Data Science For Social Good Europe

Summer Fellowship 2018

CASCAIS NOVA NOVA SCHOOL OF BUSINESS & ECONOMICS THE UNIVERSITY OF CHICAGO

Towards a Sustainable Future of Tourism in Tuscany

Fellows: Bruno Del Papa, Kaushik Mohan,
Orsi Vasarhelyi, Yanbing Wang

Project manager: Gabriele Simeone

Technical mentors: Qiwei Han, Nuno Brás

Table of Contents

Executive Summary

Tuscany, a region in central Italy, is one of the most popular tourist destinations in Europe. With a population of approximately 3.5 million[1], it is estimated by the *Istituto Regionale Programmazione Economica della Toscana* (IRPET) that around 30 million tourists visited Tuscany in 2017[2].

This record does not come without challenges. In the last few years, mass tourism has led to overcrowding in many cultural capitals across Europe. Mass tourism is often associated with an increased cost of living[3] and, on the long-run, it might lead to a loss of cultural authenticity.

On the other hand, the region's economy largely relies on tourism, which contributes to around 7% of the region's GDP[4], creating jobs and stimulating economic growth. The management of the mobility of millions of Italian and foreign tourists who visit Tuscany every year is central to any policy that attempts to balance economic prosperity with environmental and social sustainability of this unique region.

Toscana Promozione Turistica (TPT) is a regional government agency in Tuscany responsible for promoting tourism. Its main objective is to direct tourist flows throughout the region by consolidating Tuscany's presence in traditional markets and developing it in emerging markets. TPT plays an important role in the region's endeavour of seeking sustainable tourism strategies.

As part of the Data Science for Social Good Europe Summer Fellowship Programme, our project worked with TPT to understand and measure the patterns of tourist mobility by analyzing data of novel sources. Findings of this project provide data-driven insights into actionable strategies that serve as solutions to sustainable tourism management in the region.

From mobile data to insights: understanding tourism mobility

Using mobile traces data of tourists in Tuscany, provided by Vodafone Italy and prepared by our project partner CS Research, this project developed data-driven approaches to investigate tourist behaviours of different nationalities and over different seasons. Our analyses uncovered different types of tourist behaviours, their typical trajectories, and groups of locations commonly visited during the same trip. Such insights into tourist behaviours and mobility patterns pave a crucial step for our project partner, TPT, to design strategies aimed at improved segmentation of the tourism markets, and optimal spread of tourists throughout seasons and the municipalities of the region. These strategies ultimately contribute to policies that can be designed, implemented and tested by the local authorities in Tuscany.

Key findings

1. We discovered four types of tourist 'personas' across all nationalities (a persona being the name of a particular group of tourists sharing the same visit patterns): the **city-hoppers**,

[1] http://www.citypopulation.de/Italy-Toscana.html
[2] http://www.irpet.it/wp-content/uploads/2018/05/rapporto-turismo_2018-1.pdf.
[3] http://journals.sagepub.com/doi/pdf/10.5367/te.2014.0415
[4] http://www.irpet.it/the-economy?lang=en

who concentrate their trips in major cities; the **coast-lovers**, who spend most of their time along the coast; the **explorers**, who visit both the coast and inland cities; and the **countrysiders**, who spend most of their trips along the eastern border of Tuscany.

2. For tourists of a given nationality at a given season, further analysis shows distinct preferences within that group. For example, German tourists in summer mostly go to the coast, but some like to travel to inland cities while others do not.
3. When looking more closely at tourist trajectories in Tuscany, we see that they tend to vary from nationality to nationality and from season to season. For example, visiting the most popular cities, in summer, in a short amount of time is most common for non-European visitors.
4. Within the same trajectory cluster, by looking at different examples, we can see how tourists choose different forms of transportation, for example, public transportation versus driving.
5. Being geographically close to each other does not seem to be the only factor that pushes tourists to visit different municipalities during the same trip. For example, major cities with cultural heritage, such as Florence and Pisa, and certain coastal regions and islands are often visited together despite of not bordering with each other.
6. Seasons have a big influence on how much tourists move, for example, during winter months tourists move less than in any other season.

Introduction

Problem statement and project goal

Tuscany, a region in central Italy, is one of the most popular tourist destinations in Europe. Last year, for every local ~10 tourists visited the region. This phenomenon, known as mass tourism, is a challenge across major European cities and travel destinations.

Overcrowded cities typically go through a transformation that includes increased cost of living and, on the long run, a partial loss of cultural authenticity. These challenges are particularly prominent during peak seasons (mostly summer) and concentrates around the most popular tourist destinations.

Despite these challenges, 7% of the region's GDP depends on tourism, which contributes to the creation of jobs and the economic growth. As such, the region seeks improved strategies to balance economic prosperity and social and environmental sustainability.

Against this backdrop, this year's DSSG Tuscany project, "Optimising Tourism in Tuscany" analysed telecom traces of 9.6 million roaming customers[5] from Vodafone Italy, to provide our partner with data-driven insights on tourist behaviours and mobility patterns. These insights help the partner to design and implement actionable strategies aimed at an improved segmentation of the tourism markets, and a better spatial-temporal spread of tourists.

Key questions and rationale for the project

In order to provide insights into the patterns of tourist mobility, our project addressed the following questions:

1. Which types of tourists visit Tuscany?
2. What are the common tourists trajectories?
3. Where else in Tuscany could tourists go to, based on their preferences?

Investigations to answer 1. and 2. shed light on tourist preferences, especially when the analysis is further broken down by nationality (country of origin of the SIM card connecting to the roaming) and the season in which the visits occurred. With these insights, TPT will be able to enhance their current market segmentation and targeted promotion strategies in order to redirect tourist flows.

For the optimisation of tourists spread in time (usually referred to as deseasonalization of tourism) and space (tourists are encouraged to visit less known areas instead of over-crowded sites) the understanding of which tourists group exist (1.) and what trajectories they follow (2.) is, however, not enough. A key next step is to unlock the understanding of which are the locations that the majority of tourists generally visit during the same trip (3.) (co-visits).

[5] Data from foreign SIM cards connecting to the Vodafone roaming network from May 2017 to February 2018 have been anonymized, at one-minute resolution.

With the insight on co-visited locations, local authorities from those municipalities commonly visited in the same trip can coordinate their efforts in designing infrastructure and promotion strategies to best welcome tourists across these municipalities, avoiding concentration in one peak area only. A further action that our partner can take with our insights is the creation of collaborations between local businesses that provide touristic goods and services in under-visited regions with tour operators selling packages to tourists.

The Data

Our data sets

Vodafone telecom traces

The project is primarily based on data provided by Vodafone Italy and prepared by our project partner CS Research. The data contain telecom traces (IP Probe) of over 9.6 million anonymized foreign visitors roaming in Italy who connected to Vodafone Italy network from May 2017 to February 2018. Telecom traces are available at the resolution of every minute when the visitor's location changes (i.e., connected to a different tower), and at every hour mark if the person remains static (i.e., remained connected to the same tower). At each timestamp, we have the latitude and longitude of the cell tower that the visitor's device is connected to. This dataset allows us to follow the movement of the visitor within Italy, particularly across Tuscany.

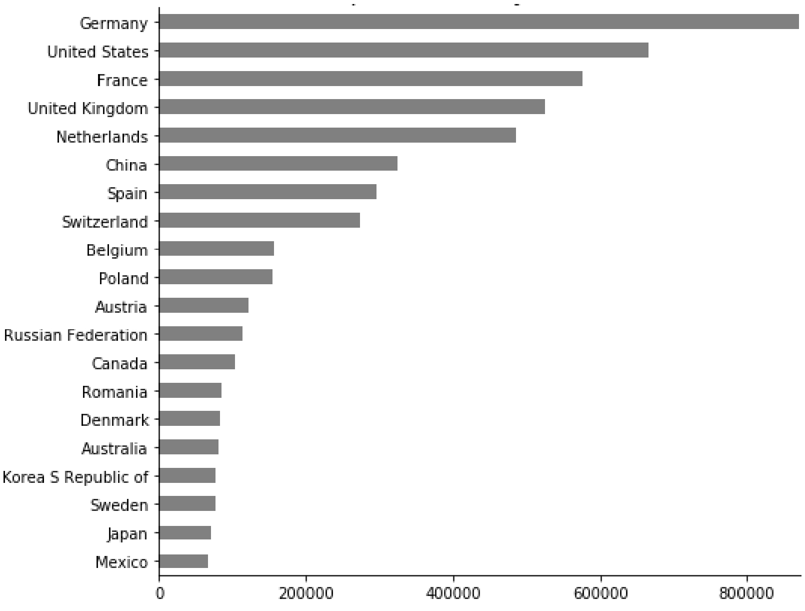
Location characteristics

We supplemented the telecom traces data with information extracted from open sources for each location (cell tower) in our dataset. Such information includes whether the location is near any points of interests, cities, and the geographic landscape (coast, forest, natural parks, etc..).

Descriptive analysis

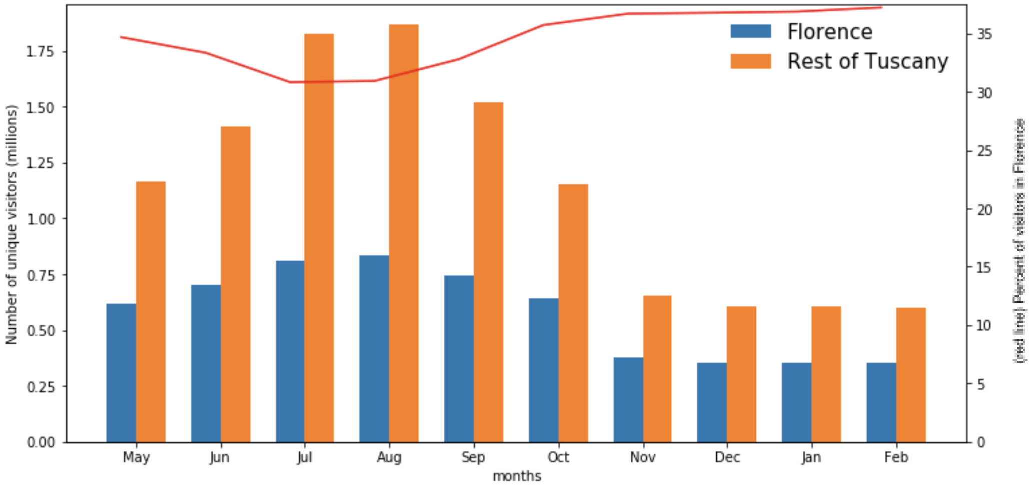
Where do tourists come from and when do they visit Tuscany?

Fig. 1 Number of unique visitors by country.



From Figure 1 on the left, we can see that Germany is the largest tourism market for Tuscany, with around 0.9 million visitors between May 2017 and February 2018. Following Germany, come the US, France, United Kingdom and the Netherlands.

Fig. 2 Number of visitors in Tuscany (bar charts) and percent of visitors in Florence (redline)



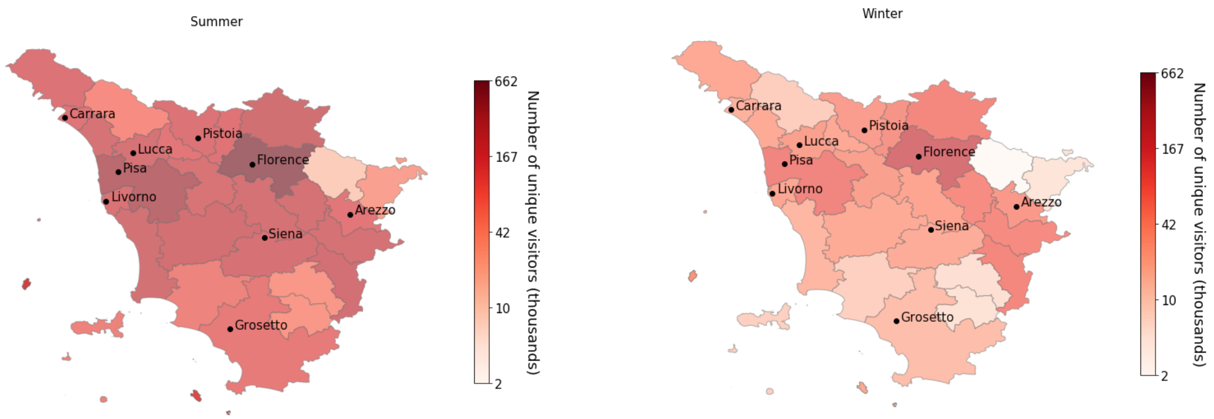
From Figure 2 above we can see how for both Florence and the rest of Tuscany, the number of visitors peak in the summer and drop in the winter.

From a closer observation of the red line in Figure 2 (which is the percent of tourists that visit Tuscany out of all tourists in Tuscany), we can see that in the peak season (summer) Florence attracts around 30% of the total amount of tourists in the region. During all other seasons though, this percentage goes up to 35%, hinting to a higher redistribution of tourism in the region in lower seasons.

Finally, the difference between the peak season and off seasons in Florence is not as prominent as in the rest of Tuscany.

How do tourists spread in the region over seasons?

Fig. 3 number of unique visitors (thousands)

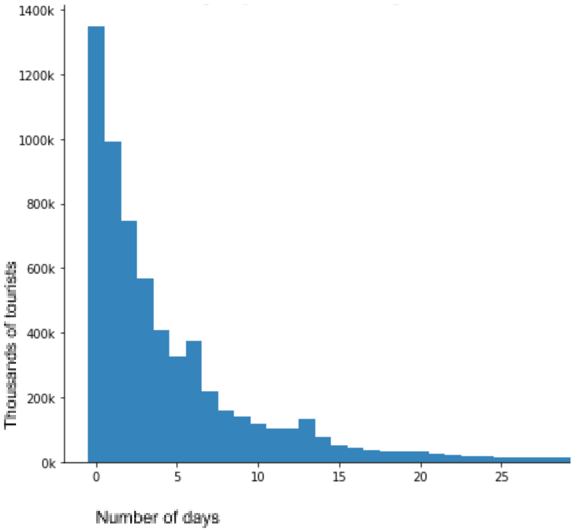


From Figure 3 we can see that tourists spread unevenly in the region. In the summer, tourists are highly concentrated in districts along the coast and surrounding major cities, whereas fewer tourists visit the northeastern and southeastern districts.

Spatial distribution of tourists are less uneven in the winter, though still highly concentrated in districts of famous cities such as Florence and Pisa.

How long do tourists stay for?

Fig. 4 Number of days tourists (in thousands) spend in Tuscany

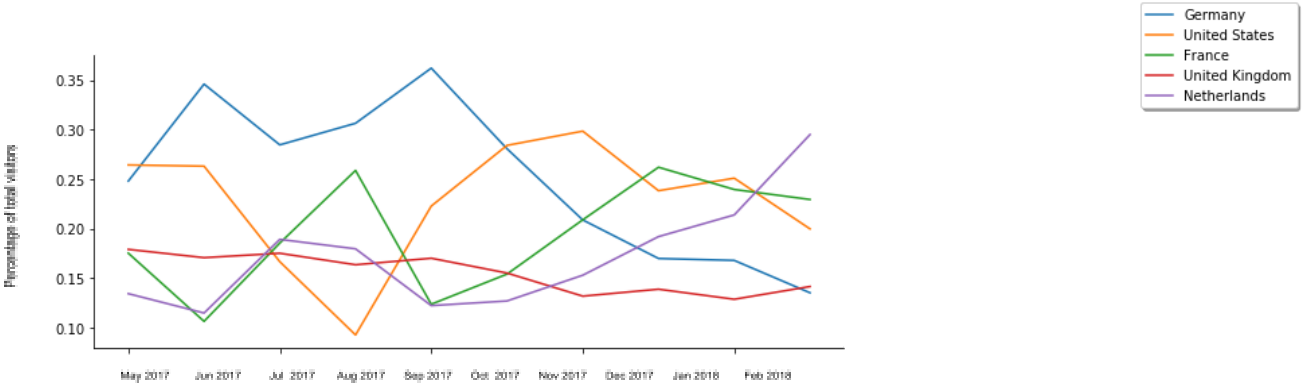


Of all tourists, the percentage of those who spent 1, 1-5 and 6-10 days are:

- 21% spent less than 1 day;
- 45% spent between 1 and 5 days;
- 15% spent between 6 and 10 days.

How do different nationalities spread over the seasons?

Fig. 5 Percentage of visitors from top countries of origins by month



From Figure 5 we see that German tourists account for the largest fractions in summer, peaking in June and September. The ratios of American tourists are the highest during pre and post-summer months while the higher ratio of French is in both winter and summer. The Dutch peak in winter, while the British are evenly distributed throughout the year.

Modeling Approaches

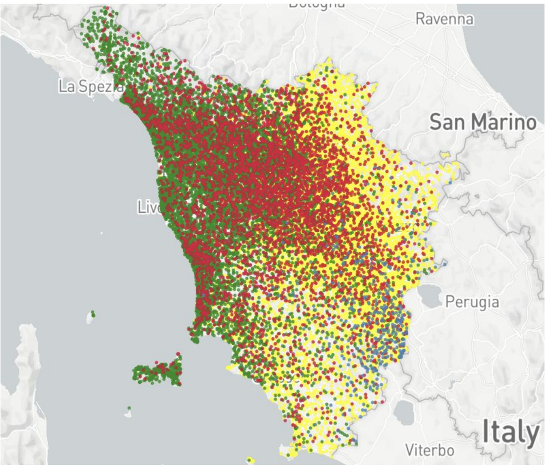
Persona clustering: K-Means

We clustered tourists based on their time-space and behavioral features. From the IP probe data, we extracted tourists' locations and time for their starting and ending points, geospatial distribution of the trip, and total time spent in and outside of Tuscany. By using open source data to extract points of interests, cities, and the geographic landscape for each of the locations, we were able to engineer features that capture tourists' preferences and behaviours.

The list of features we use for personas clustering include:

- Time of arrival in Tuscany and Italy
- Hours spent in Tuscany and rest of Italy
- Total/unique number of locations visited in Tuscany and rest of Italy
- Starting/end point in Tuscany and rest of Italy
- Location spend with most time
- Average and standard deviation of latitude/longitude
- Time spent at different landscapes
- Time spent in major cities
- Number of attractions visited

Fig. 6 Tourists personas based on k-means clustering

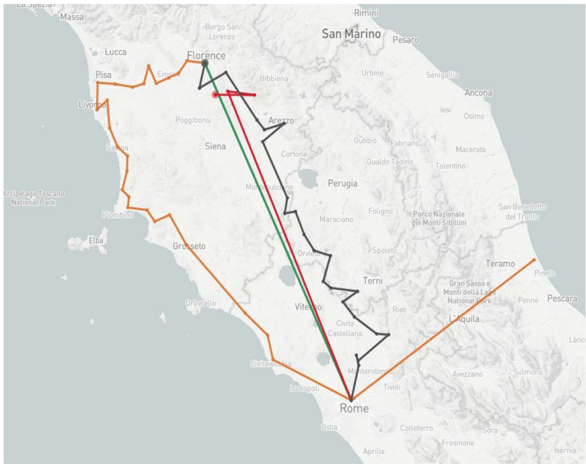


For the summer season, the results reveal four types of tourist behaviors: the city-hoppers (yellow), who concentrate their trips in major cities, and are largely consisted of non-European visitors; the coast-lovers (green), who spend most time along the coast; the explorers (red), who visit both the coast and inland cities and attractions; and the countrysiders (blue), who focus the trips mainly along the eastern border of Tuscany and likely traveled by car.

Trajectory clustering: K-Medoids

Trajectory analysis finds the most frequent paths among tourists. We clustered tourist trajectories by analyzing the sequences of municipalities they visited. Using a Longest Common Subsequence-based metric, we first computed distance between trajectories and then calculated a dissimilarity matrix between individuals. This matrix was then clustered using the k-medoids algorithm to identify different clusters and a representative trajectories for each. This methodology takes into account the order, the time, and the duration of the visits per municipalities in each itinerary.

Fig. 7 Trajectories based on K-Medoids clustering



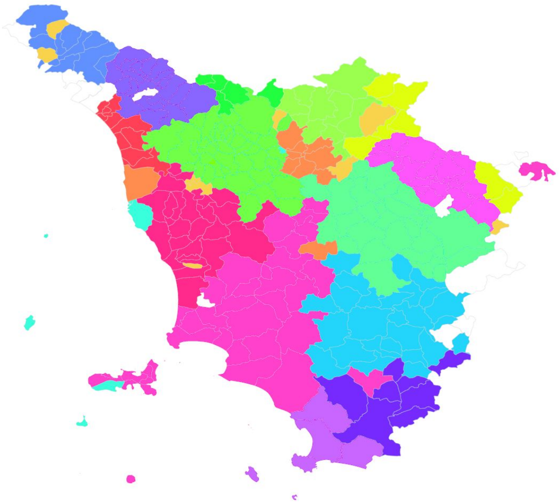
The figure on the left shows the trajectories of 4 French city-hoppers in summer 2017. After spending 1-2 days in Florence they moved on to Rome for another day or two.

Nationality-based clustering allows us to identify the most representative behaviours of tourists belonging to this cluster, which can be unpacked further when looking at random examples. In this case, we see who opted for a train (red, green), who drove through the countryside (grey) and who drove along the coast (orange).

Location clustering: geo2vec

Location clustering shows which municipalities in the region are visited together during the same trip. The location clustering is done with a combination of the geo2vec model and the k-means clustering algorithm. The geo2vec model creates an embedding matrix based on some training data (in our case, ordered sequences of municipalities visited by tourists during a trip that included Tuscany), which contains a vector of a given dimension for each of the municipalities in Italy. These vectors are then clustered in a given number of clusters. Each of these clusters contain municipalities that are commonly visited in the same trip. This approach was inspired by a previous implementation of listing embeddings for a recommendations website, and relies on the implementation of a famous natural language processing algorithm, word2vec.

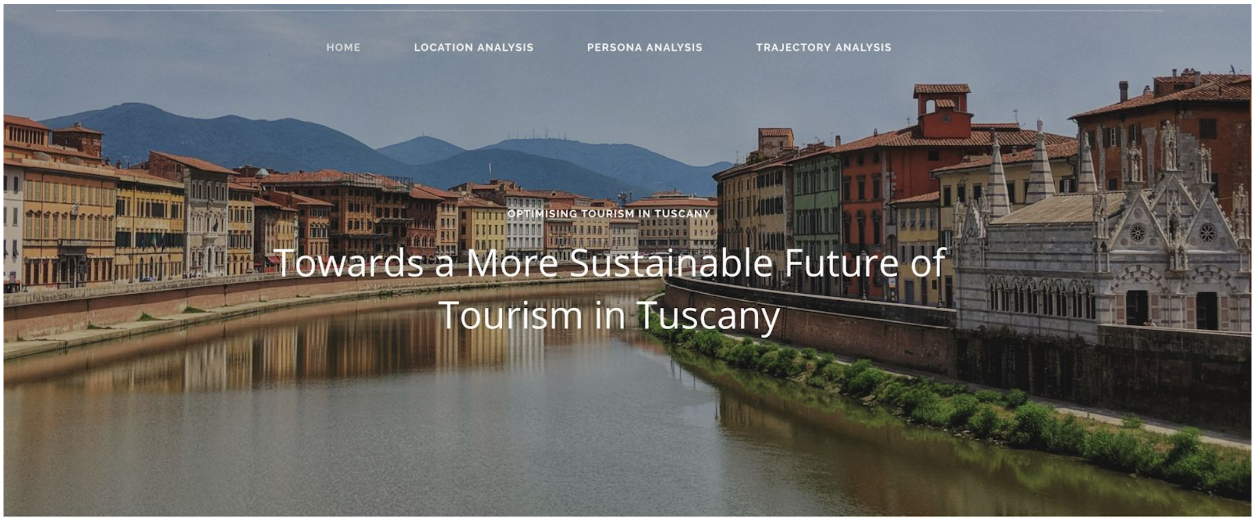
Fig. 8 Clustering of municipalities based on geo2vec embeddings



The figure on the left shows an example of location clusters (1 color, 1 cluster) for the summer months (Jun-Aug 2017). While we see that geography might play a role (all coast areas are aggregated) we are also able to identify other patterns. Livorno and some of the islands, for example, are clustered together (light blue), possibly due to cruises that sail between them. Also, the most popular cities, Florence and Pisa, belong to the same cluster (orange) despite the fact of not being geographically adjacent.

Channels of Action

Detailed results of our clustering analyses can be found at our website: http://dssg-eu.org/tuscany/



OPTIMISING TOURISM IN TUSCANY

Towards a More Sustainable Future of Tourism in Tuscany

Data Science for Social Good Europe 2018

By clustering personas, trajectories, and locations, we were able to portray a comprehensive picture of how tourists move in Tuscany across seasons and locations.

With these results, our partner will now be able to gain meaningful insights into tourists behaviours at the intersections of different dimensions (as examples, French city-hoppers in summer or German coast-lovers in post-summer). With these insights on the actual segments that exist within different combinations of nationalities and seasons in Tuscany, our partner will now be able to optimise their promotion strategies based on real data and to strike a balance between tourists preferences and local offerings of under-visited sites.

In order to find this balance, several channels can be used for interventions. First, by understanding the preferences of different types of tourists in terms of characteristics of the locations they would like to visit, our partner can design more targeted strategies to direct tourists to locations that meet their preference and interests, especially locations that are currently under-visited. This could be best done by directly influencing tourists behaviour (a new field in social science research called "digital nudging"[6]). Our partner currently co-manages Tuscany's official tourism promotion website, www.visittuscany.com. With our analysis, the content of this website could be tailored to tourists' preferences, with an improved recommender system. Learning the different tourist preferences, our partner can also carry out customized user experiments on different groups of tourists to design targeted interventions that effectively influence tourist behaviour. For instance, A/B testing of targeted promotion strategies can be implemented to examine the effectiveness on certain groups.

[6] Weinmann, Schneider & vom Broke, Digital Nudging, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2708250

In addition, outlets of promotions such as Google ads can be more effective when customized to cater to user preference.

To improve its policies, our partner could also combine insights of tourists preferences and their trajectories to diagnose why some locations might be under-visited and which types of tourists such locations could potentially attract, and design strategies accordingly. For example, under-visited locations on routes between popular locations could potentially divert heavy tourist flows if tour operators are able to enhance service offerings that suit tourist interests in such locations.

Lastly, our partner could intervene by directly working with the 28 "ambiti territoriali"[7] to improve the service delivery and the coordination of transport and welcoming structures within those municipalities which are co-visited during the same trip. As a matter of fact, by uncovering locations commonly visited together by tourists, local administrators of relevant locations can potentially coordinate actions in terms of tourism product design and service delivery. As such, synergy can be created in tourism management and development between these locations. Meanwhile, the clustering of locations also informs our partner which under-visited locations share similarities to other more frequently visited locations. Thereafter, our partner can design strategies to direct tourist flows from over-visited to under-visited locations to balance tourist flows.

[7] http://www.regione.toscana.it/entilocaliassociati/unioni-di-comuni/ambiti-territoriali