

DATA MINING TOURISM PATTERNS

Call Detail Records as Complementary Tools for Urban Decision Making

NAI CHUN CHEN¹, JENNY XIE², PHIL TINN³, LUIS ALONSO⁴,
TAKEHIKO NAGAKURA⁵ and KENT LARSON⁶
^{1,3,4,5,6}*Massachusetts Institute of Technology, United States of America*
^{1,3,4,5,6} {naichun|ptinn|alonsolp|takehiko|kll}@mit.edu
²*Wellesley College, United States of America*
²jxie2@wellesley.edu

Abstract. In this study we show how Call Detail Record (CDR) can be used to better understand the travel patterns of visitors. We show how Origin-Destination (OD) Interactive Maps can provide transportation information through CDR. We then use aggregation of CDR to show the differences between the travel patterns of visitors from different countries and of different lengths of stay. We also show that visitors move differently during event periods and non-event periods, reflecting the importance of real-time data available by CDR. From CDR, we can gain more detailed and complete information about how tourists move compared to traditional surveys, which can be used to aid smarter transportation systems and urban resource planning.

Keywords. Machine Learning; Call Detail Record; Original-Destination Matrix; Urban Design Tool.

1. Introduction

The problems of congestion and resource allocation are essential to consider in urban planning and governance. These problems are becoming more severe as resources become more scarce. For a country like Andorra where the number of foreign visitors are high compared to the number of local residents, understanding the mobility patterns of tourists becomes important. Previously data collection related to tourism consisted of travel surveys completed from a sample of tourists. Travel surveys generally do not include locations in a tourist's path and are hence inadequate to fully represent the less regular and consistent travel paths of tourists.

Call Detail Record (CDR) is data from telephone calls collected by telephone service providers for billing purposes. CDRs are digital footprints of telephone

calls, including information of the time a call was made, and the corresponding cell tower used to process the call (Agung 2016). CDR can be used to more systematically and comprehensively track the movement and location of all tourists. However, the context of how CDR analysis can be applied to tourists is less explored compared to how CDR analysis can be applied to regular residents of a city. CDR analysis, as literature shows, can take advantage of machine learning, data visualization techniques to conduct more meaningful analysis, and in this paper we will use these methodologies in the context of tourism.

Following CDR analysis methodology used by Alex (Sandy) Pentland, at MIT Media Lab research group, we first created an aggregation map of tourist's CDR by time and nationality to understand population distributions throughout Andorra for different tourists events (Montjoye 2013). Then, the movement of tourists throughout Andorra for a single day was obtained via an OD Matrix and displayed. After understanding how tourists moved through the country, the Association Map was created with the association rule algorithm and the Random Forest algorithm over a year's worth of data to determine how connected cities were to one another for various months and holidays.

2. Literature Review

The utility of CDR in the context of urban data analysis is evident as the data is spatiotemporal. This non-conventional data source offers extensive information about human interaction that were not accessible to urban designers before. CDR is shown to be a good alternative to understanding mobility patterns compared to traditional methods of data collection like travel surveys for acquiring information of travel demand of cities, which are much more limited (Toole 2015) (Calabrese 2013). Much work has been done on how to apply CDR in the context of urban design (Jiang 2016) and transportation systems (White 2004). A mixture of machine learning, data mining, and data visualization techniques are used in literature to conduct meaningful analysis based on CDR.

Data visualization is a crucial tool to better communicate understandings derived from big data. This is demonstrated by Balduini et al., in their article "City-Sensing: Fusing City Data for Visual Storytelling" (2015), where researchers show how data visualizations can be used to make large amounts of information more accessible to humans while preserving the privacy of users through aggregation.

The use of origin-destination (OD) matrices to understand travel patterns through CDR is developed by Md. Shahadat Iqbal and his research team (2014). In 2015 these methodologies were applied and generalized to be applicable to different cities, serving as a unified guide to transportation demand modelling using OD matrices derived from CDR (Toole 2015).

However, many existing literature on CDR in the context of human mobility patterns are dependent on the regularity of the travel behavior of the callers, notably in situations where the locations of homes and workplaces are consistent and recurrent. The regularity and consistency in behaviour allows researchers to create predictive models of mobility patterns through behaviour modelling (Eagle 2009). In the article "A Hierarchical Approach for Identifying User Activity Patterns from

Mobile Phone Call Detail Records” (Khan 2015), the researchers use a hierarchical analytical model where CDR is increasingly filtered by daily mobility patterns to allow more detailed analysis of cellphone users in Bangladesh. The methodology of filtering this data, however, is dependent on consistency of an individual’s daily travels (primarily between work and home).

3. Methodology

The empirical sections of this work are based on the CDR provided, in a anonymised way, by the Andorran Telecom Company and collected throughout a year. In figure 1, we can see an example of the original data frame.

DS_CDNUMORIGEN	DT_CDDATAINICI	DT_CDDATAFI	NUM_DURADA	ID_CELLA_INI	ID_CELLA_FI	ID_CDOPERADORORIGEN
0 9aff9f9d53ebd77d2cc0edc6eddb761b0c1d5ae7166ea8...	2015.01.02 00:03:50	2015.01.02 00:03:50	0	2021	2021	20801
1 9aff9f9d53ebd77d2cc0edc6eddb761b0c1d5ae7166ea8...	2015.01.02 00:54:39	2015.01.02 00:54:39	0	2021	2021	20801
4 161e38db04f53740a00aaf1734d5b07a19381567261ebf...	2015.01.02 00:01:58	2015.01.02 00:01:58	0	19091	19091	20801
5 161e38db04f53740a00aaf1734d5b07a19381567261ebf...	2015.01.02 00:00:38	2015.01.02 00:00:38	0	19091	NaN	20801
11 a1497171f3fa9f67427840a9a2793445478f5dce51a717...	2015.01.02 01:24:00	2015.01.02 01:24:00	0	9162	9162	20801

Figure 1. Original Data Frame of CDR.

From the data we are able to classify tourists based on their country of residence from the code of their cell phone carrier. We are also able to obtain the location they were at when making a call from the start cell tower associated with the call. Using the ciphered phone number we are able to trace the path each caller travels based on the changing locations of their start cell tower.

3.1. ORIGIN-DESTINATION (OD) INTERACTIVE MAP:

Based on the CDR timestamps and geo-locations that we collected, we aggregated the data by tourists’ initial cell towers’ location and subsequent cell towers’ location, in order to obtain the mobility paths of tourists. As an example, we found that at 10 am in the morning, the road from the border city of Pas de la Casa (neighboring with France) to Encamp (5 km to the capital) in direction to the Andorra La Vella, the Capital, experiences heavy movement congestion, while at 7 pm at night there is high movement congestion from Soldeu to Encamp (in direction to France), see figure 2.

Thus, this type of analysis allows us to find functional traffic problems throughout the country that vary in the day, allowing us to suggest to the government ways to implement new transportation system in specific areas of the country where improvement is needed the most, see figure 3.

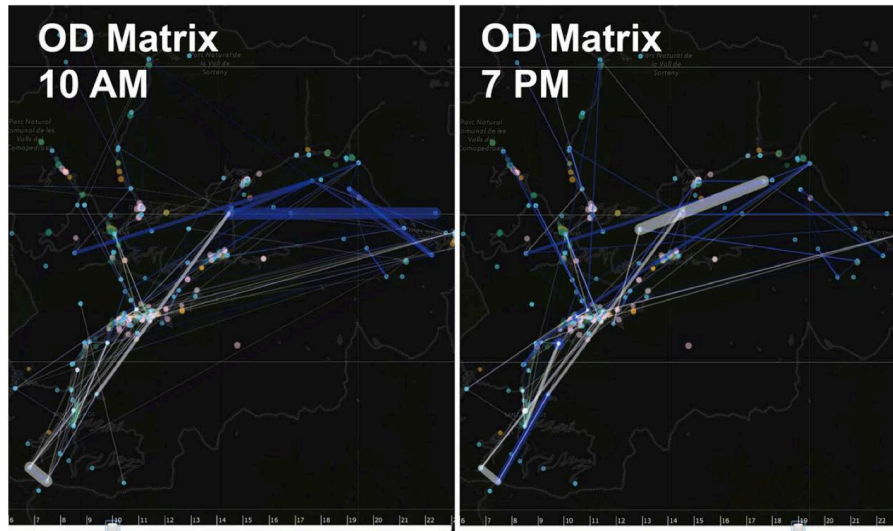


Figure 2. Origin-Destination Matrix showing changes in mobility pattern between morning and night-time, with white lines representing Spanish tourist movements and blue lines representing French tourist movements.

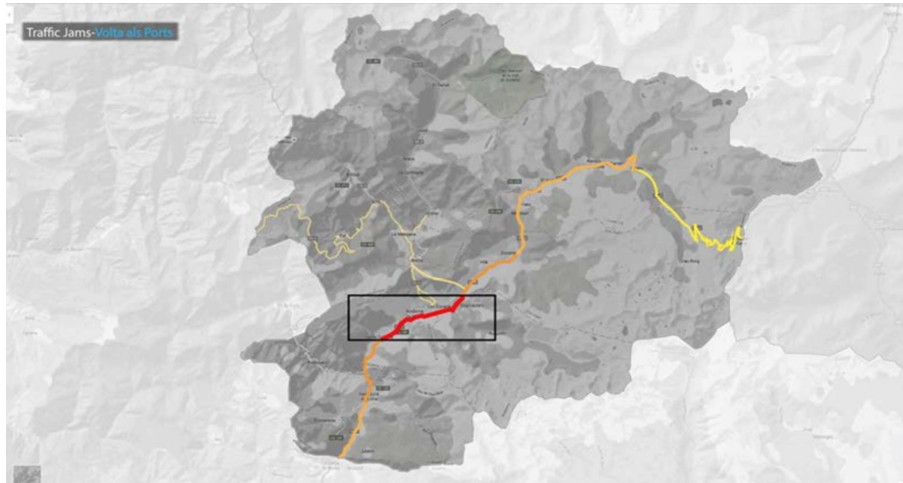


Figure 3. Traffic congestion on road segments.

4. Results

By connecting the locations of visitors as they make calls in different cities, we are able to aggregate the different paths of visitors to see which paths are most commonly taken by visitors of different countries. Our initial observations showed that tourists coming from the same country moved with remarkable consistency

within the same time of the year. Visitors from France and Spain make up to 82% of total visitors in Andorra. For French visitors the top 5 paths of travel between cities in July accounted for 54% of the total paths traveled by French visitors that month. For Spanish visitors the top 5 paths of travel between cities in July accounted for 62.7% of the total paths, see figure 4.

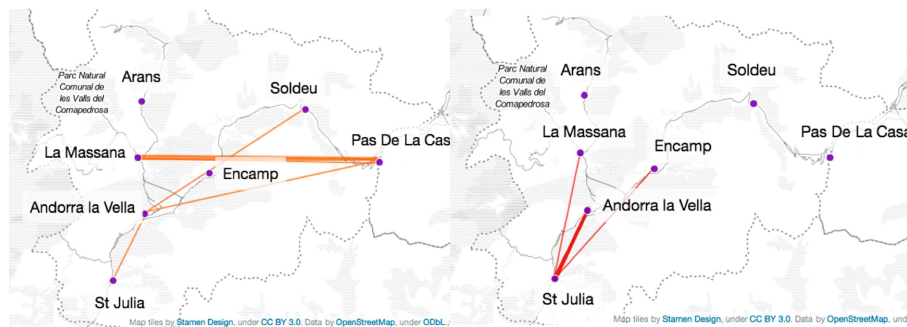


Figure 4. (Right) The orange lines and purple dots represent the top 15 paths traveled by French tourists in the summer. There is a mixture of dots and lines because some tourists only traveled within one city. (Left) The red lines and purple dots represent the top 15 paths traveled by Spanish tourists in the summer.

In the case of tourists in Andorra, one would think that the popularity of paths are highly dependent on proximity to country of residence because there are no airports or rail stations in Andorra, and hence most visitors are travelling by car or bus. To support this, we can see that the most popular path for Spanish visitors is St Julia, the westmost major city (closest to Spain), and the most popular path for French Visitors is Pas De La Casa, the eastmost major city (closest to France).

A closer observation shows that this does not cover the full story. From figure 4 we can see that though the paths of visitors from different countries are quite different, both St Julia and Centre made the top 5 for both countries, suggesting that they are major touristic cities despite proximity. In addition to that, La Massana, a city by proximity closer to Spain than to France, is much more popular for French tourists. This suggests that the popularity of cities are dependent not only on proximity but also other factors. Road transportation connectivity may be part of the reason.

We then look at the top 15 paths taken by French and Spanish visitors, see figure 4. The same seven cities made it into the top 15 paths travelled by both Spanish and French visitors, reflecting that these seven cities are likely to be the major cities for tourists in Andorra. Aside from that we can see that French visitors are more likely to move around the country, whereas Spanish visitors are more likely to stay in the west of the country.

Thus, this type of analysis allows us to find functional traffic problems the country, and allows us to suggest to the government implementation of new transportation systems in specific areas of the country that need it the most.

4.1. COMPARING VISITORS DURING EVENT PERIODS AND NON-EVENT PERIODS

Through our CDR we were also able to see how visitors during event periods traveled around the country compared to non-event periods. Tour de France is one of the biggest events in Andorra in the month of July, running for multiple days and bringing a noticeable increase of visitors.

In figure 5, we can see how Tour de France changed the paths of the visitors not from either France or Spain. We see how movement to and from St Julia becomes more common during Tour de France compared to July in general. At the same time, movement to and from Pas De La Casa becomes less common during Tour de la France compared to July in general. With this example, we learn that events can change how visitors move. This strengthens our argument that real-time information is important to understanding the transportation needs at a point in time.

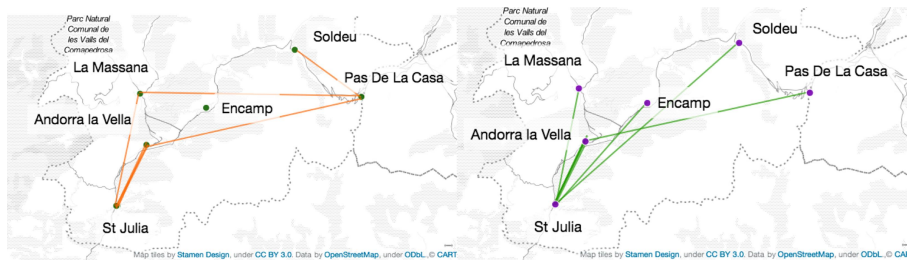


Figure 5. (Left) Top 15 paths travelled by visitors not from France or Spain during July. (Right) Top 15 paths travelled by visitors not from France or Spain during the period of Tour de La France.

5. Discussion

The present study provides evidence that the aggregation of CDR with other geolocated data can help the researchers to understand the patterns of different tourists that visit Andorra. After a validation and calibration of the methodology through a comparison with the surveys from Andorra Tourism, a series of interactive maps have been done to analyze and visualize the overlap of aggregated data, and to serve as a base for simulation-prediction tools like the Origin-Destination prediction tool presented in this paper. The simulation-prediction tools are designed as a decision making supporting tool [decision support system (DSS)] in the urban and territorial planning level.

As an example of our finds we can study the behavior of French and Spanish tourists. In figure 6 we can see that most of the Spanish tourists (the 40%) stay in Andorra La Vella (the Capital), which means that the Spanish tourists are focused on shopping and cultural events. The next likely movement is to go back to Spain (42% have St. Julia as a second step of their movement), though some of them continue to visit locations related to nature (more than the 14% have as a second movement cities related to nature and sports), which means that incentives and

new tourist and mobility policies could incentive Spanish tourists to visit other cities in the country as well, see figure 6.

On the other hand, if we focus on the flow of French tourists, we can realize that they are mostly interested in shopping (84% of French tourists in Andorra stay in Pas de la Casa, a very commercial city), and the next natural movement is to return home. Around the 16% of the French tourists have as second destination cities related to nature and sports, so once again, as with the Spaniards, this means that some incentives and new tourist and mobility policies can incentive them to visit other cities in the country, see figure 6.

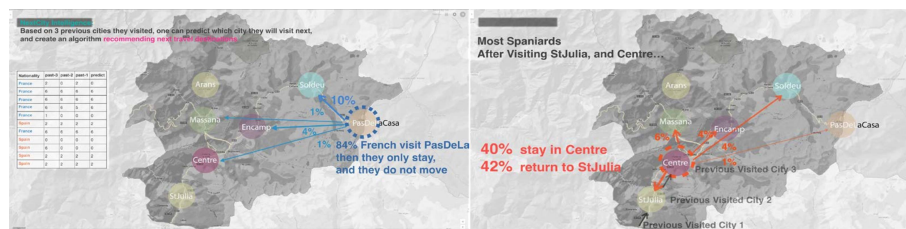


Figure 6. (Left) Next City Prediction in French Tourists. (Right) Next City Prediction in Spaniard Tourists.

5.1. URBAN DESIGN IMPLICATIONS

From all that we already have shown, we can conclude that the country of Andorra experiences a profound and focalized change of population during tourist events. This puts a large strain on resource allocation in the country. Resource allocation is a general problem in the 21st century world, owing to resource scarcity (Glenn and Warner 2016; Semertzidis 2015). In general, within cities, a significant amount of resource and energy waste stems from the rigid and centralized methods by which urban centers distribute resources. Schedules for distribution are typically static over time, blind to mismatch between supply and demand of resources, generating waste.

The use of linear regression models based on CDR data in order to predict district activity density is growing as a compelling methodology that promises to have great impact on the urban planning design. Our method of combining CDR and Open Data to inform the deployment of new mobility systems is aligned to this innovative methodology, and focuses on the optimization of the overall large-scale mobility networks. Figure 7 defines the path used by the method in order to achieve the mobility demand forecast.

The first urban design implication is that a well-stocked data gathering where a correlation between CDRs, social activities and environmental factors can easily be made is key in order to study, understand, and design a mobility system.

The second urban design implication is that in order to analyse and simulate a mobility demand system (what we call, in figure 7, the “Demand Pattern”), an iterative process of refinement is needed. This iterative process is divided into three steps:

1. Aggregate demand pattern: Based on a “real-time” origin-Destination (OD) Matrix, the pattern correlates the social and environmental influences (or impacts), so a compelling mobility behavior can be defined.

2. Decision Support for Mobility Planning: A mobility on-demand toolset is proposed as a design solution. A combination of on-demand/shared vehicles, transit service frequency adjustments, multi-modal planning and transit hub planning is modeled depending on the needs of the stakeholders, the aggregate demand pattern, and other mobility inputs.

3. Deployment simulation: Before a complex demand forecast simulation is deployed, a real-time traffic simulation (that it can be based on Google Traffic real-time data base) is recommended, so a fleet size sensitivity test can be conducted.

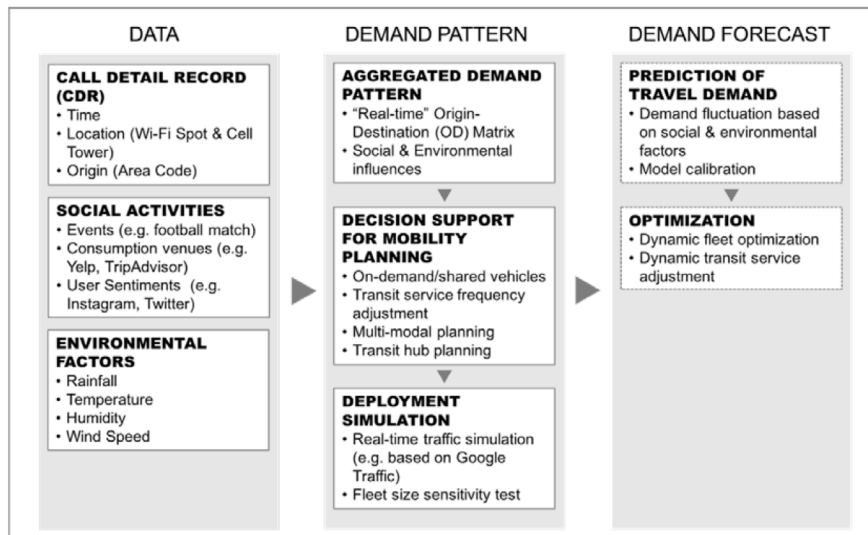


Figure 7. Method of combining CDR and Open Data to inform the deployment of new mobility systems, adjustment of existing transit services, and optimization of the overall network.

The last implication is that for defining the scale of the mobility system, a demand forecast is required. This last phase of design is focused on the calibration of the model based on the demand fluctuation dependent on social and environmental factors. As a final step of the method, the optimization of the model will bring the final numbers for the dynamic fleet optimization and the dynamic transit service adjustment.

Finally, from all that we have shown, we can conclude that the country of Andorra experiences a change in travel patterns during tourist events. This puts a large strain on resource allocation in the country, so a new mobility system based on our method is needed. In fact, resource allocation is a general problem in the 21st century world, owing to resource scarcity. Schedules for distribution are typically static over time, blind to mismatch between supply and demand of resources, and so, generates waste.

Andorra has an average volume of approximately 200,000 tourist per day, while the local population of Andorra is approximately 70,000. So a better understanding of the tourists (urban/country “users”) can help the territorial and urban decision making, spread resources in the economic ecosystem, and balance environmental impacts. A dynamical scheduling system can be suggested to solve this problem and also encourage tourists to explore more deeply the country, balancing the tourist’s environmental impact, spreading the tourist’s economic impact. We conceive that a dynamical system, updating resource allocation based on real time localized demand, is able to actively update areas of supply and demand mismatch in real time, and is thus able to optimize resource management at all times.

6. Conclusions and Further Research

In this project, we explored whether we could learn about the travel patterns of tourists in Andorra through CDR. Through this analysis, we were able to identify that by far the two largest groups of tourists in Andorra were from France and Spain. We were then able to identify different types of tourist travel plans among tourists of different countries of residence. The most popular paths for Spanish and French visitors and tourist is traveling to St Julia (Spanish 32.2%) or Pas De La Casa (French 25.6%). This pattern changes during the events, such as Tour de France.

In the future, we wish to explore this further. In collaboration with the Andorran Government and their institutions, we plan to evaluate the use of social media and CDR data to infer resource demands by testing three uses in three major urban planning resource problems:

- a) how to optimally distribute public transportation services and dynamically allocate vehicles.
- b) how to optimally distribute emergency services in event of man-made or natural disruptions.
- c) how to optimally distribute utilities (e.g. electricity) to prevent overloading the grid.

The final goal of this research is to find the elements that will allow us to design a high-performance, livable, entrepreneurial city.

References

- “The number of visitors to the Principality of Andorra in 2015 amounted to 7.851.152 people, that is 0,7% more than in 2014” : 2016. Available from <<http://all-andorra.com/tourist-number-2015/>> (accessed 8 Feb 2017).
- Agung, M. and Kistijantoro, I. A.: 2016, “High Performance CDR Processing with MapReduce”, *Journal of ICT Research & Applications Vol. 10 Issue 2*, 95-109.
- Balduini, M., Della Valle, E., Azzi, M., Larcher, R., Antonelli, F. and Ciuccarelli, P.: 2015, “CitySensing: Fusing City Data for Visual Storytelling”, *IEEE MultiMedia*, **22:3**, 44-53.
- Breiman, L.: 2001, “Random Forests”, *Machine Learning*, **45:1**, 5-32.
- Calabrese, F., Diao, M., Lorenzo, G., Ferreira, J. and Ratti, C.: 2013, Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transportation Research Part C: Emerging Technologies*, **26**, 301-313.
- Chua, A. and Servillo, L.: 2016, “Mapping Cilento: Using geotagged social media data to

- characterize tourist flows in southern Italy”, *Tourism Management*, **57**, 295-310.
- Eagle, N. and Pentland, A. S.: 2009, Eigenbehaviors: identifying structure in routine, *Behavioural Ecology and Sociobiology*, **63**(7), 1057-1066.
- Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C. and Pujolle, G.: 2014, “Estimating human trajectories and hotspots through mobile phone data”, *Computer Networks*, **64**, 296-307.
- Iqbal, M.S., Choudhury, C., Wang, P. and González, M.: 2014, “Development of origin-destination matrices using mobile phone cell data”, *Transportation Research Part C: Emerging Technologies*, **40**, 63-74.
- Jiang, S., Ferreira, J. and González, M. C.: 2016, Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore, *IEEE Transactions on Big Data*, **99**, 1-1.
- Jones, G.A. and Warner, K.J.: 2016, “The 21st century population-energy-climate nexus”, *Energy Policy*, **93**, 206-212.
- Liu, F., Janssens, D., Wets, G. and Cools, M.: 2013, “Annotating mobile phone location data with activity purposes using machine learning algorithms”, *Expert Systems with Applications: An International Journal*, **40**:9, 3299-3311.
- Monasterio, J., Salles, A., Lang, C., Weinberg, D., Minnoni, M., Travizano, M. and Sarraute, C.: 2016, Analyzing the spread of chagas disease with mobile phone data, *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.
- Montjoye, Y.A., Quoidbach, J., Robic, F. and Pentland, A.: 2013, “Predicting Personality Using Novel Mobile Phone-Based Metrics”, *Social Computing, Behavioral-Cultural Modeling and Prediction*, **7812**, 48-55.
- Semertzidis, T.: 2015, “Can Energy Systems Models Address the Resource Nexus?”, *Energy Procedia*, **83**, 279-288.
- Toole, J., Colak, S., Strut, B., Alexander, L., Evsukoff, A. and González, M.C.: 2015, “The path most traveled: Travel demand estimation using big data resources”, *Transportation Research Part C: Emerging Technologies*, **58**, 162-177.
- White, J., Quick, J. and Philippou, P.: 2004, The use of mobile phone location data for traffic information, *IEE International Conference on Road Transport Information and Control*.