Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 385; doi:10.3390/ijgi8090385

Article

Fusing Knowledge-Based and Data-Driven Techniques for the Identification of Urban Functional Regions

Emmanuel Papadakis [1]*, Song Gao [2] and George Baryannis [3]

[1] Department of Geoinformatics - Z_GIS, University of Salzburg, Schillerstr. 30, 5020 Salzburg, Austria; emmanouil.papadakis@sbg.ac.at
[2] Department of Geography, University of Wisconsin, Madison, WI 53706, USA; song.gao@wisc.edu
[3] Department of Computer Science, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK; g.bargiannis@hud.ac.uk
* Correspondence: emmanouil.papadakis@sbg.ac.at

Abstract: The problem of identifying functional regions in an urban setting has been approached in literature using two general methodologies: top-down, encoding expert knowledge on urban planning and design (e.g. into patterns) and using that knowledge for identification, and bottom-up, relying on crowdsourcing and Volunteered Geographic Information (VGI) to train learning models, using techniques such as Latent Dirichlet Allocation (LDA) topic modeling. Both approaches have their advantages but also face important limitations, with knowledge-based approaches being criticized for scalability and transferability issues and data-driven approaches for lacking interpretability and depending heavily on data quality. To mitigate these disadvantages, we propose a novel framework that fuses data and knowledge in three different ways: functional regions identified from individual approaches are evaluated against each other, knowledge from patterns is used to adjust learning model results and topic models are used to adjust pattern-based results. The proposed methodologies are demonstrated through the use case of identifying shopping-related functional regions in the Los Angeles metropolitan area. Results show that the combination of results from knowledge-based and data-driven techniques can help uncover discrepancies between the two different approaches and smoothen inaccuracies caused by the limitations of each approach.

Keywords: functional region; place; patterns; topic modeling; urban planning; Volunteered Geographic Information (VGI)

1. Introduction

Rapid urbanization has spread all over the globe in recent decades and has transformed cities worldwide, allowing them to support an ever-expanding spectrum of functions and human activities, satisfying residential, commercial, industrial and transportation needs, among others. This has created new challenges for Geographic Information Science (GIScience) that are not only limited to an exploration of land surfaces and urban space, but also involve more human-oriented notions such as places [1] and functional regions [2]. These notions are fundamental to understanding how people live and act on urban space, in order for Geographic Information Systems (GIS) to assist citizens in navigating their surroundings in everyday life [3].

In this setting, GIS need to be able to identify functional interaction patterns in an urban context, assist in understanding socioeconomic environments and eventually provide useful answers to queries such as "what can I do around here" or "where can I find places that provide this function" by relying on knowledge and data on human activity and experience. Put in simpler terms, GIS need to be

able to identify functional regions on urban space. Standard GIS technologies such as remote sensing cannot provide solutions to this problem, since they are designed to extract physical rather than human-focused characteristics of space [4,5].

To address this challenge, research has followed two mostly independent pathways. The earlier one involves a top-down methodology that begins with encoding knowledge about human activities and experience and then uses the derived knowledge models to identify and delineate functional regions on space (or, in other words, places that support particular functions). Examples include gazetteers [6], semantic spatial search engines [7] and place-based GIS [8]. More recently, the proliferation of crowdsourcing and Volunteered Geographic Information (VGI) [3] as well as the successes of data-driven and machine learning technologies has led several researchers to propose bottom-up methodologies that use collected data to train learning models that can identify regions based on similar (functionality-related) characteristics. Examples of this pathway include the use of Latent Dirichlet Allocation (LDA) topic modeling, Bayes classifiers and clustering methodologies on textual, point of interest (POI) and social network data [2,9–11].

As should be expected, both of these approaches to the problem of identifying functional regions have their advantages and disadvantages. For instance, knowledge-based approaches are easily interpretable, providing explanations behind the identification of particular functional regions, while also providing results in a machine-readable form. However, the process of acquiring and combining knowledge from relevant expert sources may be error-prone and time consuming [12]. Data-driven approaches, on the other hand, can uncover hidden patterns of human behaviour and activity from large amounts of VGI data that can be harder to discover by humans. However, their success largely depends on the availability of relevant, complete and unambiguous datasets, while their results may not always be easily interpretable [13].

In this work, we bring together the different pathways for the identification of functional regions by proposing a novel approach that combines knowledge-based and data-driven characteristics. Specifically, we propose three ways of fusing pattern-based identification based on the function-based model of place, as introduced in [14] with the data-driven extraction of functional regions exploiting POI and human activity data from social networks introduced in [10]. The overall aim of this proposal of "joining forces" is to combine the advantages of knowledge-based and data-driven approaches, while also using one approach to mitigate the drawbacks of the other. The contributions of this article are three-fold as follows:

- A critical analysis of the approaches in [14] and [10], uncovering the main advantages and disadvantages of knowledge-based and data-driven techniques for functional region identification
- A novel framework for urban functional regions identification that combines knowledge and data in three different ways: mutual evaluation, using knowledge to improve data and using data to improve knowledge
- A discussion, in the context of GIS, of the benefits of combining the interpretability offered by knowledge-based techniques with the transferability and scalability of data-driven methodologies.

The merits of the proposed methodological framework are demonstrated through the example of identifying regions offering shopping-related functionality in the Los Angeles metropolitan area. Results show that the fusion of knowledge and data allows for both a mutual evaluation mechanism to uncover discrepancies between the two methodologies and processes that adjust the results of one approach by taking into account results of the other.

The remainder of this article is organized as follows: Section 2 introduces background notions related to functional regions, including space and place, then provides a concise summary of the most prominent knowledge-based and data-driven approaches relevant to the identification of functional regions. Section 3 provides a critical analysis of approaches recently proposed in [14] and [10] which

drives the introduction of a novel framework for functional regions that fuses knowledge and data. This framework is demonstrated in Section 4 through the example of identifying "shopping plazas" in the Los Angeles metropolitan area. Finally, Section 5 discusses the benefits and lessons learned from the proposed framework, followed by concluding remarks and directions for future research in Section 6.

2. Background and Related Work

In this section we describe the notion of place and the necessary assumptions required to enable its integration within GIS, leading to the simplified variant notion of functional regions. Afterwards, we concisely describe the most prominent representatives of the knowledge-based and data-driven approaches to the problem of identifying and delineating places and functional regions.

2.1. Space, Place and Functional Regions

Space, location, place, and regions are terms widely used in literature and everyday life. Despite their subtle differences, human speech allows the interchangeable use of the aforementioned terms. Scientific literature, however, engages strict definitions and concepts when describing these fundamental terms of spatial reference. This is particularly apparent, when considering the definition of place as space infused with human meaning [1]. While easy to understand by humans, attempting to integrate such a notion within digital systems, such as GIS, is far more difficult. The need for conceptualization combined with the modeling challenges derived from the term "human meaning" has turned various scientists towards a more practical view of place, leaving the analysis of its concept as a whole in the domains of philosophy, psychology, linguistics, human geography and various social sciences.

In this work we adopt a partial and simplified view of place that facilitates formal representation, while acknowledging shortcomings related to more human-oriented aspects of place, such as emotions, purposes, opinions, affections and others. More specifically, we consider a view of place that is able to address questions about the following aspects, which are based on the idea of the essential kinds of place inference [15, p. 101]:

- **identification/classification:** a place is identified by its unique identity or categorized based on its type.
- **equipment:** a place consist of physical entities, referred as constituent elements, which form its pragmatic structure.
- **functions:** a place supports activities or actions, which in return suggest the functional context assigned to it.
- **localization:** a place has an absolute or relative location.

Such a simplified view brings place closer to the notion of a functional region. For the purposes of this work, we consider a functional region as an area found on space, projected on a specific location and equipped with particular physical entities. These entities enable particular functions, which in return attribute a unique identity or type for the region itself. For instance, Grand Park is located in the civic center of Los Angeles and consists of multiple trees, ponds, gardens, fountains, benches, food trucks, playgrounds and walkable paths. Based on its constituents, it facilitates recreation, sight seeing, social activities and sports, which in return introduce a generic type, named park, that is associated with these functions. Moving in the opposite direction, any functional region that is classified as park is expected to provide similar functions and have equipment that includes similar physical entities.

The integration of functional regions within digital systems is a long-standing challenge driven by the different perspectives of information that formal systems are capable of representing and people are able to interpret. On the one hand, GIS is a family of tools dedicated to capture, represent, process, and analyze spatial data[3]. The human mind, on the other hand, does not treat regions as mere spatial extents, but a combination of various layers of information that attribute to a semantic context,

expressed in the form of types or functions supported by the region under question. The existing approaches utilized to bridge functional regions with GIS can be summarized in two main categories: (1) knowledge-based, which follow a top-down approach, starting from semantic information (e.g. expert knowledge) of a region which is then spatially projected; and (2) data-driven, which follow the opposite, bottom-up approach, beginning from spatial data which are fed to machine learning algorithms that extract semantic information.

2.2. Knowledge-Based Approaches

The earliest and perhaps the most prevalent approach of identifying regions related to particular functions is using spatially-referenced catalogs of place names, known as digital gazetteers [6]. These encode relations between place names, space footprints, spatial categories and temporal information, to name a few. While gazetteers enable keyword-based identification of particular regions or extracting relevant place names based on space footprints, they are unable to go beyond this to support region identification based on more information than one or more keywords that represent functions and cannot resolve the inherent ambiguities.

Semantic-based approaches address this limitation by leveraging ontologies to describe geographic entities. A prominent example is SPIRIT [7], a spatial search engine which relies on a geographical ontology maintaining knowledge about place names, place types, spatial footprints and topological relationships to provide both structured and graphical spatial queries. Structured queries are in the form of a triple containing a thematic component (e.g. shopping malls), a spatial relationship (e.g. near) and a geographic component, such as a place name (e.g. London) or an imprecise region (e.g. south of England). Graphical queries allow drawing a polygon on a map to specify the geographic component. Similarly, ontological gazetteers [16] rely on knowledge graphs to enhance the capabilities of standard digital gazetteers. Knowledge contained in these graphs includes thematic information such as types, activities and hierarchies, allowing queries based on similarity and subsumption. Scheider and Purves [17] further propose the creation of semantic descriptions of places by extracting place localization knowledge from narratives, which can then be used to improve place-based search.

The aforementioned approaches greatly enhance the process of identifying relevant regions on space by relying on both thematic and spatial information. However, while they are capable of recognising and geolocating place names, they cannot explain why a particular region is relevant even without an associated place name [18]. A first step towards this direction is the definition and formalization of place reference systems [15]. These associate places with the activities and actions that are afforded by the objects contained within them and use cognitive simulations to determine whether an activity is afforded by a place. However, the complexity of these simulations hinders potential implementations of place reference systems. Also, containment of certain objects alone cannot guarantee the affordance of a particular activity: a place containing a path and a highway does not afford walking when these intersect with each other.

To address these limitations, a function-based model of place is proposed in [8,14,19], based on the assumption that place is space associated with particular functionality. In this model, places are formalized as patterns which are defined as sets of components, composition rules and functional implications (see Section 3.1 for more details). The original patterns are extracted from narratives, such as dictionary or encyclopedia definitions of a place. Subsequent work [20] enhances these with empirical knowledge derived from spatial data and shows the results of using different patterns to identify places that satisfy particular functions.

2.3. Data-Driven Approaches

A variety of data-driven algorithms have been applied to identify regions of particular characteristics on a map; to facilitate analysis they are presented here based on the type of VGI data used as source.

Adams and Janowicz [9] use unstructured text from Wikipedia articles to derive thematic signatures that can describe the place type associated with each article. LDA topic modeling is used to identify the latent structure of topics contained in each article. Then, the trained LDA model is used on each collection of articles that refer to a specific place type to infer a topic distribution for it (thematic signature).

Hobel et al. [21] rely on the OpenStreetMap dataset [1], using the tags attached by users to spatial features to understand how users describe particular spatial regions such as shopping areas. Based on these POI data, descriptions of spatial region types are extracted using an algorithm similar to image segmentation, combining descriptions of characteristics that hold for the region, such as "contains at least one shop and restaurant". The presented case study shows how a description derived from a shopping area in London can be used to identify similar areas in Vienna.

A quite popular type of VGI data used for the identification of place types and functional regions are social network activities such as check-ins. Noulas et al. [22] use frequencies of Foursquare check-in data per place category to determine which categories are more common in particular regions. This analysis is applied to $10x10km^2$ regions in New York and London to find clusters containing similar distributions of place types. Zhou and Zhang [23] similarly combine Twitter and Foursquare data to extract spatial distributions of common human activities (e.g. food and restaurants, shops and services, outdoor and recreation) and determine major hotspots. Finally, Zhi et al. [24] use a vast dataset of 15 million social media check-ins over a year to detect functional regions. Spatiotemporal structures which potentially represent the associations between functional regions and human activities are extracted and are then used to identify functional regions in the city of Shanghai.

While the aforementioned works rely on a single type of VGI data, research is increasingly focusing on synthesizing different data sources to counterbalance disadvantages of individual data sources. Yuan, Zheng and Xie [2] integrate POI data with taxi trajectory data into a novel topic model-based method to identify functional regions. Regions and functions are considered as documents and topics, respectively, while trajectory data are considered as words and POIs as metadata. Hobel, Fogliaroni and Frank [11] apply natural language processing to user comments posted in English on TripAdvisor for the historic center of Vienna to find compound names that refer to geographic features, which are then used in combination with OpenStreetMap tags to train a Bayes classifier. The area identified by the trained classifier fits with the boundaries of the city of Vienna in 1850.

More recently, Gao, Janowicz and Couclelis [10] developed a statistical framework to study functional regions of the 10 most populated US cities that relies on POI data and social network check-ins. LDA topic modeling is applied, with regions and functions are considered as documents and topics, respectively, similarly to [2]. However, POI types here are considered as words and the method takes into consideration the popularity of each POI based on the number of user check-ins. Each region on a map is associated with a number of topics and clustering techniques are used to aggregate similar regions into functional regions (more details in Section 3.2).

3. Methodology

This section begins with a critical analysis of the most recent representatives of the knowledge-based and data-driven approaches to functional region identification: the one relying on the function-based model of place, as introduced in [14] and the one applying LDA topic modeling on POI and social network datasets, as introduced in [10]. Advantages and disadvantages are highlighted for each approach, and based on this analysis, Section 3.3 proposes several methodologies that attempt to fuse knowledge-based and data-driven approaches, keeping the best of both worlds while mitigating their drawbacks.

[1] https://www.openstreetmap.org

3.1. Identification of Places using Composition Patterns

To identify places that satisfy particular functionality (in other words, identify functional regions), the approach in [14] relies on patterns created according to the function-based model of place and based on knowledge gathered from expert sources. Such sources can range from widely-acceptable descriptions of places in dictionaries or encyclopedias to specialised reports produced by experts, such as urban design standards and manuals. These sources yield information relevant to fundamental elements of the function-based model, namely components, composition rules and functional implications.

Components are at the lowest level and are constituent objects of a place that enable, enhance, hinder or block particular functions. Each component belongs to a particular class (type) of components and is associated with thematic and geometric information. Thematic properties semantically enrich a component and express properties such as "the shop was opened in 2010". Geometric properties provide a spatial description for each component, e.g. in the form of points, lines and polygons.

Composition rules express relations among components and are of four types: (1) occurrence, describing existence and population; (2) correlation, expressing relative frequency of appearance; (3) spatial relation, modeling topology; and (4) proximity, expressing distance between components. Three types of filters are also provided to apply composition rules on subsets of components based on their characteristics (according to type, thematic or geometric property).

Each function is then associated with a logical formula (named functional implication) made of composition rules that need to hold for that function to be provided. A composition pattern of a place consists of one or more of these functional implications. An example of an elaborate functional implication is that of aircraft taxiing, which is enabled if the following hold, according to [14]:

1. the occurrence rule, which describes the existence of a component of type taxiway
2. the function of housing aircraft is enabled (due to the existence of an hangar)
3. the function of repair aircraft is enabled (due to the existence of an apron)
4. the function of take off and landing of aircraft is enabled (due to the existence of a runway)
5. the spatial relation rule, which states that the spatial association of the component taxiway with each one of runway, hangar, and apron, is touch

Applying the composition pattern on data for the components of interest (e.g. from OpenStreetMap) allows the identification of regions on a map that satisfy some (or all) of the functions contained in the pattern. Regions can be scored according to the number of functions they satisfy out of a pattern, taking into account whether some are considered core or secondary. For instance, a score of zero may be attributed to a region if core functions are not provided, regardless of whether secondary ones are provided or not.

3.1.1. Advantages and Disadvantages

The approach of composition patterns considers place as a system of interrelated components, whose spatial configuration permits or prevents particular functions to hold. Therefore, it extends the "declarative" nature of functional regions, allowing a composite view based on the semantic and spatial configuration of the underlying components. For instance, a park is no longer considered as a predefined spatial footprint with assigned semantics or a set of physical entities; instead, it is a region composed by strict rules that spatially organize its containing physical entities, which in turn enable its functionality and approximate its spatial projection.

This differentiating feature has a significant impact in the formalization and integration of functional regions within GIS. On the one hand, instances of functional regions are transformed to semantically enriched spatial data (i.e. components) that are machine readable; on the other hand, the constraints and rules introduced are well-formed and easily interpretable, facilitating human understanding and reasoning. Moreover, functions are not bound to textual descriptions, but take the form of logical rules which, upon evaluation, allow the grading of the corresponding regions

according to the number of functions they support, as well as how well they are supported. This facilitates comparison of regions with different or unknown types based on how well they operate given a predefined set of functions.

In addition, since the context of regions is not represented as static text literals, patterns are easily adjustable in order to allow additional or alternate interpretations of functions based on the particular requirements of each setting: for instance, the function of walkability is realized differently in the United States than in Austria, because of the cultural background and urban structures of each country. Additionally, since the approach is built on logical rules, it requires a minimal amount of data to evaluate functional implications, hence it can perform quite well with scarce data.

However, identifying functional regions using composition patterns carries a number of limitations, most notably those of scalability and transferability in terms of the area of study. Scaling to larger areas may significantly increase the preprocessing and actual processing time: for instance, identifying parks within a city is quite efficient, however applying proximity algorithms in the scale of a continent would require several assumptions and performance optimizations to achieve reasonable efficiency.

In terms of transferability, the same pattern can be applied to different parts of the world, provided that they share similar characteristics that affect human behaviour and activities. However, there may be a need for adjustments, in order to make the composition rules or functions fit to the area of study the best way possible; this, for instance would be necessary to transfer a pattern from western to eastern world countries. In essence, transferability is made difficult because patterns rely on assumptions in order to fit the real world into well-structured hierarchical composition-based models; a semantically correct knowledge-based model that unambiguously identifies all possible connections between components and functions regardless of the area of study is virtually impossible.

Finally, while the approach is not dependent on the availability of high volumes of data, the successful discovery of functional regions requires heterogeneous data sources that need to be unambiguous and finely structured. A pattern that is built on low quality data will inevitably perform inadequately in terms of functional region identification, while the unavailability, for instance, of data in a region related to components participating in core functions within a pattern will lead to excluding this region from results.

3.2. Functional Region Extraction from POI and Human Activity Data

The methodology presented in [10] relies on a popularity-based probabilistic topic model that is trained based on VGI data on POIs and location-based social network check-ins. The key idea of LDA topic modeling for textual data is applied but with the following analogy: the type of each POI (e.g. restaurant or park) is considered a word, the region that contains these POIs is considered a document, while each urban function represents the topic, representing thematic and semantic characteristics of places. The goal is to produce a discrete probability distribution over POI types for each function.

To address the significant effect of human activity to the distribution of functions in an urban setting, the generation of the document-word frequency matrix used in the LDA topic modeling approach is modified. The occurrence of a POI type (word) within a region (document) is re-scaled according to the check-in counts for all POIs of that type in the region. The re-scaled occurrence for a POI type t given a region d is given by the following formula: $Freq_{(d,t)} = \sum Log(V_{(d,t,i)})$, where $V_{(d,t,i)}$ is the number of unique users who have checked-in (using social networks) to venue i of type t in region d.

At the end of this process each region is represented as a vector of K-dimensional latent thematic topics (POI types), with the optimal K determined experimentally. The vector essentially encodes underlying co-occurrence relationships among POI types taking social network-based human activity into account. Individual regions that are semantically similar in terms of the POI types included in the vector may be considered to be contributing to the same urban function and, hence, forming a larger, but thematically cohesive, functional region. To achieve this, clustering algorithms can be exploited.

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 385; doi:10.3390/ijgi8090385

The experiments presented in [10] apply both k-means clustering [25] and the Delaunay triangulation spatial constraints clustering [26].

3.2.1. Advantages and Disadvantages

The popularity of POIs reveals information about the trending associations of human activities with place types. Topic modeling that exploits such information has the advantage of providing a view of urban space as it is seen through the lenses of society and people. From a wider perspective, popularity-based POI topic modeling augments the traditional remote sensing view of urban environment as physical landscape with information about the distribution of urban functions and human activities, resulting into what is termed social sensing [4].

This function-based view of urban space allows the extraction of semantic signatures for spatial entities based on their functionality. These signatures can then facilitate a variety of applications. Regions can be identified given a specific functional context, as well as, being compared in terms of similarity based on their assigned functions. Furthermore, the topic extraction method is able to classify regions that are characterized as multi-functional; it reveals the likelihood of certain functions to be present in the region under question, which in turn makes it possible to discover clusters of "similar" functional regions given a context expressed as a multinomial distribution of different types of POIs.

Being a data-driven approach, the extraction of functional regions using topic modeling inherits some common benefits and drawbacks. Scalability and transferability in terms of the area of study are two of the key advantages of the approach. The discovery process of functional regions can be scaled from the boundaries of a city to a country or even a continent, without the need for constraints, assumptions or simplifications while keeping execution at efficient levels in terms of time and space requirements. In terms of transferability, the same LDA topic modeling methodology can be directly applied to any study area with limited or no required adjustments. This stems from the fact that data are used simply as numerical values, without dealing with any case-specific schemas or complex structures.

The aforementioned simplicity, however, causes a high dependency of the quality of topic modeling on the availability of significant amounts of data. The absence of POI types and human activity information may lead to an uneven distribution of functions within a wider region. Misleading classification can also arise since, for instance, POI data and social media check-ins on hotels or residences are often scarce compared to those related to restaurants or bars. Hence, while the approach can scale to larger areas and transfer to different ones, its accuracy will inevitably vary according to data availability and quality.

Another important limitation of this approach is related to interpretability. This is a common characteristic of data-driven techniques, as they tend to employ advanced formulas and parameters which are not always comprehensible by or explainable to humans. In this particular case, this translates to arbitrary boundaries for functional regions that are not necessarily linked to the actual location of the POIs that deliver the functionality in question. Additionally, in some cases the association of the probabilistic weights assigned to POI types may be difficult to be directly connected with a perceived functionality. For instance, an increased co-occurrence of shops does not necessarily mean that the region represents a shopping area, since the individual shops might be sparsely located which hinders walkability, a desirable feature associated with shopping-related functionality.

3.3. A Framework for Functional Region Identification Fusing Knowledge and Data

The analysis in Sections 3.1 and 3.2 shows that both approaches have notable benefits, but they are also restricted by important limitations. In this section, we introduce a methodology that combines the works of composition patterns and popularity-based topic modeling forming a fusion approach for identifying functional regions that keeps the best qualities of each individual approach while mitigating

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 385; doi:10.3390/ijgi8090385

underlying limitations. We present three types of fusion: mutual evaluation, data to knowledge fusion and knowledge to data fusion. These are illustrated in Figure 1.

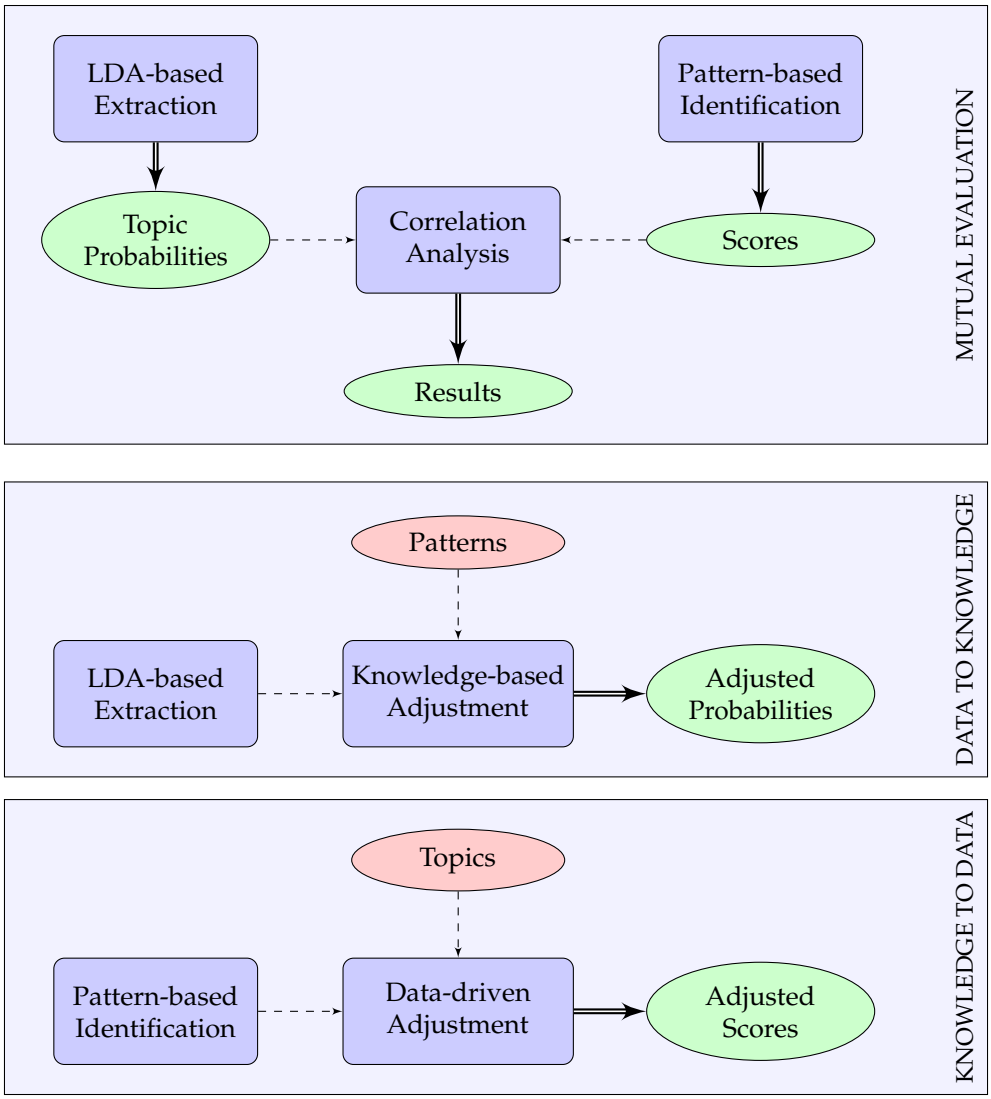

Figure 1. Overview of the proposed framework fusing knowledge-based and data-driven approaches.

The following assumptions are necessary in order to ensure consistency of the fusion processes. Both approaches, composition patterns and topic model extraction, are applied on the same data sets and within the same study area. All the numerical values used are normalized and transposed accordingly to the same spatial scale. Furthermore, the process of identifying particular functional regions differs slightly based on the particularities of each approach. For instance, identifying a functional region as "shopping plaza" translates to finding areas where "shopping mall" is the dominant POI type using LDA topic modeling along with additional POI types related to shops and restaurants (topic 67 in [10] and Table A1 in Appendix A) and without performing any clustering. The same goal translates to identifying regions that conform to a pattern containing a number of sub-functions that are linked to human activities associated with a "shopping plaza", such as shopping experience and walkability (a full pattern is provided in Tables A2 and A3 in Appendix A). Hence, comparison is performed on a dominant topic versus relevant pattern basis.

The following notations are introduced for the remainder of this section. FRT denotes the set of functional regions extracted using topic modeling, while FRP is the set of functional regions identified using composition patterns. The assigned value that represents the probability calculated using the topic model and the score calculated based on a pattern are denoted as $V_T(fr_i)$ and $V_P(fr_i)$,

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 385; doi:10.3390/ijgi8090385

respectively and are both referred to as confidence values of an individual region fr_i. Finally, the resulting set of functional regions after the fusion process is denoted as FR' and the adjusted confidence value after fusion for the region fr_i is denoted as $V_F(fr_i)$.

3.3.1. Mutual Evaluation

This methodology aims to investigate the variability between the results of the composition patterns and topic modeling processes. Initially, the fusion process measures the global correlation between the values by calculating the Pearson correlation coefficient (henceforth, R). This gives a rough indication with regard to how associated the distributions of the confidence values of the two approaches are.

To determine cases where there is significant consensus between the two approaches, we calculate an adjusted confidence value that is the product of the individual ones: $V_F(fr_i) = V_T(fr_i) * V_P(fr_i)$. By taking the product, cases of high agreement are accentuated: if both approaches have high confidence values, the product will be even higher, while in cases where a region is scored low by both approaches, the product will be even lower.

Having identified regions with high agreement, correlation analysis then focuses on the opposite case. Confidence values of each approach are compared and any regions which are scored differently by more than a threshold are identified. For each of these cases of significant disagreement, we attempt to explain the reasons that may have caused them, by looking at the individual characteristics of each approach: functional implications within the pattern and POI probabilities within the topic.

3.3.2. Data to Knowledge Fusion

This fusion process attempts to frame the functional context derived from the topic modeling extraction process in a way that conforms to the guidelines provided by the composition pattern. As opposed to the mutual evaluation case, this method does not keep equal distances between the two approaches; instead, it focuses on introducing weights that indicate how well the confidence values of the data-driven approach fit the knowledge-based individual sub-functions. This process considers the knowledge-based results as the "actual" values which are compared with (and used to adjust) the "experimental" values calculated using the data-driven approach. The goal is to inflate or deflate the "experimental" values in order to better approximate the "actual" values, taking into consideration the overall correlation of the results.

Each functional region extracted using LDA topic modeling for a particular topic is compared against the individual sub-functions contained within the composition pattern related to this topic. This is achieved using the following formula:

$$V_F(fr_i) = \begin{cases} V_T(fr_i) * (1 - R(V_T, V_P)) & V_P(fr_i) = 0 \\ V_T(fr_i) + |V_T(fr_i) - V_P(fr_i)| * R(V_T, V_P) & V_P(fr_i) > V_T(fr_i), V_P(fr_i) \neq 0 \\ V_T(fr_i) - |V_T(fr_i) - V_P(fr_i)| * R(V_T, V_P) & V_P(fr_i) < V_T(fr_i), V_P(fr_i) \neq 0 \\ V_T(fr_i) & V_T(fr_i) = V_P(fr_i) \neq 0 \end{cases}$$

In essence, this formula adjusts the probability of the topic in question proportionally to the score calculated based on the satisfaction of sub-functions in the pattern. Note that in the exceptional case where this score is equal to zero (because core sub-functions are not satisfied), the probability is adjusted according to the global correlation value across all identified regions.

3.3.3. Knowledge to Data Fusion

The third and final fusion process is the dual of the aforementioned one: the data-driven results as the "actual" values which are compared with (and used in order to adjust) the "experimental" values calculated using the knowledge-based approach. The goal here is to adjust the results of the

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 385; doi:10.3390/ijgi8090385

knowledge-based process by taking into account information derived from human activity information, which is captured through the LDA topic modeling approach. For instance, this would account for cases where a region satisfies most of the sub-functions related to shopping included in a pattern but where reported shopping-related check-ins are relatively low.

Each functional region that is identified using a composition pattern is compared against the probability value of the associated topic, calculated using LDA topic modeling. Similarly to the previous process, the weight is calculated using the following formula:

$$V_F(fr_i) = \begin{cases} V_P(fr_i) * (1 - R(V_T, V_P)) & V_T(fr_i) = 0 \\ V_P(fr_i) + |V_T(fr_i) - V_P(fr_i)| * R(V_T, V_P) & V_T(fr_i) > V_P(fr_i), V_T(fr_i) \neq 0 \\ V_P(fr_i) - |V_T(fr_i) - V_P(fr_i)| * R(V_T, V_P) & V_T(fr_i) < V_P(fr_i), V_T(fr_i) \neq 0 \\ V_P(fr_i) & V_P(fr_i) = V_T(fr_i) \neq 0 \end{cases}$$

This formula again allows the proportional adjustment of the score calculated using knowledge-based patterns considering the co-occurrence of POIs derived from the data-driven approach.

4. Demonstration

In this section, we demonstrate the application of the proposed fusion methodologies on the problem of identifying functional regions that provide functionality associated with "shopping plazas". We first show individual results of the LDA topic modeling approach, as reported in [10], and the function-based pattern approach, as reported in [14], with a slightly updated version of the included composition pattern. Then, we demonstrate the results of applying the mutual evaluation, data to knowledge and knowledge to data fusing techniques. The results are discussed in detail in Section 5.

4.1. Study Area and Data

The demonstration involves the metropolitan area of Los Angeles, California using the official boundaries provided by the U.S. Census Bureau's TIGER geographic database [2] and coordinate reference system "EPGS:3309". The POIs involved in the experiment are extracted from the online social platform Foursquare using the Foursquare developer API and represent the entries of December 2016. The total number of POIs within the study area is 14824; they are classified under 425 types and organized in 9 categories following the formal Foursquare Venue Categorization [3]. Additional data include the street network, acquired from the OpenStreetMap platform, which is classified based on the types and categories found in the OpenStreetMap Wiki [4].

4.2. Results using Individual Approaches

The topic modeling approach is demonstrated using topic 67 in [10], which is interpreted as "shopping plaza". It reflects the functional context of a region characterized by high occurrence of shopping-related POIs, such as shopping malls and accessories stores, accompanied with moderate to low numbers of restaurants or other food-oriented facilities (as shown in Table A1 in Appendix A). The LDA algorithm reported in [10] is applied on 200 regions of 4.5km radius each, properly distributed to cover approximately all of the spatial extent of the Los Angeles metropolitan area. Each of these candidate regions is then classified based on the probability of topic 67 being dominant, meaning that the candidate region is more likely to be a "shopping plaza" than any other functional region type.

---

[2] https://www.census.gov/geo/maps-data/data/cbf/cbf_msa.html
[3] https://developer.foursquare.com/docs/resources/categories
[4] https://wiki.openstreetmap.org/wiki/Key:highway

Figure 2 illustrates the spatial projection of the centroids of every candidate region, using size as a measure of scale to reflect the respective probability.

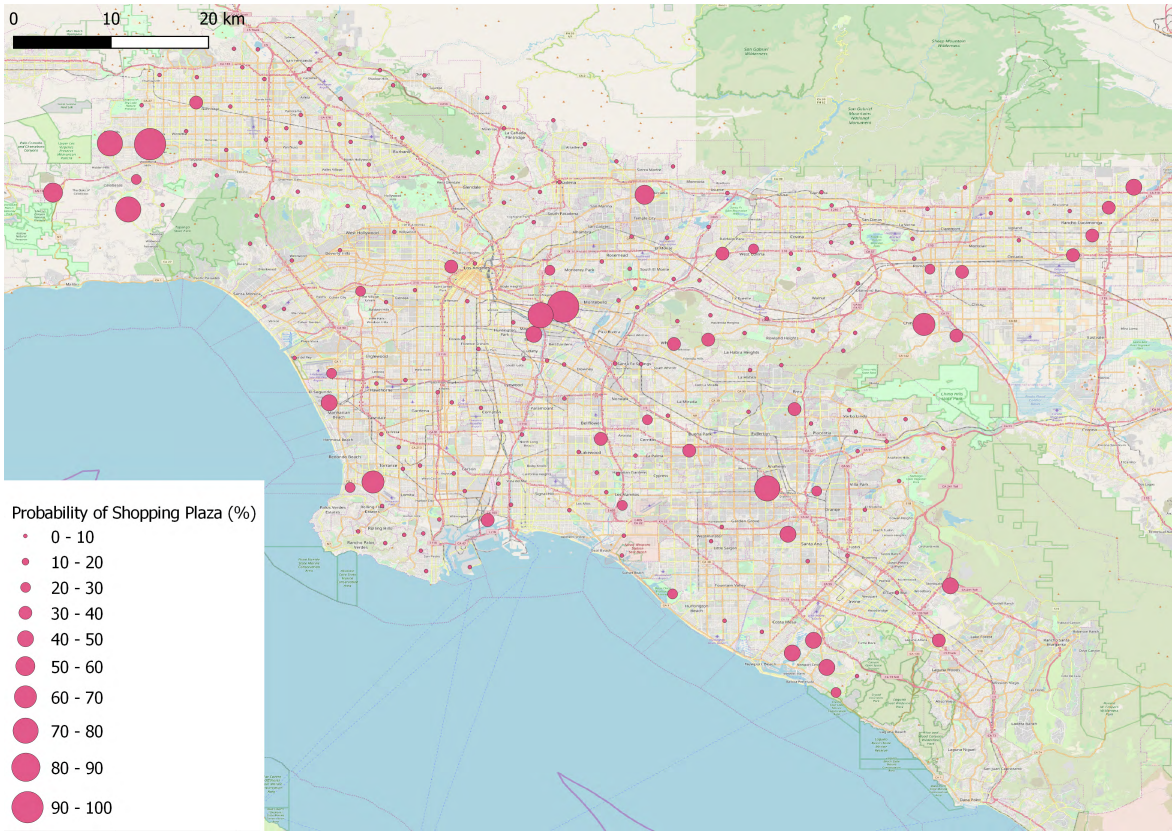

Figure 2. Shopping plazas in Los Angeles using LDA topic modeling.

For the knowledge-based approach we use the function pattern introduced in [14]. In particular, a region is considered as a candidate shopping plaza if it supports the fundamental functions of "shopping opportunities" and "walkability". Each candidate, then, is evaluated against various secondary functions, such as: "sustenance", "entertainment", "accessibility to drivers" and so on and the final score is calculated. For the purposes of the current demonstration we slightly extend the pattern in [14] with additional functions and adjust some of the existing rules. Tables A2 and A3 in Appendix A present the necessary components and the revised version of the pattern used. Figure 3 depicts the result of the approach using size-scaling for visualization purposes. Note the difference in semantics between the two approaches: every candidate region is identified and graded based on its likelihood to operate as a "shopping plaza" using the knowledge-based approach, as opposed to being one, according to the data-driven approach. This explains the higher number of regions included in Figure 3, than in Figure 2: it is more common to satisfy the essential shopping-related functions within the pattern than for the topic of " shopping plaza" to be significantly more probable than all other topics.

4.3. Results using Mutual Evaluation

The Pearson correlation coefficient value for the particular set of results is equal to 0.387. This indicates a positive association between the distributions of the confidence values of each approach. To facilitate comparison between the regions identified by each approach, Figure 4 overlays the results of both approaches over a grid (500x500 m^2). Darker hues indicate higher probability of the region being a "shopping plaza", with red and gray colours denoting results using the data-driven and knowledge-based approach, respectively.

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 385; doi:10.3390/ijgi8090385

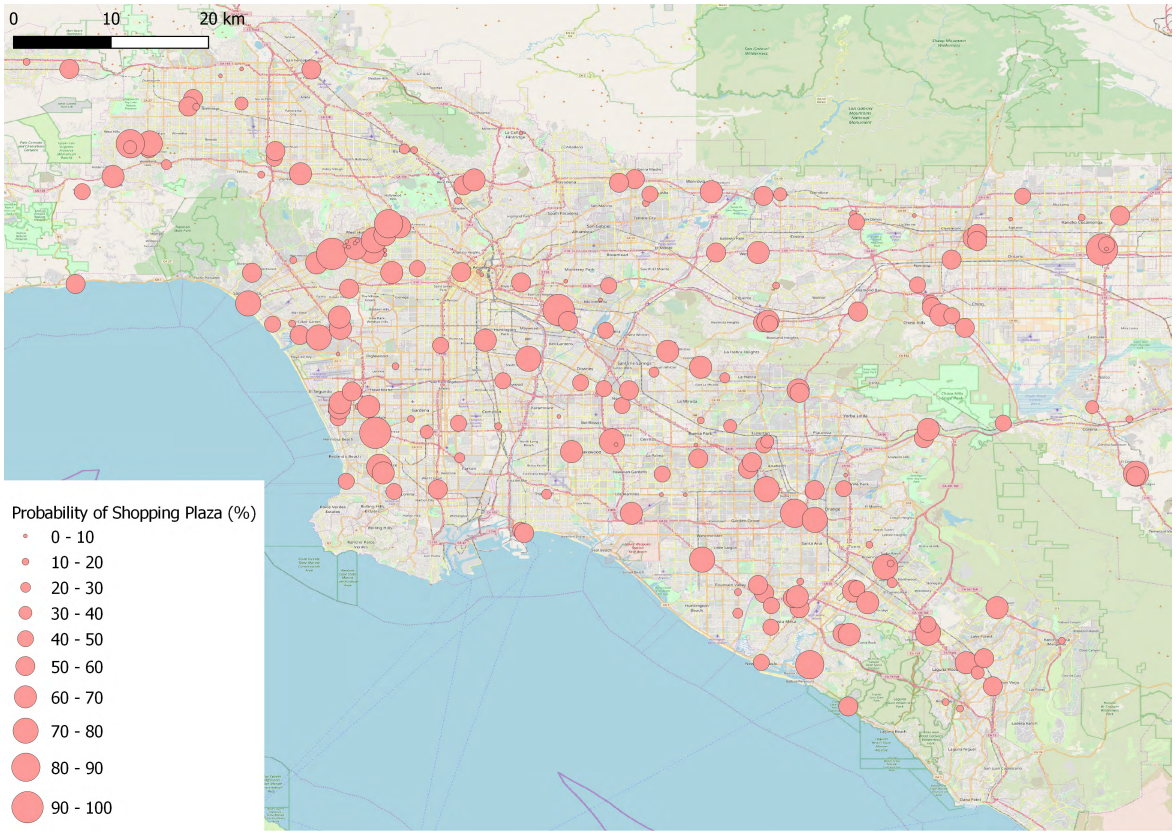

Figure 3. Shopping plazas in Los Angeles using knowledge patterns and the function-based model.

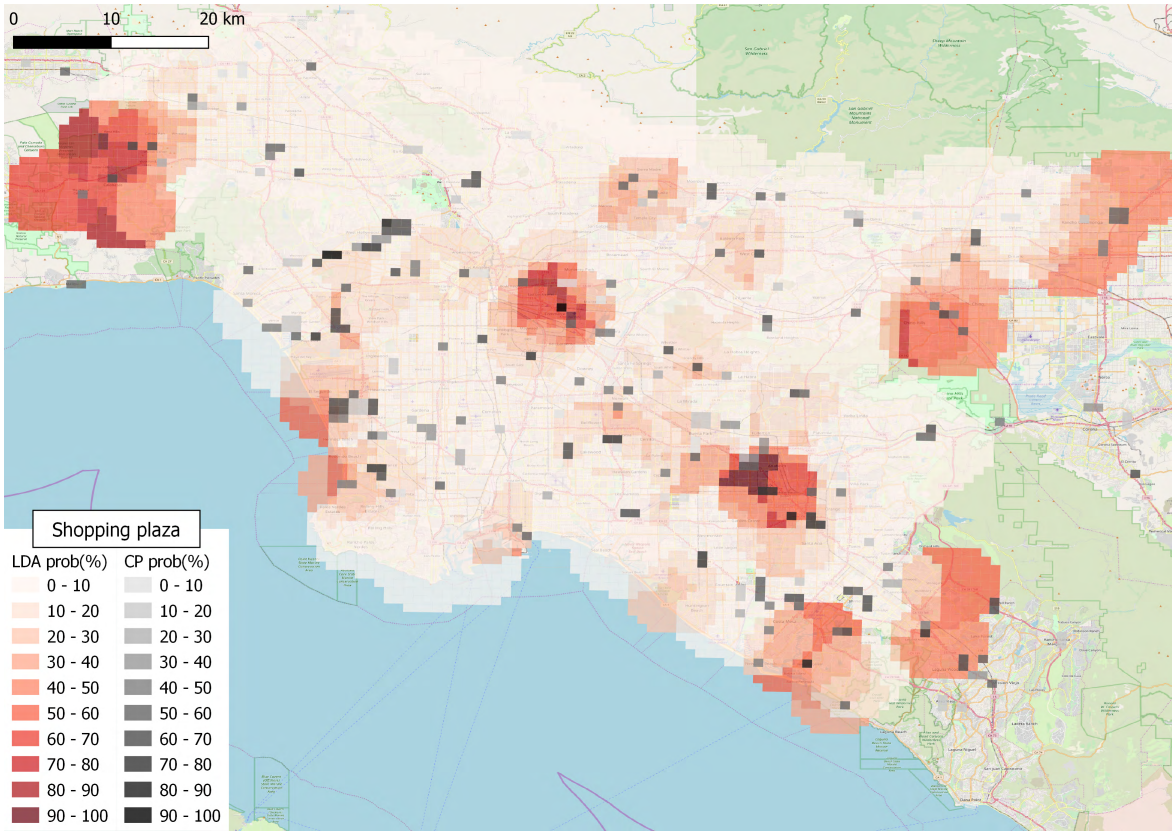

Figure 4. Distribution of shopping plazas in Los Angeles using both approaches.

Following the process described in Section 3.3.1, we then identify cases of high agreement by multiplying the individual confidence values and produce the map shown in Figure 5. Note that values are again scaled to 0-100 to facilitate comparison. Given the fact that the LDA approach alone returns less results than the pattern-based one, areas of significant agreement mainly converge around regions that have been identified by LDA.

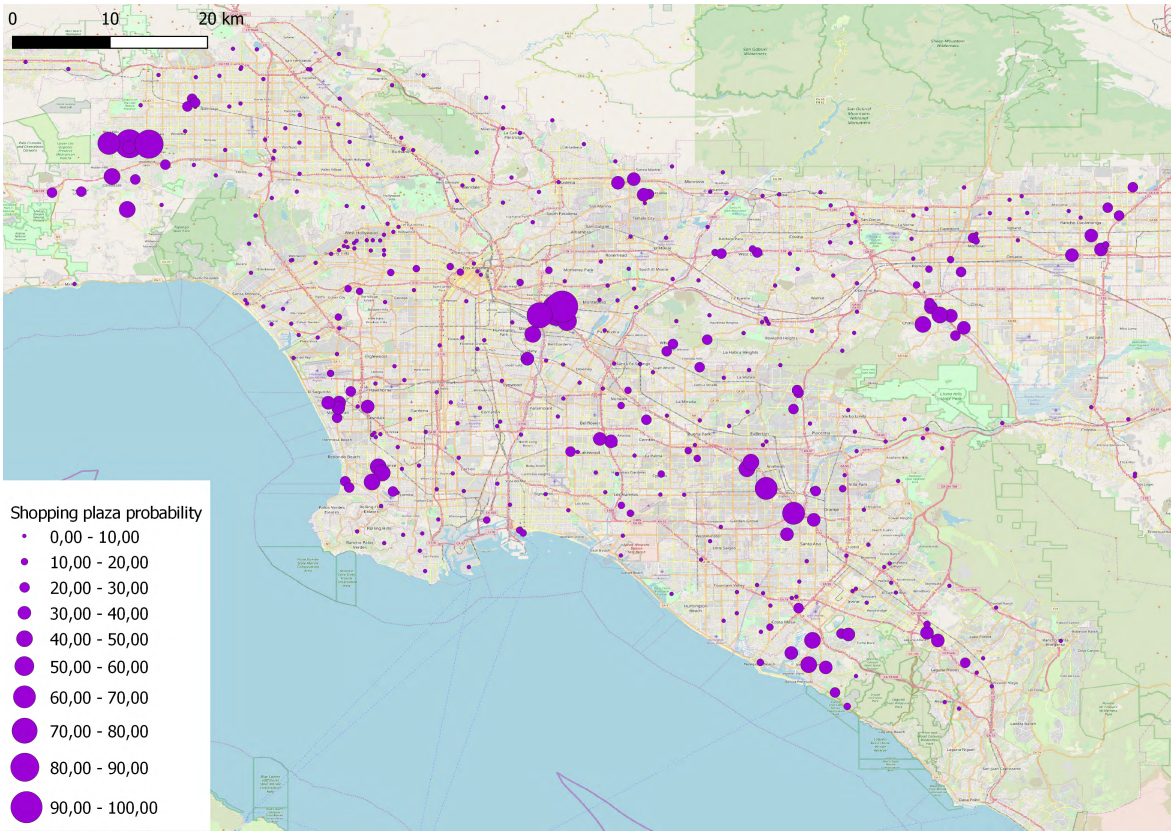

Figure 5. Shopping plazas with significant agreement between the two approaches.

Finally, we identify cases of significant disahgreement by calculating the differences between the confidence values produced using each approach; these are shown in Figure 6. The dark grey regions are cases where the knowledge-based approach attributes very low (or zero) likelihood for the region to operate as a "shopping plaza", whereas the data-driven approach gives high probability (up to 100). Confidence values of each approach are attached to these regions. On the other hand, the red regions are cases where the confidence value of the knowledge-based approach is very high (88.9 to 100), but the probability using the data-driven approach is very low (0 to 0.21). In these cases, a pie chart is provided, showing how each category of sub-functions within the pattern contributes to the confidence value. Note that the bold points within the red regions indicate the exact locations where the pattern-based method identifies a fully-functional shopping plaza. A discussion of the possible reasons behind these cases of significant disagreement is offered in Section 5.

4.4. Results using Data to Knowledge and Knowledge to Data Fusion

Starting with the confidence values calculated using LDA topic modeling, we apply the equation in Section 3.3.2 and the results are shown in Figure 7. As can be gathered from comparison with Figures 2 and 6, the regions that were previously missed are now included and all region probabilities are adjusted depending on the level of agreement or disagreement.

The opposite direction is followed for the final fusion method. Confidence values calculated using the pattern-based approach are adjusted using the formula in Section 3.3.3 in order to take into account

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 July 2019 doi:10.20944/preprints201907.0267.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 385; doi:10.3390/ijgi8090385

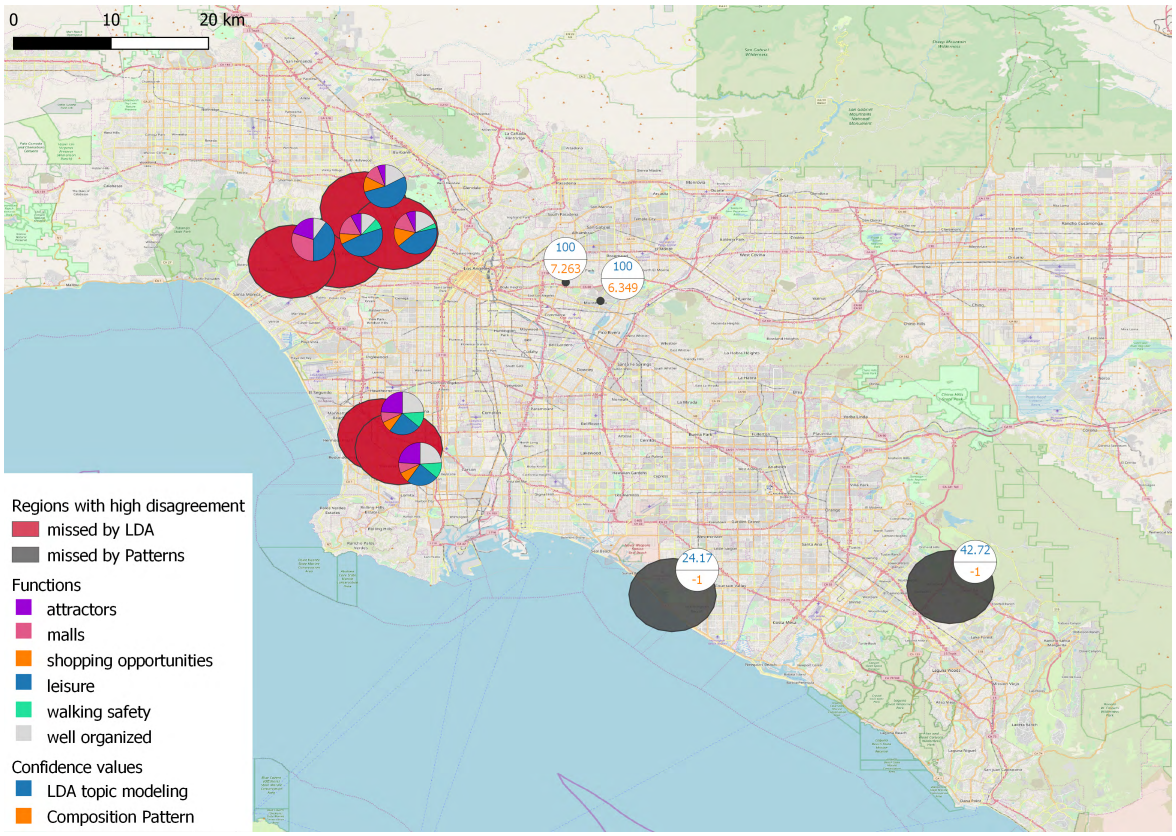

Figure 6. Regions where there is significant disagreement between approaches.

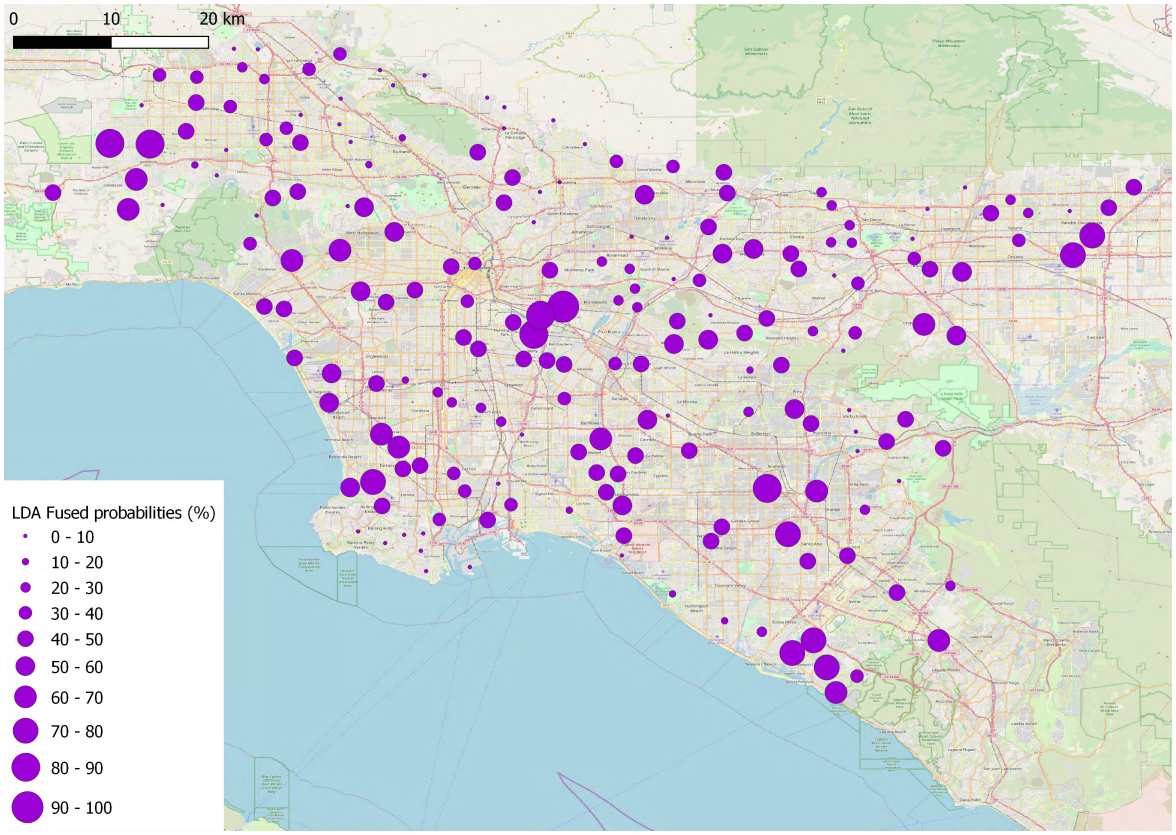

Figure 7. Adjusted results of LDA topic modeling according to knowledge-based scores.

the results of LDA topic modeling. Results are shown in Figure 8. Section 5 discusses these results in comparison to the ones in Figure 3.

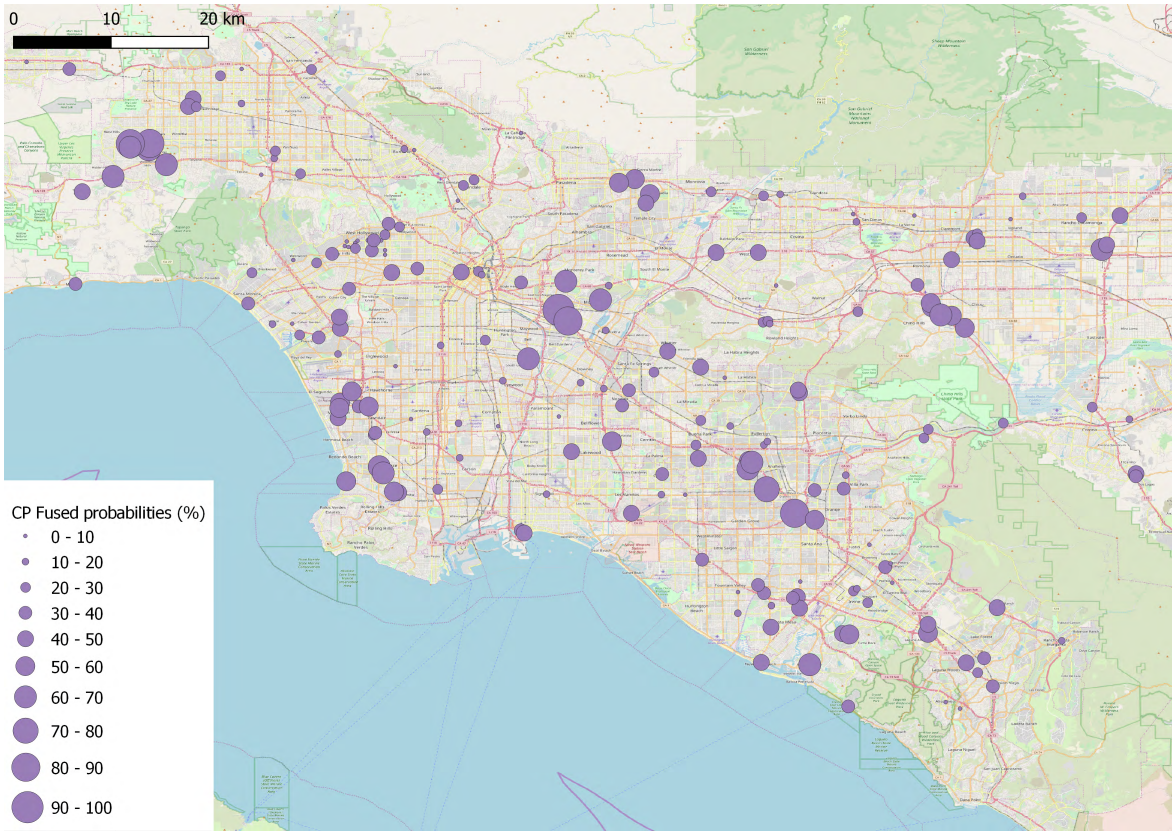

Figure 8. Adjusted results of pattern-based approach according to data-driven probabilities.

As a final result of the latter two fusing processes, we provide in Figure 9 an overall identification of regions functioning as "shopping plazas" by taking the average of the two adjustments and keeping only those regions with a probability higher than 50%, accompanied with an aggregation of the functions that can be found there. For comparison purposes, we also show the same result before applying any fusion-related adjustment in Figure 10; this includes those results from both approaches that overlap and score higher than 50%. The number of identified regions is clearly increased, while adjustments have been made to each region, with regard to their extent, attached probabilities, the location of core functionality and the distribution of sub-functions.

5. Discussion

The results in Figure 4 illustrate the different foundations of each approach, in terms of delineation of functional regions. The approach using patterns based on the function-based model of place searches for specific areas whose components and composition are capable of satisfying the supportive functions contained in the pattern. The LDA topic modelling approach, on the other hand, is capable of identifying the wider regions within which one may find the requested functionality with online social activity evidence, based on co-location of POI types and their popularity. The result of overlaying these delineations resembles the egg-yolk representation [27], which is even more evident in Figures 9 and 10. In particular, the regions identified from the data-driven approach represent an outer boundary ("egg") with the semantics that there is a chance of finding a "shopping plaza" within. The results of the knowledge-based search determine the inner boundaries of the sub-regions with the highest functionality, which resemble the core of the parent functional region ("yolk"). This lends support to the argument that the combination of knowledge and data may prove beneficial to the long-standing

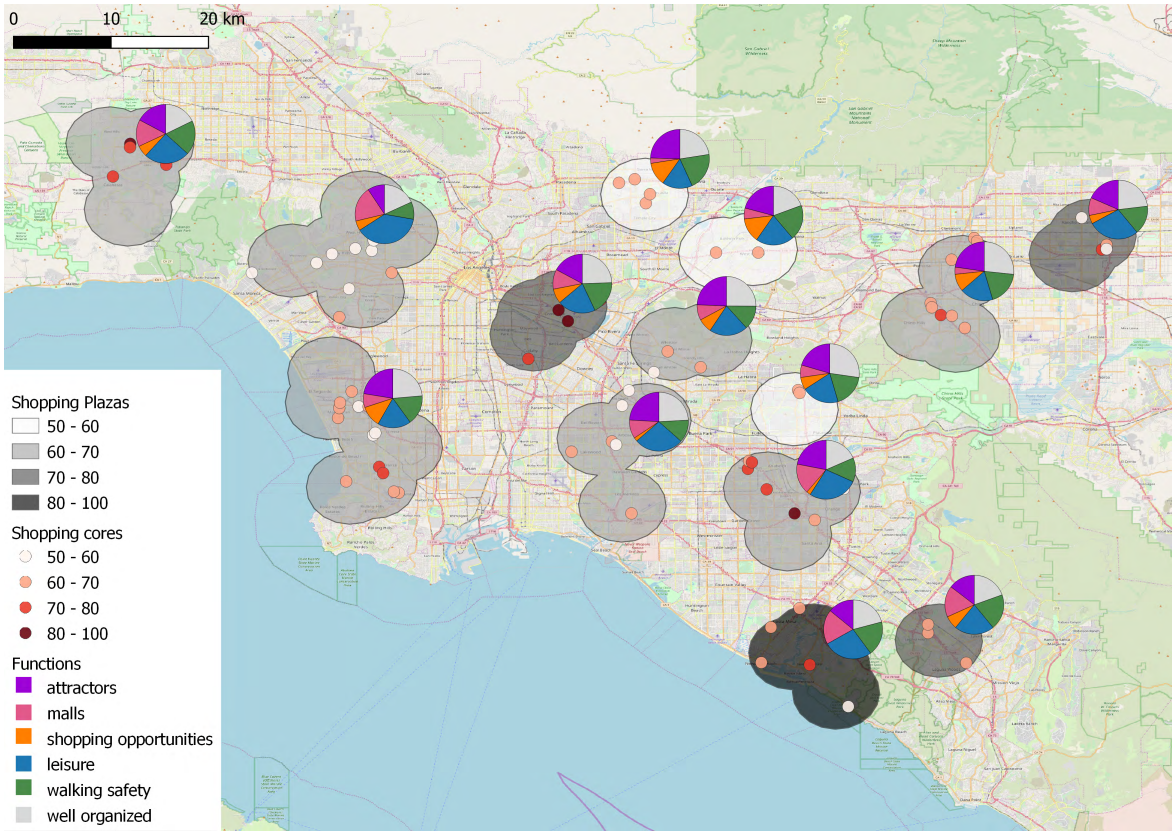

Figure 9. Final results combining data-to-knowledge and knowledge-to-data fusion.

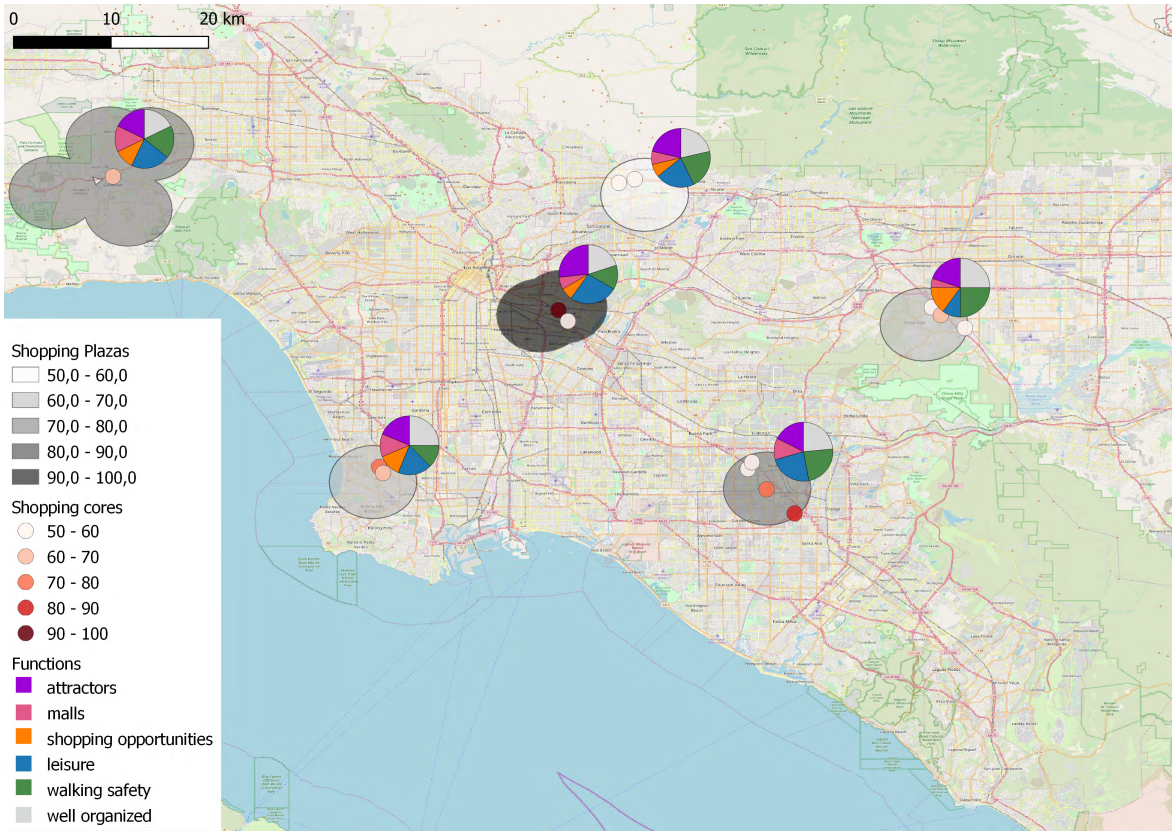

Figure 10. Results without using any of the proposed fusion methods.

problem in Geographic Information Science of delineating and modeling vaguely defined regions of which cognitive regions, functional regions and places are the most prominent examples [28,29].

A common characteristic of all methodologies to identify functional regions is the extreme difficulty (or impossibility) of acquiring ground truth, since they are dealing with highly subjective notions derived from human understanding or perception. Figures 5 and 6 indicate that the proposed mutual evaluation process can provide a useful substitute. On one hand, the results in Figure 5 help reinforce the identification of those regions that are most highly accepted as solutions, based on all the available information. In the particular example, this includes regions around East Los Angeles, Canoga Park, Torrance and Anaheim (around Disneyland Park), all of which can be argued to be widely known shopping districts.

On the other hand, Figure 6 serves as a way to detect regions that were missed by either approach due to their individual limitations. For instance, significant regions that were excluded from the LDA-based approach are those around West Hollywood and Beverly Hills: as shown in the included pie charts, all of these regions satisfy functionality directly related to shopping plazas. However, because of a higher co-occurrence (and social media popularity) of POI types related to leisure, as opposed to shopping, these regions were associated with a different topic (related to restaurants and bars). The pattern-based approach did not include regions around Sunset Beach and Northwood: while an adequate number of shopping-related POIs is contained, their spatial organization does not satisfy most (or any) of the functional implications in the pattern. Also, in the cases of Monterey Park and Montebello, while the wider area is popular and provides several shopping-related opportunities (both of which are captured by the LDA-based approach), the assumptions behind the pattern-based approach restrict its focus on a much narrower scale, hence attributing lower scores.

A comparison of Figures 2 and 7 shows that the data-to-knowledge fusion process leads to an inflation of confidence values throughout the area of study. This allows the aforementioned missed areas to be included, since the higher co-occurrence of non shopping-related POIs is counterbalanced by their spatial configuration, which, according to the defined pattern, facilitates the desired functionality. This enables the inclusion of more relevant regions, without, however, leading to an over-inclusion of results. The knowledge-to-data fusion process achieves similar results, but in the reverse direction, as evidenced by Figures 3 and 8. Confidence values are, in general, deflated, allowing a more clear identification of the most popular regions, due to the inclusion of social media data exploited by the data-driven approach.

A comparison of Figures 9 and 10 clearly shows the benefits of a functional region identification approach that fuses knowledge and data. Compared to what can be gathered by simply combining and overlaying best results from either approach, the end result in Figure 9 identifies functional regions of the type “shopping plaza” that:

- are highly functional, also explaining which particular functions mostly contribute to this, as derived from the knowledge-based aspect;
- are popular, based on the inclusion of social media information exploited by the data-driven aspect;
- are homogeneous both in terms of the POIs included and the way they are spatially organized.

The presented results indicate that trusting exclusively either of the two approaches may lead to some results being missed or some other being overly highlighted. By using the fusion methodologies, the results of one approach serve as a “bias” to challenge the “authority” of the other approach. The overall aim moving forward would be to realise fusion earlier, during the identification process and not as a post-processing step, resulting into a truly hybrid methodology. This would potentially lead to more harmonized results and provide a more realistic view that is neither entirely confined by pattern rules, nor exclusively governed by statistical analysis of data. This is a very interesting future research avenue that we fully intend to explore.

6. Conclusion

In this work, we propose a novel framework for the identification of urban functional regions that fuses two previously independent research pathways, one top-down and one bottom-up. The top-down, knowledge-driven approach relies on design patterns created based on expert knowledge on urban design and planning. The bottom-up, data-driven approach discovers semantically meaningful topics based on co-occurrence patterns of POI types, incorporating user check-ins on social networks. Three types of fusion are examined: (1) mutual evaluation, where the results of the two approaches are compared to discover cases of significant agreement and disagreement; (2) use of knowledge patterns to adjust topic probabilities produced by the data-driven approach; and (3) use of topic probabilities derived from data to adjust scores calculated using the knowledge-driven approach. The synergy between knowledge and data allows for improved results in functional region identification, as evidenced by the conducted experiment on identifying "shopping plaza" regions in the Los Angeles metropolitan area. Mutual evaluation can help identify cases where the drawbacks of either approach lead to regions included or excluded incorrectly, while using one approach to adjust the results of the other leads to improved overall accuracy.

The presented framework is a first attempt at exploring how the lines of knowledge-based and data-driven work in [14] and [10] can be brought together, by largely keeping the individual methodologies intact while using their results to either evaluate or adjust each other. In the future, we first intend to conduct additional experiments incorporating additional urban areas. We also plan to explore a tighter integration between the two methodologies with the aim of proposing a unified hybrid methodology that exploits both knowledge and data internally. For instance, knowledge (either raw or encoded in a pattern) can be used to adapt the LDA process itself, e.g. by rescaling the document-word frequency matrix, as is done using check-in data. Alternatively, VGI data can be used to adjust knowledge-based patterns, an approach similar in spirit to the empirical and probabilistic patterns proposed in [20].

Author Contributions: Conceptualization, E.P., S.G. and G.B.; methodology, E.P., S.G. and G.B..; software, E.P. and S.G.; validation, E.P. and S.G.; formal analysis, E.P., S.G. and G.B.; investigation, E.P., S.G. and G.B.; resources, E.P. and S.G.; data curation, E.P. and S.G.; writing–original draft preparation, E.P. and G.B.; writing–review and editing, E.P., S.G. and G.B.; visualization, E.P., S.G. and G.B.; supervision, E.P.; project administration, E.P.

Funding: This research is framed within the Doctoral College GIScience (DK W 1237N23), funded by the Austrian Science Fund (FWF).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. Top-15 ranked POI types for "shopping plaza" topic in [10].

Category	Probability	Category	Probability
shopping mall	0.207709	bistro	0.000105
accessories store	0.056738	dumpling restaurant	0.000096
chocolate shop	0.013896	korean restaurant	0.000090
shoe store	0.000288	german restaurant	0.000080
breakfast spot	0.000282	herbs & spices store	0.000079
gaming cafe	0.000196	airport terminal	0.000078
optical shop	0.000180	outlet store	0.000076
post office	0.000114		

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 July 2019 doi:10.20944/preprints201907.0267.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 385; doi:10.3390/ijgi8090385

Table A2. Set of components associated with a shopping plaza

Variable	Component	Filter
C_S	Shop	$Type\_Filter(\text{“}Shop\text{”})$
C_A	Amenity	$Type\_Filter(\text{“}Amenity\text{”})$
C_F	Facilities	$C_S \cup C_A$
C_{WP}	Walkable plaza	$Type\_Filter(\text{“}Surface\text{”}) \cap Prop\_Filter(\text{“}walkable\text{”}, \text{“}true\text{”})$
C_H	Motorway	$Type\_Filter(\text{“}Road\text{”}) \cap Prop\_Filter(\text{“}pedestrians\text{”}, \text{“}false\text{”})$
C_{Sr}	Service Road	$C_H \cap Prop\_Filter(\text{“}pedestrians\text{”}, \text{“}true\text{”})$
C_W	Walkable	$C_{WP} \cup C_{Sr}$
C_P	Parking place	$C_A \cap Prop\_Filter(\text{“}service\text{”}, \text{“}parking\text{”})$
C_B	Transportation node	$C_A \cap Prop\_Filter(\text{“}service\text{”}, \text{“}transportation\text{”})$
C_{An}	Anchor Store	$C_S \cap Prop\_Filter(\text{“}goods\text{”}, \text{“}various\text{”})$
C_M	Mall	$C_S \cap Prop\_Filter(\text{“}goods\text{”}, \text{“}various\text{”}) \cap Prop\_Filter(\text{“}service\text{”}, \text{“}various\text{”})$
C_{At}	Attractors	$C_M \cap C_{An}$
C_{Sb}	Basic Shop	$C_S \cap Prop\_Filter(\text{“}goods\text{”}, \text{“}basic\text{”})$
C_{Se}	Special Shop	$C_S \cap Prop\_Filter(\text{“}goods\text{”}, \text{“}special\text{”})$
C_{Su}	Uncommon Shop	$C_F \cap (Prop\_Filter(\text{“}goods\text{”}, \text{“}uncommon\text{”}) \cup Prop\_Filter(\text{“}services\text{”}, \text{“}uncommon\text{”}))$
C_{As}	Food court	$C_A \cap Prop\_Filter(\text{“}service\text{”}, \text{“}sustenance\text{”})$
C_{Ae}	Entertainment	$C_A \cap Prop\_Filter(\text{“}service\text{”}, \text{“}entertainment\text{”})$
C_{Al}	Luxury services	$C_A \cap Prop\_Filter(\text{“}service\text{”}, \text{“}health\&beauty\text{”})$
C_{Av}	Aesthetics	$C_A \cap Prop\_Filter(\text{“}service\text{”}, \text{“}visuallypleasing\text{”})$

Table A3. Composition pattern of a shopping plaza

$\mathcal{CMP}$:$\{C_F, C_S, C_A, C_H, C_{Sr}, C_W, C_{WP}, C_B, C_P, C_{An}, C_M, C_{At}, C_{Sb}, C_{Se}, C_{Su}, C_{As}, C_{Ae}, C_{Al}, C_{Av}\}$

Functional Implications

Functions ($\mathcal{F}$)	Logical Formula
$F_W(C_{Sb}, C_{At}, C_W, C_{Sr})$ (Walkability)	$Occurrence(C_W, \mathbb{N}) \wedge ((Occurrence(C_{Sb}, [5, \infty)) \wedge Proximity(C_{Sb}, C_{Sb}, (0, 500m])) \rightarrow S\_Relation(C_W, C_{Sb}, [intersects]) \vee (Occurrence(C_{At}, [1, \infty)) \rightarrow S\_Relation(C_W, C_{At}, [intersects]))$
$F_{SE}(C_{At}, C_{Sb}, C_W)$ (Shopping Experience)	$F_W \wedge (Occurrence(C_{Sb}, [5, \infty) \wedge S\_Relation(C_W, C_{Sb}, [intersects])) \vee (Occurrence(C_{At}, [1, \infty) \wedge S\_Relation(C_W, C_{At}, [contains]))$
$F_{SV}(C_{Sb})$ (Shopping Variety)	$F_{SE} \wedge Occurrence(C_{Sb}, [5, \infty))$
$F_{AT}(C_{Sb})$ (Sh. Attractiveness)	$F_{SE} \wedge Occurrence(C_{At}, [1, \infty))$
$F_{SD}(C_{Sb}, C_{Se})$ (Sh. Orientation)	$F_{SE} \wedge Correlation(C_{Sb}, C_{Se}, [2, \infty))$
$F_{SG}(C_{Se})$ (Special Goods)	$F_{SE} \wedge Occurrence(C_{Se}, \mathbb{N})$
$F_{CC}(C_{Sb}, C_{At}, C_{Su}, C_W)$ (Compatible Components)	$F_{SE} \wedge Occurrence(C_{Su}, \mathbb{N}) \rightarrow Correlation(C_{Sb} \cup C_{At}, C_{Su}, [5, \infty)) \vee Proximity(C_W, C_{Su}, [500, \infty m]))$
$F_{SO}(C_S, C_A)$ (Shopping Opportunities)	$F_{SE} \wedge Occurrence(C_A, \mathbb{N}) \rightarrow Correlation(C_S, C_A, [2, \infty))$
$F_L(C_{As})$ (Leisure)	$F_{SO} \rightarrow Occurrence(C_{As}, \mathbb{N})$
$F_E(C_{Ae})$ (Entertainment)	$F_{SO} \rightarrow Occurrence(C_{Ae}, \mathbb{N})$
$F_{LS}(C_{Al})$ (Luxury Services)	$F_{SO} \wedge Occurrence(C_{Al}, \mathbb{N})$
$F_{Resupply}(C_W, C_H)$	$F_{SE} \wedge (Occurrence(C_H, \mathbb{N}) \wedge Proximity(C_W, C_H, [0, 1000m]))$
$F_{AD}(C_W, C_P)$ (Access to Drivers)	$F_W \wedge Occurrence(C_P, [1, \infty]) \rightarrow (S\_Relation(C_W, C_P, [intersects])) \vee Proximity(C_W, C_P, [0, 200m])$
$F_{AN}(C_W, C_B)$ (Access to Non-drivers)	$F_W \wedge Occurrence(C_B, [1, \infty]) \rightarrow (S\_Relation(C_W, C_B, [intersects])) \vee Proximity(C_W, C_B, [0, 200m])$
$F_{WS}(C_H, C_W)$ (Walking Safety)	$F_W \wedge Occurrence(C_H, \mathbb{N}) \rightarrow S\_Relation(C_W, C_H, [disjoint]))$
$F_{WO}(C_S, C_A)$ (Well-Organized)	$F_{SE} \wedge Occurrence(C_A, \mathbb{N}) \rightarrow S\_Configuration(C_S, C_A, [clustered])$
$F_{VP}(C_{Av}, C_W)$ (Visually Pleasing)	$F_W \wedge Occurrence(C_{Av}, \mathbb{N}) \rightarrow (S\_Relation(C_W, C_H, [intersects]) \vee Proximity(C_W, C_{Av}, [0, 200m]))$

Scoring Function

$F_{SE} * F_W * (F_{SD} + F_{SO} + F_{SA} + F_{SG} + F_L + F_E + F_{LS} + F_{AD} + F_{AN} + F_R + F_{WS} + F_{VP} + F_{WO}) * error$

References

1. Tuan, Y.F. Space and Place: Humanistic Perspective. In *Philosophy in geography*; Springer, 1979; pp. 387–427.
2. Yuan, J.; Zheng, Y.; Xie, X. Discovering regions of different functions in a city using human mobility and POIs. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 186–194.
3. Goodchild, M.F. Geographical information science. *International journal of geographical information systems* **1992**, *6*, 31–45.
4. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers* **2015**, *105*, 512–530. doi:10.1080/00045608.2015.1018773.
5. Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.L.; Li, T.; Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science* **2014**, *28*, 1988–2007. doi:10.1080/13658816.2014.913794.
6. Hill, L.L. Core elements of digital gazetteers: placenames, categories, and footprints. International Conference on Theory and Practice of Digital Libraries. Springer, 2000, pp. 280–290.
7. Purves, R.S.; Clough, P.; Jones, C.B.; Arampatzis, A.; Bucher, B.; Finch, D.; Fu, G.; Joho, H.; Syed, A.K.; Vaid, S.; Yang, B. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science* **2007**, *21*, 717–745. doi:10.1080/13658810601169840.
8. Papadakis, E.; Blaschke, T. Place-based GIS: Functional Space. Proceedings of the 4th AGILE PhD School; Comber, L.; Malleson, N., Eds. CEUR, 2017, Vol. 2208.
9. Adams, B.; Janowicz, K. Thematic signatures for cleansing and enriching place-related linked data. *International Journal of Geographical Information Science* **2015**, *29*, 556–579.
10. Gao, S.; Janowicz, K.; Couclelis, H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **2017**, *21*, 446–467.
11. Hobel, H.; Fogliaroni, P.; Frank, A.U. Deriving the Geographic Footprint of Cognitive Regions. AGILE Conference; Sarjakoski, T.; Santos, M.Y.; Sarjakoski, L.T., Eds. Springer, 2016, Lecture Notes in Geoinformation and Cartography, pp. 67–84.
12. Boegl, K.; Adlassnig, K.P.; Hayashi, Y.; Rothenfluh, T.E.; Leitich, H. Knowledge acquisition in the fuzzy knowledge representation framework of a medical consultation system. *Artificial Intelligence in Medicine* **2004**, *30*, 1–26.
13. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning, 2017, [arXiv:1702.08608].
14. Papadakis, E.; Resch, B.; Blaschke, T. Composition of Place: Towards a Compositional View of Functional Space. *Cartography and Geographic Information Science* **2019**.
15. Scheider, S.; Janowicz, K. Place reference systems. *Applied Ontology* **2014**, *9*, 97–127. doi:10.3233/AO-140134.
16. Janowicz, K.; Keßler, C. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science* **2008**, *22*, 1129–1157.
17. Scheider, S.; Purves, R. Semantic Place Localization from Narratives. Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place; Scheider, S.; Adams, B.; Janowicz, K.; Vasardani, M.; Winter, S., Eds.; ACM: New York, 2013; pp. 16:16–16:19. doi:10.1145/2534848.2534858.
18. MacEachren, A.M. Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier. Spatial Data Handling in Big Data Era: Select Papers from the 17th IGU Spatial Data Handling Symposium 2016; Zhou, C.; Su, F.; Harvey, F.; Xu, J., Eds.; Springer Singapore: Singapore, 2017; pp. 139–155. doi:10.1007/978-981-10-4424-3_10.
19. Papadakis, E.; Resch, B.; Blaschke, T. A Function-based model of Place. International Conference on GIScience Short Paper Proceedings, 2016.
20. Papadakis, E.; Baryannis, G.; Petutschnig, A.; Blaschke, T. Function-Based Search of Place Using Theoretical, Empirical and Probabilistic Patterns. *ISPRS International Journal of Geo-Information* **2019**, *8*.
21. Hobel, H.; Abdalla, A.; Fogliaroni, P.; Frank, A.U. A Semantic Region Growing Algorithm: Extraction of Urban Settings. AGILE Conf.; Bação, F.; Santos, M.Y.; Painho, M., Eds. Springer, 2015, Lecture Notes in Geoinformation and Cartography, pp. 19–33.

22. Noulas, A.; Scellato, S.; Mascolo, C.; Pontil, M. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. The Social Mobile Web. AAAI, 2011, Vol. WS-11-02, *AAAI Workshops*.
23. Zhou, X.; Zhang, L. Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartography and Geographic Information Science* **2016**, *43*, 393–404. doi:10.1080/15230406.2015.1128852.
24. Zhi, Y.; Li, H.; Wang, D.; Deng, M.; Wang, S.; Gao, J.; Duan, Z.; Liu, Y. Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data. *Geo-spatial Information Science* **2016**, *19*, 94–105. doi:10.1080/10095020.2016.1176723.
25. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1; Le Cam, L.M.; Neyman, J., Eds. University of California Press, Berkeley, CA, USA, 1967, pp. 281–297.
26. ao, R.M.A.; Neves, M.C.; Câmara, G.; Freitas, C.D.C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science* **2006**, *20*, 797–811. doi:10.1080/13658810600665111.
27. Cohn, A.; Gotts, N. The 'Egg-Yolk' Representation of Regions with Indeterminate Boundaries. In *Geographic Objects with Indeterminate Boundaries*; Burrough, P.A.; Frank, A.U., Eds.; Taylor & Francis, 1995; pp. 171–187.
28. Mai, G.; Janowicz, K.; Hu, Y.; Gao, S.; Zhu, R.; Yan, B.; McKenzie, G.; Uppal, A.; Regalia, B. Collections of Points of Interest: How to Name Them and Why it Matters. *Spatial Big Data and Machine Learning in GIScience* **2018**, p. 29.
29. Liu, Y.; Yuan, Y.; Gao, S. Modeling the Vagueness of Areal Geographic Objects: A Categorization System. *ISPRS International Journal of Geo-Information* **2019**, *8*, 306.