



UNIVERSITÀ POLITECNICA DELLE MARCHE
Repository ISTITUZIONALE

Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy

This is the peer reviewed version of the following article:

Original

Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy / Chua, Alvin; Servillo, Loris; Marcheggiani, Ernesto; Moere, Andrew Vande. - In: TOURISM MANAGEMENT. - ISSN 0261-5177. - STAMPA. - 57:(2016), pp. 295-310. [10.1016/j.tourman.2016.06.013]

Availability:

This version is available at: 11566/236187 since: 2022-06-03T15:46:39Z

Publisher:

Published

DOI:10.1016/j.tourman.2016.06.013

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. The use of copyrighted works requires the consent of the rights' holder (author or publisher). Works made available under a Creative Commons license or a Publisher's custom-made license can be used according to the terms and conditions contained therein. See editor's website for further information and terms and conditions.

This item was downloaded from IRIS Università Politecnica delle Marche (<https://iris.univpm.it>). When citing, please refer to the published version.

(Article begins on next page)

Manuscript Number: JTMA-D-15-01046R2

Title: Mapping Cilento: Using Geotagged Social Media Data to Characterise Tourist Flows in Southern Italy

Article Type: Case Study

Keywords: Data Mining; Visual Analytics; Flow Analysis; Geotagged Social Media Data

Corresponding Author: Mr. Alvin Chua,

Corresponding Author's Institution: KU Leuven

First Author: Alvin Chua

Order of Authors: Alvin Chua; Loris Servillo; Ernesto Marcheggiani; Andrew Vande Moere

Abstract: New sources of geotagged information derived from social media like Twitter show great promise for geographic research in tourism. This paper describes an approach to analyze geotagged social media data from Twitter to characterize spatial, temporal and demographic features of tourist flows in Cilento - a regional tourist attraction in southern Italy. It demonstrates how the analysis of geotagged social media data yields more detailed spatial, temporal and demographic information of tourist movements in comparison to the current understanding of tourist flow patterns in the region. The insights obtained from our case study illustrate the potential of the proposed methodology yet attention should be paid to biases in the data as well as methodological limitations when drawing conclusions from analytical results.

Response to Reviewers

We would like to thank the reviewers for the valuable feedback on our work. We agree with most of the comments and attempt to address each to the best of our abilities. We hope that the revisions are able to meet the reviewers' expectations. In addition to the key points stated below, we have corrected grammatical and spelling errors. A substantial amount of effort has also gone into verifying that the references are consistently formatted and that missing information like the date, author names and page numbers are resolved. For the following points, we will use p. x, l. y (page x, line y) to indicate where revisions have been made.

1. Is the proposed tourist flow analysis approach a better method than previous research works dealing with the same research topic?

We do not claim that our work is an improvement on, or replacement for, other approaches to characterize tourist flows. However, it does present a substantial advancement in terms of detail when compared to previous attempts to do so with GSMD. The aforementioned points are located between p. 30, l. 581 and p. 31, l. 586 as well as p. 32, l. 629 and p. 32, l. 632.

2. How can better performance of flow analysis be assured when analytical approach goes beyond the case region and is faced with data biases as well as methodological limitations?

We acknowledge the limited discussion in the previous submission. Thus we have elaborated on the subject between p. 32, l. 632 and p. 33, l. 640.

3. It will be better to provide a comprehensive literature review on characterizing tourist flows using GSMD (or specifically geotagged Twitter data) to positioning this paper in the targeted research field, and show the significance of contribution in both the academic and practical domains.

We appreciate the constructive comments on this issue and have further elaborated on the subject between p. 11, l. 242 and p. 12, l. 263. The change includes new 10 references on the use of GSMD in analysis of tourist travel behaviour. Yet only 4 of 10 of these references, to the best of our knowledge, have attempted to characterize certain features of tourist flows.

4. It will be helpful to reorganize section 5 and provide a subsection to describe the public/business implications of the proposed approach in order to reveal its practical importance and guide the way for future adoption.

We agree with the comment and describe possible scenarios for public administrations as well as large and small organizations between p. 33, l. 640 and p. 33, l. 646.

5. How the flow map optimization is achieved and why the shortest path representation is appropriate require explanation.

We have distinguished between the description of the flow map optimization technique in Section 4.3.1 and its implementation in Section 4.3.2 to maintain a certain degree of structure in the paper. Indeed, we have overlooked the rationale for our implementation of the shortest path representation. Thus, this is added to the text at p. 18, l. 390.

*Title page with author details

Mapping Cilento: Using Geotagged Social Media Data to Characterize Tourist Flows in Southern Italy

Alvin Chua¹, Loris Servillo¹, Ernesto Marcheggiani²³ and Andrew Vande Moere³⁴⁵

¹ (alvin.chua, loris.servillo, andrew.vandemoere)@asro.kuleuven.be

² e.marcheggiani@univpm.it

³ Department of Architecture, KU Leuven, Belgium

⁴ Department of Earth and Environmental Sciences, KU Leuven, Belgium

⁵ Department D3A, University of Marche, Italy

***Highlights**

- x We introduce a novel approach to tourist flow analysis based on geotagged social media data.
- x Our approach is structured around three research questions that investigate the spatial, temporal and demographic features of tourist flows.
- x Our analysis yields more detailed spatial, temporal and demographic information of tourist movements in comparison to the current understanding of tourist flow patterns in the region.

1 **Mapping Cilento: Using Geotagged Social Media**
2 **Data to Characterize Tourist Flows in Southern**
3 **Italy**

4 **Abstract**

5 New sources of geotagged information derived from social media like Twitter show
6 great promise for geographic research in tourism. This paper describes an approach to
7 analyze geotagged social media data from Twitter to characterize spatial, temporal
8 and demographic features of tourist flows in Cilento - a regional tourist attraction in
9 southern Italy. It demonstrates how the analysis of geotagged social media data yields
10 more detailed spatial, temporal and demographic information of tourist movements in
11 comparison to the current understanding of tourist flow patterns in the region. The
12 insights obtained from our case study illustrate the potential of the proposed
13 methodology yet attention should be paid to biases in the data as well as 14
14 methodological limitations when drawing conclusions from analytical results.

15

16 **Keywords:** *Data Mining; Visual Analytics; Flow Analysis; Geotagged Social Media*
17 *Data;*

18 **1. Introduction**

19 Flow analysis is a topic of theoretical interest and practical importance in many

20 disciplines. “Flow” commonly refers to the collective movement of people or other
21 abstract concepts like energy, material and information, from a particular location to
22 another. Flow analysis is conventionally conducted to study spatial dynamics and

23 understand how the environment influences the way people move. For instance, interest
24 in modeling traffic flows emerged from the need to identify factors that lead to
25 congestion on transportation infrastructures (Nagatani, 2002). Likewise, insight into
26 routine flow patterns such as journeys between home and work is crucial for the
27 conceptualization of functional urban areas (Sykora & Mulicek, 2014), urban
28 hierarchies (Christaller, 1964) and other territorial structures for policy enactment.

29

30 Tourism plays a major role in many regional economies (Ashley, De Brine, Lehr, &
31 Wilde, 2007) and accounts for a substantial amount of human movement (Schlich &
32 Axhausen, 2003). To meet the planning demands of the tourism industry, it has become
33 increasingly important to monitor and analyze the flows of tourists (Williams, 1998).
34 Access to detailed records of travel routes enables the design of policies that prevent
35 capacity overload on the transportation infrastructure and resolve travel barriers
36 between tourist destinations (Prideaux, 2000). Similarly, attractions can be improved or
37 developed according to the preference of different tourist demographics (Lew &
38 McKercher, 2006). While techniques to study routine travel habits are well established,
39 relatively fewer methods have been developed to support the analysis of tourist flows,
40 which are comparably sporadic. One reason for the lack of progress thus far stems from
41 difficulties with data acquisition as tourism often involves travel between urban and
42 rural spaces (Christaller, 1964; Mansfeld, 1990). From a sensing perspective, this poses
43 peculiar challenges with spatial-temporal precision as well as cost in resources (See
44 Section 3).

45

46 Over the past decade, large geotagged datasets have become increasingly commonplace
47 due to the proliferation of sensor networks and portable devices like smartphones.

48 Termed “Big Data” due to the sheer volume of records that emerge from real-time
49 sensing (Kitchin, 2014), such datasets typically contain information of activities or
50 processes linked to the space and time where they occur. In the domain of “Smart City”
51 research (Kitchin, 2014), much has been accomplished with the use of “Big Data” to
52 study human movement. Smart card data from subway (Roth, Kang,
53 Batty, & Barthélemy, 2011) and bike sharing systems (Beecham, Wood, & Bowerman,
54 2014), taxi journey GPS logs (Ferreira, Poco, Vo, Freire, & Silva, 2013) as well as
55 cellular call data records (Sevtsuk & Ratti, 2009) have provided new opportunities to
56 develop greater understanding of mobility patterns in urban environments (Batty, et al.,
57 2012). In recent years, social media has exercised a powerful influence on the tourism
58 industry as people increasingly rely on virtual communities, personal blogs and
59 networks like Flickr, Twitter and Instagram for travel information (Xiang & Gretzel,
60 2010). Geotagged Twitter data in particular contains larges amounts of up-to-date
61 content for most locations worldwide (Leetaru, Wang, Cao, Padmanabhan, & Shook,
62 2013). From this perspective, the constant availability of highly granular user-generated
63 data serves as a valuable source of information to study the movements of tourist as
64 well as to understand their travel preferences.

65

66 In this paper, we describe the use of geotagged social media data to characterize the
67 spatial, temporal and demographic features of tourist flows. Based on a case study
68 situated in Cilento - a tourist venue in southern Italy, we will demonstrate how our
69 analytical approach, operationalized with geotagged Twitter data, addresses the
70 challenge of tracking large numbers of tourists across a large region. More importantly,
71 we will show how the insights we acquired provide more spatial detail than the current
72 understanding of tourist movements in the case study context, prompting a discussion

73 on the value of our approach in contrast to the methods previously utilized to analyze
74 tourist flows in Cilento.

75

76 The paper is organized in the following manner. First, we outline the case study context
77 and describe the limitations faced by existing data of tourist movements in Cilento.

78 Next we present different methods to gather data of tourist flows, elaborating on the
79 advantages and limitations of each method to determine a suitable alternative.

80 Thereafter, we explain our analytical methodology and technical details related to data
81 processing and visualization. This is followed by a report and discussion of our findings

82 structured around three research questions. Finally, we compare the insights we
83 obtained to existing knowledge of tourist movements in the case study context, and

84 discuss the relative merits of our approach.

85 **2. Case Study**

86 Cilento is a well-known tourist venue located in southern Italy where, for the last two
87 years, policy makers have engaged in a national interest project⁶ funded by European

88 and state agencies⁷ to foster the exchange of best practices in sustainable tourism
89 between developed and under-developed regions in Italy. In this particular context, the

90 objective is to develop a local strategy for tourism that encourages economic
91 development and territorial cohesion. The Cilento region comprises of 31

92 municipalities, spanning approximately 490,000 hectares. The landscape comprises of
93 different environments including a picturesque coastline and mountainous inland dotted

94 with multiple UNESCO heritage sites (e.g. Paestum, Punta Licosa, Capo Palinuro).

⁶ TOOKMC: Transfer Of Organized Knowledge Marche-Cilento

⁷ Italian Ministry of Economic Development and Department for Development and Territorial Cohesion

95 While majority of the settlements are located close to the coast, the inner boundaries of
96 Cilento mainly consist of land parcels for agriculture and nature conservation. Tourism
97 represents a significant fraction of the regional economy. In 2015, an estimated 9% of
98 the regional GDP was derived from tourism and that 11% of the workforce was
99 employed to staff the sector (WTTC, 2015). As of late, the region has suffered from
100 low economic performance, due to changing tourist demographics and spending power.
101 In light of this circumstance, a comprehensive revision to existing tourism policies is
102 required to reverse negative economic trends.

103

104 We conducted a workshop in collaboration with local policy makers and experts from
105 other institutions to familiarize with the region. This workshop spanned the duration of
106 a week involving direct observation at several coastal and inland attractions followed
107 by a seminar where participants described methods for monitoring tourist activity on
108 the ground. The insights obtained from this workshop revealed a disproportionate
109 distribution of tourist activity at the coast as opposed to the interior. Contrary to
110 prevailing policies that promote the region to other European communities, young
111 urbanites from major Italian cities are observed to account for the bulk of tourism.
112 Members of this demographic group prefer to engage in beach activities around coastal
113 resorts instead of visiting natural or heritage attractions situated inland. Other tourists
114 have expressed interest in these attractions but the absence of public transportation
115 discourages prospective travel. The result is a polarized distribution of economic
116 activity in the region, and the potential loss of jobs that rely on the inland economy.

117 **2.1. Research Questions and Data Criteria**

118 While existing information is sufficient for a broad understanding of how changes affect
119 the tourism industry, detailed knowledge of tourist flows is required to inform the
120 design of a suitable policy response. There are three specific research questions
121 (RQ) to be addressed:

122

123 RQ1. What are the meaningful tourist profiles in the region?

124 RQ2. What are the valuable patterns of tourists flows in the region?

125 RQ3. Where are the tourist attractions in the region and how do they differ?

126

127 Each RQ investigated a particular set of features related to tourist flows. RQ1 focused
128 on the demographic composition of tourists in order to determine how the movements
129 of various demographic groups differ spatially and temporally from one another. RQ2
130 investigated spatial and temporal patterns in the data for insight into tourist movements.
131 RQ3 evaluated the relative importance of locations as centers of tourist activity to
132 discover the factors that differentiate well-known attractions from those that are under
133 utilized. Table 1 summarizes the expected outcome from our analysis of tourist flows
134 in the region. The type of information as well as the corresponding level of detail for
135 each feature listed, served as a criteria to determine the quality of data we required.

136

137 Table 1. Information expected from the analysis of tourist flows in Cilento.

Feature	Expected Detail (Granularity / Segmentation)	Expected Type of Information
Demographic	Country	x Distinct demographic groups by location of origin. x Distribution of tourists among the demographic groups.
Temporal	Day	x Evolution of tourist numbers in relation to demographic groups. x Identify seasonal peaks in tourist activity.

Spatial	Neighborhood	x Spatial diffusion of different demographic groups within the region. x Describe how each demographic group moves in relation to available transportation. x Identify tourist hubs based on location popularity. x Spatial characteristics of tourist hubs in relation to transportation. x Effect of distance on movements and destination choices.
---------	--------------	---

138

139 **2.2. Limitations with Existing Data**

140 Till date, data of tourist movement is scarce and disparate due to the awkward
 141 administrative status of the region. Unlike formal administrative bodies that work with
 142 standardized protocols, member municipalities operate independently from one
 143 another. Current practices include a wide spectrum of methods ranging from aggregated
 144 tourism statistics at regional or provincial level to other proxies like records of return
 145 tourism, surveys at local tourism offices, check-ins to guest accommodations or ticket
 146 sales at popular attractions that provide coarse approximations of actual tourist
 147 numbers. A summary of the methods is outlined in Table 2. This information is
 148 extracted from a survey on the current practice of each municipal tabulated in Appendix
 149 1.

150

151 Table 2. Data currently in use for analysis of tourist flows in Cilento.

Method	Sample	Granularity		Demographic Segmentation	Num. of Municipals
		Spatial	Temporal		
Regional Tourism Statistics	From, To, Date, Number of Tourist <i>Tuscany, Campania, July, 8,041</i>	Province	Month	Country	7
Regional Transport Statistics	From, To, Date, Number of People, Purpose of Travel <i>Naples, Sarlino, July, 4,139, Leisure</i>	Province	Month	Country	2
Rail Travel Statistics	From, To, Date, Number of People <i>Sorrento, Padula, July, 43, 379</i>	Municipal	Month	None	1
Provincial Tourism Statistics	From, To, Number of Tourist <i>Capaccio, Agropoli, July, 7,964</i>	Municipal	Month	Country	7
Tourism Satisfaction Survey	From, Till, Destinations, Experience (1-5), Country of Origin, Mode of Transport, Purpose of Travel <i>21/7/14, 4/8/14, [Ascea, Palinuro, Camerota, Castellabate], [4, 4, 4, 4], IT, Rail, Leisure</i>	Municipal	Day	Country	5

Return Tourism Statistics	Address, From, Till, Country of Employment <i>Via Colombo Cristoforo 23, 02-072014, 05-08-2014, DE</i>	Street	Day	Country	2
Check-in to Guest Accommodation	Address, From, Till, Country of Origin <i>B&B Villa Degli Aranci, 16-07-2014, 18-07-2014, NL</i>	Street	Day	Country	9
Ticket Sales at Attraction	Location, Date, Number of Tourist <i>Diocesan Museum, 21-07-2014, 12</i>	Street	Day	None	9
Direct Observation in the Field	Location, Date, Number of Tourist <i>"Via Bolivar Strada exit", 22-082014, 28</i>	Irregular	Day	None	6
Reports, Publications & Periodicals	Location, Date, Number of Tourists <i>Vallo di Diano, 3/8/2013, 412</i>	Irregular	Irregular	Irregular	7
Anecdotes from Local Staff / Stakeholder	Location, Date, Comment <i>Ponte Piaggine, 09-08-2014, "Group of four interested in Nature with plans to visit Monte Motola"</i>	Irregular	Irregular	Irregular	4

152

153 As it stands, the quality of existing data is insufficient for our analysis since
154 measurements made at different spatial-temporal granularities or demographic
155 segmentations cannot be jointly analyzed without loss of precision. Regional or
156 provincial tourism statistics in particular, do not provide enough spatial detail because
157 municipals, rather than individual attractions, are specified as intended destinations.
158 Similarly, general keywords like “leisure” or “business” is often defined as the purpose
159 of travel instead of specific terms like “social visit”, “sight seeing” or
160 “shopping” which may provide clues to where and why a location is chosen. Another
161 shortcoming of these statistics is that they are compiled irregularly, and may not provide
162 timely information of the activities on the ground. Other proxies like checkins to guest
163 accommodations or ticket sales at popular attractions offer fine-grained measures of
164 tourist numbers at specific locations and regular intervals, but do not indicate how
165 tourist travel between the tracked locations or attempt to divide tourists into distinct
166 demographic segments. In some cases, information is extracted from official reports,
167 publications or periodicals that describe tourist activities. Likewise, information might
168 be obtained from anecdotes of local staff or stakeholders. As such information is
169 provided without specific reference to a geographic coordinate, spatial granularity

170 remains relatively coarse since the precision of post-hoc mapping is limited to an
171 approximated area. Though these methods are in place, tourist activity is unmonitored
172 in some municipals due to the lack of resources.

173

174 **3. Alternative Methods to Collect Tourist Flow Data**

175 Trajectories are the atomic unit in flow analysis that captures the movement of a subject
176 over space and time. In general, methods to collect trajectories can be broadly classified
177 as observational or non-observational (Shoval & Isaacson, 2009). Observational
178 methods involve tracking a subject by means of direct surveillance or remote sensing,
179 while non-observational methods rely on self-reported information to recount the
180 subject's sequence of movements.

181 **3.1. Non-Observational Methods**

182 Non-observational methods like recall diaries (Cooper, 1981) or self-administered
183 diaries (Thornton, Williams, & Shaw, 1997) are commonly employed in mobility
184 studies as they do not specifically require the use of sensors and can be tuned to obtain
185 certain information that cannot be explicitly tracked, such as the mode of travel or the
186 purpose of a trip. Nonetheless, non-observational methods face several shortcomings
187 since the quality of the data gathered depends on the subjects' efforts and collaboration.
188 In particular, the spatial-temporal precision of the gathered data tends to be
189 comparatively lower than those obtained with observational methods due to the
190 analogue data collection procedure. Furthermore, other issues like high operating cost
191 limit the regularity in which updates can be made as well as the number of subjects that
192 can be tracked. GPS devices can be introduced to compliment the information reported
193 by subjects and is shown to produce more detailed records of movement in several

194 studies of travel habits (Bakillah, et al., 2014; Mavoa, Oliver, Witten, & Badland, 2011;
195 Neuhaus, 2010), yet the financial cost of procuring GPS devices and the time required
196 to instruct subjects on their use continue to impede large-scale adoption of this
197 technology for research purposes. GPS trajectories can also be obtained from volunteers
198 who share their data publically online (Sieber, 2006) but the number of subjects who
199 participate and the locations that they are active in are relatively limited.

200 **3.2. Observational Methods**

201 Apart from direct surveillance of subjects in the field, most observational methods rely
202 on sensors to passively track movement. The primary benefit of sensors is the ability to
203 continuously collect data throughout the monitoring period. While the type of sensors
204 deployed for observation may vary, “location” and “timestamp” are standard
205 measurements that are often provided as output. Over the past decade, data from an
206 extensive range of sensors has been in use, but cellular call data records (CDR) and
207 geotagged data from social networks like Twitter, Foursquare or Flickr have emerged
208 as the primary source of trajectory information.

209

210 CDR contains information of cell phone activities generated by cellular base stations.
211 To analyze movement in CDR, space-partitioning techniques are first employed to
212 divide a given territory into subspaces determined by the locations of cellular base
213 stations. The position of a cell phone is then approximated to the location of the
214 particular base station responsible for routing its signal at a given moment in time. An
215 estimated trajectory is finally derived by chronologically ordering the locations of the
216 base stations that served a particular cell phone. While analysis of CDR has already led
217 to important discoveries of human mobility habits (Gonzalez, Hidalgo, & Barabasi,
218 2008; Song, Qu, Blumm, & Barabási, 2010), and applied in studies of tourist flows

219 (Girardin, Calabrese, Fiore, Ratti, & Blat, 2008), its proprietary value make obtaining
220 such data challenging. More importantly, the spatial granularity of CDR tends to be
221 coarse in rural environments due to sparse distribution of cellular infrastructure (Shoval
222 & Isaacson, 2009). CDR is also often anonymized for privacy preservation thus the
223 demographic information of the tracked subjects cannot be explicitly extracted.

224

225 Geotagged social media data (GSMD) is considered a valuable proxy of human
226 movement (Hawelka, et al., 2014) and was the focus of several important studies
227 concerned with how factors like social-economic status (Cho, Myers, & Leskovec,
228 2011) as well as friendship (Cheng, Caverlee, Lee, & Sui, 2011) relate to human
229 mobility. The distinct advantage of GSMD is that it provides spatial information at up
230 to street level precision, allowing for detailed trajectories to be extracted (Sun, Fan,
231 Bakillah, & Zipf, 2013; Zheng, Zha, & Chua, 2012) for a wide range of applications
232 such as point of interest classification (G. Andrienko, et al., 2013), community detection
233 (Cranshaw, Schwartz, Hong, & Sadeh, 2012; Wakamiya, Lee, & Sumiya,
234 2013) and identification of anomalous movements (Chae, et al., 2015; Gabrielli,
235 Rinzivillo, Ronzano, & Villatoro, 2014).

236

237 | Research in tourism has been quick to capitalize on this feature by utilizing GSMD for
238 | estimates of inbound tourists numbers (Barchiesi, Moat, Alis, Bishop, & Preis,
239 | 2015) as well as measures of tourist activities at specific urban (Önder, Koerbitz, &
240 | Hubmann-Haidvogel, 2014) and nature based attractions (Orsi & Geneletti, 2013;
241 | Wood, Guerry, Silver, & Lacayo, 2013). GSMD also contains user information like the
242 | subject's location of origin, which is particularly useful for building character profiles
243 | (Fuchs, et al., 2013) that can be used to tailor destinations to individual interests (Jiang,

244 | Yin, Wang, & Yu, 2013). This information is also useful for destination management,
245 | especially to identify groups of destinations that compete for the same tourists
246 | (Koerbitz & Önder, 2013). Naturally, travel-planning systems also leverage on this
247 | aspect of GSMD to recommend international travel destinations
248 | (Alowibdi, Ghani, & Mokbel, 2014), nearby attractions (Zanker, Fuchs, Seebacher,
249 | Jessenitschnig, & Stromberger, 2009) and scenic routes (Chen, Shen, & Zhou, 2011;
250 | Sun, Fan, Bakillah, et al., 2013). Very few attempts have been made to characterize
251 | tourist flows with GSMD. To our knowledge, the earliest works were concerned with
252 | identifying distinctive flow patterns between popular tourist attractions (Girardin,
253 | Fiore, Ratti, & Blat, 2008) as well as to trace the type of events tourists participate in
254 | (Vaccari, et al., 2009). More recently, GSMD was utilized to compare seasonal demand
255 | for tourist accommodations (Sun, Fan, Helbich, & Zipf, 2013), investigate how the
256 | layout of cities and their tourism infrastructures influenced tourist behavior (Kadar,
257 | 2013) and reveal how destination preferences differ between demographic groups (Vu,
258 | et al., 2015).

259

260 | Among the methods we have described in this section, GSMD satisfies our data quality
261 | criteria to the fullest extent (See Table 1). Similarly, previous attempts to characterize
262 | tourist flows with GSMD show good potential for this analytical approach to shed light
263 | on the RQs presented above (See Section 2.1). Nonetheless, there are some caveats
264 | associated with its use (Nabian, Offenhuber, Vanky, & Ratti, 2013; Shelton, Poorthuis,
265 | & Zook, 2015) and it is paramount to acknowledge that certain information may be
266 | omitted from the gathered data. In this respect, user penetration is the key shortcoming
267 | of GSMD as access to smartphones and the practice of geotagging is essentially limited
268 | to a narrow user demographic of technologically savvy individuals (Murdock, 2011).

269 Of this demographic, studies have shown that males from densely populated, urban
270 regions are significantly overrepresented (Hecht & Stephens, 2014; Riederer, Zimmeck,
271 Phanord, Chaintreau,
272 & Bellovin, 2015) while ethnic minorities are marginalized (Mislove, Lehmann, Ahn,
273 Onnela, & Rosenquist, 2011). Reliability must also be considered, as users are not
274 obliged to provide real information (Hecht, Hong, Suh, & Chi, 2011).

275 **4. Methodology**

276 Geotagged Twitter data was chosen to operationalize our analysis. Apart from the
277 distinctive advantages mentioned above, Twitter provides freely accessible
278 mechanisms to monitor activities that occur within a specified area and timeframe.
279 However, several technical requirements must be fulfilled in order to perform flow
280 analysis with geotagged tweets⁸. Firstly, tweets are not directly downloadable from a
281 web repository but must be gathered from Twitter's application programming interfaces
282 (API) based on a well-defined set of queries. Secondly, the collected tweets in raw point
283 format are ill structured for flow analysis and require processing prior to visualization.
284 Finally, common flow representations are prone to visual clutter that occludes important
285 patterns (Schulz & Schumann, 2006) and thereby require optimization for accurate
286 representation. In this section, we provide technical details that elaborate on how data
287 is collected, processed and visualized for communicable insights.

288 **4.1. Data Collection**

289 Two distinct types of data are required, namely user profiles for demographic
290 information and geotagged tweets for spatial and temporal information. We carried out

⁸ Twitter posts are colloquially referred to as tweets.

291 data collection in three phases. The first phase involves gathering geotagged tweets
292 posted within the boundaries of Cilento between 29 May 2014 and 31
293 December 2014 through Twitter’s stream API (Twitter, 2014b). In phase two, we
294 compiled a list of unique user Ids extracted from the tweets collected in phase one.
295 Finally, we iteratively queried Twitter’s REST API (Twitter, 2014a) to gather
296 geotagged tweets and user profiles that correspond to each user Id on the list compiled
297 in phase two.

298 **4.2. Data Processing**

299 Our analysis requires a data structure that supports visualization and dynamic filtering
300 of aggregated trajectories based on spatial, temporal as well as demographic constraints.
301 For this purpose, trajectories must be extracted on a daily basis for each user Id and
302 organized into demographic groups based on the information in each user profile. We
303 describe our trajectory mining and tourist detection procedure below.

304 **4.2.1. Trajectory Mining**

305 A trajectory is a time-ordered collection of geotagged tweets that traces the movement
306 of an individual through space and time. Each sequential pair of tweets in a trajectory
307 is referred to as a *pathway* and reflects the movement of an individual from a location
308 to the next. Aggregation is typically required to extract collective patterns from the data.
309 Thus trajectories are simplified by binning the origin and destination of each pathway
310 to cells in an $n \times m$ grid. This step reduces the large number of spatial variations among
311 trajectories to a representative subset where the frequencies of travel along common
312 pathways become evident.

313

314 The aggregated trajectories are expressed as a directed graph $G(V, E)$ where vertices v_i
315 $\in V$ are cells in the grid that corresponds to physical locations in the region, while edges
316 $e_i \in E$ indicate movement pathways between cells. We identify two types of edges. A
317 directed edge e_{ij} is an aggregation of pathways with origins at vertex v_i and destinations
318 at vertex v_j . A self-directed edge e_{ii} is an aggregation of pathways where both origins
319 and destinations are vertex v_{ii} . Each edge is weighted by value f , indicating the aggregate
320 number of trips between an origin and destination. Thus, $f(e_{ij})$ refers to the frequency
321 of travel between v_i and v_j .

322 **4.2.2. Tourist Detection**

323 To characterize the demographic features of tourist flows in Cilento, we must determine
324 their location of origin. We obtained a demographic breakdown of the population
325 sample by grouping individuals according to the time-zone listed on their profiles. In
326 this manner, locals were differentiated from tourists, while foreign tourists were
327 distinguished from domestic tourists. Finally, foreign tourists were broken down into
328 sub groups. Time-zone was chosen instead of content from the location field because
329 the former is formatted in a consistent manner. Twitter users choose their time-zone
330 from a list of predefined cities, but any text information can be submitted to the location
331 field regardless of its validity (Hecht, et al., 2011). We obtained the timezone
332 information by systematically querying user profiles with Twitter's REST API.
333 Missing values were expected thus the classification was improved by clustering

340 individuals based on the frequency of their activity in Italy. We define four metrics to
341 accomplish this:

342

Td Number of days an individual has been active in Italy.

Cd Number of days an individual has spent in Cilento.

all \widetilde{Td} The median number of days that
 \widetilde{Cd} individuals have been collectively active.

The median number of days spent in Cilento by all individuals collectively.

343

344 The value Td indicates the total number of days an individual has created tweets,

345 while Cd refers to the number of days that those tweets occurred within the

346 geographic boundaries of Cilento. We consider any individual who did not declare a

347 ‘time-zone’ to be domestic tourist if $Td \geq \widetilde{Td}$. Then, we define a probability index 348 p

0.75 to distinguish domestic tourists from the locals on the basis that locals spend

349 most of their time within the region. The probability index p is computed in the 350

following manner:

$$p = \begin{cases} Cd/Td & \text{if } Cd \geq \widetilde{Cd} \\ 0 & \text{otherwise} \end{cases}$$

351

352 4.3. Data Visualization

353 We developed FlowSampler (Chua, Marcheggiani, Servillo, & Vande Moere, 2014), a

354 purpose built visualization tool that enables interactive visual analysis of spatial

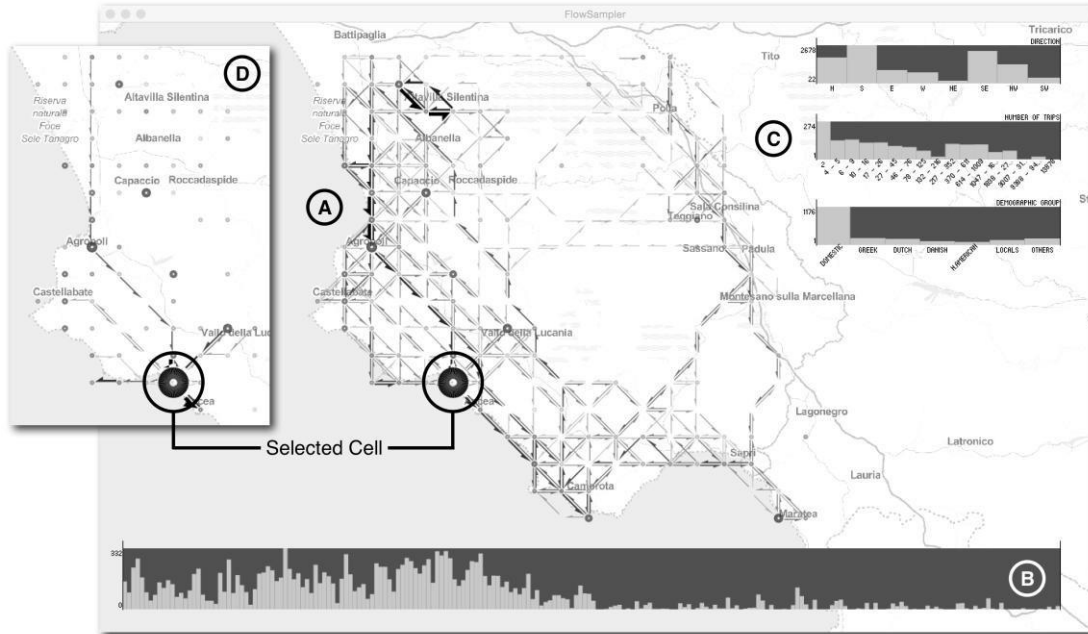
355 temporal patterns in an integrated view. As shown in Figure 1, the primary interface is

356 a flow map that depicts tourist flows among various locations in Cilento (see Figure

357 1a). The flow map can be dynamically filtered across four variables: Time (See Figure

358 1b), direction of travel, number of trips and demographic group (See Figure 1c).

359 Selecting a cell reveals the incoming and outgoing flows from that location (see
360 Figure 1d).
361



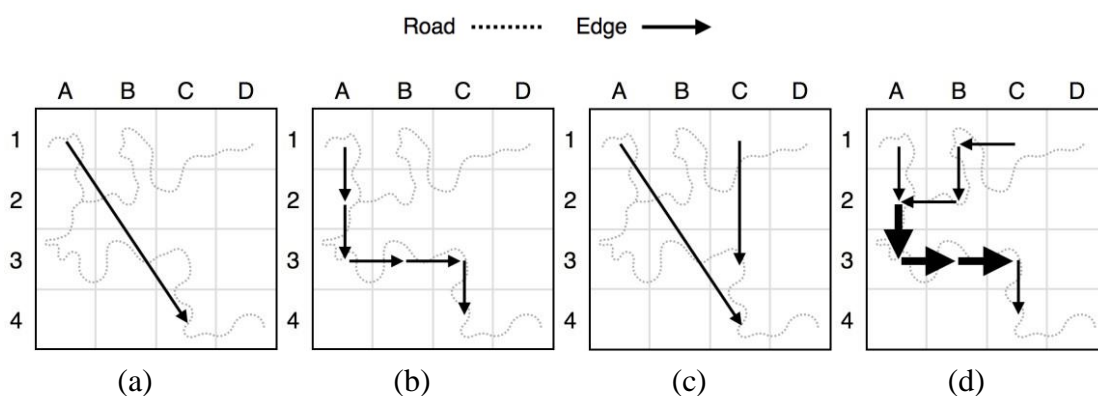
362
363 Figure 1. Components of the visualization interface. (a) Flow map geographically
364 centered on Cilento. (b) Timeline indicating the number of unique individuals posting
365 geotagged tweets per day. (c) Widgets for filtering the map based on direction of travel,
366 number of trips and demographic group. (d) Cut out of the flow map depicting both
367 incoming and outgoing flows from a selected cell.
368

369 4.3.1. Flow Map Optimization

370 A straightforward way to render flow maps is to draw arrows pointing from an origin
371 directly towards a destination. The width of the arrow is often scaled to indicate a
372 specific attribute value (e.g. Number of trips). This visual representation is easy to
373 understand when the data is sparsely distributed, but patterns become difficult to discern
374 when a large number of arrows intersect or overlap (Schulz & Schumann,

375 2006). While there are many ways to address this challenge, we employ a transition
 376 type representation (N. Andrienko & Andrienko, 2013) that divides each trajectory
 377 into segments, so that movement on a path is approximated by a sequence of
 378 transitions between adjacent discrete locations (See Figure 2). An advantage of this
 379 representation is that flows are captured as a directed graph that allows for
 380 quantitative metrics like centrality to be computed and visually compared.

381



382 Figure 2. Visualizing edges on the flow map. (a) Straight forward representation of an
 383 edge where an arrow is drawn between the origin and destination. (b) Shortest route
 384 representation of the edge shown in (a) based on the reference road network. (c)
 385 Straight forward representation of two edges. (d) Shortest route representation of two
 386 edges as illustrated in (c) where the thicker arrows depicts movement along a
 387 common path.

388 4.3.2. Route Identification

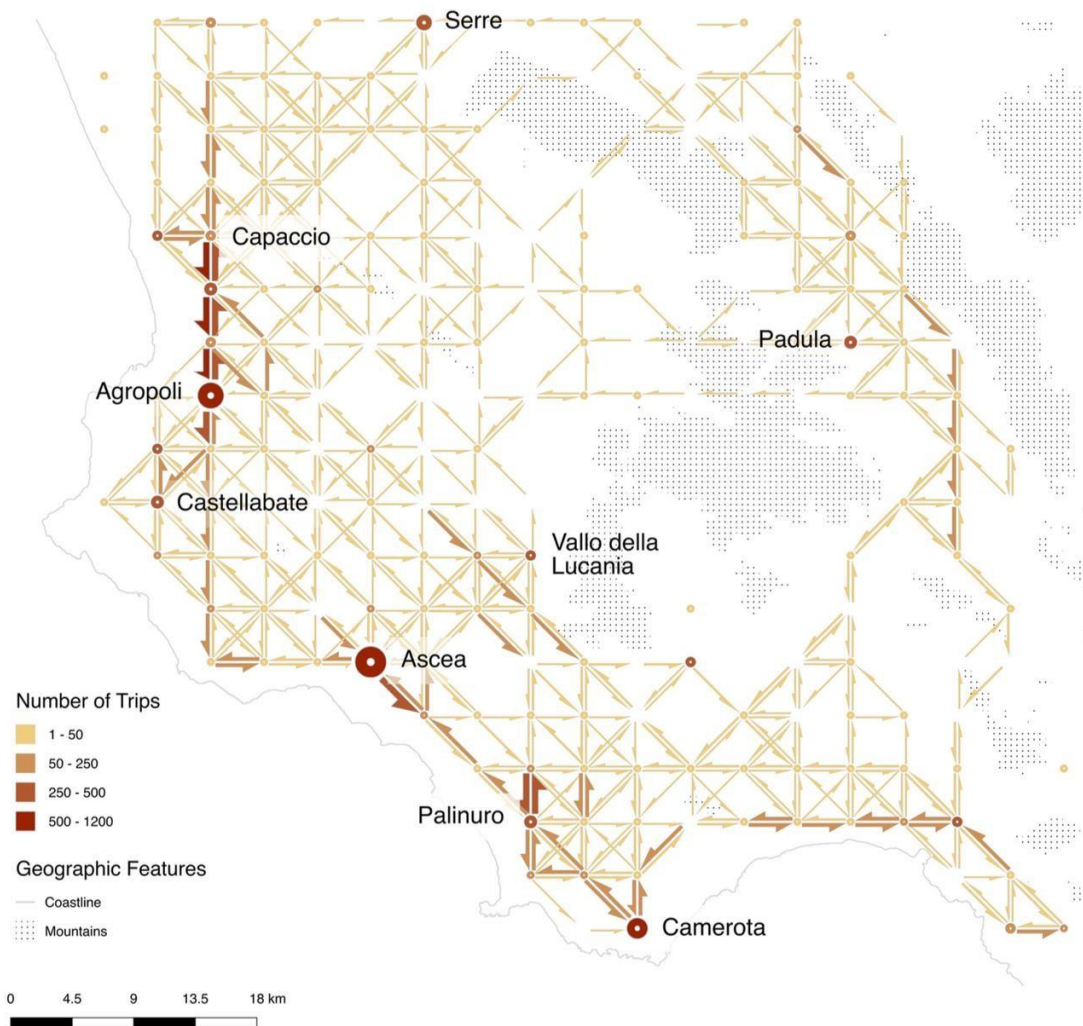
389 To identify possible travel routes, we render the shortest path between a pair of cells

390 based on a reference road network. This is preferred over more sophisticated
 391 routing

391 techniques, as it is straightforward to explain and simple for a lay audience to

392 understand. A diagrammatic explanation is shown in Figure 2. An arrow is drawn
393 between vertex A1 and C4 to represent the edge in a straightforward way (see Figure

394 2a). The same edge is represented in Figure 2b by a sequence of arrows that depict the
 395 shortest route between both cells. In this instance, the widths of the arrows are scaled
 396 according to the aggregate amount of movement along a common path. Figure 2c and
 397 2d illustrate a scenario whereby the shortest routes between two pairs of cells converge
 398 at cells A3, B3 and C3.
 399



400
 401 Figure 3. Visualization of flows derived from 3,135 individuals who have posted at least
 402 one tweet in Cilento.

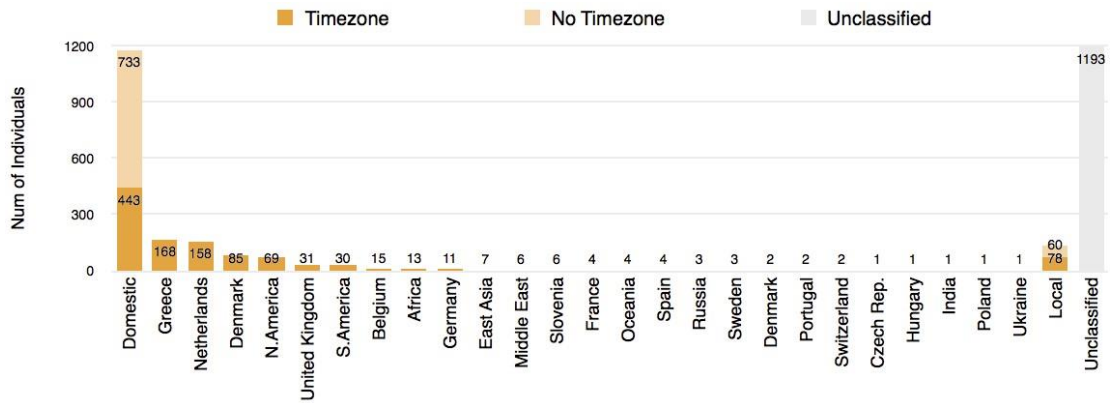
403 **5. Results**

404 Our data consisted of 72,031 geotagged tweets posted by 3,135 unique individuals. On
405 average, there were 193 (SD = 2.639) tweets per trajectory. The low standard deviation
406 indicates that majority of the trajectories contain a relatively similar number of tweets.
407 Figure 3 provides a summary of the flows in Cilento binned to a 19×20 grid. Each grid
408 cell measured 4.5×4.5 kilometers. This resolution was chosen to closely map every
409 settlement to a distinct cell. The width and color of the edges were binned to a scale
410 consisting of four value ranges. The map depicts more activity along the coastline and
411 reveals several important hubs in the region. There are several high frequency pathways
412 that indicate important flows around Ascea and Palinuro but the series of movements
413 between Capaccio and Agropoli is most distinctive.

414 **5.1. Demographic Breakdown of Tourist**

415 We detected 138 (4.4%) locals, 1176 (37.5%) domestic tourists and 628 (20%) foreign
416 tourists. 1193 (38.1%) individuals did not meet the classification conditions defined
417 above and excluded from further analysis. Figure 4 depicts the demographic breakdown
418 according to location of origin. Locations outside of Europe were aggregated into wider
419 geographic regions to simplify the classification. Our classification indicates that
420 Cilento primarily attracts domestic tourist. Correspondingly, Greek (5.4%) and Dutch
421 (5%) tourist account for more than half of foreign tourism. The discovery of tourists
422 originating from other locations suggests that the region attracts a diverse audience.

423



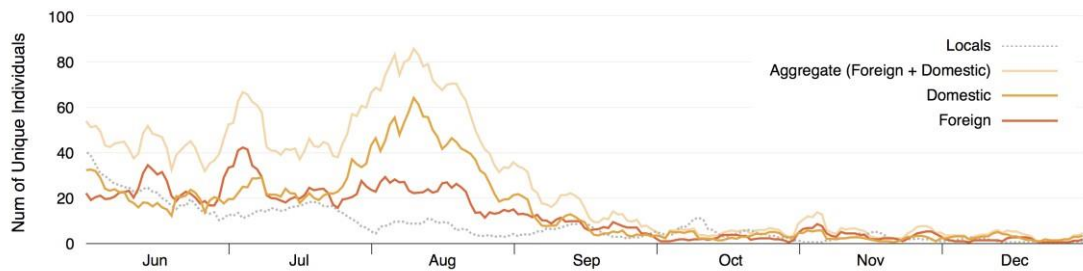
424

425 Figure 4. Demographic Breakdown of individuals in Cilento. Locations outside of
 426 Europe have been aggregated into wider geographic regions.

427 5.2. Uncovering Temporal Characteristics of Tourism

428 We compared variations in Twitter usage among demographic groups to uncover the
 429 temporal features of tourism in the region. Figure 5 indicates the aggregate number of
 430 people who posted geotagged tweets on a daily basis. The number of locals is depicted
 431 in grey, while tourist numbers are represented with three colored time series: An
 432 aggregated time series for tourists in general, and two other time series for to facilitate
 433 comparison between domestic and foreign tourists. The data shows an increase in tourist
 434 activity in late June followed by a peak in the second week of August. The trend declines
 435 by the first week of September. Incidentally, we observe a reverse trend among the
 436 locals during that period. The distinctive difference between domestic and foreign
 437 tourism is the timeframe during which they occur. This temporal pattern is characterized
 438 by the decline in foreign tourists numbers just as the presence of domestic tourist begin
 439 to escalate and peak. The bimodal distribution observed in the aggregate time series is
 440 also explained by this pattern. The highest number of tweets posted by foreign tourist is
 441 recorded during first week of July. A similar observation for domestic tourists occurs at
 442 the first week of August.

443



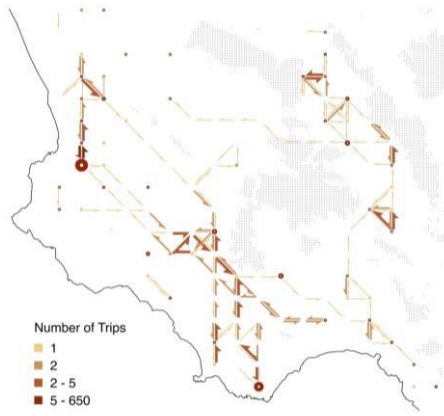
444

445 Figure 5. Temporal analysis of local and tourists twitter usage activity.

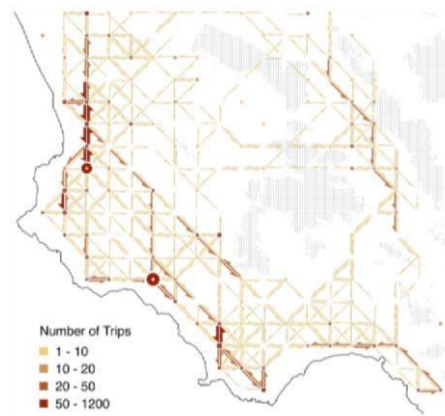
446 5.3. Spatial Topology of Tourist Flows

447 5.3.1. Circulation

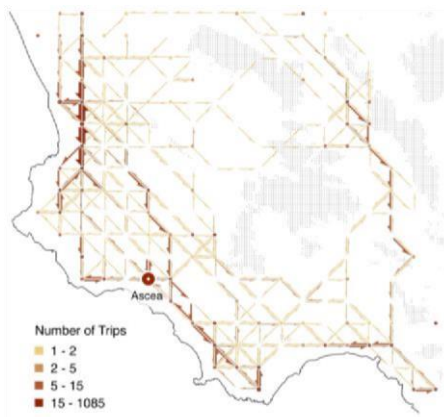
448 Circulation refers to the diffusion of flows in a system. Figure 6 presents a visual
 449 comparison of how individuals from various demographic groups move throughout the
 450 region. Figure 6a depicts the movements of locals while Figure 6b illustrates the
 451 aggregated tourist flows. Comparison between both maps reveals a substantial
 452 difference in the way individuals from both demographic groups circulate the region. In
 453 contrast to locals who primarily move inland, tourist activity tends to be situated along
 454 the coastline. The route between Capaccio and Agropoli however, appears to be equally
 455 important for both groups. Further comparisons between domestic (See Figure 6c) and
 456 foreign (See Figure 6d) tourists reveal several spatial differences. Whereas foreign
 457 tourists tend to be situated around Agropoli and Palinuro, domestic tourists are seen in
 458 Ascea. The routes taken by domestic tourists also differ substantially from foreign
 459 tourists in that they are situated further inland. Turning our attention solely on foreign
 460 tourism, we discover that Greek (See Figure 6e) and Dutch tourist, (See Figure 6f) travel
 461 to a diverse set of locations while those from Denmark (See Figure 6g) and North
 462 America (See Figure 6h) are limited to a smaller number of
 463 destinations.



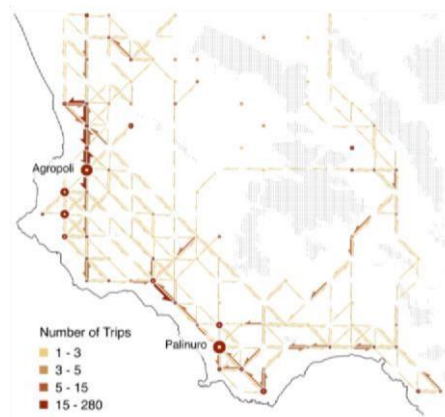
(a) Locals



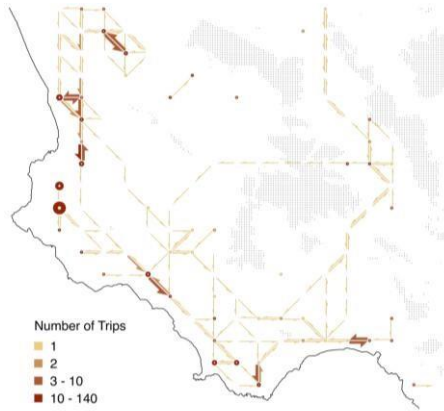
(b) All Tourist Aggregated



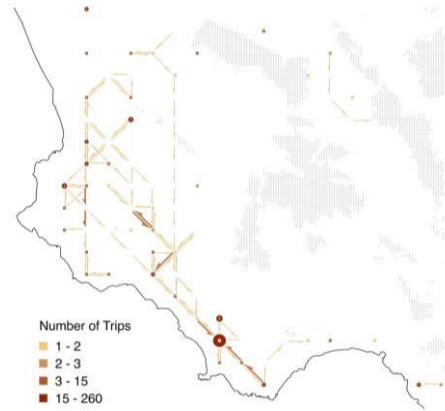
(c) Domestic Tourists



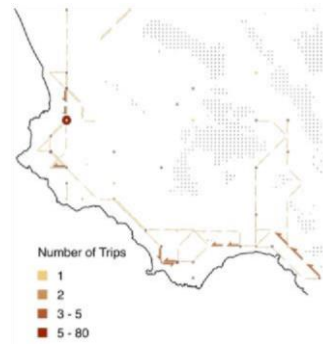
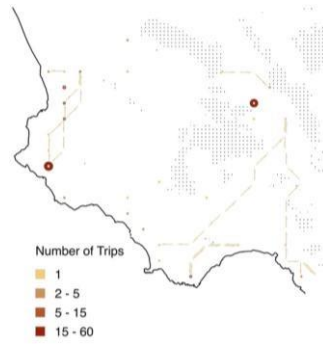
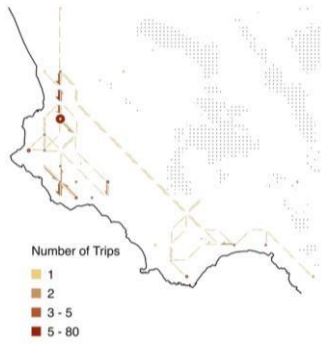
(d) All Foreign Tourists Aggregated



(e) Greek Tourists



(f) Dutch Tourist



465

(g) Danish Tourists

(h) N. American Tourists

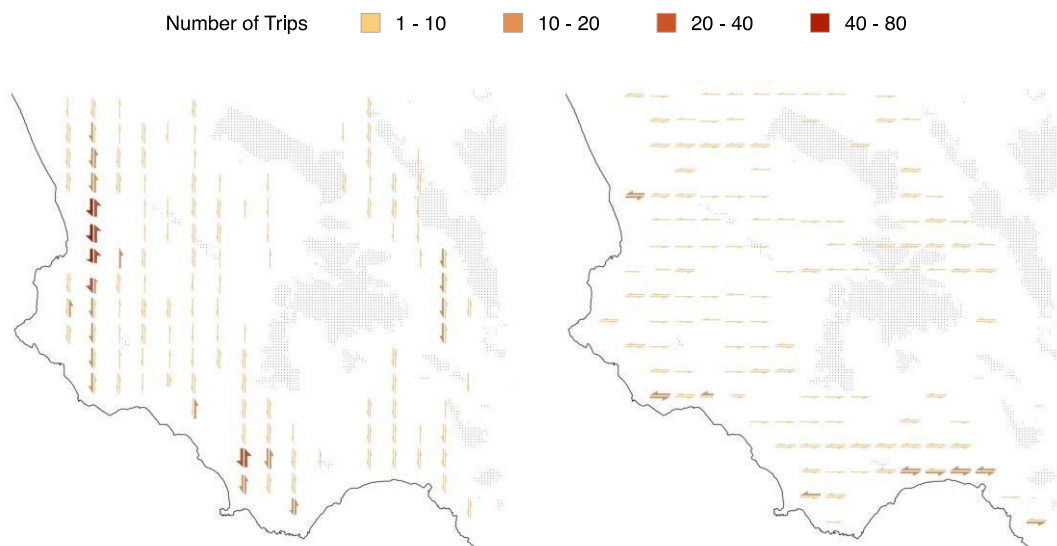
(i) Other Foreign Tourists

466 Figure 6. Comparing movements of tourists from different countries of origins.

467 5.3.2. Direction

468 Directionality is another spatial feature of movement. In Figure 7, we show the
469 aggregated tourist flows divided into four directions to compare the frequency of travel
470 per direction: N to S (See Figure 7a), E to W (See Figure 7b), NW to SE (See Figure
471 7c) and NE to SW (See Figure 7d). Self-directed flows have been removed from the
472 map, as they do not provide directional information. The maps reveal that tourists
473 primarily travel in a S-SE direction along the coastal settlements. To a lesser extent,
474 traces of N-NW bound travel in the opposite direction can be detected along the same
475 route. Filtering the map along the timeline indicates that northbound flows are traces of
476 egress that occur at the end of August when tourist depart.

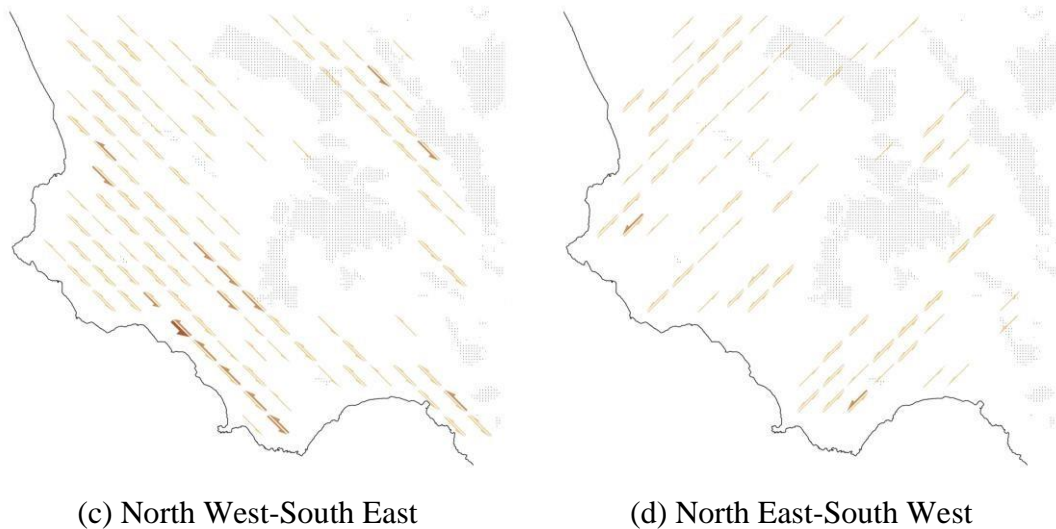
477



478

(a) North-South

(b) East-West



479
480

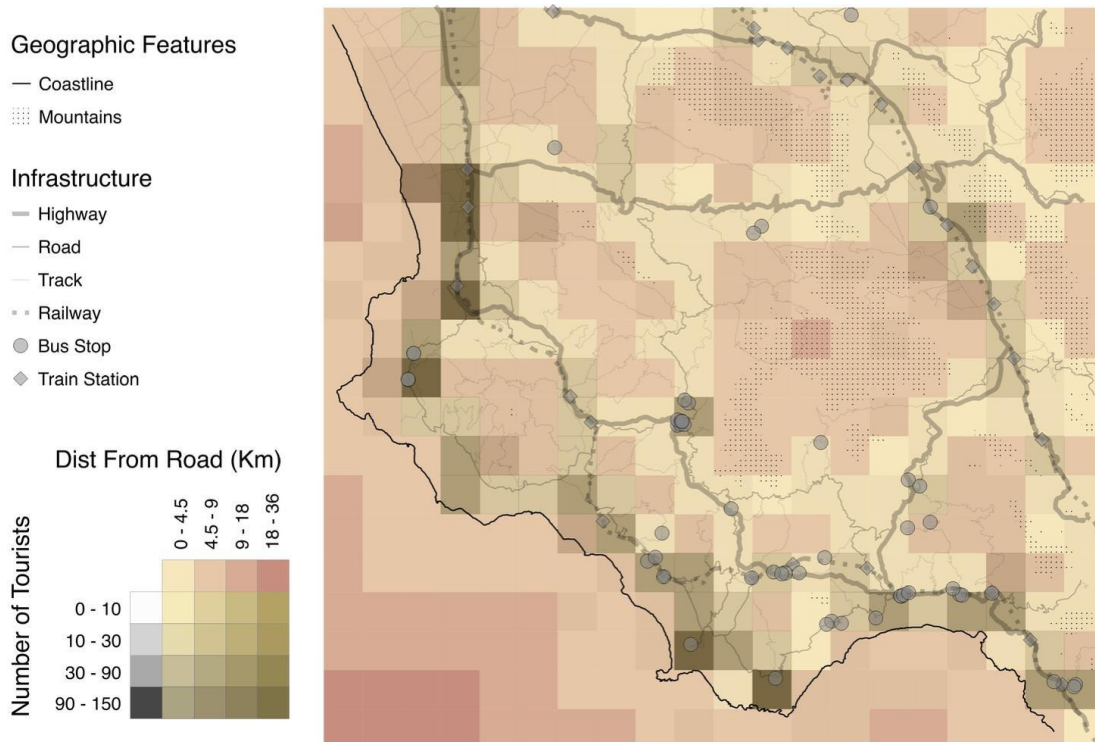
481 Figure 7. Comparing the directionality of all aggregated tourist flows along four
482 directions.

483 5.3.3. Centrality

484 The centrality of tourist attractions can be determined by its popularity among tourists
485 and is assessed in two stages. First, popular attractions were identified based on the
486 unique number of tourists recorded in each cell of the grid. Thereafter, the influence of
487 these attractions is evaluated in comparison with less popular venues on the basis of
488 connectivity; that is the geographic spread of inward and outward tourist flows to other
489 location in the region. Each cell in Figure 8 is colored in a shade of orange to indicate
490 its distance from the transport infrastructure and tinted in a shade of grey to represent
491 the unique number of tourists who posted tweets within its boundaries. Popular
492 attractions that recorded more than 90 (Q_3) unique visitors appear in the darkest shade
493 of brown – a blend of orange and grey. Unsurprisingly, these attractions are located
494 along the coastline where public transport is readily available. Additionally, we observe
495 that the number of tourists in a given cell decreases as its distance from the road or rail
496 network increases. Comparison of the connectivity between popular attractions to other
497 venues indicates an asymmetry in spread (See Figure 9). Whereas popular attractions

498 appear well connected to other locations (See Figure 9a), movement inland is limited
 499 to adjacent localities (See Figure 9b). Tracing movements over time reveals that tourists
 500 constantly travel in an “inland, coastal, inland” sequence in order to move between two
 501 disjoint inland locations.

502



503

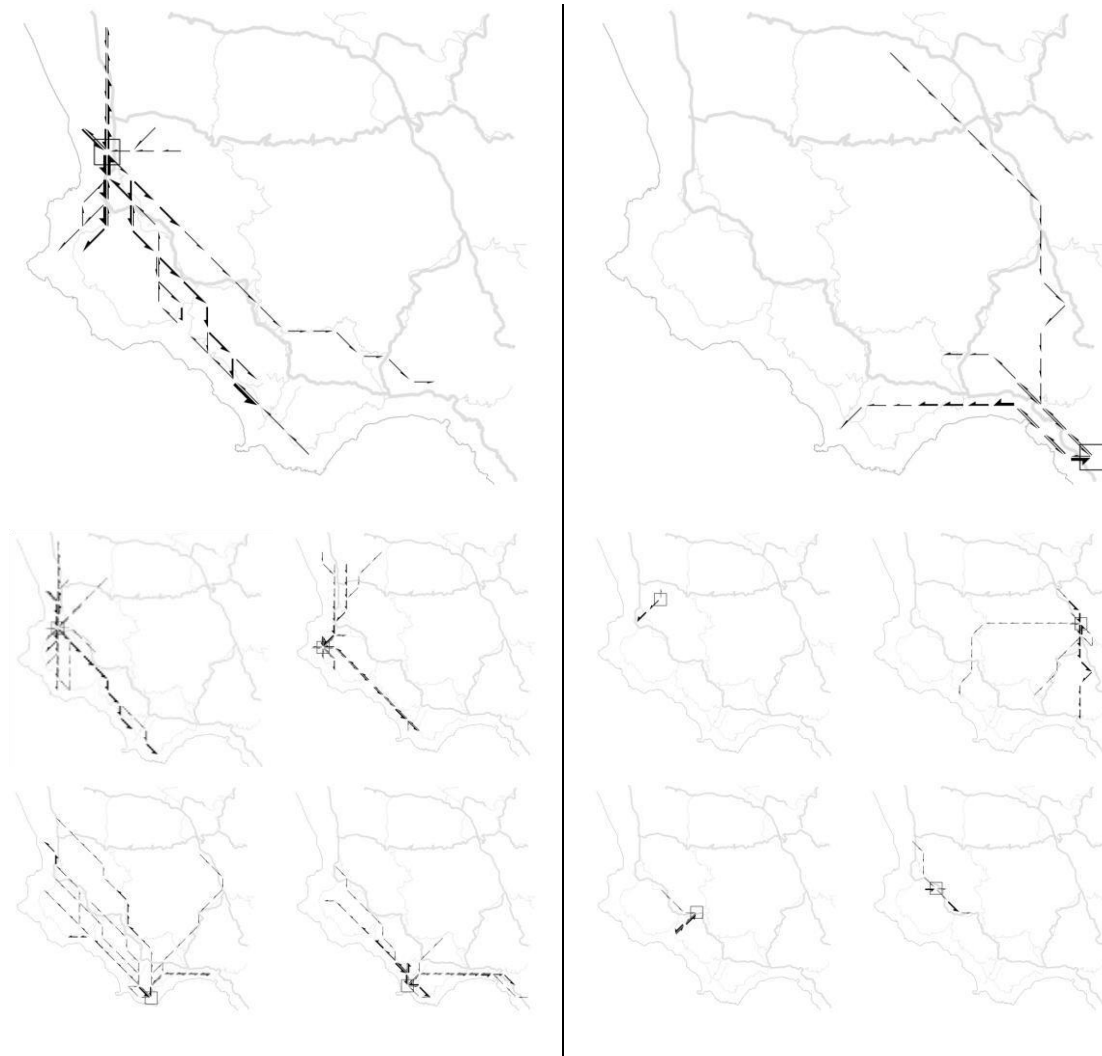
504 Figure 8. Popular locations in Cilento based on the total number of unique visitors in

505

each cell of the grid.



506



507

508

(a) Most Popular Locations

(b) Other Locations

509 Figure 9. Popular locations defined on the basis of connectivity. (a) Locations along the
 510 coastland are relatively well connected. (b) Inland flows are limited to adjacent
 511 localities.

512 **5.4. Insights**

513 In this sub-section, we review our findings structured around the initial RQs and
 514 describe our interpretation of tourist flow patterns in the region based on the
 515 information obtained.

516 **5.4.1. (RQ1) What are the meaningful tourist profiles in the region?**

517 There is a greater presence of domestic rather than foreign tourists in Cilento
518 ($\chi^2=166.4656$, $p<0.01$). According to our results, this finding is in agreement with
519 figures from our current understanding of tourist Flows (see Table 3). The breakdown
520 of foreign visitors into smaller groups indicates diverse tourist demographics. In
521 particular, the Greek and Dutch tourist account for more than half of foreign tourism
522 ($\chi^2=17.1221$, $p<0.01$). Danish and North American tourist also have a sizable presence
523 while remaining visitors originate from other locations in Europe, East Asia, South
524 America the Middle East and Africa.

525 **5.4.2. (RQ2) What are the valuable patterns of tourist flows in the region?**

526 We have observed several dominant spatial-temporal patterns in tourist flows.
527 Temporal analysis of Twitter usage reveals that foreign and domestic tourism does not
528 occur concurrently but in overlapping stages that peak in different moments in time.
529 Similarly, there is a substantial difference in the way locals and tourists circulate the
530 region ($t(352)=28.892$, $p<0.01$). Tourists generally travel along the coastline while
531 locals travel further inland. Tourists generally travel from the northwestern region of
532 Cilento heading in a southerly direction along the coastline towards destinations in the
533 south. This pattern of movement is prevalent among foreign and domestic tourists but
534 is not observed among locals who, on the contrary, travel northwards. Further
535 comparison between different groups of foreign tourist indicates two distinct patterns
536 of movement. Greek and Dutch tourists are observed to be more mobile than their
537 counterparts in that they travel longer distances while tourists from Denmark and North
538 America are limited to a smaller number of destinations. Finally, the recurrence of
539 “inland, coastal, inland” travel is particularly striking in that it reflects the importance
540 of the coastal settlements as transit hubs in the overall transport infrastructure.

541 **5.4.3. (RQ3) Where are the tourist attractions in the region and how do they**
542 **differ?**

543 We identified six popular attractions that spatially correspond to settlements in the
544 region (See Figure 9a). These attractions are located along the coastline where public
545 transport is readily available. Access to public transport appears to be paramount factor
546 that affects the popularity of attractions in Cilento since Vallo Della Lucania and
547 Padula, two UNESCO heritage sites situated inland, did not receive as many visitors as
548 Paestum, their counterpart along the coast. This pattern is prevalent among foreign
549 tourist and less pronounced with domestic tourists who were likely to have driven into
550 the region by car in on the highway situated inland. Comparatively, foreign tourists
551 travel along the coastline where the bus and train service is directly accessible.

552 **5.4.4. Interpretation**

553 Tourist flows in Cilento is likely to be defined by the physical configuration of the
554 existing transport infrastructure. Because public transportation does not extend beyond
555 the coast, tourists are required to obtain personal transport for inland travel. This insight
556 provides evidence for the relative popularity of the coastline and the lack of tourist
557 movement inland. The southward pattern of flow is likely to result from this
558 configuration as points of entry into Cilento, from regional transport hubs like Serre
559 and Capaccio are located at the northern half of the region while a scenic beach at
560 Camerota and a heritage site at Vallo di Diano are located at the southern half of the
561 region. Based on these findings, we believe that new economic opportunities can be
562 created by expanding the transport infrastructure inland. A permanent transport system
563 is costly to run. However, tailoring the service to a specific tourist demographic within
564 the duration of seasonal peaks, may lower operating cost and mitigate this issue. Since
565 foreign tourists are regarded to have greater financial spending power but are spatially

566 bounded to the coastline, a seasonal transportation service primarily catering to foreign
567 tourists is a feasible policy solution. Furthermore, tailored services like tours to inland
568 attractions can be organized and marketed to match the destination preference of
569 different tourist demographic groups. To this end, formal studies are required to
570 elaborate on how the existing transport infrastructure affects tourism.

571 **5.5. Discussion**

572 Existing knowledge of tourist flows in Cilento is derived from the analysis of regional
573 and provincial data, direct observation in the field or extracted from official reports,
574 publications and anecdotes from local staff and stakeholders. While sufficient to
575 broadly understand the context and challenges within the region, such information does
576 not meet our data quality criteria as outlined in Table 2. In this regard, our analytical
577 approach based on geotagged tweets demonstrates that GSMD is a valuable source of
578 tourist flow information. As shown in Table 3, the insights we obtained extend the
579 current understanding of tourist flow patterns. Specifically, our approach increased the
580 detail in spatial, temporal and demographic information available. In comparison to the
581 repertoire of methods that are currently in place, our approach is able to monitor tourist
582 movements across large geographic areas without the need for dedicated physical
583 sensing infrastructures. Though there were previous attempts to characterize tourist
584 flows with GSMD, these were limited to individual features of flow and generally
585 situated within cities. Conversely our analytical approach provides

584 relatively sophisticated descriptions of movement as well as profiles of tourists
585 over

586 large a region. In this respect, our analytical approach presents a substantial

587 advancement in describing tourist flows with GSMD.

588

589 Nonetheless, there are limitations to be aware of when drawing conclusions from

590 insights derived from analyzing GSMD since our approach is principally driven by

591 data provided at the social media users' discretion, without ground truth for

592 verification. The results might be somewhat misleading if biases in the data are

593 unaccounted for. As described earlier, demographic studies have shown that a large

594 part of Twitter users are young adults, and potentially represent only a partial slice of

595 the actual tourist population on the ground. Moreover, the data may not capture

596 complete travel itineraries, as mobile devices are less likely to be used when engaging in

597 certain types of activities. It is also unclear if cellular or GPS signal strength

598 affect social media usage or geotagging respectively. From this point of view, it is

599 crucial to acknowledge that the insights may overemphasize particular tourist

600 demographics, activities and attractions. We also expect the use of GSMD to
601 raise

602 privacy and ethical concerns related to the collection of data without direct
603 consent

604 from social media users. On this subject, it should be expressed that the data
605 gathered

606 includes only information that users explicitly disclose. Furthermore, the data is

aggregated in a way that all traces of individual trajectories are removed.

607

608 Our work constitutes an integrated approach for tourist flow analysis with limited

609 consideration of related computational methods. Trajectories are mined as observed

607 movements between origin and destinations without the actual travel paths in
608 | between. We introduced route identification to address this aspect of movement yet

609 | this information is purely inferred based the on shortest path between two locations that
610 | might be very different from the actual path taken. In this respect, more sophisticated
611 | routing techniques based on criteria such as well known points of interest could be
612 | implemented to obtain higher information accuracy. To make large analysis
613 | computationally tractable, we aggregated the data by dividing the observation area into
614 | grid of cells. As a consequence, the finest unit of analysis occurs on the level of cells.
615 | In this respect, movements within cells, although represented in our visualization, were
616 | not further investigated. Naturally, these cells contain urban areas within its boundaries
617 | and can be further subdivided for finer breakdowns of flow patterns. Alternatively,
618 | quadtree or kd-tree partitioning can be applied to derive grids that distinguish between
619 | urban and rural areas, based on a supplemental dataset like the whereabouts of buildings
620 | to better reflect urban density. Accordingly, our approach to demographic classification
621 | can also be improved. Considering that a substantial number of Twitter users could not
622 | be classified on the basis of time-zone, localization algorithms can be employed to
623 | determine their location of origin on the basis of content from the “location” field on
624 | their individual profiles.

625

626 | Despite these limitations, we would like to point out that our approach provides equally
627 | valuable and alternative insights that are complimentary to the current understanding of
628 | tourist flows in Cilento, derived from existing data sources. Additionally, we do not
629 | claim that it is an improvement on, or replacement for, other approaches to characterize
630 | tourist flows. While our findings are specific to Cilento, much of what has been found
631 | may be transferred to other regional destinations though the local context should always
632 | be taken into account. In this regard, future attempts at similar analysis may consider
633 | expanding the source data beyond geotagged tweets to include other GSMD like

634 geotagged photos, to capture a wider spectrum of the tourist demographics and
635 activities. Likewise, official tourism statistics maybe incorporated into the analytical
636 procedure to uncover potential biases in the data. Our work has focused on the spatial,
637 temporal and demographic features of tourist flows. Yet the embedded text content in
638 each tweet has been left out. In this regard, semantic and sentiment analysis may offer
639 new perspectives and provide rich contextual information for certain activities that
640 tourist engage in. Since GSMD is generally accessible through public APIs, large
641 datasets can be obtained at relatively low costs. Thus, similar analysis can be
642 undertaken by smaller organizations that have limited resources at their disposal when
643 tools for data collection, processing and visualization exist. Correspondingly, our
644 analytical approach can be useful for public administration and large tourism
645 enterprises to develop indicators for sustainable planning of territorial resources and
646 benchmarking markets respectively.

647

648 Table 3. Comparison of information from the analysis of GSMD to the current
649 understanding of tourist flows.

Findings Based on Geotagged Social Media Data		Current Understanding of Tourist Flows
Features	Insights	
Demographic	<ul style="list-style-type: none"> x Largely patronized by domestic tourists (67%). x Foreign tourists (33%) originate from many locations but Greek (9%), Dutch (8%), North American (4%) and Danish (4%) are most prominent. The remaining (8%) originate from other locations in Europe, East Asia, South America the Middle East and Africa. 	<ul style="list-style-type: none"> x More domestic (63%) than foreign (37%) tourists. x Anecdotes suggest that majority of the tourist are Dutch and Greek.
Spatial Circulation	<ul style="list-style-type: none"> x Limited mobility inland. x Tourists transit along the coastline to travel long distances. 	No Information

	Directionality	<ul style="list-style-type: none"> x Tourists primarily travel in a southerly direction passing through the coastal settlements on the road or rail network. x Southward flow due to the configuration of transport where transit hubs are located in the north of the region while the tourist attractions are located in the south. 	<ul style="list-style-type: none"> x No formal studies of tourist movements till date. x Anecdotal evidence regarding mode of transport suggest that domestic tourists drive to their destinations while foreign tourist journey to Cilento by train, where they alight at either Capaccio or Vallo di Lucania. Subsequent trips towards various locations in the region are then made by bus.
	Centrality	<ul style="list-style-type: none"> Popular tourist attractions are located along the coast and have immediate access to the transport infrastructure. As a result, these locations are better connected than those situated inland. 	<ul style="list-style-type: none"> x Currently no consensus on any form of ranking. x Anecdotes suggest that individual municipalities claim to be more important than others.
Temporal	<ul style="list-style-type: none"> Analysis reveals a bimodal trend where foreign and domestic tourist activities occur over different durations and peak at separate moments in time. 	<ul style="list-style-type: none"> Official tourists season begins on the 2nd week of May till the end of August. 	

650

651 **6. Conclusion**

652 We have described a set of findings from studying tourist flows through the lens of
653 GSMD. Our approach - developing an analytical technique to collect and investigate
654 the spatial, temporal and demographic features of tourist flows, enables relatively
655 sophisticated descriptions of tourist movement, as well as the demographic profiles of
656 tourist groups. However, biases in the data as well as methodological limitations should
657 be considered when drawing conclusions from analysis of GSMD.

658 Nonetheless, this is the first large-scale observational study of tourist flows that to our
659 knowledge attempt to provide a comprehensive description of tourist profiles and their
660 associated movement.

661 **7. References**

662 Alowibdi, J. S., Ghani, S., & Mokbel, M. F. (2014). VacationFinder: a tool for
663 collecting, analyzing, and visualizing geotagged Twitter data to find top
664 vacation spots. In *Proceedings of the 7th ACM SIGSPATIAL International*

665 *Workshop on Location-Based Social Networks*, 9-12.

666 Andrienko, G., Andrienko, N., Fuchs, G., Raimond, A.-M. O., Symanzik, J., &
667 Ziemlicki, C. (2013). Extracting semantics of individual places from movement
668 data by analyzing temporal patterns of visits. In *Proceedings of the*
669 *1st ACM SIGSPATIAL International Workshop on Computational Models of*
670 *Place*, 9-15, Orlando, FL, USA.

671 Andrienko, N., & Andrienko, G. (2013). Visual analytics of movement: An overview
672 of methods, tools and procedures. *Information Visualization*, 12(1), 3-24.

673 Ashley, C., De Brine, P., Lehr, A., & Wilde, H. (2007). *The role of the tourism sector*
674 *in expanding economic opportunity*: John F. Kennedy School of Government,
675 Harvard University.

676 Bakillah, M., Lauer, J., Liang, S., Zipf, A., Jokar Arsanjani, J., Loos, L., & Mobasheri,
677 A. (2014). Exploiting big VGI to improve routing and navigation services. In H.
678 A. Karimi (Ed.), *Big Data Techniques and Technologies in*
679 *Geoinformatics* (pp. 177-192). Boca Raton, FL, USA: CRC Press, Taylor &
680 Francis Group.

681 Barchiesi, D., Moat, H. S., Alis, C., Bishop, S., & Preis, T. (2015). Quantifying
682 international travel flows using Flickr. *PloS one*, 10(7), 1-8.

683 Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz,
684 M., Ouzounis, G., & Portugali, Y. (2012). Smart cities of the future. *The*
685 *European Physical Journal Special Topics*, 214(1), 481-518.

686 Beecham, R., Wood, J., & Bowerman, A. (2014). Studying commuting behaviours
687 using collaborative visual analytics. *Computers, Environment and Urban*
688 *Systems*, 47, 5-15.

689 Chae, J., Cui, Y., Jang, Y., Wang, G., Malik, A., & Ebert, D. S. (2015). Trajectorybased
690 visual analytics for anomalous human movement analysis using social
691 media. In *Proceedings of the EuroVis Workshop on Visual Analytics*, 43-47,
692 Cagliari, Sardinia, Italy.

693 Chen, Z., Shen, H. T., & Zhou, X. (2011). Discovering popular routes from trajectories.
694 In *Proceedings of the 27th International IEEE Conference on Data Engineering*
695 *(ICDE) 2011* 900-911.

696 Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring millions of footprints
697 in location sharing services. In *Proceedings of the 5th International AAAI*
698 *Conference on Weblogs and Social Media*, 81-88, Barcelona, Spain.

699 Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: user movement
700 in location-based social networks. In *Proceedings of the 17th ACM*
701 *SIGKDD International Conference on Knowledge Discovery and Data*
702 *Mining*, 1082-1090, San Diego, CA, USA.

703 Christaller, W. (1964). Some considerations of tourism location in Europe: the
704 peripheral regions - underdeveloped countries - recreation areas. *Papers in*
705 *Regional Science*, 12(1), 95-105.

706 Chua, A., Marcheggiani, E., Servillo, L., & Vande Moere, A. (2014). FlowSampler:
707 visual analysis of urban flows in geolocated social media data. In L. M. Aiello
708 & D. McFarland (Eds.), *Social Informatics* (pp. 5-17) Springer.

709 Cooper, C. (1981). Spatial and temporal patterns of tourist behaviour. *Regional*
710 *Studies*, 15(5), 359-371.

711 Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. M. (2012). The livelihoods project:
712 utilizing social media to understand the dynamics of a city. In

713 *Proceedings of the 6th International AAAI Conference on Weblogs and Social*
714 *Media*, 58-65, Trinity College, Ireland.

715 Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual exploration of
716 big spatio-temporal urban data: a study of new york city taxi trips. *IEEE*
717 *Transactions on Visualization and Computer Graphics*, 19(12), 2149-2158.

718 Fuchs, G., Andrienko, G., Andrienko, N., & Jankowski, P. (2013). Extracting personal
719 behavioral patterns from geo-referenced tweets. In *Proceedings of the*
720 *16th AGILE Conference on Geographic Information Science*, Leuven,
721 Belgium.

722 Gabrielli, L., Rinzivillo, S., Ronzano, F., & Villatoro, D. (2014). From tweets to
723 semantic trajectories: mining anomalous urban mobility patterns. In J. Nin &
724 D. Villatoro (Eds.), *Citizen in Sensor Networks* (pp. 26-35) Springer.

725 Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., & Blat, J. (2008). Digital footprinting:
726 uncovering tourists with user-generated content. *IEEE Pervasive*
727 *Computing*, 7(4), 36-43.

728 Girardin, F., Fiore, F. D., Ratti, C., & Blat, J. (2008). Leveraging explicitly disclosed
729 location information to understand tourist dynamics: a case study. *Journal of*
730 *Location Based Services*, 2(1), 41-56.

731 Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual
732 human mobility patterns. *Nature*, 453(7196), 779-782.

733 Hawelka, B., Sitko, I., Beinart, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C.
734 (2014). Geo-located Twitter as proxy for global mobility patterns.
735 *Cartography and Geographic Information Science*, 41(3), 260-271.

736 Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart:

737 the dynamics of the location field in user profiles. In *Proceedings of the*
738 *SIGCHI Conference on Human Factors in Computing Systems*, 237-246,
739 Vancouver, Canada.

740 Hecht, B., & Stephens, M. (2014). A tale of cities: urban biases in volunteered
741 geographic information. In *Proceedings of the 8th International AAAI*
742 *Conference on Weblogs and Social Media*, 197-205, Ann Arbor, Michigan,
743 USA.

744 Jiang, K., Yin, H., Wang, P., & Yu, N. (2013). Learning from contextual information
745 of geo-tagged web photos to rank personalized tourism attractions.
746 *Neurocomputing*, 119, 17-25.

747 Kadar, B. (2013). Differences in the spatial patterns of urban tourism in Vienna and
748 Prague. *Urbani izziv*, 24(2), 96-111.

749 Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*,
750 79(1), 1-14.

751 Koerbitz, W., & Önder, I. (2013). Destination benchmarking with geotagged
752 photographs. In Z. Xiang & I. Tussyadiah (Eds.), *Information and*
753 *Communication Technologies in Tourism 2014* (pp. 201-211) Springer.

754 Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the
755 global Twitter heartbeat: the geography of Twitter. *First Monday*, 18(5).

756 Lew, A., & McKercher, B. (2006). Modeling tourist movements: a local destination
757 analysis. *Annals of Tourism Research*, 33(2), 403-423.

758 Mansfeld, Y. (1990). Spatial patterns of international tourist flows: towards a
759 theoretical framework. *Progress in Human Geography*, 14(3), 372-390.

760 Mavoa, S., Oliver, M., Witten, K., & Badland, H. M. (2011). Linking GPS and travel
761 diary data using sequence alignment in a study of children's independent
762 mobility. *International Journal of Health Geographics*, 10(1), 1-10.

763 Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011).
764 Understanding the demographics of Twitter users. In *Proceedings of the 5th*
765 *International AAAI Conference on Weblogs and Social Media*, 554-557,
766 Barcelona, Spain.

767 Murdock, V. (2011). Your mileage may vary: on the limits of social media.
768 *SIGSPATIAL Special*, 3(2), 62-66.

769 Nabian, N., Offenhuber, D., Vanky, A., & Ratti, C. (2013). Data dimension: accessing
770 urban data and making it accessible. *Proceedings of the Institution of Civil*
771 *Engineers - Urban Design and Planning*, 166(1), 60-75.

772 Nagatani, T. (2002). The physics of traffic jams. *Reports on Progress in Physics*,
773 65(9), 1331.

774 Neuhaus, F. (2010). UrbanDiary - A tracking project capturing the beat and rhythm of
775 the city: using GPS devices to visualise individual and collective routines within
776 Central London. *The Journal of Space Syntax*, 1(2), 336.

777 Önder, I., Koerbitz, W., & Hubmann-Haidvogel, A. (2014). Tracing tourists by their
778 digital footprints. The case of Austria. *Journal of Travel Research*, 1-8.

779 Orsi, F., & Geneletti, D. (2013). Using geotagged photographs and GIS analysis to
780 estimate visitor flows in natural areas. *Journal for Nature Conservation*, 21(5),
781 359-368.

782 Prideaux, B. (2000). The role of the transport system in destination development.
783 *Tourism management*, 21(1), 53-63.

784 Riederer, C., Zimmeck, S., Phanord, C., Chaintreau, A., & Bellovin, S. M. (2015). "I
785 don't have a photograph, but you can have my footprints." – Revealing the
786 demographics of location data. In *Proceedings of the 9th International AAAI*
787 *Conference on Weblogs and Social Media*, University of Oxford, Oxford, UK.

788 Roth, C., Kang, S. M., Batty, M., & Barthélemy, M. (2011). Structure of urban
789 movements: polycentric activity and entangled hierarchical flows. *PloS one*,
790 6(1).

791 Schlich, R., & Axhausen, K. W. (2003). Habitual travel behaviour: evidence from a six-
792 week travel diary. *Transportation*, 30(1), 13-36.

793 Schulz, H. J., & Schumann, H. (2006). Visualizing graphs - A generalized view. In
794 *Proceedings of the 10th International Conference on Information*
795 *Visualization*, 166-173, London, England.

796 Sevtsuk, A., & Ratti, C. (2009). Does urban mobility have a daily routine? Explorations
797 using aggregate mobile network data. In *Proceedings of the 11th*
798 *International Conference on Computers in Urban Planning and Urban*
799 *Management*, Hong Kong, China.

800 Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: rethinking
801 urban socio-spatial inequality using user-generated geographic information.
802 *Landscape and Urban Planning*, 142, 198-211.

803 Shoal, N., & Isaacson, M. (2009). Methodological aspects of measurement and
804 visualization of tourists' spatial behavior. In S. Page (Ed.), *Tourist Mobility*
805 *and Advanced Tracking Technologies* (Vol. 19, pp. 28-43) Routledge.

- 806 Sieber, R. (2006). Public participation geographic information systems: a literature
807 review and framework. *Annals of the Association of American Geographers*,
808 96(3), 491-507.
- 809 Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in
810 human mobility. *Science*, 327(5968), 1018-1021.
- 811 Sun, Y., Fan, H., Bakillah, M., & Zipf, A. (2013). Road-based travel recommendation
812 using geo-tagged images. *Computers, Environment and Urban Systems*.
- 813 Sun, Y., Fan, H., Helbich, M., & Zipf, A. (2013). Analyzing human activities through
814 volunteered geographic information: using Flickr to analyze spatial and
815 temporal pattern of tourist accommodation. In J. M. Krisp (Ed.), *Progress in*
816 *Location-Based Services* (pp. 57-69) Springer.
- 817 Sykora, L., & Mulicek, O. (2014). Functional analysis of urban systems: identification
818 of small and medium sized towns and their territorial arrangements. In L.
819 Servillo (Ed.), *TOWN. Small and medium sized towns in their functional*
820 *territorial context. Scientific Report* (pp. 113-161). Luxembourg: ESPON.
- 821 Thornton, P., Williams, A., & Shaw, G. (1997). Revisiting time-space diaries: an
822 exploratory case study of tourist behaviour in Cornwall, England.
823 *Environment and Planning A*, 29(10), 1847-1867.
- 824 Twitter. (2014a). REST APIs. retrieved November 2014, from
825 <http://dev.twitter.com/rest/public>.
- 826 Twitter. (2014b). The streaming APIs. retrieved November 2014, from
827 <http://dev.twitter.com/streaming/overview>.

828 Vaccari, A., Liu, L., Biderman, A., Ratti, C., Pereira, F., Oliveirinha, J., & Gerber, A.
829 (2009). A holistic framework for the study of urban traces and the profiling of
830 urban processes and dynamics. In *Proceedings of the 12th International IEEE*
831 *Conference on Intelligent Transportation Systems.*, 1-6.

832 Vu, H. Q., Li, G., Law, R., & Ye, B. H. (2015). Exploring the travel behaviors of
833 inbound tourists to Hong Kong using geotagged photos. *Tourism management*,
834 *46*, 222-232.

835 Wakamiya, S., Lee, R., & Sumiya, K. (2013). Social-urban neighborhood search based
836 on crowd footprints network. In A. Jatowt, E. P. Lim, Y. Ding, A.
837 Miura, T. Tetzuka, G. Dias, K. Tanaka, A. Flanagan & B. T. Dai (Eds.), *Social*
838 *Informatics* (pp. 429-442) Springer.

839 Williams, S. (1998). Tourism, geography and geographies of tourism. In *Tourism*
840 *Geography* (2 ed., pp. 3-25) Psychology Press.

841 Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to
842 quantify nature-based tourism and recreation. *Scientific reports*, *3*(2976), 1-
843 7.

844 WTTC. (2015). Travel & tourism economic impact 2015. World. retrieved November
845 2015, from [http://www.wttc.org/research/economic-research/economic-](http://www.wttc.org/research/economic-research/economic-impact-analysis/)
846 [impact-analysis/](http://www.wttc.org/research/economic-research/economic-impact-analysis/).

847 Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information
848 search. *Tourism management*, *31*(2), 179-188.

- 849 Zanker, M., Fuchs, M., Seebacher, A., Jessenitschnig, M., & Stromberger, M. (2009).
850 An automated approach for deriving semantic annotations of tourism products
851 based on geospatial information. In *Proceedings of the International*
852 *Conference on Information and Communication Technologies in Tourism,*
853 211-221, Amsterdam, Netherlands.
- 854 Zheng, Y.-T., Zha, Z.-J., & Chua, T.-S. (2012). Mining travel patterns from geotagged
855 photos. *ACM Transactions on Intelligent Systems and Technology*, 3(3).

Appendix 1. Survey of the existing practice in each municipal of Cilento.

Municipals of Cilento	What are the existing methods used to gauge tourist demographic?	What are the existing methods used to determine tourist movements?	What are the existing methods used to detect tourist hotspots?	How is the popularity of tourist attractions benchmarked?
Agropoli	ISTAT* or on site interviews at public activities and receptions.	ISTAT* and surveys at tourism offices.	Information on the internet.	Opinion polls
Alfano	None	None	ISTAT* or EPT [§]	None
Ascea/Velia	EPT [§]	EPT [§]	Chamber of Commerce Periodical	None
Camerota	ISTA and EPT [§]	ISTAT*, EPT [§] and the Campania Region periodical.	Sectorial publications	Registration at the entrance of cultural and natural attractions.
Campora	None	Surveys at tourism offices when time permits.	None	None
Cannalonga	None		None	None
Casaletto Spartano	Ticket sales at local attractions.	Ticket sales at local attractions.	None	Ticket sales at local attractions.
Castelcivita	Observation among the caves situated within our municipality.	ISTAT*	ISTAT*	Ticket sales or registration at local attractions.
Castellabate	ISTAT*, surveys at tourism offices and check-ins to travel accommodations.	EPT [§]	EPT [§]	Ticket sales at the castle of the Abate and the museum of Sacra.
Centola/ Palinuro	Check-ins at travel accommodations.	On site interviews at public activities and receptions.	None	Surveys at various travel accommodations.
Ceraso	None	None	Observation	None
Cuccaro Vetere	None	ISTAT*	None	None
Gioi	None	None	None	None
Ispani	ISTAT* and check-ins to travel accommodations.	EPT [§]	None	None
Laurito	None	None	None	None
Moio della Civitella	None	Return tourism through database of second homes.	None	None
Montano Antilia	None	None	None	None.
Montesano sulla Marcellana	Ticket sales at two museums in the city of Montesano.	None	Interviews at two museums in the city of Montesano	None
Morigerati	WWF Periodicals	WWF Periodicals	None	None
Pertosa	Observation	Check-ins at travel accommodations.	None	None
Piaggine	None	None	Formal discussions organized by the Cilento Park Authority.	None
Rofrano	Check-ins at travel accommodations.	None	None	None
Roscigno	None	None	Surveys at tourism offices.	None
Rutino	None	None	None	None
San Rufo	ISTAT*, Chamber of Commerce Periodical and Reports from Cilento Park Authority.	None	Surveys at tourism offices, on site interviews at public activities and receptions.	Surveys at tourism offices, on site interviews at public activities and receptions.
Santo Angelo a Fasanella	Tour registrations at the caves of St. Angelo a Fasanella	Observations and on site interviews.	None	None
Sapri	ISTAT* and EPT [§]	Formal discussions with local stakeholders.	None	Formal discussions with local stakeholders.
Serramezzana	EPT [§] , observations and checkins at travel accommodations.	EPT [§] and reports by tour operators.	EPT [§]	None
Teggiano	Ticket sales at the Diocesan museum.	Ticket sales at the Diocesan museum.	None	Ticket sales at the Diocesan museum.
Torraca	None	None	None	None

Tortorella	EPT [§] , check-ins at travel accommodations.	EPT [§]	None	Ticket sales at local attractions.
------------	--	------------------	------	------------------------------------

858

859 * Istituto Nazionale di Statistica, Italian National Institute of Statistics

860 § Ente Provinciale per il Turismo, Sarlano Provincial Agency for Tourism

Figure 1
[Click here to download high resolution image](#)

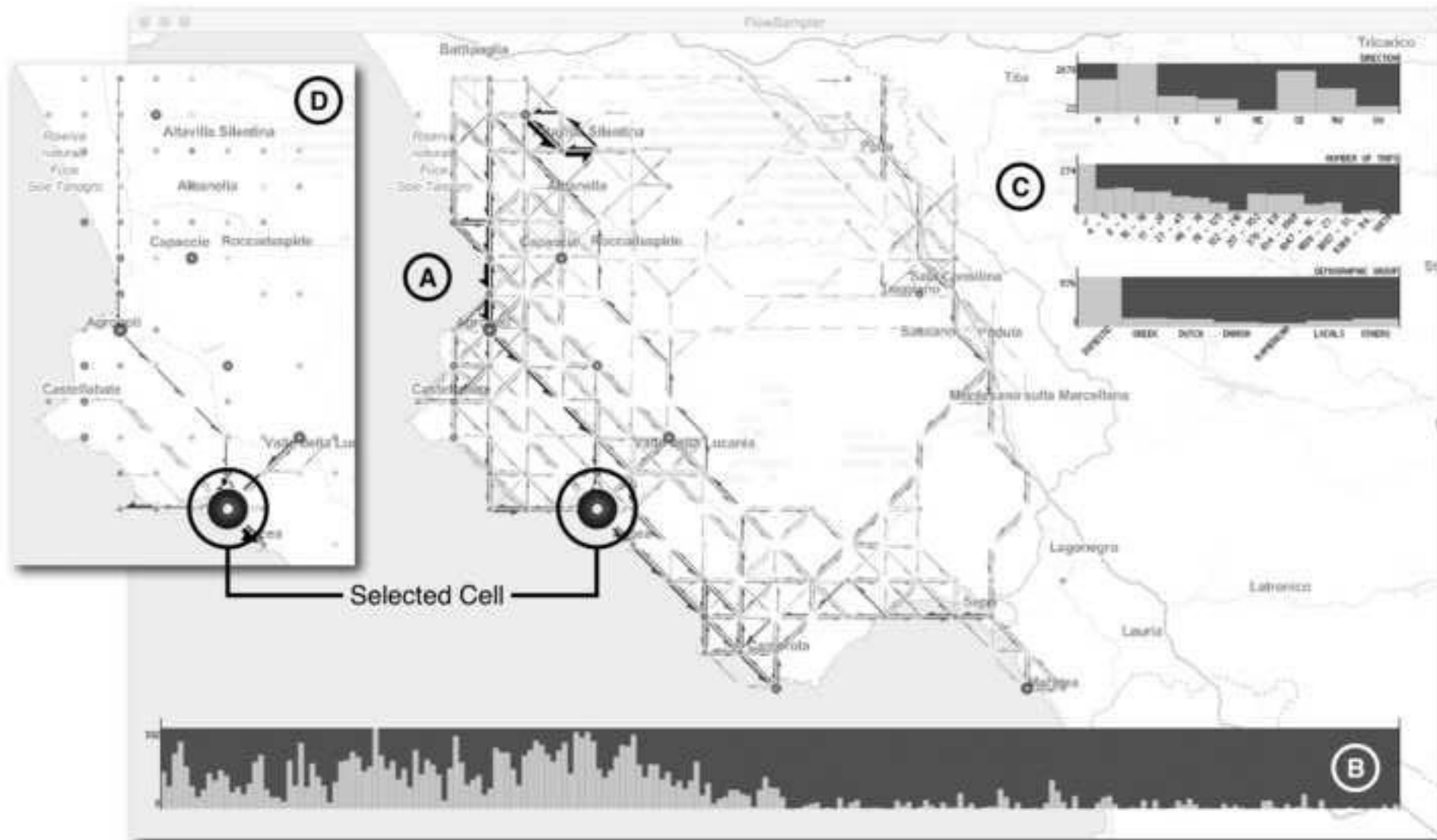
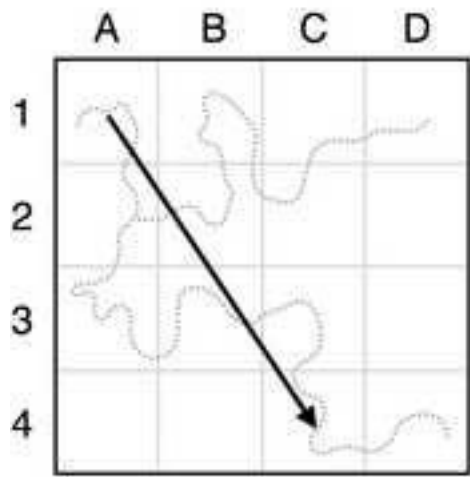
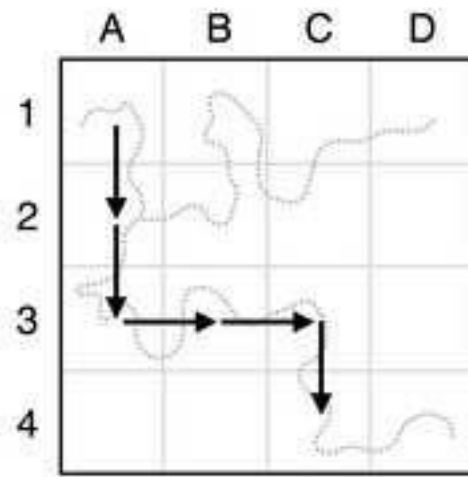


Figure 2
[Click here to download high resolution image](#)

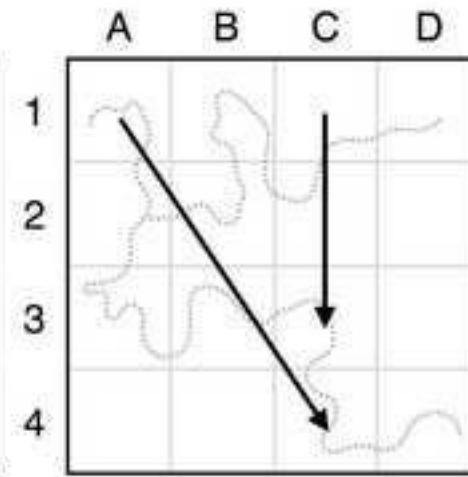
Road Edge →



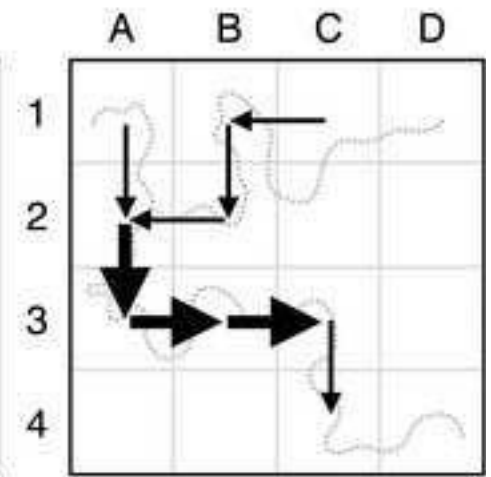
(a)



(b)



(c)



(d)

Figure 3
[Click here to download high resolution image](#)



Figure 4
[Click here to download high resolution image](#)

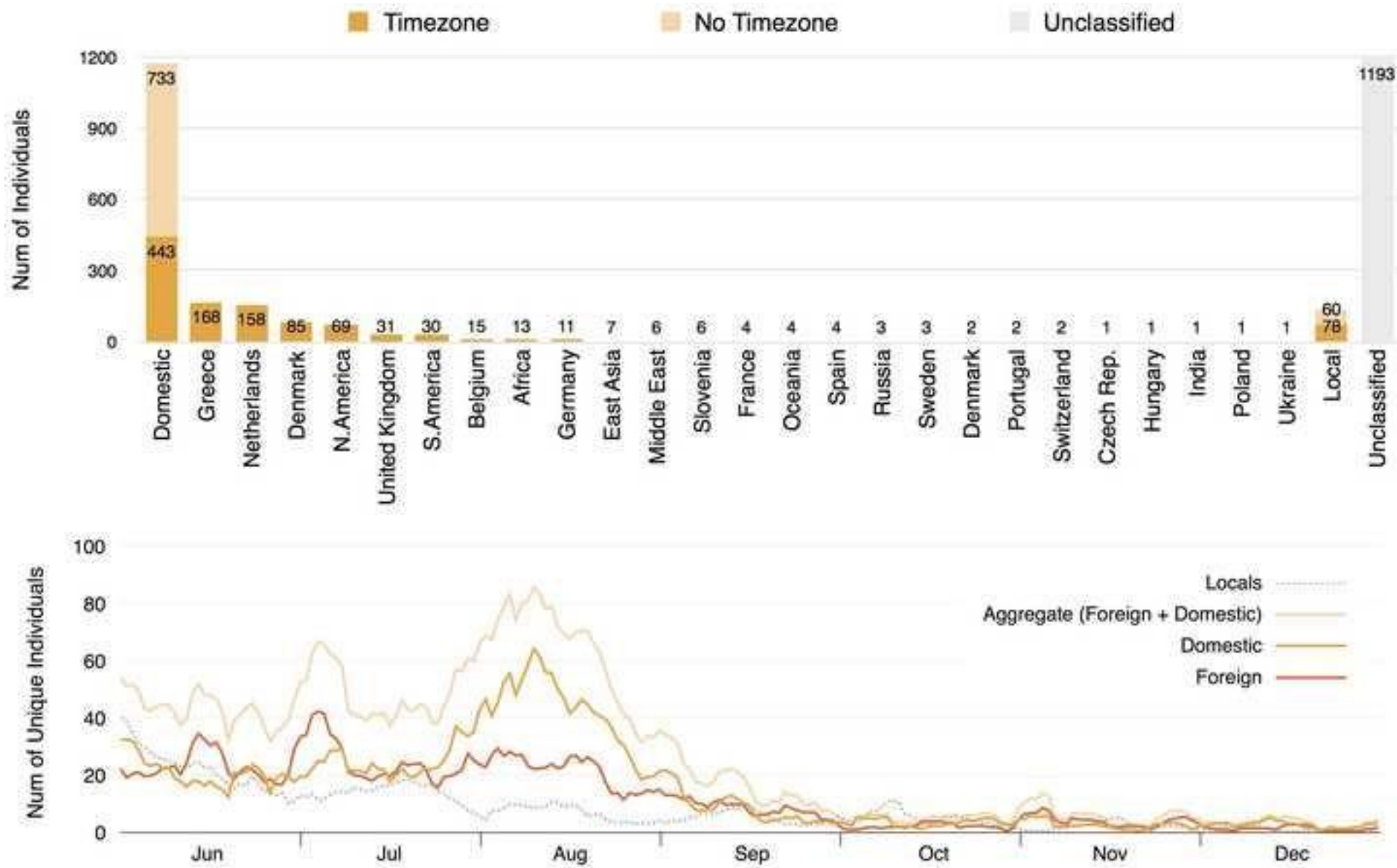


Figure 5
[Click here to download high resolution image](#)

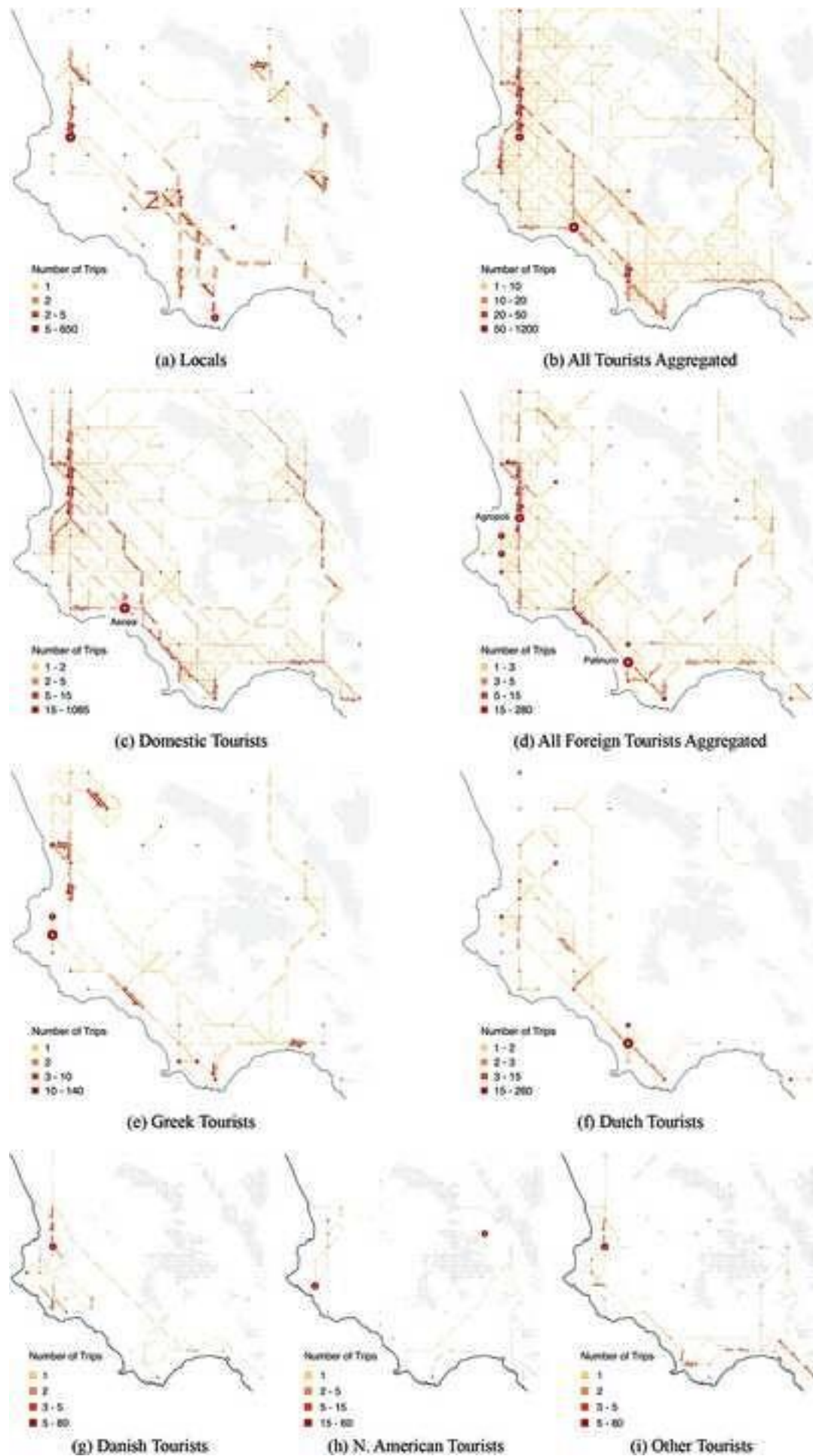


Figure 6

[Click here to download high resolution image](#)

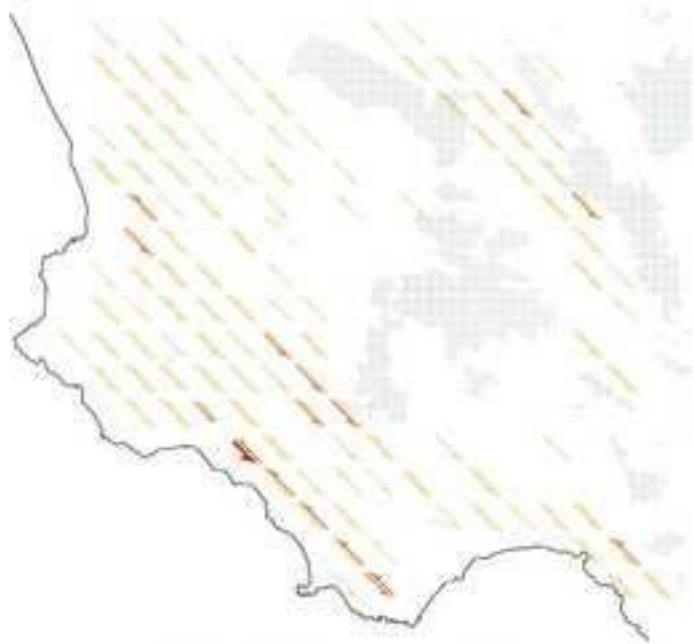
Number of Trips 1 - 10 10 - 20 20 - 40 40 - 80



(a) North-South



(b) East-West



(c) North West-South East



(d) North East-South West

Figure 7
[Click here to download high resolution image](#)

Geographic Features

- Coastline
- ▨ Mountains

Infrastructure

- Highway
- Road
- Track
- · Railway
- Bus Stop
- ◆ Train Station

Dist From Road (Km)

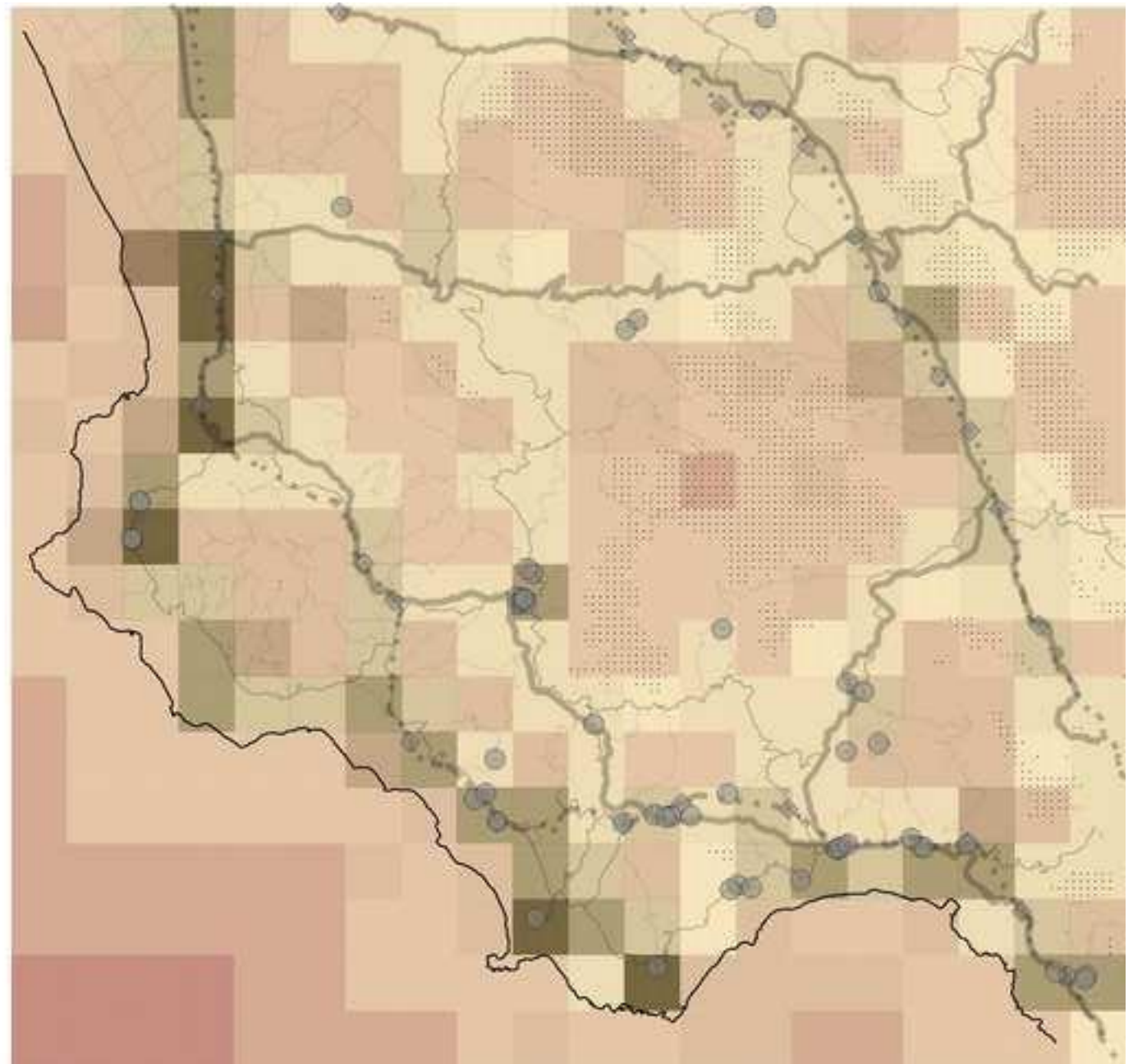
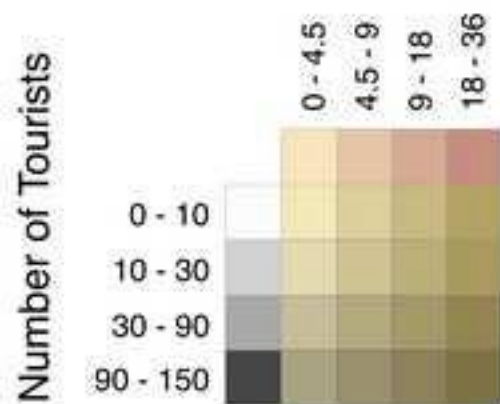
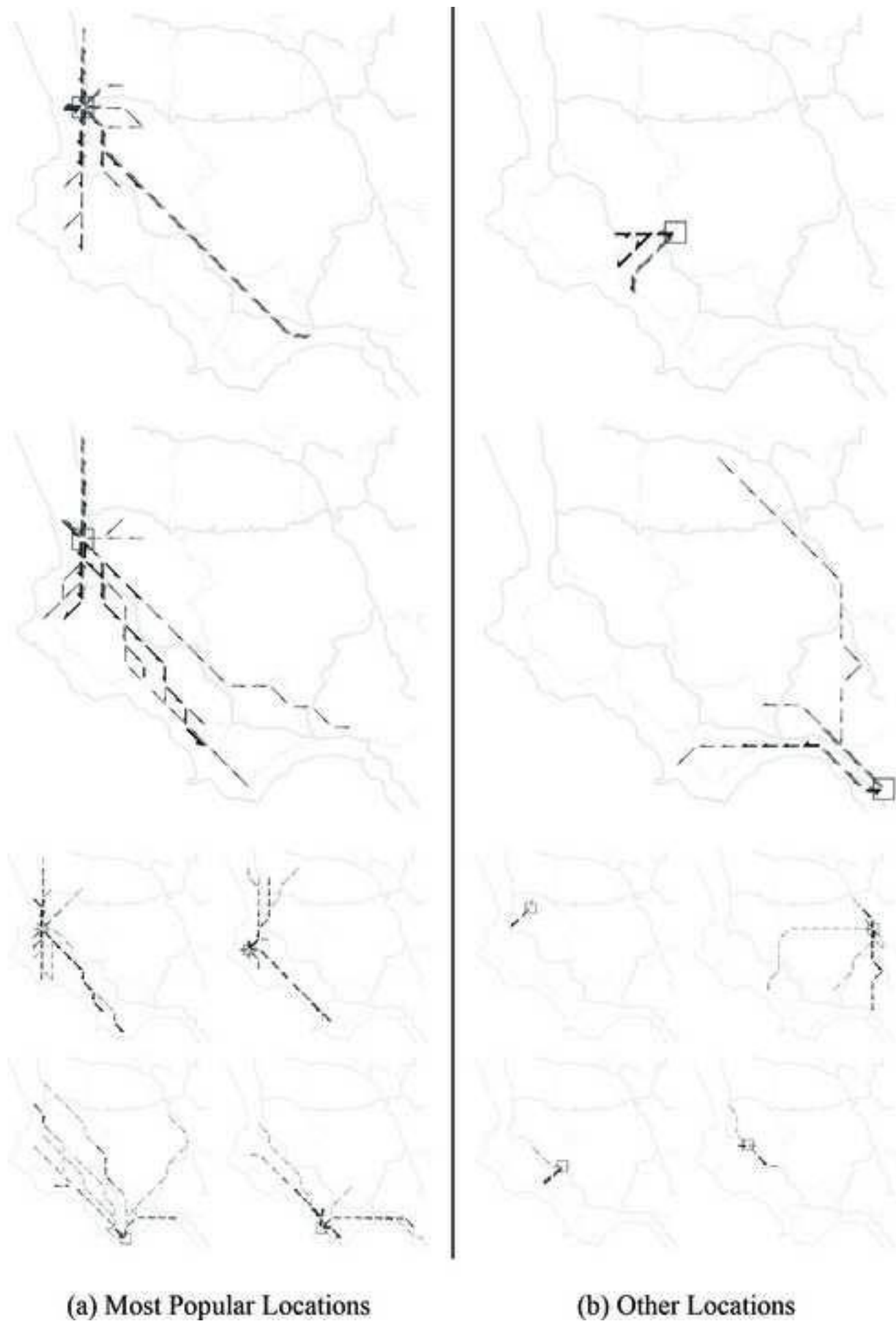


Figure 8
[Click here to download high resolution image](#)



***Author Biography**

Alvin Chua
PhD Candidate
Research [x] Design, Department of Architecture, KU Leuven, Belgium

Loris Servillo
Post-Doctoral Researcher
Department of Architecture, KU Leuven, Belgium

Ernesto Marcheggiani
Professor
Division of Forest, Nature and Landscape, KU Leuven, Belgium

Andrew Vande Moere
Professor
Research [x] Design, Department of Architecture, KU Leuven, Belgium

***Author Photo (to accompany biography)**

Alvin Chua



Loris Servillo



Ernesto Marcheggiani



Andrew Vande Moere

