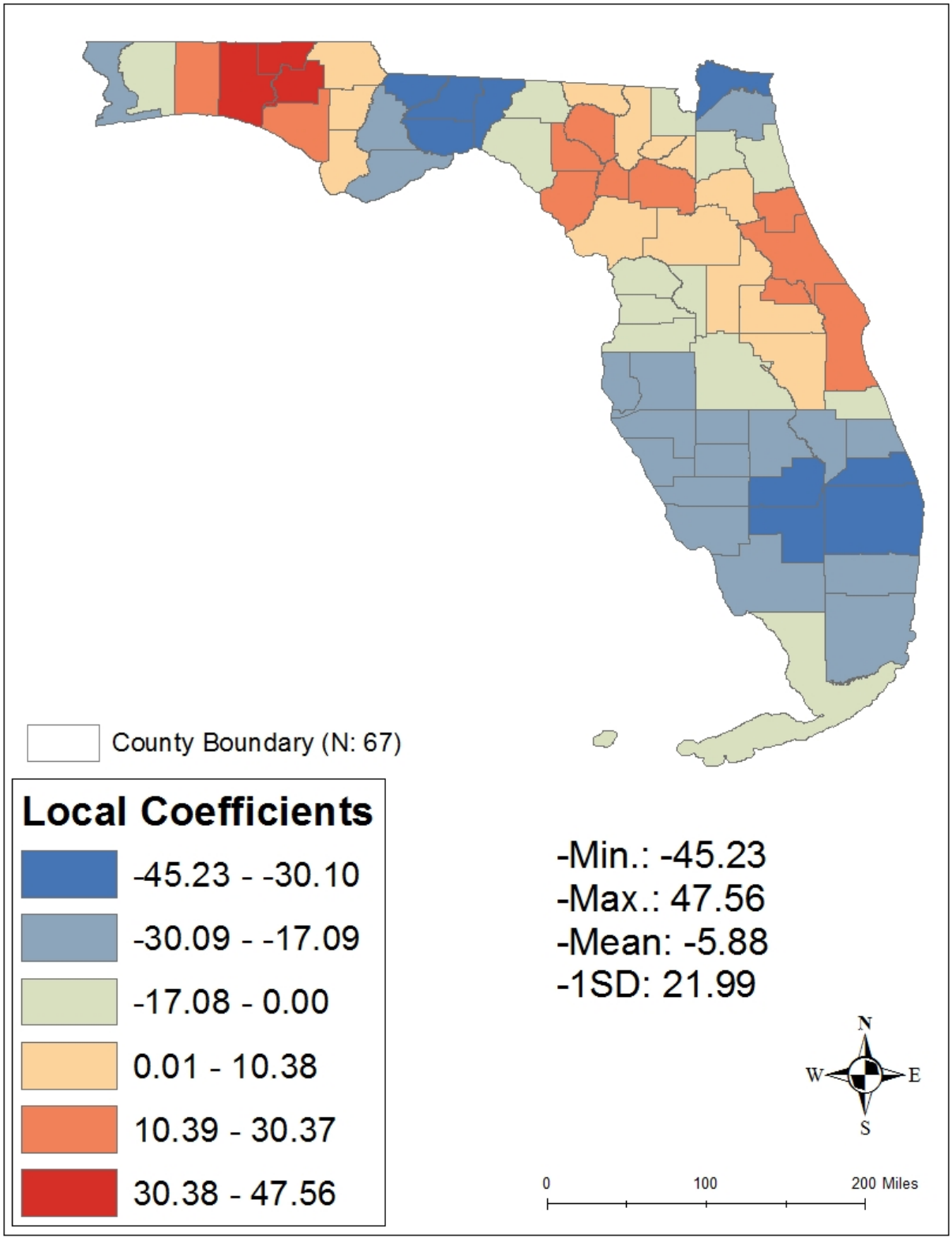


Figure 2.57 Local correlation between tourism resources and highway systems (county)

2.9 Conclusion

Task 2 describes procedures for identifying tourism resources on the basis of county- and census tract-level resource patterns. Twelve tourism resource factors were identified: “Park Tourism,” “Theme Park Tourism,” “Urban Tourism,” “Recreational Boating/Outdoor Recreation,” “Beach Tourism,” “Camping Tourism,” “Event Tourism,” “Aquarium/Zoo Tourism,” “Sports Tourism,” “Cultural/Heritage Tourism,” “Amusement Park/Casino Tourism,” and “Garden Tourism.” Each tourism resource factor had different spatial pattern. Local hot spots and cold spots for each tourism resource factor were also identified. Lastly, important local variations in the relationships between tourism resources and highway accessibility/availability were explored and visualized.

Chapter 3. Tourism supply components in Florida

Task description

Identify tourism supply components in Florida using social media data to cross-validate and enhance data obtained in Task 2:

- Generate a database of social media data generated by travelers to Florida and Floridian travelers visiting Florida attractions (TripAdvisor reviews)
- Employ a geotagging engine to generate latitude and longitude data for tourist trip origins and destinations;
- Use the database to estimate the distribution of travel distances at a census tract level for all locations;
- Combine the tourism supply components (Task 3 product) to the GIS layers obtained in Task 2.

Deliverable: A written report of the outcomes and the GIS layers.

3.1 Introduction

Understanding the spatial patterns of tourism resources in Florida is important for measuring the potential for attracting tourists (Formica & Uysal, 2006) and planning the provision of effective transportation systems. This task concentrates on data collection and analysis that are necessary for generation of the tourist trip matrix and further forecasting the spatial pattern of visitors' travel to the Florida area. First, social media data allow estimation of the relative number of tourists travelling to and between Florida attraction points. Second, the county and census tract levels tourism supply indices created in Task 2 allow cross-validation of the generated tourist trip matrix (to be done in Task 4). Hence, Task 3 focused on social media data collection (TripAdvisor reviews of Florida hotels, attraction points, and restaurants), population of the review data with geotags reflecting tourist origins at a city-scale granularity and destinations (at a census tract granularity), collecting the auxiliary data such as locations of the major airports to estimate the travel origins for the international visitors, descriptive data analysis, combining the Task 2 and Task 3 data layers, and pilot runs of the cross-validation process.

Purpose of Task 3

Identify tourism supply components in Florida using social media data to cross-validate and enhance data obtained in Task 2.

Objectives

1. Generate a database of social media data generated by travelers to Florida and Floridian travelers visiting Florida attractions (TripAdvisor reviews)
2. Employ a geotagging engine to generate latitude and longitude data for tourist trip origins and destinations;
3. Use the database to estimate the distribution of travel distances at a census tract level for all locations;
4. Combine the tourism supply components (Task 3 product) to the GIS layers obtained in Task 2.

3.2 Methodology

The overall process of validating a series of tourism resource indices involves several steps. Figure 3.1 presents an overall flowchart for Task 3. In step 1, we collected the TripAdvisor data for validating the tourism supply index from Task 2. The TripAdvisor data include reviews, ratings, and number of hotels, tourism attractions, restaurants, and rentals. In step 2, we employed the Google geocoding engine to generate latitude and longitude data for each tourism property (i.e., hotels, restaurants, tourism attractions, and rentals) from the TripAdvisor data. In step 3, we converted location data to both census tract and county levels and constructed a tourism supply component dataset based on the TripAdvisor data. Lastly, in step 4, we analyzed the spatially varying relationship between the tourism supply index obtained from Task 2 and tourism supply components from TripAdvisor to identify how strongly those two datasets are correlated. In other words, whether a series of tourism supply indices from Task 2 is reliable depends on the correlation between them. The detailed explanation for data and analysis methodology is described in the data and methodology sections.

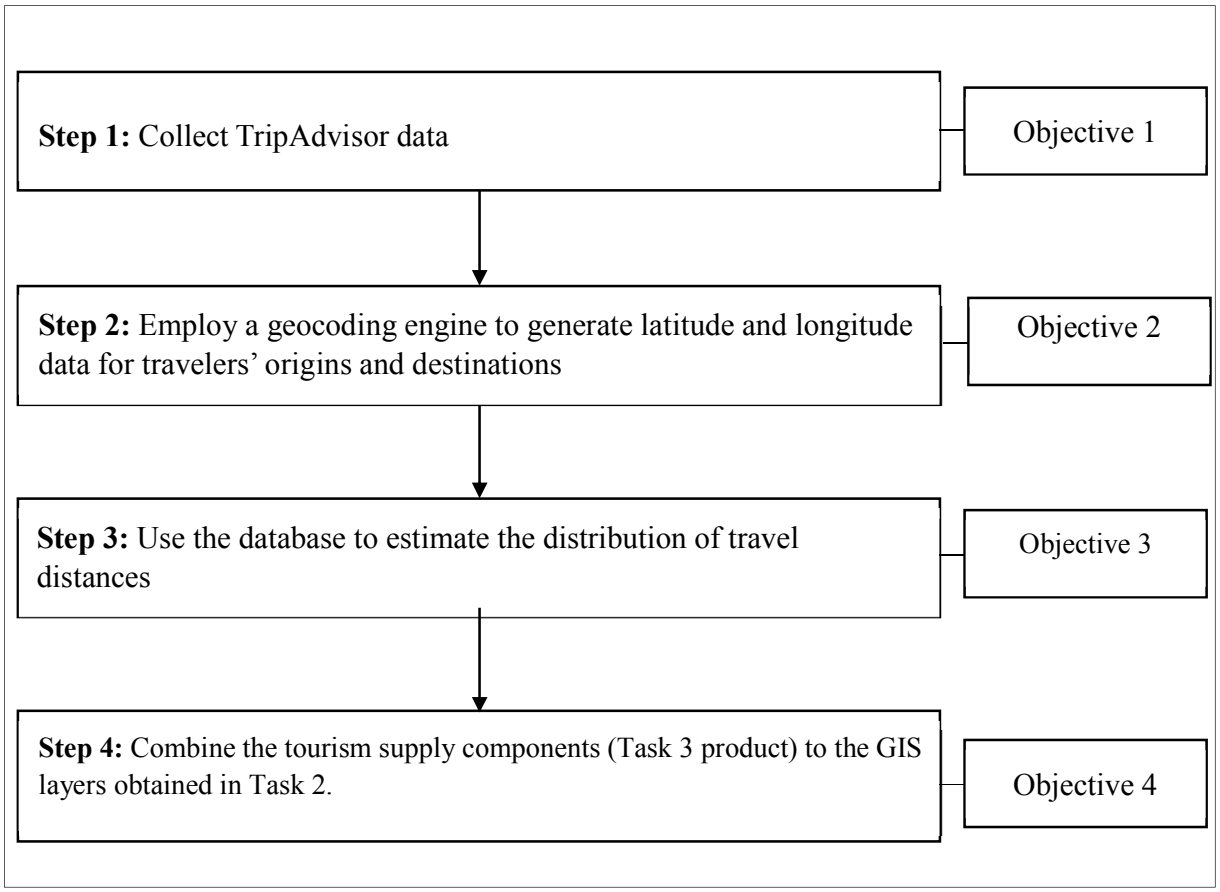


Figure 3.1 Flowchart for Task 3

3.3 TripAdvisor data collection

The TripAdvisor reviews were collected for the entire Florida starting from 2008 until October 2019. TripAdvisor is arguably the largest travel information platform at present, with more than 460 million average monthly unique visitors and over 830 million user generated reviews on 8.6 million hotels, restaurants, and attractions worldwide (TripAdvisor, 2019). TripAdvisor is the channel where travel customers provide various dimensions regarding their travel experience by rating, ranking, and reviewing the facilities and activities they have attended. For this reason, a remarkable number of studies in tourism have been using TripAdvisor as data sources for tourism research (Taecharungroj & Mathayomchan, 2019).

The TripAdvisor data include several major components of a travel property review, such as the property type (attraction, hotel, restaurant and rental properties), ID, name, address, review ratings, and the total number of reviews (Figure 3.2). Based on this identifiable information, we retrieved the following main variables as for the analysis of tourism supply in a region.

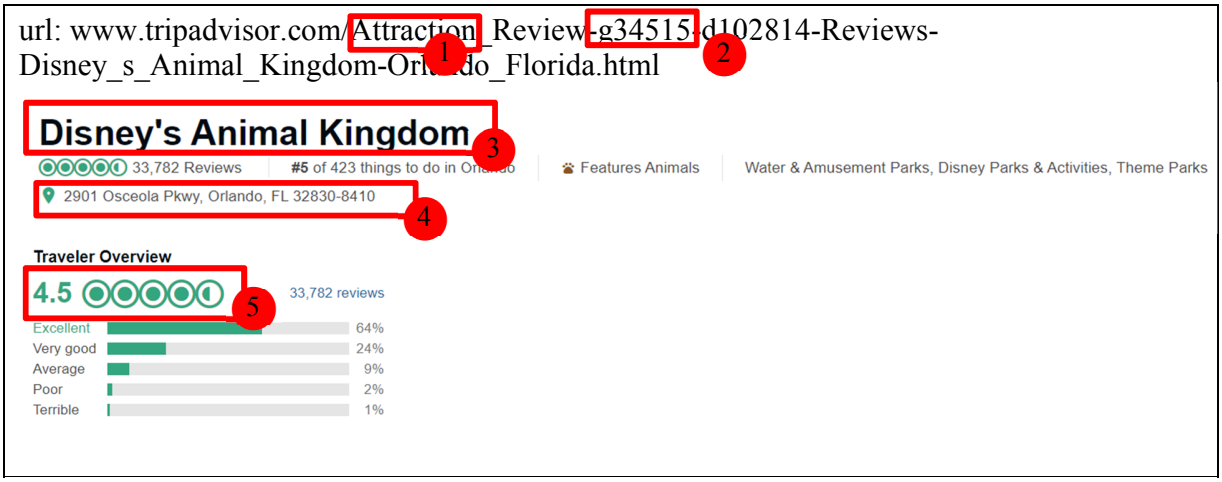


Figure 3.2 Example of an attraction review. The elements are as follows: 1: property type (hotel, attraction, restaurant, rental); 2: TripAdvisor assigned property ID; 3: property name; 4: street address; 5: detailed rating score.

The TripAdvisor data collection was done in the following way:

- The IDs for users staying in Florida hotels were collected and filtered to leave only those who volunteered to leave their place of living (with at least the city granularity for Florida residents, state resolution for the US visitors and country resolution for the international visitors);
- For the filtered IDs, all reviews (attraction, hotel, restaurant and rental properties) were collected;
- The collected reviews were filtered to leave only Florida data;
- Property latitude and longitude was collected from TripAdvisor;
- Users' self-described locations were resolved to latitude and longitude using Google geotagging engine.

In total, data on 51,525 Florida tourism properties were collected: 71.2% of the properties were, 18.3% were attractions, 9.8% were hotels, and 0.7% were rental properties; the latter were excluded from further analysis due to their insignificant share. Figure 3.3 shows distribution of the properties and their review ratings, the latter to be used in Task 4.

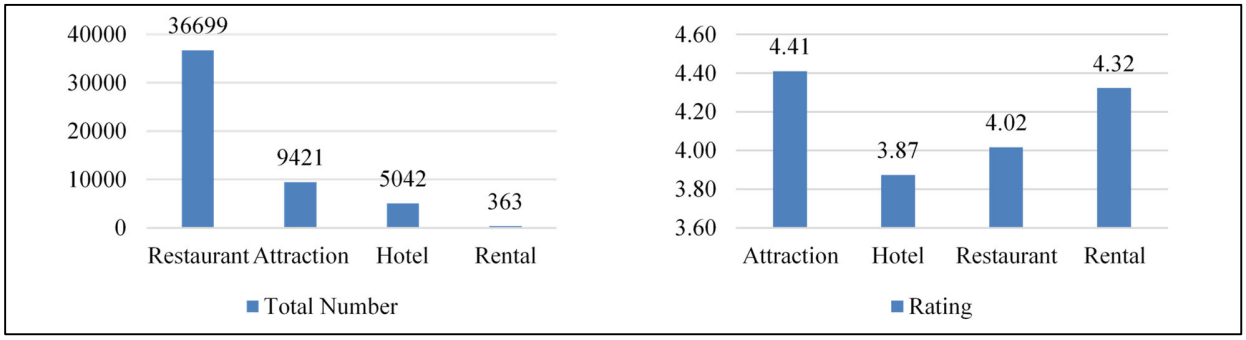


Figure 3.3 Summary of tourism properties in Florida

In total, 2,622,713 reviews were collected from TripAdvisor. After removal of the reviewers without origin information, the number of reviews was reduced to 2,162,249. In total, there were 250,844 reviewers reviewing 51,525 properties.

3.4 Geotagging

Latitude and longitude of the properties were scraped from TripAdvisor's property page (Figure 3.4). User's location was scraped from the review data and then resolved to latitude and longitude using Google API.

Figure 3.4 Scraping property's latitude and longitude

Based on users' self-described locations, we categorized the origins and corresponding reviewers into four types: the in-state Floridians, near-state visitors (Alabama and Georgia), out-state domestic US tourists and international tourists (Table 3.1, see Section 7.3 for details).

Among total 250,844 reviewers:

- 61,308 (24.4%) were in-state visitors, contributing to a total of 921,231 (42.6%) reviews on destination properties in Florida
- 12,973 (5.2%) were near-state visitors from Alabama or Georgia, accounting for the 91,589 reviews of Florida destination properties
- 131,005 (52.2%) were domestic tourists except the previous two groups. The out-state tourists made 856,400 (39.6%) reviews
- 45,558 (18.2%) were international tourists, contributing to 293,029 (13.6%) reviews.

Table 3.1 provides a summary of locations for collected reviews.

Table 3.1 Summary of the traveler origins and locations based on the collected reviews

origin	N locations	N reviewers	N reviews	toponymal resolution	tourist type	travel mode assumption
FL	518	61,308	921,231	place-level	In-state	car
GA, AL	605	12,973	91,589	place-level	near-state	car
Other USA	52	131,005	856,400	state-level	out-state	flight
international	185	45,558	293,029	country-level	international	flight
Total	1,360	250,844	2,162,249			

The toponymal resolutions of the total 1,360 origin locations were extracted and treated in different levels:

- 518 FL locations and 605 near-state locations with relatively high-resolution, in place/municipality level
- 52 out-state locations with medium-resolution, in state level
- 185 international locations with low resolution, in country level

The top origins of the reviewers of each types are showed in Table 3.2 (see Section 7.3 for details).

In addition, auxiliary geotagging data were collected or generated to support the generation of travel distances in Objective 3 and Task 4 (see Section 7.1 and 7.2 for details):

- **Places to County lookup tables**: the origin locations of the visitors were organized in places or municipality level (city, township, village or CDP). We generated lookup tables to interchange the geolocations from place to county level. The county values were assigned based on the centroids of places. It is noteworthy that most places are completely within single county, yet with few exceptions overlapping on multiple counties. We minimalized the error by manually reviewing and assigning those places to the county where the majority coverage located.

Table 3.2 Top origin locations of different types of visitors

#	country	reviews	state	reviews	near-state	reviews	in-state	reviews
1	UK	133,559	NY	82,422	Atlanta, GA	22,076	Orlando, FL	61,388
2	Canada	63,878	PA	54,314	Birmingham, AL	3,291	Tampa, FL	47,354
3	Brazil	7,130	OH	50,664	Savannah, GA	3,284	Miami, FL	45,997
4	Australia	7,077	IL	48,794	Marietta, GA	2,155	Jacksonville, FL	31,080
5	Ireland	5,419	NJ	43,636	Huntsville, AL	1,614	Fort Lauderdale, FL	27,920
6	Germany	5,161	TX	42,342	Valdosta, GA	1,477	Sarasota, FL	25,719
7	Netherlands	5,104	NC	41,422	Mobile, AL	1,423	Naples, FL	24,585
8	Sweden	3,527	MI	38,288	Augusta, GA	1,127	Fort Myers, FL	22,349
9	France	3,426	MA	36,929	Columbus, GA	1,125	St. Petersburg, FL	15,746
10	Switzerland	3,381	TN	35,180	Alpharetta, GA	1,083	Boca Raton, FL	14,414
11	Italy	3,265	VA	32,912	Cumming, GA	1,040	Ocala, FL	13,750
12	Mexico	2,958	CA	27,800	Dothan, AL	972	Cape Coral, FL	13,655
13	Spain	2,751	IN	26,466	Macon, GA	948	West Palm Beach, FL	13,493
14	Argentina	2,622	SC	26,277	Brunswick, GA	892	Tallahassee, FL	13,466
15	Norway	2,562	MD	23,042	Athens, GA	882	Gainesville, FL	12,516
16	Denmark	1,758	WI	21,059	Roswell, GA	875	St. Augustine, FL	12,153
17	Belgium	1,646	MO	20,196	Madison, AL	854	The Villages, FL	12,035
18	Trinidad and Tobago	1,529	MN	19,357	St Simons, GA	854	Clearwater, FL	11,207
19	New Zealand	1,448	CT	18,940	Montgomery, AL	836	Port St. Lucie, FL	10,008
20	Bahamas	1,440	KY	17,336	Canton, GA	805	Lakeland, FL	9,956

- **Distance to the nearest major airport**: we calculated the road distance from each county to its nearest major airport using Google Distance Matrix API (https://developers.google.com/maps/documentation/distance-matrix/start). P-L airports based on FAA airport categories (Large hub that accounts for at least 1% of total U.S. passenger enplanements) in Florida were chosen as the major airports, namely MCO, MIA, FLL and TPA. These four major airports account for about 70% of total Florida passenger enplanements.

3.5 Distribution of travel distances

The distribution of travel distances between any possible origins to destinations is meant to offer distance reference to further trip generation and OD matrix establishment in Task 4. For each type of tourists, their distribution of travel distances varies and we used different methods to calculate (see Section 7.2 for details):

- **For in-state tourists**, their distributions of travel distances were based on **Florida County to Census Tract travel distances and times**, where the road distance and travel time were generated by Google Distance Matrix API. The calculations and estimations were based on the center of population (2010 census) in each county and each census tract.
- **For near-state tourists,** their distributions of travel distances were based **on AL/GA County to FL Census Tract distances,** where the distances were great-circle distances calculated using the Haversine formula based on centers of population (2010 census) of each county in Alabama/Georgia and of each census tract in Florida.
- **For out-state tourists,** their distributions of travel distance were based **on State to FL county distances,** where the distances were great-circle distances calculated using the Haversine formula based on centers of population (2010 census) of each state and of each Florida county.
- **For international tourists,** their distributions of travel distances were based **on Country to FL county distances,** where the distances were great-circle distances calculated using the Haversine formula based on centers of population of each country and of each Florida county.

3.6 Combining the tourism supply components with GIS layers (Chapter 2)

First, latitude and longitude data of tourism supply components were converted to the census tract level using the spatial join function in ArcGIS (10.4.1). The spatial join function was used to combine the attributes of different features based on their spatial relationship. In a spatial join, the target features were the point data of tourism supply components, and the join features were the census tract locational data. The following parameters of the spatial join were used: 1) the intersect match: the features in the join features are matched if they spatially intersect a target feature and 2) the merge rules: average the rating values and sum the number of reviews and properties. Through the above process, we generated the tourism supply component data at the census tract level. Finally, the census tract data of tourism supply components were joined to the GIS layer obtained from Task 2 based on census tract codes.

Second, we generalized the census tract level tourism supply components to the county level using the dissolve function in ArcGIS (10.4.1). The dissolve function creates a new feature by merging polygons

with the same geo-reference such as the Federal Information Processing Standards (FIPS) county codes. Using county-level FIPS codes, we merged the census tract-level polygons into county-level polygons.

Finally, the county data of tourism supply components were joined to the GIS layer obtained from Task 2 based on county codes. Overall, we combined the tourism supply components and the GIS layers obtained in Task 2 at the county and census tract levels. The spatial distribution of the combined data is shown on Figures 3.4 – 3.6.

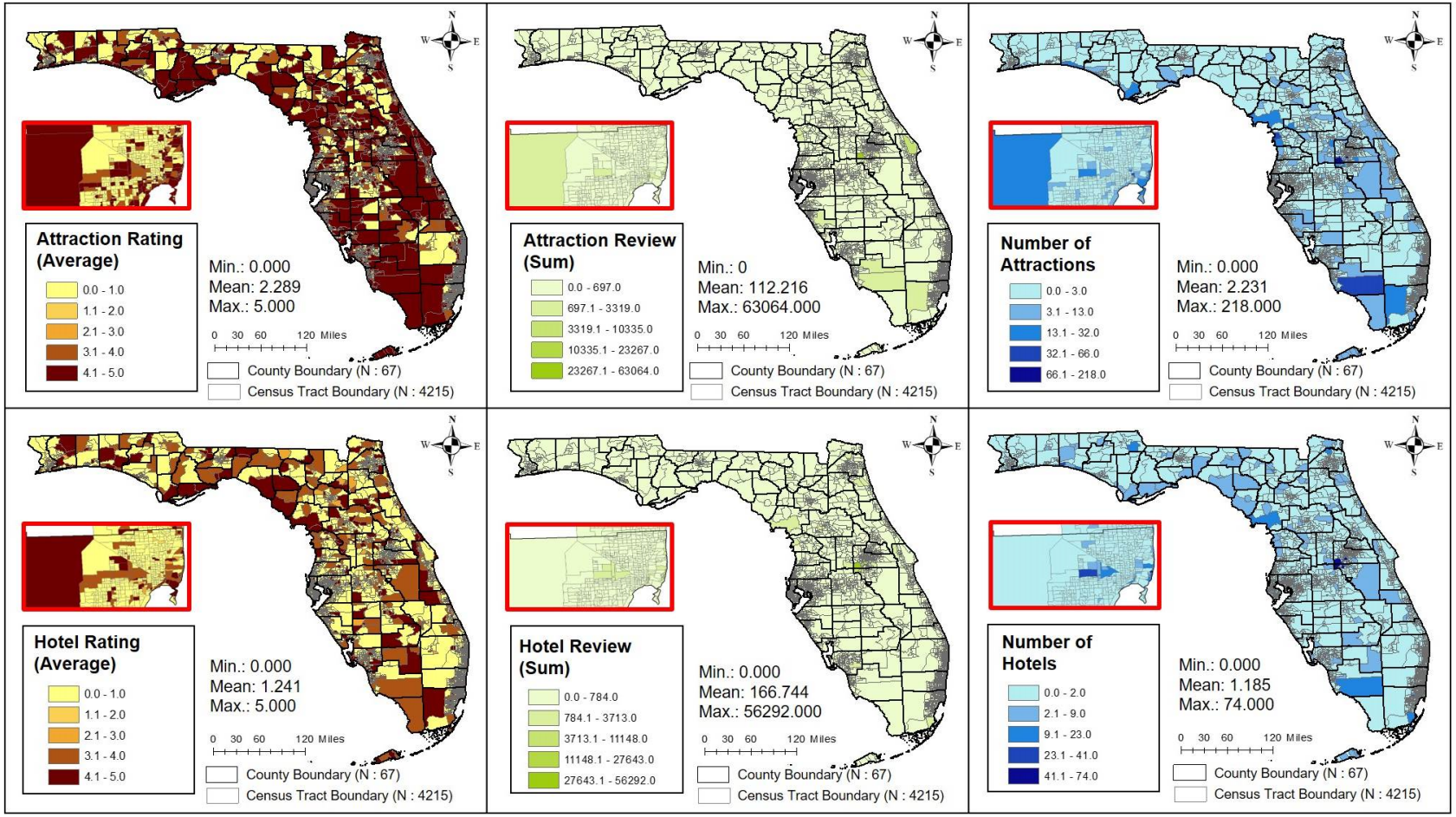


Figure 3.5 Spatial Distribution of Values from TripAdvisor (Top: Attractions; Bottom: Hotels)

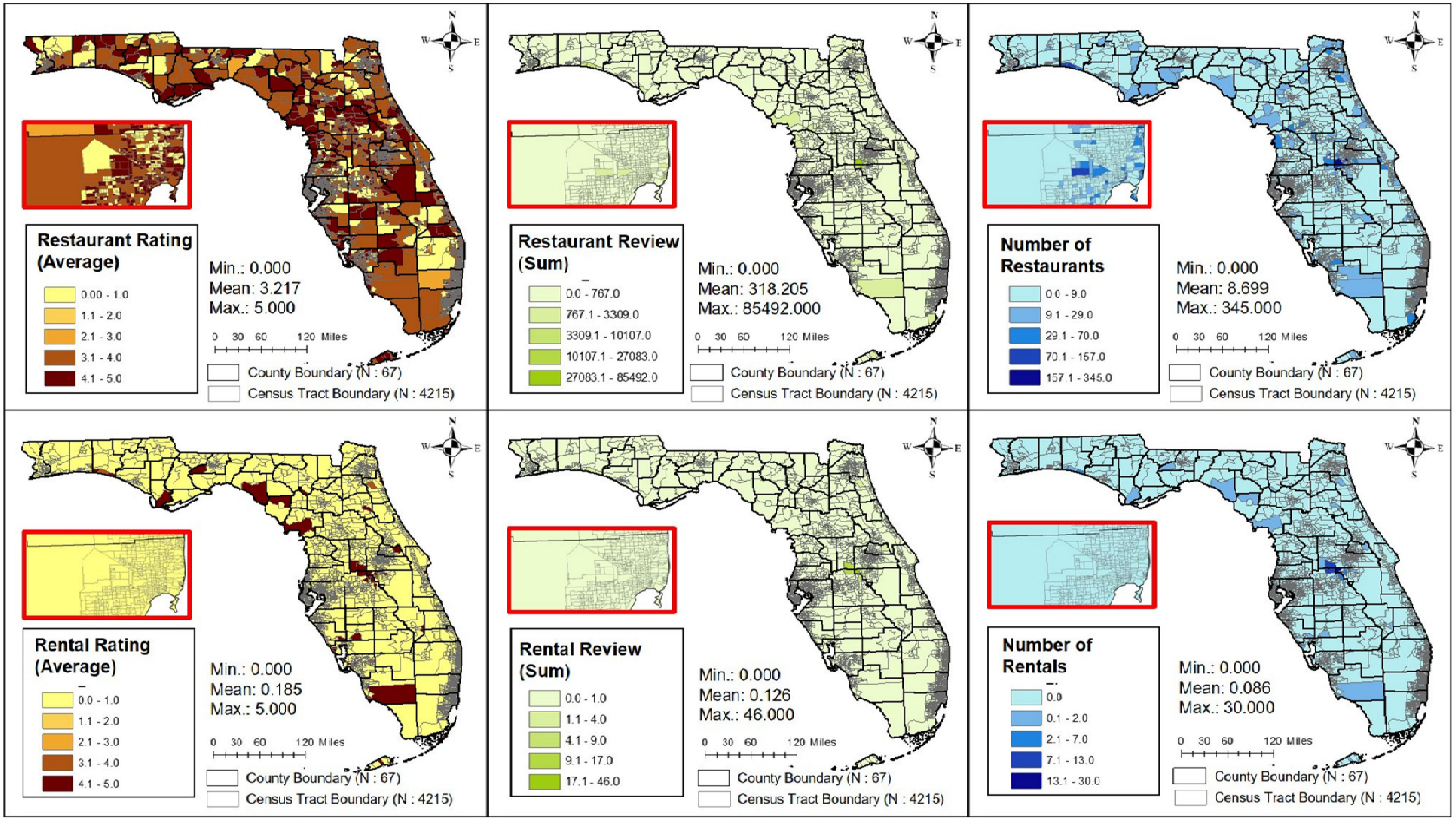


Figure 3.6 Spatial Distribution of Values from TripAdvisor (Top: Restaurants; Bottom: Rentals)

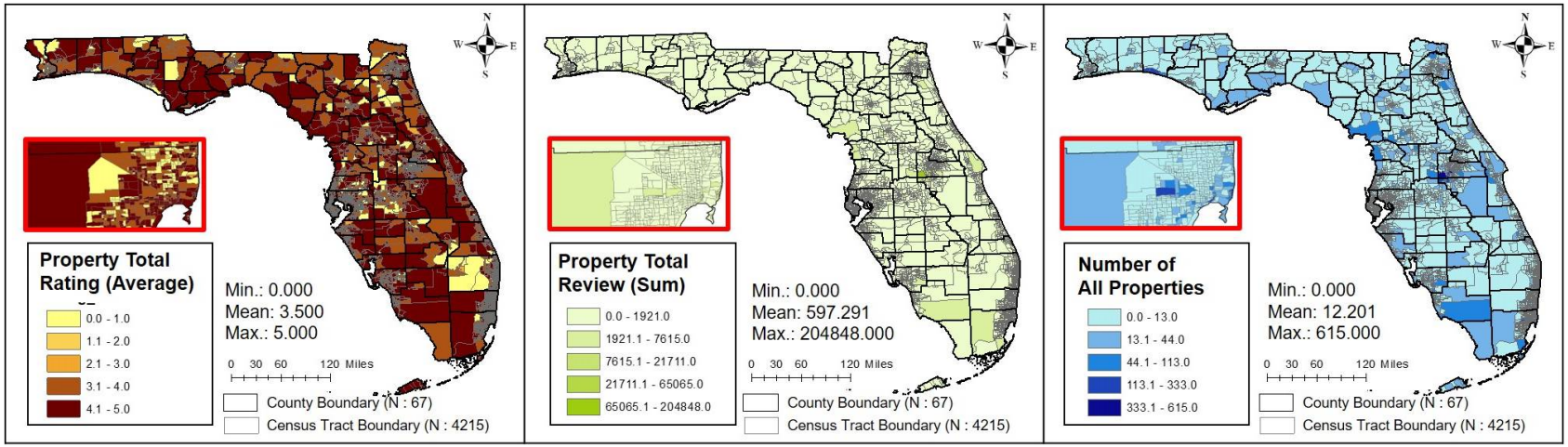


Figure 3.7 Spatial Distribution of Values from TripAdvisor (Property Total)

3.7 Data storage

The collected data was stored in Google drive with the following access:
https://drive.google.com/open?id=11iD3c05UpRwWqexqesDzI1q7-5yGeW84
Each folder contains a Readme file with metadata describing the data and access.

3.7.1 Location reference data (folder 0_Location Reference Data):

ref_fl_cty_pop2010_center: Florida county population (2010 census), centers coordinates;

ref_fl_tract_pop2010_center: Florida census tract population (2010 census pop), centers coordinates;
Source: US Census Bureau (https://www.census.gov/geographies/reference-files/time-series/geo/centers-population.html).
ref_fl_pla2cty: lookup table between places GEOID and county GEOID in Florida;

ref_al_ga_pla2cty: lookup table between places GEOID and county GEOID in Alabama and Georgia;
Source: API of the Federal Communications Commission (https://geo.fcc.gov/api/census/#!/block/get_block_find)

3.7.2 Distance reference data (folder 0_Distance Reference Data):

ref_fl_cty2airport_network: Florida county distance to nearest major airport. Note: zero or negative values - the airport is within the county; negative values additionally indicate the annual passenger volume in the largest airports.

ref_fl_cty2tract_links: Florida County to tract travel road links (based on Google road distance API).

ref_fl_cty2tract_links_extra_missing_tracts: missing Florida County to tract travel road links due to a failed Google API – e.g., the original centroids were on water or on offshore islands.

ref_fl_cty2tract_dis_mat: Florida county to tract **road distance** matrix (distance in miles); based on Google road network estimation by Distance Matrix API (https://developers.google.com/maps/documentation/distance-matrix/start).

ref_fl_cty2tract_time_mat: Florida county to tract **road travel time** matrix (travel time in minutes); based on Google road network estimation by Distance Matrix API (https://developers.google.com/maps/documentation/distance-matrix/start)

3.7.3 Processed visitor home location data (folder 1_VisitRecords):

final_0_homelocation_summary: home locations for the reviewers; location distributions for each type of reviewers

final_0_visit_record_tract_level: review records of each reviewers indicating their type of tourists, their origin locations, their reviewing date and the location (census tract level) of the property they reviewed.

Chapter 4. Estimating tourism flows using the tourism supply components alone

Task description

- Use a geographically weighted regression (GWR) model to fit tourism visitations (productions) to tourism resources (attractions).
- Forecast tourism visitations for counties with no current visitation data

Deliverable: Upon completion of Task 4, the University shall submit to the Research Center at research.center@dot.state.fl.us a written report of the outcomes and the GIS layers. The Submission will follow the same format as specified in Task 2.

4.1 Introduction

Understanding tourist flows in Florida is important for estimating the number of visitors and planning the provision of effective transportation systems. Tourism is at its very core a distinct geographical phenomenon, involving the movement of tourists from one place – their places of origin or generating regions – to one or more destinations via a complex web of multimodal transportation network (Kang et al., 2014). Thus, tourist flows are spatial interactions between locations and are affected by a variety of push and pull factors (Marrocu & Paci, 2013). Specifically, destination places have tourism resources that generate tourists' demand (Li et al., 2017). Depending on the type of tourism resources, visitation varies across counties in Florida. However, it is hard to accurately estimate tourists' visitation over time because tourists can visit destinations through various transport modes, such as airplanes and cars. Thus, we develop a tourism flows model to fit tourism visitations to tourism resources, then forecast tourism visitations for counties with no current visitation data. Ultimately, we suggest an alternative to estimate visitation of all counties.

Objectives

The purpose of Task 4 is to build a model of tourism flows and estimate visitation for counties where the visitation data is missing. To achieve the task, two objectives are identified.

- Use of a geographically weighted regression (GWR) model to fit tourism visitations (productions) to tourism resources (attractions)
- Forecast tourism visitations for counties with no current visitation data

Note that visitation data is provided by the social media in this task and hence is interpreted as relative (to other counties) rather than absolute visitation counting. Relative visitation is convertible to absolute visitation through visitation estimates coming from e.g. Visit Florida as will be discussed in Task 6.

4.2 Methodology

The overall process of building a tourism flow model and estimating visitation involves several steps. Figure 4.1 presents an overall flowchart for Task 4. In step 1, we generated the origin-destination (OD) matrix and transformed the matrix into a visitation data set of a county unit. In step 2, we developed two geographically weighted regression (GWR) models to fit visitations to tourism resources. In step 3, we generated the equation based on the results of the GWR to forecast the number of visitations. In step 4, we estimated tourism visitations for counties with no visitation data based on the equation generated in step 2. The detailed explanation for data and analysis methodology is described in the data and methodology sections.

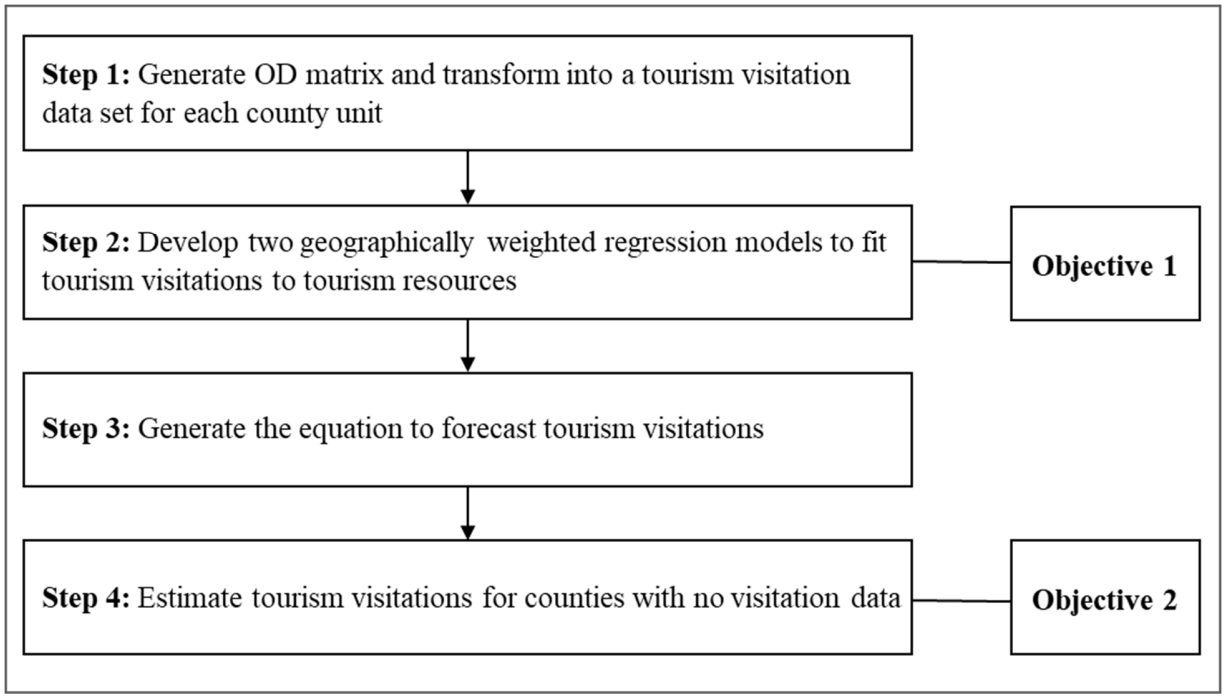


Figure 4.1 Flowchart for Task 4

4.3 Data

Four datasets were employed in Task 4. First, the tourism supply index obtained from Task 2 was used as tourism resource data. Specifically, the tourism supply index was created based on 35 tourism resource-related variables and 12 tourism resource factors in Task 2. Second, the number of tourism properties (attractions, hotels, restaurants, and rentals), obtained from TripAdvisor and constructed in Task 3, was also used as tourism resource data. Third, the number of beach access points constructed in Task 2 was also used as tourism resource data. Specifically, the beach is Florida's most attractive tourism resource, so the number of beach access points can improve the performance of the GWR model, which estimates the number of visitations. Fourth, a county-level tourism visitation data set was created based on the travel frequency matrix, which was formed by origin and destination trips, obtained from Task 3. Visitation data include visitors from other countries, other states in the United States, and other counties in Florida. All three datasets were constructed based on a county unit and used in developing tourism flows models and estimating tourist visitations. Specifically, the construction and transformation of the visitation data set derived from Task 3 is explained in the following.

4.3.1 Data source - TripAdvisor

As described in Task 3, we collected all TripAdvisor reviews visiting Florida starting from 2008 until October 2019. Each review possibly provides the following relevant information after geotagging:

- User ID;
- Self-report origin location name and address;
- Review rating;
- Review date;

- Visited property name;
- Visited property type;
- Visited property location and coordinates.

We selected the users who volunteered their places of living (origins) and the property locations they visited in Florida. After filtering, 2,162,249 individual review records left by 250,844 visitors were retrieved (see the report of Task 3, section 3 for the detail of data collection).

4.3.2 Define the zone system

4.3.2.1 Origin Zone

Based on visitors' self-described origin locations, visitors to Florida can be segmented into four types: Floridians, near-state visitors from Alabama and Georgia, domestic visitors and international travelers. The sizes of each visitor group and their review records are shown in Table 4.1.

Table 4.1 Visitor origins and types summary

Origin	# Visitors	# Review records	Travel mode assumption
Floridian	61,308	921,231	car
Near-State visitors (AL, GA)	12,973	91,589	car
Other domestic visitors	131,005	856,400	flight
International visitors	45,558	293,029	flight
Total	250,844	2,162,249	

It is noteworthy that the self-reported origins were defined at heterogeneous and sometimes broad spatial resolutions, namely, the toponyms could be in place (city, township, village or CDP), county, state, and nation level. In particular, the place-level toponyms were not aligned to the Standard Hierarchy of Census Geographic Entities system (census blocks – census tracts – counties – states – nations). Hence, a new Zone System for visitors to Florida was defined:

- Transform the place-level toponyms into county level to align to the Standard Hierarchy of Census Geographic Entities system for consistency of analysis.
- The counties were assigned based on the centroids of places. Given that boundaries of counties and places may overlap, no perfect match accuracy can be guaranteed. We minimized the error by manually reviewing and assigning those places to the county where the majority coverage they are located (refer to *ref_fl_pla2cty* and *ref_al_ga_pla2cty* for GEOID Look-up Tables)
- Traffic analysis zones (TAZs) of origins for different types of visitors were defined in different resolutions: international visitor – nation level; domestic visitors – state level; near-state and Floridian visitors – county level (see Table 4.2).

Table 4.2 Origin zone resolution and coding

Visitor type	Origin resolution	Coding format	# Origins
Floridian	County level	12099 (County FIPS)	67
Near-State visitors (AL, GA)	County level	13067 (County FIPS)	211
Other domestic visitors	State level	US30 ("US" + State FIPS)	52
International visitors	Nation level	ABW533 (ISO code + UNSD code)	185

4.3.2.2 Destination Zone

Given that the coordinates properties are known, the destination zones can be easily coded with either a county or census tract granularity. For the sake of simplicity and ease of following analysis, the destination zones of visitors from broader origins were correspondingly coded in broader resolution, and vice versa (Table 4.3).

Table 4.3 Destination zone resolution and coding

Visitor type	Available resolution	Resolution for analysis	Coding format	# Desti-nations
Floridian	Census tract	Census Tract	12031010301 (Tract FIPS)	3458
Near-State visitors (AL, GA)	Census tract	Census Tract	12031010301 (Tract FIPS)	2027
Other domestic visitors	Census tract	County	12031 (County FIPS)	66
International visitors	Census tract	County	12031 (County FIPS)	66

4.3.3 Trip observation

For each individual user of TripAdvisor, their review records compose a trajectory of visitation footprints during the observed time frame, which can be found in the following way.

Step 1. Split Trips

Following the statistics on the average length of stay of different groups of visitors from 'Visit Florida' annual report (Floridians less than 2 nights, near-state visitors less than 4 nights, domestic visitors less than 7 nights and international visitors less than 14 nights) (Visit Florida, 2019), we split the consecutive trajectory of review records of each individual into separate trips assuming their trip durations aligning with statistical stay length. Any two review records with a time gap larger than the threshold of the typical stay of its kind were regarded as two separate trips, otherwise as in the same trip.

Step 2. Identify the Day of the Trip

In a multiple-day trip, we marked the first day as the *'arrival day'*, which contains an O-D trip from visitor's origin to the destination. Likewise, the last day is the *'leaving day'* of the trip inevitably generating a D-O trip from the last destination back to home. Any day except the previous two was regarded as the *'within-trip day'*.

Step 3. Generate Trip Record

(1) OD trip for short-distance visitors

For short-distance visitors (Floridians and near-state visitors), if there is only one visited zone on the *'arrival day'*, that visited zone is no doubt to be the destination and thus marks a corresponding O-D trip record; otherwise, the most far visited zone from the origin is regarded as the final destination. Vice versa for the *'leaving day'* of the trip to generate a D-O trip record.

Any other movement between TAZs (between days or on '*within-trip day' or* the remaining movements on *'arrival day'* and *'leaving day'*) are marked as D-D trips (within destination).

(2) OD trip for long-distance visitors

For long-distance visitors (domestic and international visitors), if there is only one visited zone on the *'arrival day'*, that visited zone is no doubt to be the destination and thus marks a corresponding O-D trip record; otherwise, the zone with the nearest distance to a major airport is regarded as the first stop of the trip, generating a O-D trip in that regard. Vice versa for the *'leaving day'* of the trip to generate a D-O trip record.

Any other movement between TAZs (between days or on '*within-trip day' or* the remaining movements on *'arrival day'* and *'leaving day'*) are marked as D-D trips (within destination).

Step 4. Generate OD, DO and DD Matrices

The OD trip records, DO and DD trips are aggregated together to generate respective OD, DO and DD Matrices. It is noteworthy that the OD and DO matrices are directed asymmetric matrices while the DD matrix is a undirected symmetric one, given that we are proactively assume there is a travel flow between any within destination stops, inevitably inflating the frequency of DD travels.

- The OD matrix of Floridian visitors was at the county * census tract level (67*3458 matrix), vice versa for DO matrix
- The OD matrix of near-state visitors was at the county * census tract level (211*2027 matrix), vice versa for DO matrix
- The DD matrices of Floridian and near-state visitors were at census tract level (3458 *3458 matrices)

- The OD matrix of domestic visitors was at the state * county level (52*66 matrices), vice versa for DO matrix
- The OD matrix of international visitors was at the nation * county level (185*66 matrices), vice versa for DO matrix
- The DD matrices of domestic and international visitors were at county level (67 *67 matrices)

4.3.4 Visitation dataset construction

Visitation data were generated and constructed for each county in Florida based on review records from in-state Floridian, near-state visitors and from other states and abroad (Table 4.4). It is noteworthy that the concept of visitation here is review counts from social media, an alternative of the traditional visitor arrivals.

Table 4.4 Visitation to each county in Florida (social media data)

County	International	Domestic	Near-State	Floridian	Total	County	International	Domestic	Near-State	Floridian	Total
Alachua	153	1,508	444	5,687	7,792	Lee	1,896	13,974	810	20,011	36,691
Baker	1	18	11	50	80	Leon	193	1,402	519	5,384	7,498
Bay	250	4,392	2,745	2,854	10,241	Levy	41	273	85	1,061	1,460
Bradford	8	56	10	166	240	Liberty	-	5	-	19	24
Brevard	1,367	7,704	1,164	12,704	22,939	Madison	4	61	17	151	233
Broward	7,354	24,287	1,480	24,282	57,403	Manatee	931	4,415	391	6,537	12,274
Calhoun	1	1	3	17	22	Marion	177	1,570	308	5,158	7,213
Charlotte	352	1,705	115	4,016	6,188	Martin	150	1,261	123	3,494	5,028
Citrus	232	1,128	174	3,291	4,825	Miami-Dade	20,168	32,951	2,216	23,960	79,295
Clay	27	309	66	1,029	1,431	Monroe	3,231	21,440	1,957	20,312	46,940
Collier	2,674	11,118	537	12,221	26,550	Nassau	164	2,403	927	3,127	6,621
Columbia	37	508	199	1,082	1,826	Okaloosa	208	4,142	1,654	2,958	8,962
DeSoto	16	72	11	241	340	Okeechobee	8	110	4	361	483
Dixie	1	16	9	110	136	Orange	38,342	77,614	5,998	51,113	173,067
Duval	577	5,840	2,112	11,811	20,340	Osceola	3,386	6,878	1,067	8,517	19,848
Escambia	217	3,575	1,346	3,284	8,422	Palm Beach	2,320	14,266	973	22,265	39,824
Flagler	108	1,125	218	2,632	4,083	Pasco	200	1,382	166	3,848	5,596
Franklin	71	589	428	799	1,887	Pinellas	4,667	22,600	1,744	27,292	56,303
Gadsden	4	97	25	180	306	Polk	708	2,424	418	6,551	10,101
Gilchrist	1	14	6	125	146	Putnam	9	145	34	500	688
Glades	3	16	3	46	68	St. Johns	844	6,325	1,641	14,388	23,198
Gulf	9	217	213	265	704	St. Lucie	241	1,421	198	3,367	5,227
Hamilton	1	5	2	59	67	Santa Rosa	49	842	248	792	1,931
Hardee	7	29	2	139	177	Sarasota	1,720	8,888	781	14,800	26,189
Hendry	48	107	11	266	432	Seminole	208	1,688	223	5,179	7,298
Hernando	81	636	78	1,770	2,565	Sumter	42	503	54	1,775	2,374
Highlands	61	429	27	1,272	1,789	Suwannee	6	60	20	241	327
Hillsborough	2,307	13,477	1,343	20,392	37,519	Taylor	12	106	42	309	469
Holmes	3	41	23	71	138	Union	-	-	-	2	2
Indian River	252	1,971	178	4,037	6,438	Volusia	1,191	8,600	1,505	14,911	26,207
Jackson	18	159	54	397	628	Wakulla	40	195	82	468	785
Jefferson	9	31	11	109	160	Walton	161	3,214	1,126	1,410	5,911
Lafayette	5	9	2	38	54	Washington	2	83	19	131	235
Lake	329	1,791	194	5,979	8,293	**Total**	**97,903**	**324,221**	**38,594**	**391,813**	**852,531**

4.4 Method

4.4.1 GWR models

Two geographically weighed regression (GWR) models were developed with the goal of estimating tourist visitations from tourism resources of the counties. To reduce the influence of outliers, the counties with few visitation records (below 500; see Table 4.5) were removed from the GWR model; those counties account for less than 0.5% of total visitations. Then, two GWR models (Table 6) were created to analyze the influence of the tourism resource index (model 1) or the number of properti4.es (i.e., attractions, hotels, and restaurants), the tourism resource index, and the number of beach access points (model 2) on visitation.

Table 4.5 Counties with fewer than 500 visitations

FIPS	County	Visitations	FIPS	County	Visitation
3	Baker	80	59	Holmes	138
7	Bradford	240	65	Jefferson	160
13	Calhoun	22	67	Lafayette	54
27	DeSoto	340	77	Liberty	24
29	Dixie	136	79	Madison	233
39	Gadsden	306	93	Okeechobee	483
41	Gilchrist	146	121	Suwannee	327
43	Glades	68	123	Taylor	469
47	Hamilton	67	125	Union	2
49	Hardee	177	133	Washington	235
51	Hendry	432			

Table 4.6 GWR models 1 and 2

Model	Dependent variable		Independent variables	
	Variable	Operational definition	Variable	Operational definition
1	Visitation	Number of visitations	Tourism supply index	Tourism supply index
2	Visitation	Number of visitations	Attractions	Number of attractions
			Hotels	Number of hotels
			Restaurants	Number of restaurants
			Beaches	Number of beach access points
			Tourism supply index	Tourism supply index

4.4.2 GWR analysis

A GWR is a spatial regression technique which can estimate local spatial heterogeneity between variables. GWR assumes that relationships between variables may differ from location to location (Fotheringham et al., 2002). In Task 4, GWR was employed to explore the spatially varying relationships between the dependent variable (visitation) and independent variables (the tourism supply index and numbers of attractions, hotels, restaurants, and beach access points) from Tasks 2 and 3. If the GWR results show that visitation and tourism-related variables are highly correlated, then tourism-related variables are reliable in estimating visitation where the observational data are missing. The proposed GWR model is as follows:

$$Visit_i = \beta_{i0}(u_i, v_i) + \beta_{ik}(u_i, v_i)Tourism_{ik} + \varepsilon_i \quad (4.1)$$

where $Visit_i$ refers to the number of visitations at county i; (u_i, v_i) is the coordinate of the centroid at county i; and $\beta_{ik}(u_i, v_i)$ is the local regression coefficient for the independent variable k at county i. A bi-square kernel function was utilized to consider the different sizes of each county in Florida (Fotheringham et al., 1998). This function determines a specific number of neighbors used to maximize the model fit. The spatial weight (w_{ij}) for the bi-square function is estimated as follows:

$$w_{ij} = [1 - (d_{ij} / b^2)] \text{ when } d_{ij} \leq b, \ w_{ij} = 0 \text{ when } d_{ij} > b \quad (4.2)$$

where d_{ij} is the Euclidean distance from the regression point i and the property j, and b is the bandwidth (Fotheringham et al., 1998). An iterative statistical optimization was applied to mitigate the corrected Akaike Information Criterion (AICc). Lastly, the local coefficients and adjusted R^2 values from GWR models were mapped to visualize the relationships between visitation and tourism-related independent variables (i.e., the tourism supply index and numbers of attractions, hotels, restaurants, and beach access points). ArcGIS 10.4.1 was used for GWR analysis.

4.5 Results

4.5.1 Descriptive statistics

Table 4.7 shows the descriptive statistics for the dependent variable and independent variables in the county level. The average value of the tourism supply index was 1.2 (median: 0.9) and ranged from 0.1 to 5.9. In terms of property variables, the number of attractions ranged from 1.0 to 783.0 with a mean of 140.4 (median: 51.0); the number of hotels ranged from 0.0 to 505.0 with a mean of 74.5 (median: 25.0); the number of restaurants ranged from 2.0 to 3769.0 with a mean of 547.3 (median: 233.0); and the number of beach access points ranged from 0.0 to 182.0 with a mean of 32.6 (median: 1.0). Figure 4.2 displays the distribution of each variable.

Table 4.7 Descriptive statistics of dependent variables in models 1 and 2 (per county)

Variable		Minimum	Maximum	Mean	Median	Std. Deviation
Tourism supply index		0.1	5.9	1.2	0.9	1.1
TripAdvisor	Attractions	1.0	783.0	140.4	51.0	195.8
	Hotels	0.0	505.0	74.5	25.0	103.2
	Restaurants	2.0	3769.0	547.3	233.0	807.4
	Beaches	0.0	182.0	32.6	1.0	53.0

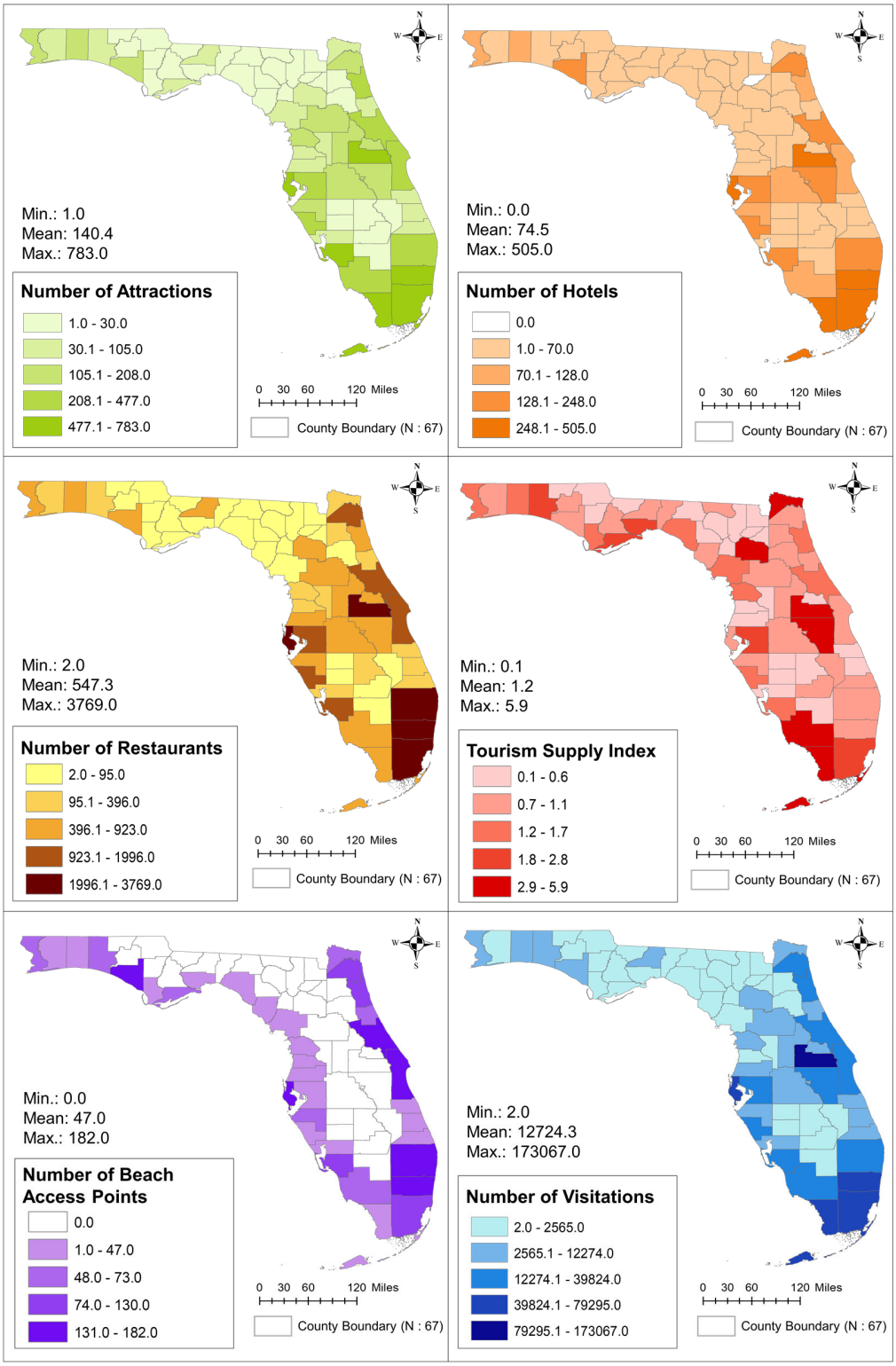


Figure 4.2 Spatial distributions of variables in models 1 and 2

4.5.2 Model 1: Visitation and Tourism Supply Index

The result of the GWR model 1 is summarized in Table 4.8. An ordinary least squares (OLS)-based regression model was first developed to investigate the global relationship between variables before conducting the GWR model. This step is a common approach to demonstrate the utility of GWR model (Charlton & Fotheringham, n.d.) To examine whether the GWR model exhibits better model performance than the OLS model, the values of adjusted R^2 and AICc from the OLS and GWR models were compared. As shown in Table 4.8, the adjusted R^2 increased from 0.13 (OLS model) to 0.18 (GWR model), and the AICc index decreased from 1074.82 (OLS model) to 1070.02 (GWR model). These findings suggest that the GWR model could offer better model performance than the OLS model to develop an equation for estimating visitation.

For the OLS model 1, the overall model was statistically significant (Joint F-Statistic: 7.86, p-value < 0.05). The tourism supply index was significant at a 95% level. Specifically, the coefficient for the tourism supply index (9,577.12, p-value < 0.05) indicated that counties with a higher level of tourism supply index have a higher number of visitations.

For the GWR model 1, the local R^2 ranged from a minimum of 0.00 to a maximum of 0.39 with a mean of 0.19. The spatial autocorrelation of residuals (Moran's I: -0.10, p-value: 0.38) and the local condition index (from 2.50 to 4.40) within a threshold of 30 indicate spatial randomness and the appropriateness of running a GWR model. The local condition index ranged from a minimum of 2.50 to a maximum of 4.40, indicating that there is no local collinearity issue. The local coefficients for tourism supply index ranged from -652.51 to 25,018.02 with a mean of 11,851.77. This variability in the local parameter estimates indicates spatial variability, which represents spatially varying relationships between tourism supply and visitation throughout Florida. Maps in Figure 4.3 show the spatial distribution of the local coefficients for the tourism supply index and local R^2 in the GWR model 1. Specifically, counties with strong positive local coefficients were observed mainly in the central regions of Florida, whereas counties with negative local coefficients were identified in the northwestern regions of Florida. The GWR model exhibited various values of the local R^2, which indicated that the exploratory power of the GWR model was not the same throughout Florida. In the results of the GWR model 1, counties with a significant negative correlation between the number of visitations and the tourism supply index can be interpreted largely in three ways. First, there is no balance between the demand for visitors and the supply of tourism resources. Second, since the visitation data were generated based on TripAdvisor, most users were concentrated in younger age groups, which may not reflect visits by relatively diverse age groups. Third, simply the number or area of tourism resources does not reflect the number of visitations. For example, a single amusement park in a county can have a large number of visitors, while a large campground in the county can have a small number of visitors.

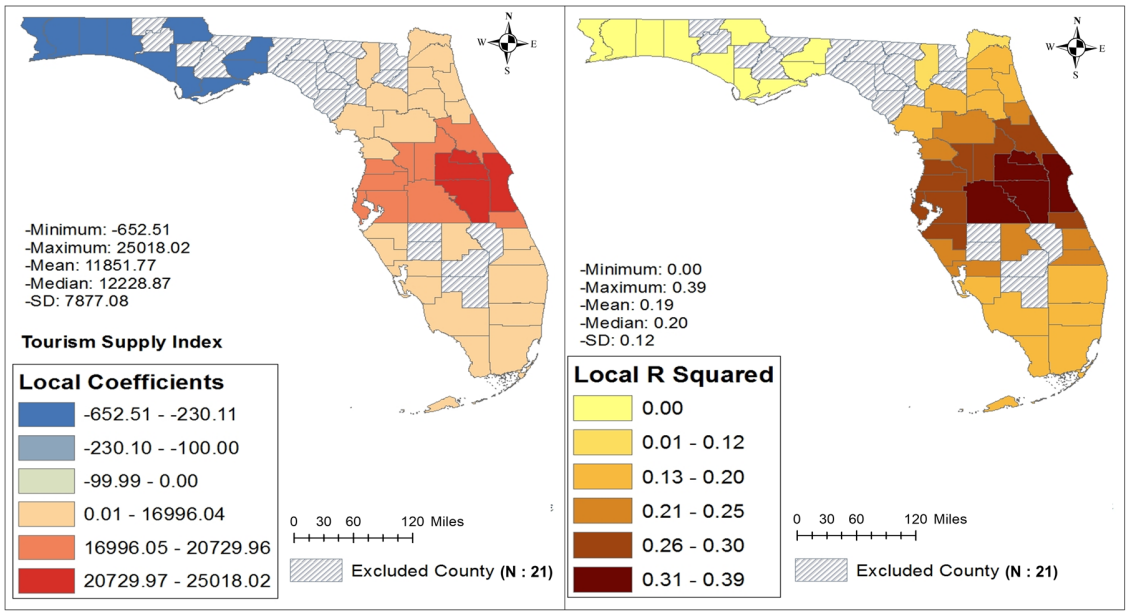


Figure 4.3 Spatial distribution of local coefficients for the tourism supply index and local R^2

4.5.3 Model 2: Visitation and tourism properties (attractions, hotels, restaurants, and beaches)

The initial OLS analysis explained 67% of the variance in the dependent variable, visitation, with all properties and tourism supply index variables included, but the model exhibited high multicollinearity for attractions (VIF: 11.0). Thus, this variable was excluded. Furthermore, the variables, the tourism supply index did not have a significant effect and was removed. As a result, the final model had three significant variables – hotels, restaurants, and beaches – and explained 67% of the variability in visitation. The results indicated that hotels, restaurants, and beaches were significant at a 95% level. The coefficients for hotels (120.41, p-value < 0.05), restaurants (16.33, p-value < 0.05), and beaches (50.31, p-value < 0.05) indicated that counties with more hotels, restaurants, and beaches have a higher number of visitations.

The results of the GWR model 2 are summarized in Table 4.8. To examine whether the GWR model exhibits better model performance than the OLS model, the adjusted R^2 and AICc values from the OLS and GWR models were also compared. As shown in Table 4.8, the adjusted R^2 increased from 0.70 (OLS model) to 0.75 (GWR model), and the AICc index decreased from 1028.06 (OLS model) to 1023.21 (GWR model). These findings suggest that the GWR model could offer better model performance than the OLS model to develop an equation for estimating visitation.

For the GWR model 2, the local R^2 values ranged from a minimum of 0.69 to a maximum of 0.94 with a mean of 0.82. The spatial autocorrelation of residuals (Moran's I: -0.23, p-value: 0.30) and the local condition indexes (from 6.36 to 11.30) within a threshold of 30 indicated spatial randomness and the appropriateness of running a GWR model. Based on the average local coefficients, all variables, hotels (mean: 113.76), restaurants (mean: 24.03), and beaches (mean: 58.04), were positively associated with visitation. Specifically, the local coefficients of the independent variables ranged from 16.62 to 246.74 with a mean of 113.76 (hotels), 9.21 to 46.39 with a mean of 24.03 (restaurants), and -8.21 to 143.51 with a mean of 58.04 (beaches).

Figure 4.4 illustrates the distribution of the local coefficients for the significant independent variables and local R^2. Specifically, counties with strong positive local coefficients for hotels and restaurants were observed mainly in the central and southern regions of Florida, whereas those with the strong positive local coefficients for beaches were identified in the southern regions. Two counties (Franklin and Wakulla) showed the negative local coefficients with regard to beaches, indicating that the number of beach access points was not positively correlate with the number of visitations. This result might be explained for two reasons. First, the visitation data based on the number of reviews from TripAdvisor may not accurately reflect the number of visitations to certain beach access points. Second, even if there are many beach access points, the number of visitations can actually be small, as there are fewer tourist-related facilities such as hotels and restaurants.

This variability in the local parameter estimates indicates spatial variability, which represents spatially varying relationships among variables across counties in Florida. The results of the GWR model 2 demonstrated that the number of visitations can be explained well with the numbers of hotels, restaurants, and beaches, indicating that these variables are suitable for estimating the number of visitations across counties in Florida.

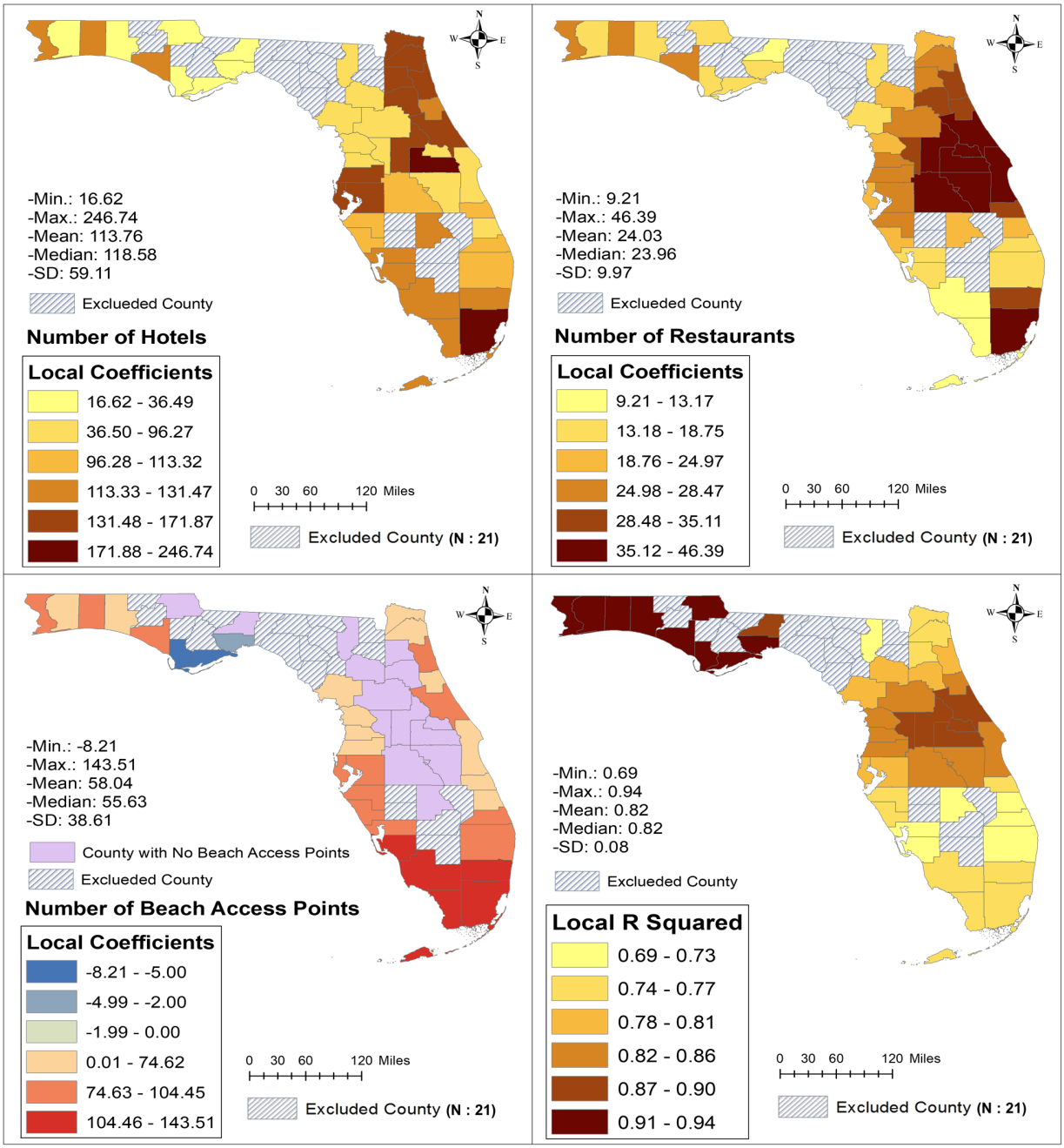


Figure 4.4 Spatial distribution of local coefficients for the independent variables and local R^2

Table 4.8 Results of GWR Models

Variables	Model 1 (Tourism supply index)				Model 2 (TripAdvisor properties)			
	OLS	GWR Coefficient			OLS	GWR Coefficient		
	β	Min.	Mean	Max.	β	Min.	Mean	Max.
Intercept	4336.17	-8970.50	2479.44	17213.73	-1778.48	-17549.45	-6139.26	3982.98
Tourism supply index	9577.12*	-652.51	11851.77	25018.02				
Hotels					120.41*	16.62	113.76	246.74
Restaurants					16.33*	9.21	24.03	46.39
Beaches					50.31*	-8.21	58.04	143.51
Local R^2	0.15	0.00	0.19	0.39	0.72	0.69	0.82	0.94
Adjusted R^2	0.13		0.18		0.70		0.75	
Condition Index		2.50	3.07	4.40		6.36	8.52	11.30
AIC_c	1074.82		1070.02		1028.06		1023.21	

* $p < .05$

4.5.4 Estimating tourism visitations

Based on the results of the GWR model 2, which showed a higher R^2 and better model performance, we created the equation to estimate tourism visitations for counties where the observational data were missing. The equation for estimating visitation is as follows:

$$\text{Visitation}_i = \beta_0 + \beta_1\text{Hotel}_i + \beta_2\text{Restaurant}_i + \beta_3\text{Beach}_i + \varepsilon \quad (4.3)$$

where Visitation_i refers to the number of visitations at county i; β_0 is the intercept; β_1 is the local coefficient for the independent variable, the number of hotels; β_2 is the local coefficient for the independent variable, the number of restaurants; β_3 is the local coefficient for the independent variable, the number of beach access points; and ε is the error term. The numbers of hotels, restaurants, and beaches were significant predictors that estimate visitation. In other words, the numbers of hotels, restaurants, and beaches greatly contributed to the number of visitations. As shown in Figure 5, after estimating the number of visitations to counties with visitation below 500, the final visitation data were mapped.

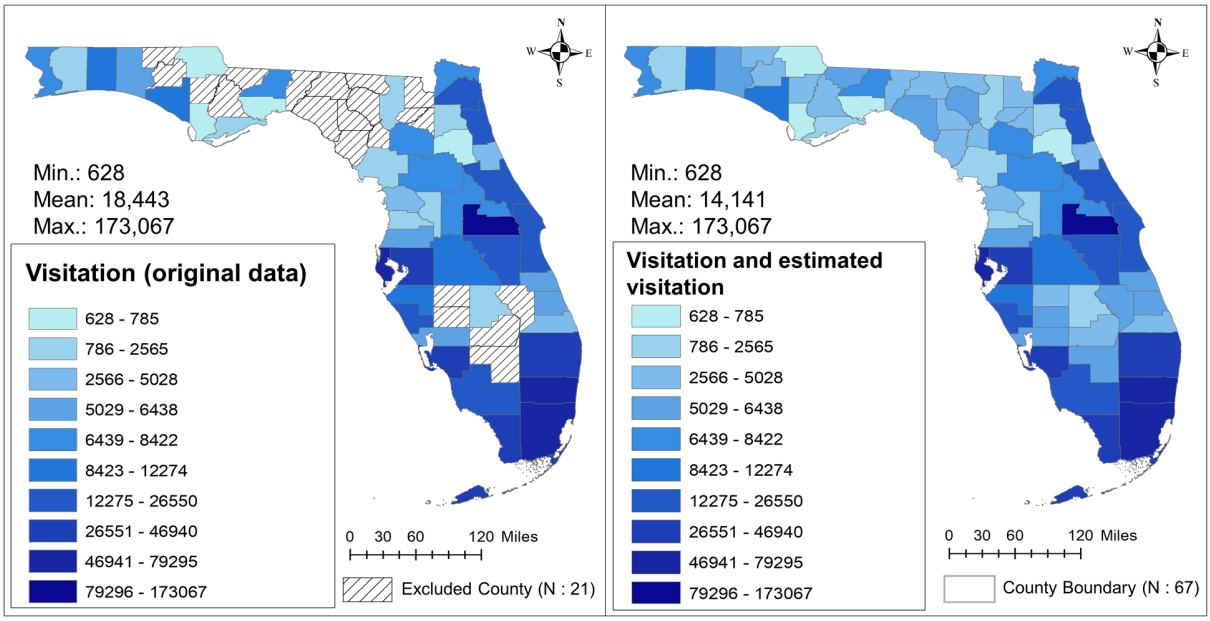


Figure 4.5 Spatial distributions of visitation (original data) and a combination of existing and estimated visitation data sets

Lastly, based on the equation developed, the number of visitations for all counties in Florida were calculated and the results between the original visitation data and the estimated visitation data were compared (see Figure 4.6). The number of visitations increased in most counties, where the number of visitations was significantly lower before. The minimum number of visitations increased from 2 to 3,473, and the average number of visitations increased from 12,724 to 26,947 in Florida. Although there were differences between the map with the actual visitation data and the map with the estimated visitation data, the difference between the average numbers of visitations was 14,223 which was lower than one standard deviation, 28,992.15. In summary, as a result of the GWR model, which is based on data about tourism properties closely related to the number of visitations, we derived the equation to estimate the number of visitations by county. This equation can later be used to estimate the number of visitations when data on the number of visitors per county is missing.

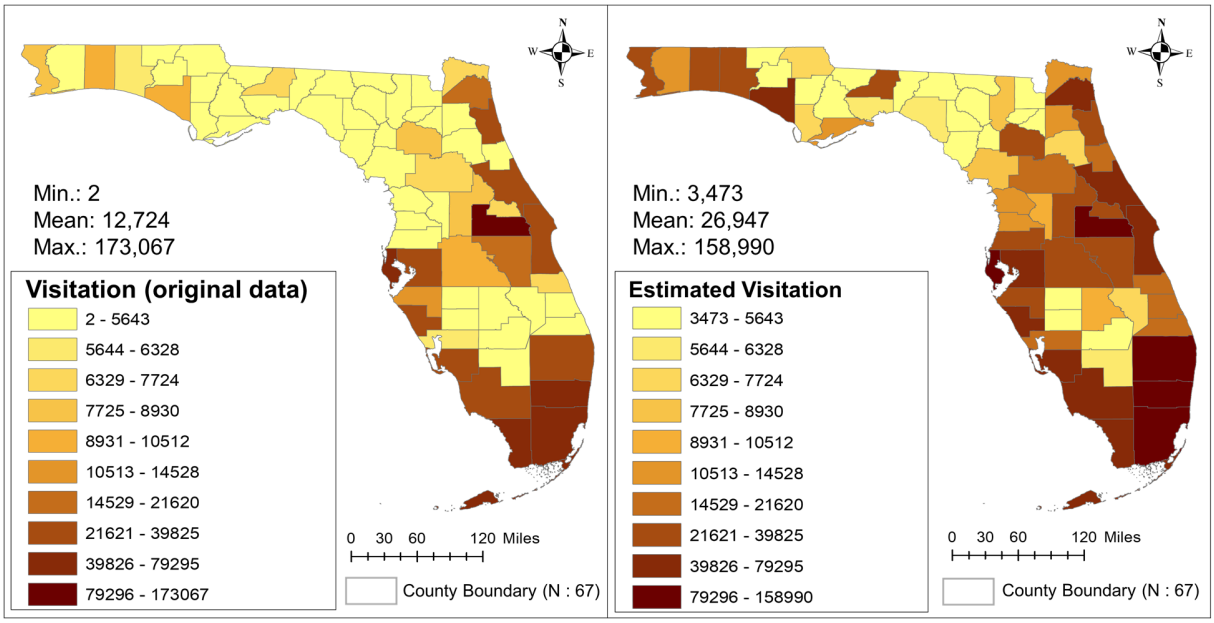


Figure 4.6 Spatial distribution of observations and predictions for visitation

4.6 Conclusion

Task 4 describes a procedure for estimating the number of visitations by county, through the tourism properties data obtained from Tasks 2 and 3. In this task, we applied GWR, which considers the spatial variability of tourism-related variables, to create the equation for estimating the number of visitations in each county. The results proved that the numbers of hotels, restaurants, and beach access points were significant predictors for estimating the number of visitations, while the tourism supply index was not highly correlated with the number of visitations. Thus, tourism properties data were reliable to estimate the number of visitations. As tourism is a top economic driver in Florida (Visit Florida, 2019), considering the exact number of visitors is important in the effective management and planning of a transportation system. Through the equation for estimating the number of visitations in Task 4, the Florida Department of Transportation (FDOT) can estimate visitation by the county without actual visitation data in the future, estimate their impact on the Florida transportation system, and develop transportation systems from macro- and micro- perspectives.

Chapter 5. Estimating tourism flow in Florida using social media and cell phone data and cross-validation

Task description

The research team validates Task 4 results using a comprehensive independent travel dataset. Specifically, census tract granularity generalized travel data estimated from the cell phone provided by AirSage is used to adjust estimated visitor flows. Additionally, Visit Florida tourism data is used to cross-check results against the real travel data.

- Outline the validation methodology and acceptable parameters.
- Validate spatial visitation patterns.
- Validate temporal visitation patterns.
- Calibrate and revalidate visitation projections as necessary.

Deliverable: Upon completion of Task 5, the University shall submit to the Research Center at research.center@dot.state.fl.us validated tourism visitation data consistent with the approved validation methodology. The submission will be in a form of (1) a written report and (2) GIS layers sent to FDOT following the same format as specified in Task 2

5.1 Introduction

One issue that has been frequently challenged in big data analytics, particularly in social media data analysis, is its lack of representativeness. Scholars have been argued that social media data are presumably biased towards the social media users in nature, and only stand for a fraction of the whole population. Therefore, it is vital to validate the data quality of social media collected from TripAdvisor in Task 3 and used in the prediction model in Task 4 with an additional independent dataset. The predication results derived from Task 4 are also in need of validation regarding its estimation performance as well. The independent datasets used for cross-validation include the mobile phone signal tower data provided by AirSage (www.airsage.com) and official tourism statistics from Florida statewide destination management organization (DMO) VisitFlorida (www.visitflorida.com).

5.2 Methodology

The goal of Task 5 is to validate the reliability of social media data retrieved from TripAdvisor in Task 3 with an extra dataset and validate the performance of the prediction model constructed in Task 4. There are the following objectives:

1. Outline the validation methodology and acceptable parameters.
2. Validate spatial visitation patterns.
3. Validate temporal visitation patterns.
4. Calibrate and revalidate visitation projections as necessary.

The following Figure 5.1 presents methodological framework of Task 5.

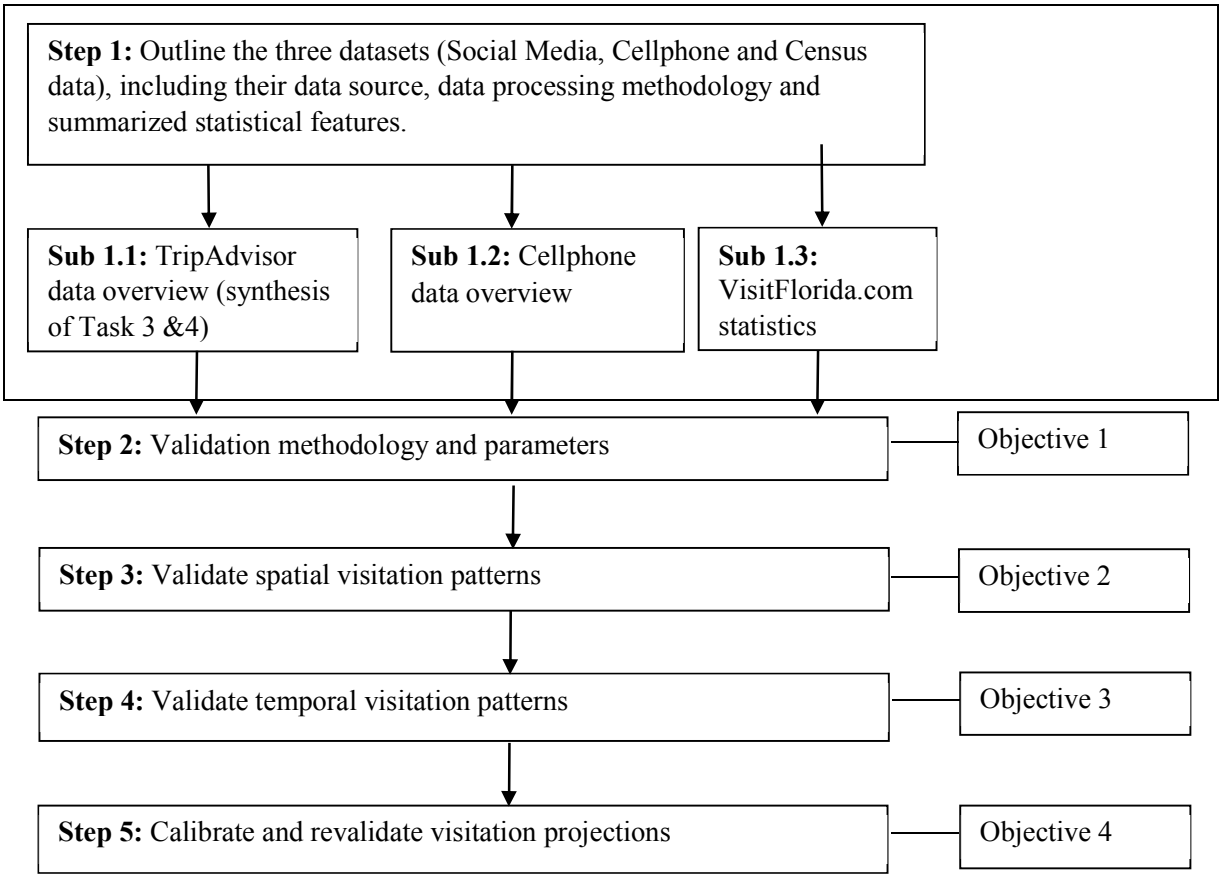


Figure 5.1 Framework of Task 5

5.3 Data review

5.3.1 TripAdvisor social media data (synthesis of Task 3 & 4, updated)

5.3.1.1 Original data snapshot

We collected TripAdvisor reviews of all properties in Florida registered with TripAdvisor with the timeframe from January 2003 to October 2019. The relevant variables collected include (1) Reviewer's ID; (2) Reviewer's home address (self-reported); (3) Reviewer's total review numbers; (4) Reviewed property ID; (5) Property type; (6) Property location coordinates; and (7) Review date.

During the social media data cleansing process, we (1) filtered out the abnormally active reviewers ranking in top 5%[1]; (2) selected the properties relevant to tourism activities only (hotel, restaurant, and attraction) and discarded the others; (3) used Google location API to geotag the home address of the reviewers; and (4) classified the visitors into three groups based on their origins, that is, Floridian, domestic and international. The home locations were kept with at least a city granularity for Floridian, a state resolution for domestic visitors, and a nation resolution for the international visitors. Data points with unidentified, uncertain, or low-granularity home locations were discarded.

The raw dataset comprised 2,622,713 reviews; after data cleansing, the number of reviews was reduced to 2,162,249. These reviews were generated by 250,844 reviewers visiting 51,525 properties in Florida. The temporal distribution of collected reviews is illustrated below in Figure 5.2 and mostly represents changes in TripAdvisor platform popularity.

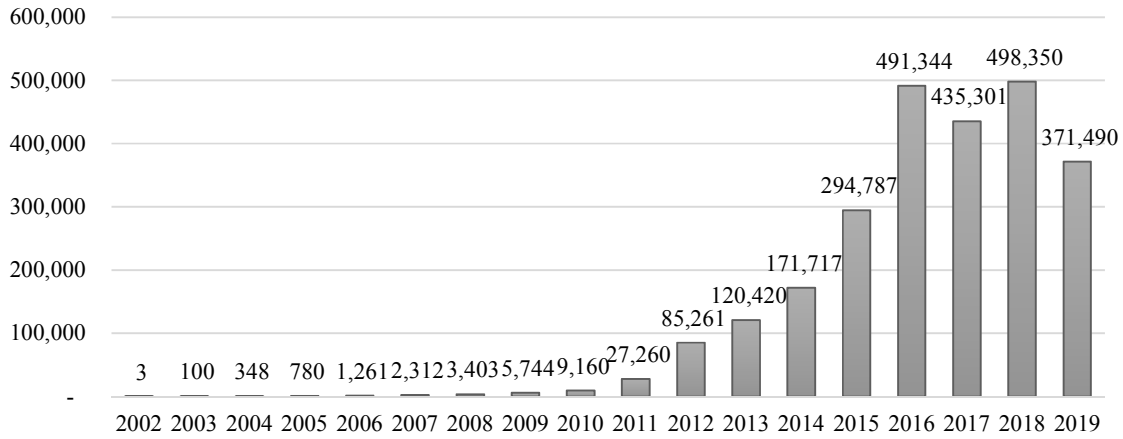


Figure 5.2 Temporal distribution of collected reviews

Based on users' self-described locations, the reviewers were grouped into (1) 61,308 (24.4%) Floridians, contributing to a total of 921,231 (42.6%) reviews; (2) 143,978 (57.4%) domestic visitors, making up 856,400 (39.6%) reviews; and (3) 45,558 (18.2%) international tourists, contributing to 293,029 (13.6%) reviews as detailed in Table 5.1.

[1] The top 5% reviewers (approximately equivalent to those who posted over 1,000 reviews in total) were presumed to mostly represent professional bloggers or promoting robots rather than genuine travelers.

Table 5.1 Summary of reviews and reviewer origins

Reviewer origin	N reviewers	% of total	N reviews	% of total	N reviews p/person	N of origins	Resolution
Florida	61,308	24.4%	921,231	42.6%	15.03	67	county
Domestic	143,978	57.4%	947,989	43.8%	6.58	54	state
International	45,558	18.2%	293,029	13.6%	6.43	185	nation
Total	250,844	100%	2,162,249	100.0%	8.62	306	

The identified top home origins of Floridian, domestic and international visitors are presented in Tables 5.2, 5.3 and 5.4 respectively. Note that the data includes only English speaking visitors as Spanish and Portuguese data was not collected at the time Task 5 started. Preliminary analysis of Spanish and Portuguese data (section 2.1.4) concluded minor effect on Floridian and domestic data segments; international segment needs updating when relevant data is fully collected.

Table 5.2 Top origin nations

#	Nation	N reviews
1	UK	133,559
2	Canada	63,878
3	Brazil	7,130
4	Australia	7,077
5	Ireland	5,419
6	Germany	5,161
7	Netherlands	5,104
8	Sweden	3,527
9	France	3,426
10	Switzerland	3,381
11	Italy	3,265
12	Mexico	2,958
13	Spain	2,751
14	Argentina	2,622
15	Norway	2,562
16	Denmark	1,758
17	Belgium	1,646
18	Trinidad & Tobago	1,529
19	New Zealand	1,448
20	Bahamas	1,440

Table 5.3 Top origin states

State	N reviews
NY	82,422
GA	71,622
PA	54,314
OH	50,664
IL	48,794
NJ	43,636
TX	42,342
NC	41,422
MI	38,288
MA	36,929
TN	35,180
VA	32,912
CA	27,800
IN	26,466
SC	26,277
MD	23,042
WI	21,059
MO	20,196
AL	19,967
MN	19,357

Table 5.4 Top origin counties

County	N reviews
Orange	72,941
Palm beach	71,740
Broward	69,387
Miami-Dade	64,569
Hillsborough	62,606
Pinellas	57,209
Lee	50,169
Sarasota	48,620
Brevard	37,649
Duval	35,168
Volusia	28,186
Collier	27,172
Polk	23,189
Pasco	17,460
Lake	17,104
Marion	15,818
St. Johns	14,905
Seminole	14,662
Alachua	14,340
Leon	13,468

5.3.1.2 Data preparation

For each user of TripAdvisor, their review records compose a trajectory of visitation footprints during the observed time frame, and their trip segments can be identified with the following procedures:

Step 1. Split trips

Based on statistics of the average length of stay of different groups of visitors from Visit Florida annual report (Floridians less than 2 nights, domestic visitors less than 7 nights and international visitors less than 14 nights – see Visit Florida, 2019), we split the consecutive trajectory of review records of each individual into separate trips assuming their trip durations aligning with statistical stay length. Any two review records with a time gap longer than the threshold of the respective typical stay were regarded as two separate trips.

Step 2. Identify the days of the trip

In a multiple-day trip, we marked the first day as the *'arrival day'*, which contains an O-D trip from visitor's origin to the destination. Likewise, the last day is the *'leaving day'* of the trip to generate a D-O trip from the last destination back to the home location. Any day except the previous two was regarded as the *'within-trip day'*.

Step 3. Filter out short-distance trips

According to the National Tourism Resources Review Commission and the U.S. Travel Association definition of a tourist, a tourist travels at least 50 miles' one-way distance (McIntosh et al. 1998). Therefore, all trips and within-trip relocations shorter than 50 miles were filtered out.

Step 4. Generate trip record

(1) Trip generation for Floridians.

For Floridians, if there is only one visited zone (over 50 miles from home) on the 'arrival day', that visited zone is labeled as destination and the corresponding vector forms an O-D trip record. If there are multiple visited locations within one trip, the farthest from the origin visited zone is regarded as the final destination and the corresponding vector forms an O-D trip record. The D-O trip record is defined in a similar way for the 'leaving day' of the trip. Any other movement between TAZs are marked as D-D trips (within destination).

(2) Trip generation for domestic and international visitors.

We assume that the majority of the out-state visitors use airplanes as their primary transport mode. For long-distance visitors (domestic and international visitors), if there is only one visited zone on the 'arrival day', that visited zone is labeled as destination and the corresponding vector forms an O-D trip record. If there are multiple visited locations within one trip, the closest zone to a major airport is regarded as the first stop of the trip and the corresponding vector forms an O-D trip record. The D-O trip record is defined in a similar way for the 'leaving day' of the trip. Any other movement between TAZs are marked as D-D trips (within destination).

Step 5. Generate OD, DO and DD matrices

The OD, DO and DD trips records were aggregated to generate respective OD, DO and DD matrices. Notice that the OD and DO matrices are directed asymmetric matrices while the DD matrix is an undirected symmetric matric. The OD, DO and DD matrices are described in Table 5.5.

Table 5.5 Summary of the travel flow matrices (social media data)

	OD /DO matrix	DD matrix
Floridians	• Primary data structure in this report: county*county level (67*67 matrix) • Alternative data structure: county*tract level (67*3458 matrix)	• Primary data structure in this report: county*county level (67*67 matrix) • Alternative data structure: tract *tract level (3458*3458 matrix)
Domestic	state * county level (54*67 matrix)	county*county level (67*67 matrix)
International	nation * county level (185*67 matrices)	county*county level (67*67 matrix)

5.3.1.3 Overview of travel patterns

From the trip generation and OD matrices of Floridian, domestic and international visitors, we can depict the basic travel flow patterns of those three groups. Again, note that the data includes only English speaking visitors as Spanish and Portuguese data was not collected at the time Task 5 started. Preliminary analysis of Spanish and Portuguese data concluded minor effect on Floridian and domestic data segments; international segment needs updating when relevant data is fully collected.

The origins

The distribution of the origin of the visitors was highly resembling the distribution of the review origins. Tables 5.6 – 5.8 represent distribution of origins for international, domestic, and Florida travelers.

Table 5.6 Top origin countries of international visitors

#	country	# of review	# of trips
1	UK	133,559	40,296
2	Canada	63,878	23,150
3	Brazil	7,130	2,707
4	Australia	7,077	2,398
5	Ireland	5,419	1,936
6	Germany	5,161	1,706
7	Netherlands	5,104	1,449
8	Sweden	3,527	914
9	France	3,426	1,198
10	Switzerland	3,381	997
11	Italy	3,265	1,314
12	Mexico	2,958	1,173
13	Spain	2,751	1,005
14	Argentina	2,622	1,097
15	Norway	2,562	760
16	Denmark	1,758	515
17	Belgium	1,646	522
18	Trinidad and Tobago	1,529	664
19	New Zealand	1,448	448
20	Bahamas	1,440	707

Table 5.7 Top origin states of domestic visitors

#	state	# of review	# of trips
1	NY	82422	34242
2	GA	71622	30639
3	PA	54314	19783
4	OH	50664	18393
5	IL	48794	18124
6	TX	42342	16726
7	NJ	43636	16661
8	NC	41422	15850
9	MA	36929	14162
10	MI	38288	13589
11	TN	35180	12853
12	VA	32912	12605
13	CA	27800	11633
14	SC	26277	10457
15	IN	26466	9280
16	MD	23042	8943
17	AL	19967	7955
18	WI	21059	7237
19	CT	18940	7194
20	MO	20196	7114

Table 5.8 Top origin counties of Floridian visitors

#	Origin	Name	# of Trips	Population
1	12099	Palm beach	15309	1320134
2	12011	Broward	15048	1748066
3	12095	Orange	14447	1145956
4	12057	Hillsborough	13325	1229226
5	12086	Miami-dade	12986	2496435
6	12103	Pinellas	10156	916542
7	12115	Sarasota	9578	379448
8	12071	Lee	9218	618754
9	12031	Duval	8711	864263
10	12009	Brevard	6898	543376
11	12127	Volusia	4724	494593
12	12105	Polk	4618	602095
13	12021	Collier	4587	321520
14	12001	Alachua	4249	247336
15	12073	Leon	4017	275487
16	12083	Marion	3773	331298
17	12101	Pasco	3273	464697
18	12111	St. Lucie	3155	277789
19	12109	St. Johns	3134	190039
20	12069	Lake	3108	297052

The destinations

Orlando was the top destination for all three type of visitors. The ranking of other top destinations however differed. Floridian visitors preferred visiting Everglades and St. Petersburg-Clearwater area. For domestic visitors, Miami and Ft. Lauderdale areas were ranked #2 and 3, closely followed by Everglades and St. Petersburg-Clearwater areas. Finally, international visitors traveled to Miami area with much smaller representation of other Florida areas. For detail, see Tables 5.9, 5.10 Overall, the counties with major cities and popular attractions were universally favored by all three traveler types, which is evidenced by high correlation between the destination visitations of each visitor group (Figure 5.3).

Table 5.9 Top destinations for int'l, domestic, and Floridian visitors, absolute numbers

Destination	County Name	Int'l	Domestic	Floridian	Total	General area
12095	Orange	38,342	83,612	32,014	153,968	Orlando
12086	Miami-Dade	20,168	35,167	9,787	65,122	Miami
12011	Broward	7,354	25,767	8,149	41,270	Fort Lauderdale
12087	Monroe	3,231	23,397	14,196	40,824	Everglades
12103	Pinellas	4,667	24,344	11,055	40,066	St. Pete-Clearwater
12057	Hillsborough	2,307	14,820	8,779	25,906	Tampa
12071	Lee	1,896	14,784	8,871	25,551	Fort Myers
12099	Palm beach	2,320	15,239	7,145	24,704	Palm beach
12021	Collier	2,674	11,655	5,488	19,817	Naples
12109	St. Johns	844	7,966	9,947	18,757	St. Augustine
12127	Volusia	1,191	10,105	7,101	18,397	Daytona
12097	Osceola	3,386	7,945	5,746	17,077	Kissimmee
12115	Sarasota	1,720	9,669	5,610	16,999	Sarasota
12009	Brevard	1,367	8,868	4,907	15,142	Cape Canaveral
12031	Duval	577	7,952	5,804	14,333	Jacksonville
12005	Bay	250	7,137	1,664	9,051	Panama City
12081	Manatee	931	4,806	2,264	8,001	Manatee county
12091	Okaloosa	208	5,796	1,612	7,616	Destin
12033	Escambia	217	4,921	1,729	6,867	Pensacola
12105	Polk	708	2,842	2,978	6,528	Lakeland
12001	Alachua	153	1,952	3,761	5,866	Gainesville
12073	Leon	193	1,921	3,627	5,741	Tallahassee

Table 5.10 Top destinations for international, domestic, and Floridian visitors (percentages)

Destination	County Name	Int'l	Domestic	Floridian	Total	General area
12095	Orange	39.2%	23.0%	17.0%	23.7%	Orlando
12086	Miami-Dade	20.6%	9.7%	5.2%	10.0%	Miami
12011	Broward	7.5%	7.1%	4.3%	6.4%	Fort Lauderdale
12087	Monroe	3.3%	6.4%	7.5%	6.3%	Everglades
12103	Pinellas	4.8%	6.7%	5.9%	6.2%	St. Pete-Clearwater
12057	Hillsborough	2.4%	4.1%	4.7%	4.0%	Tampa
12071	Lee	1.9%	4.1%	4.7%	3.9%	Fort Myers
12099	Palm beach	2.4%	4.2%	3.8%	3.8%	Palm beach
12021	Collier	2.7%	3.2%	2.9%	3.1%	Naples
12109	St. Johns	0.9%	2.2%	5.3%	2.9%	St. Augustine
12127	Volusia	1.2%	2.8%	3.8%	2.8%	Daytona
12097	Osceola	3.5%	2.2%	3.0%	2.6%	Kissimmee
12115	Sarasota	1.8%	2.7%	3.0%	2.6%	Sarasota
12009	Brevard	1.4%	2.4%	2.6%	2.3%	Cape Canaveral
12031	Duval	0.6%	2.2%	3.1%	2.2%	Jacksonville
12005	Bay	0.3%	2.0%	0.9%	1.4%	Panama City
12081	Manatee	1.0%	1.3%	1.2%	1.2%	Manatee county
12091	Okaloosa	0.2%	1.6%	0.9%	1.2%	Destin
12033	Escambia	0.2%	1.4%	0.9%	1.1%	Pensacola
12105	Polk	0.7%	0.8%	1.6%	1.0%	Lakeland
12001	Alachua	0.2%	0.5%	2.0%	0.9%	Gainesville
12073	Leon	0.2%	0.5%	1.9%	0.9%	Tallahassee

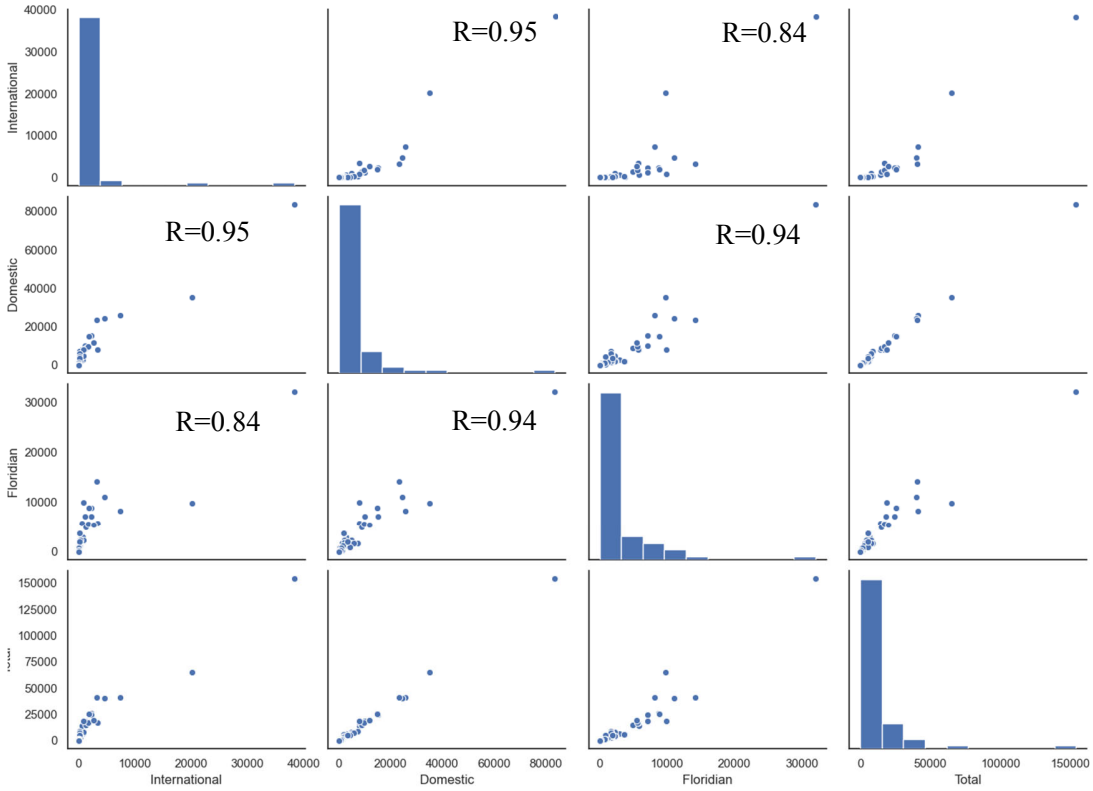


Figure 5.3 Bivariate correlation matrix of total number of visits to different Florida destinations, for Floridian, domestic, and international tourists

5.3.1.4 Preliminary data from Spanish and Portuguese language reviews on TripAdvisor

The analysis of origin distribution of international visitors from English language TripAdvisor reviews presented above exhibited significant deviation from Visit Florida in under-representation of two major origin markets in South America, specifically, Brazil and Argentina. We, therefore, collected two additional datasets from TripAdvisor Spanish and Portuguese websites as complementary data sources. The approaches for data collection and methods in the data process were identical to those applied in the English dataset, and the summary of the original and prepared data is shown below in Tables 5.11 – 5.14.

Table 5.11 TripAdvisor Spanish language dataset summary

Origin	N reviewers	% of total	N reviews	% of total	N reviews p/person	N of origins	Resolution
Florida	1,607	4.1%	10,058	5.5%	6.26	1	state-level
Domestic	1,482	3.8%	5,411	3.0%	3.65	59	state-level
Int'l	36,059	92.1%	166,477	91.5%	4.62	79	country-level
Total	39,148	100.0%	181,946	100.0%	4.64	139	

Table 5.12 Top origin countries of intentional visitors (Spanish language dataset)

#	Country	N reviews	%of int'l	N reviewers	%of int'l	N reviews p/person	N Trips	%of int'l
1	Argentina	93,522	56.18 %	18,249	50.61%	5.12	33,575	53.75%
2	Chile	14,013	8.42%	3,100	8.60%	4.52	5,053	8.09%
3	Mexico	9,397	5.64%	2,591	7.19%	3.63	3,824	6.12%
4	Colombia	9,480	5.69%	2,386	6.62%	3.97	3,876	6.20%
5	Spain	8,993	5.40%	2,383	6.61%	3.77	3,383	5.42%
6	Peru	6,349	3.81%	1,444	4.00%	4.40	2,423	3.88%
7	Uruguay	5,323	3.20%	1,118	3.10%	4.76	1,957	3.13%
8	Venezuela	5,020	3.02%	1,196	3.32%	4.20	2,141	3.43%
9	Ecuador	3,036	1.82%	806	2.24%	3.77	1,344	2.15%
10	Costa Rica	2,867	1.72%	687	1.91%	4.17	1,222	1.96%
11	Panama	1,503	0.90%	365	1.01%	4.12	639	1.02%
12	Dominican	1,118	0.67%	272	0.75%	4.11	467	0.75%
13	Guatemala	989	0.59%	267	0.74%	3.70	449	0.72%
14	Bolivia	768	0.46%	160	0.44%	4.80	293	0.47%
15	Paraguay	673	0.40%	173	0.48%	3.89	314	0.50%

Table 5.13 TripAdvisor Portuguese language dataset summary

Origin	N reviewers	% of total	N reviews	% of total	N reviews p/person	N of origins	Resolution
Florida	686	1.6%	8,774	3.4%	12.79	1	state-level
Domestic	728	1.7%	4,499	1.7%	6.18	42	state-level
Int'l	42,602	96.8%	244,651	94.9%	5.74	77	country-level
Total	44,016	100.0%	266,072	100.0%		120	

Table 5.14 Top origin countries of intentional visitors (Portuguese language dataset)

#	Country	N reviewers	%of int'l	N reviews	%of int'l	N reviews p/person	N Trips	%of int'l
1	Brazil	41,170	96.64%	237,685	97.15%	5.77	71,654	96.70%
2	Portugal	511	1.20%	1,991	0.81%	3.90	725	0.98%
3	Canada	136	0.32%	760	0.31%	5.59	272	0.37%
4	UK	103	0.24%	597	0.24%	5.80	216	0.29%
5	Spain	79	0.19%	385	0.16%	4.87	130	0.18%
6	Italy	59	0.14%	278	0.11%	4.71	93	0.13%
7	Argentina	56	0.13%	319	0.13%	5.70	106	0.14%
8	France	41	0.10%	280	0.11%	6.83	77	0.10%
9	Australia	38	0.09%	194	0.08%	5.11	64	0.09%
10	Germany	35	0.08%	132	0.05%	3.77	46	0.06%
11	Mexico	30	0.07%	110	0.04%	3.67	52	0.07%
12	Netherlands	29	0.07%	185	0.08%	6.38	61	0.08%
13	Chile	26	0.06%	106	0.04%	4.08	38	0.05%
14	Ireland	24	0.06%	104	0.04%	4.33	46	0.06%
15	Switzerland	18	0.04%	178	0.07%	9.89	44	0.06%

The data retrieved from Spanish language TripAdvisor showed that Argentina is the largest origin market to Florida (53.75% of total visit trips), followed by Chile, Mexico, Colombia, Spain and Peru, a majority of South American markets identified in this regard. Additionally, from the Portuguese language dataset,

Brazil was recognized as the dominant origin market (96.70% of total international trips). This complementary data should provide additional insights on the landscape of the Florida international tourism market. Note a small number of domestic and Floridian travelers in both datasets, which supports our earlier assumption that English language TripAdvisor data is representative of domestic and Floridian visitors and Spanish and Portuguese data can be disregarded for these segments.

Note that the temporal distributions of both Spanish and Portuguese datasets (Figures 5.4 and 5.5, accordingly) had the number of reviews decreasing starting in 2016 – 2017. Our personal contacts in Brazil and Mexico did not confirm our initial hypothesis that this decrease reflects migration of users to other platforms; instead, they suggested that it is reflective of the initial spike in travel activity followed by a lesser interest towards the destination, especially after the 2015 economic crises in those countries. An additional factor could be related to differences in data collection timing with English language dataset (all sets were using same properties database, which could change by the time of Portuguese and Spanish data collection). Finally, the total visit basis of Spanish and Portuguese seemed to be inconsistent with English, as they appeared to be disproportionately larger than that of English-speaking reviewers. Overall, further research on Spanish and Portuguese datasets is needed prior to merging them to English language international visitor dataset. Because of that, only Floridian and domestic English-speaking visitor data were used in the follow-up validations.

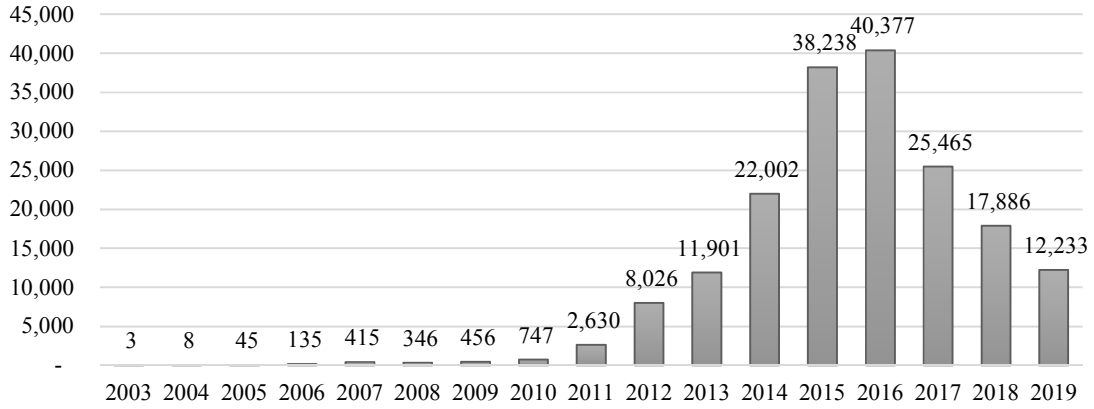


Figure 5.4 Temporal distribution of collected reviews (Spanish language)

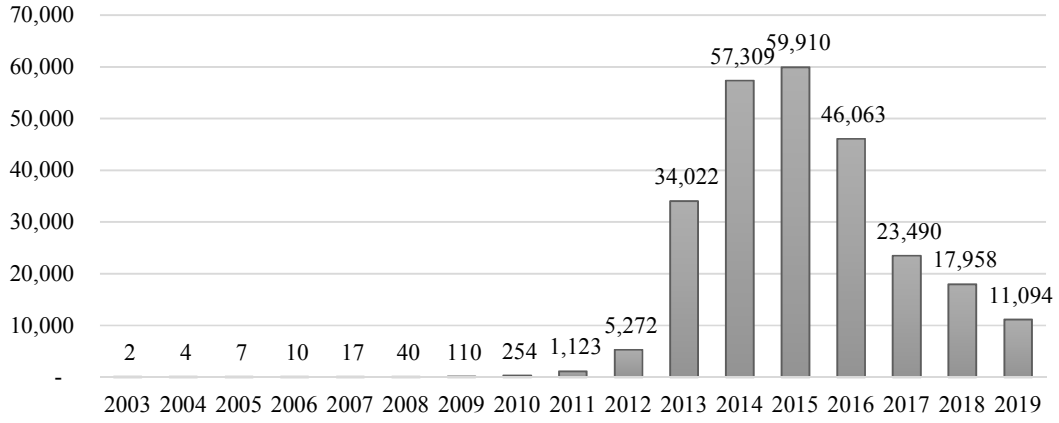


Figure 5.5 Temporal distribution of collected reviews (Portuguese language)

5.3.2 AirSage cellphone data

5.3.2.1 Original data snapshot

The original mobile phone data was provided by AirSage (www.airsage.com) with the following configurable parameters:

(1) data timeline: the year and month of study. The data was collected from October 2018 and September 2019.

(2) day aggregations: selected days in the week across which the study area movement is to be aggregated. The configuration in the data was in four categories: Total Weekday (Tuesday – Thursday), Friday, Saturday and Sunday.

(3) time aggregations: The hours within a day across which the study area movement is to be aggregated. 24 hours and/or time of dayparts to reflect peak travel periods.

(4) trip purpose: this configuration aggregates trips by predicted travel purpose, including the following classes: Home to Work (HW), Work to Home (WH), Home to Other (HO), Other to Home (OH), Work to Other (WO), Other to Work (WO) and Other to Other (OO).

The original data was organized as a list of edges representing possible trips to and between Florida census blocks, for different day/time aggregations and trip purposes. Spatially, each record inside the dataset includes trip "origin" and "destination" attributes {US Census Block, TAZ ID}, and also the device home location {US Census Block}. A sample of the packaged trip matrix links is shown below in Table 5.15.

Table 5.15 Original data sample

Origin Zone	Destination Zone	Home Zone	Day Aggregation	Type	Time of day	Count	Month
120330040004	120539703002	120330040004	Mon to Thu	HW	H08:H09	1.931982	201810
010030106001	120030106001	120330026022	Fri	WO	H16:H17	0.543333	201810
120330026021	010030114052	120050023002	Sun	OO	H23:H24	0.228322	201810

The trips between a pair of analysis TAZs are segmented based on the optional attributes requested by the analyst (Day aggregation, Trip type, Time of the day, and Month). Table 5.16 describes each of those attributes in detail.

Table 5.16 Trip matrix attributes

Attribute Field	Description	Sample Value
Origin Zone	The TAZ_ID where the trips began (in census block)	120330040004
Destination Zone	The TAZ_ID where the trips ended (in census block)	120030106001
Home Zone	The TAZ_ID where the trip-maker home located (in census block)	120330026022
Day Aggregation	The day aggregation for which the trips are being reported	Mon to Thu
Type	Trip types by travel purpose	HW
Time of day	The daytime part aggregation for which trips are being reported	H15: H19
Count	The number of estimated extrapolated person-trips made between the analyses TAZs “Origin Zone” and “Destination one” with the above attributes.	1.931982
Month	The year and month when the data collected	201810

The original dataset contains over 8,000,000,000 records, stored in 12 monthly files. Files are archived in the folder named Task_5_Cellphone_Data\5.0_PhotoData_Raw.

5.3.2.2 Data preparation

Step 1. Select tourism-related trips

We assumed that the Other type of trips with at least 50-mile distance are for tourism purpose and hence selected HO, OH, and OO trip types for further processing. HO and OH trip types represent travels between one’s home and travel destinations, while OO trip segments are the movements between different segments of a trip.

Step 2. Aggregate trips

The trips were aggregated based on the origin census tract, destination census tracts, and home census tracts, regardless of their travel times and weekdays. Note that census tract attributes are easily retrieved from census block attributes in respective TAZ_ID fields by truncating the last digit. For example, the corresponding census tract of census block 120330026022 is 12033002602.

Trip counts were summated according to individual Origin, Destination, and Home TAZs, and the rearranged aggregation data samples appear to be as follows (Table 5.17):

Table 5.17 Aggregated trips samples

O Tract	D Tract	H Tract	Count	Type	Month
12003010100	12053120540	12003010100	13948.987624	HO	201810
12053024400	12003010200	12003010200	23854.975887	OH	201810
12001000300	12086000200	08230283200	72.891743	OO	201810

Step 3. Generate OD and DD trip databases

Databased were separately generated and restored based on trip types:

(1) OD (origin to destination) database represents all trips from Home to Other Place (designated HO in AirSage data), aggregated based on origin and destination census tracts. The trip makers in the OD database were mostly Floridians. A filtering process was applied to retain the tourism trip as defined by at least 50-mile distance with all other trips discarded. The trips were then aggregated on a monthly and annual basis hence creating two different databases (Table 5.18). The larger monthly base dataset was used for validation and temporal analysis. Note that the DO trips mirror the OD database; accordingly, no separate datasets were created.

(2) DD (destination to destination) database represents all trips from Other to Other Places (designated OO in AirSage data). These trips were assumed to represent segments of a single trip. In contrast to OD database, the DD database include not only Floridians, but also domestic visitors with home locations in other states. A similar filtering process was applied to retain only the trips of at least 50 miles (Table 17).

Table 5.18 OD and DD databases

	N records	Time field	Data file
Database OD 1*	1,676,426	month variable aggregated	od_fulllist_50mile_mon.csv
Database OD 2*	3,073,371	month variable kept	od_fulllist_50mile.csv
Database DD**	10,082,451	month variable kept	dd_fulllist_50mile_mon.csv

* Data storage directory: Task_5_Cellphone_Data\5.2_Travel_Flow\od_list

** Data storage directory: Task_5_Cellphone_Data\5.2_Travel_Flow\dd_list

Step 4. Dataset enhancement

Unlike OD or DO databases where the Home TAZ is identical to either the Origin or Destination TAZ, the DD trip records were constructed based on three different TAZs: Origin, Destination, and Home, inevitably generating a much larger volume of records (over 10,000,000) the majority of which representing very small number of trips (mean = 2.36, median = 1.5). We, therefore, aggregated the DD trip records on a level of Home States. Hence, the DD records were reorganized to represent the Origin Census Tract, Destination Census Tract and the Home State (see a sample of data in Table 5.19). The reduced DD dataset contains a total of 6,014,888 DD trip count records (mean=3.96; median=2.16).

Table 5.19 A sample of DD trips aggregated based on the home state of a traveler

O Tract	D Tract	H State	Count	Month
12001000302	12067960200	12	7.378022	201903
12001000302	12069030306	12	0.894590	201903
12001000302	12069030411	12	1.519209	201903

We generated several variants of the database depending on the different operations for time fields and visitor origins. Table 5.20 shows the information of the database variants and their features*.

Table 5.20 DD databases

Name	N records	monthly /annual	Home zone resolution	Visitors	Mean N trip	Data file
Database 1	10,082,451	month	Tract	All	2.36	dd_fulllist_50mile_mon.csv
Database 2	6,014,888	month	State	All	3.96	dd_fulllist_50mile_agg_state_mon.csv
Database 3	3,614,707	annual	State	All	6.59	dd_fulllist_50mile_agg_state_.csv
Database 4	2,238,975	annual	State	Floridian	7.43	dd_fulllist_50mile_agg_state_floridian.csv
Database 5	1,375,732	annual	State	Domestic	5.21	dd_fulllist_50mile_agg_state_domesitc.csv

* Data storage directory: Task_5_Cellphone_Data\5.2_Travel_Flow\dd_list

Step 5. Generate OD, DO and DD matrices

The OD, DO and DD trips records were aggregated to generate respective OD, DO and DD matrices. Note that the OD and DO matrices mostly represent Floridians with small representation of Georgia and Alabama residents living near the border of Florida. The DD matrices, on the other hand, contain travel flows generated by both Floridians and out-state domestic visitors. The data structures of the OD, DO and DD matrices are summarized as in Table 5.21.

Table 5.21 Summary of the travel flow matrices (cellphone data)

	OD /DO matrix	DD matrix
Floridians	• Primary data structure: tract *tract level (3458 *3458 matrix) • Alternative data structure: flexible in county*county, county*tract or tract*county level	• Primary data structure: tract *tract level (3458 *3458 matrix) • Alternative data structure: flexible
Domestic	/	• Primary data structure: tract *tract level (3458 *3458 matrix) • Alternative data structure: flexible

5.3.2.3 Overview of travel patterns

(1) The origins

Origins of Floridians

The majority of the trips in OD/DO databases were made by Floridians (19,765,100 trip counts, 94.4%), with a small percentage of adjacent Alabama (2.42%) and Georgia (1.29%) residents. The origins of Floridians visitors are shown in Table 5.22 and Figure 5.6.

Table 5.22 Top origin census tracts (left) and counties (right) of Floridian visitors

#	Origin Tract	N Trips	County	Origin County	Trip Counts	County Name	Population
1	12057002300	75,609	Hillsborough	12099	1,535,091	Palm Beach	1320134
2	12115002712	45,670	Sarasota	12057	1,441,702	Hillsborough	1229226
3	12057013503	45,018	Hillsborough	12086	1,213,415	Miami-Dade	2496435
4	12057011902	36,242	Hillsborough	12031	1,166,759	Duval	864263
5	12111382106	35,316	St. Lucie	12095	1,139,352	Orange	1145956
6	12083001004	33,911	Marion	12011	872,741	Broward	1748066
7	12111382108	32,950	St. Lucie	12071	833,261	Lee	618754
8	12021011202	29,584	Collier	12103	819,085	Pinellas	916542
9	12095015103	29,436	Orange	12105	761,525	Polk	602095
10	12095017108	27,707	Orange	12009	677,390	Brevard	543376
11	12017451601	26,730	Citrus	12083	560,229	Marion	331298
12	12111382111	26,621	St. Lucie	12111	542,193	St. Lucie	277789
13	12109020902	26,389	St. Johns	12127	514,058	Volusia	494593
14	12057000201	26,152	Hillsborough	12101	445,792	Pasco	464697
15	12051000200	26,052	Hendry	12021	444,507	Collier	321520
16	12007000200	25,684	Bradford	12115	414,533	Sarasota	379448
17	12083002502	25,569	Marion	12073	404,376	Leon	275487
18	12021011102	24,526	Collier	12001	360,658	Alachua	247336
19	12095018900	23,382	Orange	12081	351,702	Manatee	322833
20	12111382113	23,075	St. Lucie	12117	334,855	Seminole	422718

In general, the areas with higher trip generations were those with higher population, such as Miami metropolitan area, Tampa Bay area and Orlando metropolitan area; those areas were all ranking among the top origins. Nevertheless, the trip generations were not necessarily merely determined by the population of the origin, as the correlation between population and trip generation was 0.76. More factors such as socio-economic indicators influencing the trip generations are worthy of investigation in this regard.

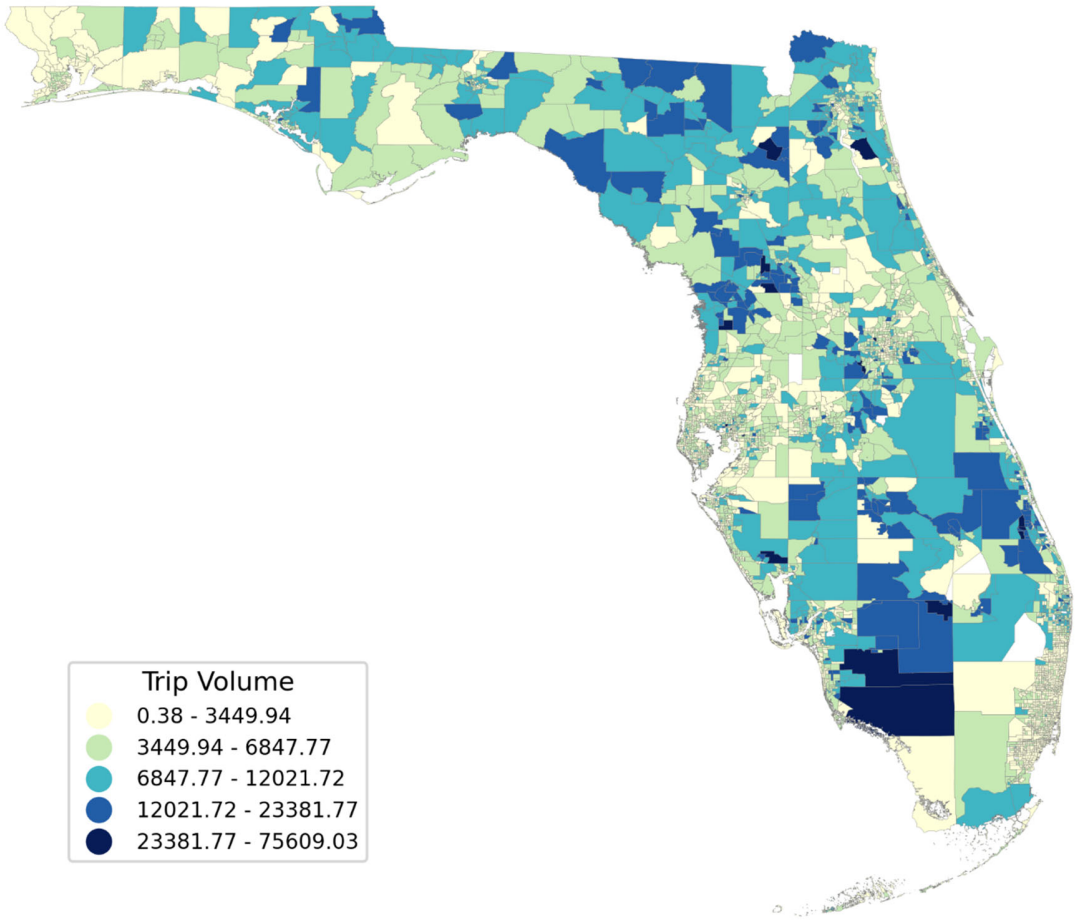


Figure 5.6 Heatmap of origin tracts with trip generations

Origins of domestic visitors

The DD matrix of domestic visitors shows their in-state movements. The state with the largest visitor generation was Georgia, followed by New York, Alabama, California, Texas and North Carolina. The top-ranking states statistics and the overall distribution are shown in Table 5.23 and Figure 5.7.

Table 5.23 Top origin states of domestic visitors

Origin	Name	# Trips	% of Domestic

13	Georgia	834,620	11.6%
36	New York	661,042	9.2%
06	California	368,015	5.1%
48	Texas	364,626	5.1%
37	North Carolina	326,272	4.6%
34	New Jersey	277,725	3.9%
39	Ohio	271,976	3.8%
42	Pennsylvania	270,474	3.8%
01	Alabama	266,281	3.7%
51	Virginia	266,275	3.7%
17	Illinois	261,418	3.6%
45	South Carolina	228,062	3.2%
25	Massachusetts	203,044	2.8%
26	Michigan	200,719	2.8%
47	Tennessee	175,951	2.5%
22	Louisiana	158,392	2.2%
18	Indiana	150,297	2.1%
24	Maryland	141,362	2.0%
21	Kentucky	132,887	1.9%

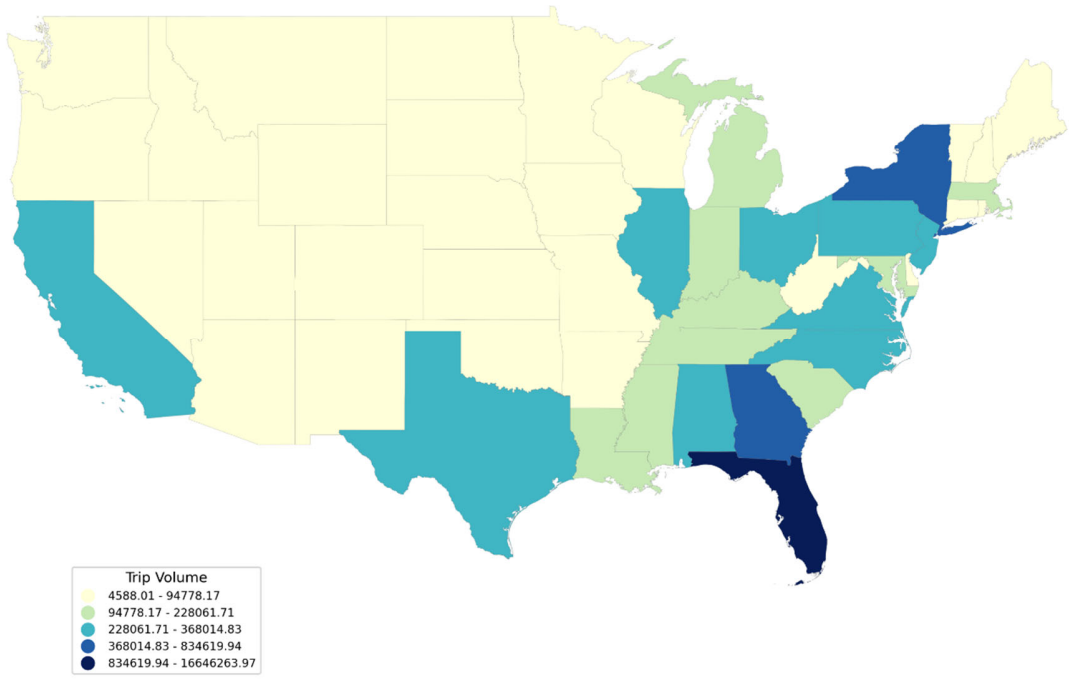


Figure 5.7 Top origin states in term of trip generation

(2) The destinations

Destinations of Floridians

Table 5.24 and Figure 5.8 summarize information on the destination visitations of the Floridians from the OD/DO databases. The most popular destinations were located in Orange, Miami-Dade and Hillsborough counties and belonged to Orlando, Miami and Tampa metropolitan areas.

Table 5.24 Top destination census tracts and counties of Floridian visitors

#	Destination Tract	N Trips	County
1	12086005103	311,934	Miami-Dade
2	12095016906	269,731	Orange
3	12095015103	220,786	Orange
4	12095014200	193,991	Orange
5	12095010200	185,624	Orange
6	12095014807	180,902	Orange
7	12095012304	170,915	Orange
8	12057011902	159,733	Hillsborough
9	12095017001	157,470	Orange
10	12095012403	156,178	Orange
11	12095017103	152,876	Orange
12	12057013503	151,174	Hillsborough
13	12095013605	144,300	Orange
14	12086001003	137,103	Miami-Dade
15	12095013511	132,103	Orange
16	12095014608	116,347	Orange
17	12097041900	112,403	Osceola
18	12086009905	108,514	Miami-Dade
19	12095013701	103,669	Orange
20	12095990000	96,408	Orange

Destination County	Trip Counts	County Name
12095	4,222,721	Orange
12086	2,573,859	Miami-Dade
12057	1,878,658	Hillsborough
12011	911,854	Broward
12099	690,837	Palm Beach
12105	677,824	Polk
12031	622,775	Duval
12103	549,218	Pinellas
12097	534,319	Osceola
12117	534,262	Seminole
12127	470,827	Volusia
12071	458,814	Lee
12083	455,840	Marion
12009	441,198	Brevard
12081	346,670	Manatee
12115	313,698	Sarasota
12001	298,339	Alachua
12069	296,098	Lake
12021	278,867	Collier
12111	274,508	St. Lucie

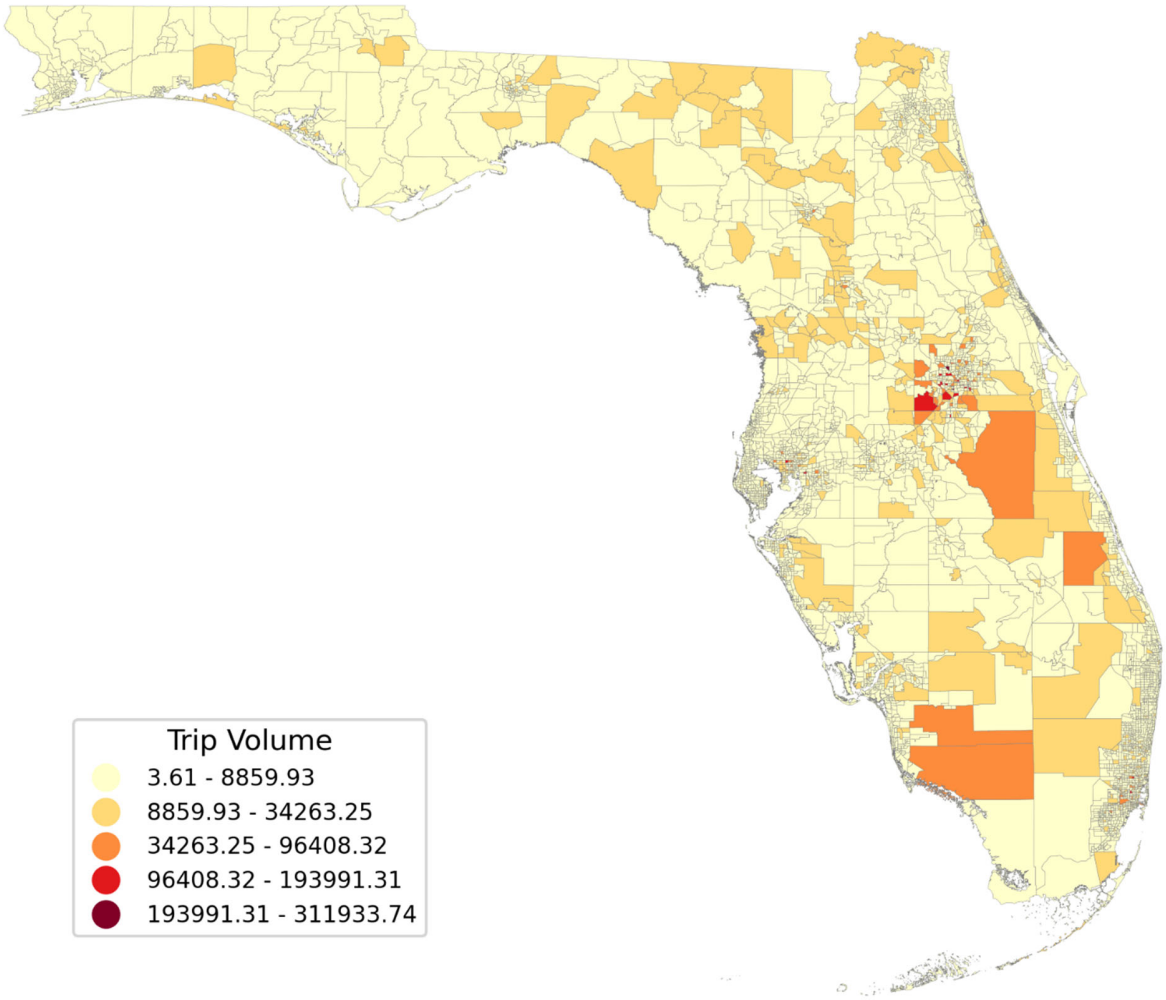


Figure 5.8 Heatmap of destination tracts with visitations

Destination of domestic visitors

The in-state movements of domestic visitors were represented by the DD matrix. The 'destination' in a single travel flow is not necessarily the actual destination of that domestic visitor but represents only one trip segment of their entire trip to Florida. Therefore, we did not summarize the destination statistics of domestic visitors similar to that for Floridians.

(3) Travel networks

OD Travel network of Floridians

The OD network was easily retrieved based on the origins and destinations of Floridians. Table 5.25 and Figure 5.9 show top travel flows with the busiest travel flows identified between the following locations: Palm Beach to Miami, Tampa to Orlando and Jacksonville to Orlando.

Table 5.25 Floridian top OD travel flows between counties

O County	D County	Travel Direction	N Trips
12099	12086	12099-12086	806,358
12057	12095	12057-12095	699,457
12031	12095	12031-12095	339,657
12103	12095	12103-12095	326,103
12071	12086	12071-12086	325,180
12031	12057	12031-12057	295,452
12095	12057	12095-12057	284,670
12086	12099	12086-12099	234,129
12009	12095	12009-12095	227,097
12086	12095	12086-12095	224,680
12083	12095	12083-12095	222,944
12111	12086	12111-12086	173,937
12099	12095	12099-12095	171,169
12021	12086	12021-12086	167,268
12101	12095	12101-12095	145,765
12011	12095	12011-12095	143,016
12127	12095	12127-12095	129,908
12001	12095	12001-12095	110,988
12111	12011	12111-12011	106,188
12105	12103	12105-12103	106,016
12071	12011	12071-12011	104,958
12095	12103	12095-12103	101,965

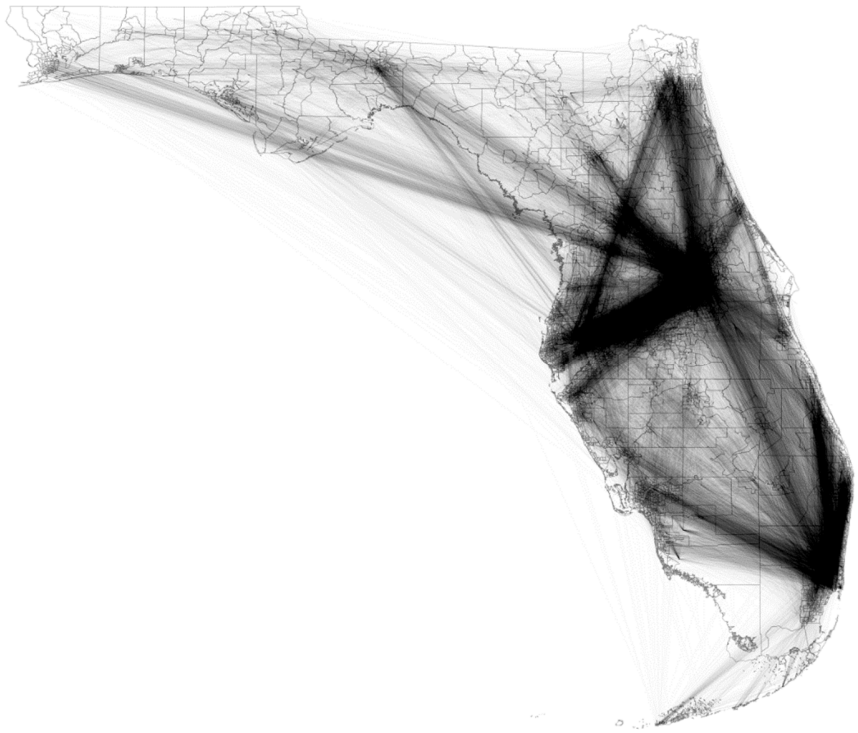


Figure 5.9 OD travel network of Floridians. The visualization represents top 5% of links

DD Travel networks

The DD travel network was constructed based on in-Florida DD travel flows. Each travel link starts from the beginning zone of the travel segment and ends in the final zone. The DD travel network patterns of Floridians and domestic visitors are shown in Figures 5.10 and 5.11 respectively.

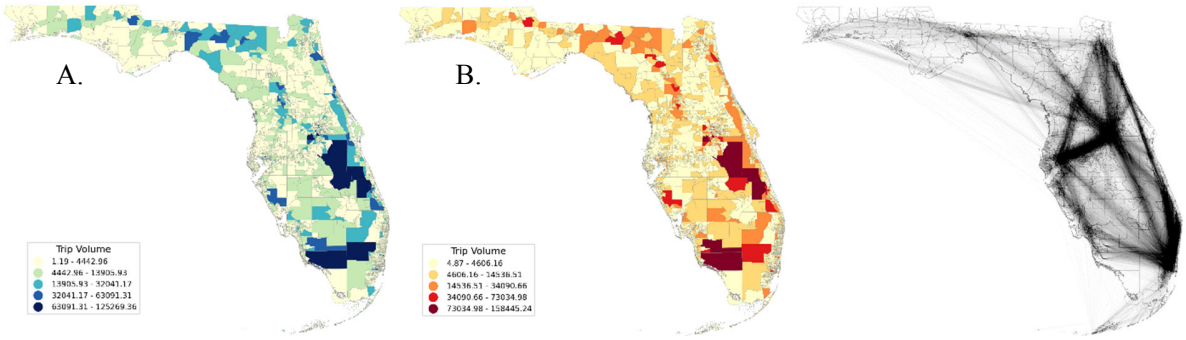


Figure 5.10 In-Florida DD movement of Floridian tourists: A. starting zones; B: ending zones; C: top 5% of trip links

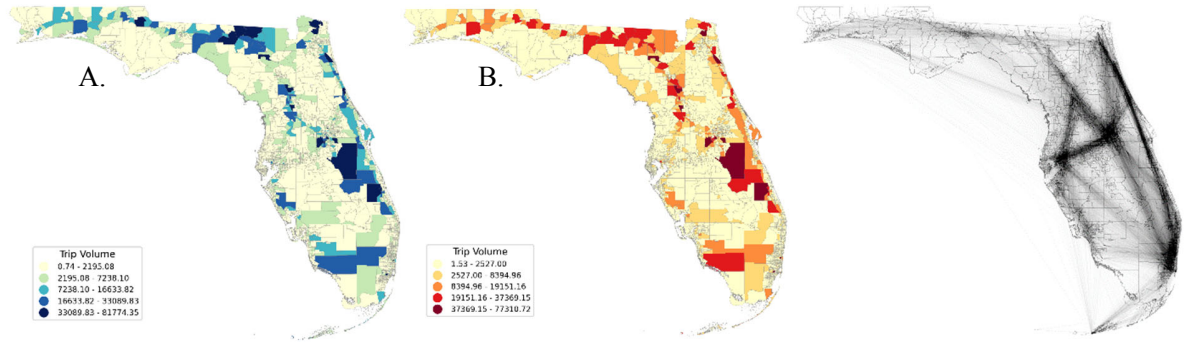


Figure 5.11 In-Florida DD movement of domestic tourists: A. starting zones; B: ending zones; C: top 5% of trip links

Notice that the overall movements of both Floridians and domestic tourists is somewhat aligned with the road network of Florida, as the most frequent starting zones and ending zones are located along the major highways. In addition, the distributions of starting zones and ending zones are similar (Table 5.26), implying that the DD movements largely resulted from reciprocal traffic, with each TAZ receiving and generating similar number of visitors. We conclude that the DD links are poorly reflecting the tourists' motivations in destination choice and concentrate on analysis of OD travel flows are prior to analyzing the DD network.

Table 5.26 Top TAZs in Floridian DD movements

#	Starting Zone	N Trips
1	12097043800	125,269
2	12111382200	116,381
3	12095015103	99,642
4	12095016906	92,881
5	12086005103	86,937
6	12061050904	84,736
7	12011980000	84,238
8	12093910102	83,846
9	12095017001	83,690
10	12111380800	83,059

#	Ending Zone	N Trips
1	12097043800	158,445
2	12111382200	127,246
3	12086005103	95,631
4	12021011102	93,587
5	12095015103	91,527
6	12061050904	90,697
7	12095017001	87,985
8	12095016906	87,951
9	12021011202	85,604
10	12095017103	83,264

5.3.3 VisitFlorida statistics

Original data snapshot

The VISIT FLORIDA research department is researching preferences and travel patterns of Florida's visitors. The annual Florida Visitor Study summary is the premier reference guide for statistics on visitors to Florida (https://www.visitflorida.org/resources/research). These data largely rely on conventional survey tools such as questionnaires and interviews. The released statistics cover from 2009 to the third quarter of 2019 timeframe (the third quarter of 2019 is not released at the point of this study) and includes quarterly statistics on domestic, overseas (international except Canada), and Canadian visitors. The data also provides additional statistics on the top 10 origin countries and the top 15 origin states.

Overview of key statistics

The Visit Florida statistics provides structured data regarding visitors' origins on a state and nation scales, as well as the overall number of visitors to Florida on an annual basis. No detailed destination visitation data are available. Floridian travel data is also not provided. We obtained the annual and seasonal visitation statistics to Florida from 2015 to 2018 (Table 5.27), and the top origin information regarding domestic and international visitors (Tables 5.28, 5.29).

Table 5.27 Overall annual and seasonal visitation to Florida

Year	Quarter	Total Arrivals	YoY	domestic	YoY	Int'l	YoY
2018	Q1	34,771,000	9.0%	30,612,000	10.0%	4,159,000	2.5%
2018	Q2	31,506,000	5.0%	27,968,000	5.7%	3,538,000	-0.8%
2018	Q3	31,261,000	12.2%	28,118,000	14.0%	3,143,000	-1.5%
2018	Q4	29,441,000	2.7%	25,898,000	3.2%	3,543,000	-0.5%
2018	Total	126,979,000	7.2%	112,596,000	8.2%	14,383,000	0.0%
2017	Q1	31,890,000	5.2%	27,833,000	6.1%	4,057,000	-0.9%
2017	Q2	30,017,000	7.1%	26,449,000	7.8%	3,568,000	2.3%
2017	Q3	27,854,000	3.0%	24,663,000	3.8%	3,191,000	-2.4%
2017	Q4	28,662,000	7.0%	25,100,000	8.5%	3,562,000	-2.4%
2017	Total	118,423,000	5.6%	104,045,000	6.5%	14,378,000	-0.9%
2016	Q1	30,321,000	6.5%	26,226,000	9.5%	4,095,000	-9.4%
2016	Q2	28,029,000	5.9%	24,542,000	8.6%	3,487,000	-10.0%
2016	Q3	27,030,000	5.6%	23,761,000	6.3%	3,269,000	0.6%
2016	Q4	26,794,000	3.1%	23,143,000	3.3%	3,651,000	1.7%
2016	Total	112,174,000	5.3%	97,672,000	7.0%	14,502,000	-4.8%
2015	Q1	28,482,000	7.0%	23,960,000	7.4%	4,522,000	4.5%
2015	Q2	26,478,000	8.3%	22,604,000	10.1%	3,874,000	-0.9%
2015	Q3	25,605,000	7.5%	22,356,000	9.3%	3,249,000	-3.6%
2015	Q4	25,990,000	10.1%	22,399,000	12.7%	3,591,000	-3.5%
2015	Total	106,555,000	8.2%	91,319,000	9.8%	15,236,000	-0.6%

Table 5.28 Top 15 origin states of domestic visitors

	State	2018	% of Domestic	2017	% of Domestic	2016	% of Domestic	2015	% of Domestic
1	Georgia	11935176	10.60%	9988320	9.60%	10353232	10.60%	8675305	9.50%
2	New York	10021044	8.90%	8843825	8.50%	9474184	9.70%	8857943	9.70%
3	Texas	4841628	4.30%	5722475	5.50%	5078944	5.20%	5570459	6.10%
4	Ohio	4841628	4.30%	5514385	5.30%	5078944	5.20%	4565950	5.00%
5	Pennsylvania	5742396	5.10%	5514385	5.30%	4883600	5.00%	4657269	5.10%
6	Tennessee	4616436	4.10%	4682025	4.50%	3711536	3.80%	2739570	3.00%
7	New Jersey	4729032	4.20%	4473935	4.30%	4297568	4.40%	3926717	4.30%
8	North Carolina	5292012	4.70%	4057755	3.90%	4785928	4.90%	3470122	3.80%
9	Missouri	3040092	2.70%	3953710	3.80%	2148784	2.20%	1735061	1.90%
10	Illinois	5517204	4.90%	3953710	3.80%	4004552	4.10%	3835398	4.20%
11	Alabama	5404608	4.80%	3641575	3.50%	4004552	4.10%	3470122	3.80%
12	Maryland	2927496	2.60%	3537530	3.40%	2441800	2.50%	2739570	3.00%
13	Michigan	4278648	3.80%	3433485	3.30%	3223176	3.30%	3926717	4.30%
14	Virginia	3265284	2.90%	3017305	2.90%	2344128	2.40%	3196165	3.50%
15	Indiana	3603072	3.20%	2913260	2.80%	2637144	2.70%	3013527	3.30%

Table 5.29 Top 10 countries of international visitors

	Country	2018	% of Int'l	2017	% of Int'l	2016	% of Int'l	2015	% of Int'l
1	Canada	3512000	24.4%	3447000	24.0%	3345000	23.1%	3797000	24.9%
2	UK	1498000	10.4%	1496000	10.4%	1587000	10.9%	1696000	11.1%
3	Brazil	1218000	8.5%	1110000	7.7%	1030000	7.1%	1475000	9.7%
4	Argentina	646000	4.5%	767000	5.3%	757000	5.2%	722000	4.7%
5	Colombia	598000	4.2%	556000	3.9%	538000	3.7%	565000	3.7%
6	Mexico	468000	3.3%	439000	3.1%	479000	3.3%	453000	3.0%
7	Germany	398000	2.8%	465000	3.2%	477000	3.3%	499000	3.3%
8	Venezuela	386000	2.7%	402000	2.8%	445000	3.1%	404000	2.7%
9	France	323000	2.2%	315000	2.2%	312000	2.2%	313000	2.1%
10	China	288000	2.0%	308000	2.1%	309000	2.1%	300000	2.0%

5.4 Data validation

5.4.1 Validation methodology (Objective 1)

Three data sources (social media, cell phones, and Visit Florida surveys) have provided information on Florida travel with each of the sources having its own unique advantages and disadvantages.

The social media data contain holistic information about all groups of visitors, and allow retrieval of origins, destinations and travel network of Floridians, domestic and international visitors. The timeframe of social media is long enough for time series analysis. One demerit of social media is its course granularity in terms of trip origin, given that the home locations of the social media users are self-reported and presumably contain somewhat noisy data. Note that data from English, Spanish and Portuguese speaking international travelers were not totally consistent; while the additional analysis is required, it should be similar to that for the domestic travelers.

The cellphone data vastly outperform the other two sources in terms of its geographical resolutions. All trips, origins and destination information alike can be tracked down to a census tract (and potentially to a census block) level. However, the cellphone data in this study covers one year only and excludes international visitors' information to be retrieved from this database.

The survey data from Visit Florida is conventionally regarded as the official statistics and provides visitors' travel profiles such as demographic features, trip companies and stay length, etc., which are

largely unavailable in social media and cellphone data. Nevertheless, the survey data only incorporate generic domestic and international visitors' arrivals and origins with no detailed data on visited destinations. Travel information regarding Floridians is unavailable.

A summary of the features and comparative details of the three data sources is illustrated below in Table 5.30.

Table 5.30 Three data source features and details

	Origin	Destination	Network	Geo Resolution	Timeframe	Time Frequency
Social media data						
Floridian	Yes	Yes	Yes	County - County (Tract)	2003-2019	Monthly
Domestic	Yes	Yes	Yes	State - County	2003-2019	Monthly
Int'l	Yes	Yes	Yes	Nation - County	2003-2019	Monthly
Cellphone data						
Floridian	Yes	Yes	Yes	Tract - Tract	2018.10-2019.9	Monthly
Domestic	Yes	No	No	State - Not applicable	2018.10-2019.9	Monthly
Int'l	No	No	No	Not applicable	2018.10-2019.9	Monthly
Survey data						
Floridian	No	No	No	Not applicable	Not applicable	Not applicable
Domestic	Yes	Partial	No	State - Region	2015-2018	Annually
Int'l	Yes	Partial	No	Nation - Region	2015-2018	Annually

Based on the data field availability and geographical resolution consistency, the validation methodology was as follows. To validate the origins of Floridians, the spatial distributions retrieved from social media and cellphone data were compared. The correlation coefficient (Pearson's r) between the log-transformed paired data was used to estimate match between different data sources. Similarly, to validate the origins of domestic visitors, the destination of Floridians, and the travel network of Floridians, their respective representations in different databases were used as shown in Table 5.31.

Table 5.31 Cross-validation of different data sources

	Group	Data source	Resolution	Time Aggregate
Origin	Floridian	Social Media * Cellphone	County	Overall
	Domestic	Social Media * Cellphone * Survey	State	Overall
Destination	Floridian	Social Media * Cellphone	County (Tract)	Overall
Network	Floridian	Social Media * Cellphone	County - County	Overall

5.4.2 Spatial validation (Objective 2)

5.4.2.1 Validation of trip origins

Floridians

The validation of the origins of Floridian travels was based on the data from social media and cellphone, using a county level of resolution. The data points of the top origins from both datasets are illustrated in Table 5.32. The number of trips from same origins estimated from social media and cellphone data is highly correlated: Pearson's r= 0.93, p<0.001 (Figure 5.12). The preliminary estimation implies that 1 trip counts from social media approximate 100 trip counts from cellphone data.

Table 5.32 Top origin counties of Floridians

Origin	Name	N Trips (Cellphone)	N Trips (Social)
12099	Palm beach	1531156	15309
12057	Hillsborough	1435614	13325
12086	Miami-Dade	1205709	12986
12031	Duval	1147412	8711
12095	Orange	1128544	14447
12011	Broward	866250	15048
12071	Lee	830158	9218
12103	Pinellas	816220	10156
12105	Polk	757789	4618
12009	Brevard	674611	6898
12083	Marion	556081	3773
12111	St. Lucie	540339	3155
12127	Volusia	509854	4724

Origin	Name	N Trips (Cellphone)	N Trips (Social)
12101	Pasco	443524	3273
12021	Collier	443330	4587
12115	Sarasota	413620	9578
12073	Leon	380699	4017
12001	Alachua	355862	4249
12081	Manatee	350799	2323
12117	Seminole	332619	2635
12017	Citrus	316901	2244
12055	Highlands	283674	1026
12097	Osceola	268120	1811

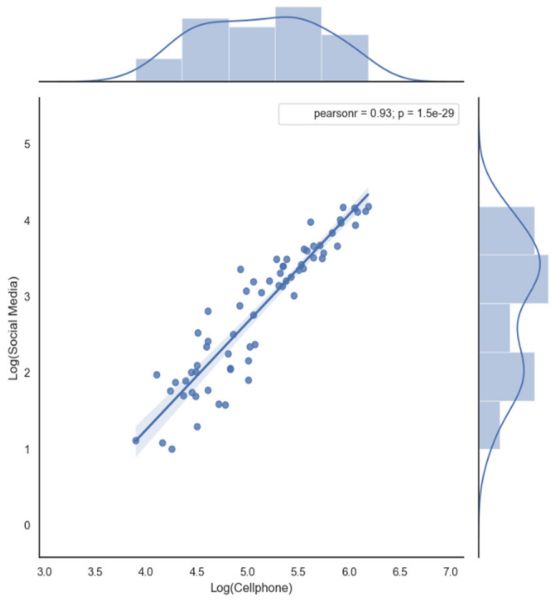


*Figure 5.12 Correlation of log (social media) * log(cellphone) trip origin counts. Only Floridian travelers. R=0.93*

Domestic travelers

Validation of the origins of domestic visitors was based on the data from social media, cellphone, and survey data, on a state level of resolution. Given that the cellphone data reflected only the 2018 – 2019 visitors we used the 2018 annual Visit Florida survey data (Visit Florida, 2019) for validation. The data

on the top 15 origin states provided in the Survey was compared with data from the other two datasets (Table 5.33) and demonstrated high correlation between the respective data points (see details in Figure 5.13). The preliminary estimation implies that 1 trip count from the TripAdvisor data is equivalent to 100 trips from the cellphone data and 2000 trip counts from Visit Florida survey, hence providing the base to translate social media and cellphone record data to real visitation data.

Table 5.33 Top origin states of domestic visitors

Origin State	N Trips (Cellphone)	N Trips (Social)	N Trips (Survey 2018)
Georgia	834620	30639	11935176
New York	661042	34242	10021044
California	368015	11633	4503840
Texas	364626	16726	4841628
North Carolina	326272	15850	5292012
New Jersey	277725	16661	4729032
Ohio	271976	18393	4841628
Pennsylvania	270474	19783	5742396
Alabama	266281	7955	5404608
Virginia	266275	12605	3265284
Illinois	261418	18124	5517204
Massachusetts	203044	14162	3152688
Michigan	200719	13589	4278648
Tennessee	175951	12853	4616436
Indiana	150297	9280	3603072
Maryland	141362	8943	2927496

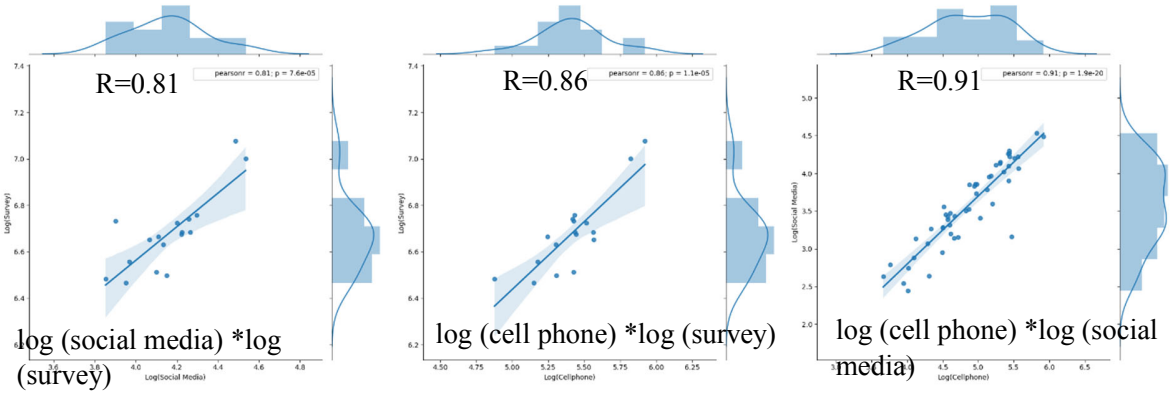


Figure 5.13 Correlations of origin trip counts estimated from three datasets

5.4.2.2 Validation of destinations

The validation of the destination choices of Floridian travelers was based on data from social media and cellphone, on a county level of resolution. The comparative data for the top destinations from both datasets are found in Table 5.34. The number of trips estimated from social media is highly correlated to that from cellphone data: Pearson's r= 0.89, p<0.001 (Figure 5.14). The preliminary estimation implies that 1 trip counts from social media approximate 100 trip count from cellphone data.

Table 5.34 Top destination counties of Floridians

Destination	Name	N Trips (Cellphone)	N Trips (Social)
12095	Orange	4222721	32014
12086	Miami-Dade	2573859	9787
12057	Hillsborough	1878658	8779
12011	Broward	911854.1	8149
12099	Palm beach	690836.9	7145
12105	Polk	677824.3	2978
12031	Duval	622774.8	5804
12103	Pinellas	549217.5	11055
12097	Osceola	534319.4	5746
12117	Seminole	534261.6	1893
12127	Volusia	470827.4	7101
12071	Lee	458814.3	8871

Destination	Name	N Trips (Cellphone)	N Trips (Social)
12083	Marion	455840.1	2334
12009	Brevard	441197.5	4907
12081	Manatee	346669.6	2264
12115	Sarasota	313697.7	5610
12001	Alachua	298339.3	3761
12069	Lake	296097.9	2263
12021	Collier	278866.8	5488
12111	St. Lucie	274508	1733
12101	Pasco	225522.6	981
12073	Leon	209989.7	3627

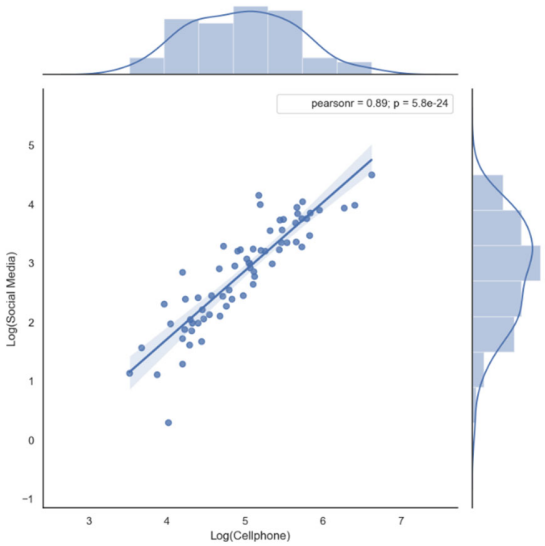


*Figure 5.14 Correlation of log (Social media) * log(cellphone). Only Floridian travelers. R=0.89*

5.4.2.3 Validation of travel network

The validation on the travel network of Floridians was based on data from social media and cellphone data, on a county-to-county level of resolution. The number of trips in the OD travel flow from both datasets for the top network links are shown in Table 5.35. The number of trips in the corresponding links are highly correlated: Pearson's r= 0.72, $p<0.05$ (Figure 5.15). The preliminary estimation implies that 1 travel estimated from the social media approximates 180 travel travels estimated from the cellphone data.

Table 5.35 Top destination counties of Floridians

Travel Direction	N Trips (Cellphone)	N Trips (Social)
12099-12086	806,358	1619
12057-12095	699,457	3606
12031-12095	339,657	2187
12103-12095	326,103	2286
12071-12086	325,180	622
12031-12057	295,452	496
12095-12057	284,670	1314
12086-12099	234,129	1109
12009-12095	227,097	-
12086-12095	224,680	3748
12083-12095	222,944	668
12111-12086	173,937	150
12099-12095	171,169	3264
12021-12086	167,268	519
12101-12095	145,765	833
12011-12095	143,016	3504
12127-12095	129,908	-
12001-12095	110,988	818
12111-12011	106,188	212
12105-12103	106,016	665
12071-12011	104,958	556
12095-12103	101,965	1936

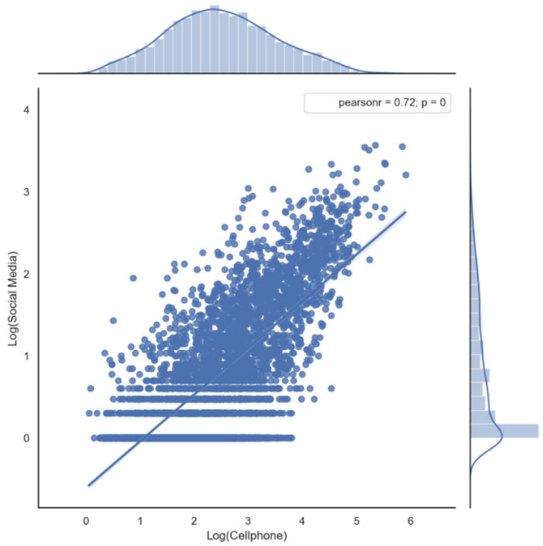


*Figure 5.15 Cross-plot of log (Social media) * log(cellphone). Only Floridian travelers. R=0.72*

5.4.3 Temporal validation (Objective 3)

The temporal validation was conducted between social media data and survey data, given that the timeframe of these datasets was long enough for temporal analysis. The time frequency was chosen on a seasonal level, from 2015 Q1 to 2018 Q4 (Table 5.36). From the temporal distributions of the two data sources (Figure 5.16), we observed a consonant seasonal pattern that each first season receives the largest volume of arrivals, while the fourth season the lowest. The correlation test also shows that the two datasets are highly correlated to each other: Pearson's $r=0.82$, $p<0.05$. Note that TripAdvisor data should be adjusted for changes in popularity of the platform, potentially resulting in a higher correlation between datasets.

Table 5.36 Tourist arrivals comparison

Season	Arrivals (Survey)	Arrivals (Social)
2015Q1	28,482,000	29,987
2015Q2	26,478,000	30,212
2015Q3	25,605,000	33,872
2015Q4	25,990,000	33,011
2016Q1	30,321,000	53,911
2016Q2	28,029,000	51,421
2016Q3	27,030,000	49,944
2016Q4	26,794,000	41,252
2017Q1	31,890,000	54,904
2017Q2	30,017,000	48,802
2017Q3	27,854,000	40,187
2017Q4	28,662,000	38,276
2018Q1	34,771,000	63,583
2018Q2	31,506,000	57,250
2018Q3	31,261,000	56,850
2018Q4	29,441,000	44,220

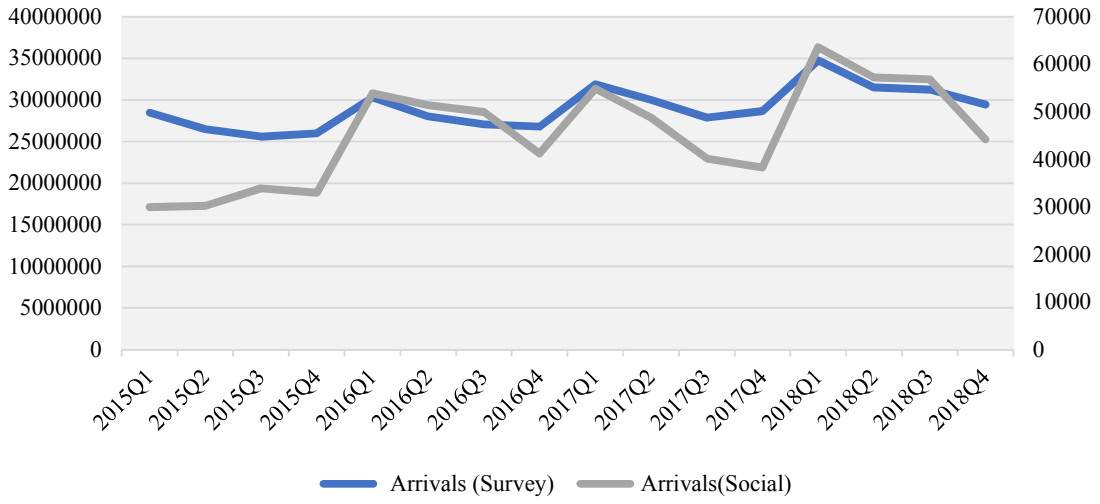


Figure 5.16 Temporal distribution of arrivals from social media and survey data

Chapter 6. Implications for modeling

Task description

The Research Team will prepare and test the tourism flow methodology to evaluate tourism travel impacts to the transportation system. Subtasks to this work include:

- Building on the experience obtained in tasks 2-5 and previous research (Task l), provide recommendations on the improved methodology of data collection, processing, validation, and warehousing;
- Based on the results of tasks 4 and 5, estimate the impact (in terms of contribution to traffic flow) of tourism travel on Florida's transportation system;
- Estimate the potential impact of various levels of accessibility/availability on Florida's tourism-related enterprises. This work will build relationships between the number and size of tourist enterprises, as well as levels of tourism-related industry clusters (e.g., lodging, food/beverage, event/entertainment and others), with the level of highway accessibility/availability for each county.

Deliverable: Upon completion of Task 6, the University sha11 submit to the Research Center at research.center@dot.state.fl.us a written Technical memorandum of the research methodology, findings and analyses.

6.1 Destination choice model

6.1.1 Introduction

Gravity models have been the most common form of trip distribution model for decades and are arguably still the most used form in practice. However, destination choice models (DCM) have gained an increasing replacement for gravity models to improve the accuracy of the trip distribution estimation, given their advantages in the incorporation of additional variables, as well as reflecting more complex statistical assumptions (e.g., capturing spatial autocorrelation) (Bernardin et al., 2009). Destination choice models are even more advantageous over gravity models for longer distance personal travel and multinucleated travel regions, and have therefore been widely incorporated in statewide travel models (e.g., Arizona, California, Idaho, Iowa, Maryland, New Hampshire, Ohio, Oregon, Tennessee, Wisconsin, etc.) and many metropolitan area models alike (e.g., South Bend, Evansville, and Columbus, Indiana; Ann Arbor, Michigan; Burlington, Vermont; Knoxville and Chattanooga, Tennessee; Charlottesville, Virginia; Charleston, South Carolina; and Jacksonville, Florida).

The destination choice model is a type of trip distribution or spatial interaction model which is formulated as discrete choice models, typically logit models. The formulations of destination choice models are flexible and extensible to include a wider range of explanatory variables and thus provide a better behavioral basis for trip distribution than the traditional gravity models. Typically, a destination choice model incorporates additional variables beyond size/attractions, impedance/friction factors, and constants/k-factors.

The most common destination choice model nowadays is some form of the random utility model, usually a multinomial logit (MNL) model. A typical logit destination choice model for the probability that destination *j* is chosen from origin *i* ($P_{j|i}$) is expressed as:

$$P_{j|i} = \frac{e^{U_{j|i}}}{\sum_{j'} e^{U_{j'|i}}} \quad (6.1)$$

where $U_{j|i}$ is the systematic utility of destination *j* given origin *i*, which can be written as follows.

$$U_{ij} = \beta_1 \cdot X_i + \beta_2 \cdot Y_j + \beta_2 \cdot Z_{ij} \quad (6.2)$$

In formulation (2), the utility of a chosen destination depends on (a) origin-specific variables X_i, which do not vary between destinations, (b) destination-specific variables Y_j, which vary between destinations, and (c) origin-destination interactive variables Z_{ij}, which differ based on different origins and destinations. This is the simplest representation of destination choice utility.

6.1.2 Model design

Rather than present an eclectic model, we estimate the optimized model with iterating development. *Model 1* is a simple model based on origin socioeconomic variables; *Model 2* incorporates further interaction variables between origin and destinations; *Models 3* and *4* focus on tourism resources and facilities in destinations; finally, *Model 5* is the comprehensive model including all possible variables in the destination choice model.

Model 1 (socioeconomic)

The traditional theory in tourism regarding the destination choice as a 'push-pull' process, where the decision is not merely influenced by the pull attractiveness of the destinations, but also the push drivers from the origins or tourist themselves. Hence some socioeconomic features of the origin are likely to influence the decision in choosing destinations, and we include the two factors, population $pop_i$and average income $inc_i$of the origins as the independent variables in ***Model 1***. The two factors were transformed with a logarithmic function following Train (2009) suggestion that the representative utility needs to be specified with parameters inside a log operation if the destination choice model is not sensitive to the level of zonal aggregation. Thus, the utility function U_{ij} in ***Model 1*** is presented as:

$$U_{ij} = \beta_1 \log(d_{ij}) + \beta_2 \log(pop_i) + \beta_3 \log(inc_i) \quad (6.3)$$

Where d_{ij} is the travel distance between origin *i* to destination *j*. Travel distance is a typical representation of the travel impedance in any utility function, an essential measurement of the generalized cost of the travel, also possibly measured by travel time, traffic congestion time, etc. A convenient measure of impedance is the inclusive value, or log sum, of the mode choice model (de Jong et al., 2007).

Model 2 (land use)

Traffic analysis zones (TAZs) are not homogeneous in land use patterns. Urban areas with more sufficient accommodation and entertainment facilities are more likely to attract visitors. Mishra et al. (2013) found that interaction terms between the origin and the destination land usage were significant for their destination choice model for Maryland. We include two control variables in ***Model 2*** regarding the urbanized zone effects from origin to destinations uu_{ij} and ru_{ij}. The control variables were calculated based on $urban_i$, which indicates whether the TAZ is an urbanized zone.

$$uu_{ij} = \begin{Bmatrix} 1, & if\ urban_i \wedge urban_j \\ 0, & otherwize \end{Bmatrix} \quad (6.4)$$

$$ru_{ij} = \begin{Bmatrix} 1, & if\ !urban_i \wedge urban_j \\ 0, & otherwize \end{Bmatrix} \quad (6.5)$$

Thus, the utility function U_{ij} in ***Model 2*** is presented as:

$$U_{ij} = \beta_1 \log(d_{ij}) + \beta_2 \log(pop_i) + \beta_3 \log(inc_i) + \beta_4 uu_{ij} + \beta_5 ru_{ij} \quad (6.6)$$

Model 3 (tourism simplified)

The traditional socio-economic variables and land-use feature in urban development do not completely reflect why tourists travel to particular destinations. The attractiveness of destinations is more likely to lie in the tourism resources and facilities where tourists can perform certain leisure and tour activities. To this end, ***Model 3*** incorporates two tourism-specific variables of destinations from TripAdvisor database collected in Task 2 and Task3, namely, tourism attractions $attraction_j$ and hotels $hotel_j$. Tourism attractions include venues commonly visited such as beaches, theme parks, museums, indoor and outdoor recreation facilities, and the parameter $attraction_j$ is to represent the tourism attractiveness of the TAZs

based on their diverse tourism resources and activities appealing to tourists. $Hotel_j$ variable is the reflection of the accommodation capacity of each TAZ in Florida. The two variables were calculated on census tract level and logarithmically transformed.

Note that the attractiveness of tourism resources is not only revealed by the number of attractions, but also the quality of its appeal. Hence there are two optional calculation of the parameter $attraction_j$:

$$attraction_j = \log(number\ of\ attractions) \quad (6.7)$$

$$attraction_j = \log(number\ of\ reviews\ related\ to\ attractions) \quad (6.8)$$

where the review numbers related to attractions on TripAdvisor in (8) are more accurately reflect the scale of attraction, especially those extremely popular attractions like Disneyland. Nevertheless, both of them are used in the model to test their performance.

The calculation of the parameter $hotel_j$ is likewise optional, either on the number of hotels, or the number of beds, or the number of reviews related to hotels. We adopted the following four optional values for the parameter $hotel_j$:

$$hotel_j = \log(number\ of\ hotel) \quad (6.9)$$

$$hotel_j = \log(number\ of\ hotel\ rooms) \quad (6.10)$$

$$hotel_j = \log(number\ of\ reviews\ related\ to\ hotel) \quad (6.11)$$

$$hotel_j = \log(number\ of\ airbnd\ rooms) \quad (6.12)$$

We conducted a paired correlation test to see if these measurements are replaceable to each other. The results showed in Figure 6.1 indicate that the three possible measurements for $hotel_j$ are highly correlated to each other (r=0.91 for hotel number vs. hotel review numbers, r=0.85 for the correlation of the hotel room numbers and the hotel review numbers). The Airbnb however is rather weekly related to the hotel: for example, the correlation between the Airbnb room numbers and the hotel room numbers r=0.37. Therefore, we selected the $\log(number\ of\ reviews\ related\ to\ hotel)$ as the measurement for $hotel_j$ and also added $airbnd_j$ as an independent variable in Model 3.

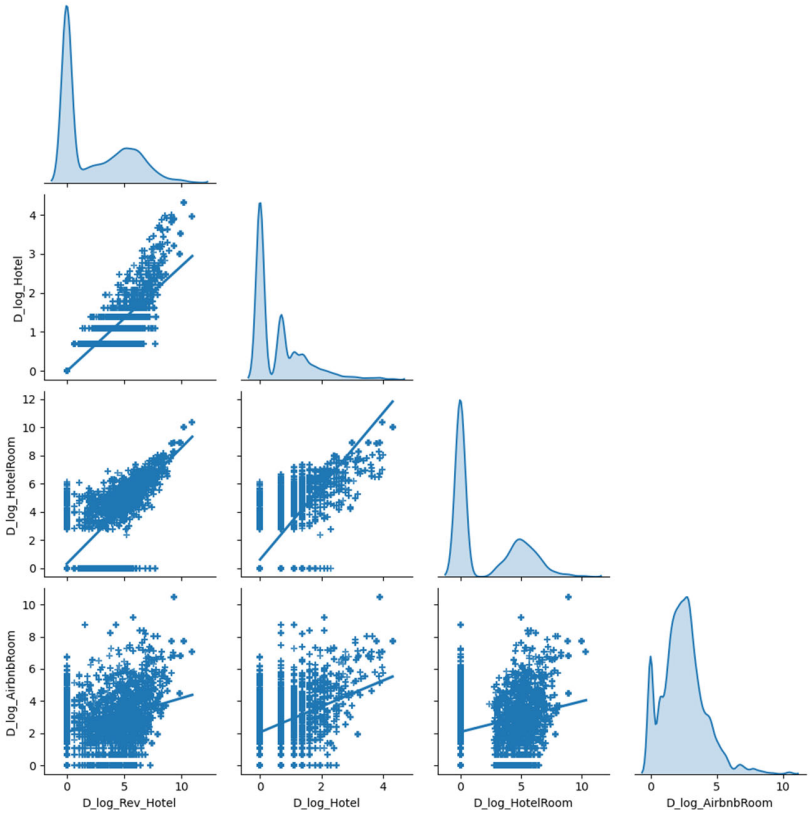


Figure 6.1 Correlations between optional hotel measurements and Airbnb measurement. Notice that a large number of census tracts have no hotel rooms, which is compensated by the Airbnb offering

Finally, the utility function U_{ij} in ***Model 3*** is presented as:

$$U_{ij} = \beta_1 \log(d_{ij}) + \beta_2 \log(pop_i) + \beta_3 \log(inc_i) + \beta_6 attraction_j + \beta_7 hotel_j \quad (6.13)$$

Model 4 (tourism extended)

Model 4 is an extended model based on ***Model 3***, where two additional destination features specifically related to Florida were taken into account. According to a report from Florida's Office of Economic and Demographic Research, two major tourism activities in Florida are theme park attendance and coastal/cruise activities. Whether a destination zone is a theme park or coastal related is highly likely to influence the tourism resource and facility supply in the destinations and thus impact the tourists' decision-making in choosing Florida destinations (See Figure 6.2).

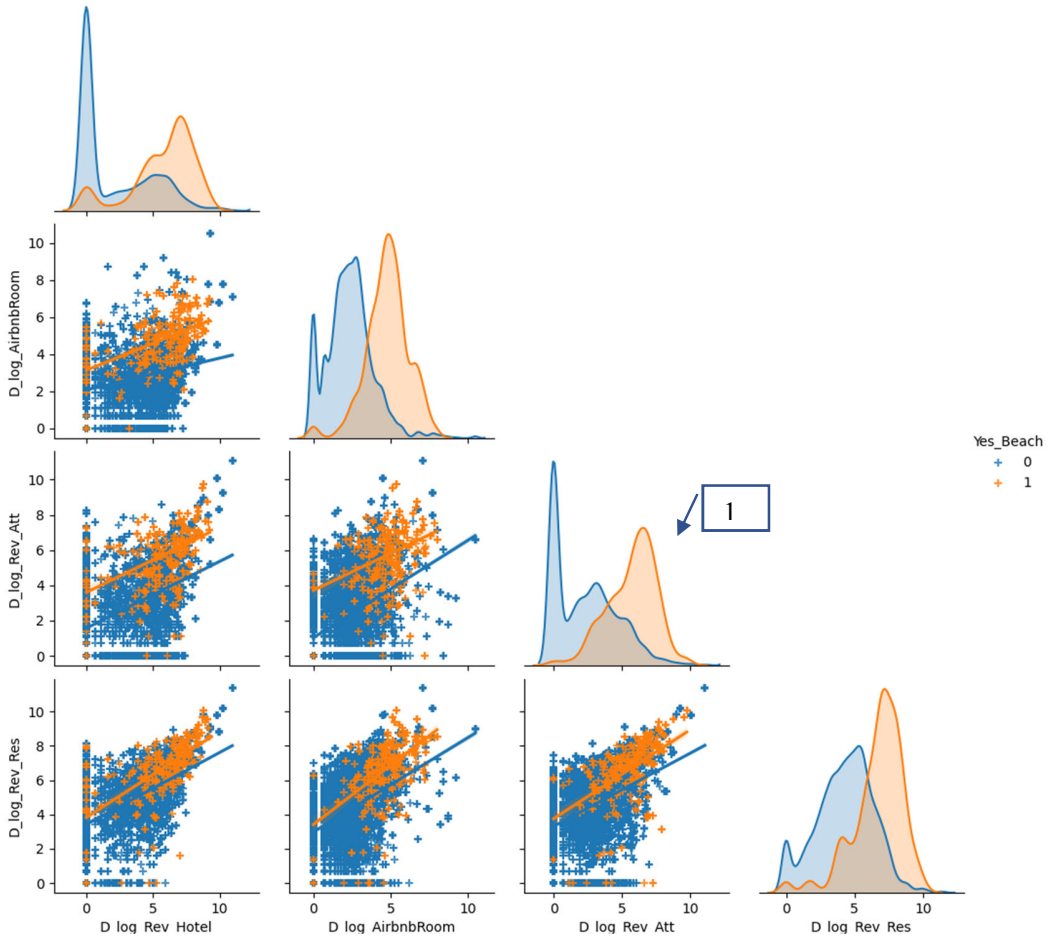


Figure 6.2 Distribution of tourism facilities in beach and non-beach destinations. Notice that the destinations with beach accesses (orange) are have significantly more accommodation facilities and attraction points. Also notice a large number of non-beach attraction points with no or very few reviews (box 1 on the figure)

We consequently generated two parameters, $themepark_j$ and $costal_j$ to indicate if the destination has such featured attractions or products.

$$yes\_themepark_j = \begin{cases} 1, & if\ zone\ contains\ any\ theme\ park \\ 0, & otherwize \end{cases} \text{ and}$$

$$yes\_costal_j = \begin{cases} 1, & if\ zone\ has\ beach\ access \\ 0, & otherwize \end{cases} \qquad (6.14)$$

The utility function U_{ij} in ***Model 4*** is presented as:

$$U_{ij} = \beta_1 d_{ij} + \beta_2 \log(pop_i) + \beta_3 \log(inc_i) + \beta_6 attraction_j + \beta_7 hotel_j + \beta_8 airbnd_j + \beta_9 yes\_themepark_j + \beta_{10} yes\_costal_j \qquad (6.15)$$

Model 5 (comprehensive)

Model 5 is a comprehensive model presumptively to incorporate all possible variables mentioned in the previous four models. The utility function U_{ij} in ***Model 5*** is presented as:

$$U_{ij} = \beta_1 d_{ij} + \beta_2 \log(pop_i) + \beta_3 \log(inc_i) + \beta_4 uu_{ij} + \beta_5 ru_{ij} + \beta_6 attraction_j + \beta_7 hotel_j + \beta_8 airbnd_j + + \beta_9 yes\_themepark_j + \beta_{10} yes\_costal_j \quad (6.16)$$

6.1.3 Data

6.1.3.1 Observed Choice Data

The most common source for Observed Choice Data is household travel surveys. It is possible to retrieve tourism-related travel choice when the travel purposes are elaborated as leisure or recreation in certain surveys. It is also one advantage of survey data to collect the socio-demographic features of travelers and their motivations for the trips. Nevertheless, the survey work is costly and time-consuming, with relatively small coverage of the population.

In this study, we adopted the passively collected data from cellular phones, restored in the data structure of origin-destination pairs. It is expected that cell phone data was able to capture almost all origin-destination pairs that did take place, with refined granularity on census tract level. The data collection and cleaning process have been detailed in the Report of Task 5, Section 2.2. Cross-validated by social media data and Visit Florida survey data, the Observed Choice Data of tourists from cellphone data has validated to be reliable and representative.

Note that there were seven categories of travel purposes in the original cellular data. We selected only the HO (home to other) as the possible representation of leisure and tourism-related traffics, approximately 18.63% of the entire traffic counts. Also, traffic less than 50-mile distance was removed in the later modeling analysis, in conformity with the definition of tourist. Summary of the data is presented in Table 6.1.

Table 6.1 Summary of the cellphone data

	Total Traffic	HO Traffic	Rate	OO Traffic	OH Traffic	HW, WH, WO, OW traffic
Total Traffic Counts	2,641,291,009	492,044,579	18.63%	1,273,802,105	452,410,978	339,284,311
Total Pairs	7,071,479	3,543,072	50.10%	5,593,686	3,550,178	3,164,185
Average Traffic Count	373.51	138.87	37.18%	227.72	127.43	107.23

The finalized dataset was restored as ***6.3.1_Observed_Choice_Data_Tract.csv***, an O-D matrix structure of the observed origin-destination flow counts on census tract level. The alternative data set is ***6.3.1_Observed_Choice_Data_County.csv***, which reorganized the OD matrix on the county level.

6.1.3.2 Explanatory Data

In addition to observed choice data, destination choice models need information on possible origins, destinations, as well as origin-destination interaction to estimate and predict the parameters. These data often are called Explanatory Data, or size term data (Table 6.2).

The socioeconomic data (population, median household income) of origins in Model 1were retrieved from the 2018 ACS 5-Year Estimates data. The urbanization development indicator used to construct the interaction of origin and destination regarding their land use in Model 2 was retrieved from the 2010 census data. These cleansed variables can be found in the data file: ***6.3.1_ED_Census_Survey_tract.xlsx***

In Model 3, the attraction and hotel variable information was collected from TripAdvisor. The data collection and data processing has been explained thoroughly in Report of Task 3 and 4, and the cleansed data variables employed in the destination choice model can be found in the file:
6.3.1_ED_Social_Media_tract.xlsx

In Model 4, the additional tourism resource data were collected from data sources such as STR (Smith Travel Research), UFGC (University of Florida GeoPlan Center), and FDEP (Florida Department of Environmental Protection). The detailed survey data sources and data collection have been explained in the Report of Task 2. The selective cleansed data variables are restored in the file:
6.3.1_ED_Tourism_Resource_Survey_tract.xlsx

The impedance is required in any destination choice model, which is commonly calculated as travel time, travel distance, or travel costs. We used the travel distance between zones as the impedance in the destination choice models. The travel distances were calculated as haversine distance between origin centroids to destination centroids. All 1,572,611 OD-distance links are restored in the file:
6.3.1_IMP_Travel_Distance_tract2tract_link.csv

Compared with haversine distance between ODs, travel time and travel distance based on real road networks are alternatives, arguably more ideal measurement of impedance to represent the generalized

travel costs. Hence, we have generated two additional origin-destination impedance databases accordingly, which were estimated by Google API. The potential price to request all the OD pairs calculations on the Google platform would be beyond the budget of the project (approximately $90,000). We thereby only convert the travel time matrix and real travel distance matrix on the county to census tract dimensions (67*4245). The matrix files are restored in ***6.3.1_IMP_Travel_Distance_cty2tract.csv*** and ***6.3.1_IMP_Travel_Time_cty2tract.csv***

Table 6.2 List of explanatory variables

Data	Source	Resolution	Category	Model
Population	Census data	Census tract	Origin	Model 1
Median Income	Census data	Census tract	Origin	Model 1
Urbanization	Census data	Census tract	Origin-Destination	Model 2
Attraction	Social Media	Aggregate to census tract	Destination	Model 3
Hotel	Social Media/survey	Aggregate to census tract	Destination	Model 3
Theme Park	Survey	Calculated on census tract	Destination	Model 4
Beach Access	Survey	Calculated on census tract	Destination	Model 4
Impedances	Conversion	Calculated on census tract	Origin-Destination	Model 1/2/3/4

6.1.3.3. Data Adjustment and Sampling

The probability variable (OD_Prob), the probability of travel from a given origin to certain destination was calculated based on the formular $P_{ij} = \frac{Traffic_{from\ i\ to j}}{\sum Traffic_{from\ i}}$, where $\sum Traffic_{from\ i}$is the the total traffic counts originated from the origin zone *i* (O_Total). The shortcoming of such measurement is that it possibly exaggerates the travel probability from zones with low total travel volume heading to limited destinations. We hereby adopted the Bayes' theorem to recalculate the probability as conditional probability, where the travel probability between Oi to Dj should be mediated by the county traffic where Oi located $County_A(O_i \in County_A)$, and the conditional probability is calculated as:

$$P\_Bay_{ij} = P(O_i\ in\ County_A) \cdot P\big(County_A\ to\ D_j\big) = \frac{\sum Traffic_{from\ i}}{\sum Traffic_{from\ A}} \cdot \frac{Traffic_{from\ A\ to\ D_j}}{\sum Traffic_{to\ D_j}} \quad (6.17)$$

Which is restored as variable OD_Prob_Bay in the database.

The overall dataset prepared for the destination choice model is restored in the file as ***DCM_all.*****csv. A** typical entry of the data point contains the following variables and fields as presented in Table 6.3:

Table 6.3 OD travel attributes for destination choice model

	Variable	Description	Sample Value
Basic information	O_Tract	The census tract FIPS where the trips began (Origin)	"12095017001"
	O_County	The county FIPS where the trips began	"12095"
	O_Total	Total traffic counts originated from the origin zone	6218.375158
	OD_Rank	$Rank_{d\|o}$, the rank of the destination from a given origin	496
	O_AREA	Land area coverage of the origin zone (square meters)	27308074
	O_UR	Land use type of the origin zone (urban, rural or mixed)	"U"
	O_LAT	Latitude of the population centroid of origin zone	28.4437111
	O_LON	Longitude of the population centroid of origin zone	-81.4468718
	D_Tract	The census tract FIPS where the trips ended (Destination)	"12071010406"
	D_County	The county FIPS where the trips ended	"12071"
	D_Total	Total traffic counts ending in the destination zone	10595.88937
	D_Rank	$Rank_d$, the overall rank of the destination in volume	331
	D_UR	Land use type of the destination zone (urban, rural or mixed)	"M"
	D_LAT	Latitude of the population centroid of destination zone	26.5327906
	D_LON	Longitude of the population centroid of destination zone	-82.0342966
Variables for model	Count	Observed traffic from origin to destination	2.629819943
	OD_Prob	Probability to certain destination from a given origin	0.000422911
	OD_Prob_Bay	Bayes Probability to certain destination from a given origin	0.00144558
	OD_DIS	Distance from origin to destination (mile)	137.565101
	O_AVR_INCOME	Mean household income in origin zone (dollars)	91937
	O_POP	Population of origin zone (estimated on 2018)	6722
	O_log_Distance	Logarithm of distance	4.931340259
	O_log_Income	Logarithm of mean income	11.42886972
	O_log_Pop	Logarithm of population	8.813289762
	OD_Urban	OD interaction indicator, whether travel interurban areas	0
	OD_Rural	OD interaction indicator, whether travel from rural to urban	1
	N_Rev_Att	Number of reviews regarding attractions in the destination	91
	N_Attraction	Number of tourism attractions in the destination	4
	N_Rev_Hotel	Number of reviews regarding hotels in the destination	0
	N_Hotel	Number of hotels in the destination	0
	N_AirbnbRoom	Number of Airbnb rooms in the destination	357
	N_HotelRoom	Number of hotel rooms in the destination	0
	D_log_Rev_Att	Logarithm of attraction reviews	4.521788577
	D_log_Att	Logarithm of attractions	1.609437912
	D_log_Rev_Hotel	Logarithm of hotel reviews	0
	D_log_Hotel	Logarithm of hotels	0
	D_log_AirbnbRoom	Logarithm of Airbnb rooms	5.880532986
	D_log_HotelRoom	Logarithm of hotel rooms	0
	Yes_Amuse	Whether the destination has any theme parks	0
	Yes_Beach	Whether the destination has any beaches	0

6.1.4 Model evaluation, census tract level

A subset of the original dataset was selected for estimation modeling, given that the raw data volume was relatively too large (1,572,611) and with a certain proportion of noise. The selected sample only contains the traffic to the top 100 destinations, originated from zones whose traffic makes up more than 1% of the corresponding travel flow. 45,025 OD travel flows were included in the estimation models, starting from 1174 unique Origin census tracts ending in 100 possible top Destination census tracts.

Usually, to run a destination choice model, the survey data is used to train the Multinomial Logit Model; hence, the individual trips are used as individual records with an assigned destination of each trip. The dependent variable is a discrete choice of the destination. However, we uniquely have the observed data of the travel flow frequency in a form of OD pairs (the dependent variable is the frequency of OD flow), which can be used *instead* of the model, at least, for Floridian travelers – see Table 6.4. The model however has its own benefits such as predictive capabilities. In this section, we make preliminary investigation of different formulations of the models formulated in section 6.3.

Table 6.4 OD data structure sample

O_Tract	O_log_Income	O_log_Pop	D_Tract	D_log_Rev_Att	D_log_Att	D_log_Rev_Hotel	D_log_Hotel	D_log_HotelRoom	Yes_ThemePark	Yes_Beach	Count	OD_Urban	OD_Rural	OD_distance	OD_log_dis	OD_Prob
12057006502	11.1	7.6	12095014200	6.628	2.079	7.537	1.609	6.992	0	0	35.41	1	0	80.1	4.3956	0.023607
12057006502	11.1	7.6	12095010200	5.976	2.197	4.644	0.693	4.511	0	0	7.54	1	0	84.1	4.4442	0.005030
12057006502	11.1	7.6	12095017001	9.243	4.682	10.227	4.317	9.992	1	0	34.33	1	0	76.5	4.3499	0.022889
12057006502	11.1	7.6	12097040802	6.605	2.773	9.316	3.892	8.904	0	0	29.70	1	0	63.2	4.1626	0.019799
12057006502	11.1	7.6	12095016906	0.000	0.000	2.773	0.693	0.000	0	0	54.33	1	0	80.4	4.3995	0.036224
12057006502	11.1	7.6	12095015103	2.708	1.099	0.000	0.000	4.663	0	0	59.72	1	0	84.0	4.4422	0.039816

Therefore, the following models were all evaluated with multivariate linear regression models, where the dependable variables were defined as the probable travel volume:

$$Volume_{ij} = P\_Bay_{ij} \cdot Population_i \quad (6.18)$$

Basic model

The basic model only took the distance impedance as the variable for destination choice, where $Volume_{ij} = \beta_1 d_{ij}$. The parameters of this model m_0 show (table below) that distance is a significant predictor variable with a negative impact.

MODEL FIT:
$\chi^2(1) = 101397.8624$, $p = 0.0000$
Pseudo-R^2 (Cragg-Uhler) = 0.0161
AIC = 349608.2651, *BIC* = 349634.4101

	Est.	S.E.	t val.	p	
(Intercept)	8.4352	0.1044	80.8297	0.0000	***
OD_DIS	-0.0180	0.0007	-27.1128	0.0000	***

Model 1

Model 1 took the socioeconomic variables from origin zones in addition to the distance impact, where $Volume_{ij} = \beta_1 d_{ij} + \beta_2 \log(inc_i)$. The parameters of this model m1 are as expected, where distance is still with significant negative coefficient while the income level of the origin plays a positive role. Travels from higher-income level zones are likely to be higher than those with lower levels.

MODEL FIT:
$\chi^2(2) = 120688.5359$, $p = 0.0000$
Pseudo-R^2 (Cragg-Uhler) = 0.0191
AIC = 349470.1893, *BIC* = 349505.0492

	Est.	S.E.	t val.	p	
(Intercept)	-2.5117	0.9301	-2.7004	0.0069	**
O_log_Income	0.9864	0.0833	11.8442	0.0000	***
OD_DIS	-0.0177	0.0007	-26.5723	0.0000	***

Model 2

Model 2 added Origin-Destination interactions into the model, especially the different types of land use between origins and destinations. Tourism facilities are expected to be more developed in urbanized areas and possibly to attract more travelers to such destinations. The model 2 is formulated as:

$$Volume_{ij} = \beta_1 d_{ij} + \beta_2 \log(inc_i) + \beta_3 uu_{ij} + \beta_4 ru_{ij} \quad (6.19)$$

The parameters of this model m2 show that the signs are largely as expected, where travel directions from urbanized origins to urbanized destinations would strongly influence the travel pattern. The interaction between rural origins and urban destinations somehow is not as robust as the inter-urban travels (though with positive coefficient), and arguably could be ignored in the following models.

MODEL FIT:
$\chi^2(4) = 125272.1538$, $p = 0.0000$
$R^2 = 0.0199$
AIC = 349440.8420, *BIC* = 349493.1319

	Est.	S.E.	t val.	p	
(Intercept)	-4.0763	1.0131	-4.0236	0.0001	***
OD_DIS	-0.0177	0.0007	-26.6794	0.0000	***
O_log_Income	0.9586	0.0837	11.4516	0.0000	***
OD_Urban	1.9494	0.3985	4.8913	0.0000	***
OD_Rural	0.8841	0.5137	1.7210	0.0853	.

Model 3

Model 3 is a tourism-specific model, assuming tourism attractions and accommodation facilities are two major motives of tourist travels. The model formula thus is as:

$$Volume_{ij} = \beta_1 d_{ij} + \beta_6 attraction_j + \beta_7 hotel_j + \beta_8 airbnd_j \quad (6.20)$$

The parameters resulted from the estimation show that tourism attractions are of the positive coefficient to the travel pattern, indicating that destinations with more attractions are more likely to attract tourist flows. It is interesting to find that accommodation facilities play negative roles in the model, and the explanation to such findings remains to be discussed with in-depth analysis.

MODEL FIT:
$F(4,39691) = 190.5923$, $p = 0.0000$
$R^2 = 0.0188$
Adj. $R^2 = 0.0187$

Standard errors: OLS

	Est.	S.E.	t val.	p	
(Intercept)	9.5085	0.1519	62.6137	0.0000	***
OD_DIS	-0.0186	0.0007	-25.1936	0.0000	***
D_log_Att	0.4176	0.0830	5.0314	0.0000	***
D_log_Rev_Hotel	-0.0678	0.0284	-2.3900	0.0169	*
D_log_AirbnbRoom	-0.4663	0.0492	-9.4755	0.0000	***

Model 4

Model 4 is the extensive tourism-specific model, where whether destinations featured with coastal and theme park was taken into consideration. The model is formulated as:

$$Volume_{ij} = \beta_1 d_{ij} + \beta_6 attraction_j + \beta_7 hotel_j + \beta_8 airbnd_j + \beta_9 yes\_themepark_j + \beta_{10} yes\_costal_j \quad (6.21)$$

The parameters resulted in the estimation modeling show that both the theme park and beach features are as expected to be positive impacts. Theme park feature has a remarkable influencing coefficient to stimulate tourists' willingness to travel. The beach feature in comparison is a minor influencer.

MODEL FIT:
$F(6,39689) = 198.3650$, $p = 0.0000$
$R^2 = 0.0291$
Adj. $R^2 = 0.0290$

Standard errors: OLS

	Est.	S.E.	t val.	p	
(Intercept)	9.5044	0.1527	62.2618	0.0000	***
OD_DIS	-0.0192	0.0007	-25.7241	0.0000	***
D_log_Att	0.4107	0.0829	4.9518	0.0000	***
D_log_Rev_Hotel	-0.0696	0.0282	-2.4674	0.0136	*
D_log_AirbnbRoom	-0.4721	0.0490	-9.6409	0.0000	***
Yes_Amuse	14.5696	0.7136	20.4178	0.0000	***
Yes_Beach	0.8298	0.4610	1.8000	0.0719	.

Model 5

Model 5 is a comprehensive model to incorporate all possible explanatory variables in the estimation, where the formula is constructed as:

$$Volume_{ij} = \beta_1 d_{ij} + \beta_2 \log(inc_i) + \beta_3 uu_{ij} + \beta_4 ru_{ij} + \beta_6 attraction_j + \beta_7 hotel_j + \beta_8 airbnd_j + \beta_9 yes\_themepark_j + \beta_{10} yes\_costal_j \quad (6.22)$$

MODEL FIT:
$F(9,39686) = 145.7551$, $p = 0.0000$
$R^2 = 0.0320$
Adj. $R^2 = 0.0318$

Standard errors: OLS

	Est.	S.E.	t val.	p	
(Intercept)	-2.8401	1.2168	-2.3340	0.0196	*
OD_DIS	-0.0189	0.0007	-25.1887	0.0000	***
O_log_Income	0.9387	0.0896	10.4821	0.0000	***
OD_Urban	1.9188	0.6914	2.7752	0.0055	**
OD_Rural	1.1610	0.7712	1.5054	0.1322	
D_log_Rev_Att	0.1313	0.0321	4.0915	0.0000	***
D_log_Rev_Hotel	-0.0536	0.0272	-1.9695	0.0489	*
D_log_AirbnbRoom	-0.4322	0.0476	-9.0828	0.0000	***
Yes_Amuse	14.4174	0.7127	20.2288	0.0000	***
Yes_Beach	0.7066	0.4635	1.5245	0.1274	

To sum up, distance is a robust impedance variable in traffic estimation models, with an indispensable yet slight negative coefficient. Income level in origin zones is also an important socio-economic variable, indicating that financial condition is a positive variable in tourists' travel decisions. Origin-destination interaction especially travels between inter-urban areas, also plays a positive role in generating tourism and leisure-oriented travels.

As for tourism elements, tourism attractions and products appear to be the stimulus for travel traffics as expected, and destinations with theme parks are extremely powerful in appealing tourists. Accommodation facilities like hotels and Airbnb, on the other hand, have not shown a positive influence on bolstering tourism travel flows. One plausible explanation to such findings is that the observation data was based on the Floridian population, who are less likely to need accommodations during in-Florida travels. Nevertheless, these specific variables are worthy of investigation in follow-up refined models, especially in the Multinomial logistic regression models for destination choice.

The results show that the comprehensive model is consistent to the previous stepwise models, and the weak explanatory variables in previous tests appear to be vain predictors in the final model. Overall, however, model performance at a census tract level was concluded to be inadequate due to significant

measurement errors in trilaterated cell phone locations. Hence, the data were aggregated to a zip code level and models were re-evaluated. The next section provides modeling results at a zip code level

6.2Model evaluation, zip code level

6.2.1 Introduction

The destination choice models in Section 6.1, while providing informative results regarding the explanatory factors of travel flows between certain origins and destinations, underperforms in explaining the overall variability. The primary reason is locational errors in cell phone data which become critical in urban areas with the census tract resolution. The same data taken at a county level is significantly more robust (see Section 5.4), however the OD travel flows on county level are too generalized to capture the transportation patterns. Therefore, in this section we introduce an intermediate zip-code level to test the destination choice model for OD travel flows. This level is based on the ZIP Code Tabulation Areas (ZCTAs) of the U.S. Postal Services and hence must fall within the US national boundaries.

6.2.2 Data transformation and format

The census tracts are not necessarily entirely contained within a single census tract. Hence, data transformation of cellphone data from census tracts to zip code was completed are following:

- Travel origin counts from a certain zip code were prorated based on its population proportion to the matching census tract area. That is, if a census tract is divided between the zip codes a and b, the travels originating from this census tract is be distributed between those zip codes proportionally to the zip code population;
- Visit counts to certain zip code destination area were prorated based on its coverage proportion to the mapping census tract area. That is, if a census tract is divided between the zip codes a and b, the travels ending in this census tract is distributed between those zip codes proportionally to the zip code areas.

For example, there are 100 visits from census tract 12133970301 to 12133970200 according to the cell phone records. The population of the origin census tract 12133970301 is distributed between the zip codes 32428, 32438, and 32466 as 79.17%, 0.74%, and 20.09%. Similarly, the area of the destination census tract 12133970301 is distributed between the zip codes 32425, 32427, and 32462 as 25.99%, 33.38%, and 40.63%. Then, the travel flow between the census tracts 12133970301 and 12133970200 will distribute between the nine zip codes as following (Table 6.5):

Table 6.5 Census tract to zip code data transformation: distribution of 100 visits between the zip codes based on their population and area.

O_ZIP	Population %	D_ZIP	Area %	Visit Counts
32428	79.17	32425	25.99	**20.58**
32428	79.17	32427	33.38	**26.43**
32428	79.17	32462	40.63	**32.17**
32438	0.74	32425	25.99	**0.19**
32438	0.74	32427	33.38	**0.25**
32438	0.74	32462	40.63	**0.30**
32466	20.09	32425	25.99	**5.22**
32466	20.09	32427	33.38	**6.71**
32466	20.09	32462	40.63	**8.16**
				100.00

The data transformation of social media data to the zip code level was done using a similar algorithm: travel counts *from* certain zip code origin area were prorated based on its population proportion to the matching "place" area. Note that the tourists' origin information (home place) on social media was self-reported on 'Place' level. **Places**, or Census Bureau Places, encompass both 'Incorporated Places' such as cities, towns and villages, as well as 'Census Designated Places (CDPs)'. Generally, places cover a larger area than zip code areas. Here in the transformation, we interpolated places into zip code level, and data loss or inaccuracy is more likely to happen than the transformation from census tract to zip code, which is more of an aggregation process. Since the destinations in the social media data were known with high accuracy, visit counts *to* the certain zip code destination did not need to be transformed.

Thus, we aggregated the cellphone data into 970 zip-code origin areas and 970 zip-code destination areas. Similarly, we aggregated the social media data into 954 zip-code origin areas and 838 zip-code destination areas .

Table 6.6 Data structures of cellphone and social media data in zip-code level

	Unique zip code destination	Unique zip code origins	Unique zip code OD travel flows
Cellphone data	970	970	580,022
Social media data	954	838	305,829

We enriched the zip code level trip data with the socio-economic attributes such as population, house unit, median household income, as well as tourism-related attributes such as hotel number, attraction number, etc. The final trip database hence contains the following fields:

- O_ZIP: origin zip code
- D_ZIP: destination zip code
- OD_ZP_Count: OD visit counts
- O_ZPOP: population in origin zip code
- O_ZHU: house units in origin zip code

- O_ZAREA: area coverage in origin zip code
- N_Rev_Att: number of attraction reviews in destination zip code
- N_Rev_Hotel: number of hotel reviews in destination zip code
- N_BeachAccess: number of beach accesses in destination zip code
- N_ThemePark: number of theme parks in destination zip code
- N_AirbnbRoom: number of Airbnb rooms in destination zip code
- N_HotelRoom: number of hotel rooms in destination zip code
- Yes_ThemePark: whether the destination zip code has theme parks
- Yes_Beach: whether the destination zip code has beach access
- O_ZLAT: centroid latitude of origin zip code
- O_ZLON: centroid longitude of origin zip code
- D_ZLAT: centroid latitude of destination zip code
- D_ZLON: centroid longitude of destination zip code
- D_Total: total visits to the destination zip code
- OD_ZDIS: distance between origin and destination zip codes
- O_Total: total visits starting from the origin zip code

6.2.3 Data validation

After removing the short-distance trips (recall that only the trips at least 50-mile-long were defined as tourism travel), the trips from and to the top zip codes were distributed as shown in Table 6.7. To validate the travel flow data at a zip code level, we conducted a correlation test on the OD travel flows between cellphone data and social media data, the result shows that the destination visits are strongly correlated (R=0.72); while the correlation of origins is acceptable (R= 0.53). Compared with the similar cross-validation we conducted in Section 5.3, we noticed that the correlation is decreased at a more refined analysis unit. Overall, we found that the social media and cellphone data are more consistent and more reliable when used to predict the destination counts at a fine resolution, but less reliable for the trip origins. The main reason is the social media origin information is self-reported at a Place level, which we interpolate to the zip code level. Some inaccuracy is inevitable in this process.

Table 6.7 Top origin and destination zip code travel counts

Top destination zip code areas		
ZIP	OD Flow Count	County Name
33125	276,875	Miami-Dade
32839	258,750	Orange
32809	192,537	Orange
32819	188,824	Orange
32810	184,357	Orange
33614	177,933	Hillsborough
32835	168,434	Orange
32812	145,503	Orange
32801	142,948	Orange
32822	141,585	Orange
33147	123,649	Miami-Dade
34787	120,748	Orange
32803	119,662	Orange
34741	118,771	Orange
33612	103,357	Hillsborough
32805	98,845	Orange
33012	88,291	Miami-Dade
33167	87,751	Miami-Dade
32824	87,430	Orange
32821	83,971	Orange

Top origin zip code areas		
ZIP	OD Flow Count	County Name
34953	97,324	St. Lucie
33603	85,121	Hillsborough
33612	68,206	Hillsborough
33614	63,414	Hillsborough
33411	60,621	Palm Beach
33825	59,628	Highlands
32174	58,890	Volusia
32907	58,747	Brevard
32304	57,246	Leon
34952	56,785	St. Lucie
34983	56,780	St. Lucie
33458	53,355	Palm Beach
34972	53,076	Okeechobee
32210	53,057	Duval
32137	53,033	Flagler
33440	52,964	Hendry
32244	52,242	Duval
32114	51,947	Volusia
32608	51,813	Alachua
34997	51,231	Martin

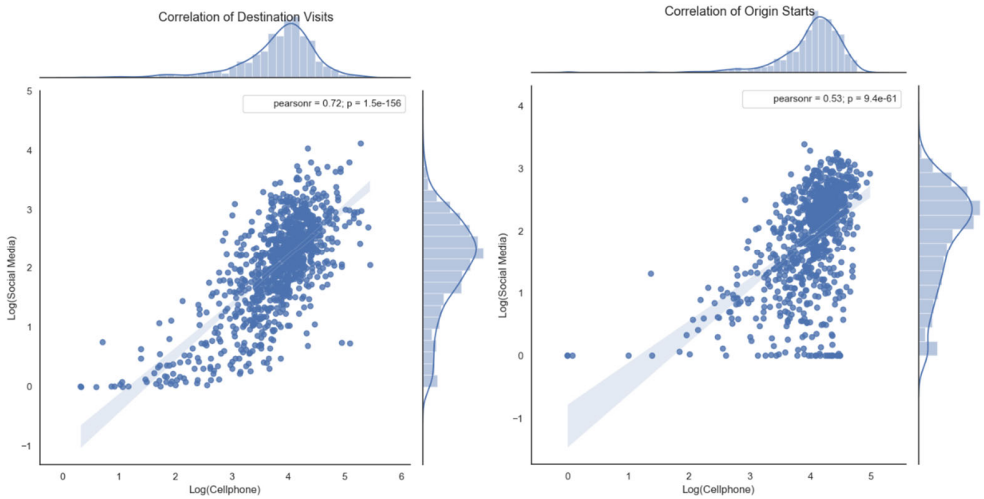


Figure 6.3. Correlation between the cell phone and social media estimations of trip counts for destinations (left) and origins (right) at a zip-code level.

6.2.4 Model evaluation

Unlike the model evaluation at the census tract data, which used a sample of data, this section is using the entire dataset for model evaluation reflecting a smaller number of zip codes compared with the number of census tracts (compare 1,572,611 census tract travel flows to 580,022 zip-code travel flows). The dependent and independent variables in the presented models are similar to those in used in Section 6.1.4, except for the urban-rural interaction variables which are not available at the zip-code level. In addition, the dependent variable $Volume_{ij}$ variable is using the actual OD counts, since the data in the model is no longer a sample but the total observed number of trips.

Basic model

The basic model is using only one independent variable, distance impedance:

$$\log(Volume_{ij}) = \beta_1 \log(d_{ij}) \qquad (6.23)$$

The parameters of this model m0 show that distance is a significant predictor variable with a negative impact:

```
MODEL FIT:
F(1,580020) = 82326.1592, p = 0.0000
R² = 0.1243
Adj. R² = 0.1243

Standard errors: OLS
--------------------------------------------------------------------
                             Est.     S.E.      t val.        p
----------------------- --------- -------- ----------- --------
(Intercept)                3.0001   0.0077    388.1604   0.0000
log10(OD_DIS+1)           -1.0231   0.0036   -286.9254   0.0000
--------------------------------------------------------------------
```

Model 1

Model 1 is using three independent variables, distance impedance, median household income, and local population:

$$\log(Volume_{ij}) = \beta_1 \log(d_{ij}) + \beta_2 \log(inc_i) + \beta_3 \log(Pop_i) \qquad (6.24)$$

The distance still has a negative coefficient while the local population acts as a strong positive factor in stimulating travels. Interestingly, the income level of the origin has a negative coefficient.

```
MODEL FIT:
F(3,580018) = 63003.1265, p = 0.0000
R² = 0.2458
Adj. R² = 0.2458

Standard errors: OLS
-------------------------------------------------------------------
                             Est.     S.E.      t val.        p
------------------------ --------- -------- ----------- --------
(Intercept)                2.6332   0.0079    334.2578   0.0000
log10(O_POP + 1)           0.4144   0.0016    266.5565   0.0000
log10(OD_DIS + 1)         -1.0878   0.0033   -327.9651   0.0000
log10(O_INC + 1)          -0.2562   0.0016   -163.6419   0.0000
-------------------------------------------------------------------
```

Model 2

Model 2 is a tourism-specific model, assuming tourism attractions and accommodation facilities are two major drivers of tourist travels:

$$\log(Volume_{ij}) = \beta_1 \log(d_{ij}) + \beta_4 \log(attraction_j) + \beta_5 \log(hotel_j) \quad (6.25)$$

The model shows that both tourism attractions and hotels have positive coefficients as expected, indicating that destinations with more attractions and accommodation facilities are more likely to attract tourist flows.

```
MODEL FIT:
F(3,580018) = 46711.6446, p = 0.0000
R² = 0.1946
Adj. R² = 0.1946

Standard errors: OLS
-------------------------------------------------------------------
                                Est.     S.E.      t val.        p
--------------------------- --------- -------- ----------- --------
(Intercept)                   2.7557   0.0075    366.0751   0.0000
log10(N_Rev_Hotel + 1)        0.0631   0.0006     98.7110   0.0000
log10(N_Rev_Att + 1)          0.0830   0.0008    103.8243   0.0000
log10(OD_DIS + 1)            -1.0520   0.0034   -307.1207   0.0000
-------------------------------------------------------------------
```

Model 3

Model 3 is an extensive tourism-specific model which takes into account coastal and theme parks:

$$\log(Volume_{ij}) = \beta_1 \log(d_{ij}) + \beta_4 \log(attraction_j) + \beta_5 \log(hotel_j)$$

$$+\beta_6 yes\_themepark_j + \beta_7 yes\_beach_j \quad (6.26)$$

The model coefficients show that the theme park feature has a major influence on the decision to travel, yet the beach variable has a negative coefficient. It is noticeable that beach is not as appealing as theme parks in Floridian's travel decision making, given that the beach feature was a minor influencer in the former modeling of Section 6.4. Nevertheless, the explanation of beach proximity to be a negative factor in tourist travel flow asks for further investigation.

```
MODEL FIT:
F(5,580016) = 31774.0294, p = 0.0000
R² = 0.2150
Adj. R² = 0.2150

Standard errors: OLS
-------------------------------------------------------------------------
                                  Est.     S.E.      t val.        p
--------------------------- --------- -------- ----------- --------
(Intercept)                     2.7440   0.0075    367.5261   0.0000
log10(N_Rev_Hotel + 1)          0.0719   0.0006    112.1896   0.0000
log10(N_Rev_Att + 1)            0.0860   0.0008    104.6162   0.0000
log10(OD_DIS + 1)              -1.0590   0.0034   -312.2708   0.0000
Yes_Beach                      -0.2454   0.0025    -96.6303   0.0000
Yes_ThemePark               0.1514   0.0019     78.0431   0.0000
-------------------------------------------------------------------------
```

Model 4

Model 4 is a comprehensive model that incorporates all explanatory variables:

$$\log(Volume_{ij}) = \beta_1 \log(d_{ij}) + \beta_2 \log(inc_i) + \beta_3 \log(Pop_i) + \beta_4 \log(attraction_j) + \beta_5 \log(hotel_j) + \beta_6 yes\_themepark_j + \beta_7 yes\_beach_j \quad (6.27)$$

The comprehensive model is consistent with the previous stepwise models:

```
MODEL FIT:
F(7,580014) = 42643.1487, p = 0.0000
R² = 0.3398
Adj. R² = 0.3398

Standard errors: OLS
--------------------------------------------------------------------
                                 Est.     S.E.      t val.        p
---------------------------- --------- -------- ----------- --------
(Intercept)                     2.3760   0.0075    317.1207   0.0000
log10(O_ZPOP + 1)               0.4212   0.0015    289.4873   0.0000
log10(O_ZMINC + 1)             -0.2616   0.0015   -178.5676   0.0000
log10(N_Rev_Hotel + 1)          0.0745   0.0006    126.7118   0.0000
log10(N_Rev_Att + 1)            0.0848   0.0008    112.4585   0.0000
log10(OD_ZDIS + 1)             -1.1257   0.0031   -361.0978   0.0000
Yes_Beach                      -0.2584   0.0023   -110.9276   0.0000
Yes_ThemeAmusePark              0.1594   0.0018     89.5900   0.0000
--------------------------------------------------------------------
```

Similar to the outcomes of the census tract level model, distance is a robust impedance variable in traffic estimation models, with a small negative coefficient. Income level in origin zones is also an important socio-economic variable, indicating that financial condition is a positive variable in tourists' travel decisions. Origin-destination interaction especially travels between inter-urban areas, also plays a positive role in generating tourism and leisure-oriented travels. Between the tourism elements, tourism attractions and products increase travel traffic as expected, and destinations with theme parks are extremely powerful in appealing tourists. Accommodation facilities such as hotels and Airbnb, on the other hand, have not shown a positive influence on bolstering tourism travel flows. The results from Section 5 suggest that the main reason for that is that the accommodation facilities and attractions are multicollinear variables; we suggest to include either accommodations or attractions into the model, possibly differentiation between Floridians (attractions are the preferred variable) and long-distance visitors (accommodations are preferred).

6.3 Data collection and warehousing

6.3.1 Cell phone data

Cell phone data format and preprocessing is described in detail in Task 5 (section 2.2.1). The data used in the project represented monthly averages for the number of travel between the census tracts for 12 consecutive months, for workdays and weekdays, with aggregation over time of the day. Generally, such a dataset is very expensive to obtain and too large to process efficiently. In the project, the utility of this dataset was to validate the travel data based on the social media and to explore the possibility of the social media data downscaling. In the future, the cell phone data could be used only periodically to validate possible changes in the social media downscaled data; we also advise to use more generalized data to control costs.

6.3.2 Social media data

The social media data format and preprocessing are described in detail in Task 5 (section 2.1.1). We recommend periodical (on an annual base) update of the social media collection to keep the OD matrix used in destination choice model current. We also recommend adding data collected in the Portuguese and Spanish languages to better represent travelers coming from the Latin American countries. This collection was not planned in the original proposal.

6.3.3 Industry and socioeconomic data

Tourism resource and tourism industry data were collected from a variety of sources, including the Florida Department of Economic Opportunity (FDEO), the University of Florida GeoPlan Center (UFGC), and the geospatial database recently collected by FDOT which describes the location and capacity of all major tourism resources. In particular, the FDOT includes 62 categories of tourism resources, which were from a variety of agencies, organizations, companies, or educational institutions. The twelve tourism resource indices constructed in Task 2 were also used in Task 6 for tourist flows modeling. The industry data include food/beverage (e.g., number of restaurants and drinking places) and lodging/accommodation (e.g., number of hotels and Airbnb properties) industries. Population and median income data were collected from the U.S. Census Bureau. To reflect the local patterns and relationships, all geographic data were collected and aggregated at the census tract level in Florida.

Table 6.8 shows a comprehensive list of datasets, including source and date. The detailed information about tourism resources and tourism resource indices has been explained in the Report of Task 2.

Table 6.8 Data sources

Dataset	Source	Date
Hotel Rooms	STR, UFGC, WS	2014
Airbnb Rooms	AIRDNA	2016
Amusement and Theme Parks	UFGC, FDEO	2016
Restaurants	GRI, FRLA	2016
Drinking places	GRI, FRLA	2016
Other Tourism Resource Data	UFGC, DEO, FAROC, FDOT, FDEP, GRI, FRLA, ESRI, BAR, NPS	2012-2016
Tourism Resource Index	Task 2 of This Project	2019
Population and Median Income	USCB	2018

Note: AirDNA: AIRDNA Inc.; BAR: Bureau of Archaeological Research; ESRI: Environmental Systems Research Institute; FAROC: Florida Association of RV Parks and Campgrounds; FDOT: Florida Department of Transportation; FDEP: Florida Department of Environmental Protection; FRLA: Florida Restaurant and Lodging Association; GRI: Geographic Research Inc.; NPS: National Park Service; STR: Smith Travel Research; UFGC: University of Florida GeoPlan Center; USCB: U.S. Census Bureau; WS: Web-scraping program.

6.4 Relationship between tourist flow and traffic flow

6.4.1 Introduction

As tourism is at its very core a distinctly geographical phenomenon, involving the movement of tourists from one place – their places of origins or generating regions – to one or more destinations via a complex web of multimodal transportation networks (Kang et al., 2014), understanding the correlation between tourist flow and the overall traffic flow is critical for tourism development and transportation planning. Florida is the largest tourism destination worldwide, receiving over 100 million visitors annually and contributing over $89 billion to the state's economy (Lee et al., 2019); importantly, Florida tourism occurs throughout the year. So, the purpose of task 6.4 is to understand the correlation between annual tourist flow and traffic flow in Florida. To achieve the purpose, we (1) examined the correlation between tourist flow and traffic flow and (2) explored and visualized the association between tourist flow and traffic flow. The findings will enable FDOT to understand the role of tourists in the overall traffic flow in Florida.

6.4.2 Variable definition

The variable of traffic flow was defined as the annual traffic volume at the census tract. To generate traffic flow, the annual average daily traffic (AADT) data in 2019 measured by FDOT was converted to the census tract level using the spatial join function in ArcGIS (v. 10.7.1). The spatial join function was used to combine the attributes of different features based on their spatial relationship. In the spatial join, the target features were the AADT data, and the join features were the census tract locational data. The following parameters of the spatial join were used: (1) the intersect match: the features in the join features are matched if they spatially intersect a target feature and (2) the merge rules: sum the number of traffic counts. Based on the above process, we generated the AADT data at the census tract level. Finally, the AADT data were converted to the annual traffic volume by multiplying each value by 365.

The variable of tourist flow was defined as annual tourist flow by census tract, which was measured based on the cell phone data in previous tasks. Figures 6.3 and 6.4 show the distribution of traffic flow and tourist flow in Florida.

6.4.3 Data analysis

To examine the relationship between tourist flow and traffic flow, Pearson's correlation was used. To explore and visualize the association between tourist flow and traffic flow, the ordinary least squires (OLS)-based global regression model (OLS model) and the spatial geographically weighted regression (GWR)-based local regression model (GWR model) were developed using SPSS (version 20.0) and ArcGIS (10.7.1) software, respectively.

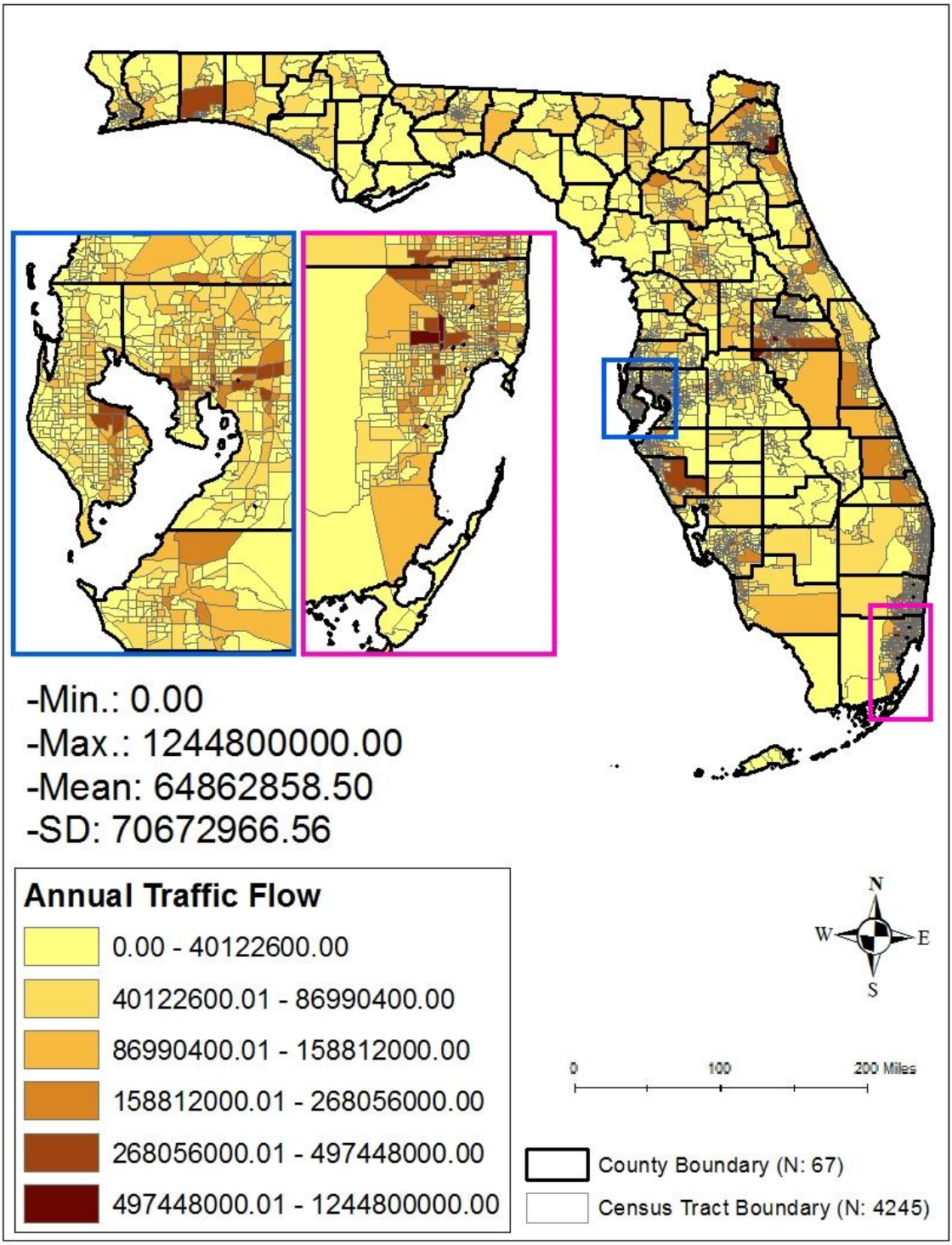


Figure 6.4 Spatial distribution of traffic flow in Florida (annual traffic volume), FDOT data

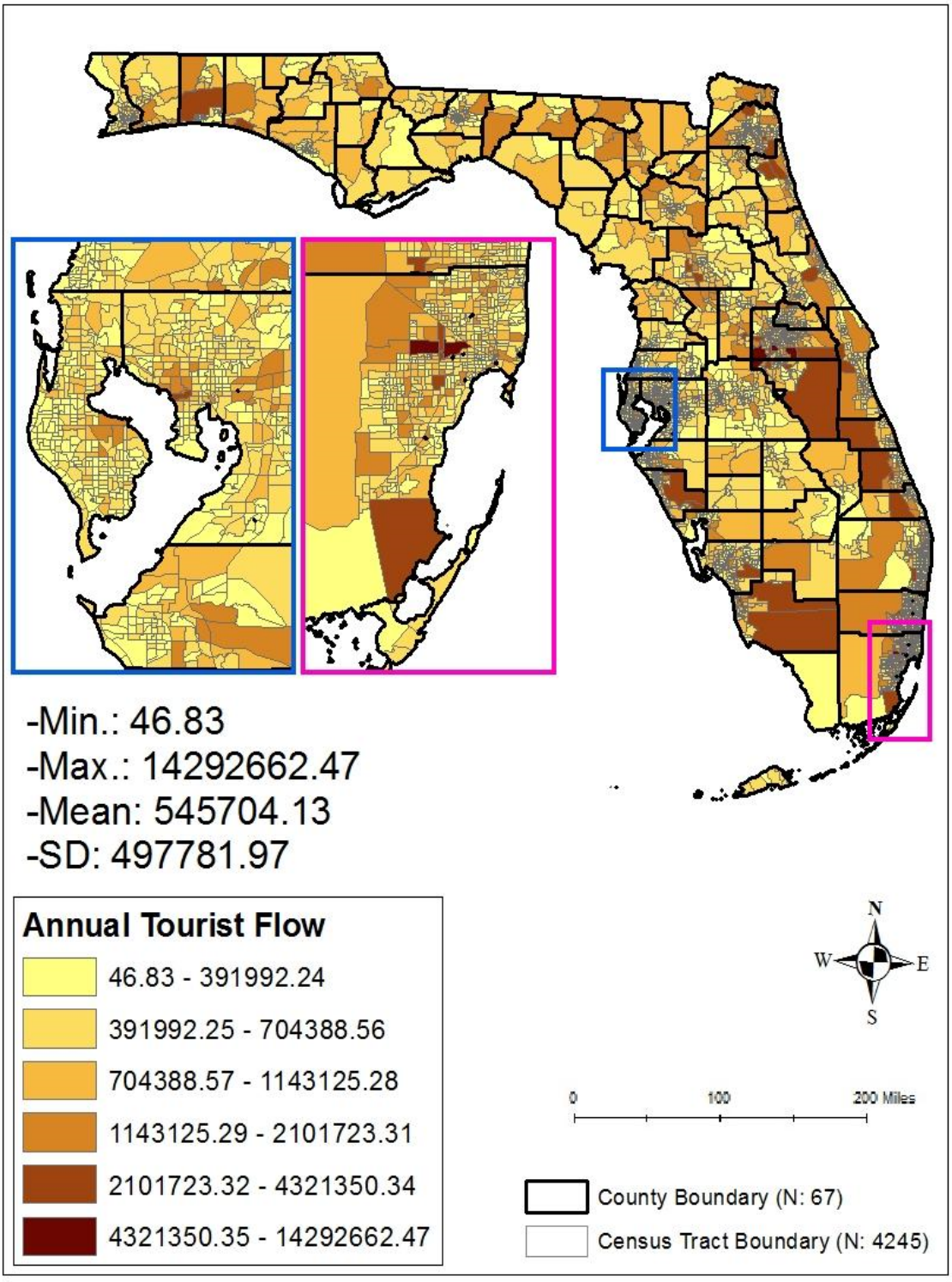


Figure 6.5 Spatial distribution of tourist flow in Florida (annual tourist visitation), cell phone data

6.4.4 Results

The tourist flow from cell phone data and traffic flow from FDOT are positively correlated (Pearson's r= 0.65, p<0.01 – see Table 6.9). The coefficient of determination $R^2 = 0.41$ indicates a moderate goodness-of-fit, that is, a 40% of the variation in the number of cars on the road is in fact explainable by the tourist related traffic. The spatial statistical model (Table 6.10) explicates the variability of that finding over different parts of Florida. In other words, census tracts with a higher level of tourist flow have a higher level of traffic flow, representing that tourist flow has a significant impact on increasing traffic flow. For the GWR model, the value of local R^2 ranged from 0.24 to 0.53 with a mean of 0.45. The local coefficient of tourist flow ranged from 55.26 to 145.94 with a mean of 107.00, indicating that census tracts with a higher tourist flow have a higher traffic flow. Specifically, an increase of one visitation can yield an average annual increase of 107 traffic volumes. Such variability in the local coefficient indicates spatial non-stationarity, which presented spatially heterogeneous association between tourist flow and traffic flow in Florida. Figures 6.5 and 6.6 reveal maps of the distribution of coefficients for tourist flow and local R^2 respectively.

Table 6.9 Correlation between tourist flow and traffic Flow

		Tourist Flow	Traffic Flow
Tourist Flow	Pearson Correlation	1	.648**
	Sig. (2-tailed)		0.000
	N	4215	4215
Traffic Flow	Pearson Correlation	.648**	1
	Sig. (2-tailed)	0.000	
	N	4215	4215

**. Correlation is significant at the 0.01 level (2-tailed).

Table 6.10 Results of OLS and GWR models

Variable	OLS estimate	GWR estimate			
		Min.	Mean	Max.	Range
Intercept	14,646,467.36	-8,512,817.05	9,844,101.75	31,184,867.39	36,697,684.44
Tourist Flow	92.02**	55.26	107.00	145.94	90.68
Local R^2	0.41	0.24	0.45	0.53	0.29
Condition Index		2.03	2.80	3.69	3.66
AIC_c	162,031.16		161,483.75		

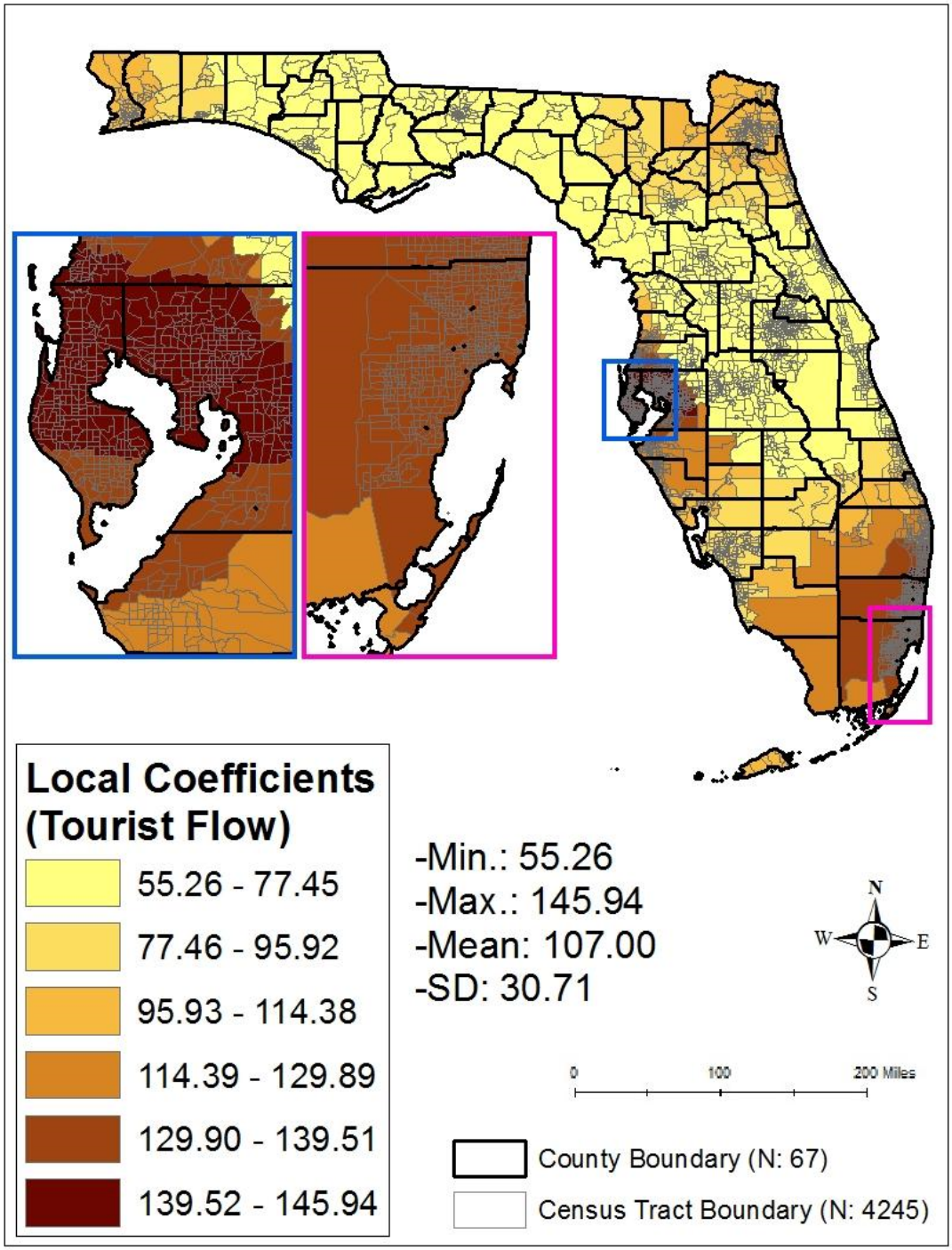


Figure 6.6 Spatial distribution of local coefficients for tourist flow

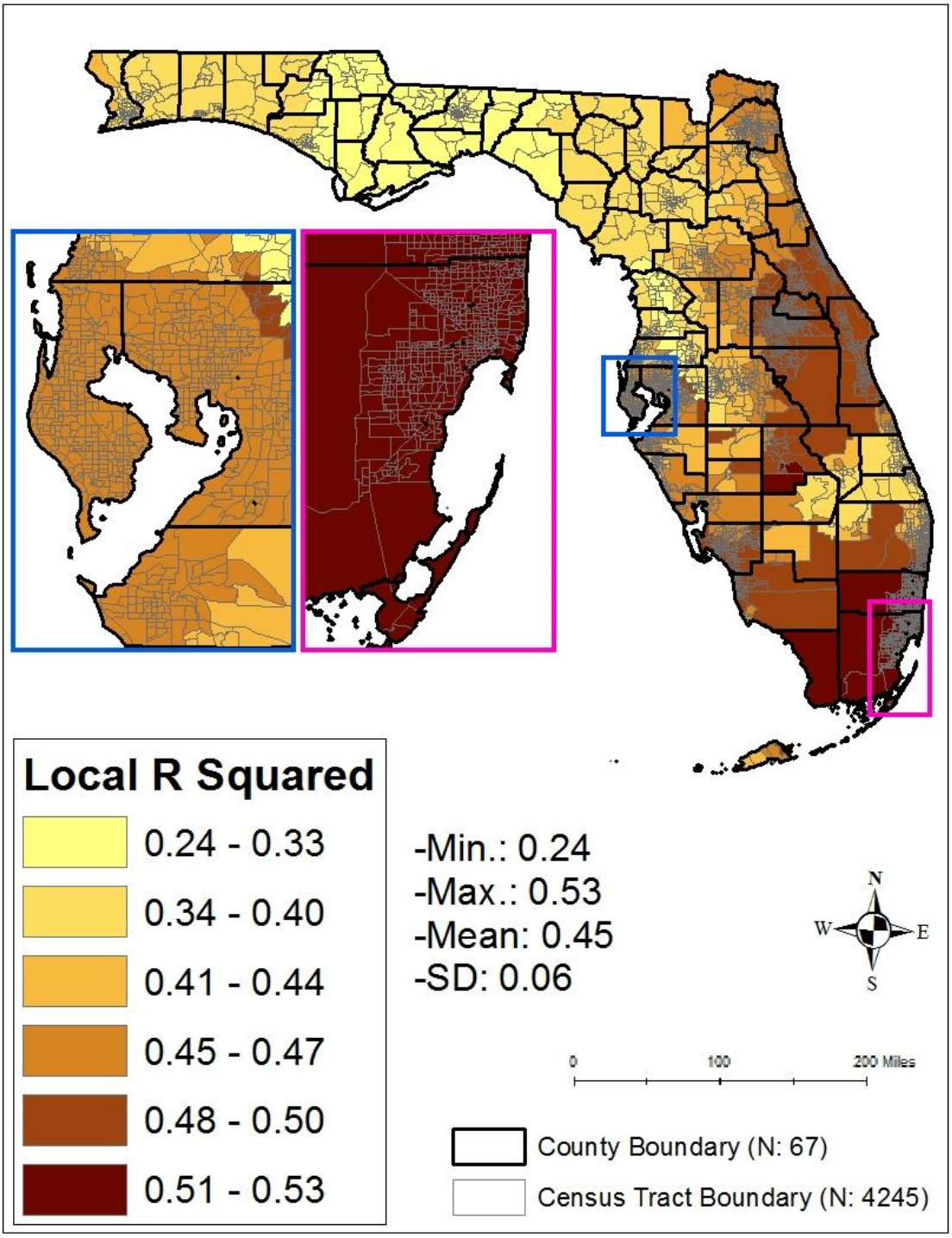


Figure 6.7 Spatial distribution of local R^2

6.5 Tourist flows model 1

6.5.1 Introduction

Tourist flows modeling is important for estimating the number of visitors. Tourist flows are spatial interactions between locations and affected by a variety of variables, including tourism resources, tourism industries and traffic flows. Thus, identifying the key determinants of tourist flows and related tourist flow modeling is critical for forecasting tourism visitation. The purpose of Task 6.5 is to build a model of tourist flows. To achieve the task, two objectives were identified.

- Identify the key determinants of tourist flows in Florida
- Explore and visualize the spatially varying relationships between identified key determinants and tourist flows in Florida.

Notice that Model 1 considers only the supply side. Next Model 2 in the next section builds on Model 1 considering the demand and supply sides both.

6.5.2 Variable definition

The dependent variable was the annual tourist flow, which was operationally defined as log-transformed annual traffic volume based on cell phone data in previous tasks. The independent variables included tourism resources, tourism industries and transportation. All dependent and independent variables and their operational definitions are summarized in Table 6.11.

6.5.3 Data analysis

Various software programs, including ArcGIS (v. 10.7.1), the ArcGIS Spatial Statistics Tool extension, SPSS (v. 20.0), and GWR (v. 4.0), were employed for the data analysis. First, Pearson's correlation was used to explore collinear variables. The strongest identified correlations ($r > 0.70$, $p<0.01$) were among Factor 2 (Urban Tourism) Supply Index, hotel industry, Airbnb industry, food industry and beverage industry. Thus, the variables of hotel industry, Airbnb industry, food industry and beverage industry were excluded to avoid the potential multicollinearity issue.

Second, a multiple regression analysis was performed using OLS to investigate the relationship between the tourist flows and the tourism resources and transportation-related variables. However, some variables such as Factor 1 (Water/Park-based Tourism and Recreation) Supply Index, Factor 3 (Recreational Boating Tourism) Supply Index, Factor 7 (Horse/Race Tract) Supply Index, Factor 8 (Aquarium/Zoo Tourism) Supply Index, Factor 9 (Theme Park/Casino Tourism) Index, Factor 10 (Sport Tourism) Supply Index, Factor 11 (Garden Tourism) Supply Index, Factor 12 (MICE Tourism) Supply Index, and Highway Accessibility were found to be not statistically significant and excluded. Therefore, the variables of Factor 2 (Urban Tourism) Supply Index, Factor 4 (Beach Tourism) Supply Index, Factor 5 (Golf Tourism) Supply Index, Factor 6 (RV and Camping) Supply Index and Traffic Flow were finally employed for regression analyses. Figures 6.7 – 6.11 show the distribution of dependent and finalized independent and control variables for regression analyses.

Third, the dependent variable and identified independent and control variables were utilized in the GWR to explore local relationships between the independent and dependent variables. While employing the GWR, a bi-square kernel was employed with adaptive bandwidth selecting the window width which

maximizes model fit. To explore spatially varying relationships among variables, local coefficients and R^2 from GWR were mapped.

Finally, the statistical diagnostics, such as R^2 and AIC from the OLS and GWR, were compared to demonstrate the utility of GWR-based spatial models.

Table 6.11 Operationalization of dependent, independent, and control variables

Variable	Operationalized definition
Dependent variable	
Tourist Flow	Log-transformed annual traffic count based on cellphone data for census tract (CT)
Tourism resource variables (IV)	
Factor 1 (Water/Park-based Tourism and Recreation) Supply Index (SI)	CT-based standardized score for factor 1
Factor 2 (Urban Tourism) SI	CT-based standardized score for factor 2
Factor 3 (Recreational Boating Tourism) SI	CT-based standardized score for factor 3
Factor 4 (Beach Tourism) SI	CT-based standardized score for factor 4
Factor 5 (Golf Tourism) SI	CT-based standardized score for factor 5
Factor 6 (RV and Camping) SI	CT-based standardized score for factor 6
Factor 7 (Horse/Race Tract) SI	CT-based standardized score for factor 7
Factor 8 (Aquarium/Zoo Tourism) SI	CT-based standardized score for factor 8
Factor 9 (Theme Park/Casino Tourism) SI	CT-based standardized score for factor 9
Factor 10 (Sport Tourism) SI	CT-based standardized score for factor 10
Factor 11 (Garden Tourism) SI	CT-based standardized score for factor 11
Factor 12 (MICE Tourism) SI	CT-based standardized score for factor 12
Tourism Industry variables (IV)	
Hotel Industry	Log-transformed number of hotel rooms for CT
Airbnb Industry	Log-transformed number of Airbnb rooms for CT
Food Industry	Log-transformed number of restaurants for CT
Beverage Industry	Log-transformed number of drinking places for CT
Transportation variables (CV)	
Highway Availability	Log-transformed total distance of highway systems for CT
Traffic Flow	Log-transformed annual traffic volume for CT

Note. IV: independent variable; CV: control variable

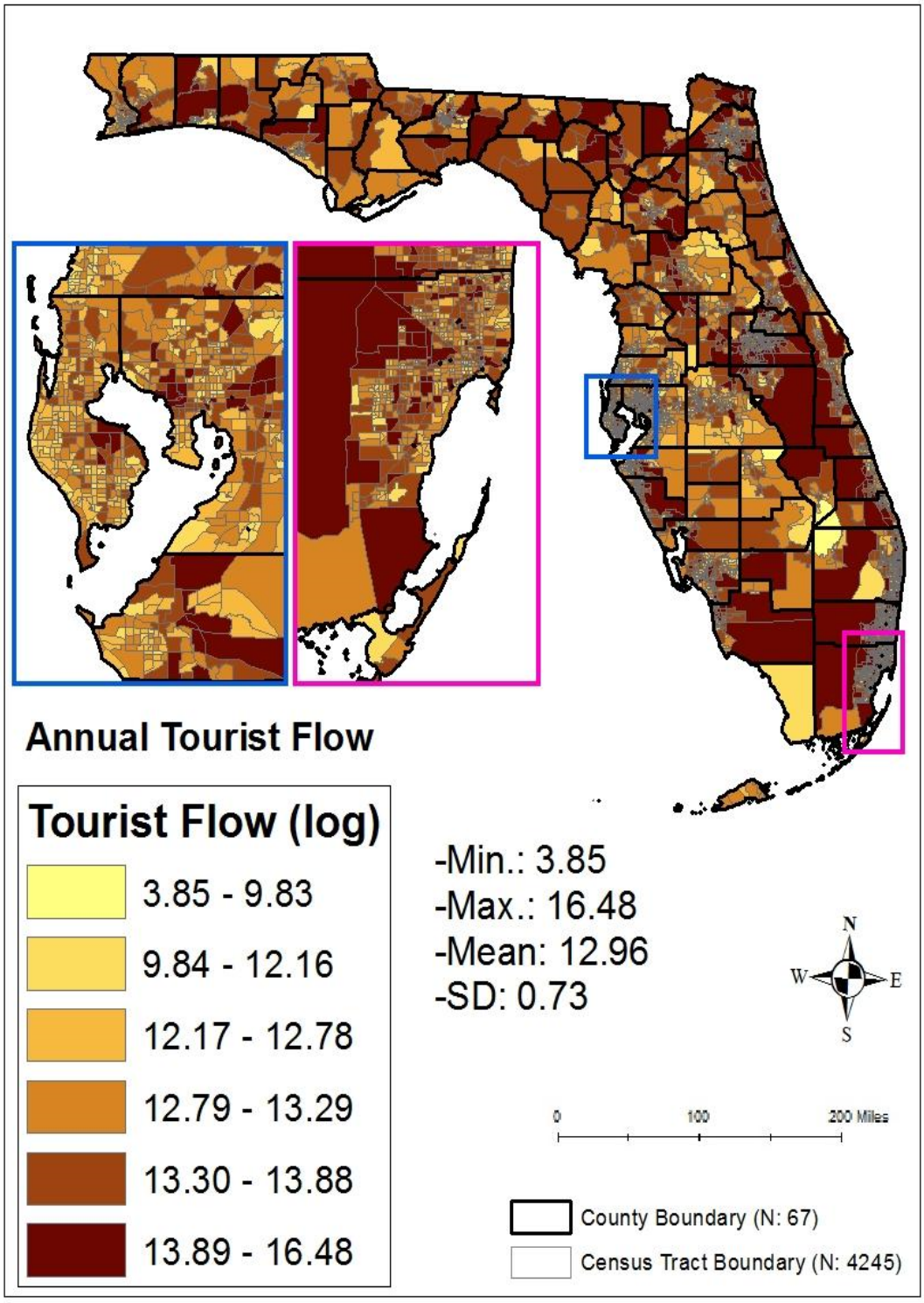


Figure 6.8 Spatial distribution of tourist flow in Florida

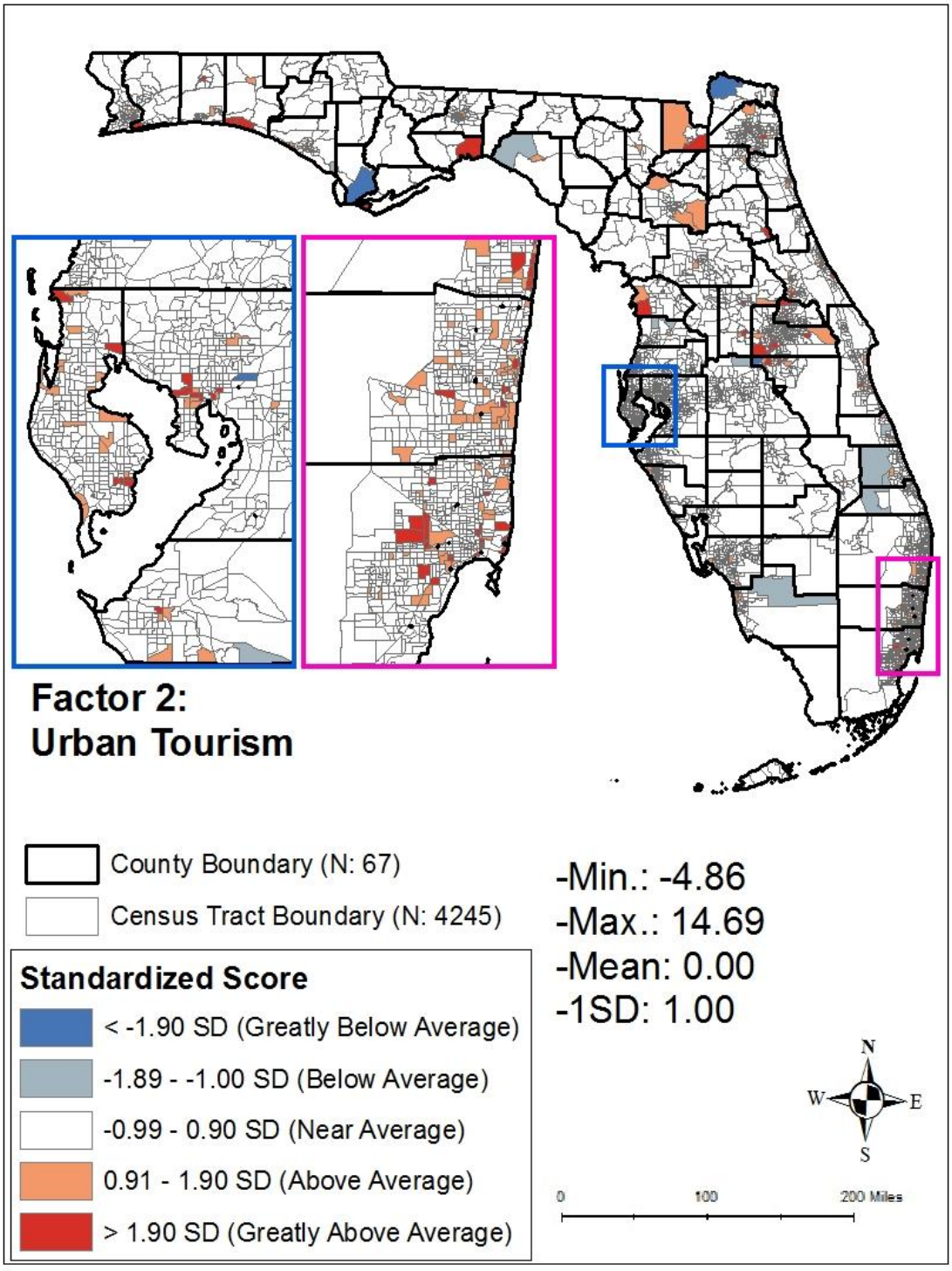


Figure 6.9 Spatial distribution of Factor 2 (Urban Tourism) supply index in Florida

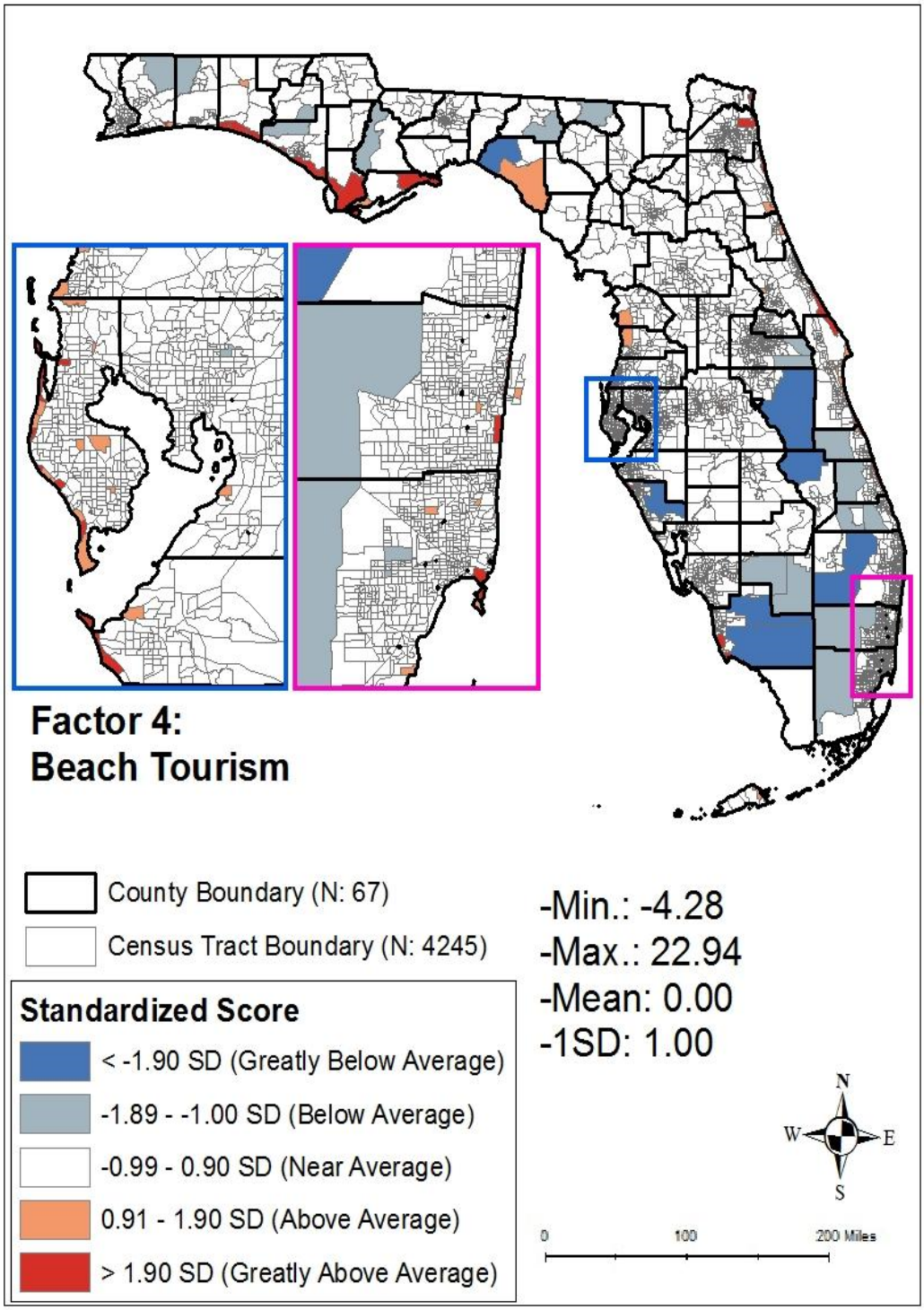


Figure 6.10 Spatial distribution of Factor 4 (Beach Tourism) supply index in Florida

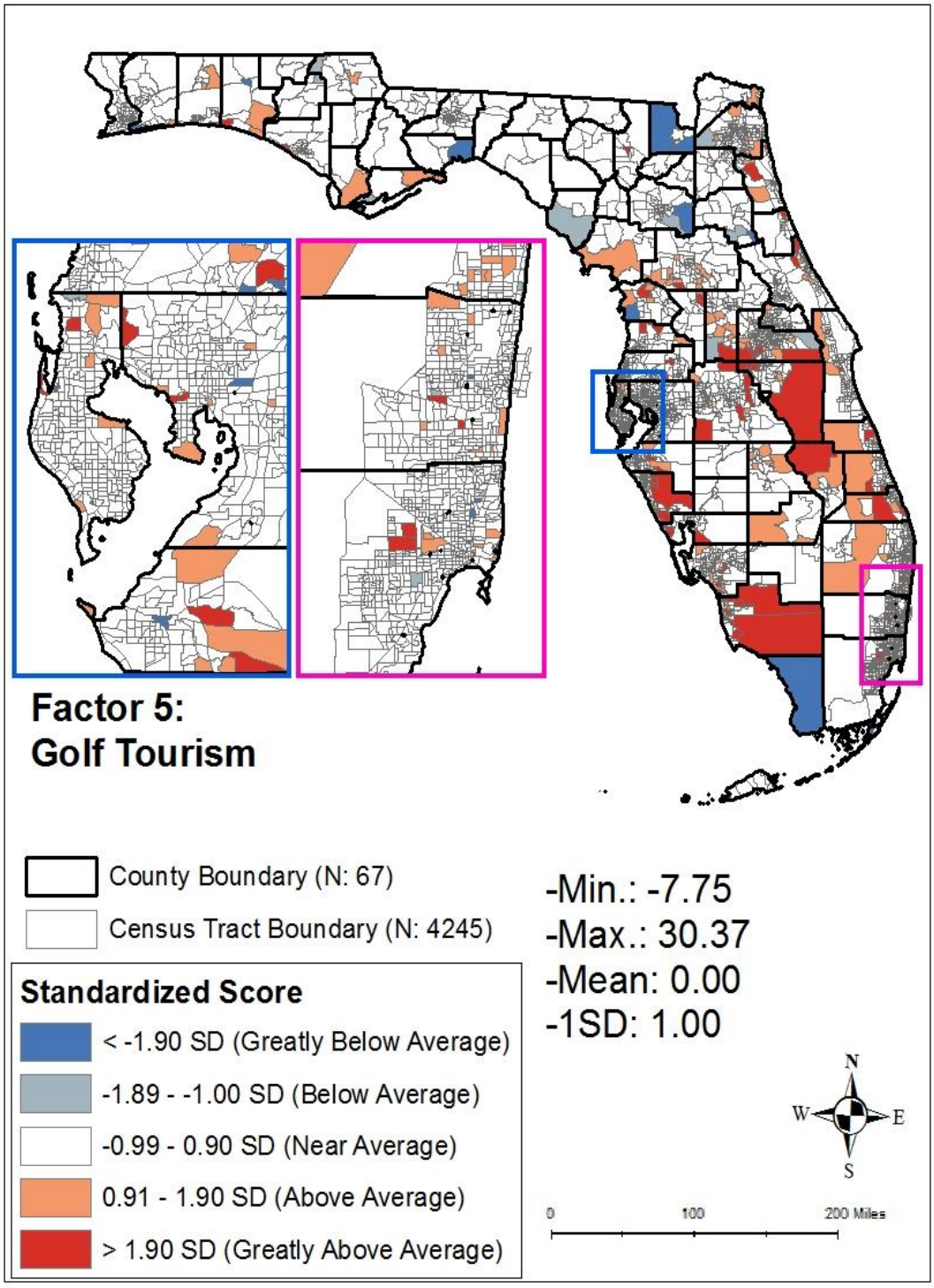


Figure 6.11 Spatial distribution of Factor 5 (Golf Tourism) supply index

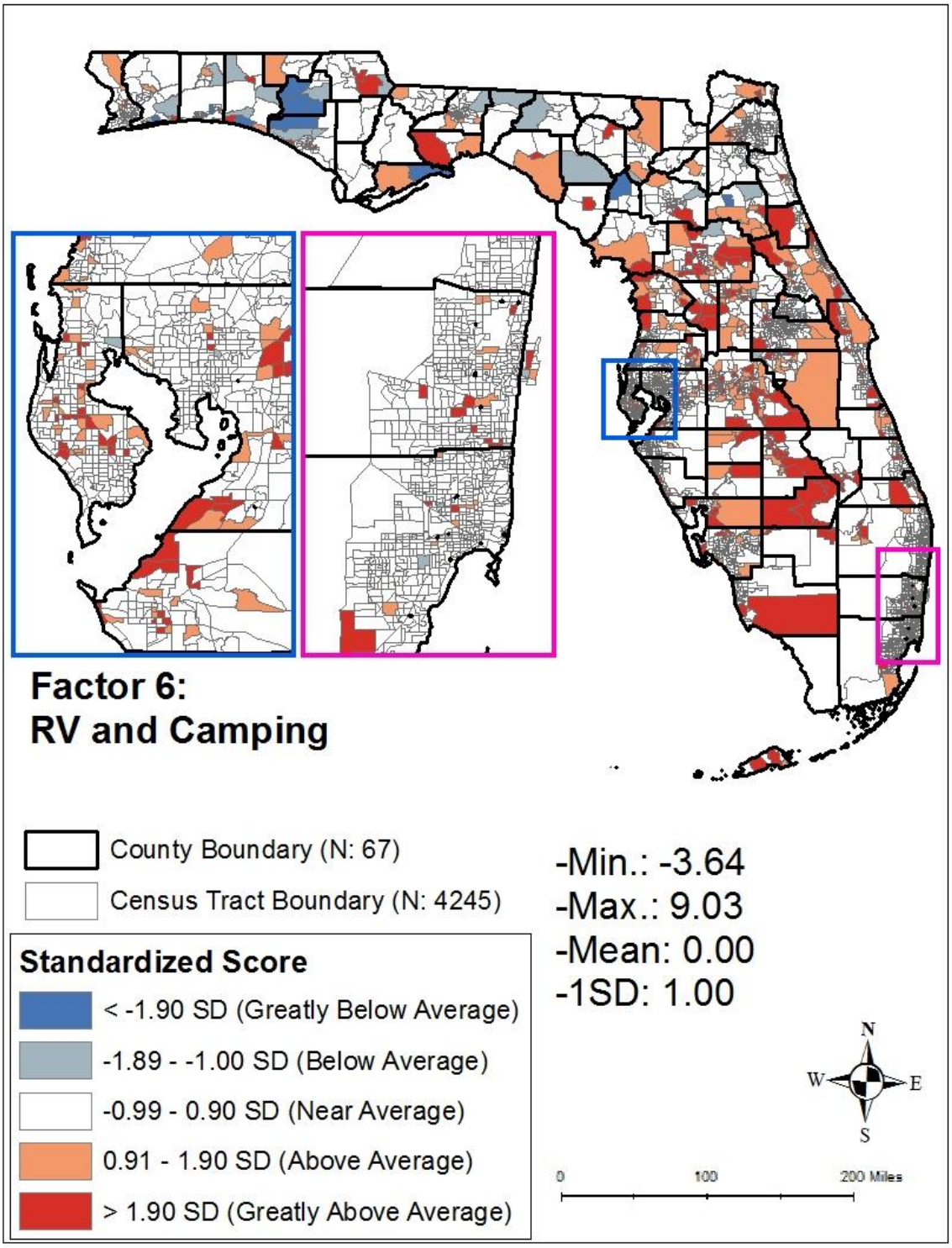


Figure 6.12 Spatial distribution of RV and Camping Supply Index in Florida

6.5.4 Results

Table 6.12 summarizes the outcomes of the OLS and GWR models. For the OLS model, the value of R^2 (0.37) indicated a moderate goodness-of-fit. All independent (Factor 2, 4-6) and control (traffic flow) variables were statistically significant ($p<0.05$). Specifically, the coefficients on the Factor 2: Urban Tourism (0.11), Factor 4: Beach Tourism (0.01), Factor 5: Golf Tourism (0.13), Factor 6: RV and Camping (0.08), and traffic flow (0.40) were of the predicted sign and positively associated with the tourist flows. In other words, census tracts with higher levels of urban tourism, beach tourism, golf tourism, and RV/Camping-related tourism resources had higher levels of tourist flow, indicating that specific tourism resources and traffic flow may significantly influence the tourist flow.

For the GWR model, the value of local R^2 ranged from 0.26 to 0.79 with a mean of 0.43. The values of the local condition index ranged from 9.97 to 24.36, which showed a lack of local collinearity issue in the model (local condition index<30). The local coefficients of the exploratory and control variables ranged from 0.08 to 0.38 with a mean of 0.19 for Factor 2 (Urban Tourism), -0.30 to 0.19 with a mean of 0.01 for Factor 4 (Beach Tourism), 0.02 to 0.40 with a mean of 0.19 for Factor 5 (Golf Tourism), 0.03 to 0.20 with a mean of 0.10 for Factor 6 (RV and Camping), and 0.08 to 0.39 with a mean of 0.18 for Traffic Flow. Figures 6.12 – 6.16 visualize the distribution of local coefficients for the significant tourism resources and local R^2 in Florida. Such variability in the local parameter estimates indicates significant spatial variability or spatially varying association between variables across Florida.

Table 6.12 Results of OLS and GWR models

Variable	OLS estimate	GWR estimate			
		Min.	Mean	Max.	Range
Intercept	6.05*	6.31	9.85	11.61	5.30
Factor 2 (Urban Tourism) Supply Index	0.11*	0.08	0.19	0.38	0.30
Factor 4 (Beach Tourism) Supply Index	0.01*	-0.30	0.01	0.19	0.49
Factor 5 (Golf Tourism) Supply Index	0.13*	0.02	0.19	0.40	0.38
Factor 6 (RV and Camping) Supply Index	0.08*	0.03	0.10	0.20	0.17
Traffic Flow	0.40*	0.08	0.18	0.39	0.31
Local R^2	0.37	0.26	0.43	0.79	0.53
Condition Index		9.97	20.12	24.36	14.39
AIC_c	6,501.23		6,463.50		

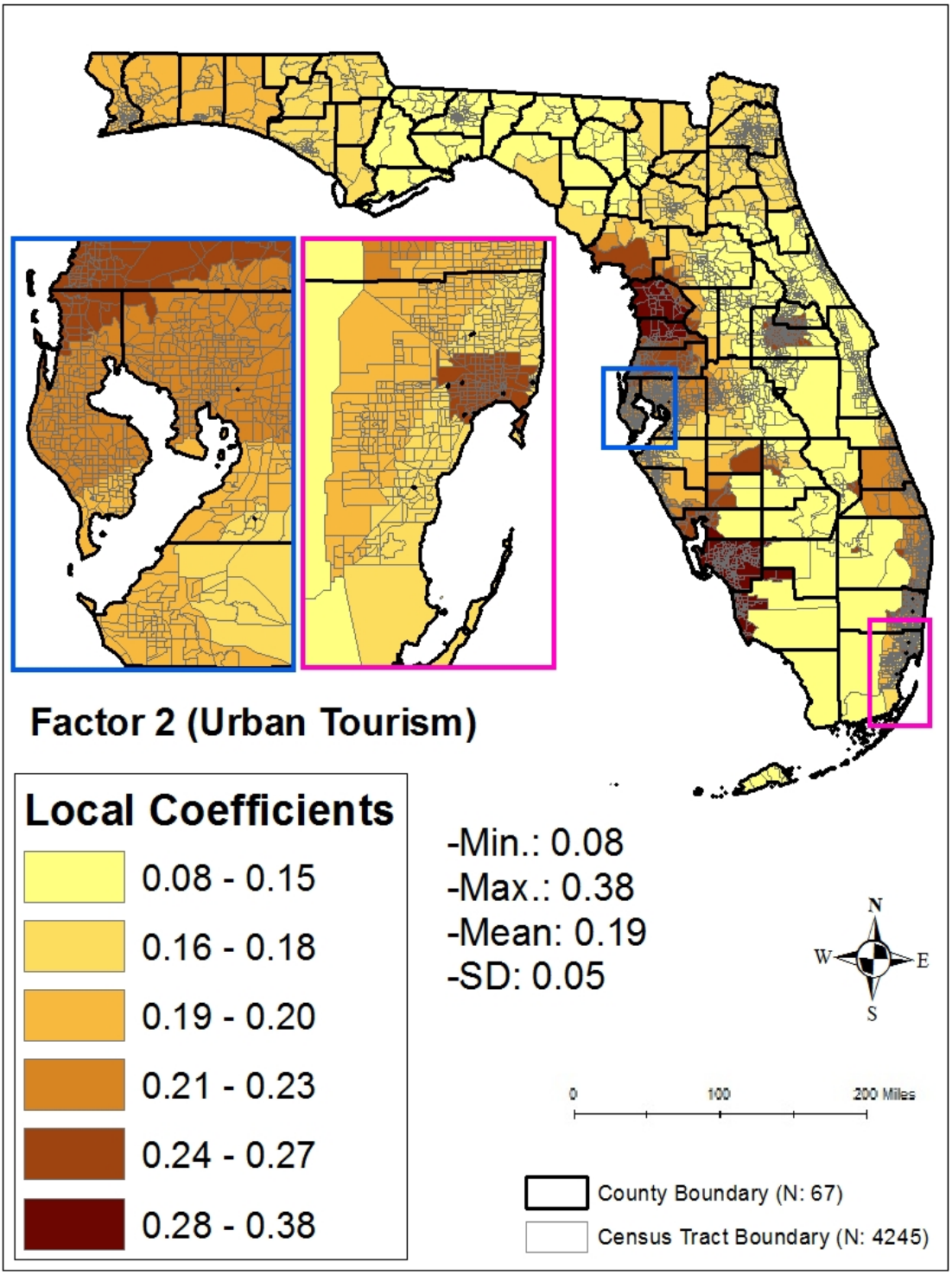


Figure 6.13 Spatial distribution of local coefficients for Factor 2 (Urban Tourism) in Florida

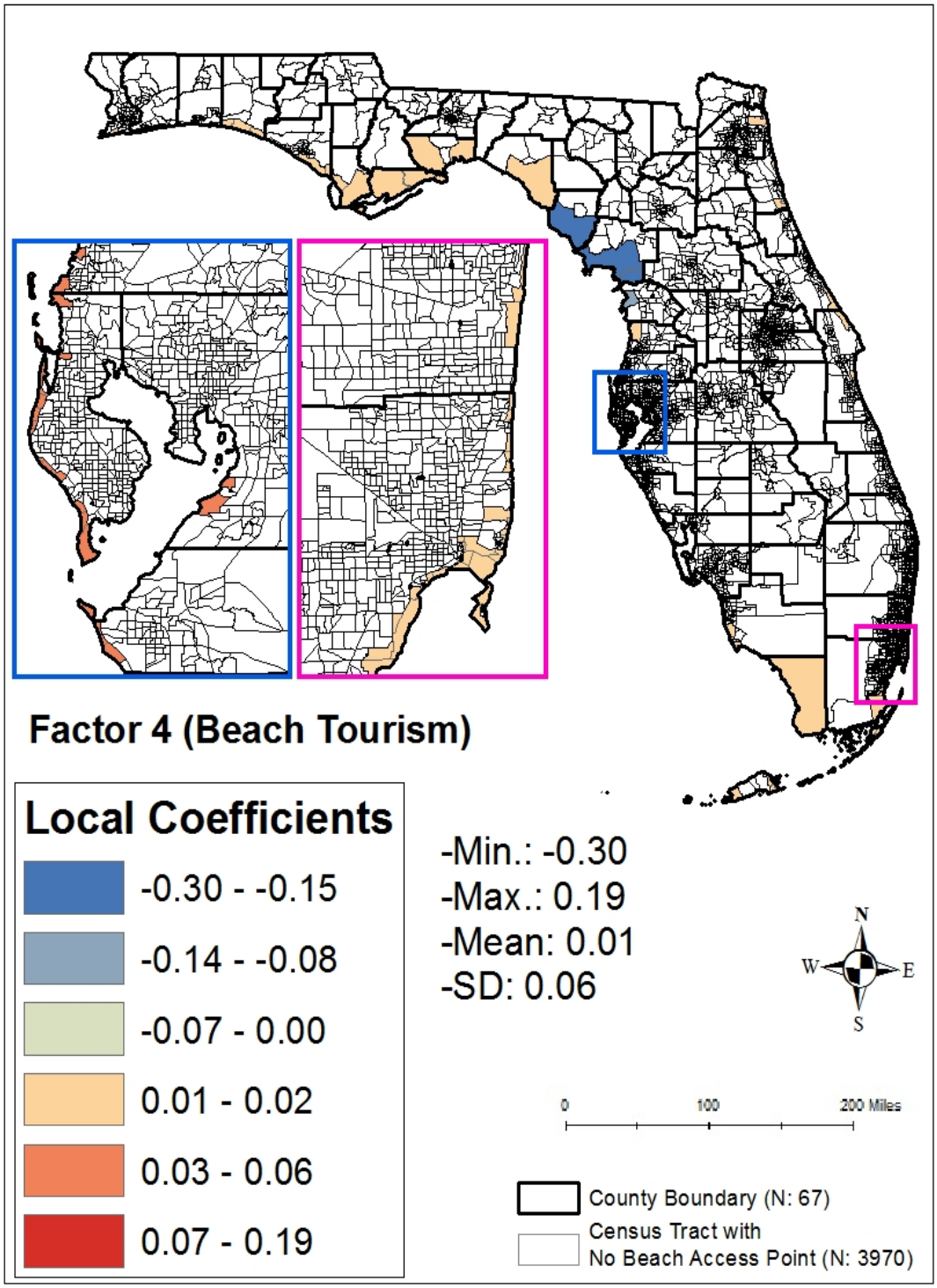


Figure 6.14 Spatial distribution of local coefficients for Factor 4 (Beach Tourism) in Florida

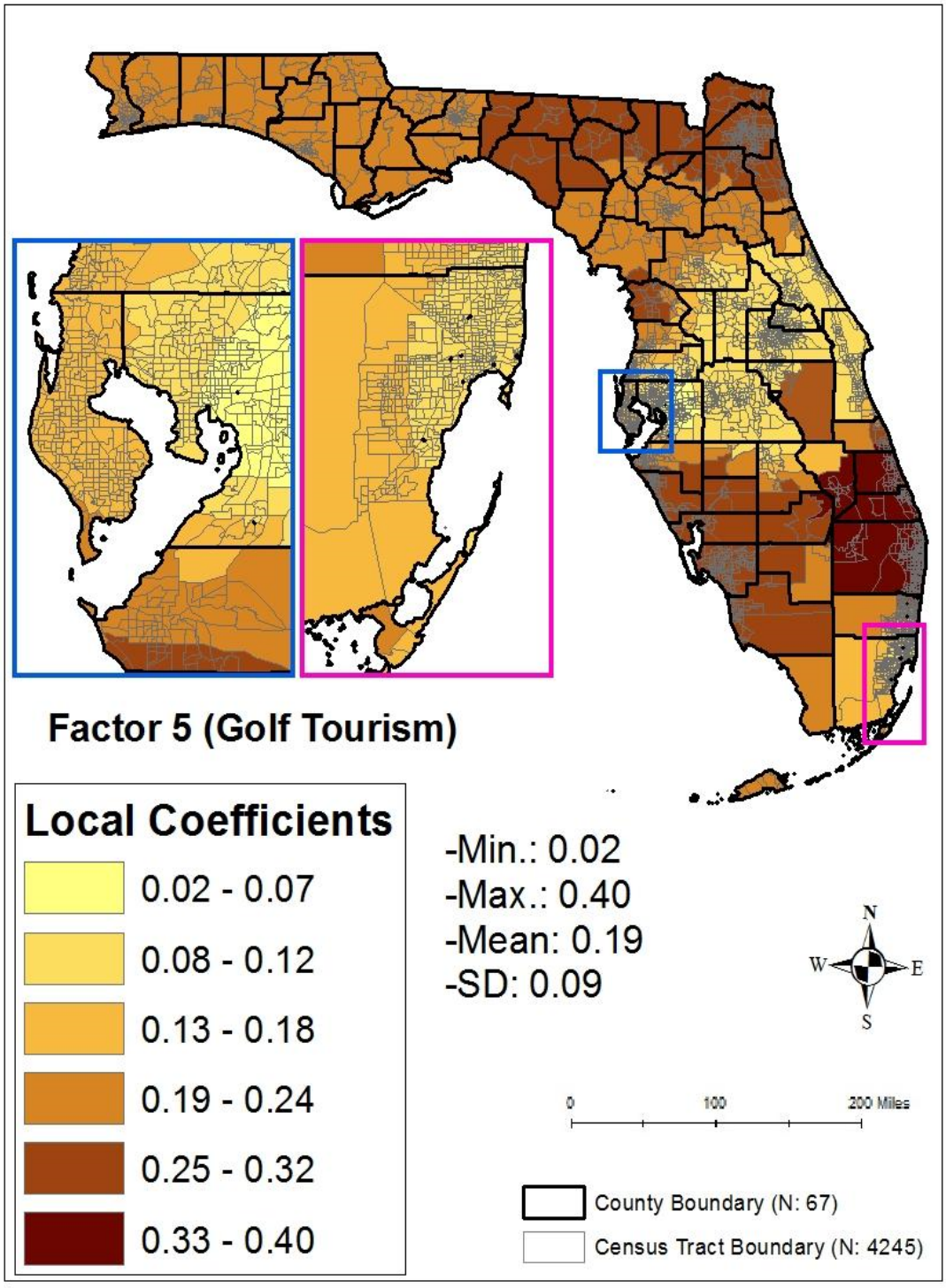


Figure 6.15 Spatial distribution of local coefficients for Factor 5 (Golf Tourism) in Florida

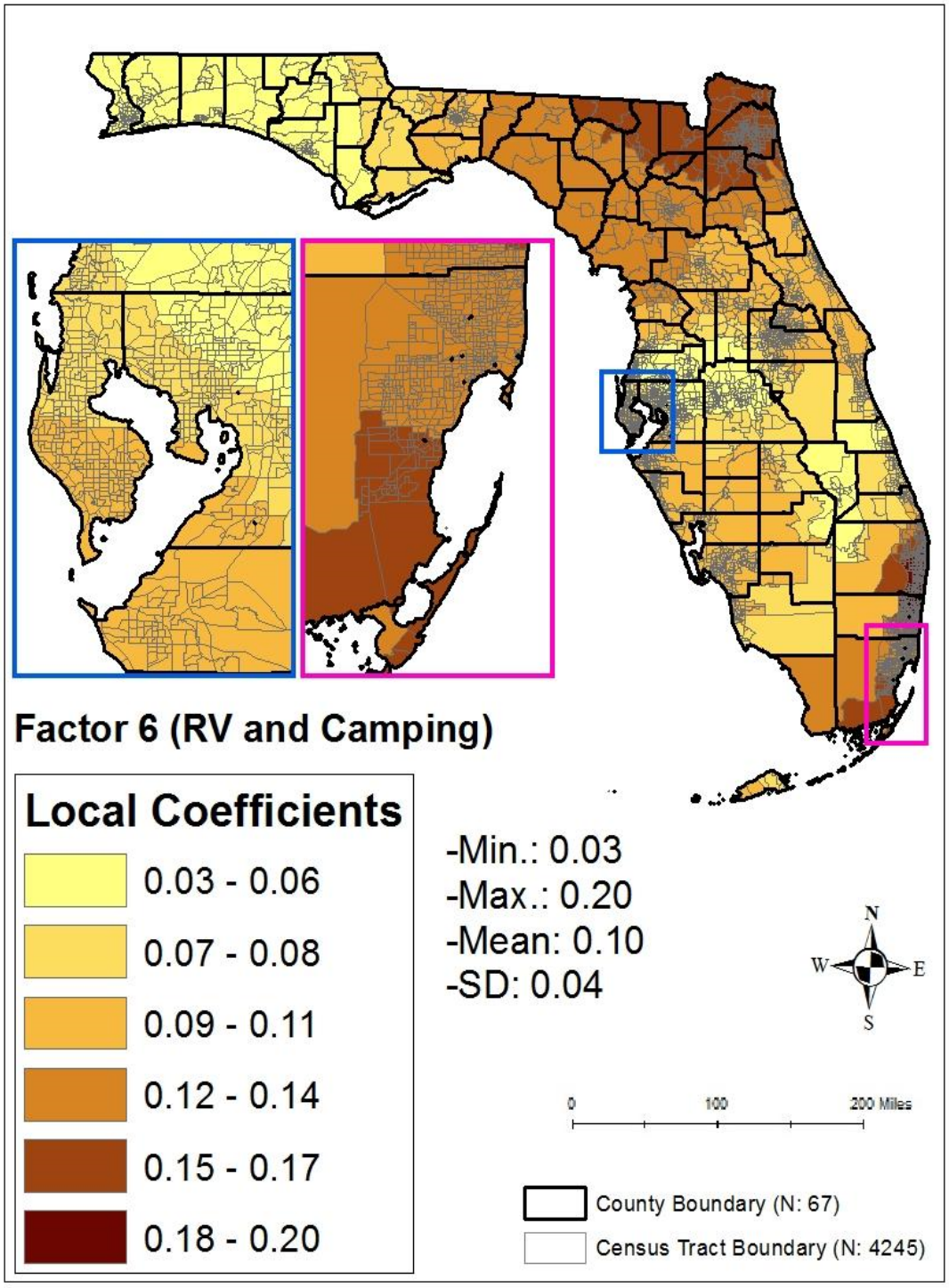


Figure 6.16 Spatial distribution of local coefficients for Factor 6 (RV and Camping) in Tourism

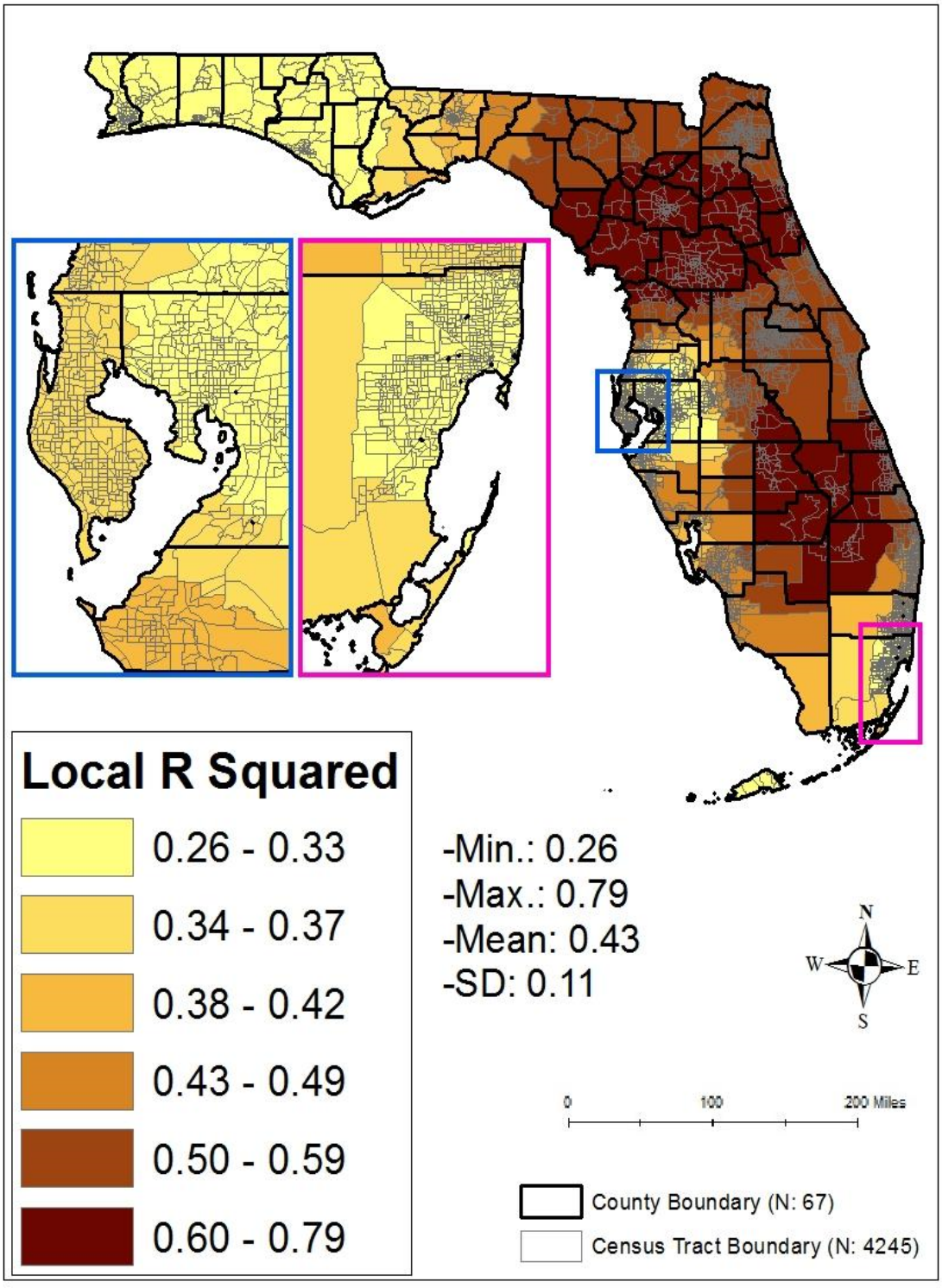


Figure 6.17 Spatial distribution of local R^2

6.6 Tourist flows model 2

6.6.1 Variable description

In this task, we considered both demand and supply aspects of tourist destinations to estimate the number of visitations. The number of Airbnb rooms and the number of hotel rooms can be regarded as the tourism supply side, and the number of reviews of tourist attractions can be normally used as the tourism demand side (Dogru et al., 2020). Furthermore, the average distance from origins to destinations can also be considered to estimate visitations since travel distance significantly affects travel decisions (Kah et al., 2016). Therefore, (1) reviews of attractions, (2) OD average distance, (3) Airbnb rooms, and (4) hotel rooms were defined as the independent variables. In addition, we used the latitude variable to account for shorter travel to reach northern Florida and Panhandle. For the dependent variables, this model considered the actual number of tourist visits by zip code obtained from cellphone data. All variables except for the latitude variable were logarithmically transformed to make data conform to normality (Feng et al., 2014). Table 1 summarizes all variables and operational definitions used in this Task.

Table 6.13 Model 2 variables

Dependent variable		Independent variables	
Variable	**Operational definition**	**Variable**	**Operational definition**
Visitation	Log (number of destination visits+1)	Reviews of attractions	Log (number of reviews of attractions+1)
		OD average distance	Log (average distance from origins to each destination+1)
		Airbnb rooms	Log (number of Airbnb rooms+1)
		Hotel rooms	Log (number of hotel rooms+1)
		Latitude	Latitude of an individual zip code

6.6.2 GWR Analysis

The proposed GWR model is as follows:

$$Visitation_i = \beta_{i0}(u_i, v_i) + \beta_{ik}(u_i, v_i)Tourism_{ik} + \beta_i(u_i, v_i)Latitude_i + \varepsilon_i \quad (6.28),$$

where $Visitation_i$ refers to the number of visitations at zip code i; (u_i, v_i) is the coordinate of the centroid at zip code i; $\beta_{ik}(u_i, v_i)$ is the local regression coefficient for the independent variable k at zip code i; and $\beta_i(u_i, v_i)$ is the local regression coefficient for the control variable, latitude, at zip code i. Lastly, the local coefficients and adjusted R^2 values from GWR models were mapped to visualize the relationships between visitations and independent variables (i.e., reviews of attractions, OD average distance, Airbnb rooms, and hotel rooms). First, the ordinary least squares (OLS) regression was used to identify statistically significant variables before conducting GWR and investigate the global relationship between variables. Then, ArcGIS GWR 4.0 was used for GWR analysis.

6.6.3 GWR model results

The results of the GWR model are summarized in Table 6.14. To examine whether the GWR model exhibits better model performance than the OLS model, the values of adjusted R^2 and AICc from the OLS

and GWR models were compared. The adjusted R^2 increased from 0.47 (OLS model) to 0.78 (GWR model), and the AICc index decreased from 3425.22 (OLS model) to 2778.51 (GWR model). These findings suggest that the GWR model offers better performance than the OLS model for estimating visitations.

For the OLS model, the overall model was statistically significant (Joint F-Statistic: 180.66, p-value < 0.01), and all the independent variables were statistically significant at 99% and 95% levels. Specifically, the coefficient for the OD average distance (1.13, p-value < 0.01) indicated that zip codes with a higher average distance from origins to the destination had a higher number of visitations. For the GWR model, the local R^2 ranged from a minimum of 0.48 to a maximum of 0.97 with a mean of 0.79. The spatial autocorrelation of residuals (Moran's I: -0.001, p-value: 0.849) indicates spatial randomness and the appropriateness of running a GWR model.

Maps in Figure 6.18 R^2 in the GWR model. The local coefficients of the independent variables ranged from -0.15 to 0.44 with a mean of 0.09 (reviews of attractions), -3.21 to 23.17 with a mean of 6.79 (OD average distance), -0.48 to 1.00 with a mean of 0.19 (Airbnb Rooms), and -0.14 to 0.146 with a mean of 0.10 (Hotel rooms). Based on the average local coefficients, all variables were positively associated with visitation. This result means that the number of visitations is positively affected by the independent variables included in the model (i.e., reviews of attractions, OD average distance, Airbnb rooms, and hotel rooms). The variability in the local parameter estimates indicates spatial variability, which represents spatially varying relationships between visitation and the independent variables throughout Florida. Furthermore, the GWR model exhibited various values of the local R^2 which indicated that the exploratory power of the GWR model was not stationary throughout Florida.

As shown in Table 6.14, the numbers of Airbnb rooms and the average OD distance were significant predictors of visitations. Specifically, the local coefficients of the OD average distance ranged from -3.21 to 23.17 with a mean of 6.79, indicating that the number of visitations tends to increase as the average distance from the origin to the destination increases. Specifically, zip codes with strong positive local coefficients were observed mainly in the central and southeastern regions of Florida. This result is different from traditional trade models in which trade is proportional to the economic sizes and inversely proportional to the geographic distance between two units. The finding of this project can be explained by the fact that travelers do not mind to travel long distances to visit prominent tourist destinations (LaMondia at al., 2010). This means that tourists visiting Florida may travel long distances to visit many prominent tourist destinations in Florida. Some zip codes showing negative local coefficients on the maps show that tourism supply (the number of Airbnb rooms, hotel rooms, and tourist attractions) may mismatch tourism demand (tourist visits). For example, the number of Airbnb rooms is relatively small compared to the number of visitations in northwest zip codes in Florida.

Overall, the results of the GWR model demonstrated that the number of visitations can be explained well with the number of reviews of attractions, the average distance from origins to destinations, the number of Airbnb rooms, and the numbers of hotel rooms, indicating that these variables are suitable for estimating the number of visitations across zip codes in Florida.

Table 6.14 Results of the GWR model

Variables	OLS	GWR Coefficient		
	β	Min.	Mean	Max.
Intercept	3.13	-371.41	-49.70	93.53
Reviews of attractions	0.12**	-0.15	0.09	0.44
OD average distance	1.13**	-3.21	6.79	23.17
Airbnb rooms	0.22**	-0.48	0.19	1.00
Hotel rooms	0.11**	-0.14	0.10	0.146
Latitude	-0.06*	-4.73	0.86	11.27
Local R^2	0.48	0.48	0.79	0.97
Adjusted R^2	0.47		0.78	
AIC_c	3425.22		2778.51	

Note: $**p < .01$; $*p < .05$

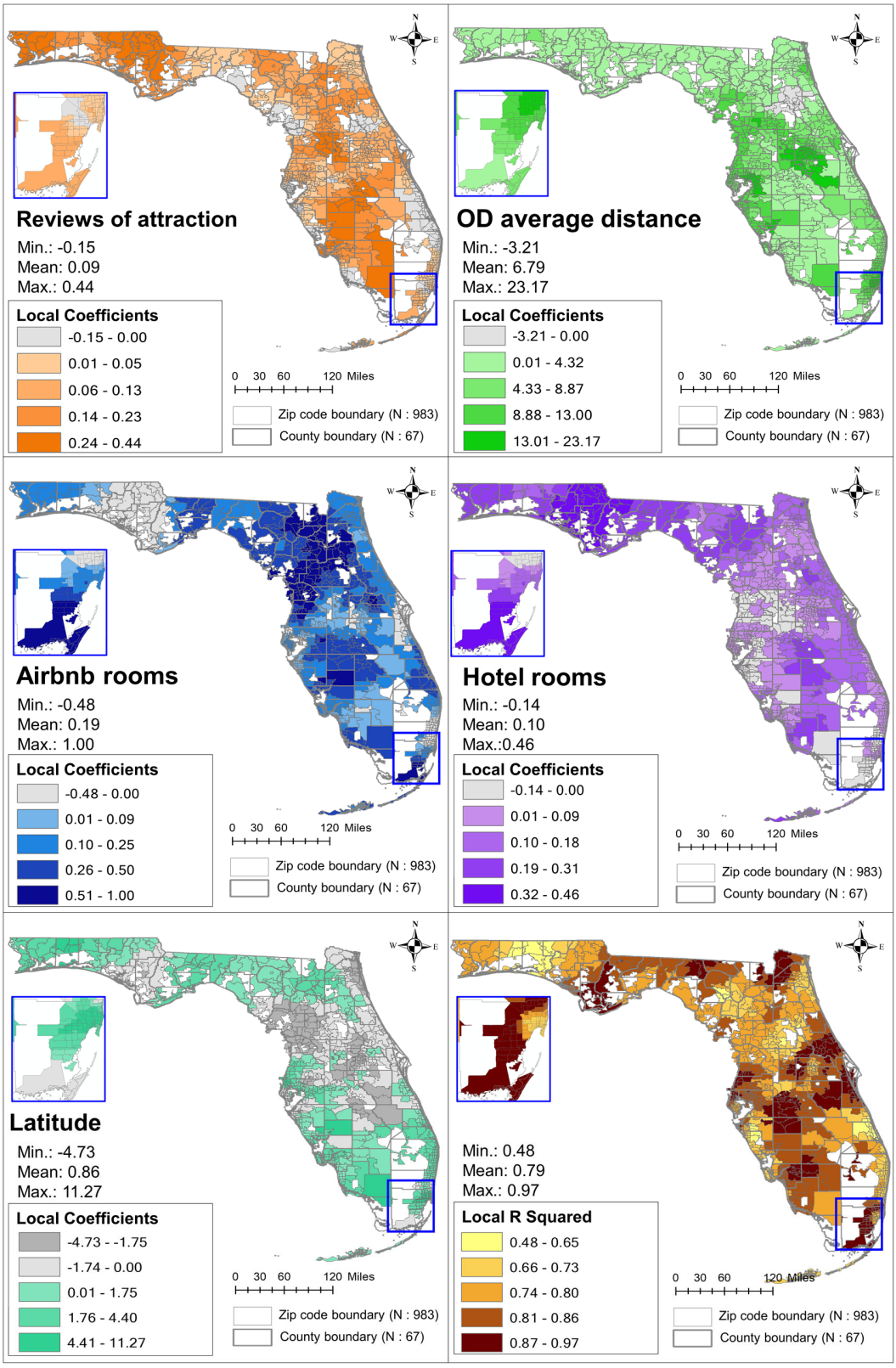


Figure 6.18 Spatial distribution of local coefficients for the independent variables and local R^2

6.7 Conclusions and recommendations

Task 6 describes a procedure of testing tourist flow methodologies to assess the impact of tourist travel on the transportation system and identify key determinants of tourist flows. In this task, we 1) applied a destination choice model to assess trip distributions through comprehensive variables including tourism resources in destinations and interaction variables between origins and destinations 2) measured the relationship between tourist flows and traffic flows, and 3) developed a tourist flows model to derive key predictors that significantly affect tourist flows. The outcomes proved that tourist flows significantly affect traffic flows, and the impact of tourist flows spatially varies depending upon census tracts. Furthermore, urban tourism, beach tourism, golf tourism, and RV/Camping-related tourism resources were significant predictors for estimating tourist flows. As tourism has a significant impact on the transportation system in Florida, considering key factors that affect tourist travel is important in the effective management and planning of the transportation system. In Task 6, a tourist flows model for predicting tourism visitation that has a significant impact on Florida's transportation system was constructed based on the all outcomes of the previous tasks of this project. The outcomes of Task 6 can be applied to forecast tourism visitation and assess the impact of tourism on the transportation system, both of which have a significant effect on building sustainable and efficient transportation policies and plans.

Overall, the project led us to formulate the following conclusions:

- Innovative data sources, specifically cell phone data and social media are able to implement tourist flows in transportation travel demand modelling;
- Data collection process is established and tested;
- There is a good correlation between tourism flow estimations coming from the social media, cell phone data, and Visit Florida surveys;
- A set of explanatory models for tourism flows was developed and analyzed; model performance was inadequate at a census tract level exhibiting measurement errors. At least a zip code spatial resolution is recommended;
- The new data come with limitations which are important to understand; those limitations also justify using multiple data sources to compensate for those limitations;
- Significant variables for prediction of tourist flows are identified which include distance, income level in origin, and tourism attractions and products, especially theme parks.

In addition, we provide the following recommendations:

- We presented several visitation model working at different scales. Final decision on model selection should be made by the FDOT team;
- Practical work on including the tourist flows into FDOT needs to be start with building a pilot model;
- Periodical update of the travel flow database is needed with data coming from the social media updated on an annual base and cell phone data – on a 5-year base;
- An investigation of the effect of extreme events such as hurricanes and COVID-19 pandemic on tourism travel is valuable.

During the research, we identified several areas that needed additional research. Particularly, those areas include:

- Data on South America travelers. South America comprise 25% of the total number of foreign visitors. While data collection on those travelers was not planned in this project, we collected the social media data and found discrepancies between the social media data and Visit Florida on Brazilian and Argentinian travelers;
- Appropriate time resolution for the model needs a separate analysis;
- Data storage and update protocol need to be established;
- Extreme event effect on tourist evacuation traffic needs a separate research; social media and cell phone data are very useful in this respect.

References

Ahas, R., Aasa, A., Mark, Ü., Pae, T., & Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management*, *28*(3), 898–910. https://doi.org/10.1016/j.tourman.2006.05.010

Andreoni, A., & Postorino, M. N. (2006). A multivariate ARIMA model to forecast air transport demand. *Proceedings of the Association for European Transport and Contributors, 1-14.*

Anvari, S., Tuna, S., Canci, M., & Turkay, M. (2016). Automated Box-Jenkins forecasting tool with an application for passenger demand in urban rail systems. *Journal of Advanced Transportation*, *50*(1), 25–49. https://doi.org/10.1002/atr.1332

Bastianoni, S., Pulselli, R. M., Romano, P., & Pulselli, F. M. (2008). Dynamics and evolution of urban patterns: The evidence of the Mobile Landscape project. *WIT Transactions on Ecology and the Environment*, *114*, 253–260. https://doi.org/10.2495/DN080261

Belsley, D.A. (1991). Conditioning diagnostics: collinearity and weak data in regression. Wiley, New York

Bernardin Jr, V. L., Koppelman, F., & Boyce, D. (2009). Enhanced destination choice models incorporating agglomeration related to trip chaining while controlling for spatial competition. Transportation research record, 2132(1), 143-151.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM.*

Bostrom, R. (2006). Kentucky Statewide Travel Model (KYSTM). *Combined 25 Kentucky-Tennessee Model Users Group Meeting*. Bowling Green KY.

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis Forecasting and Control.* (Holden-Day, ed.). https://doi.org/10.2307/3008255

Cai, G., Lee, K., & Lee, I. (2016). A framework for mining semantic-level tourist movement behaviours from geo-tagged photos. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9992 LNAI, 519–524. https://doi.org/10.1007/978-3-319-50127-7_44

California High-Speed Rail Authority. (2016). California High-Speed Rail Ridership and Revenue Model.

Cankurt, S., & Subaşi, A. (2016). Tourism demand modelling and forecasting using data mining techniques in multivariate time series: A case study in Turkey. *Turkish Journal of Electrical Engineering and Computer Sciences*, *24*(5), 3388–3404. https://doi.org/10.3906/elk-1311-134

Çelebi, D., Bolat, B., & Bayraktar, D. (2009). Light rail passenger demand forecasting by artificial neural networks. *2009 International Conference on Computers and Industrial Engineering, CIE 2009*, 239–243. https://doi.org/10.1109/iccie.2009.5223851

Chareyron, G., Da-Rugna, J., & Branchet, B. (2013). Mining tourist routes using Flickr traces. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013*, 1488–1489. https://doi.org/10.1145/2492517.2500307

Charlton, M., & Fotheringham, A. S. (n.d.). *Geographically Weighted Regression: A Tutorial on using GWR in ArcGIS 9.3*. https://doi.org/10.1016/B978-008044910-4.00447-8

Chavas, J. P., Stolu, J., & Sellar, C. (1989). On the commodity value of travel time in recreational activities. *Applied Economics*, *21*(6), 711–722. https://doi.org/10.1080/758520269

Chen, C. F., Lai, M. C., & Yeh, C. C. (2012a). Forecasting tourism demand based on empirical mode decomposition and neural network. *Knowledge-Based Systems*, *26*, 281–287. https://doi.org/10.1016/j.knosys.2011.09.002

Chen, K. Y., & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, *28*(1), 215–226. https://doi.org/10.1016/j.tourman.2005.12.018

Chen, M. S., Ying, L. C., & Pan, M. C. (2010). Forecasting tourist arrivals by using the adaptive network-based fuzzy inference system. *Expert Systems with Applications*, *37*(2), 1185–1191. https://doi.org/10.1016/j.eswa.2009.06.032

Chen, S. C., Kuo, S. Y., Chang, K. W., & Wang, Y. T. (2012b). Improving the forecasting accuracy of air passenger and air cargo demand: The application of back-propagation neural networks. *Transportation Planning and Technology*, *35*(3), 373–392. https://doi.org/10.1080/03081060.2012.673272

Chinnakum, W., & Boonyasana, P. (2017). Modelling Thailand tourism demand: A dual generalized maximum entropy estimator for panel data regression models. *Thai Journal of Mathematics*, *15* (Special issue on entropy in econometrics), 67–78. http://thaijmath.in.cmu.ac.th

Chu, F. L. (2008). A fractionally integrated autoregressive moving average approach to forecasting tourism demand. *Tourism Management*, *29*(1), 79–88. https://doi.org/10.1016/j.tourman.2007.04.003

Chu, F. L. (2014). Using a logistic growth regression model to forecast the demand for tourism in Las Vegas. *Tourism Management Perspectives*, *12*, 62–67. https://doi.org/10.1016/j.tmp.2014.08.003

Chua, A., Servillo, L., Marcheggiani, E., & Moere, A. Vande. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management*, *57*, 295–310. https://doi.org/10.1016/j.tourman.2016.06.013

Chung, H. C., Chung, N., & Nam, Y. (2017). A social network analysis of tourist movement patterns in blogs: Korean backpackers in Europe. *Sustainability (Switzerland)*, *9*(12). https://doi.org/10.3390/su9122251

Claveria, O., & Torra, S. (2014). Forecasting tourism demand to Catalonia: Neural networks vs. time series models. *Economic Modelling*, *36*, 220–228. https://doi.org/10.1016/j.econmod.2013.09.024

Costinett, P., & Stryker, A. (2007). Calibrating the Ohio Statewide Travel Model. *11th TRB National Transportation Planning Applications Conference, Daytona Beach, FL, May 6-10.* https://www.trbappcon.org/2007conf/files/085 Costinett 1 final.pdf

Çuhadar, M., Cogurcu, I., & Kukrer, C. (2014). Modelling and Forecasting Cruise Tourism Demand to İzmir by Different Artificial Neural Network Architectures. *International Journal of Business and Social Research*, *4*(3), 12–28. https://doi.org/10.18533/ijbsr.v4i3.431

Date, S. (2019). The Akaike Information Criterion. https://towardsdatascience.com/the-akaike-information-criterion-c20c8fd832f2

Davis, A. W., McBride, E. C., Janowicz, K., Zhu, R., & Goulias, K. G. (2018). Tour-Based Path Analysis of Long-Distance Non-Commute Travel Behavior in California. *Transportation Research Record*, *2672*(49), 1–11. https://doi.org/10.1177/0361198118778926

De Jong, G., Daly, A., Pieters, M., & Van der Hoorn, T. (2007). The logsum as an evaluation measure: Review of the literature and new results. Transportation Research Part A: Policy and Practice, 41(9), 874-889.

De Oliveira Santos, G. E., Ramos, V., & Rey-Maquieira, J. (2012). Determinants of multi-destination tourism trips in Brazil. *Tourism Economics*, *18*(6), 1331–1349. https://doi.org/10.5367/te.2012.0170

Dogru, T., Mody, M., Line, N., Suess, C., Hanks, L., & Bonn, M. (2020). Investigating the whole picture: Comparing the effects of Airbnb supply and hotel supply on hotel performance across the United States. Tourism Management, 79, 104094. https://doi.org/https://doi.org/10.1016/j.tourman.2020.104094

Engle, R. F., & Granger, C. W. J. (1987). Co-Integration and Error Correction : Representation , Estimation , and Testing. *Econometrica*, *55*(2), 251–276.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry, 26(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Fesenmaier, D. R., & Lieber, S. R. (1988). Destination diversification as an indicator of activity compatibility: An exploratory analysis. *Leisure Sciences*, *10*(3), 167–178. https://doi.org/10.1080/01490408809512187

Formica, S., & Uysal, M. (2006). Destination attractiveness based on supply and demand evaluations: An analytical framework. *Journal of Travel Research*, *44*(4), 418-430.

Fotheringham, A. S., Charlton, M. E., & Brunsdon, C. (1998). Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis. *Environment and Planning A: Economy and Space*, *30*(11), 1905–1927. https://doi.org/10.1068/a301905

Fotheringham, A., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons (Vol. 13).

Ghalehkhondabi, I., Ardjmand, E., Young, W. A., & Weckman, G. R. (2019). A review of demand forecasting models and methodological developments within tourism and passenger transportation industry. *Journal of Tourism Futures*, *5*(1), 75–93. https://doi.org/10.1108/JTF-10-2018-0061

Goh, C., & Law, R. (2011). The methodological progress of tourism demand forecasting: A review of related literature. *Journal of Travel and Tourism Marketing*, *28*(3), 296–317. https://doi.org/10.1080/10548408.2011.562856

Grosche, T., Rothlauf, F., & Heinzl, A. (2007). Gravity models for airline passenger volume estimation. *Journal of Air Transport Management*, *13*(4), 175–183. https://doi.org/10.1016/j.jairtraman.2007.02.001

Gunn, C. A. (1972). *Vacationscape: Designing tourist regions*. Austin: Bureau of Business Research, University of Texas at Austin.

Gunter, U., & Önder, I. (2016). Forecasting city arrivals with Google Analytics. *Annals of Tourism Research*, *61*, 199–212. https://doi.org/10.1016/j.annals.2016.10.007

Hofer, K., Haberl, M., & Fellendorf, M. (2016, October). Travel Demand Modelling of Touristic Trips in the Province of Salzburg. In European Transport Conference 2016. http://abstracts.aetransport.org/paper/index/id/4937/confid/21%0Ahttps://trid.trb.org/view/1440484

Hong, W. C., Dong, Y., Chen, L. Y., & Wei, S. Y. (2011). SVR with hybrid chaotic genetic algorithms for tourism demand forecasting. *Applied Soft Computing Journal*, *11*(2), 1881–1890. https://doi.org/10.1016/j.asoc.2010.06.003

Hwang, Y. H., Gretzel, U., & Fesenmaier, D. R. (2006). Multicity trip patterns. Tourists to the United States. *Annals of Tourism Research*, *33*(4), 1057–1078. https://doi.org/10.1016/j.annals.2006.04.004

Ibrahim, Y., Nanthakumar, & Loganathan. (2010). Forecasting International Tourism Demand in Malaysia Using Box Jenkins Sarima Application. *South Asian Journal of Tourism and Heritage*, *3*(2), 50–60.

Jang, S., & Kim, J. (2018). Remedying food policy invisibility with spatial intersectionality: A case study in the Detroit Metropolitan Area. *Journal of Public Policy & Marketing, 37*(1), 167-187.

Jang, S., Kim, J., & Zedtwitz, M. (2017). The importance of spatial agglomeration in product innovation: A microgeography perspective. *Journal of Business Research, 78*, 143-154.

Johnson, R. J., Gregory, D., Pratt, G., & Watts, M. (2000). *The dictionary of human geography*. Oxford, England: Blackwell.

Kah, J. A., Lee, C.-K., & Lee, S.-H. (2016). Spatial–temporal distances in travel intention–behavior. Annals of Tourism Research, 57, 160–175. https://doi.org/https://doi.org/10.1016/j.annals.2015.12.017

Kang, S., Kim, J., & Nicholls, S. (2014). National tourism policy and spatial patterns of domestic tourism in South Korea. Journal of Travel Research, 53(6), 791-804.

Kidd, A. M., D'Antonio, A., Monz, C., Heaslip, K., Taff, D., & Newman, P. (2018). A GPS-Based Classification of Visitors' Vehicular Behavior in a Protected Area Setting. *Journal of Park and Recreation Administration*, *36*(1), 69–89. https://doi.org/10.18666/jpra-2018-v36-i1-8287

Kim, J., & Nicholls, S. (2016). Influence of the measurement of distance on assessment of recreation access. *Leisure Sciences, 38*(2), 118-139.

Kim, J., Thapa, B., Jang, S., & Yang, E. (2018). Seasonal spatial activity patterns of visitors with a mobile exercise application at Seoraksan National Park, South Korea. *Sustainability, 10*(7), 1-21.

LaMondia, J., Snell, T., & Bhat, C. R. (2010). Traveler Behavior and Values Analysis in the Context of Vacation Destination and Travel Mode Choices: European Union Case Study. Transportation Research Record, 2156(1), 140–149. https://doi.org/10.3141/2156-16

Lau, G., & McKercher, B. (2006). Understanding Tourist Movement Patterns in a Destination: A GIS Approach. *Tourism and Hospitality Research*, *7*(1), 39–49. https://doi.org/10.1057/palgrave.thr.6050027

Lee, Y., Pennington-Gray, L., & Kim, J. (2019). Does location matters? Exploring the spatial patterns of food safety in a tourism destination. *Tourism Management*, 71, 18-33.

Lew, A., & McKercher, B. (2006). Modeling tourist movements: A local destination analysis. *Annals of Tourism Research*, *33*(2), 403–423. https://doi.org/10.1016/j.annals.2005.12.002

Li, Y., Hu, C., Huang, C., & Duan, L. (2017). The concept of smart tourism in the context of tourism information services. *Tourism Management*, *58*, 293–300. https://doi.org/https://doi.org/10.1016/j.tourman.2016.03.014

Lin, C. J., Chen, H. F., & Lee, T. S. (2011). Forecasting Tourism Demand Using Time Series, Artificial Neural Networks and Multivariate Adaptive Regression Splines:Evidence from Taiwan. *International Journal of Business Administration*, *2*(2). https://doi.org/10.5430/ijba.v2n2p14

Lin, C.-J., & Lee, T.-S. (2013). Tourism Demand Forecasting: Econometric Model based on Multivariate Adaptive Regression Splines, Artificial Neural Network and Support Vector Regression. *Advances in Management and Applied Economics*, *3*(6), 1–18.

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

Lovingood, R. E., & Mitchell, L. E. (1989). A regional analysis of South Carolina Tourism. *Annals of Tourism Research*, *16*(3), 301-317.

Lue, C. C., Crompton, J. L., & Fesenmaier, D. R. (1993). Conceptualization of multi-destination pleasure trips. *Annals of Tourism Research*, *20*(2), 289–301. https://doi.org/10.1016/0160-7383(93)90056-9

Lue, C. C., Crompton, J. L., & Stewart, W. P. (1996). Evidence of cumulative attraction in multidestination recreational trip decisions. *Journal of Travel Research*, *35*(1), 41–49. https://doi.org/10.1177/004728759603500107

Lundgren, J. O. J. (1984). Geographic concepts and the development of tourism research in Canada. *GeoJournal*, *9*(1), 17–25. https://doi.org/10.1007/BF00518314

Luo, D., Xu, H., Zhen, Y., Dilkina, B., Zha, H., Yang, X., & Zhang, W. (2017). Learning mixtures of markov chains from aggregate data with structural constraints (Extended abstract). *Proceedings - International Conference on Data Engineering*, 35–36. https://doi.org/10.1109/ICDE.2017.24

Marrocu, E., & Paci, R. (2013). Different tourists to different destinations. Evidence from spatial interaction models. *Tourism Management*, *39*, 71–83. https://doi.org/10.1016/j.tourman.2012.10.009

McCullough, R. (1998). Evolution of Statewide Modeling in Florida: the Teamwork Approach. *Statewide Travel Demand Forecasting Conference Proceedings*. Transportation Research Board.

McIntosh, R., Goeldner, C., Ritchie, J.R. B. (1998). Tourism: Principles, Practices, Philosophies. Seventh Edition. New York: John Wiley & Sons

McKean, J. R., Johnson, D. M., & Walsh, R. G. (1995). Valuing time in travel cost demand analysis: an empirical investigation. *Land Economics*, *71*(1), 96–105. https://doi.org/10.2307/3146761

McKercher, B. (1998). The Effect of Market Access on Destination Choice. *Journal of Travel Research*, *37*(1), 39–47. https://doi.org/10.1177/004728759803700105

McKercher, B., & Chow So-Ming, B. (2001). Cultural distance and participation in cultural tourism. *Pacific Tourism Review*, *5*(1–2), 23–32. https://www.ingentaconnect.com/contentone/cog/ptr/2001/00000005/f0020001/art00005

McKercher, B., & Lau, G. (2008). Movement patterns of tourists within a destination. *Tourism Geographies*, *10*(3), 355–374. https://doi.org/10.1080/14616680802236352

McKercher, B., & Lew, A. A. (2004). Tourist Flows and the Spatial Distribution of Tourists. In A. A. Lew, C. M. Hall, & A. M. Williams (Eds.), *A Companion to Tourism* (pp. 36–48). https://books.google.com/books?hl=en&lr=&id=nePEB4e3Y0EC&oi=fnd&pg=PA36&ots=6283t_K4H9&sig=KPW5JhYOKlooYp5NCDOnarey-5g

McKercher, B., & Wong, D. Y. Y. (2004). Understanding tourism behavior: Examining the combined effects of prior visitation history and destination status. *Journal of Travel Research*, *43*(2), 171–179. https://doi.org/10.1177/0047287504268246

Mishra, S., Wang, Y., Zhu, X., Moeckel, R., & Mahapatra, S. (2013). Comparison between gravity and destination choice models for trip distribution in Maryland (No. 13-3196).

Morris, J. M., Dumbie, P. L. & Wigan, M. R. (1979). Accessibility indicators for transportation planning. *Transportation Research A, 13*, 91–109.

Nellett, R., Banninga, G., Johnson, C., Witherspoon, L., & Whiteside, L. (1999). Michigan's Statewide Travel Demand Model. *Statewide Travel Demand Forecasting Conference Proceedings*, 100–115. Transportation Research Board.

Nguyen, T.-L., Shu, M.-H., Huang, Y.-F., & Hsu, B.-M. (2013). Accurate forecasting models in predicting the inbound tourism demand in Vietnam. *Journal of Statistics and Management Systems*, *16*(1), 25–43. https://doi.org/10.1080/09720510.2013.777570

Nicholls, S. (2001). Measuring the accessibility and equity of public parks: A case study using GIS. *Managing Leisure, 6*(4), 201–219.

Nishad, A., & Abraham, S. (2017). Semantic trajectory analysis for identifying locations of interest of moving objects. *2017 International Conference on Networks and Advances in Computational Technologies, NetACT 2017*, 257–261. https://doi.org/10.1109/NETACT.2017.8076776

Noersasongko, E., Julfia, F. T., Syukur, A., , P., Pramunendar, R. A., & Supriyanto, C. (2016). A Tourism Arrival Forecasting using Genetic Algorithm based Neural Network. *Indian Journal of Science and Technology*, *9*(4). https://doi.org/10.17485/ijst/2016/v9i4/78722

Oh, C. O., & Morzuch, B. J. (2005). Evaluating time-series models to forecast the demand for tourism in Singapore: Comparing within-sample and postsample results. *Journal of Travel Research*, *43*(4), 404–413. https://doi.org/10.1177/0047287505274653

Outwater, M. L., Bradley, M., Ferdous, N., Junction, W. R., Bhat, C., Pendyala, R., … Frequency, T. (2015). A Tour-Based National Model System To Forecast Long-Distance Passenger Travel in the

United States. *TRB 94th Annual Meeting Compendium of Papers 15-4322*. https://trid.trb.org/view/1338557

Pai, P. F., Hong, W. C., & Lin, C. S. (2005). Forecasting tourism demand using a multifactor support vector machine model. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3801 LNAI, 512–519. https://doi.org/10.1007/11596448_75

Pai, P. F., Hung, K. C., & Lin, K. P. (2014). Tourism demand forecasting using novel hybrid system. *Expert Systems with Applications*, *41*(8), 3691–3702. https://doi.org/10.1016/j.eswa.2013.12.007

Pan, B., Wu, D. C., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, *3*(3), 196–210. https://doi.org/10.1108/17579881211264486

Pan, B., Xiang, Z., Law, R., & Fesenmaier, D. R. (2011). The dynamics of search engine marketing for tourist destinations. *Journal of Travel Research*, *50*(4), 365–377. https://doi.org/10.1177/0047287510369558

Peng, G., Liu, Y., Wang, J., & Gu, J. (2017). Analysis of the prediction capability of web search data based on the HE-TDC method – prediction of the volume of daily tourism visitors. *Journal of Systems Science and Systems Engineering*, *26*(2), 163–182. https://doi.org/10.1007/s11518-016-5311-7

Pizam, A., & Sussmann, S. (1995). Does nationality affect tourist behavior? *Annals of Tourism Research*, *22*(4), 901–917. https://doi.org/10.1016/0160-7383(95)00023-5

Pourabdollahi, Z., Tillery, R., Gawade, M., & Hill, T. (2017). Statewide Tourism Travel Demand Forecasting: A Behavior-Based Modeling Framework for the State of Florida. *Transportation Research Board 96th Annual Meeting, Washington DC, Jan 8-12*. https://trid.trb.org/view/1438404

Proussaloglou, K. E., & Popuri, Y. D. (2004). Enhancing State and MPO Transportation Planning Using National Household Travel Survey Add-On Data : The Wisconsin Experience. https://www.researchgate.net/publication/228703538

Rafidah, A., Shabri, A., Nurulhuda, A., & Suhaila, Y. (2017). A Wavelet Support Vector Machine Combination Model for Singapore Tourist Arrival to Malaysia. *IOP Conference Series: Materials Science and Engineering*, *226*(1), 274–279. https://doi.org/10.1088/1757-899X/226/1/012077

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, *75*, 197–211. https://doi.org/10.1016/j.trc.2016.12.008

Rosselló-Nadal, J., Riera-Font, A., & Cárdenas, V. (2011). The impact of weather variability on British outbound flows. *Climatic Change*, *105*(1), 281–292. https://doi.org/10.1007/s10584-010-9873-y

Samsudin, S., Saad, P., & Shabri, A.. (2010). Hybridizing GMDH and least squares SVM support vector machine for forecasting tourism demand. *International Journal of Research and Reviews in Applied Sciences*, *3*(3), 274–279. https://www.cabdirect.org/cabdirect/abstract/20113107913

Semeida, A. M. (2014). Derivation of travel demand forecasting models for low population areas: the case of Port Said Governorate, North East Egypt. *Journal of Traffic and Transportation Engineering (English Edition)*, *1*(3), 196–208. https://doi.org/10.1016/S2095-7564(15)30103-3

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1. https://doi.org/10.2307/1912017

Sivrikaya, O., & Tunç, E. (2013). Demand forecasting for domestic air transportation in Turkey. *Open Transplantation Journal*, *7*(1), 20–26. https://doi.org/10.2174/1874447820130508001

Smallwood, C. B., Beckley, L. E., & Moore, S. A. (2012). An analysis of visitor movement patterns using travel networks in a large marine park, north-western Australia. *Tourism Management*, *33*(3), 517–528. https://doi.org/10.1016/j.tourman.2011.06.001

Smith, S. (1987). Regional analysis of tourism resources. *Annals of Tourism Research*, *14*(2), 254-273.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting-A review of recent research. *Tourism Management*, *29*(2), 203–220. https://doi.org/10.1016/j.tourman.2007.07.016

Spotts, D. (1997). Regional analysis of tourism resources for marketing purposes. *Journal of Travel Research, 35*(3), 3-15.

Taecharungroj, V., & Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*, *75*, 550–568. https://doi.org/https://doi.org/10.1016/j.tourman.2019.06.020

Talen, E., & Anselin, L. (1998). Assessing spatial equity: An evaluation of measures of accessibility to public playgrounds. *Environmental and planning A, 30*(4), 595–613

Tang, J., Sriboonchitta, S., & Yuan, X. (2015). Forecasting inbound tourism demand to China using time series models and belief functions. *Studies in Computational Intelligence*, *583*, 329–341. https://doi.org/10.1007/978-3-319-13449-9_23

Tideswell, C., & Faulkner, B. (1999). Multidestination travel patterns of international visitors to Queensland. *Journal of Travel Research*, *37*(4), 364–374. https://doi.org/10.1177/004728759903700406

Tideswell, C., & Faulkner, B. (2009). Identifying antecedent factors to the traveler's pursuit of a multidestination travel itinerary. *Tourism Analysis*, *14*(3), 177–190. https://www.ingentaconnect.com/content/cog/ta/2003/00000007/F0020003/art00001

Train, K. E. (2009). Discrete choice methods with simulation. Cambridge university press.

Transportation Research Board. (1998). Tourism Travel and Transportation System development. In *NCHRP Report 419*. https://doi.org/10.1002/9781118883341.ch18

Transportation Research Board. (2004). Integrating Tourism and Recreation Travel with Transportation Planning and Project Delivery: A Synthesis of Highway Practice.

Transportation Research Board. (2005). Statewide Travel Demand Modeling: A Peer Exchange. In *Transportation Research E-Circular*. https://trid.trb.org/view/760295

Transportation Research Board. (2007). Determination of the State of the Practice in Metropolitan Area Travel Forecasting: Findings of the Surveys of Metropolitan Planning Organizations. In *Transportation Research Board*. https://trid.trb.org/view/1599347

Transportation Research Board. (2012). Long-Distance and Rural Travel Transferable Parameters for Statewide Travel Forecasting Models. In *Long-Distance and Rural Travel Transferable Parameters for Statewide Travel Forecasting Models*. https://doi.org/10.17226/22661

TripAdvisor. (2019). Media Center. https://tripadvisor.mediaroom.com

Tussyadiah, I. P., & Fesenmaier, D. R. (2007). Interpreting tourist experiences from first-person stories: A foundation for mobile guides. *Proceedings of the 15th European Conference on Information Systems, ECIS 2007*, 2259–2270. https://www.researchgate.net/publication/221408840

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. New York, NY.: Springer Science & Business Media.

Varagouli, E. G., Simos, T. E., & Xeidakis, G. S. (2005). Fitting a multiple regression line to travel demand forecasting: The case of the prefecture of Xanthi, Northern Greece. *Mathematical and Computer Modelling*, *42*(7–8), 817–836. https://doi.org/10.1016/j.mcm.2005.09.010

Visit Florida. (2019). The Power of Tourism.

Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, *29*(1), 507–529. https://doi.org/10.1108/IJCHM-05-2015-0249

Wu, Q., Law, R., & Xu, X. (2012). A sparse Gaussian process regression model for tourism demand forecasting in Hong Kong. *Expert Systems with Applications*, *39*(5), 4769–4774. https://doi.org/10.1016/j.eswa.2011.09.159

Xiong, C., & Zhang, L. (2013). Deciding whether and how to improve statewide travel demand models based on transportation planning application needs. *Transportation Planning and Technology*, *36*(3), 244–266. https://doi.org/10.1080/03081060.2013.779473

Xu, X., Law, R., Chen, W., & Tang, L. (2016). Forecasting tourism demand by extracting fuzzy Takagi–Sugeno rules from trained SVMs. *CAAI Transactions on Intelligence Technology*, *1*(1), 30–42. https://doi.org/10.1016/j.trit.2016.03.004

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, *46*, 386–397. https://doi.org/10.1016/j.tourman.2014.07.019

Yang, Y., Pan, B., & Song, H. (2014). Predicting Hotel Demand Using Destination Marketing Organization's Web Traffic Data. *Journal of Travel Research*, *53*(4), 433–447. https://doi.org/10.1177/0047287513500391

Yue, E., & Ksaibati, K. (2018). Using Tourism-Based Travel Demand Model to Estimate Traffic Volumes on Low-Volume Roads. *International Journal of Traffic and Transportation Engineering*, *7*(4), 71–77. https://doi.org/10.5923/j.ijtte.20180704.01

Yun, H. J., Kang, D. J., & Lee, M. J. (MJ). (2018). Spatiotemporal distribution of urban walking tourists by season using GPS-based smartphone application. *Asia Pacific Journal of Tourism Research*, *23*(11), 1047–1061. https://doi.org/10.1080/10941665.2018.1513949

Zhang, L., Xiong, C., & Hetrakul, P. (2011). State Highway Administration RESEARCH Report Feasibility And Benefits Of Advanced Four-Step And Activity-Based Travel Demand Models For Maryland. https://trid.trb.org/view/1313965

Appendix 1. Preliminary methodology for trip distribution of tourist flows

Introduction

Gravity models have been the most common form of trip distribution model for decades and are arguably still the most used form in practice. However, destination choice models (DCM) have gained an increasing replacement for gravity models to improve the accuracy of the trip distribution estimation, given its advantages in incorporation of additional variables, as well as reflecting more complex statistical assumptions (e.g., capturing spatial autocorrelation) (Bernardin et al., 2009). Destination choice models are even more advantageous over gravity models for longer distance personal travel and multinucleated travel regions, and have therefore been widely incorporated in statewide travel models (e.g., Arizona, California, Idaho, Iowa, Maryland, New Hampshire, Ohio, Oregon, Tennessee, Wisconsin, etc.) and many metropolitan area models alike (e.g. South Bend, Evansville, and Columbus, Indiana; Ann Arbor, Michigan; Burlington, Vermont; Knoxville and Chattanooga, Tennessee; Charlottesville, Virginia; Charleston, South Carolina; and Jacksonville, Florida).

Destination choice models are a type of trip distribution or spatial interaction model which are formulated as discrete choice models, typically logit models. The formulations of destination choice models are flexible and extensible to include a wider range of explanatory variables and thus provide a better behavioral basis for trip distribution than the traditional gravity models. Typically, a destination choice model incorporates additional variables beyond size/attractions, impedance/friction factors and constants/k-factors.

Why to choose destination choice models over gravity models for tourist flows?

- Gravity models have intrinsic limitations for tourism travels and tourist flows:

a. gravity model may still suitable for trip distribution in mono-centric urban regions where accessibility to transit plays little to no role in choice of destination, while tourist flows take place in multiple-centric situation as there are more than one dominant attraction destination for tourist to choose

b. gravity model may respond illogically to changes in service where improved accessibility to a given destination may cause a disproportionate increase in total trips

- Destination choice models have advantages in overcoming the mentioned limitations:

a. With appropriate specifications of utility, consistency between changes in levels of service and changes in demand can be assured in destination choice models.

b. functional form of the destination choice utility is very flexible, thus a term can be added to the utility equation, statistically estimated from observed data, and interpreted in terms of equivalent minutes of travel time; a much more data-based and intuitive measure of the impact the any possible factors on a person's travel choice.

- Therefore, it is convincing that destination choice model is a much more appropriate approach to be applied in tourism travel demand model, given that there are numerous factors that may influence tourist's destination-choosing decisions. In addition, destination choice models have been standard and ubiquitous in tour-based and activity-based models in current practice. The tourism module for FLSWM is largely based on tourist behavior during their travel, and that make destination choice model a well fit for this study.

Theoretical and Mathematical Foundations of DCM

Destination choice models are derived from theoretical foundations in entropy maximization and random utility theory. Some of its basic assumptions, their functional forms and parameter estimation requirements are explained below.

A general spatial interaction model (trip distribution model) attempts to address the problem how trips between locations in space (typically traffic zones) are to be predicted, given limited information concerning these interactions (illustrated in Figure A-1).

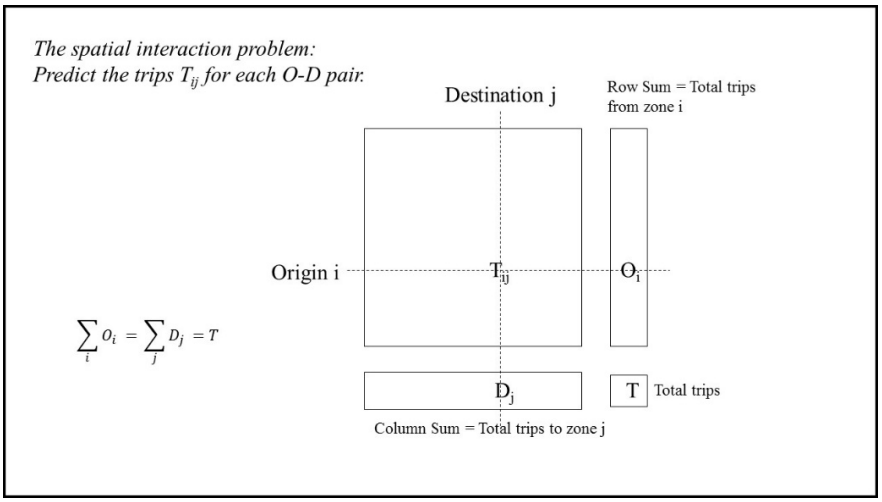


Figure A-1 Spatial distribution of local coefficients for Factor 4 (Beach Tourism) in Florida

The most common destination choice model nowadays is some form of random utility model, usually a multinomial logit (MNL) model. A typical logit destination choice model for the probability that destination *j* is chosen from origin *i* ($P_{j|i}$) is expressed as:

$$P_{j|i} = \frac{e^{U_{j|i}}}{\sum_{j'} e^{U_{j'|i}}} \qquad \text{(A-1)}$$

where $U_{j|i}$ is the systematic utility of destination *j* given origin *i*, which can be written as follows:

$$U_{j|i} = \beta \times TravelImpedance_{ij} + \ln(Size_j) \qquad \text{(A-2)}$$

In formulation (A-2), the utility of a destination depends on (a) the impedance or spatial separation between the trip origin and the destination, and (b) the size or attractions at the destination. This is the simplest representation of destination choice utility. The impedance term is often referred to as the *qualitative* utility component, while the size or attraction term is referred to as the *quantitative* component.

(a) Impedance can be measured by distance, auto travel time, or a generalized cost, among other possible measures of spatial separation. A convenient measure of impedance is the inclusive value, or logsum, of the mode choice model (de Jong et al., 2007). The coefficient β of the impedance variable(s) can be generic (i.e., the same for all decision-makers), or it can vary for certain types of travelers.

For destination choice models, apart from impedance, *qualitative* utility component can also include accessibility, psychological boundaries, and other destination qualities, as well as traveler attributes. For example, γ is an indicator variable that take value if the trip-maker exhibits a certain characteristic (e.g. she is a part-time worker), and takes value 0 otherwise, and the *qualitative* utility would be presented:

$$U_{j|i} = \beta_1 \times D_{ij} + \beta_2 \times D_{ij} \times \gamma \qquad \text{(A-3)}$$

(b) The attraction variable is commonly referred to as the size term. It measures the activity opportunities at each destination. In the case of a work location model, the size term is typically employment. In the case of a tourism attraction location model, the size term can be option of tourism facilities. For many other trip purposes, the size term is typically a linear combination of different types of employment, for example:

$$Size_j = \alpha_1 \times RetailEmp + \alpha_2 \times ServiceEmp + \alpha_3 \times ProductionEmp \qquad \text{(A-4)}$$

The size term always enters the utility function in log form. The log formulation is necessary so that the choice probability of a destination is directly proportional to the number of opportunities at the destination.

$$Pr_{j|i} = \frac{exp(U_{j|i})}{\sum_k exp(U_{k|i})} = \frac{exp(\beta \times Imp)}{\sum_k exp(U_{k|i})} \times S_j \qquad \text{(A-5)}$$

A corollary of the size term log specification is that the choice probabilities are invariant with respect to the scale of the size term. That is, the choice probabilities remain the same when the entire size term is multiplied by an arbitrary factor *f*:

$$Pr_j = \frac{exp(U_j + \ln(S_j))}{\sum exp(U_k + \ln(S_k))} = \frac{exp(U_j) \times S_j}{\sum exp(U_k) \times S_k} = \frac{exp(U_j) \times fS_j}{\sum exp(U_k) \times fS_k} \qquad \text{(A-6)}$$

Data for DCM

The flexibility of destination choice models comes at a cost. While it is possible to represent the selection of trip destinations more rigorously, destination choice models tend to require more data and data with higher fidelity than traditional gravity models.

There are two types of data that are relevant for destination choice models. The observed choice data describe origin-destination flows that have been observed in a survey, by counting or by passive data collection. The explanatory data, on the other hand, refer to input data that describe either destinations or characteristics of the decision maker who chooses the destination.

How big data can change the DCM?
The growing availability of big origin-destination (OD) data and other large-scale sources of OD data have provided the impetus for incorporating these data directly into trip distribution models, and in some cases entirely replace the model. There are generally two methods for using travel demand models together with observed OD data:

- The first approach uses travel demand models (usually of more traditional, aggregate designs) to pivot off of OD matrices developed from a wide array of data sources, including mobile phone data, automated passive count (APC) data, and traffic counts.

  In this study, to build up a DCM of tourist flows, we will collect two sets of big OD data. The first one is a one-year cellphone OD data from AirSage. There is possibility that the tourist data and resident data could be separated, and we could retrieve a clear dataset representing the OD trips of tourists. The other dataset is tourist OD data collected from TripAdvisor reviews of Florida attractions posted by tourists. This collection of social media incorporates practically the whole assembly of OD data of tourists who reported their travel experiences on TripAdvisor.
  Both datasets have potential to entirely replace the current model as they have large enough data points to represent the OD trips of tourists overall. The choice probabilities thus are formulated by observed data as follows:

  $$Pr_{j|i} = \frac{T_{ij}}{O_i}$$

  where:
  T_{ij} = observed trip to destination j from given origin i.
  O_i = Total observed trips from origin i
  In such case, the volume of the trips to a particular destination zone is no longer based on the estimation of utility function but replied upon the observed big data of true tourist travel records.

- The second approach instead uses these OD matrices to develop fixed factors or constants which are incorporated into the travel model; this approach is more attractive for activity-based demand simulation models, but it can also be applied with aggregate trip-based travel models (how social media and big data change DCM in this way is elaborated in the next section).

Adjustment of DCM for tourist flows

Data adjustment

Data	Traditional Destination choice models	New data potential in this study
Observed choice data		
Source	Travel survey or traffic counts	Social media/cellphone
Volume	Small	Large
Group	All population	Tourist only/all population
Spatial resolution	Zone	Zone/block/spot
Temporal resolution	Daily/weekly/monthly	AM/PM/MD/NT
Randomness	Analytic	Analytic/simulation
Behavior		
Trip-chaining	No	Maybe
Inter-personal interactions	No	Maybe
Explanatory data		
Traveler attribute:		
Gender	Yes	Yes
Income	Supported census data	
Education	Yes	No
Travel type	Yes	Yes
Previous visit experience	No	Maybe
Traffic attribute:		
Accessibility	Supported by transportation data	Supported by transportation data
Impedance	Supported by transportation data	Supported by transportation data
Destination attribute:		
Tourism attraction number	No	Yes
Tourism attraction type	No	Yes

Formulation adjustment

- More factors add to *qualitative* utility based on tourist destination theories:

$$U_{j|i} = \beta_1 \times D_{ij} + \beta_2 \times D_{ij} \times \gamma_2 + \beta_3 \times D_{ij} \times \gamma_3 + \cdots + C \qquad \text{(A-7)}$$

  Where C stands for any possible qualitative attribute, such as tourist previous visit experience, different travel purpose, income, education etc.

- The *quantitative* utility merely based on tourism resources and facilities as the representation of tourism attractiveness

$$Size_j = \alpha_1 \times Accommadation + \alpha_2 \times Attraction + \alpha_3 \times Serivce + \cdots + A \quad \text{(A-8)}$$

  Where A stands for any possible type of tourism resources, the type of tourism resources will be based on the analysis result from task 2 and 3.