Expert Systems With Applications 140 (2020) 112892

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Land use discovery based on Volunteer Geographic Information classification

Fernando Terroso-Saenz [a,*], Andrés Muñoz [a]

Catholic University of Murcia, Spain

ARTICLE INFO

Article history:
Received 3 March 2019
Revised 8 July 2019
Accepted 19 August 2019
Available online 20 August 2019

Keywords:
Urban computing
Volunteer Geographic Information (VGI)
Land usage
Supervised classification

ABSTRACT

Nowadays, cities are dynamic ecosystems where urban changes occur at a very fast pace. Hence, social sensing has become a powerful tool to uncover the actual land-use of a metropolis. However, current solutions for land-use discovery based on user-generated data usually rely on an information retrieval mechanism applied on a textual corpus. This causes ad-hoc place labelling with limited semantic meaning. In this line, the present work introduces a novel data-driven methodology that extends existing solutions by means of a classifier based on a pre-defined hierarchy of land categories. Two types of social networks –text-based and venue-based platforms– are utilized to train the classifier, which is then applied to infer the use of the land based on text data in areas where venue data are not available. The approach has been evaluated by using large datasets comprising two large cities, showing an accuracy above 90% in predicting the land-use categories.

1. Introduction

During the last years, rural areas have witnessed the endless transference of population to cities so that it is projected that 68% of the population will live in metropolises by 2050.[1] Consequently, cities can be seen as complex and dynamic systems imposing new problems to city administrators in order to come up with efficient urban services like air pollution monitoring or multi-modal transportation services. In this context, it is paramount to understand the dynamics and organization of a city to properly design the aforementioned services in an efficient manner. In particular, this paper focuses on the application of social sensing techniques to automatically discover the real use of land in cities. This information is paramount for municipalities so as to monitor the evolution of different areas of the city and define an efficient urban planing, providing more comfortable urban environments (Jensen & Cowen, 1999).

Several technologies and techniques to gather and analyse urban data have been developed in the last years within the *Urban Computing* paradigm (Salim & Haque, 2015; Zheng, Capra, Wolfson, & Yang, 2014; Zheng, Mascolo, & Silva, 2017). One of the most promising directions to capture such urban data comes from social sensing (Liu et al., 2015). The ubiquitous use of personal mobile devices with location-based capabilities has entitled citizens to generate an unprecedented amount of spatio-temporal data which are stored in different online social network (OSN) platforms like Twitter,[2] Facebook,[3] Flickr[4] and other web services (Chiara Renso & Stefano, 2013). All the wealth of user-driven geo-tagged information has giving raise to the term *Volunteer Geographic Information* (VGI) (Goodchild, 2007).

A prominent course of action to use such digital footprints has focused on better understanding and describing the land usage of a city. That is, the characterization of the geographical areas of a city in different *categories* such as residential, recreational or business. In this case, VGI documents are used as an alternative data source to uncover the actual land use of a city. Unlike official land-use data based on surveys and satellite imagery, which tend to be rather out-dated, VGI platforms allow composing more timely and updated land-use maps of an urban environment.

Whereas a plethora of approaches has been successfully applied to VGI in land-use detection (Cranshaw, Schwartz, Hong, & Sadeh, 2012; Hollenstein & Purves, 2010; Lansley & Longley, 2016; Rudinac, Zahálka, & Worring, 2017), some limitations can be observed though. On the one hand, most proposals use different information retrieval and topic modelling techniques in order to analyze

* Corresponding author.
E-mail addresses: fterroso@ucam.edu (F. Terroso-Saenz), amunoz@ucam.edu (A. Muñoz).

[1] https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

[2] https://twitter.com.

[3] https://www.facebook.com.

[4] https://www.flickr.com.

https://doi.org/10.1016/j.eswa.2019.112892
0957-4174/

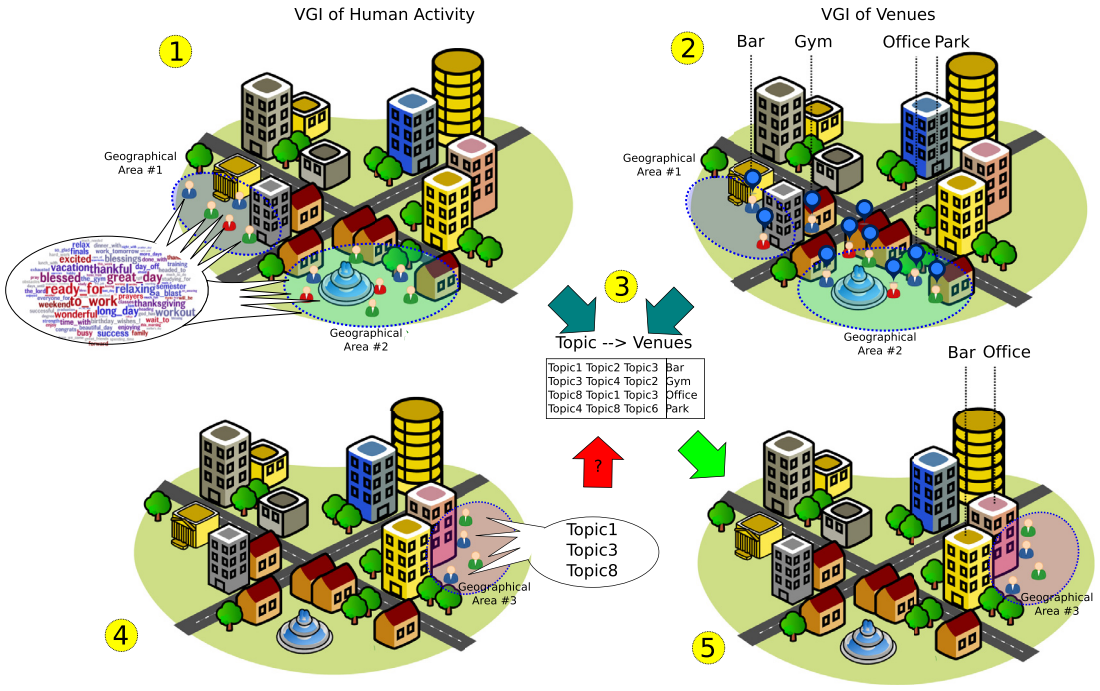


Fig. 1. Approach overview. In Step 1 topics of interest are extracted from text-based VGI sources in areas 1 and 2. Similarly, in Step 2 places of interest are extracted from venue-based VGI sources in the same areas. In Step 3 a classifier is developed based on the association between the topics and the venues. In Step 4 topics of interest are extracted from new areas where venue data are not available. Finally, in Step 5 land-use information is discovered using the classifier. Note that the data are simplified for example purpose.

the textual information contained in VGI documents (e.g., the text written by a user in a tweet or the tags describing a photo in Flicker). Next, the outcome of these techniques is used to label the regions of the target city describing their usage. This procedure tends to generate ad-hoc labels which actually hamper the interoperability of the generated land-use maps.

On the other hand, despite the large number of available VGI sources, many existing solutions do not fully leverage them. In this sense, most proposals rely on a unique VGI source in order to perform the land-use mapping. This usually results in a quite biased analysis depending on the selected data source. Besides, other proposals only use a subset of the documents contained in the VGI feeds and do not benefit from all the data available in that source.

All of this has led us to present a novel approach to classify the land usage of a city. As shown in Fig. 1, the introduced solution combines information from two different types of VGI platforms, a text-based and a venue-based VGI platform. The former stands for platforms where contributors post geo-tagged documents mainly containing textual information (like most OSNs) whereas the latter represents platforms where users report information about existing venues of a city. A clear example of this second type of platforms are Location-based Social Networks (LBSNs).

More specifically, the documents from the text-based platform are used to extract a set of features from the textual content published by users at different *places* of the city (step 1 of Fig. 1). These places are discovered by means of a clustering process. Then, the venue-based platform is used to identify the most relevant land-use categories at each place (step 2 of Fig. 1). For this task, the category hierarchy of the venue-based platform is considered. Next, a supervised learning task is performed to develop a learning function that associates the input textual content's features of a place to its relevant categories (step 3). Finally, this classifier is used to uncover the land-use categories for new areas of a city on the basis of the topics discussed by users at such areas (steps 4 and 5 of Fig. 1).

The contributions in the state of the art through the present solution are twofold. Firstly, the resulting classifier allows *using* the land categories of the venue-based platform in geographical areas where there is a lack of documents of such a platform. Hence, our approach *extends* the coverage of VGI venue-based platforms to new urban regions. Furthermore, since place labelling is based on a predefined category hierarchy, we avoid the aforementioned ad-hoc categorization problem. The obtained results allow predicting the category of use of the urban land with an accuracy of over 90% in the two large cities (New York and San Francisco) used as test cases.

The remainder of the paper is structured as follows. Section 2 provides an overview about land use discovery based on social sensing. Next, Section 3 is devoted to describing in detail the logic structure and the processing stages of the proposed system. Then, Section 4 discusses the main results of the performed experiments. Finally, the main conclusions and the future work are summed up in Section 5.

2. Related work

When it comes to discover the land use of an urban area with social sensing, it can be established three different characteristics to catalogue the existing literature: the type of incoming VGI, the method to detect areas of interest and the place labelling procedure (see Table 1). A review on each of these characteristics is given next. Finally, this section ends with a review on multi-labelling classification techniques.

2.1. Type of incoming VGI

As far as the type of incoming VGI is concerned, a sheer number of proposals make use of OSN documents for place labelling. In that sense, Twitter is one of the most popular platforms in this context (Frias-Martinez & Frias-Martinez, 2014; Hollenstein & Purves, 2010; Kovacs-Gyri et al., 2018; Lansley & Longley, 2016; Mahmud, Nichols, & Drews, 2014; Yuyun, Akhmad Nuzir, & Julien Dewancker, 2017). Other interesting works make use of the Foursquare platform as input data

Table 1
Key features of existing land-use detection approaches.

Ref.	Input Datasource	Region Identification Method	Land-labelling	
			Method	Target labels
(Hollenstein & Purves, 2010)	Twitter	Kernel Density Estimator	Terms aggregation	User-generated terms
(Yuyun et al., 2017)	Twitter	*k*-means algorithm	Rank method	Foursquare hierarchy
(Lansley & Longley, 2016)	Twitter	Land-use Database	Latent Dirichlet Allocation (LDA)	LDA topics
(Kovacs-Gyri et al., 2018)	Twitter	OSM polygons	Sentiment analysis	Sentiment scores
(Mahmud et al., 2014)	Twitter	Geocoding API	Nave Bayes Multimodal	Ad-hoc labels
(Frias-Martinez & Frias-Martinez, 2014)	Twitter	Self-Organizing Maps	Time series analysis	Ad-hoc labels
(Yang et al., 2016)	Foursquare	Spectral clustering	Hierarchical classifier	Cultural features
(Gao et al., 2017)	Foursquare	*k*-means algorithm	Latent Dirichlet Allocation (LDA)	LDA topics
(Zhou & Zhang, 2016)	Foursquare	Spatial tessellation	Support Vector Machine	Foursquare hierarchy
(Rudinac et al., 2017)	Foursquare & Flickr	*k*-means algorithm	Extra-tree classifier	Regions statistics
(Wartmann et al., 2018)	Flickr & Hiking Blog	Location-based service	Terms aggregation	User-generated terms
(Lenormand et al., 2015)	CDRs	Community detection algorithm	Time series analysis	Ad-hoc labels
(Noulas et al., 2013)	Foursquare & CDRs	DBSCAN	Multi-class classifier	Foursquare hierarchy
(Quercia & Saez, 2014)	Foursquare & Municipality OD	Neighbourhood areas	Logistic Bayesian classifier	Deprivation level
(Spyratos et al., 2017)	Foursquare & Municipality OD	Building-block areas	Geometric allocation	Foursquare hierarchy
(García-Palomares et al., 2018)	Twitter & Municipality OD	Transport areas	Cadastral-data analysis	Ad-hoc labels
(McKenzie et al., 2018)	Rental Listing	Random Forest	-	-
Our proposal	Foursquare & Flickr	HDBSCAN	Multi-label classifier	Foursquare hierarchy

(Gao, Janowicz, & Couclelis, 2017; Yang, Zhang, & Qu, 2016; Zhou & Zhang, 2016). This second type of proposals generally obtain the Foursquare check-in posts embedded in tweets. Since not all the Foursquare check-ins are published in Twitter, solutions based on this type of VGI usually suffer from rather biased results.

Flickr has been also used for land-use discovery. One interesting example can be found in Wartmann, Acheson, and Purves (2018) where authors also take into account alternative user-generated data sources like blogs. In Rudinac et al. (2017), Flickr photos are combined with Foursquare check-ins as input data. Although our approach uses the same two platforms as input feeds, in the cited work Foursquare is used to extract information about users visiting certain venues. On the contrary, our approach leverages Foursquare in order to obtain information about venues themselves.

Other interesting data sources for land-use discovery are the Call Detail Records (CDRs) provided by communication firms in an isolated manner (Lenormand et al., 2015) or merged with other VGI (Noulas, Mascolo, & Frias-Martinez, 2013). The key drawback of CDR data is that it is not usually freely available mainly due to privacy legislation. Some works have also combined static land-use Open Data (OD) from municipalities with VGI from Foursquare (Quercia & Saez, 2014; Spyratos, Stathakis, Lutz, & Tsinaraki, 2017) or Twitter (García-Palomares, Salas-Olmedo, Moya-Gómez, Condeo-Melhorado, & Gutiérrez, 2018). Finally, other user-generated data sources like rental listings have also been studied (McKenzie, Liu, Hu, & Lee, 2018).

2.2. Region identification method

Regarding the method to detect the target regions of a city, we can clearly distinguish two courses of action.

On the one hand, several approaches depend on different third-party services for region partitioning. These are the cases of Wartmann et al. (2018), which make use of a location-based service to identify the target landscapes, and Lansley and Longley (2016), which use a national land-use database. Likewise, Kovacs-Gyri et al. (2018) depend on the Open Street Map's (OSM)[5] pre-defined polygons to define the target areas and Mahmud et al. (2014) rely on the Google's geo-coding API.[6] In Quercia and Saez (2014), spatial aggregation is carried out on the basis of the official-census areas of the city and García-Palomares et al. (2018) follow a similar approach but using the transport zones established by the city transport authority. Although these are very reliable methods, they are not able to rapidly adapt to changes in the dynamics of a city.

On the other hand, a second trend proposes different data mining techniques applied to VGI for region tessellation. A foremost approach is based on clustering algorithms applied to the incoming documents' coordinates. In this context, Rudinac et al. (2017) and Gao et al. (2017); Yuyun et al. (2017) apply the *k*-means technique, whereas an ad-hoc spectral clustering algorithm is proposed in Yang et al. (2016) and DBSCAN is used in Noulas et al. (2013). Besides, Self-Organizing Maps have been also used in Frias-Martinez and Frias-Martinez (2014) whereas some other works rely on other non-parametric methods such as the Kernel Density Estimator (KDE) (Hollenstein & Purves, 2010). In McKenzie et al. (2018) authors opt for a supervised learning approach using Random Forest (RF). A lightweight spatial division based on grilled cells is proposed in Zhou and Zhang (2016). Going beyond the spatial features of the documents, Lenormand et al. (2015) establish a region aggregation based on user-based community detection algorithms.

Our proposal makes use of HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) (Campello, Moulavi, & Sander, 2013), a novel density-based clustering algorithm. HDBSCAN provides remarkable advantages with respect to the aforementioned clustering algorithms, since it allows for non-round clusters with different density along with noise classification. Furthermore, HDBSCAN has already been successfully applied in several scenarios related to geographical data (Cao, Liu, Li, Wang, & Qin, 2018; Jenson, Reeves, Tomasini, & Menezes, 2017; Lopez-Ramirez & Siordia, 2016).

[5] http://www.openstreetmap.org/.
[6] https://developers.google.com/maps/documentation/geocoding/start.

2.3. Land-labelling method

The third key feature for land-use discovery is the procedure for labelling the previously-generated spatial zones with their corresponding category. As it can be seen from the two rightmost columns in Table 1, a large variety of proposals exist in this field.

One common line of work has made use of the Latent Dirichet Allocation (LDA) method to analyze the textual corpus of the incoming VGI documents and extract its relevant topics. Next, each region of interest is labelled with its dominant topic (Gao et al., 2017; Lansley & Longley, 2016). Other works have proposed even more simple solutions based on just the aggregation of user-generated terms to make up the regions labels (Hollenstein & Purves, 2010; Wartmann et al., 2018). As it was discussed in Section 1, these mechanisms compose ad-hoc labels for region tagging which turn out in a lack of interoperability among the solutions.

Another interesting method to categorize the regions consists in analyzing their temporal dynamics (Frias-Martinez & Frias-Martinez, 2014; Lenormand et al., 2015). By studying the temporal features of the documents posted by users in each cluster, it is possible to infer their land use (e.g., regions with a high frequency of documents posted in the evening are likely residential or nightlife areas). The key drawback of this type of works is that they only rely on the temporal attributes of the VGI documents discarding the textual content. Therefore, they do not fully profit the information contained in the documents.

A third course of action focuses on using a classifier in order to infer the most suitable label for each region. In this context, Mahmud et al. (2014) propose a Naive Bayes Multimodal solution in order to identify the home location of VGI contributors, and therefore it can distinguish between residential and non-residential land use of a city. A methodology to analyze the deprivation level of certain areas of a city by means of a Logistic Bayesian classifier is stated in Quercia and Saez (2014). Moreover, Rudinac et al. (2017) propose an extra-tree classifier to identify the most discriminant statistical features of city regions. A completely different approach is proposed in Kovacs-Gyri et al. (2018) where sentiment analysis is used to label city parks with city dwellers' experience. In a more simple way, Spyratos et al. (2017) just follow a geometric allocation heuristic in order to adjust the actual placement of the Foursquare venues at a building block level, whereas García-Palomares et al. (2018) carry out a preliminary analysis on the use of previously uncovered regions using cadastral data. Besides, Yuyun et al. (2017) propose a rank method in order to assign to each cluster the category of its most visited venue.

Our proposal presents certain similarities with some works regarding the land labelling procedure. In particular, Zhou and Zhang (2016) train a Support Vector Machine (SVM) model to associate the most likely Foursquare category to each single VGI document. Although our work also classifies each VGI document on the basis of the Foursquare hierarchy, it takes a step further by aggregating such outcome in order to discover the underlying land use of city regions. Furthermore, we apply a multi-class classifier that allows uncovering more than one Foursquare category per document. Yang et al. (2016) make use of a hierarchical classifier but the goal in that work is to infer the cultural features (e.g., language, religion and the like) of the city regions instead of their business characteristics like ours.

Finally, Noulas et al. (2013) make use of a multi-class classifier in order to detect the most popular activity of an area on the basis of the Foursquare hierarchy. Two key differences exist between this work and ours. To begin with, while Noulas et al. (2013) take as independent variables certain features related to the CDR data like the frequency of outgoing calls of a communication tower, our classifier takes as input the topics discussed by users located at the regions. Secondly, our approach proposes a multi-label classifier, instead of a multi-class one, in order to label each region with different categories.

This multi-labelling feature turns out quite instrumental in urban environments. In this sense, the present work is based on the hypothesis that modern urban areas tend to have a mixture of usages avoiding zones with a single type of land use. As a matter of fact, shopping malls or department stores, which are now in almost any city, are surrounded by a large number of many different venues ranging from restaurants or theaters to museums. On the contrary, nowadays it is difficult to find urban areas comprising only venues of a single type. Consequently, in many situations the actual use of urban regions can not be properly described by a single label but a palette of them instead.

2.4. Multi-labelling classification

The key goal of multi-labelling classification is to tag an instance into a set of labels instead of a single one like in the single-label classification (Herrera, Charte, Rivera, & Del Jesus, 2016). In this case, we can distinguish two types of mechanisms in order to carry out this classification.

One the one hand, a common approach to deal with multi-labelled classification problems focuses on applying different methods that transform a single multi-label problem into multiple single-label problems. Hence, three well-established mechanisms can be referred:

- Binary Relevance (BR) (Godbole & Sarawagi, 2004). This technique takes as input a particular base classifier. Next, it transforms a multi-label classification problem with $\mathcal{L}$ labels into $\mathcal{L}$ different single-label binary classification problems using the base classifier with an one vs. all strategy. Hence, the prediction outcome is the union of all per-label classifiers.
- Classifier Chains (CC) (Read, Pfahringer, Holmes, & Frank, 2009). This method also takes as input a base single-label estimator. Then, it constructs a bayesian conditioned chain of base classifiers so that the multi-label problem is transformed to a multi-class problem where each label combination is a separate class and the size of the chain is equal to the number of labels.
- Label Powerset (LP) (Tsoumakas & Vlahavas, 2007). In this case, LP solves a multi-label problem by considering it as a multi-class one where each label combination is a separate class. Thus, it uses the base single-label classifier to solve the problem.

On the other hand, *algorithm adaptation methods* intend to generalize existing classification algorithms to handle multi-label data. In that sense, one popular technique is Multilabel k Nearest Neighbours (MLkNN) (Zhang & Zhou, 2007). This is a variation of the popular clustering algorithm kNN in order to deal with multi-label challenges. In particular, MLkNN builds kNN models to find the nearest examples to a test class and then applies a Bayesian inference to select assigned labels. Unlike the previous techniques, this one does not require a base single-label classifier.

Regarding applications of multi-label classification, a common area where this technique has been used is text categorization (Charte, Rivera, del Jesus, & Herrera, 2015; Liu & Chen, 2015). More specifically, Liu and Chen (2015) evaluate different multi-label classification techniques for sentiment analysis extracted from textual documents. Besides, Charte et al. (2015) use the MLkNN mechanism to tag questions posted in electronic forums. Beyond textual documents, multi-label classification has also been applied to multimedia content (Briggs et al., 2012; Wang, Wang, & Ji, 2015). More in detail, Wang et al. (2015) apply this classification

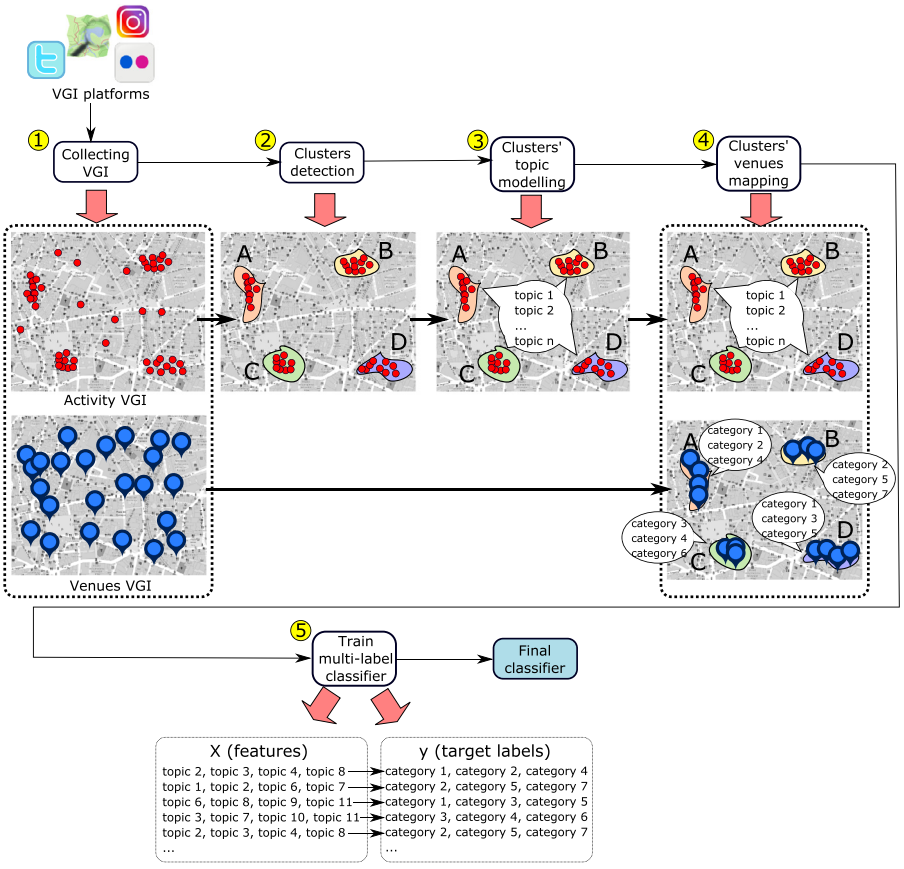


Fig. 2. Main stages of the methodology.

to assign emotion labels for audio and visual features. Moreover, Briggs et al. (2012) make use of these techniques in order to identify bird species from sound records. Finally, another field of interest where multi-label models have been used is biology. For example, Chou, Wu, and Xiao (2011) make use of this type of models to infer the potential location of a protein.

3. Land-use categorization based on VGI classification

This section is devoted to put forward the proposed mechanism of land use classification based on VGI.

3.1. Methodology overview

The proposed methodology is developed as follows. Firstly, it uncovers the most important spatial areas of a city. Next, it identifies the topics of interest (e.g., sports, politics, leisure, etc.) discussed by users located at these areas. Then, we identify the land use of each of these areas as the most popular categories of the venues located at each area (e.g., bars, shops, restaurants, etc.). Finally, a classifier associates the discussed topics with a particular land use. Consequently, it is possible to infer the land use of a spot on the basis of the topic of interest discussed by its visitors.

Fig. 2 sums up the key steps of the methodology. The upcoming sections explain each step in detail. The methodology notation used in these sections is summarized in Table 2.

3.2. VGI collection

As Fig. 2 shows, the first step of the proposed methodology focuses on collecting the necessary VGI from different sources. In that sense, two types of VGI must be gathered, one about citizens moving around the target city and other about the actual land use of such a city.

Table 2
Methodology notation.

Symbol	Meaning
d	Raw user-centered VGI document
$\mathcal{D}$	Dataset of raw user-centered documents
v	Venue VGI document
$\mathcal{V}$	Dataset of raw venue documents
cl	Density-based cluster representing city landmark
$\mathcal{CL}$	Dataset of density-based clusters
cl_l	Cluster representing city landmark (category labelled)
$\mathcal{CL}_l$	Dataset of category-labelled clusters
d_f	Filtered user-centered VGI document
$\mathcal{D}_f$	Dataset of filtered user-centered documents
d_f^s	Filtered user-centered VGI document with topics
$\mathcal{D}_f^s$	Dataset of filtered user-centered documents with topics
$\mathcal{S}$	Set of topics from the textual corpus of $\mathcal{D}_f^s$

3.2.1. User-centered dataset

Firstly, we collect data about users located at the target urban area. This type of VGI is related to preferences, activities, opinions and so forth expressed by users in their documents published at several platforms. Depending on the platform under consideration, the collected VGI documents take different formats like geotagged *tweets* in case of Twitter, posts in Facebook or photographs in Flickr.

Consequently, in order to provide a uniform view of these documents, a user-based document $d \in \mathcal{D}$ is characterized by a user u who actually generated the document, the spatial location l at which d was posted as a pair $\{x, y\}$ of coordinates,

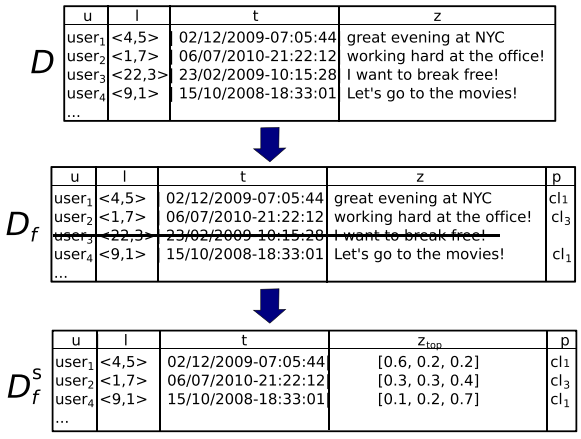


Fig. 3. Transformation of the user-centered documents.

the timestamp of the submission t and the textual corpus of the document z. Thus, a user-based document is defined as a tuple $d =< u, l, t, z >$.

3.2.2. Venue-centered dataset

The second type of collected VGI is about the venues located at the target city. Location-based Social Networks (LBSNs) like Foursquare[7] allow users to report existing business by means of *check-in* posts. One key feature of these posts is that users are entitled to specify the category of the reported venue. This category points out the *nature* of the business at multiple granularity levels (e.g. bar, grocery store, restaurant, etc). As it was put forward in Section 3.1, we use such categories in order to specify the land use of the different regions of a city.

As a result, for each venue $v \in \mathcal{V}$, we gather its name n, its category c and its spatial location l as a pair $\{x, y\}$ of coordinates. Hence, a venue-based document is defined as a tuple $v =< n, c, l >$.

3.3. Cluster detection and document cleaning

Once we have collected the required data, the next step is to detect the most dynamic spatial regions of the city in terms of human activity. These dynamic regions will be the target of our mechanism to infer their underlying land use.

Since this type of regions generally contains a large number of published VGI documents (Jiang, Ma, Yin, & Sandberg, 2016), we apply a density-based spatial clustering over the documents in $\mathcal{D}$ to uncover those regions. As a result, a set of clusters $\mathcal{CL}$ is created where each cluster $cl \in \mathcal{CL}$ represents a single city landmark (see step 2 in Fig. 2). For each cluster, we keep its identifier id and a list of locations $\{l^1_{cl}, l^2_{cl}, \ldots, l^n_{cl}\}$ composing its spatial region's convex hull h. As we will see in Section 3.6, we need this convex hull in order to retrieve the venues that perfectly fit into the cluster's spatial area. All in all, a cluster is defined by a tuple $cl =< id, h >$.

Once the cluster set $\mathcal{CL}$ has been composed, we use it to filter the documents in $\mathcal{D}$. In particular, we remove the documents that are not contained in any cluster. In that sense, we consider that a document d is part of a cluster cl when its location l is enclosed in the convex hull h of the cl. This results in a new set of filtered documents $d_f \in \mathcal{D}_f$ comprising the documents contained in any cluster $\mathcal{CL}$. Each filtered document includes the cluster p that its belongs to, so it is defined as a tuple $d_f =< u, l, t, z, p >$. The upper and middle tables in Fig. 3 show a small example of this transformation where one document of $\mathcal{D}$ is not included in $\mathcal{D}_f$ as it does not belong to any cluster.

3.4. Clusters' topic modelling

The third step of the methodology is to infer the different topics of interest of the documents located at the landmarks discovered in the previous stage (see step 3 in Fig. 2).

In the existing literature, we can find a palette of information retrieval techniques for topic modelling based on matrix factorization (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer, Foltz, & Laham, 1998) or probabilistic generative models (Blei, Ng, & Jordan, 2003; Teh, Jordan, Beal, & Blei, 2006). In general, these techniques allow to (1) uncover the relevant set of topics associated to a textual corpus, and (2) for each document of the corpus, establish its related topics from the set.

Bearing this idea in mind, we generate the topics associated to the documents $\mathcal{D}_f$ by means of one of the most popular mechanisms for topic modelling, the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). Thus, we feed this model with the corpus comprising the textual content z of all the documents $\mathcal{D}_f$. Before doing that, we clean such corpus by removing stop words and applying a word stemming process to set the rest of words in their root form.

The outcome of the model is a set of n_{top} topics of the aforementioned corpus, $\mathcal{S} = \{s_1, s_2, .., s_{n_{top}}\}$. These are the top-$n_{top}$ themes of interest discussed by users at the target city according to the documents in D_f.

3.5. Documents' features extraction

Once the LDA model has been used to detected the topics of interest, it can also infer the weight of these topics in a single document according to its textual content. These weights measure the relationship of the document with the topics and each one usually ranges from 0 (no relationship) to 1 (strong relation).

By making use of this inference mechanism, each document $d_f \in \mathcal{D}f$ gives raise to a new document $d^s_f \in \mathcal{D}^s_f$ which basically replaces the raw textual content z with its topic weights z_{top}. Hence, the new document can be regarded as a tuple $d^s_f = \{u, l, t, z_{top}, p\}$.

The bottom table of Fig. 3 shows an example of this transformation when n_{top} is set to 3. It can be seen that the weights of the first document for each of the 3 topics are [0.6, 0.2, 0.2]. This means that this document is much more related to the first topic than to the second and third one. On the contrary, the second document is slightly more related to the third uncovered topic as its weight distribution is [0.3, 0.3, 0.4].

It is worth noticing that, by means of this procedure, we are able to transform the raw textual content of the documents to a set of numeric features. These numeric features are suitable to be taken as input by a classification algorithm as we will see in Section 3.7.

3.6. Cluster venue mapping

This step focuses on detecting the land usage of the landmarks $\mathcal{CL}$. For this goal, we make use of the venues $\mathcal{V}$ collected at Section 3.2.2.

For each cluster $cl \in \mathcal{CL}$, we obtain the subset of venues $\mathcal{V}_{cl} \subset \mathcal{V}$ that fit into the cluster spatial area defined by the cluster's convex hull h. On the basis of this subset, we get the n_{cat} most frequent venue categories of the cluster, $c_{cl} = \{c_1, c_2, .., c_{n_{cat}}\}$. As a result, we extend each cluster $cl \in \mathcal{CL}$ by appending this category ranking. Thus, each *labelled* cluster $cl_l \in \mathcal{CL}_l$ is regarded as a tuple $cl_l = \{id, h, c_{cl}\}$.

[7] https://foursquare.com.

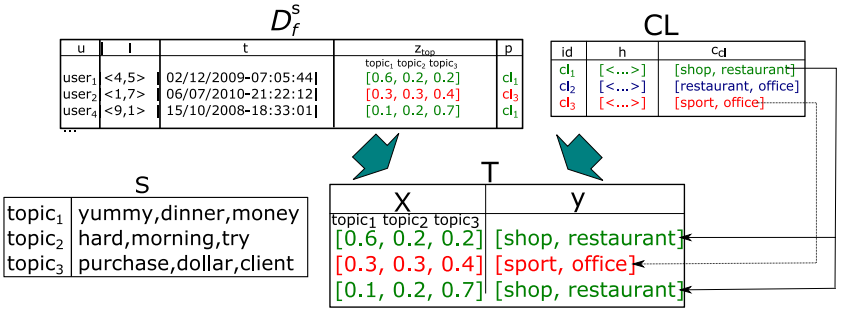


Fig. 4. Example of generation of the training dataset $\mathcal{T}$.

3.7. Classifier generation

All the previous sections are devoted to prepare the necessary data to compose a multi-label classifier model in charge of inferring the land use of new areas of interest in the target city.

This classifier takes as input a VGI document and, on the basis of its textual content z, it infers the n_{cat} land categories of the region comprising such a document. To do that, the content z of the input document should be firstly translated to a numeric vector z_{top} comprising the weights of the set of topics $\mathcal{S}$ in the document as described in Section 3.5.

In order to fit this model, a training dataset $\mathcal{T}$ is created comprising the topics' weights of all the documents in $\mathcal{D}_f^s$. More in detail, the attribute z_{top} of each document in $d_f^s \in \mathcal{D}_f^s$ gives raise to a new sample in $\mathcal{T}$. These attributes are the input variables of the model ($\mathcal{X}$). As for the output variable (y) of each sample, it is the top-n_{cat} categories of the cluster associated to the document d_f^s. As explained in Section 3.6, this information is contained in the dataset $\mathcal{CL}$. Consequently, the predefined labels that the model might potentially generate as output amounts to the set comprising the different land categories contained in all the clusters of $\mathcal{CL}_l$. In this manner, the resulting model will infer n_{cat} land categories coming from this set.

Fig. 4 shows an example of this process where three documents in $\mathcal{D}_f^s$ generate three different samples in $\mathcal{T}$ considering that each cluster in $\mathcal{CL}$ comprises its top-2 land use categories. As we can see, the first and third samples (shown in green in the figure) are linked to categories *shop & restaurant*. These are the categories of the cluster cl_1 which, in turn, comprises the original documents in $\mathcal{D}_f^s$. For the sake of completeness, the set of topics $\mathcal{S}$ is also depicted where each topic contains its top-3 keywords.

Finally, the predictions of the classifier are aggregated by means of a count procedure in order to infer the top n_{cat} of the target region. This mechanism is described in Algorithm 1. As we can see, when we want to label a new region r of a city which might have been previously detected by means of a clustering process, the mechanism takes all its VGI documents $\mathcal{D}^r$ as input. Next, the topics associated the textual corpus z of each document $d \in \mathcal{D}^r$ is extracted (line 3 of Algorithm 1). Then, such topics feed the classifier m that infers their top n_{cat} land use categories $\mathcal{C}_r^d$ (line 4). After that, these topics are aggregated into a single list $\mathcal{C}_r$. Finally, the mechanism selects the n_{cat} most frequent categories from such a list (line 6) which are eventually returned as the predicted output.

Algorithm 1: General mechanism for land use labelling based on multi-label prediction.

Input: Set of VGI documents of a untagged region r, $\mathcal{D}^r$, multi-label classifier m, its associated LDA model LDA_m and the number of categories to detect n_{cat}.
Output: Set of the top n_{cat} categories of region r, $\mathcal{C}_r$.

1 $\mathcal{C}_r \leftarrow \emptyset$
2 **for each** $d \in \mathcal{D}^r$ **do**
3 $\quad z_{top} \leftarrow LDA_m(d.z)$
4 $\quad \mathcal{C}_r^d \leftarrow m(z_{top})$
5 $\quad \mathcal{C}_r \leftarrow \mathcal{C}_r + \mathcal{C}_r^d$
6 $\mathcal{C}_r \leftarrow select(\mathcal{C}_r, n_{cat})$
7 **return** $\mathcal{C}_r$

Having explained the whole methodology, we now show how it is instantiated in a real-world scenario. The details are put forward in Section 4.

4. Evaluation of the methodology

In this section we provide a comprehensive evaluation of the methodology using real data.

4.1. Implementation details

This evaluation has been implemented using python 3.6 as programming language integrating scikit-learn,[8] nltk,[9] gensim[10] and scikit-multilearn (Szymański & Kajdanowicz, 2017) as third-party libraries. More specifically, we have imported from scikit-learn the single-label classifiers used in the proposal. The nltk library has been applied in the word stemming and stop words removal stage. From gensim, we have used its inner implementation of LDA. Finally, scikit-multilearn has provided the set of mechanisms to develop the multi-label solution.

The evaluation tests have been launched on a PC with the Operating System Windows 7 Professional Edition, 16GB of RAM and CPU Intel Core i7 3,40GHz.

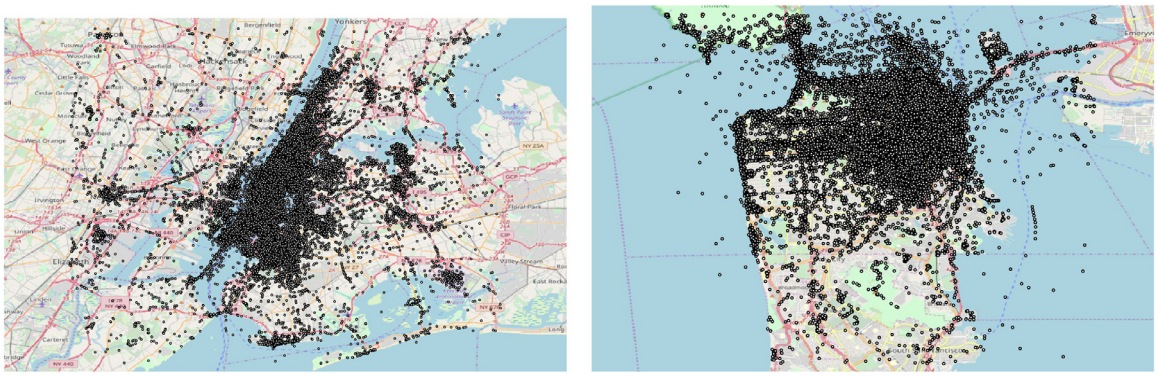
4.2. Datasets

For this study we collected data from two large metropolises, New York (NY) and San Francisco (SF), two of the most dense areas of the world in terms of population and business presence. For each city, the two types of OSN datasets mentioned in Section 3.2 have been collected.

4.2.1. User-centered dataset

For the user-centered dataset, we have used the documents contained in the Yahoo Flickr Creative Commons 100M public repository (Thomee et al., 2016). In this way, we just kept for each city the geo-tagged documents from the repository that fit into

[8] https://scikit-learn.org/stable/.
[9] https://www.nltk.org/.
[10] https://radimrehurek.com/gensim.



(a) NY Dataset (b) SF Dataset

Fig. 5. Spatial distribution of user-centered documents.

Table 3
Evaluation datasets.

Feature	San Francisco (SF)	New York (NY)
Time period	01/01/2008 → 31/12/2009	01/01/2008 → 31/12/2009
Covered area (km^2)	600	1214
Geo-tagged docs/users	264655/3934	288633/5670

Table 4
Target categories from the Foursquare hierarchy.

Categories	Arts & Entertainment (A&E)
	Shop & Service (S&S)
	Food (F)
	Outdoors & Recreation (O&R)
	Professional & Other Places (P&O)
	Travel & Transport (T&T)
	Residence (R)
	College & University (C&U)
	Nightlife Spot (NS)
	Event (E)

the spatial polygon defined for each city in OSM covering a two-year period. Fig. 5 depicts the spatial distribution of the documents in both cities and Table 3 summarizes the details of these two datasets.

4.2.2. Venue-based dataset

In order to obtain this dataset we have made use of Foursquare. This platform was created in 2008 and has attracted more than 50 million users.[11] In this platform, which was initially launched as a game, users can check-in at venues to inform about their whereabouts. In addition to that, users can inform about new venues that are not included in the Foursquare catalogue.

One of the key features of this platform is that users can specify the category of the reported venues on the basis of a predefined taxonomy of categories.[12] This taxonomy classifies venues (like bars, restaurants or grocery stores) based on their nature of activity. More in detail, it takes the form of a four-level category with 10 root categories and more than 300 categories at the second level.

For this study, we have used such root categories in order to label city regions (see Table 4). The rationale behind this decision is that the deeper we go in the hierarchy, the more sparse are the categories. This might cause the resulting classifier to not generate land-use labels descriptive enough.

In order to access such Foursquare data, we developed an ad-hoc crawler to collect venues of NY and SF contained in the same spatial regions used for the Flickr dataset. Fig. 6 shows the distribution of categories in each city. As it can be seen, such distributions are quite similar in both cities with a remarkable number of venues related to *Food* and *Shop & Service* categories.

4.3. Cluster detection and venues mapping

Once we have collected the datasets, we apply the HDBSCAN algorithm to the Flickr dataset to uncover the target regions in both use cases.

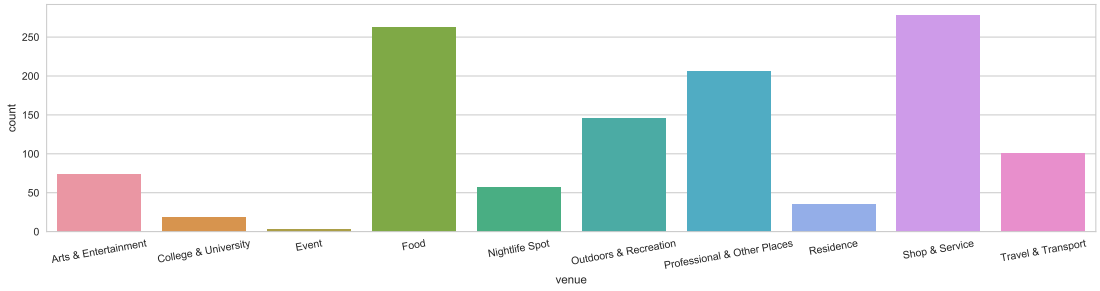
In order to correctly perform this step, we should carefully set the HDBSCAN *minSize* parameter that determines the smallest size grouping that must be considered a proper cluster. For its configuration, we made use of the silhouette score (Rousseeuw, 1987). In brief, this cluster validity measurement is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. Fig. 7 shows the evolution of such a score using different values for *minSize*.

According to such score, the most suitable *minSize* value for both cities is 50m. After launching HDBSCAN with this value in both cities, 1249 preliminary clusters where created in NY and 471 in SF. However, we observed a meaningful number of *spurious* clusters that do not represent potential regions of interest of the cities. Such clusters were actually generated by spam users that posted a large number of Flickr documents at the same spatial location during a short period of time giving raise to a false density of documents in such a location.

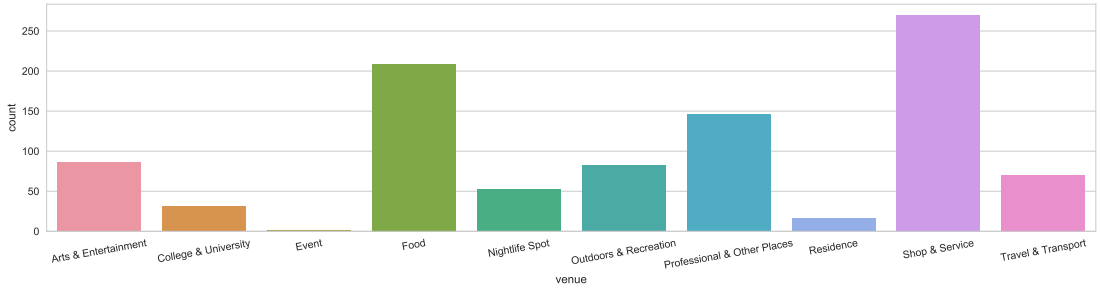
Consequently, we removed clusters comprising documents from less than 20 different users and covering a temporal range below 30 days. In this way we ensure that the resulting clusters actually stand for meaningful spatial regions of a city as they are visited by many different people during a large time period. As a result of this filtering procedure, we finally identified 248 regions in NY and 217 in SF.

Next, we mapped the obtained venues from Foursquare to each region. In this case, we opted for setting the parameter n_{cat} to 3,

[11] https://foursquare.com/about.
[12] https://developer.foursquare.com/docs/resources/categories.

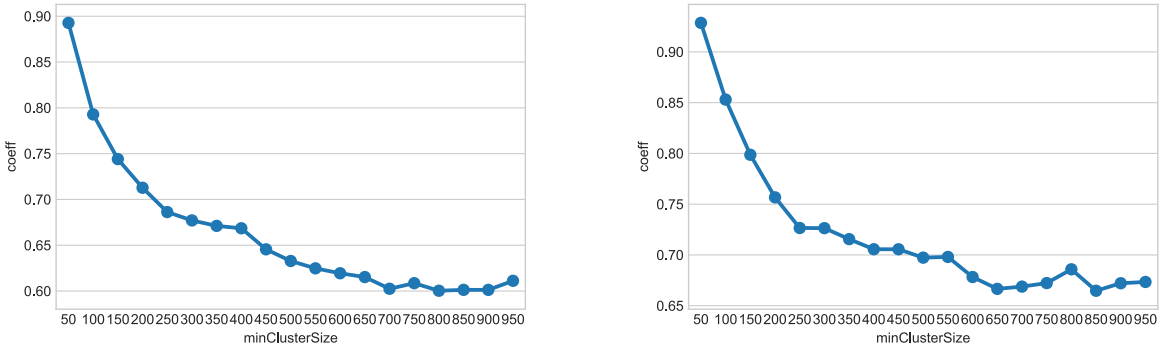


(a) NY Venues Dataset



(b) SF Venues Dataset

Fig. 6. Distribution of venue categories.



(a) NY Dataset (b) SF Dataset

Fig. 7. Cluster validity study based on the Silhouette score.

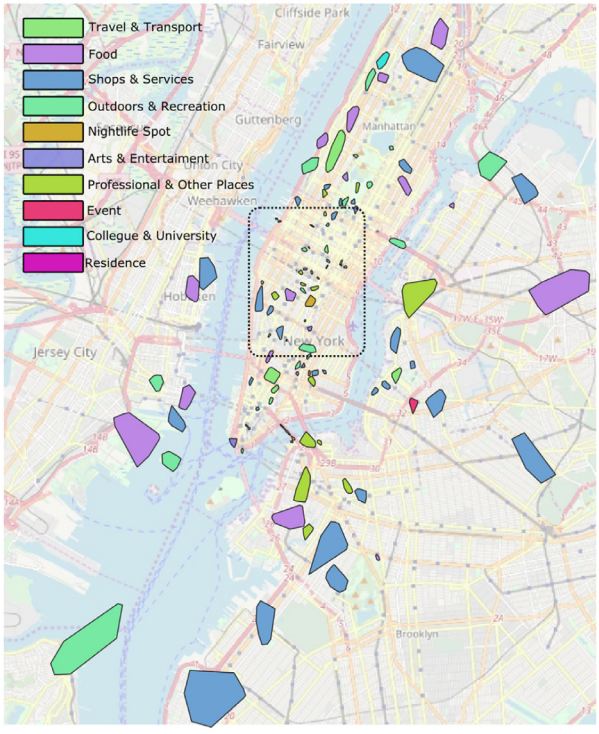
so we extracted the 3 most frequent categories within each cluster as put forward in Section 3.6. The resulting labelled regions are shown in Fig. 8. Observing this figure, we encountered two interesting phenomena.

In the first place, clusters at outskirts cover spatial areas larger than clusters at the city center (see Fig. 8a,c). This is because the density of Flickr documents is much higher at the center of cities, and therefore HDBSCAN is capable of distinguishing a large variety of clusters. On the contrary, the density of documents at suburbs is much lower and this makes HDBSCAN to *fuse* documents into single groups.

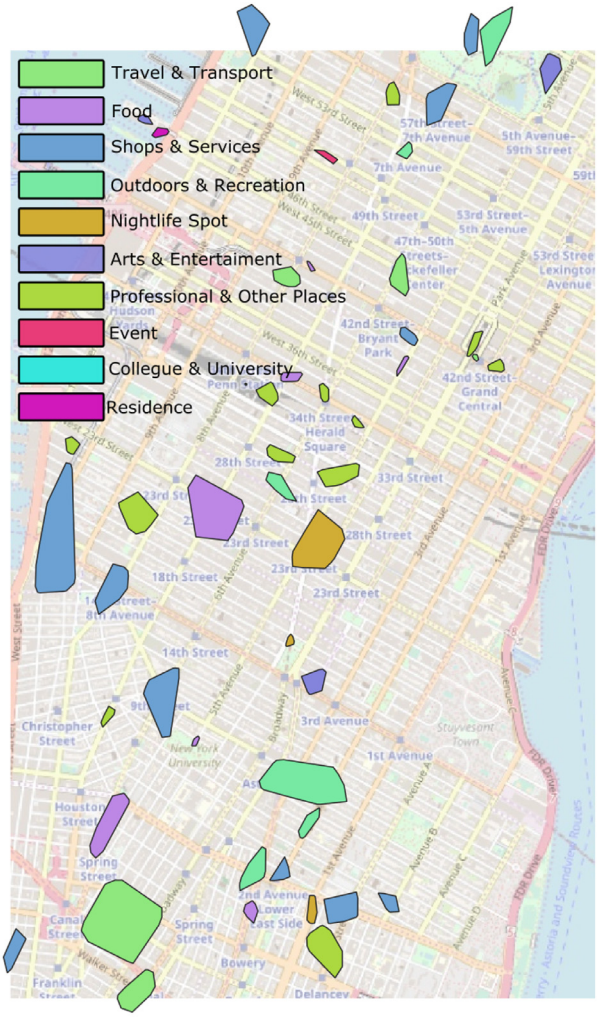
Secondly, the aforementioned external large clusters of both cities are mainly dominated by the *Food, Shop & Service* and *Travel & Transport* categories (see Fig. 8a,c). This is consistent with the natural distribution of modern cities where shopping malls –comprising a large variety of restaurants and shops– or transport hubs are located at the city outskirts. Nevertheless, city-center clusters exhibit a more varied palette of categories like *Nightlife Spot, Professional & Other Places* or *Outdoors & Recreation* (see Fig. 8b,d). This is also consistent with the urban planning of modern cities where many company headquarters are located at downtown areas along with varied and highly attractive cultural and leisure offerings.

We studied deeply the relationships among categories associated to clusters. To do that, we constructed a label graph per city comprising the target categories in Table 4 as nodes. Then, based on the co-existence of two labels in the same cluster, we connected both of them in the graph. Besides, the graph edges were weighted based on the number of samples labeled with these two labels. Fig. 9 shows the two resulting graphs.

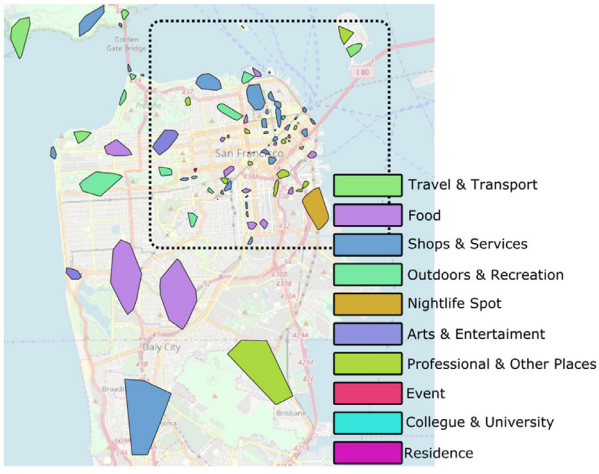
As it can be observed from these graphs, the strongest relationship in both cities occur between *Travel & Transport* and either *Professional & Other Places* or *Shop & Services* categories. The rationale of this dependence may be the fact that big transport hubs like international airports are usually surrounded by business parks or include a large number of shops in their premises. Another remarkable interconnection occurs between the *Nightlife Spot* and *Outdoors & Recreation* categories. This gives insight into the fact that most venues like clubs or bars are usually located near parks areas. This may be due to the comfortable



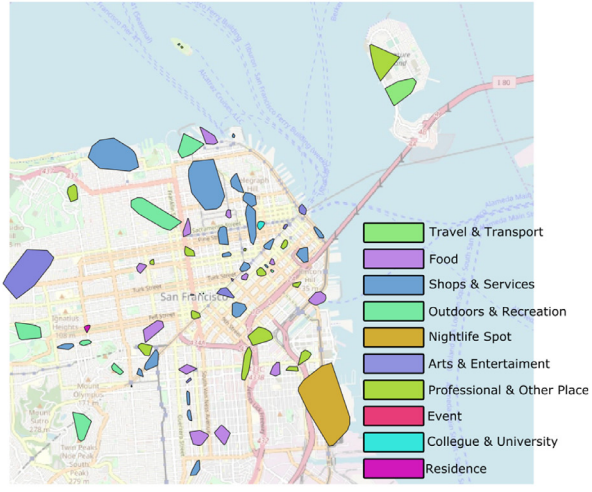
(a) NY Clusters



(b) NY city-center clusters

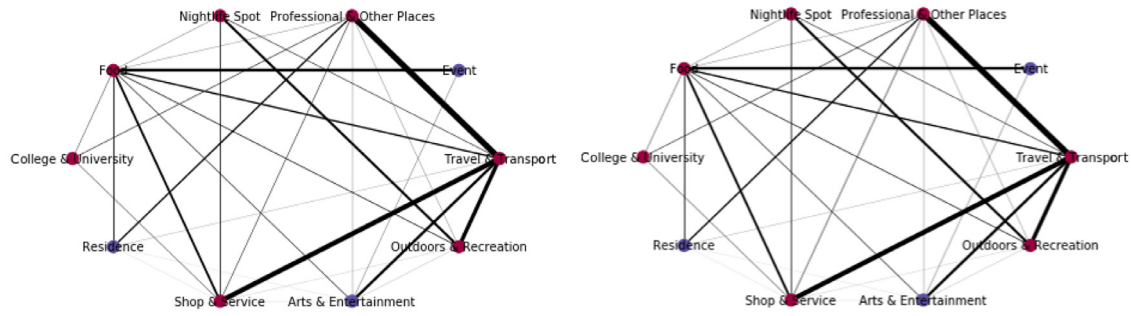


(c) SF Clusters



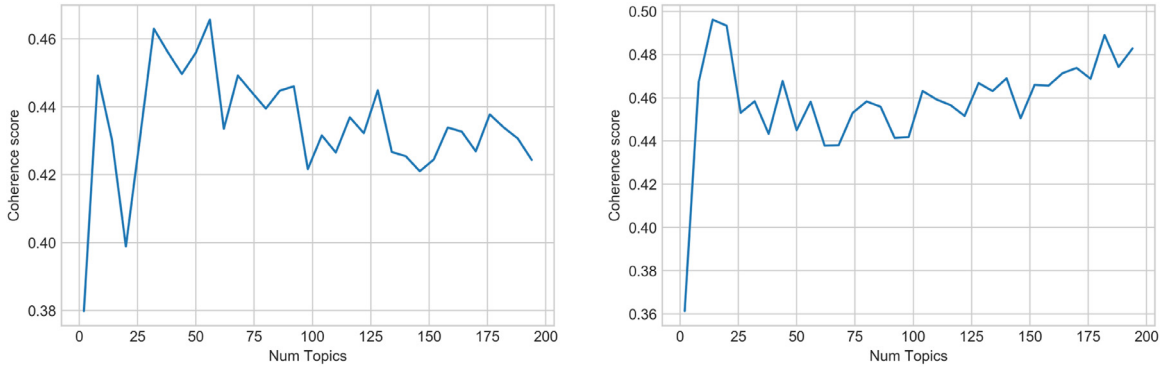
(d) SF city-center clusters

Fig. 8. Final clusters for both cities labelled with their most popular venue categories. Each cluster is represented by its convex hull and coloured depending on its popular category. (a) Global view of all NY clusters. The city center is marked with a rectangle. (b) Detail of clusters in the NY city center. (c) Global view of all SF clusters. The city center is marked with a rectangle. (d) Detail of clusters in the SF city center.



(a) NY (b) SF

Fig. 9. Category graphs of the target cities. Edges are weighted based on co-occurrence of categories in the same cluster. The thicker the edge, the higher the co-existence of two categories in the same cluster.



(a) NY (b) SF

Fig. 10. Coherence value for each generated LDA model per number of topics and city.

environment which this type of areas usually provides in an urban context.

Finally, the clusters generation also allowed to filter the Flickr documents as described in Section 3.3. As a result, we kept 32,441 out of 288,633 in NY (89% of removed documents) and 82,790 out of 264,655 documents at SF (72% of removed documents). These filtered documents constitute the $\mathcal{D}_f$ dataset actually used to train and test the multi-label classifier.

4.4. Cluster topic modelling

In order to extract the latent topics from $\mathcal{D}_f$ with an LDA model, as put forward in Section 3.4, one paramount parameter of such a model is the target number of topics n_{top}. Such parameter must be set beforehand.

Our approach to finding the optimal n_{top} was based on the coherence measurement for textual corpus stated in Röder, Both, and Hinneburg (2015). In brief, it is proposed a framework for coherence definition based on correlation of words. Consequently, we built several LDA models with different values of number of topics and selected the one that gave the highest coherence value. Fig. 10 shows the different coherence values for each generated LDA model. From these results, we concluded that, in the case of NY, choosing 56 topics was the most coherent option whereas 14 topics was the most suitable value of n_{top} for SF. These were the numbers of topics contained in the dataset $\mathcal{S}$.

Using the aforementioned values, we generated the training dataset $\mathcal{T}$ as explained in Section 3.7. As for NY, such dataset comprised 32,441 rows (one per filtered document) and 59 columns (56 related to the topic weights plus the top 3 venue categories). Concerning SF, $\mathcal{T}$ was defined as a 82790x17 matrix comprising the filtered documents as rows along with the 14 topic weights and the 3 venues categories per document.

4.5. Classifier generation

Once we composed the training dataset, we focused on generating the target single-label (or multi-class) classifier. This type of classifier, which only defines a dependent variable as output, acts as the building brick to compose the final multi-label model.

4.5.1. Selection of the base classifier

For the selection of the base classifier, we studied the three well-known single-label models as candidates, namely Random Forest (RF), Logistic Regression (LR) and Support Vector Machines (SVM).

These three models have been widely and successfully used in many different domains and scopes (Genuer, Poggi, Tuleau-Malot, & Villa-Vialaneix, 2017; Hosmer Jr, Lemeshow, & Sturdivant, 2013; Nayak, Naik, & Behera, 2015). Furthermore, each one is based on very different approaches in order to carry out the classification task (James, Witten, Hastie, & Tibshirani, 2013). Hence, RF is mainly based on ensembles of decision trees using if-then rules that sample the input space. SVM works by defining hyperplanes to separate data into different classes. Finally, LR defines a logistic model describing the relationship between one dependent binary variable and one or more independent variables. These three approaches work well with tabular data (Genuer et al., 2017). It is worth mentioning that this is the format of our training dataset $\mathcal{T}$ described in Section 3.7.

Table 5
F1-scores of the single-label model per target category in NY. The best score per venue is shown in bold. The rightmost column shows the average score per model whereas the last row shows the average score per category.

Model	Categories										Avg.
	A&E	S&S	F	O&R	P&O	T&T	R	C&U	NS	E	
RF	**0.70**	**0.66**	**0.63**	**0.73**	**0.75**	**0.72**	**0.71**	**0.73**	**0.67**	**0.90**	0.71
LR	0.23	0.37	0.39	0.47	0.42	0.42	0.41	0.49	0.41	0.83	0.41
SVM	0.65	0.64	0.62	0.72	0.73	0.69	0.68	0.69	0.64	**0.90**	0.68
Avg.	0.53	0.56	0.55	0.64	0.63	0.61	0.60	0.64	0.57	0.88	0.60

Table 6
F1-scores of the single-label model per target category in SF. The best score per venue is shown in bold. The rightmost column shows the average score per model whereas the last row shows the average score per category.

Model	Categories										Avg.
	A&E	S&S	F	O&R	P&O	T&T	R	C&U	NS	E	
RF	**0.71**	**0.63**	**0.68**	**0.71**	**0.60**	**0.69**	0.42	**0.68**	**0.73**	**0.66**	**0.67**
LR	0.25	0.20	0.44	0.10	0.24	0.18	0.10	0.25	0.26	0.26	0.25
SVM	0.61	0.55	0.60	0.62	0.48	0.61	**0.44**	0.58	0.63	0.60	0.59
Avg.	0.52	0.46	0.57	0.48	0.44	0.49	0.32	0.50	0.54	0.51	

We trained and evaluated each model with the dataset $\mathcal{T}$ following a 5-fold cross-validation approach. Furthermore, we launched several times each model with different configurations so as to discover the optimal parameter settings. In order to evaluate the classifiers, we use the F1 score as measurement. This score is calculated following the present formula:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

where

$$recall = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$precision = \frac{True\ positives}{True\ positives + False\ positives}$$

In this case, we only used the top-1 category of each sample in $\mathcal{T}$ as dependent variable to train the models. Furthermore, it is worth noticing that true positives, false negatives and false positives are defined at document level instead of a region level. This is because the key task of single-label classifier is to accurately classify documents. Then, the outcome of the multi-label classifier is used to generate the final prediction of venues at region level.

Tables 5 and 6 show the scores of each candidate when configured with its optimal parameters for each city. Besides, Appendix A shows the optimal configuration of the models.

It is observed that RF clearly overcame the other two classifiers in both cities. In particular, it obtained the highest F1-score for all categories but *Residential* in SF (see Table 6) with an average F1-score of 0.71 in NY and 0.67 in SF. In the light of these results, we eventually selected RF as the base model to compose the multi-label classifier in the two cities.

4.5.2. Selection of the multi-label composer

In order to compose the final multi-label classifier, we considered the four multi-label approaches already discussed in Section 2.4. As for the base classifier, we evaluated these four mechanisms on the basis of the F1 score. For this evaluation, we used the same datasets than the ones applied for the base classifier selection described in the previous section. Hence, the label distribution is the same than the one in Fig. 6.

Tables 7 and 8 show the results in NY and SF. These two tables represent the accuracy of the candidates with respect each single category. Since they are multi-label classifiers a true positive occurs when a predicted category is within the top-3 categories associated to the input document.

From these results, we could see that BR, CC and LP exhibited quite similar F1 scores at document level with an average value of 0.82 in SF and around 0.83 in case of NY. On the contrary, the MLkNN classifier achieved slightly poorer results than their competitors in both cities. In order to clarify the most suitable model, we opted for evaluate the accuracy of the candidates at a region level following the procedure described at Section 3.7.

As we can see from Fig. 11, BR, CC and LP also had very similar results in terms of region prediction. For example, the three candidates correctly detected around 0.93 of the categories regardless of their relevance in both cities (see column *totalCats* in Fig. 11). Besides, in terms of percentage of covered venues, the three approaches achieved similar results in the two cities as well (see column *%venues* in Fig. 11). In brief, roughly 0.74 of the NY venues and around 0.79 of the SF ones had their associated category in some of the predicted categories of the model.

However, if we took into account the relevance of the categories, LP was the most accurate model in order to predict the first and third region categories in NY (see columns *category1* and *category3* in Fig. 11a). In particular, LP predicted around 35% of the top-1 categories and 40% of the third most important categories. As for SF, BR also achieved the best results for the first and third categories of the regions. In detail, BR was capable of prediction 45% of the top-1 categories and 41% of the top-3 categories (see columns *category1* and *category3* in Fig. 11b).

All in all, we eventually chose the LP-based model for NY and BR-based model for SF as the final predictors for each city.

4.5.3. Inter-city evaluation

Finally, we evaluated the classifiers generated in a city using the documents of the other city. The rationale of this inter-city evaluation is to study the suitability of exporting models across cities. Fig. 12 shows the results of this study.

The results show that directly exporting a model from one city to another meaningfully degrades its accuracy. More in detail, when we used the BR-based model from SF in NY (SF-NY bars in Fig. 12) the accuracy of the model decreased more than 50% with

Table 7
F1-scores of the multi-label candidates per target category in NY. The best score per venue is shown in bold. The rightmost column shows the average score per model whereas the last row shows the average score per category.

Model	Categories										Avg.
	A&E	S&S	F	O&R	P&O	T&T	R	C&U	NS	E	
BR	0.75	0.84	**0.93**	0.77	**0.86**	0.76	0.73	**0.68**	0.69	0.90	0.83
CC	0.75	0.84	**0.93**	**0.79**	**0.86**	**0.77**	0.72	**0.68**	0.69	**0.92**	0.83
LP	**0.76**	**0.85**	**0.93**	0.76	**0.86**	**0.77**	**0.74**	**0.68**	**0.71**	0.91	**0.84**
MLkNN	0.69	0.80	0.91	0.74	0.82	0.71	**0.74**	0.61	0.62	0.87	0.80
Avg.	0.74	0.83	0.93	0.77	0.85	0.75	0.73	0.66	0.68	0.90	0.83

Table 8
F1-scores of the multi-label candidates per target category in SF. The best score per venue is shown in bold. The rightmost column shows the average score per model whereas the last row shows the average score per category.

Model	Categories										Avg.
	A&E	S&S	F	O&R	P&O	T&T	R	C&U	NS	E	
BR	**0.76**	**0.93**	**0.89**	0.70	**0.78**	**0.75**	**0.69**	0.55	**0.75**	**0.53**	**0.82**
CC	0.75	0.92	0.88	0.70	**0.78**	0.74	**0.69**	0.57	0.74	**0.53**	**0.82**
LP	0.75	0.92	0.88	**0.71**	0.77	**0.75**	**0.69**	**0.58**	0.74	0.51	**0.82**
MLkNN	0.71	0.91	0.87	0.65	0.77	0.70	0.63	0.52	0.69	0.32	0.79
Avg.	0.74	0.92	0.88	0.69	0.78	0.74	0.68	0.56	0.73	0.47	0.81

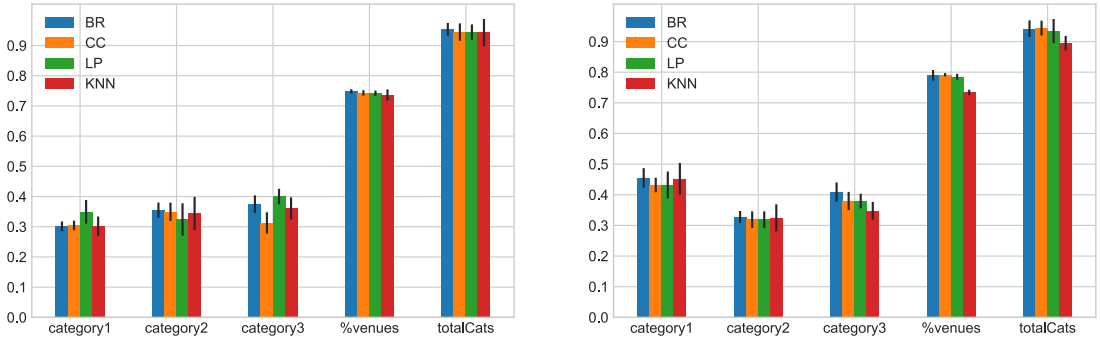


Fig. 11. Accuracy of the multi-label models at region level. Columns *category1/2/3* represent the percentage of correctly predicted top 1, 2 and 3 categories of the clusters. Column *%venues* depicts the percentage of venues of the target region whose category is included in the prediction outcome. Column *totalCats* shows the rate of correctly predicted categories regardless of their relevance.

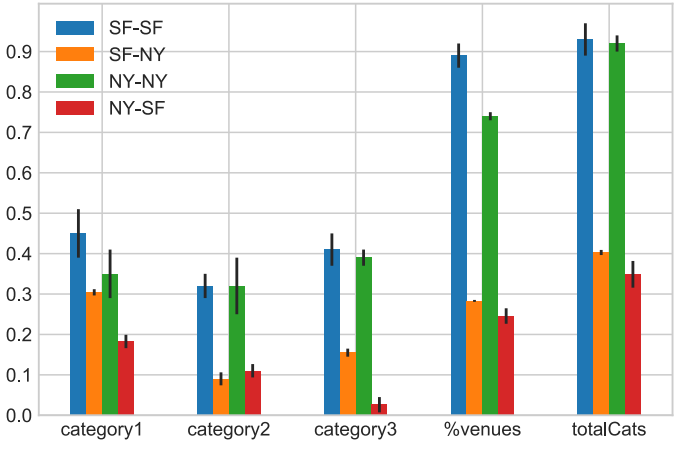


Fig. 12. Accuracy the final models in different cities. Each key of the figure legend stands for the training city and the target city (e.g., SF-NY bar depicts the evaluation of the SF model using the VGI documents from NY).

respect its *home* city (SF-SF bars). A similar behaviour can be observed in case of the LP-based model of NY.

4.6. Lessons learned

Some interesting remarks can be drawn after analyzing the experimentation results:

- Firstly, the comparative among the four multi-label candidates has shown that there is a quite high variability in the potential predictability of a land-use category. According to Tables 7 and 8, the F1 score of the four competitors ranged from 0.90 for the *Event* category to 0.56 for the *College & University* category. This might be due to the fact that categories with high F1 score are composed of venues that somehow *induce* homogeneous textual content. This, in turn, gives raise to topics strongly related to such venues' categories. For example, *Event* venues (e.g. conventions halls, festivals facilities, parades and so forth) will generally imply users to post VGI documents with quite similar textual content. On the contrary, more heterogeneous environments such as university campuses are likely to generate more diverse VGI documents.
- Secondly, the final multi-label models have exhibited a quite high accuracy in both cities. As a matter of fact, the methodology was able to detect above 90% of the regions' land-use categories in both cities. Since these models only take as

input the latent topics discussed by users in the target regions, this proves the fact that there is a close correlation between the textual content of VGI documents and the actual location context of such documents in terms of land use.

- Next, the complexity of the final model meaningfully varies depending on the urban area under consideration. For example, the NY model took 56 independent variables as input whereas the SF model took 14. Since these are the number of latent topics previously detected by the LDA, we can conclude that such model complexity is directly related to the variety of the textual content in the VGI documents.
- Finally, the inter-city evaluation confirms the fact that models strongly depends on the particular urban area where they have been generated and they cannot easily applied in other cities. Due to the fact that such models are based on latent topics, we can state that topics can not be smoothly exported from one city to another.

5. Conclusion and future work

The present work proposes a new methodology based on the analysis of information of two different types of Volunteer Geographic Information (VGI), namely text-based social networks and location-based social networks. A classifier model is generated by combining the information extracted from these sources, which can then be used to infer new knowledge related to the venues of city areas where there is only information available from text-based social network. The results show that the proposed multi-label classifiers were able to correctly predict more than 90% of the top-3 land-use categories of the regions in the two use cases of the cities of New York and San Francisco.

In this scope, VGI sources are giving birth to a novel, cheap and effective alternative to gain knowledge on the current state of affairs in the cities. Among the applications of such information, it is particularly noteworthy the automatic discovery of the land usage in any metropolis. This information could be of great help to local authorities when reorganizing urban planning and providing new services to the neighbourhood, as well as for citizens when looking for updated information about the area they are located in.

Current approaches on the use of VGI sources for land-use discovery suffer from two debilities: lack of interoperability on the labelling of the generated land-use maps and under-utilization of information sources, only focusing on a sole text-based online social network. In that sense, the proposed methodology alleviate these two problems. Firstly, the problem of interoperability is overcome by using a well-known, fine-grained categorization of venues provided by the adopted location-based social network. Secondly, we make use of two different types of online social networks feeds (text-based and location-based) instead of a single one in order to develop the final classifier.

Future studies could fruitfully explore this work further by three research lines. Firstly, the inclusion and analysis of images and videos in the user-centered documents (e.g., photographs in Flickr) could help to detect new topics and in a more accurate manner. Thus, image recognition techniques will be explored to include them in the cluster topic modelling stage. Secondly, a comparison with available open data about real estate property information (e.g, the Spanish property information web[13]) can be designed as an alternative to validate the results obtained by our proposal, and at the same time these results may be used by the local authorities to keep updated their records. Finally, the use of deep neural networks for Natural Language Processing (NLP) is demonstrating to be a promising research line for identifying topics in texts (Goldberg, 2016; Young, Hazarika, Poria, & Cambria, 2018). The integration of these deep neural networks in our proposal could be helpful to advance further the performance of VGI classification.

[13] http://www.catastro.meh.es/esp/acceso_infocat.asp.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Fernando Terroso-Saenz: Data curation, Formal analysis, Investigation, Methodology, Software, Writing - original draft, Writing - review & editing. **Andrés Muñoz:** Funding acquisition, Supervision, Methodology, Project administration, Writing - original draft, Writing - review & editing, Validation, Visualization.

Acknowledgements

This work was partially supported by the Fundación Séneca del Centro de Coordinación de la Investigación de la Región de Murcia under Project 20813/PI/18, and by Spanish Ministry of Science, Innovation and Universities under grant TIN2016-78799-P (AEI/FEDER, UE).

Appendix A. Single-label classifier configuration

(Table 9).

Table 9
Optimal setting of the single-label models for their evaluation.

Model	Parameter	NY value	SF value	Meaning
RF	$n_{estimators}$	3250	7750	The number of trees in the forest.
	max_{deep}	1000	30	Maximum deep of the each tree
LG	C	166.81	1	Inverse of regularization strength
	penalty	L1	L1	Norm used in the penalization
SVM	C	1000	1000	Penalty parameter of the error term
	kernel	RBF	RBF	Type of kernel
	gamma	1	1	Kernel coefficient

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X. Z., Raich, R., Hadley, S. J., … Betts, M. G. (2012). Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America, 131*(6), 4640–4650.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in knowledge discovery and data mining* (pp. 160–172). Berlin, Heidelberg: Springer Berlin Heidelberg.

Cao, C., Liu, Z., Li, M., Wang, W., & Qin, Z. (2018). Walkway discovery from large scale crowdsensing. In *Proceedings of the 17th ACM/IEEE international conference on information processing in sensor networks*. In *IPSN '18* (pp. 13–24). Piscataway, NJ, USA: IEEE Press. doi:10.1109/IPSN.2018.00009.

Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Quinta: A question tagging assistant to improve the answering ratio in electronic forums. In *IEEE EUROCON 2015 - international conference on computer as a tool (EUROCON)* (pp. 1–6). doi:10.1109/EUROCON.2015.7313677.

Chiara Renso, E. Z., & Stefano, S. (2013). *Mobility data - modeling, management, and understanding*. Cambridge. doi:10.1017/CBO9781139128926.

Chou, K.-C., Wu, Z.-C., & Xiao, X. (2011). Iloc-euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PloS One, 6*(3), e18258.

Cranshaw, J., Schwartz, R., Hong, J., & Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, & Z. Tufekci (Eds.), *Proceedings of the sixth international conference on weblogs and social media, ICWSM*. The AAAI Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using twitter activity. *Engineering Applications of Artificial Intelligence, 35*, 237–245. doi:10.1016/j.engappai.2014.06.019.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS, 21*(3), 446–467. doi:10.1111/tgis.12289.

García-Palomares, J. C., Salas-Olmedo, M. H., Moya-Gómez, B., Condeo-Melhorado, A., & Gutiérrez, J. (2018). City dynamics through twitter: Relationships between land use and spatiotemporal demographics. *Cities, 72*, 310–319. doi:10.1016/j.cities.2017.09.007.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., & Villa-Vialaneix, N. (2017). Random forests for big data. *Big Data Research, 9*, 28–46. doi:10.1016/j.bdr.2017.07.003.

Godbole, S., & Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In H. Dai, R. Srikant, & C. Zhang (Eds.), *Advances in knowledge discovery and data mining* (pp. 22–30). Berlin, Heidelberg: Springer Berlin Heidelberg.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research, 57*, 345–420. doi:10.1613/jair.4992.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221. doi:10.1007/s10708-007-9111-y.

Herrera, F., Charte, F., Rivera, A. J., & Del Jesus, M. J. (2016). Multilabel classification. In *Multilabel classification* (pp. 17–31). Springer.

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science, 2010*(1), 21–48. doi:10.5311/JOSIS.2010.1.3.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*: 398. John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*: 112. Springer.

Jensen, J. R., & Cowen, D. C. (1999). Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric Engineering and Remote Sensing, 65*, 611–622. doi:10.1002/9780470979587.ch22.

Jenson, S., Reeves, M., Tomasini, M., & Menezes, R. (2017). Mining location information from users' spatio-temporal data. In *2017 IEEE smartworld, ubiquitous intelligence computing, advanced trusted computed, scalable computing communications, cloud big data computing, internet of people and smart city innovation (smartworld/scalcom/uic/atc/cbdcom/iop/sci)* (pp. 1–7). doi:10.1109/UIC-ATC.2017.8397519.

Jiang, B., Ma, D., Yin, J., & Sandberg, M. (2016). Spatial distribution of city tweets and their densities. *Geographical Analysis, 48*(3), 337–351. doi:10.1111/gean.12096.

Kovacs-Gyri, A., Ristea, A., Kolcsar, R., Resch, B., Crivellari, A., & Blaschke, T. (2018). Beyond spatial proximityclassifying parks and their visitors in London based on spatiotemporal and sentiment analysis of Twitter data. *ISPRS International Journal of Geo-Information, 7*(9). doi:10.3390/ijgi7090378.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. doi:10.1080/01638539809545028.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems, 58*, 85–96. doi:10.1016/j.compenvurbsys.2016.04.002.

Lenormand, M., Picornell, M., Cantú-Ros, O. G., Louail, T., Herranz, R., Barthelemy, M., … Ramasco, J. J. (2015). Comparing and modelling land use organization in cities. *Royal Society Open Science, 2*(12). doi:10.1098/rsos.150449.

Liu, S. M., & Chen, J.-H. (2015). A multi-label classification based approach for sentiment classification. *Expert Systems with Applications, 42*(3), 1083–1093. doi:10.1016/j.eswa.2014.08.036.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social sensing: A New approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*(3), 512–530. doi:10.1080/00045608.2015.1018773.

Lopez-Ramirez, P., & Siordia, O. S. (2016). Multi-scale extraction of regular activity patterns in spatio-temporal events databases: A study using geolocated tweets from Central Mexico. *International conference on giscience short paper proceedings*: 1. doi:10.21433/B31101b1r10h.

Mahmud, J., Nichols, J., & Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology, 5*(3), 47:1–47:21. doi: 10.1145/2528548.

McKenzie, G., Liu, Z., Hu, Y., & Lee, M. (2018). Identifying urban neighborhood names through user-Contributed online property listings. *ISPRS International Journal of Geo-Information, 7*(10). doi:10.3390/ijgi7100388.

Nayak, J., Naik, B., & Behera, H. (2015). A comprehensive survey on support vector machine in data mining tasks: Applications & challenges. *International Journal of Database Theory and Application, 8*(1), 169–186.

Noulas, A., Mascolo, C., & Frias-Martinez, E. (2013). Exploiting foursquare and cellular data to infer user activity in urban environments. In *2013 ieee 14th international conference on mobile data management: 1* (pp. 167–176). doi:10.1109/MDM.2013.27.

Quercia, D., & Saez, D. (2014). Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing, 13*(2), 30–36. doi:10.1109/MPRV.2014.31.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier chains for multi-label classification. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 254–269). Springer.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on web search and data mining*. In *WSDM '15* (pp. 399–408). New York, NY, USA: ACM. doi:10.1145/2684822.2685324.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. doi:10.1016/0377-0427(87)90125-7.

Rudinac, S., Zahálka, J., & Worring, M. (2017). Discovering geographic regions in the city using social multimedia and open data. In L. Amsaleg, G. Gumundsson, C. Gurrin, B. Jónsson, & S. Satoh (Eds.), *MultiMedia modeling, lecture notes in computer science: 10133* (pp. 148–159). Cham: Springer International Publishing. doi:10.1007/978-3-319-51814-5_13.

Salim, F., & Haque, U. (2015). Urban computing in the wild: A survey on large scale participation and citizen engagement with ubiquitous computing, cyber physical systems, and Internet of Things. *International Journal of Human-Computer Studies, 81*, 31–48.

Spyratos, S., Stathakis, D., Lutz, M., & Tsinaraki, C. (2017). Using Foursquare place data for estimating building block use. *Environment and Planning B: Urban Analytics and City Science, 44*(4), 693–717. doi:10.1177/0265813516637607.

Szymański, P., & Kajdanowicz, T. (2017). A scikit-based Python environment for performing multi-label classification. arXiv:1702.01460v1.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association, 101*(476), 1566–1581. doi:10.1198/016214506000000302.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., … Li, L.-J. (2016). YFCC100M: The new data in multimedia research. *Communications of the ACM, 59*(2), 64–73. http://doi.acm.org/10.1145/2812802.

Tsoumakas, G., & Vlahavas, I. (2007). Random k-Labelsets: An ensemble method for multilabel classification. In J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Machine learning: Ecml 2007* (pp. 406–417). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, S., Wang, Z., & Ji, Q. (2015). Multiple emotional tagging of multimedia data by exploiting dependencies among emotions. *Multimedia Tools and Applications, 74*(6), 1863–1883. doi:10.1007/s11042-013-1722-3.

Wartmann, F. M., Acheson, E., & Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: An interdisciplinary approach. *International Journal of Geographical Information Science, 32*(8), 1572–1592. doi:10.1080/13658816.2018.1445257.

Yang, D., Zhang, D., & Qu, B. (2016). Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Transactions on Intelligent Systems and Technology, 7*(3), 30:1–30:23. http://doi.acm.org/10.1145/2814575.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine, 13*(3), 55–75. doi:10.1109/MCI.2018.2840738.

Yuyun, Akhmad Nuzir, F., & Julien Dewancker, B. (2017). Dynamic land-use map based on twitter data. *Sustainability, 9*(12). doi:10.3390/su9122158.

Zhang, M.-L., & Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition, 40*(7), 2038–2048. doi:10.1016/j.patcog.2006.12.019.

Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology, 5*(3), 38:1–38:55. http://doi.acm.org/10.1145/2629592.

Zheng, Y., Mascolo, C., & Silva, C. T. (2017). Guest editorial: Urban computing. *IEEE Transactions on Big Data, 3*(2), 124–125. doi:10.1109/TBDATA.2017.2699838.

Zhou, X., & Zhang, L. (2016). Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartography and Geographic Information Science, 43*(5), 393–404. doi:10.1080/15230406.2015.1128852.