Exploring different combinations of data and methods for urban land use analysis: a survey

Renato Andrade [1,2]
renatoandrade@dei.uc.pt

Ana Alves [1,2]
ana@dei.uc.pt

Carlos Bento [1]
bento@dei.uc.pt

[1] CISUC, Centre for Informatics and Systems, University of Coimbra, Polo II, 3030-290 Coimbra, Portugal
[2] Coimbra Institute of Engineering, Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal

Abstract

Modern planning and management of urban spaces is an essential topic for smart cities and depends on up-to-date and reliable information on land use and functional roles of the places that integrate urban areas. In the last years, driven by increased availability of geo-referenced data from social or embedded sensors and remote sensing (RS) images, various methods become popular for land use analysis. This paper addresses the various methods employed in this context, as well as needed data and respective categorization, best applications of each method, and their comparison. We focus on approaches based on data from RS images, open maps and various categories of crowdsourced data. We identify and discuss the way these approaches use Data Mining (DM) and Machine Learning (ML). From our initial study we concluded that even using the same methods and the same kind of data, results depend on spatial configuration of the data, accordingly to the specificity of each region. The work described in this paper is intended to provide relevant contributions to method selection for knowledge discovery for city planning and management, taking into consideration available data and the pros and cons of each technique.

1 Introduction

With the recent and rapid development of cities, concerns with sustainability opened a new way for an essential field in recent studies: smart growth. In general, it is an effort for better management of natural resources, by reducing and controlling its consumption [S+16]. Because of this, the needs for urban land use planning and efficient management of urban areas have evidently become important [L+17]. These points are directly connected with the design and development of smart cities, converging to a common objective, which attempt to create a high quality of life for people in a more sustainable world. With attentions turned to urban spaces, land use analysis become an essential topic in this context.

Currently, urban spaces have also gained focus because of issues related to urban expansion, traffic control, well-being, population activity monitoring, construction projects, environmental preservation, hazard and pollution analysis, economic analysis, as well as public health care and others essential topics, which are all around smart growth and smart cities. These subjects need essentially fine-grained maps to design and manage the work [L+17, Z+17b]. However, as urban areas change, maintaining maps and information about infrastructures and functional zones up-to-date is a challenge that research teams and public administration face daily, given the complexity of modern urban systems [Z$^+$17b, Z$^+$17a].

Land cover is when a given region may be considered forest, agriculture, impervious surfaces, wetlands, other types of landscape and even water types, which includes open water or wetlands. In contrast, land use is high-related as the way as landscape is utilized by people, i.e., for conservation, development or mixed use [Ser18]. In addition, functional zones are the basic unit of urban areas. The concept of functional zones refers to spaces where human activities occur. The same functional zone could support a variety of functions depending on the types of land use. These types include residential, commercial or industrial use, business, etc. Moreover, the same functional zone could be used in various human activities, such as living, shopping, eating, recreation, among others [GJC17]. For Land Use and Land Cover (LULC) analysis, many authors usually adopt methods based in image interpretation, extracting information from image objects, which are scene components or meaningful entities in a given image [B$^+$14] (e.g., a tree, a house, a car parking or a vehicle).

Understanding how people use and interact with functional zones, and how these areas usually change becomes essential for land use analysis [Z$^+$17b, GJC17]. Because most researchers focus on land cover objects rather than large-scale functional zones, maps of this type of urban unit are hardly available. Besides functional zones being spatially larger than objects, they are also semantically different from them. For example, while a residential area is a functional zone, a building belongs to land cover objects. Because of these two types of units are in different semantic layers, traditional object-based methods cannot classify functional zones [Z$^+$17a]. Nevertheless, there are many studies based on RS images to provide a classification of urban areas through morphological analysis, by extracting spectral and textural characteristics for the representation of information for a given region [XM18]. This approach has been evolving significantly in recent years, since it allows somehow revealing land cover information related to the morphology of the area, given the presence, shape, size and even spatial distribution of buildings, including open spaces [BW06, GIM11, Q$^+$12].

In the field of land use and functional region analysis, when discussing about image interpretation, low-level semantic features can be described as information that comes with data such as physical properties (e.g. color and texture); and high-level semantic features are directly related to specific "knowledge" for each user and application [L$^+$17, Z$^+$17a]. Semantic gab usually refers to the disparity of features identified between low-level and high-level features. Using only low-level semantic features is probably low-accurate because different objects may have the same physical properties and identical objects may have different attributes. Adding high-level semantic features, referring to various attributes of the object given by the human operator in the classification, will probably archive better results. For example, a set of RS images where land cover objects (e.g., buildings), can be recognized based on low-level description: In this case, high-level information provides good features for functional type classification, such as residential, commercial and industrial areas [L$^+$17]. Moreover, the relationship between urban landscapes and how people use them is essential for identifying functional zones, considering that land use patterns are also affected by indoor lifestyles and other factors as well [XM18, Y$^+$17a] . Because remote sensing images do not provide high-level semantic features, it is necessary to aggregate other data sources to provide this possibility.

Using traditional remote sensing models is also difficult to classify land use with typical thematic features[Z$^+$17a]. Because of this concern, many authors suggest the use of a combination of different data types and methods [L$^+$17, Z$^+$18b]. In addition, driven by rapid technological development in recent years, several methods have emerged, based on new capabilities added by advances in Geographic Information Systems (GIS), Geospatial Big Data, RS images, and others [Y$^+$17a]. As a proposal for better classifying urban landscapes, some authors have suggested the use of data such as social information, socioeconomic features, Points Of Interest (POI) data, or location-based social network data along with remote sensing images to enable the construction of a more robust model [L$^+$17, Z$^+$17b, GJC17, XM18].

The main objective of this paper is to explore a set of scientific studies to address the most recent data types and techniques frequently used for knowledge discovery in context of LULC and urban functional regions. For this, a table is adopted as a method of comparison, highlighting, for each work, the type of data used and the set of methods, as well as their purposes.

The remainder of this paper is structured as follows: Section 2 discusses the key topics for classifying and extracting information for LULC analysis and identification of urban functional regions: (1) features extraction; (2) the most common data types and (3) the techniques adopted in recent studies. Section 3 presents a systematic comparison to promote a better understanding of the types of data used in conjunction with common methods, highlighting the purpose of each technique. Finally, section 4 presents the conclusions and suggestions of future work for this subject.

2 Knowledge Discovery in LULC and Urban Functional Regions

Given the importance of up-to-date information related to LULC and urban functional regions, many efforts have recently been made on this topic, making popular different types of data and various methods. In this section, we will discuss the most common categories of data and the most frequent methods employed by authors in this subject

2.1 Features Extraction

A feature is characterized as an attribute that represents certain information about a given object [Q^+12]. For effects of image classification, a feature can be referred as a pattern extracted from objects in the images. Features have an essential role in the field of data analysis in general and it is not different when talking specifically about image analysis. Therefore, during a typical image analysis flow, various processing techniques are usually applied on the work dataset before getting features (e.g. binarization, thresholding, resizing, normalization, etc.), making possible and, sometimes, simplifying the application of different methods for the extraction phase [K^+14]. Features extraction process often results in a set of attributes useful for image recognition and classification.

Due to the availability of a very large number of RS images driven by the recent development of modern technologies, new opportunities have emerged to extract urban LULC information with a high level of detail. However, features extracted from such category of images are very heterogeneous and highly complex due to the proximity of a mixture of artificial urban land and semi-natural surfaces. Often, the same type of land use can be characterized by a set of physical properties or land cover materials. On the other hand, different categories of land use may have the same physical characteristics [Z^+18a]. Thus, features extracted from remote sensing images are implicitly presented as patterns in which some kinds of object or identical low-level features are often related to different categories of land use.

Some authors, e.g. [ZZZ15], propose the introduction of high-level semantic features in image classification to improve accuracy. These semantic features, that can be extracted from user data for example, are a set of attributes linked to a given object, introduced by a human operator according to the type of use given to it. Such approach can be adopted in cases where land cover objects are identified based on the low-level features extracted from a set of RS images (e.g., buildings), and various functional types (e.g., commercial, residential and industrial areas) can be extracted from high level features, obtained from user data, e.g. activities on Location-Based Social Network (discussed in the next subsection).

When talking about features extraction, as mentioned before, a central question is related to the gap between low-level and high-level semantic features. Because filling this gap is not a trivial task, some studies, e.g. [XM18], recommend to employ socioeconomic information as a complementary data, to help improve results. Socioeconomic information can be used to fill some space between high-level and low-level semantic features, by adding extra information which was not available before. Moreover, socioeconomic information can be extracted from a variety of sources, including crowdsourced data, which often allow the possibility of classifying urban functional regions also from a human activity perspective.

2.2 Data

In this specific field, many techniques can be used based on different data types. An important task for researchers is improving the accuracy of the results generated by these methods. The integration of features extracted from various data types can to some extent show better results. In this section, we present the main data types frequently used for urban functional region extraction and LULC classification. The data types presented in this subsection were used in at least two studies, among the set of works analyzed during the paper preparation.

2.2.1 Remote sensing (RS) images

Several methods used to update urban land use and land cover maps, are based on the interpretation of aerial photos and field surveys, which are time-consuming and difficult. Due to the recent development of remote sensing technologies, a large amount of RS images is available through sensors installed in aircrafts or satellites [HZS18]. In addition, RS images are present in scientific datasets, in some cases provided by universities [Dur15], research centers [Z^+17b], government agencies [D^+19], among other organizations. This type of data is often useful for extracting land use information and generally adds the ability to identify lots of land used for various purposes (e.g. residential, commercial, or industrial). This identification is usually based on the physical properties of objects, with different characteristics, such as spatial distribution, color, texture, shape, etc. [L^+17, HZS18].

Using RS images, Yao et al. [Y+17b], proposed a model to classify urban land use, by combining object-based images and a convolutional neural network (CNN), considering irregular land-parcel level. Remote sensing images were also utilized by J. Song et al. [S+18], combined with Points of Interest (POI) and road network data.

2.2.2 Crowdsourced data

Due to the use of the most diverse type of applications, there are large amounts of data related to several domains. Some of them, mainly mobile apps, provide data such as **POI** [GIM11, Y+12, GJC17, Y+17a, L+18, Z+19a]; **text messages** [FMFM14, LL16, J+17, XM18, Z+19a, GP+18]; **check-in activities from Location-Based Social Networks (LBSN)** [C+11, GJC17, GP+18, XM18]; **collaborative mapping data** [AV15, L+17, Z+17b]. Around the world, there are 7000 million check-ins on Foursquare, 500 million tweets are posted and more than 80 million of photos are uploaded to Instagram, daily [GP+18]. These rich and diversified sources of data potentially provide information on human activities and socioeconomic information, which have been the central idea of many studies to indicate urban functions [XM18].

2.2.3 Taxi Trajectories

Beyond the mentioned categories of data, many works have also used with some frequency, data such as taxi trajectories [L+18, Y+15]. Taxi trajectories can easily provide pick-up and drop-off points, trip lengths and time of each trip. However, these points often do not represent the exact locations where users have their activities [G+16]. In the most cases, passengers exit their taxi a bit far from their final destination. Also, because of the information provided by taxi trajectories does not contain accurate indication of the passengers purposes in their activities, it is challenging to deal only with this kind of information. That is the main reason for combining taxi trajectories data with other data types, e.g. building blocks, points of interest or LBSN user information, to provide better results.

2.2.4 Building Blocks

Building blocks is often referred as street blocks and although it represents a different category from taxi trajectories, for example, it is often used as complementary information in many studies. Building blocks information is normally provided by local administration [Z+18b], but it is possible to extract this kind of data from remote sensing (RS) images. This technique was utilized for example, by Liu et al. [L+18], and the obtained building blocks were combined with social network records, taxi trajectories and POIs to characterize mixed-use buildings. Huang et al. [HZS18], also employed building blocks together with remote sensing images for urban land use mapping.

2.3 Methods

As mentioned previously, there are several methods used for spatial data analysis to provide knowledge discovery in this context. However, some of these methods are most commonly used in recent scientific studies. The following is a set of techniques in this category, which represent the methods most frequently encountered. The methods covered in this subsection have been used in at least two scientific studies among the set of works analyzed during the paper preparation stage.

2.3.1 Object-Oriented Classification (OOC)

Terms such as "object-oriented" and "object-specific" were adopted by a research area often referred as object-based image analysis (OBIA). This scientific area emerged after the first commercial software designed specifically for the design and analysis of "image objects", rather than individual pixels, based in remote sensing images [B+14]. By these concepts, scene components or entities are distinguishable objects in a given image, e.g. a tree, a house or a vehicle. Using an object-oriented approach, according to a specific user definition, pixel-based images are segmented into objects. Within each image object, the user-defined homogeneity is obtained during the segmentation process. To avoid a large growth of user-stablished heterogeneity, a pair of adjacent objects is merged at each step of the process. The process is interrupted if smaller growth exceeds the scale parameter [BW06]. Currently, this is one of the most popular methods, extracting land use patterns through the physical features of ground objects from images [Y+17a]. Although many studies use this method, object-oriented classification can only reveal land cover information based in low-level semantic features, where spatial relationships among ground objects are not considered.

2.3.2 Latent Dirichlet Allocation (LDA)

When analyzing abundant textual descriptions to discover thematic features and their respective structures, a large number of studies use probabilistic topic models and among them, the most common is LDA [GJC17]. It is an unsupervised model that works in a generative and probabilistic way implementing a bag-of-words approach, which means that the order of words in the document is not applicable. In LDA, the main idea is to represent documents as a distributed probability of latent topics, where each topic is a distribution over words. To simplify the concept, the probabilistic topic model, including LDA, can be generically described as a random mixture of topics [16]. LDA model is often used to extract socioeconomic information from crowdsourcing data, providing explicit descriptions of human activities, as implemented in [XM18].

2.3.3 K-means

Among many clustering algorithms, K-means is one of the most common in data mining [16]. As a type of unsupervised learning, K-means clustering is used for unlabeled data, i.e., when the categories are not defined. This algorithm works locating groups in the data, by using a parameter k that represents the number of groups. The clustering process is iterative and at each iteration the data points are assigned to one of k groups, based on their attributes, i.e., feature similarity is used for clustering the data points [Tre16]. Clustering techniques have been successfully applied by many authors for various proposes, such as defining areas or regions [J^+15], classifying features extracted from social media data [L^+17], analyzing correlations between points of interest and zones [Y^+17a], aggregating similar formal regions in terms of region topic distributions [Y^+12], etc.

2.3.4 Hierarchical Semantic Cognition (HSC)

HSC is a bottom-up Bayesian method with a hierarchical structure to classify urban functional zones [Z^+17a, Z^+18b]. It consists of four semantic levels: functional zones, patterns of spatial objects, categories of objects and visual features. In this model, using conditional probabilities, each level characterizes a relation between two semantic layers. Thus, the first level can for example model the relationship between functional zones and patterns of spatial objects. Typically, different objects generally have different distributions of visual features in the same spatial object pattern, whereas in the same object type, different patterns of spatial objects may exist and have small differences related to their distributions of visual features. HSC is used for LULC and functional zone classification, by using data such as remote sensing images, POIs and roadblocks.

2.3.5 Random Forest (RF)

RF is a bagging ensemble learning algorithm that works by building multiples decision trees, where each one is based on a random subsample of the training dataset [Z^+17b]. The model provides its results based on the class voted by most trees. As a tree-based ensemble method, this classifier can handle a high accuracy than single decision trees, such as Regression Tree (CART) or C4.5. In addition, in many cases, without the need to adjust numerous parameters, RF overcomes popular models, such as SVM. In this context, it is well established in the literature and is widely used in land use and in the classification of functional regions. Random Forest is used for example in [Z^+17a] and [Z^+19a].

2.3.6 Support Vector Machine (SVM)

In a general way, Support Vector Machine can be described as a supervised learning method that work as a discriminative classifier. It creates a hyperplane or a set of hyperplanes that allow classifying the inputs in a high-dimensional space, by spearing them. The algorithm outputs an optimal hyperplane, based on training data. This hyperplane is an N-dimensional space, where ”N” is the number of features used for training. For example, a hyperplane created for a two-dimensional space is a line splitting it into two different parts [Pat17]. It is a model based on the principle of structural risk minimization [ST17]. This method is used, for example, as a classifier in scene classification, to predict scene labels. The main idea of this technique is to train in kernel space, a linear learning classifier, able to overcome the problem of pattern classification [16], considering generalization and performance optimization. SVM was chosen for example, by Liu et al.[L^+17] for identifying urban land use types. The authors adopted SVM because it was suggested in previous studies (e.g. [ZD15] and [LZZ15]) that when working with high-dimensional features, this method have high-efficiency level as a classifier

2.3.7 Deep Convolutional Neural Network (DCNN)

One common approach for LULC classification is to use methods per field to directly extract or classify low-level features on the physical properties of images. These methods can add some advantages over the per-pixel or object-based methods. However, per-pixel object-based and per-field land use and land cover classification techniques are based on manual feature descriptors and shallow architectures, and cannot work with complex land-use images to capture fine features [HZS18]. Because this type of image is used for generalization, none of these methods reaches the level of accuracy generally required by practical applications. Land use can be described at many levels in LULC scheme, including the intensities of pixels, edges, objects, parts of object and parcels of land. Deep architectures can efficiently represent all these levels. Through the deep learning process, a group of machine learning algorithms aims to model high-level abstractions employing deep architectures, which is a composition of multiple nonlinear transformations. Deep learning model is a high promising approach to handle urban LULC classification problems, since it can model hierarchical representations of features that describe urban LULC schemes. DCNNs consist of several convolutional layers and can learn high level abstract features from the original pixel values of the images[ZZD16]. Among many deep learning methods, the DCNN technique has achieved a high level of performance in land use classification, based on remote sensing images.

3 Discussion

The use of traditional models for urban functional regions and LULC classification based on remote sensing images is not an easy task since the physical properties of regions have become sophisticated with the increased complexity of urban systems. As we can see in table 1, Many authors, e.g. [GJC17] and [L+18], focused on efficient fusion of the most diverse data types such as LSBN user activities, taxi trajectories, and POIs, along with images to improve results. Although several studies have focused on similar objectives in the same context, they often use different methods and data. For example, for land use and functional zone mapping, [Z+18b] and [HZS18] used both remote sensing images and road blocks, but the first applied different methods to segment and classify images and the latter adopted a convolutional neural network (CNN)-based approach. Similarly, many have suggested various approaches, including the use of crowdsourced data together with methods such as LDA [GJC17], Place2vec [Z+19a], Bayesian-based models [L+18], etc.

As we also can observe in table 1, the most common data type in this field are POIs. This category of data has been widely used by many authors, usually because it is directly related to the use of urban space by humans and can reveal peoples behavior. POIs are also often related to LBSN activities and therefore, both are used together in many studies, e.g. [L+17] and [Z+17b]. Moreover, various other types of crowdsourced data generated by citizens in their daily routines are frequently used in this context. Crowdsourced information is available in large amounts for many countries, encouraging their use in cases where datasets containing different information is not easily available, such as up-to-date urban plans, global positioning system (GPS)-based data (e.g. taxi trajectories) or urban topology at the building level.

Another type of data commonly used since a long time, mainly for land cover interpretation, is remote sensing images. It represents an essential piece in many cases and still used today. Although remote sensing images have some limitations in terms of land use, they can be combined with other types of data. This type of image was adopted in many cases, such as [D+19], [FZS19] and [Z+19b], partly due to recent advances in remote sensing technologies and their availability in a vast group of equipment, such as moderns airplanes, satellites or unmanned aerial vehicles. Another great example of use for remote sensing images, could be seen in [L+18], where authors used it to extract building footprints to be used together with POIs, taxi trajectories and LBSN data.

Regarding methods, due to the unavailability of ground truth data, many researchers have adopted the use of unsupervised techniques, where clustering techniques such as spectral clustering and KNN are frequent. However, the most commonly observed method in the group of studies we analyzed, is K-means clustering, given its simplicity and effectiveness for tasks such as grouping POIs, functional regions or even land parcels. Clustering methods are successfully employed in many cases, such as [L+17] and [Dur15]. In the first one, the authors used K-means to group features into different classes, while in the last one, it was reported that using this method together with some measurements calculated for each feature, as a hybrid data-weight method, they archived satisfactory results for classifying land cover data.

In cases where labeled data is available, many researchers have adopted supervised approaches. The most common for this context is Random Forest algorithm. One probable reason for that is the effectiveness of the technique as a classifier, in terms of balance between results and performance for purposes such as classifying

Table 1: Common data types and methods utilized for knowledge discovery in context of LULC and urban functional regions, during the last 5 years. This table is available online at `http://tiny.cc/0v7w5y`

Author	Year	Data	Method	Objective
V. Frias-Martinez, E. Frias-Martinez [FMFM14]	2014	Twitter activity	Self-Organizing Maps (SOM)	Land segmentation with geolocated data, for characterization based on its usage pattern
			Spectral clustering	Detect urban land uses
S. Zhan, S. Ukkusuri, F. Zhu [ZUZ14]	2014	LBSN user activities, Points of Interest (POI)	Laplacian Score (LS)	Feature selection
			Clustering (various algorithms)	Land use inference
			Naïve Bayes; support vector machine (SVM); random forest (RF)	Classify land use
Jiang et al. [J+15]	2015	Points of Interest (POI), Aggregate census employment data, Boundaries of towns	POI Matching algorithm	Map POI from one source to another
			Bayesian networks; tree-based learners; instance-based learners; rule-based learners	Classify Points of Interest
			Maximum likelihood estimation (MLE)	Estimate disaggregated land Use
S. Durduran [Dur15]	2015	Land cover dataset (remote images + extra attributes)	Multi-resolution segmentation	Image segmentation
			K-means	Define data belongin to each class
			Central tendency measures	Get the central tendency measure of each class
			k-nearest neighbor (K-NN), extreme learning machine (ELM), support vector machine (SVM)	Detection of urban land cover
Yuan et al. [Y+15]	2015	Points of Interest (POI), Taxi trajectories, Pubic transit records	Dilatation	Remove unnecessary details for map segmentation
			Subfields-based parallel thinning algorithm	Extract the skeleton of the road segments
			Two-pass algorithm	Generate segmented regions
			Latent Dirichlet Allocation (LDA)	Discovery region topics using mobility patterns based on mobility semantics and location semantics
			DirichletMultinomial Regres- sion (DMR)	
X. Zhang, S. Du, Q. Wangb [ZDW17]	2017	Remote sensing images, Points of Interest (POI)	Hierarchical semantic cognition (HSC)	Classify urban functional zones
			Multiresolution Segmentation	Segment remote sensing images
			Random Forest (RF)	Label categories of land use image objects
			ISO- DATA algorithm	Automatically cluster spatial object patterns
Zhang et al. [Z+17]	2017	OSM road network, Remote sensing images, Points of Interest (POI), LBSN users posts	Cellular automata model	Generate the urban land use parcels
			Random Forest (RF)	Land use classification
			Object-based classification	Classify preprocessed remote sensing images
			Gray-Level Co-occurrence Matrix (GLCM)	Calculate texture attributes
X. Liu et al. [L+17]	2017	OSM road network, Remote sensing images, Points of Interest (POI), LBSN user activities	Scale invariant feature transform (SIFT)	Extract features from remote sensing images
			K-means	Classify features
			Probabilistic latent semantic analysis (pLSA)	Identify latent semantic features
			Latent Dirichlet Allocation (LDA)	
			Support vector machine (SVM)	Classify urban land use types
S. Gao, K. Janowicz, H. Couclelis [GJC17]	2017	Points of Interest (POI), LBSN user activities	Latent Dirichlet allocation (LDA)	Generate summaries of thematic place topics
			K-means	Group semantically similar regions
			Delaunay triangulation spatial constraints	
			Ward clustering	Identify topological and hierarchical relations
Yao et al. [Y+17a]	2017	Points of Interest (POI), Traffic analysis zones (TAZ)	Greedy Algorithm	Construct the TAZ-based documents
			Word2Vec	Extract POI vectors
			K-Means	Group TAZs
			Random Forest (RF)	Land use classification
Yao et al. [Y+17b]	2017	Remote sensing images	TF-IDF algorithm	Transform the word frequencies into semantic features
			Random Forest (RF)	Classify urban land use patterns
			Google Inception v5	Detect land use patterns
H. Xing, Y. Meng [XM18]	2018	Points of Interest (POI), Text messages, Building-level blocks	Latent Dirichlet Allocation (LDA)	Calculate semantic information from crowdsourced data (text messages)
			Random Forest (RF)	Classify functional regions
J. Song et al. [S+18]	2018	Remote sensing images, Points of Interest (POI), Road network	Example-based feature extraction	Produce a binary built-up/non-built-up land cover map
			Multi-resolution segmentation	Image segmentation
			Object-based classification	Urban Land Cover Classification
Huang et al. [HZS18]	2018	Remote sensing images, Road blocks	Skeleton-based decomposition method	Decompose multispectral image
			Semi-transfer deep convolutional neural network	Land use mapping
Liu et al. [L+18]	2018	LBSN user activities, Remote sensing images, Taxi trajectories, Points of Interest (POI)	Inverse Distance Weight (IDW) function	Construct the relationships of different data types
			Kernel density estimation	Infer buildings' mixed-use functions
			A modified Bayesian model	Calculate the probability of purposes of passengers based on taxi data and POIs
X. Zhang, S. Du, Q. Wang [ZDW18]	2018	Remote sensing images, Road blocks	Multiresolution segmentation	Segment blocks
			Hierarchical semantic cognition (HSC)	Bottom-up classification (land covers and functional zones)
			Inverse hierarchical semantic cognition (IHSC)	Optimize classification results
Zhang et al. [Z+18]	2018	Remote sensing images	Object-based convolutional neural network	Urban land use classification
Deng et al. [D+19]	2019	Remote sensing images	Space-time fusion algorithm (ESTARFM)	Fuse original data pairs at two periods
			Multiresolution segmentation	Segment the images
			Support vector machine (SVM)	Extract land use and land cover types
Flores et al. [FZS19]	2019	Remote sensing images	ResNet-50 DCNN	Extract the deep features from images
Zhai et al. [Z+19a]	2019	Points of Interest (POI), Origin-destination (OD) datasets	K-means	Group POIs and cluster neighborhood areas
			POI frequency analysis	Annotate the function of each region
			Random Forest (RF)	Evaluate and compare the model accuracy
			Place2vec	Extract and classify urban functional regions
Zhang et al. [Z+19b]	2019	Remote sensing images	Joint Deep Learning (JDL)	Land use and land cover classification

urban land use and functional regions. However, there are other characteristics for justifying the use of this method. Random Forest was adopted for example, in [Z^+17a] and [ZUZ14]. Particularly, in the second, the authors chose a supervised approach based on RF because they see the algorithm in a practical, as powerful and scalable way for datasets containing a large number of features.

As listed in table 1, it was observed the occurrence of several different methods among the analyzed studies, beyond those described before. These techniques were employed for various different purposes. Some examples of them include: Nave Bayes, Extreme Learning Machine (EML), Word2Vec, Skeleton-based decomposition, Multiresolution segmentation, Place2vec, Joint Deep Learning (JDL) and many others. Although there are cases in which different methods are utilized for similar objectives, the datasets are often different, making very difficult to compare results and conclusions in such cases.

4 Conclusions and future work

In this paper, we talked about knowledge discovery on urban land use and land cover, addressing the importance of functional regions in this context. Moreover, we analyzed several scientific studies related to this topic, making it possible to discuss about the main challenges related to features selection. We also approached the main data types and the methods most frequently used in this specific field. During our analysis, we compared various works based on the types of data and the methods that were selected. We think this comparison is a source of new challenges, which we believe are essential to be considered in future work. In various cases, even using the same methods, for different regions, different authors arrived at different results and conclusions. Thus, we conclude that the results vary according to the method used, but also depend on the dataset and specificities of each region, due to factors such as construction patterns, population density and geography of the areas. Nevertheless, considering geographic data analysis as a specific topic of data analysis is important to remember that the results are directly related with data quality and granularity, but in this context, when using crowdsourced data for example, the spatial distribution of the data is also an essential factor to take into account.

Moreover, another consideration relates to the availability of data. During the study, we found the use of various data sources, and some of them are only available for some countries or regions. A very clear example of this situation is Weibo data, which is only available for China and building-level blocks, that is usually provided by public administration and is hardly available in various other locations. This limitation makes impossible or difficult to reproduce some studies in different locations.

In this research field, when talking about land use, a growing concern is related to the improvement of the accuracy of results, and therefore many authors have proposed the use of different types of data, together with remote sensing images, pursuing this goal. However, the use of innovative types of data, in many cases, did not result in a higher level of accuracy, compared to approaches that only use remote sensing images. This statement does not mean that combining data from multiple sources is not an important path to follow. From this observation we conclude that, depending on the chosen methodology, this wealth of data can improve the results obtained using remote sensing images or in cases where only one category of data is not enough to provide acceptable results.

Although many approaches have often been based on the same data types, the methods adopted by the authors are frequently different. In this paper, we presented the techniques that are used in at least two works among those chosen for our study. As a purpose for future work, we believe that implementing these methods using specific datasets, to allow a quantitative comparison of results obtained for each one, can provide a deep and solid base for our colleagues and future researchers interested in this area of knowledge. Also, based on the complexity of features selection when talking about spatial data analysis, we see an opportunity of research focused on this specific topic.

References

[AV15] J. J. Arsanjania and E. Vaz. An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. *International Journal of Applied Earth Observation and Geoinformation*, 35(PB):329–337, 2015.

[B^+14] T. Blaschke et al. Geographic Object-Based Image Analysis - Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87:180–191, 2014.

[BW06] R. Brennan and T. L. Webster. Object-oriented land cover classification of lidar-derived surfaces. *Canadian Journal of Remote Sensing*, 32(2):162–172, 2006.

[C+11] Z. Cheng et al. Exploring Millions of Footprints in Location Sharing Services. *The International Conference on Weblogs and Social Media*, 2010(Cholera):81–88, 2011.

[D+19] Z. Deng et al. Land use/land cover classification using time series Landsat 8 images in a heavily urbanized area. *Advances in Space Research*, 2019.

[Dur15] S. S. Durduran. Automatic classification of high resolution land cover using a new data weighting procedure: The combination of k-means clustering algorithm and central tendency measures (KMC-CTM). *Applied Soft Computing Journal*, 35:136–150, 2015.

[FMFM14] V. Frias-Martinez and E. Frias-Martinez. Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237–245, 2014.

[FZS19] E. Flores, M. Zortea, and J. Scharcanski. Dictionaries of deep features for land-use scene classification of very high spatial resolution images. *Pattern Recognition*, 89:32–44, 2019.

[G+16] L. Gong et al. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science*, 43(2):103–114, 2016.

[GIM11] B. Gong, J. Im, and G. Mountrakis. An artificial immune network approach to multi-sensor land use/land cover classification. *Remote Sensing of Environment*, 115(2):600–614, 2011.

[GJC17] S. Gao, K. Janowicz, and H. Couclelis. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21(3):446–467, 2017.

[GP+18] J. C. García-Palomares et al. City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities*, 72(September 2017):310–319, 2018.

[HZS18] B. Huang, B. Zhao, and Y. Song. Urban land-use mapping using a DCNN with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 214(April):73–86, 2018.

[J+15] S. Jiang et al. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53:36–46, 2015.

[J+17] M. Jendryke et al. Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai. *Computers, Environment and Urban Systems*, 62:99–112, 2017.

[K+14] G. Kumar et al. A detailed review of feature extraction in image processing systems. *International Conference on Advanced Computing and Communication Technologies, ACCT*, pages 5–12, 2014.

[L+17] X. Liu et al. Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science*, 31(8):1675–1696, 2017.

[L+18] X. Liu et al. Characterizing mixed-use buildings based on multi-source big data. *International Journal of Geographical Information Science*, 32(4):738–756, 2018.

[LL16] G. Lansley and P. A. Longley. The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58:85–96, 2016.

[LZZ15] J. Lilleberg, Y. Zhu, and Y. Zhang. Support vector machines and Word2vec for text classification with semantic features. *Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2015*, pages 136–140, 2015.

[Pat17] S. Patel. Chapter 2: SVM (Support Vector Machine) Theory. url: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72, 2017. Accessed: 2019-06-10.

[Q+12] Z. Qi et al. A novel algorithm for land use and land cover classification using RADARSAT-2 polarimetric SAR data. *Remote Sensing of Environment*, 118:21–39, 2012.

[S$^+$16] R. Susanti et al. Smart Growth, Smart City and Density: In Search of The Appropriate Indicator for Residential Density in Indonesia. *Procedia - Social and Behavioral Sciences*, 2016.

[S$^+$18] J. Song et al. Mapping Urban Functional Zones by Integrating Very High Spatial Resolution Remote Sensing Imagery and Points of Interest. *Remote Sensing*, 10(11):1737, 2018.

[Ser18] NOAA National Ocean Service. What is the difference between land cover and land use? url: https://oceanservice.noaa.gov/facts/lclu.html, 2018. Accessed: 2019-03-13.

[ST17] D. N. Sotiropoulos and G. A. Tsihrintzis. Machine Learning Paradigms. In *Machine Learning Paradigms: Artificial Immune Systems and their Applications in Software Personalization*, pages 107–129. Springer International Publishing, Cham, 2017.

[Tre16] A. Trevino. Introduction to k-means clustering. url: https://www.datascience.com/blog/k-means-clustering, 2016. Accessed: 2019-03-20.

[XM18] H. Xing and Y. Meng. Integrating landscape metrics and socioeconomic features for urban functional region classification. *Computers, Environment and Urban Systems*, 72(February):134–145, 2018.

[Y$^+$12] J. Yuan et al. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

[Y$^+$15] N. J. Yuan et al. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):712–725, 2015.

[Y$^+$17a] Y. Yao et al. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec. *International Journal of Geographical Information Science*, 31(4):825–848, 2017.

[Y$^+$17b] Y. Yao et al. Sensing urban land-use patterns by integrating Google Tensorflow and scene-classification models. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(2W7):981–988, 2017.

[Z$^+$17a] X Zhang et al. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 132:170–184, 2017.

[Z$^+$17b] Y. Zhang et al. The combined use of remote sensing and social sensing data in fine-grained urban land use mapping: A case study in Beijing, China. *Remote Sensing*, 9(9), 2017.

[Z$^+$18a] C. Zhang et al. An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sensing of Environment*, 216(June):57–70, 2018.

[Z$^+$18b] X. Zhang et al. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. *Remote Sensing of Environment*, 212(Dec.):231–248, 2018.

[Z$^+$19a] W. Zhai et al. Beyond Word2vec: An approach for urban functional region extraction and identification. *Computers, Environment and Urban Systems*, 74(August 2018):1–12, 2019.

[Z$^+$19b] C. Zhang et al. Joint Deep Learning for land cover and land use classification. *Remote Sensing of Environment*, 221(November 2018):173–187, 2019.

[ZD15] X. Zhang and S. Du. A Linear Dirichlet Mixture Model for decomposing scenes: Application to analyzing urban functional zonings. *Remote Sensing of Environment*, 169:37–49, 2015.

[ZUZ14] X. Zhan, S. V. Ukkusuri, and F. Zhu. Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. *Networks and Spatial Economics*, 14(3-4):647–667, 2014.

[ZZD16] L. Zhang, L. Zhang, and B. Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.

[ZZZ15] Y. Zhong, Q. Zhu, and L. Zhang. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 53(11):6207–6222, 2015.