Please cite as: Han, Q., Abreu Novais, M. and Zejnilovic, L. (2021), "Toward travel pattern aware tourism region planning: a big data approach", *International Journal of Contemporary Hospitality Management*, Vol. 33 No. 6, pp. 2157-2175. https://doi.org/10.1108/IJCHM-07-2020-0673

Toward travel pattern aware tourism region planning: A big data approach

Qiwei Han*Leid Zejnilović† Margarida Abreu Novais ‡

Purpose The purpose of this paper is to propose and demonstrate how Tourism2vec, an adaptation of a natural language processing technique *Word2vec*, can serve as a tool to investigate tourism spatio-temporal behavior and quantifying tourism dynamics.

Design/Methodology/Approach Tourism2vec, the proposed destination-tourist embedding model that learns from tourist spatio-temporal behavior is introduced, assessed, and applied. Mobile positioning data from international tourists visiting Tuscany are used to construct travel itineraries, which are subsequently analyzed by applying the proposed algorithm. Locations and tourist types are then clustered according to travel patterns.

Findings Municipalities that are similar in terms of their scores of their neural embeddings tend to have a greater number of attractions than those geographically close. Moreover, clusters of municipalities obtained from the K-means algorithm do not entirely align with the provincial administrative segmentation.

Research limitations/implications

Mapping locations that are typically visited together and discerning patterns of spatio-temporal behavior is of great significance for tourism region planning and management. The major limitations of this paper are related to the type of data used and the subjective interpretation inherent to unsupervised learning.

Originality/value Through the proposed Tourism2Vec, this paper contributes to the discussion that promotes conceptualizing tourism regions in a way that includes tourists' actions. It also contributes to the existing knowledge on tourism spatio-temporal behavior by offering a macro perspective on a topic that has traditionally been investigated through small scale studies.

Keywords— tourism region planning, Tourism2vec, tourism spatio-temporal behavior, travel patterns, mobile positioning data, Big Data

Paper type— Research paper

* *qiwei.han@novasbe.pt*
Nova School of Business and Economics, Universidade Nova de Lisboa
† *leid.zejnilovic@novasbe.pt*
Nova School of Business and Economics, Universidade Nova de Lisboa
‡ *m.abreunovais@griffith.edu.au*
Department of Tourism, Sport and Hotel Management, Griffith University

1. Introduction

Tourism regions are destinations designated by the respective governmental tourism authorities having common natural and historical characteristics of tourist activity (Ritchie and Zins, 1978; Czernek, 2013). Historically, these regions often emerged as cultural or geographic areas with prominent tourism resources, established to attract more tourists by offering them a discrete set of tourism experiences, as well as to attract non-local investors in tourism. Tourism authorities increasingly take a proactive role in the development and management of tourism regions to bolster growth and respond to dynamic markets (Pearce, 1989). Given the growing complexity of such a task, systems approach is becoming a requisite to ensure the sustainability of economic growth as well as the social development and preservation of the environment (Fyall and Garrod, 2019; Baggio, 2020). This approach is also needed in order to enable the cooperation among stakeholders within regions and to promote tourists as co-creators in these regions (Piriou, 2019; Jovicic, 2019). For instance, changes in tourists' preferences and behaviours may significantly change the types and volumes of hospitality services needed within a region. Accordingly, by carving up the territory into contiguous tourism regions with distinct identities and a manageable range of attractions, governmental authorities can maximize its tourism potential.

Given that tourism regions can sometimes be managed more for administrative convenience, these do not always provide a coherent tourist experience to visitors (March and Wilkinson, 2009; Tosun and Jenkins, 1996). Furthermore, existing conceptualisations and discussions of tourism regions highlight a number of issues. First, the contentious nature of tourism authorities which often have competing interests at different spatial scales may complicate the development and management of tourism regions (Dredge and Jenkins, 2003). As noted by March and Wilkinson (2009), when a sub-region of a designated tourism region offers a differentiated tourist experience from the overall planning strategy of the region as a whole, it may become relatively neglected. Thus, partnerships in planning, marketing and resource management for tourism regions must decentralize practices to coordinate and integrate the varied interests among tourism authorities (Tosun and Jenkins, 1996). Second, the conceptualization of the tourism region has evolved from geographical terms to go beyond spatial configuration, which involves various (economic, social, or environmental) dimensions of tourism (Baidal, 2004; Baggio, 2020). Instead, a new regional perspective that focuses on the tourist function of different regions is proposed to assimilate the multi-dimensional meanings of territory (Zhong and Zhang, 2012). In line with such views is the idea that tourism destinations with diverse "actors from different geographical locations and with distinct typologies are in a better position to achieve a higher innovation performance" (Brandao *et al.*, 2018, p.1). Third, as tourism regions reflect social and cultural aspects of tourism destinations, tourists are also considered key players in the co-production of the tourist experience (Saarinen, 2004; March and Wilkinson, 2009). As such, the tourism region should be co-constructed with all the stakeholders in tourism, including tourists(Piriou, 2019). Last, the tourism region is seen as a dynamic, ongoing socially constructed spatial unit, and so the tourism region should be expected to evolve and change over time (Saarinen, 2004).

Overall, it is clear that the spatial organization of tourism regions needs to reflect its evolving and socially constructed nature as well as the interactions with other socio-spatial units (Saarinen, 2004). Accordingly, a tourism region may join the tourist activities into a multi-level network of tourism destinations to reflect its function as the social system (March and Wilkinson, 2009). In addition, Piriou

(2019) proposes that the formation of networks of tourism destinations should account for tourist mobility given that activities are practiced by tourists. The author further argues that a territory cannot be considered a tourism region if it lacks tourist mobility that links the various destinations. Therefore, investigating tourist spatio-temporal behavior and understanding the travel patterns become an imperative element in the design of a travel pattern aware tourism region.

The overall aim of this paper is to propose a travel pattern aware tourism region planning to characterize the socio-spatial meanings of tourism destinations by bringing existing knowledge from natural language processing, a sub-field of computer science and artificial intelligence, to the investigation of tourism. It introduces and applies for this purpose Tourism2vec, an adaptation of a widely used natural language processing technique - Word2vec - for the investigation of tourism spatio-temporal behavior. Also, an application of the proposed Tourism2vec is demonstrated on signaling data, used together with geospatial data on the geographical boundaries of locations as well as tourist attractions within a destination.

2. Related Work

2.1. *Conceptualization of tourism region*

Tourism research has long been interested in tourism destinations, their development and management in order to attract more tourists as well as investment while maintaining the regions' economic, social, and environmental sustainability (see *e.g.,* (Ritchie and Zins, 1978; Rigall-I-Torrent and Fluvià, 2011; Baggio, 2020)). Early studies look at this issue mostly from a human geography perspective to consider the tourist region as a homogeneous and contiguous area comprising a group of places with significant characteristics of tourist activity (Britton, 1991; Gordon and Goodall, 2000; Terkenli, 2002). Although "tourism regions" and "tourism destination" are often used interchangeably, Tosun and Jenkins (1996) argued that these two concepts are not necessarily identical. The tourism destination is "a geographical area containing a critical mass of development" with a discrete set of attractions (Smith, 1989), while tourist region is largely defined in geographical terms regarding the spatial combination of attractions as well as their related facilities (Pearce, 1989). The spatial structures at a varying range of scales have been investigated to understand regional tourism development from a technical and static perspective (Saarinen, 2004). For example, Jansen-Verbeke (1995) defined the tourism region at the economic or cultural aggregation levels to study the European tourism market instead of the national statistics.

Saarinen (1998) proposed the discourse of region theory in which a region's social and geographical meanings are combined with traditional territorial representations of the region. The discourse of the region reveals the conceptualization of the tourism region as a cultural and social construction. Accordingly, the tourism region is defined by social and cultural aspects instead of by the more traditional physical or administrative delimitations, which are not necessarily recognized by tourists or potential tourists (Saarinen, 2004). This suggests that the planning of the tourism region should be adjusted to encompass various dimensions of tourism using the decentralized regionalization approach (Baidal, 2004; Tosun and Jenkins, 1996).

As regional tourism development increasingly involves more stakeholders, the conceptualization of tourism regions also entails the notion of tourists as active players in the co-production of the tourist experience (March and Wilkinson, 2009). In fact, tourism region is a concept that can exist simply in the mind of the tourists (Gretzel, 2018). It may be formed as networks of places designed by tourists according to their mobility (Piriou, 2019). In other words, tourist spatio-temporal behavior influenced by voluntary actions may shape the network of destinations through tourist flows (Dredge and Jenkins, 2003). The information about where and how tourists move may offer an opportunity to critically evaluate the understanding of tourism regions, recognizing their dynamic and complex natures (Baggio and Sainaghi, 2011). Moreover, Piriou (2019, Ch.10) showed that networks of places built on tourist mobility provide information on their socio-spatial relationships that would constitute a tourism region.

2.2. *Tracking tourism behavior*

Tourists can be segmented using a variety of approaches. In the case of movement studies, tourism scholars have examined how country of origin (Tiru *et al.*, 2010), length of stay, number of visits (East *et al.*, 2017; Lau and McKercher, 2006; McKercher and Lau, 2009), party size (East *et al.*, 2017; Zhao *et al.*, 2018), and the geographical locations of visits (Versichele *et al.*, 2014) affect the way tourists behave in space and time. Overall, the approaches taken depend on the available data (Md Khairi *et al.*, 2018), but the possibilities are also numerous for profiling and segmenting tourists (McKercher *et al.*, 2012). A major weakness with the existing studies investigating spatio-temporal behavior of tourists is related to the sample in terms of time frame of tracked movements, geo-coordinates of the geographical locations, or number of individuals involved in the study.

In addition to the size limitation, the vast majority of published studies have tracked tourists' movements for a limited time only (Versichele *et al.*, 2014). This is a result of the dominance of data types such as surveys and data derived from GPS and Bluetooth, which typically enable the collection of data for mere hours or a limited number of days. More recently, user-generated content data and mobile phone data have allowed expanding data sets beyond the typical time frames to include extended periods of up to years. Notwithstanding some initial efforts, as noted recently (Su *et al.*, 2020), little research has focused on the long-term perspective of spatio-temporal behavior. While short-term tourist-flow studies are useful in identifying potential problems related to crowding and conflicts, a long-term approach to understanding tourist movement can reveal trends of movement, and even predict future scenarios (Su *et al.*, 2020). Ultimately, this will offer insights that can lead to action that is in line with more sustainable development of destinations.

2.3. *Neural embedding model*

The notion of word embedding was firstly introduced by Bengio *et al.* (2003) in an attempt to overcome the issue of dimensionality of word representations in Natural Language Processing (NLP). Bengio *et al.* (2003) proposed representing words as vectors in a real-valued and low dimensional space. More specifically, word embedding takes a large unannotated corpus of words as the input and associates each word in the vocabulary with a *word vector* as the output based on neural networks that involve multiple layers, such as input/output layer and several hidden layers. This allows researchers to learn and compare the meaning of words by simply calculating the distance between word vectors.

The seminal work by Mikolov *et al.* (2013a) introduced *Word2vec*, a three-layer neural network model that produces word embeddings to capture syntactic and semantic properties of words, while vastly improving the training speed through several approximation techniques. Given the fact that it can be trained on large-scale data within reasonable time constraints, Word2vec quickly gained traction and its use was extended to many other domains to generate vector representation for different types of sequence data. A variety of adaptations of Word2vec for studying human mobility patterns for different purposes, including understanding traffic flows (Zhu *et al.*, 2019), enhancing the understanding of individuals' preferences (Chen *et al.*, 2019; Feng *et al.*, 2017; Pang and Zhang, 2017), and recommendation of location or point of interest (Chang *et al.*, 2018; Liu *et al.*, 2016; Wang *et al.*, 2018; Zhao *et al.*, 2017).

3. Tourism2vec Model for Learning Location and Tourist Type Representation

This section details the model - Tourism2vec - which adapts the neural embedding model that learns tourists' travel patterns based on their movements. Figure 1 overviews the pipeline of learning, applying and evaluating the latent representation of locations and tourist types from travel itineraries using Tourism2vec. First, each travel itinerary is augmented with its tourist context, including the season of the visit and the nationality of the tourist. This tourist context is considered to be indicative of travel patterns and is thus added to the itineraries as the *global context* (Grbovic and Cheng, 2018). Then, Tourism2vec is applied to tourists' travel itineraries at the municipality level to generate embeddings for locations and tourist types in the latent space. Last, these embeddings are leveraged and the similarities between the municipalities and tourist types are explored with the clustering technique to categorize the municipalities into clusters in terms of tourists' common travel preferences.

In particular, Tourism2vec was adopted to learn a d-dimensional continuous vector v_{l_i} for each unique location l_i, such that locations co-visited by tourists in their travel itineraries would be embedded closely together in the latent space. Formally, given a set of S itineraries obtained from N tourists, each itinerary is defined as a sequence $s = (l_1, \ldots, l_M) \in S$ of M locations visited by the tourist. The context of location is denoted l_i as $C(l_i) = \{l_{i-j}, l_{i-j+1}, \ldots, l_{i-1}, l_{i+1}, \ldots, l_{i+j-1}, l_{i+j}\}$, which includes the surrounding locations in the same itinerary, where the size of the

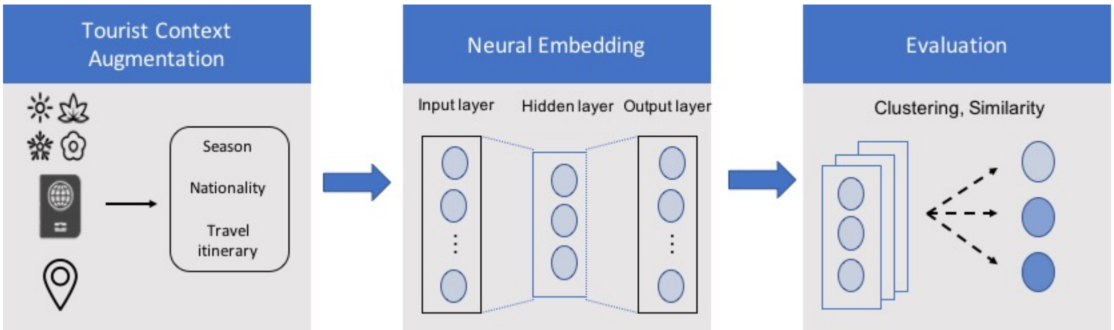


Fig. 1: **Tourism2vec Pipeline Overview**

left and right side of the context window is j, respectively. Then, the probability for observing a location $l_c \in C(l_i)$ within the context window given location l_i is computed using the softmax function:

$$P(l_c|l_i) = \frac{\exp(v_{l_i}^T v'_{l_c})}{\sum_{l-j}^{l+j} \exp(v_{l_i}^T v'_l)} \tag{1}$$

where v_l and v'_l are the input and output vector representations of location l. Figure 2 is a graphic representation of how the context window was shifted from the origin location to the destination location to learn the neural embedding of location l_i. This is expected to help capture the spatio-temporal correlation between these locations. The neural embedding model aims to maximize the average log probability over the entire set S of travel itineraries:

$$L = \sum_{s \in S} \sum_{l_c \in C(l_i)} \log P(l_c|l_i) \tag{2}$$

Given the time-consuming aspect of directly optimizing the objective function L, the negative sampling technique, which was introduced by Mikolov *et al.* (2013b), was used to improve the optimization efficiency. More specifically, for each location l_i, two sets of location pairs were generated, D_p and D_n, respectively, where D_p

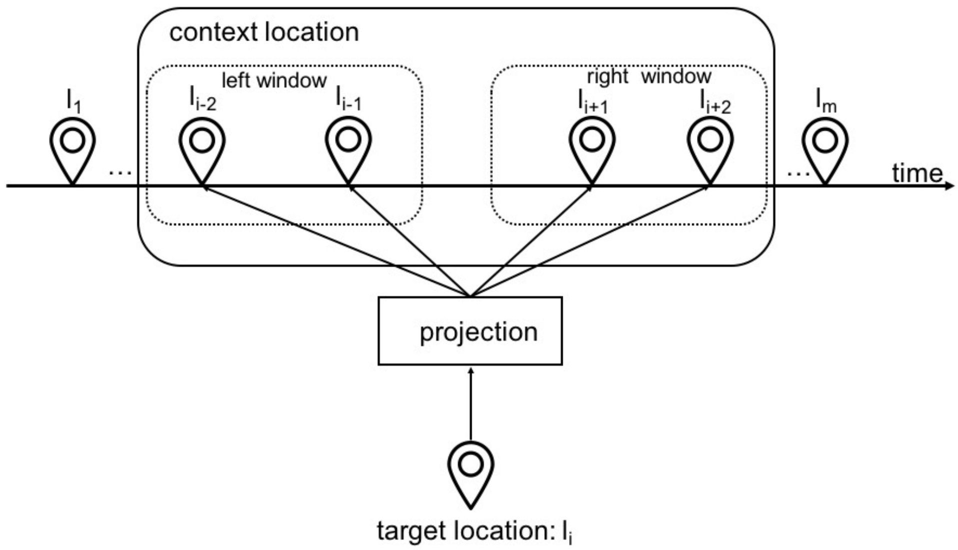


Fig. 2: **Sliding context window for learning neural embeddings of locations using neighboring location in the same travel itinerary (left/right window size is set to 2 for illustrations)**

contains *positive pairs* of location l_i and one from its context $l_c^+ \in C(l_i)$ (*i.e.*, locations visited by the same tourist that occurred before and after visit location l_i within the context window), and D_n contains *negative pairs* of location l_i and K random sampled locations $l_c^- \notin C(l_i)$ that are not within the context window. Then negative sampling approximates the softmax by transforming the objective function into the sigmoid function and can be solved by stochastic gradient descent:

$$\underset{\theta}{\operatorname{argmax}} \sum_{(l_i, l_c^+) \in D_p} \log \sigma(v_{l_c^+}^T v_{l_i}) + \sum_{(l_i, l_c^-) \in D_n} \log(1 - \sigma(v_{l_c^-}^T v_{l_i})) \tag{3}$$

where $\sigma(x) = (1/1 + \exp(-x))$ and the Tourism2vec model learns parameters θ to represent vectors $v_{l_c^+}$, $v_{l_c^-}$ and v_{l_i} for location l_c^+, l_c^- and l_i, respectively.

After generating the vector representations for all locations learned from tourists' travel itineraries, these vectors can be leveraged for various downstream tasks.

4. The application of Tourism2Vec: Investigating Tourist's Travel Patterns in Tuscany

4.1. *Context, data description, and preprocessing*

The context of this study is Tuscany, a region of central Italy with a population of 3.72 million and an area of 22,985 km^2. With over 44 million yearly overnight stays in official accommodation, Tuscany is Italy's second top tourist region and one of the top 20 most visited regions within the European Union (Eurostat, n.d.). As Tuscany offers a wide variety of tourism resources and experiences including cultural heritage (Popp, 2012), natural landscape (Randelli *et al.*, 2014; Ferrari *et al.*, 2016), dining (Bertella, 2011), the local tourism authority faces challenges in effectively promoting the entire region. Thus, creating and developing tourism regions within Tuscany becomes necessary and useful to attract different tourist types.

Three types of data were used for this study. The first includes a mobile signaling dataset provided by a European mobile-phone carrier (referred to hereinafter as EURMO) including anonymized logs of signaling traces of mobile phones with foreign SIM cards connected to EURMO's network infrastructure in Tuscany between May 2017 and February 2018. Given the purpose of this study, it was decided to aggregate tourists' travel itineraries at the municipality level, which allows an examination of tourists' travel patterns consistent with administrative and statistical practices.

The second includes the ESRI shapefile for the Tuscany region from DIVAGIS (Hijmans *et al.*, n.d.). This was used to identify the boundaries of administrative subdivisions for all municipalities which were then used to replace the geo-coordinates of cell towers by the ID for the municipality in which these cell towers are located. The third and last data type includes a list of 792 tourist attractions in the region including their names, category and geo-coordinates, which was obtained by crawling Tuscany's official tourism website (https://www.visittuscany.com/).

Several preprocessing steps were performed to ensure the veracity and validity of data for this study. Firstly, four types of visitors were removed: short-term visitors (duration of stay in the region of less than 6 hours), long-term visitors (duration of stay in the region of more than one month) static visitors, and cross-over visitors. Secondly, for computational and interpretation reasons, the data were further reduced by including only those tourists from the top 10 markets: Germany (DE), United States (US), France (FR), United Kingdom (UK), Netherlands (NL), China (CN), Spain (ES), Switzerland (CH), Belgium (BE), and Poland (PL). After the data cleaning process, nearly 4 million tourists remained, which represent 70% of all international tourists.

Seasonality and nationality of tourists were also taken into consideration in this study. The data were divided into four seasons: pre-summer (May), summer (June-August), fall (September-October), and winter (November-February) and the country of origin of individuals' SIM cards was used as a proxy for tourists' nationality.

4.2. *Construction of travel itineraries*

Travel itineraries were constructed by organizing visited locations in temporal order. More specifically, one location sequence for each tourist was created, in which each location is a municipality that has been visited by the tourist for at least one hour. If the tourist travels within the same municipality, it would be recorded only once in the sequence. However, when a tourist travels to another municipality and then back to the previous one, the sequence would have three records to capture this flow.

Tourist types (from the combination of the nationality and season) were augmented to the travel itineraries as the metadata. This enabled the Tourism2vec model to learn the latent representation of municipalities and tourist types in the same latent space. The constructed travel itineraries were drawn from a "vocabulary" with size of 273 municipalities and 40 tourist types (the combination of 10 nationalities and 4 seasons) resembling a corpus of documents, in which each "document" represents the itinerary for one tourist containing a sequence of "words" that represent the municipalities the tourist has visited, augmented by the specific tourist type. In total, 3.58 million travel itineraries were generated, each representing a sequence of visited municipalities pertaining to each tourist with the known type of nationality and season.

4.3. *Model configuration*

Given its context distinct from NLP tasks, the Tourism2vec model is also differentiated from Word2Vec in terms of the configuration details. Goldberg and Levy (2014) argued that much of the improved performance from the neural embedding model can be attributed to the fine-tuning of algorithmic elements (known as hyper-parameters) inherent in the model. Thus hyper-parameter configurations were customized by replacing the default values for NLP tasks with the ones suitable to our context as follows.

The negative sampling distribution controls the degree to which negative pairs of locations (l_i, l_c^-) are sampled proportionally to their frequency from the smoothed unigram distribution. Mikolov *et al.* (2013b) originally set the value as 0.75 for NLP tasks and instead the value -0.5 was chosen in order to allow the negative sampling to be more likely to sample unpopular locations as negative pairs (Caselles-Dupré *et al.*, 2018).

The embedding size defines the dimensionality of the neural embedding and directly impacts its quality. The commonly used dimensionality is set to 300, following the original work in Mikolov *et al.* (2013a). However, given that the vocabulary size in our study is merely 313, the neural embedding size as 32 was chosen in order to achieve expressive embeddings while controlling for the model complexity.

Following Caselles-Dupré *et al.* (2018), a hyperparameter search strategy for the number of epochs (0 to 200 with step of +10) is adopted to fine-tune the Toursim2vec model for the task of predicting the next visited location. More specifically, travel itineraries were split into three subsets (training, validation, and test sets). For each travel itinerary, the model initially fits the first $(n-1)$ locations in the training set. The pairs of locations randomly drawn from $((n-1)$-th, n-th) element in the validation set is evaluated to identify the number of epochs that yield the best results. Last, final results were reported by performing predictions on pairs of locations drawn similarly in the test set. Then, the top N closest locations to the last location in the training itinerary are generated using a nearest-neighbor search method. In practice, minibatch of 100k travel itineraries were randomly sampled for training and 10k validation/testing pairs were randomly sampled without replacement for 10 times and evaluated using Hit Ratio at N (HR@N), as the ratio of n-th location in validation/test pairs matches any location in the list of N predicted locations.

Figure 3 shows the average Hit Ratio obtained from the Tourism2Vec model trained with varying number of epochs for recommending top 3, 5 and 10 locations as the predicted next visited location. The model performs best in all three metrics when the number of epochs is 120, and thus the model is retrained using all travel

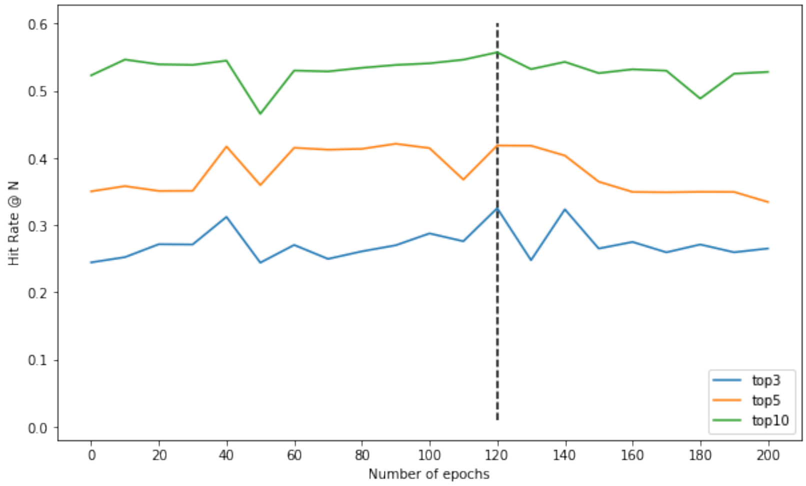


Fig. 3: **Identify the optimal number of epochs (120) through hyperparameter search strategy**

itineraries with 120 epochs.

4.4. *Clustering analysis*

A clustering technique using the neural embeddings of municipalities learned from travel itineraries was performed in order to divide them into homogeneous subgroups whereby municipalities within the same group are more similar than those in other groups. In particular, the widely used K-means algorithm was performed on the embeddings of municipalities to produce k different clusters of municipalities by minimizing the intra-cluster variance according to the Euclidean distance.

The K-means algorithm typically requires choosing K, the number of clusters *a priori*. However, given the unsupervised nature of the task, finding optimal K without ground truth labels often relies on either *ad hoc* visual inspection or strong modelling assumptions. Instead, v-fold cross-validation, an automatic approach determining the "best" number of clusters is use (Nisbet *et al.*, 2009). More specifically, the algorithm divides neural embeddings of all municipalities into v folds,

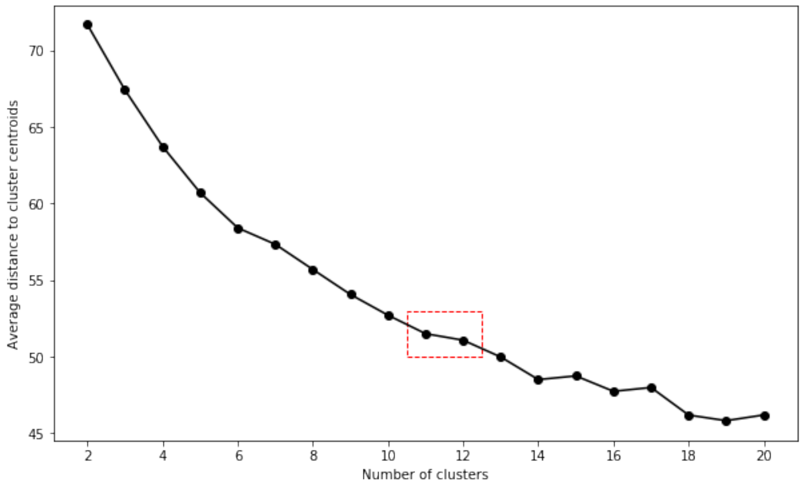


Fig. 4: **Identify the optimal number of clusters (11) through v-fold cross validation**

with equal-sized randomly drawn subsamples. Then K-means algorithm is applied to training samples containing municipalities in $v-1$ folds, and predict test samples of those in the v-th fold to calculate the within-cluster distances. The analysis is iterated with a varying number of clusters until a (K + 1)-clusters solution that computes an average distance measure reaches almost equal performance to that obtained for K clusters (1% threshold is set for differences in the distance for K cluster solution and K+1 cluster solution).

Figure 4 depicts the changes in average distances of municipalities in testing samples to cluster centroids with the number of clusters set by the K-means algorithm from 2 to 20. When the number of clusters is set to 11, the change in average distance of locations in the testing sample compared to the same metric when the number of clusters is set to 12 is 0.8%. Finally, K-means clustering is carried out with neural embeddings to all municipalities using 11 clusters for the following analysis.

4.5. *Similarity analysis*

The Tourism2vec model was applied on the travel itineraries and the neural embeddings with the size of (313, 32) were obtained, in which each municipality and tourist type is represented as a vector of 32 real values. Using the neural embeddings of municipalities, cosine similarities (which may be indicative of relationships between them) were calculated. Table 1 lists the top 5 most similar municipalities for a set of municipalities of interest, including those in major cities, coastal areas, countrysides, and islands. Meanwhile, the geographic distances for these municipalities were calculated and the ranking of proximity is reported in the parentheses. Noticeably, in general, municipalities that are located closely in terms of their neural embeddings similarity score also tend to be geographically close to each other. For example, 61.3% of the top 5 most similar municipalities are also among the top 5 geographically closed ones. This implies that the Tourism2vec model understands the spatial correlations between the municipalities from tourists' travel itineraries because tourists are more likely to visit the nearby municipalities on the same trip for convenience. Interestingly, it was found that for some municipalities, geographically disparate municipalities are located closely in the latent space. One notable example is the municipality of Livorno, which has Capraia Isola as the most similar municipality in the latent space, although they are geographically disparate. This municipality turns out to be an island that is only accessible only through the Livorno port. This finding suggests that the Tourism2vec model can uncover the inherent similarity between municipalities to create travel patterns aware tourism regions going beyond the geographic proximity. These results are in line with recent research using similar approaches. For instance, Zhao *et al.* (2018) generally observed that tourists choose future attractions relatively close to the those previously visited, but also found exceptions for particular attractions. In a different yet related study, Su *et al.* (2020) found that tourists had a more concentrated movement when compared with that of local residents.

To further understand how the Tourism2vec model captures tourism activities in the neural embeddings of municipalities, Figure 5 shows the average number of attractions from all municipalities, in comparison with the average number of attractions from those listed as the top 3(5) most similar municipalities in terms of similarity score of their neural embeddings in the latent space, as well as from those located among the top 3(5) geographically closest municipalities. Interestingly, municipalities identified as the top 3(5) most similar in the latent space were found to have on average 48%(15%) more attractions than those top 3(5) municipalities that are geographically close, and that the difference is statistically significant ($p < 0.001$). This supports the usefulness of learning neural embeddings of municipalities for tourism region planning, because the Tourism2vec model tends to place those with more attractions to be located close to most municipalities in the latent space.

4.6. *Tourism region construction*

Figure 6a depicts 11 clusters of municipalities based on neural embeddings, in comparison with the 10 administrative provinces that these municipalities belong to, as shown in Figure 6b. It was found that municipalities that are geographically close tend to be clustered together, which implies that neural embeddings encode geographical similarity. However, clusters obtained from the K-means algorithm do not entirely align with the provincial segmentation. Instead, these clusters seem to suggest that similarities of municipalities are due to the similarities in tourist preferences. For example, major cities such as Florence, Siena, Pisa, and Livorno (which are known for tourism attractions) belong to the same cluster, even though these cities belong to several different provinces, implying that these cities may be co-visited by many tourists in their itineraries. Moreover, although the Tuscany islands are under the same jurisdiction, their visiting patterns are different according to the clustering results. The islands closest to the Livorno port, Capraia and Gorgona, are clustered with Livorno port, while Elba and other southern islands are in a separate cluster, indicating a distinctive type of tourists who visit them. As such, the tourism authority may consider coordinating the creation of tourism region that facilitates the tourist flows between these cities across the administrative boundaries by which the Tuscany Tourist Board used to manage the marketing of tourism activities.

5. Discussion and Conclusions

5.1. *Conclusions*

There is an ongoing shift in the understanding and conceptualization of tourism regions, with calls for regional delineations that include tourism actions and conceptualizing of regions as networks of places according to tourists' spatio-temporal behaviour (Piriou, 2019, Ch.1). Also, there is increasing awareness of the importance of machine learning and big data for informing tourism and hospitality management (Li *et al.*, 2018; Mariani *et al.*, 2018). The argument advanced in this paper is that the development of regional networks and collaborations and dynamic, spatio-temporal aware tourism region planning can be informed by a big data approach. The example of the application of this approach to the Tuscany region in region demonstrated that the approach is possible and can also reveal non-obvious similarities among non-neighbouring geographical regions.

5.2. *Theoretical implications*

This paper offers several theoretical contributions to the tourism and hospitality literature, especially to the strands related to data-driven and spatio-temporal behavior aware planing in tourism. The first contribution is the advancement of tourism regions planning as an implicit co-creation process with tourists, informed by big data approaches. Most studies of tourism regions start with the qualities of places and the interpretations that tourists associate with them to make choices (Piriou, 2019). The proposed method reverses the process; similarities of places are derived based on the observed tourist visits in relation to the tourist types. Hence, the tourists participate in the region planning process by making their choices what to visit, in which order, and for how long. As such, neural embeddings for municipalities that are directly derived from tourist travel patterns may reflect potential for the development and management of these municipalities as the tourism region.

The second contribution is methodological. The paper proposes a machine learning approach, Tourism2vec, as a tool for tourism region planning informed by tourist spatio-temporal behavior. Temporal sequences of visited locations in tourists' travel itineraries are augmented with the contextual information of tourist category defined as a combination of the nationality and traveling season (*e.g.,* US-Summer). That information from millions of tourists is input into a neural network that is trained to represent locations and tourist types as vectors in a latent space. It enables the use of a range of tools to evaluate the quality of these representations and provides insights for tourist travel preferences toward the destinations given different tourist types. The method offers inference of the relationship between locations beyond geography, considering tourists' actions. The example in which the methods is applied to the itinerary of tourists in Tuscany shows that the municipalities that are most similar in the latent space have on average more attractions than those that are geographically close. Overall, the paper demonstrates how neural embeddings can be used and post-processed, applying similarity analysis and clustering methods to further extend the analysis and enrich the interpretation.

This paper also contributes to the burgeoning demonstration of the potential of big data approaches in studying spatio-temporal behavior of tourists (Caldeira and Kastenholz, 2020; Reif and Schmücker, 2020; Shi *et al.*, 2020) and broader issues in tourism and hospitality (Pillai and Sivathanu, 2020; Ravi *et al.*, 2019). In particular, it shows the strengths of using mobile data to better understand how tourists move within and across destinations. Given the widespread use of mobile phones for travelling purposes (Jamal and Habib, 2019), data gathered from these devices have great potential in capturing the behaviour of a wide range of travellers. Furthermore, given its granularity, volume and frequency, mobile data are suitable to uncover insightful patterns of movement.

While mobile positioning data are becoming increasingly popular and available, to date most academic work is built on call detail records, which records mobility information only on a user's active phone usage, such as calls, text, or sometimes data connections. This paper is among the very few recent scholarly undertakings *e.g.,* Zhao *et al.* (2018) that show the applicability of mobile positioning data, in particular signalling data, for studying spatio-temporal mobility, and in particular with the focus of informing tourism and hospitality management.

5.3. *Practical implications*

To discuss the practical use of the methods for co-construction of tourism regions, a framework is proposed and presented in Figure 7. The framework follows a systemic approach that is commonly used for studying tourism regions and destinations planning (Pearce, 1989; Valeri and Baggio, 2020). It includes identifying stakeholders and their interactions in creating a tourist experience, and allows capturing dynamic structure of regions (Leiper, 1979; Piriou, 2019). Figure 7 shows five general stakeholder types: i)tourists, ii) large hospitality firms, iii) medium and small hospitality firms (food, accommodation, entertainment, transportation, and others), iv) tourism authorities (agencies in charge of tourism promotion and executive branches of the government), and v) policy makers (government). The crux of the framework is the co-creation of tourism regions that emerges from the interactions of the five entities. Tourists determine the demand for hospitality services, by making choices of visiting places and using (or not) the services, effectively determining the structures of the region (Piriou, 2019). In the framework presented, tourists' choices are observed at the level of locations, and tourists participate by generating travel patterns based on their interests. The size of the firms is an indicator of their capabilities to influence the demand and create the identity of regions. Large hospitality firms (e.g., large hotel chains) may be better positioned to invest own capital or may have easier access to it than smaller firms. Large firms may also be better positioned to capture the data from the tourists, given their shares. Hence, as shown in Figure 7, the use of the insights generated by the proposed methods may differ somewhat. For example, small and medium hospitality firms may have a greater predilection or even a stronger need for regional collaboration than large firms, which may achieve the same effect on their own. As Beritelli (2011) and Brandao *et al.* (2018) show, cooperative behavior in the tourism community is not easy to achieve. The role of tourism authorities may therefore extend beyond foreign market promotions abroad, to developing networks in regions constructed by tourist travel patterns and bolstering cooperative behavior. Finally, policy makers may use the insights to consider formalizing or planning the development of tourism regions, creating infrastructure to trigger new patterns and tourist preferences or paving the way for investments.

5.4. *Limitations and future research*

Most of the limitations of this study and the proposed Tourism2Vec method are data-related. The first significant limitation relates to the potential lack of accuracy of the location data, which is dictated by the availability of antennas and the capacity of the mobile infrastructure management to service a connected device. The second limitation of this paper concerns the representativeness of the sample investigated, as a single mobile operator typically captures only a fraction of the entire tourist population subject to EURMO's roaming partners from their home country. Nevertheless, the scale and reliability of the data available from even a single operator dwarfs other data collection methods. Another potential limitation is the inherent interpretability due to the nature of the Tourism2vec model and subsequent clustering analysis, which can be largely considered as unsupervised learning. As results obtained from the neural network and clustering are given meaning, there is a need to complement the insights with further validation, either with tourism experts or through qualitative analysis. Last, the data do not allow for the exploration of how spatio-temporal movement is affected by elements such as socio-demographic characteristics, motivations, or past experience. Future research should explore the possibility of combining different methods, qualitative and quantitative, or different data that can capture the level of detail required to better understand tourists' movements and their intentions.

Overall, this paper presents a methodological, analytical, and practical addition to the analytical and methodological toolkit for tourism region management. Tourism2Vec and complementary big data approaches can be useful tools for assisting all the stakeholders in the tourism system to better understand tourists' behavior. This enhanced understanding about tourist spatio-temporal behavior at scale can in turn inform more effective and travel-pattern-aware conceptualizations of tourism regions and contribute to sustainable regional growth.

References

Baggio, R. (2020), 'Tourism destinations: A universality conjecture based on network science', *Annals of Tourism Research* **82**, 102929.

Baggio, R. and Sainaghi, R. (2011), 'Complex and chaotic tourism systems: Towards a quantitative approach', *International Journal of Contemporary Hospitality Management* **23**(6), 840–861.

Baidal, J. A. I. (2004), 'Tourism Planning in Spain: Evolution and Perspectives', *Annals of Tourism Research* **31**(2), 313–333.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003), 'A neural probabilistic language model', *Journal of Machine Learning Research* **3**, 1137–1155.

Beritelli, P. (2011), 'Cooperation among prominent actors in a tourist destination', *Annals of Tourism Research* **38**(2), 607 – 629.

Bertella, G. (2011), 'Knowledge in food tourism: the case of lofoten and maremma toscana', *Current Issues in Tourism* **14**(4), 355–371.

Brandao, F., Carlos, C. and Dimitros, B. (2018), 'Tourism innovation networks: A regional approach', *European Journal of Tourism Research* **18**, 33–56.

Britton, S. (1991), 'Tourism, Capital, and Place: Towards a Critical Geography of Tourism', *Environment and Planning D: Society and Space* **9**(4), 451–478.

Caldeira, A. M. and Kastenholz, E. (2020), 'Spatiotemporal tourist behaviour in urban destinations: a framework of analysis', *Tourism Geographies* **22**(1), 22–50.

Caselles-Dupré, H., Lesaint, F. and Royo-Letelier, J. (2018), Word2vec applied to recommendation: Hyperparameters matter, *in* 'Proceedings of the 12th ACM Conference on Recommender Systems', Vancouver, Canada, pp. 352–356.

Chang, B., Park, Y., Park, D., Kim, S. and Kang, J. (2018), Content-aware hierarchical point-of-interest embedding model for successive poi recommendation, *in* 'Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence', Stockholm, Sweden, pp. 3301–3307.

Chen, C., Liao, C. and Wang, Y. (2019), 'Trip2vec: a deep embedding approach for clustering and profiling taxi trip purposes', *Personal and Ubiquitous Computing* **23**(1), 53–66.

Czernek, K. (2013), 'Determinants of cooperation in a tourist region', *Annals of Tourism Research* **40**(1), 83–104.

Dredge, D. and Jenkins, J. (2003), 'Destination place identity and regional tourism policy', *Tourism Geographies* **5**(4), 383–407.

East, D., Osborne, P., Kemp, S. and Woodfine, T. (2017), 'Combining GPS &

survey data improves understanding of visitor behaviour', *Tourism Management* **61**, 307–320.

Eurostat (n.d.), 'Top 20 tourism destinations (NUTS 2 regions) in terms of nights spent in tourist accommodation establishments, 2019', Available at: https://bit.ly/2K8alee.

Feng, S., Cong, G., An, B. and Chee, Y. M. (2017), POI2Vec: Geographical latent representation for predicting future visitors, *in* 'Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence', AAAI, San Francisco, CA, pp. 102–108.

Ferrari, G., Jimenez, J. M. and Secondi, L. (2016), 'Agriculture first: Assessing european policies and scientific typologies to define authentic agritourism and differentiate it from countryside tourism', *Tourism Management Perspectives* **20**(3), 251–264.

Fyall, A. and Garrod, B. (2019), 'Destination management: a perspective article', *Tourism Review* **75**(1), 165–169.

Goldberg, Y. and Levy, O. (2014), 'word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method', *arXiv:1402.3722* .

Gordon, I. and Goodall, B. (2000), 'Localities and Tourism', *Tourism Geographies* **2**(3), 290–311.

Grbovic, M. and Cheng, H. (2018), Real-time personalization using embeddings for search ranking at airbnb, *in* 'Proceedings o fthe 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', London, United Kingdom, pp. 311–320.

Gretzel, U. (2018), 'From smart destinations to smart tourism regions', *Journal of Regional Research* (42), 171–184.

Hijmans, R. J., Guarino, L. and Mathur, P. (n.d.), 'DIVA-GIS', Available at: http://www.diva-gis.org/gdata.

Jamal, S. and Habib, M. A. (2019), 'Investigation of the use of smartphone applications for trip planning and travel outcomes', *Transportation Planning and Technology* **42**(3), 227–243.

Jansen-Verbeke, M. (1995), 'A regional analysis of tourist flows within Europe', *Tourism Management* **16**(1), 73–82.

Jovicic, D. Z. (2019), 'From the traditional understanding of tourism destination to the smart tourism destination', *Current Issues in Tourism* **22**(3), 276–282.

Lau, G. and McKercher, B. (2006), 'Understanding tourist movement patterns in a destination: A GIS approach', *Tourism and hospitality research* **7**(1), 39–49.

Leiper, N. (1979), 'The framework of tourism: Towards a definition of tourism, tourist, and the tourist industry', *Annals of Tourism Research* **6**(4), 390 – 407.

Li, J., Xu, L., Tang, L., Wang, S. and Li, L. (2018), 'Big data in tourism research: A literature review', *Tourism Management* **68**, 301–323.

Liu, X., Liu, Y. and Li, X. (2016), Exploring the context of locations for personalized location recommendations, *in* 'Proceedings of the Twenty-Fifth International

Joint Conference on Artificial Intelligence', New York, NY, pp. 1188–1194.
March, R. and Wilkinson, I. (2009), 'Conceptual tools for evaluating tourism partnerships', *Tourism Management* **30**(3), 455–462.
Mariani, M., Baggio, R., Fuchs, M. and Höpken, W. (2018), 'Market segments based on the dominant movement patterns of tourists', *International Journal of Contemporary Hospitality Management* **30**(12), 3514–3554.
McKercher, B. and Lau, G. (2009), 'Methodological considerations when mapping tourist movements in a destination', *Tourism analysis* **14**(4), 443–455.
McKercher, B., Shoval, N., Ng, E. and Birenboim, A. (2012), 'First and repeat visitor behaviour: GPS tracking and GIS analysis in Hong Kong', *Tourism Geographies* **14**(1), 147–161.
Md Khairi, N. D., Ismail, H. N. and Syed Jaafar, S. M. R. (2018), 'Tourist behaviour through consumption in Melaka World Heritage Site', *Current Issues in Tourism* **22**(5), 1–19.
Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013a), 'Efficient estimation of word representations in vector space', *arXiv:1301.3781* .
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013b), Distributed representations of words and phrases and their compositionality, *in* 'Proceedings of the Twenty-seventh Conference on Neural Information Processing Systems', Lake Tahoe, NV, pp. 3111–3119.
Nisbet, R., Elder, J. and Miner, G. (2009), *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press, chapter 7, pp. 121–150.
Pang, J. and Zhang, Y. (2017), Deepcity: A feature learning framework for mining location check-ins, *in* 'Proceedings of the Eleventh International AAAI Conference on Web and Social Media', Montreal, Canada, pp. 652–655.
Pearce, D. (1989), *Tourism Development*, 2 edn, Longman, Harlow, UK.
Pillai, R. and Sivathanu, B. (2020), 'Adoption of ai-based chatbots for hospitality and tourism', *International Journal of Contemporary Hospitality Management* **32**(10), 3199–3226.
Piriou, J. (2019), *The Tourist Region: A Co-Construction of Tourism Stakeholders*, John Wiley Sons, Inc., Hoboken NJ, USA.
Popp, M. (2012), 'Positive and negative urban tourist crowding: Florence, italy', *Tourism Geographies* **14**(1), 50–72.
Randelli, F., Romei, P. and Tortora, M. (2014), 'An evolutionary approach to the study of rural tourism: The case of tuscany', *Land Use Policy* **38**(3), 276–281.
Ravi, L., Subramaniyaswamy, V., Vijayakumar, V., Chen, S. G., Karmel, A. and Devarajan, M. (2019), 'Hybrid location-based recommender system for mobility and travel planning', *Mobile Networks & Applications* **24**(4), 1226–1239.
Reif, J. and Schmücker, D. (2020), 'Exploring new ways of visitor tracking using big data sources: Opportunities and limits of passive mobile data for tourism', *Journal of Destination Marketing & Management* **18**, 100481.
Rigall-I-Torrent, R. and Fluvià, M. (2011), 'Managing tourism products and desti-

nations embedding public good components: A hedonic approach', *Tourism Management* **32**(2), 244–255.

Ritchie, J. B. and Zins, M. (1978), 'Culture as determinant of the attractiveness of a tourism region', *Annals of Tourism Research* **5**(2), 252–267.

Saarinen, J. (1998), The Social Construction of Tourism Destination: The Transformation Process of Saariselka Resort in Finnish Lapland, *in* G. Ringer, ed., 'Destinations: Cultural Landscapes of Tourism', Routledge., London and New York, pp. 154–73.

Saarinen, J. (2004), ''Destinations in change': The transformation process of tourist destinations', *Tourist Studies* **4**(2), 161–179.

Shi, J., Xin, L. and Liu, Y. (2020), 'Simulation of tourists' spatiotemporal behaviour and result validation with social media data', *Transportation Planning and Technology* **43**(7), 698–716.

Smith, L. J. (1989), *Tourism Analysis: A Handbook*, Longman, Harlow, UK.

Su, X., Spierings, B., Dijst, M. and Tong, Z. (2020), 'Analysing trends in the spatiotemporal behaviour patterns of mainland Chinese tourists and residents in Hong Kong based on Weibo data', *Current Issues in Asian Tourism* **23**(12), 1542–1558.

Terkenli, T. (2002), 'Landscapes of Tourism: Towards a Global Cultural Economy of Space?', *Tourism Geographies* **4**(3), 227–254.

Tiru, M., Kuusik, A., Lamp, M.-L. and Ahas, R. (2010), 'LBS in marketing and tourism management: measuring destination loyalty with mobile positioning data', *Journal of Location Based Services* **4**(2), 120–140.

Tosun, C. and Jenkins, C. (1996), 'Regional planning approaches to tourism development: the case of Turkey', *Tourism Management* **17**(7), 519–531.

Valeri, M. and Baggio, R. (2020), 'Italian tourism intermediaries: a social network analysis exploration', *Current Issues in Tourism* .

Versichele, M., De Groote, L., Bouuaert, M. C., Neutens, T., Moerman, I. and Van de Weghe, N. (2014), 'Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium', *Tourism Management* **44**, 67–81.

Wang, H., Shen, H., Ouyang, W. and Cheng, X. (2018), Exploiting poi-specific geographical influence for point-of-interest recommendation, *in* 'Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence', Stockholm, Sweden, pp. 3877–3883.

Zhao, S., Zhao, T., King, I. and Lyu, M. R. (2017), Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation, *in* 'Proceedings of the 26th International Conference on World Wide Web Companion', ACM, pp. 153–162.

Zhao, X., Lu, X., Liu, Y., Lin, J. and An, J. (2018), 'Tourist movement patterns understanding from the perspective of travel party size using mobile tracking data: A case study of Xi'an, China', *Tourism Management* **69**, 368–383.

Zhong, S. and Zhang, J. (2012), 'Why People Travel to Different Regions: a New

Tourism Research Framework from Geographical Perspective', *Procedia Environmental Sciences* **12**, 408–412.

Zhu, M., Chen, W., Xia, J., Ma, Y., Zhang, Y., Luo, Y., Huang, Z. and Liu, L. (2019), 'location2vec: a situation-aware representation for visual exploration of urban locations', *IEEE Transactions on Intelligent Transportation Systems* **20**(10), 1–10.

Municipality	Top 5 most similar municipalities (rank of geographic distance)
Major cities	
Florence	Bagno a Ripoli (4), Sesto Fiorentino (1), Impruneta (3), Scandicci (5), Campi Bisenzio (6)
Siena	Monteriggioni (1), Monteroni d'Arbia (2), Sovicille (3), Castelnuovo Berardenga(4), Rapolano Terme (16)
Pisa	San Giuliano Terme (1), Vecchiano (4), Cascina (2), Collesalvetti (3), Ponsacco (17)
Prato	Campi Bisenzio (3), Serravalle Pistoiese (20), Pistoia(21), Quarrata (8), Calenzano (5)
Lucca	Capannori (1), Altopascio (14), Borgo a Mozzano (5), San Giuliano Terme (6), Vecchiano (7)
Arezzo	Civitella in Val di Chiana (2), Laterina (8), Terranuova Bracciolini (20), Pergine Valdarno (10), San Giovanni Valdarno (29)
Pistoia	Serravalle Pistoiese (3), Pieve a Nievole (14), Montecatini-Terme (9), Quarrata (10), Agliana (4)
Carrara	Massa (1), Fosdinovo (2), Seravezza (4), Pietrasanta (11), Forte dei Marmi (6)
Coastal area	
Livorno	Capraia Isola (113), Pisa (6), Collesalvetti (1), Rosignano Marittimo (4), Ponsacco(12),
Piombino	Campiglia Marittima (1), Suvereto (3), Follonica (4), San Vincenzo (2), Rio Marina (6)
Countryside	
Greve in Chianti	Radda in Chianti (5), Castellina in Chianti (11), San Casciano in Val di Pesa (3), Tavarnelle Val di Pesa (2),Impruneta (4)
Asciano	Rapolano Terme (1), Siena (7), Trequanda (6), Monteroni d'Arbia (2), San Giovanni d'Asso(4)
San Quirico d'Orcia	Castiglione d'Orcia (3), Pienza (1), Montalcino (4), Radicofani (13), Monteroni d'Arbia (18)
Reggello	Figline e Incisa Valdarno (5), Rignano sull'Arno (3), San Giovanni Valdarno (10), Bagno a Ripoli (8), Pelago (2)
Islands	
Portoferraio	Marciana (6), Capoliveri (1), Campo nell'Elba (3), Porto Azzurro (2), Marciana Marina (4)
Isola del Giglio	Monte Argentario (1), Orbetello (2), Capalbio (4), Rio Marina (18), Magliano in Toscana (3)

Table 1: **List of top 5 most similar municipalities given the highest cosine similarity score of the neural embeddings. The ranks of the municipality in terms of the geographic proximity are reported in the parentheses.**

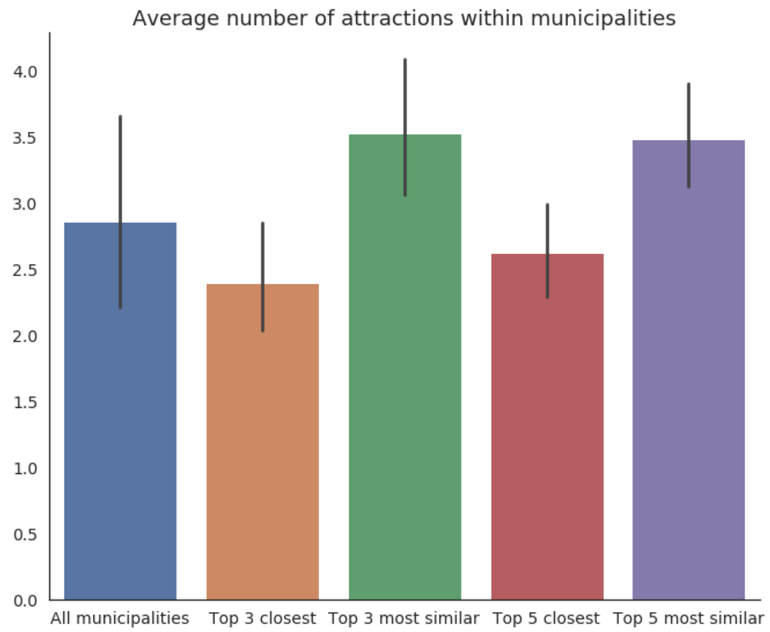
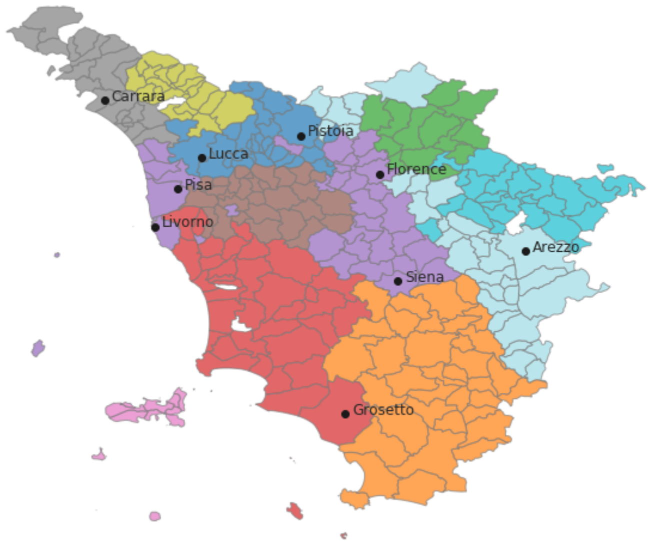


Fig. 5: **Average number of attractions within all municipalities and those that are among the top 3(5) most similar in the latent space and top 3(5) geographically closest.**



(a) Clusters of municipalities based on neural embeddings

(b) Provinces of Tuscany region

Fig. 6: **Location clustering using K-means algorithm to group municipalities into 11 clusters, in comparison with 10 provinces of Tuscany**

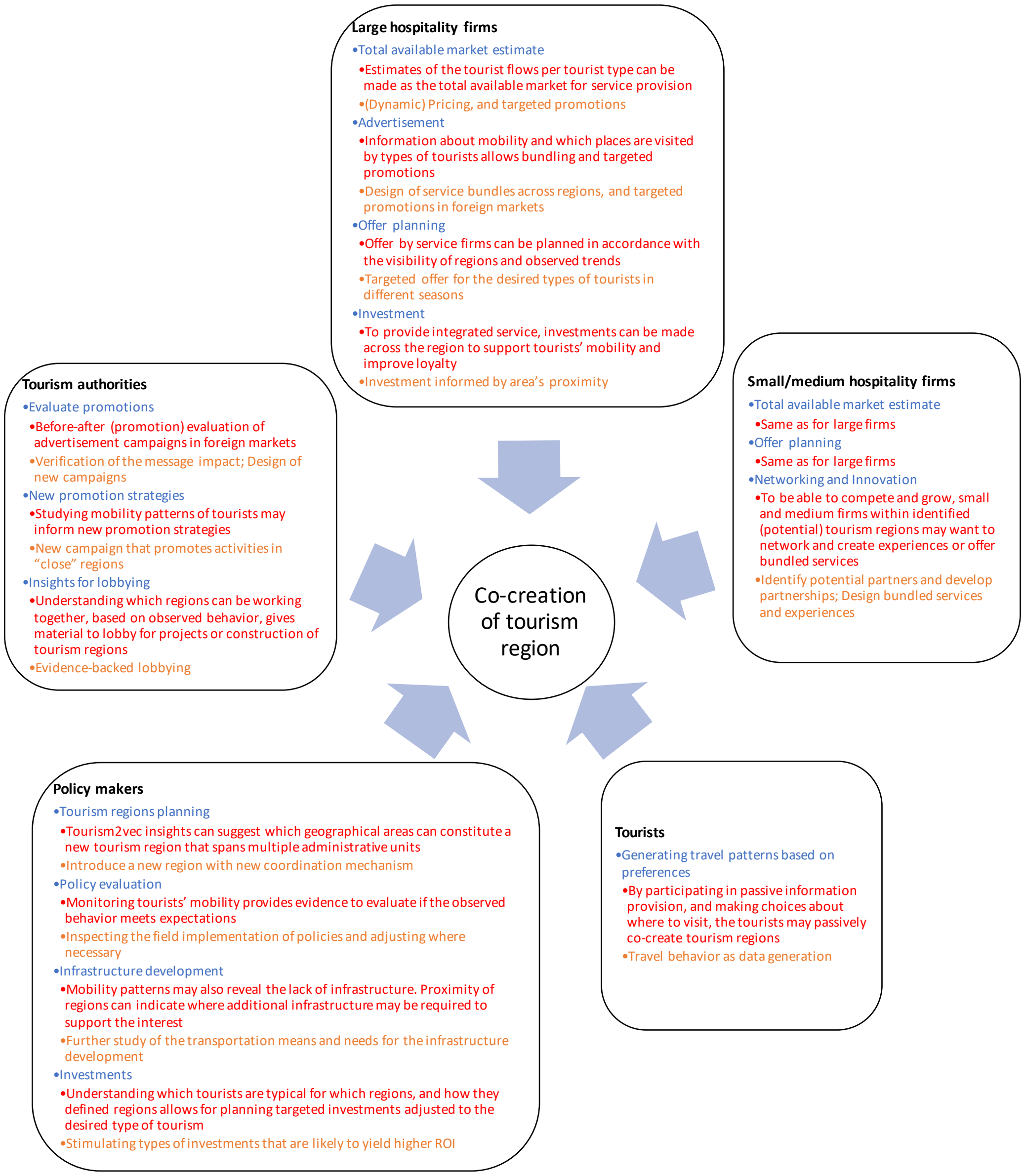


Fig. 7: **The role of multiple stakeholders in the co-creation of tourism region. Each stakeholder is characterized by its functions (in blue), description of the functions(in red) and actions (in orange)**