ISPRS Int. J. Geo-Inf. **2015**, *4*, 2109-2130; doi:10.3390/ijgi4042109

OPEN ACCESS

ISPRS International Journal of
Geo-Information
ISSN 2220-9964
www.mdpi.com/journal/ijgi

Article

Extracting Urban Land Use from Linked Open Geospatial Data

Gloria Re Calegari *, Emanuela Carlino, Diego Peroni and Irene Celino

ICT Center of Excellence For Research, Innovation, Education and industrial Labs partnerships (CEFRIEL), Politecnico di Milano, *via* Fucini 2, 20133 Milano, Italy;
E-Mails: carlino@cefriel.it (E.C.); diego.peroni@cefriel.com (D.P.); irene.celino@cefriel.com (I.C.)

* Author to whom correspondence should be addressed; E-Mail: gloria.re@cefriel.com;
Tel.: +39-02-23954266.

Academic Editors: Jochen Schiewe and Wolfgang Kainz

Received: 30 June 2015 / Accepted: 7 October 2015 / Published: 20 October 2015

Abstract: The ever-increasing availability of linked open geospatial data provides an unprecedented source of geo-information to describe urban environments. This wealth of data should be turned into actionable knowledge: for example, open data could be used as a proxy or substitute for closed or expensive information. The successful employment of linked open geospatial data can pave the way for innovative solutions to smart city problems. In this paper, we illustrate a set of experiments that, starting from linked open geospatial data, execute a knowledge discovery process to predict urban semantics. More specifically, we leverage geo-information about points of interests as input in a classification model of land use at a moderate spatial resolution (250 meters) over wide urban areas in Europe. We replicate our experiments in different European cities—Milano, München, Barcelona and Brussels—to ensure the repeatability and generality of our approach, and we explain the experimental conditions, as well as the employed datasets to guarantee reproducibility. We extensively report on quantitative and qualitative evaluation results, to judge the validity, as well as the limitations of our proposed approach.

Keywords: urban land use; linked open geo-spatial data; points of interest; smart cities

1. Introduction

Urban space digitization, caused by the ever-increasing pervasiveness of information and communication technologies, has led to a rich ecosystem of information producers and information consumers [1]. This wealth of data, however, risks hiding the latent added value that the smart management of such information can bring to cities. The open question remains "what is the relevant information to achieve my objective?": information consumers could be flooded and get lost in big data without solving their tasks. In this picture, linked open geospatial data can play an important role, providing rich semantic geo-information related to urban environments. As happens in other contexts, turning rough information into actionable knowledge is the key challenge [2].

Moreover, the rise of the so-called data-driven economy shows that ownership and transfer of data lead to several legal and financial considerations [3]. Furthermore, in the context of urban data, different sources come at different costs: while some information is created and updated almost for free, as a consequence of other activities (e.g., vehicle mobility generates GPS traces, mobile phone activity collects people location over time [4]), other datasets still remain very expensive to produce and maintain, like those requiring manual intervention (e.g., demographics data, which requires a human-based census activity) or semi-automatic processing. The datasets' costs are related to the entire data-driven value chain, from production to distribution, from processing to consumption [5].

Would it be possible to use one or more inexpensive datasets as a "proxy" for more expensive data sources? In other words, would it be possible to (semi)automatically generate or revise an outdated dataset, which otherwise would require costly human work, on the basis of the content of other up-to-date information sources? This is the challenge that the data management community has to face today. Our research aims to answer the above questions in the context of land use planning.

Urban planning is the discipline that deals with the improvement of people's and communities' welfare by creating more convenient, sustainable and attractive places. When applied to land use, urban planning regards the management and the modifications to the environment, and it is usually aimed at regulating the types of activities allowed to be accommodated in specific areas. In the context of cities, monitoring the changes of land use is of utmost importance to drive and support a sustainable urbanization process [6].

Detecting and reporting land use modifications is far from being a trivial task. Indeed, it usually requires an expensive and partially manual process to collect, integrate and make sense of urban information to derive useful knowledge to support the planning activities. In Europe, the CORINE initiative (*cf.* http://www.eea.europa.eu/publications/COR0-landcover) provided a shared classification and procedure to support local and regional undertakings that periodically produce or update land use maps, usually starting from remote sensing images.

The goal of this paper is to experiment on whether free linked open geospatial data related to urban environments can be used as a "proxy" for expensive land use geo-information sources with the meaning explained above, similarly to what [7–10] did by using social media, mobile phone data or GPS traces.

In the following, we present our innovative solution for extracting urban land use and support smart cities' planning activities. We illustrate a set of classification experiments to predict CORINE land use that employs cities' point of interest (POI) geo-information derived from OpenStreetMap, similar to [11,12], but with a finer spatial resolution.

In order to ensure the repeatability and generality of our approach, we replicate our experiments in different European cities. As CORINE maps are available for the whole of Europe, we select four cities that show similar characteristics in terms of the number of inhabitants and the urban area size. We choose Milano in Italy, München in Germany, Barcelona in Spain and Brussels in Belgium.

The CORINE initiative, coordinated by the European Environment Agency (*cf.* http://www.eea.europa.eu/) of the European Union, provides a multi-level taxonomy (named "land cover nomenclature", last updated in 2006) to classify the land use in various categories (water *vs.* land, land with plant cover *vs.* land without plant cover, *etc.*); this taxonomy is then used to describe the actual land use of the European territory, generating thematic maps at a quite fine-grained resolution (100 meters). Producing and updating CORINE maps is a long and expensive process [13]: images acquired by remote sensing are used as the main source to derive land cover information, through a series of photo-interpretation, ortho-correction and various quality assurance activities. CORINE maps are made available as raster data; for example, the December 2013 update is available online for download (*cf.* http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3). The procedures standardized and suggested by CORINE are then also implemented at the local and regional level, with a highly varying update frequency that often depends not only on the local urban planning strategy, but also on the availability of human and economic resources to complete the process.

Because of those costs, it would be beneficial to find alternative or additional "inexpensive" solutions to the land use update and classification process. In other words, we would like to explore the possibility to exploit datasets that are open and/or come for free and support producing or updating other expensive datasets, like CORINE thematic maps. In this paper, we experimentally test our hypothesis, by extracting urban land use from linked open geospatial data.

The most popular, as well comprehensive source on the so-called Geospatial Web is probably OpenStreetMap (OSM; *cf.* http://www.openstreetmap.org/), the free, editable and user-generated spatial data collection, also known as "the Wikipedia of maps". OpenStreetMap consists of a large spatial knowledge base, which is increasingly being collected and curated via volunteered geographic information efforts (VGI [14]), *i.e.*, through cooperative efforts that follow a citizen scientist or a crowdsourcing approach [15]. It includes a very rich characterization of the Earth in terms of spatial features (points and polygons), which are further described by a set of semi-controlled key-value pairs (e.g., `amenity:restaurant` or `leisure:park`). The Semantic Web community, by analyzing the use of those key-value pairs, created LinkedGeoData (LGD [16]) by mapping the OSM characterization into an ontology and re-publishing the OSM dataset as linked data [17].

The remainder of the paper is as follows: Section 2 illustrates the data preparation and pre-processing; Section 3 details the methodology followed in the experiments; Sections 4–6 present our experimental results, with a detailed discussion in Section 7; Section 8 aims at comparing our findings with state-of-the-art approaches; finally, in Section 9, we draw some conclusions.

2. Data Preparation

In this section, we illustrate the mix of manual choices and automated approaches that we go through in our experiments. We explain the selection and spatial characterization of the four urban areas, and then, to better understand the following sections, we give details on the input and output variables of our classification experiments.

2.1. Spatial Resolution of the Selected Urban Areas

The spatial resolution, as well as the spatial characterization of the urban areas of the four selected cities should be the same to test the generality of our approach. To provide for a reasonable and uniform spatial resolution, we adopt a regular grid of square cells to divide each urban space.

To have comparable land extents in the different cities, we select an area of 625 square kilometers for each city; the grid is therefore constituted by 10,000 square cells of 250 meters. This area includes both the whole metropolitan area and the surrounding villages and lands. As explained in the following, those urban grids are used in our experiments as a reference and uniform spatial characterization for the input and output datasets; in other words, each city provides 10,000 samples to train and test the urban land use classifiers.

2.2. Pre-Processing of CORINE Data

CORINE information is provided as raster data for the whole European territory. Therefore, we need to (1) focus on the selected cities and (2) project CORINE information onto the city grids. Both activities can be implemented through the use of a GIS, like QGIS.

The former activity is realized by extracting the geographic portion from the Europe-wide CORINE raster image, using the bounding box of the city grids. The latter activity applies an intersection operation between the CORINE layer and the city grid; the result is a vector layer that, for each grid cell of the city, specifies its composition with respect to the CORINE taxonomy. For example, a cell could be a mix of 60% dense residential area (Category 111, continuous urban fabric), 15% streets (Category 122, road and rail networks and associated land) and 25% park (Category 141, green urban areas). Because of the relatively small scale of the cell grid, we simplify the land use information by taking into account for each grid cell only its predominant land use, *i.e.*, the CORINE category that covers the largest share of the cell area (in the example above, that cell would have Category 111 as the predominant land use, since it represents 60% of the cell surface).

The CORINE land use categories are described with a quite detailed multi-level taxonomy including more than 40 classes at the deepest hierarchy level [13]. For the purpose of our classification experiments, however, we prefer to reduce the number of expected output classes with respect to the original 40+ CORINE categories. Indeed, usually, classification algorithms handle the multi-class problem as multiple binary classification problems: using the one-against-one approach, $k(k-1)/2$ binary classifiers are trained, where k is the number of classes (*cf.* [18] for SVM multi-class classification). The selection of the suitable land use categories is done as follows.

Given that each cell is described by a vector of percentages belonging to each land use, we perform K-means clustering [19] on these vectors, in order to understand how these values naturally group together and, so, which kind of land uses could be put together. We repeat the clustering on all cities, obtaining comparable results, both in terms of number of cluster (convergence on $k = 5$) and in terms of cluster composition; therefore, we can consider this clustering robust, general and suitable for describing the urban environment.

Analyzing the composition of the resulting clusters in terms of land uses, we find out that each cluster is characterized by a well-defined set of land uses, to which we can give a meaningful explanation: Cluster 1 identifies the dense residential areas (corresponding to CORINE Category 111), Cluster 2 the sparse residential areas (Category 112), Cluster 3 the industrial and commercial areas (Categories 12x and 13x), Cluster 4 the agricultural areas (Categories 2xx) and Cluster 5 the parks and the natural areas (Categories 14x, 3xx, 4xx and 5xx).

As a result of this step, we decide to group the CORINE categories that are characteristic for the clusters as our output classification. The grid cells of the selected cities are mapped to those five classes by taking into consideration their predominant land use (as described above). The cell distributions across the five classes for Milano, Brussels, München and Barcelona are shown in Figure 1.

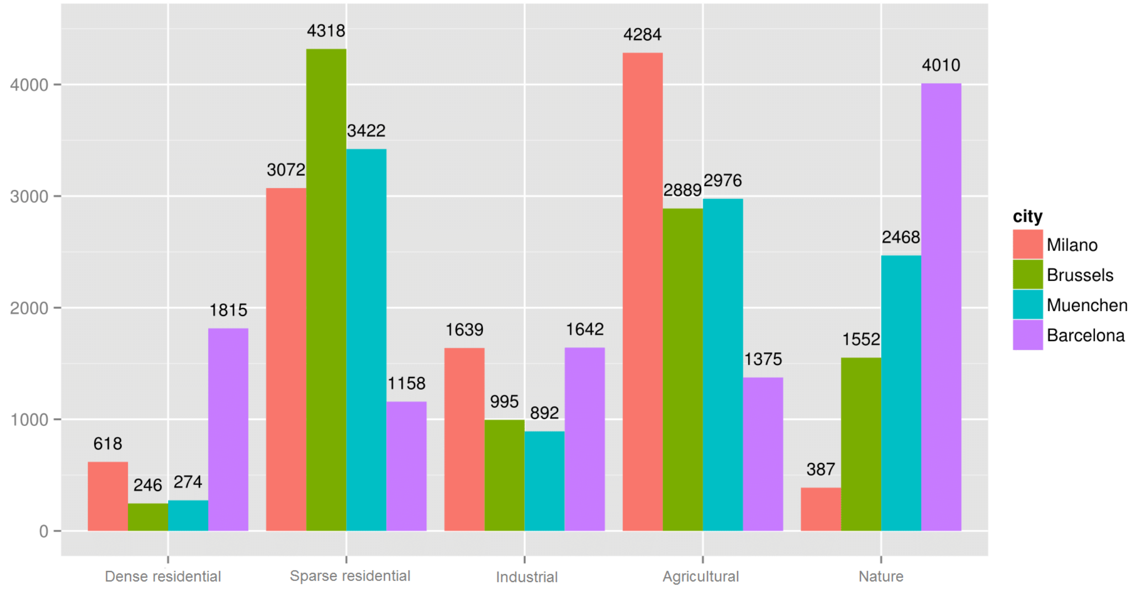

Figure 1. Cell distributions in the selected cities (2013).

While we tried to define classes as balanced as possible in terms of cardinality, by looking at Figure 1, it is evident that our cities show clear differences. This reflects the intrinsic structure and nature of a city and its surroundings: for example, the Barcelona area has a lot of natural lands ("Nature" class), whereas the Milano area lacks this class and abounds with arable lands and rice fields ("Agricultural" class).

2.3. Pre-Processing of Linked Open Geospatial Data

Because of the very wide geographic coverage, as well as the moderate spatial resolution, we select OpenStreetMap/LinkedGeoData as the source for our geo-information variables to be used as predictors in the classification experiments.

In order to characterize the spatial grid cells with data from OSM/LGD, we turn the qualitative and spatial information into a set of quantitative variables. To this end, we first select a set of 50 POI categories that can characterize the urban landscape. Listing 1 reports the full list of the selected LGD ontology concepts used in our experiments. Then, for each POI category, we create a numerical variable that describes the cell grids.

We need to describe each grid cell in terms of its surrounding environment. Initially, we thought of using POI density as a quantitative measure, but because of the very moderate spatial resolution, this was an unsuitable dimension. Therefore, we adopted a common way of thinking and analyzing the urban space, for example in the real estate business: when choosing where to buy or rent a house, we usually evaluate the distance to the closest shops, public transport, school, *etc*.

```
lgdo: Artwork         lgdo: Doctors        lgdo: Hospital      lgdo: Parking                lgdo: Restaurant
lgdo: Attraction      lgdo: DrinkingWater  lgdo: Hotel         lgdo: Pharmacy               lgdo: School
lgdo: Bank            lgdo: EmergencyThing lgdo: Kindergarten  lgdo: PlaceOfWorship         lgdo: SportThing
lgdo: Bar             lgdo: Farm           lgdo: Kiosk         lgdo: Police                 lgdo: Supermarket
lgdo: BicycleParking  lgdo: FastFood       lgdo: Library       lgdo: PostOffice             lgdo: SwimmingPool
lgdo: BookShop        lgdo: Fountain       lgdo: Marketplace   lgdo: PowerThing             lgdo: Taxi
lgdo: BusStop         lgdo: FuelStation    lgdo: Museum        lgdo: Pub                    lgdo: Theatre
lgdo: Cinema          lgdo: Furniture      lgdo: Newsagent     lgdo: PublicBuilding         lgdo: TourismInformation
lgdo: Clothes         lgdo: Gym            lgdo: Nightclub     lgdo: PublicTransportThing   lgdo: University
lgdo: DepartmentStore lgdo: HistoricThing  lgdo: Park          lgdo: RailwayStation         lgdo: WaterwayThing
```

Listing 1: Point of interest (POI) categories from LinkedGeoData used in our experiments. The *lgdo:* prefix abbreviates the LinkedGeoData ontology namespace.

```
PREFIX lgdm: <http://linkedgeodata.org/meta/>
PREFIX geom: <http://geovocab.org/geometry#>
PREFIX ogc: <http://www.opengis.net/ont/geosparql#>
SELECT ?mydistance              # select the closest distance
WHERE {
                          ?poi a @@POI_CATEGORY@@, lgdm:Node ;        # get POIs belonging to a given category
       geom:geometry [ ogc:asWKT ?wkt ] .       # get POI's spatial representation
  BIND ( bif:st_point(@@LONGITUDE@@, @@LATITUDE@@) AS ?mypoint ) # the grid cell center is a point with coordinates
  BIND ( bif:st_distance(?wkt, ?mypoint) AS ?mydistance )   # compute the distance between the POI and the grid cell
  FILTER ( ?mydistance < @@MAX_DISTANCE@@ )       # get only POIs in a given range (e.g., max 25 km)
                     }
ORDER BY ASC(?mydistance)            # order results by distance
LIMIT 1                  # take the first (closest) one
```

Listing 2: Template for GeoSPARQL queries over LinkedGeoData to get the distance from a grid cell to the closest POI of a given category.

As a consequence, each cell is described by 50 quantitative dimensions, one for each POI category, that represent the distance from the cell center to the closest POI of a given category. To compute those dimensions, instead of querying OpenStreetMap, we used the semantically-annotated version of

LinkedGeoData that automatically maps the OSM various key-value pairs into the LGD concepts of Listing 1. The query to LinkedGeoData follows the GeoSPARQL [20] query template shown in Listing 2; with 50 POI categories and four European cities, each one with 10,000 cells, in total, we execute two million queries to derive the classification predictors.

3. Land Use Classification Experiments Preliminaries

Our experiments are designed to train predictive models that, taking as input predictor variables derived from the LinkedGeoData POI dataset, are able to forecast urban land use. More specifically, since CORINE is a taxonomy, we are facing a classification problem, in that we would like to classify the urban environment according to its land use.

Several classification algorithms exist, and we tested a number of them—linear, quadratic and logistic statistical classifiers [21] and random forest, neural networks and support vector machine (SVM) supervised machine learning techniques [22]—to understand the best possible fit to address our problem, similarly to [23] for remote sensing images' feature prediction. Unlike that work, however, in our case, we observed that the performance of simple statistical classifiers was not comparable to that of more complex supervised machine learning techniques and, in particular, that SVM [24] achieved the best results.

In the following sections, we only illustrate the results obtained using SVM. The classifiers were trained with 10-fold cross-validation, and its parameters were tuned with the so-called grid search optimization [25]. The optimization attempts to maximize the sensitivity of each class and to minimize the differences between classes. The evaluation of our experiments (*cf.* Sections 4–6) is presented in terms of overall classification accuracy, sensitivity and specificity of the different output classes, confusion matrix and error maps.

Our experiments are designed as a series of consecutive steps, as follows: first, we classify the land use of each single city separately, training a model for each city and predicting unseen data of the same city (city-specific model selection); then, we generalize our methodology by creating a single model suitable for predicting multiple cities and trained using some previous knowledge about all of the cities involved (cross-city model selection with some background knowledge); finally, we make our model even more general, by predicting a city using the models trained using multiple different cities, *i.e.*, without any previous knowledge about the city to be predicted (cross-city model selection without background knowledge).

All materials related to our experimental protocol (datasets, queries, scripts, visualizations) are made available on this paper's companion website at http://swa.cefriel.it/geo/ijgi.html.

4. City-Specific Model Selection

Assuming that each urban area has its own distinctive characteristics, we decided to start our experiments by analyzing each city separately in order to build a model able to describe each specific urban pattern.

We trained a model for each of our selected cities (Milano, München, Barcelona, Brussels) using the SVM algorithm tuned with the best parameters discovered in the grid search optimization phase (as described in Section 3) with a 10-fold cross-validation.

The most common and intuitive metric used to evaluate the performance of a classifier is undoubtedly the overall accuracy, which measures the number of items correctly classified divided by the total number of items. The values of overall accuracy in Table 1 suggest that SVM is a robust classifier, as it performs very well on all four different experiments (similar values for the four cities and all accuracies greater than 0.80). Cohen's kappa coefficient is another statistic used to measure inter-rater agreement for categorical items, which is generally a quite robust measure, as it takes into account the agreement occurring by chance. Kappa coefficients in Table 1 indicate that there is a substantial agreement between the SVM prediction and the real land uses (values greater than 0.74).

A deeper investigation of how the classifiers behave on the different classes could be also performed by analyzing the sensitivity and the specificity indexes as listed in Table 1. This table reveals the presence of inter-class differences, more evident in the sensitivity index, which shows a high variability with values ranging from a minimum of 0.75 to a maximum of 0.90. On the contrary, specificity is always very high and greater than 0.90. This means that, on the one hand, our classifier is able on average to predict some classes better than others, and on the other hand, it correctly identifies cells that do not belong to a given class.

By looking at Table 1 from an inter-city perspective, we can observe, firstly, that the sensitivity values of the same class are quite different from each other (except for the "Dense residential" class) and, secondly, that the worst predicted class is not always the same ("Nature" class for Milano, "Industrial" class for München and "Sparse residential" and "Agricultural" classes for Barcelona). If we go back to Figure 1, we can notice a correspondence between low cardinality of the class and low value of sensitivity (for example the "Nature" class in Milano has only 387 samples and a sensitivity of 0.75). Although SVM turns out to be a very robust algorithm, this behavior suggests us that having an unbalanced dataset can have a notable impact on the model predictive power.

Table 1. Overall accuracy, kappa coefficient, sensitivity and specificity of the five classes (dense residential, sparse residential, industrial, agricultural, nature).

	Overall	**Kappa**	**Sensitivity**					**Specificity**				
	Acc.	**Coeff.**	**Dense**	**Sparse**	**Ind.**	**Agric.**	**Nat.**	**Dense**	**Sparse**	**Ind.**	**Agric.**	**Nat.**
Milano	0.85	0.78	0.85	0.84	0.81	0.87	0.75	0.99	0.93	0.95	0.92	0.99
München	0.87	0.82	0.88	0.87	0.77	0.90	0.86	0.99	0.93	0.98	0.95	0.96
Barcelona	0.84	0.79	0.90	0.77	0.84	0.78	0.86	0.98	0.97	0.97	0.97	0.90
Brussels	0.82	0.74	0.88	0.82	0.81	0.80	0.84	0.99	0.87	0.98	0.91	0.97

To better analyze those results, we deepen our investigation on each city; for reasons of space, hereafter, we report only our considerations about München, which is the city on which, on average, our classifier performs better (87% of overall accuracy and highest values of sensitivity), and Brussels, which is the one with the lowest overall accuracy (82%). The other two cities show very similar behavior, and the interested readers can find all of the confusion matrices and the error maps for the four cities on the companion website.

By looking at the (normalized) confusion matrices (see Table 2), we can analyze how the prediction errors spread across all of the classes. For example, if we look at the predicted "Dense residential" class of Table 2a, we can see that 88% of cells are correctly predicted, and the remaining cells are mainly incorrectly labeled as the "Sparse residential" class (7%); this is reasonable, as they both describe residential areas. The "Industrial" class turns out to be the most difficult to predict, as its correctness rate is not so high (78%), and the error equally spreads between the other three classes (10% "Sparse residential" class, 6% "Agricultural" class and 5% "Nature" class).

The worst city, Brussels (Table 2b), shows a similar behavior, except for the "Sparse residential" and "Agricultural" classes, which have the highest misclassification rate, 11% and 17%, respectively.

Besides this quantitative analysis, we performed a qualitative inspection of the classification errors, by plotting the misclassified cells of München on the map; our goal is to identify possible patterns in the spatial distribution of errors. We verified that all cities exhibit the same behavior, both in terms of spatial displacement of errors and of misclassification type.

Table 2. Confusion matrix of München and Brussels with the five classes (dense residential, sparse residential, industrial, agricultural, nature). City-specific model with 50 predictors and all of the observations. (**a**) Best city: München; (**b**) Worst city: Brussels.

(a)

		Real				
		Dense	**Sparse**	**Ind.**	**Agric.**	**Nature**
Predicted	**Dense**	**0.88**	0.07	0.02	0.00	0.03
	Sparse	0.01	**0.87**	0.03	0.04	0.05
	Ind.	0.01	0.10	**0.78**	0.06	0.05
	Agric.	0.00	0.06	0.02	**0.88**	0.04
	Nature	0.00	0.07	0.02	0.04	**0.87**

(b)

		Real				
		Dense	**Sparse**	**Ind.**	**Agric.**	**Nature**
Predicted	**Dense**	**0.89**	0.06	0.04	0.00	0.01
	Sparse	0.00	**0.83**	0.02	0.11	0.04
	Ind.	0.01	0.09	**0.82**	0.06	0.02
	Agric.	0.00	0.17	0.02	**0.78**	0.03
	Nature	0.00	0.09	0.01	0.04	**0.86**

Figure 2 shows the spatial distribution of misclassified cells in München. In Figure 2a, errors are shown as black cells on the entire München area, which is otherwise colored according to the correctly predicted land use classes. It is evident that all of the errors lie on the "boundaries" between the areas with homogeneous land use. Figure 2b zooms into a portion of the map to better visualize the error types: the color of the small square at the center of each misclassified cell visually represents the (mis) predicted class and can thus be compared to the background color representing the correct class.

We discover that the classifier always mistakes a cell's class with the class of one of its adjacent cells. This is indeed reasonable: the cells on those "boundaries" are more likely to be made up of mixed land uses, while in our analysis, we considered the predominant land use only (*cf.* Section 2). Furthermore, the smaller and the more "globular-shaped" the homogeneous areas, the more the errors.

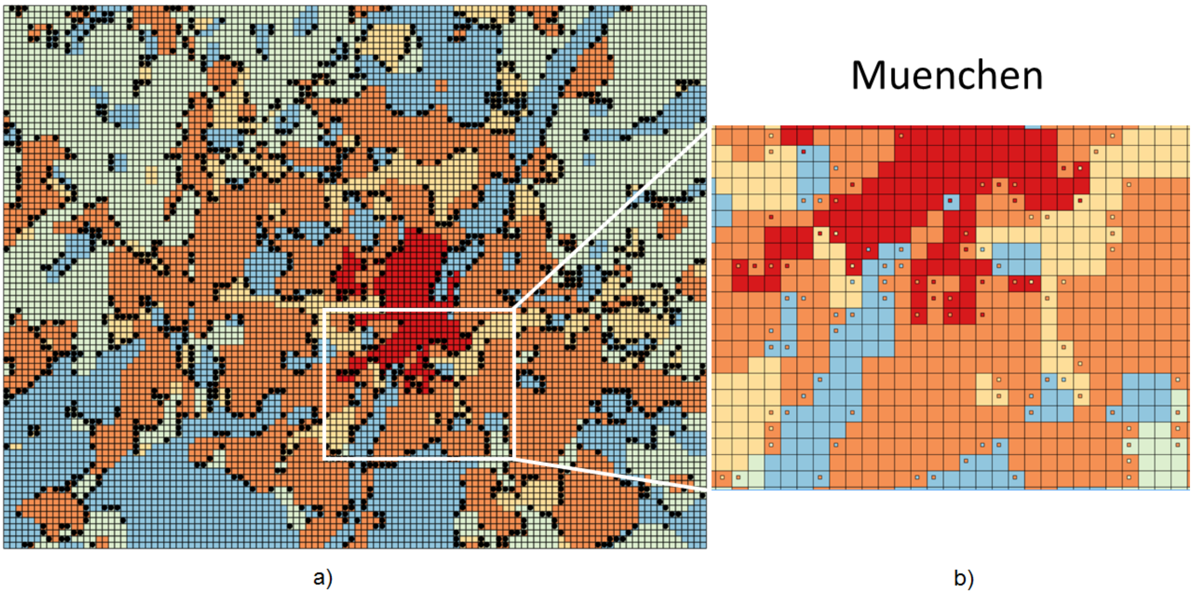

Figure 2. Spatial distribution of misclassified cells in München. "Dense residential" = red; "Sparse residential" = orange; "Industrial" = yellow; "Agricultural" = green; and "Nature" = blue. (**a**) Misclassified cells (black dots) on the entire München area; (**b**) Error types in a zoomed area.

5. Cross-City Model Selection with Some Background Knowledge

Since we obtained satisfactory results in predicting the land use of a single city in terms of overall accuracy, Cohen's kappa coefficient and sensitivity, we continue our investigation to verify whether it would be possible to build a single cross-city model, using some background knowledge of each city considered in the analysis.

The rational of using the background knowledge is that those classification models could be used to update land use maps, e.g., to identify specific areas in which the land use could have changed, to focus the expensive manual maintenance of land use maps only where it is actually needed.

To check if the cross-city model is able to predict the land use classification of unknown cells, we create different subsets of our original datasets: a training set from all four cities, on which the model is built with 10-fold cross-validation, and a test set for each city to check the algorithm precision.

We use two sampling strategies to verify how much information is needed during the training phase: on the one hand, we keep 200 cells for each class and for each city, obtaining a training set of 4000 observations (1000 for each city) and three test sets of 9000 units each, *i.e.*, the difference between the total 10,000 cells of each city and the 1000 cells used in training (balanced on cities, balanced on classes (BCi.BCl)); on the other hand, we use a third of the original 40,000 cells as the training set, respecting the original proportion of five classes across cities, and the remaining cells as the test set (stratified on cities, stratified on classes (SCi.SCl)).

Naturally, it is reasonable to expect that the higher the number of cells used in training, the more reliable the prediction model on unknown cells. Our expectations are confirmed by looking at Table 3, which shows the prediction overall accuracy and Cohen's kappa coefficient for each city and for each sampling strategy: they are always higher in the SCi.SCl experiment.

Table 3. Overall accuracies and Kappa coefficients obtained from a cross-city model with background knowledge training (using all 50 predictors). BCi.BCl, balanced on cities, balanced on classes; SCi.SCl, stratified on cities, stratified on classes.

	BCi.BCl		**SCi.SCl**	
	Overall Acc.	**Kappa Coeff.**	**Overall Acc.**	**Kappa Coeff.**
Milano	0.59	0.45	0.75	0.63
München	0.70	0.60	**0.80**	**0.72**
Barcelona	0.67	0.57	0.78	0.71
Brussels	0.65	0.51	0.77	0.67

Furthermore, in this second set of experiments, the classification achieves the best performance on München, so we continue to offer in-depth considerations on this city, by analyzing the confusion matrix in Table 4a, which shows the model prediction accuracy across all of the classes.

Table 4. Confusion matrix of München and Milano with the five classes (dense residential, sparse residential, industrial, agricultural, nature); result obtained with a cross-city model with background knowledge sampled with SCi.SCl and 50 predictors. (**a**) Best city: München; (**b**) Worst city: Milano.

(a)

		Real				
		Dense	**Sparse**	**Ind.**	**Agric.**	**Nature**
Predicted	**Dense**	**0.71**	0.21	0.04	0.00	0.04
	Sparse	0.01	**0.79**	0.06	0.06	0.08
	Ind.	0.00	0.14	**0.72**	0.07	0.07
	Agric.	0.00	0.05	0.03	**0.85**	0.07
	Nature	0.00	0.09	0.02	0.07	**0.81**

(b)

		Real				
		Dense	**Sparse**	**Ind.**	**Agric.**	**Nature**
Predicted	**Dense**	**0.75**	0.21	0.02	0.00	0.02
	Sparse	0.06	**0.71**	0.11	0.10	0.02
	Ind.	0.01	0.12	**0.70**	0.15	0.02
	Agric.	0.00	0.09	0.09	**0.79**	0.03
	Nature	0.00	0.11	0.07	0.13	**0.69**

By looking at the diagonal elements, which represent the percentage of cells correctly classified, we can see that the model predicts very well the agricultural areas (85%), while it is less precise on classifying cells that belong to the "Dense residential" class (71%). We can justify this worse result by analyzing how the errors on this class spread on the other classes: 21% of the cells incorrectly predicted as the "Dense residential" class belong to "Sparse residential" class, and again, this is reasonable.

In addition, Table 4b shows the detailed results of Milano, the city with the lowest global accuracy (75%). The results are similar, with values on the diagonal slightly lower.

6. Cross-City Model Selection without Any Background Knowledge

Given the satisfactory results obtained in predicting the land use of a city using a model that includes some background information about the city itself, we decide to further test the repeatability and generality of our approach by predicting a city without any previous knowledge, *i.e.*, without data from that city used in the training phase.

Therefore, we train a classifier with three of the four cities available, and we predict the fourth one, which is completely discarded in the modeling phase (*i.e.*, we predict München using a model trained on Milano, Barcelona and Brussels).

6.1. Classification with Five Levels

First, we train an SVM model using all of the predictors (50 POI distances) and all of the observations (30,000 cells, 10,000 for each city used in the training). As we will see in details in the following, the overall accuracy values, obtained predicting all 10,000 cells of the fourth city, are quite low, ranging from 14% to 40%. This means that the classifier has poor predictive power and overfitting could have occurred (the model does not generalize well on the test set).

To address this issue, we design further experiments reducing the number of both predictor variables and training samples.

As regards the number of observations, we adopt three under-sampling strategies: (1) we sample our data in a balanced way, randomly selecting 200 observations for each land use class and for each city (balanced on cities and balanced on classes (BCi.BCl)), therefore using 3000 observations out of the original set; (2) we select a third of the original 30,000 observations in a stratified way according to both land use classes and cities (stratified on cities and stratified on classes (SCi.SCl)); (3) we adopt a hybrid solution, sampling one third of the original 30,000 observations in a stratified way according to classes and in balanced way with respect to cities (balanced on cities and stratified on classes (BCi.SCl)). We test those different solutions in order to face the issues identified in the single city experiment (*cf.* Section 4) about the influence of the cardinality of the classes used in the training set on the evaluation results (the lower the cardinality, the lower the performance indexes).

As regards variable selection, we rank all of the predictors in terms of their information gain, calculated according to the Shannon entropy [26], which measures the data heterogeneity with respect to the land use classes. Then, we select the top five and the top 11 variables, based on apparent discontinuities on information gain values. This procedure aims at selecting only the most informative predictors, avoiding model overfitting.

The selected variables, listed in Listing 3, represent the distances to the closest POI of the following categories: public facilities (schools, pharmacies and banks), transportation (bus stops, railway and fuel stations) and amenities (restaurants, cinemas and shops).

Figure 3 illustrates the distribution of the top five variables for each city: indeed, it seems that those POI-specific distances do not show a relevant difference between the four cities. This similarity could suggest that a cross-city model built on these five predictors could more suitably describe the patterns of different cities.

```
1. lgdo: Pharmacy                  7. lgdo: Restaurant
2. lgdo: Bank                      8. lgdo: School
3. lgdo: BusStop                   9. lgdo: FuelStation
4. lgdo: RailwayStation}           10. lgdo: BookShop}
5. lgdo: Supermarket}              11. lgdo: Cinema}
6. lgdo: PublicTransportThing}
```

Listing 3: Ranking of the predictors according to the information gain. The lgdo: prefix abbreviates the LinkedGeoData ontology namespace.

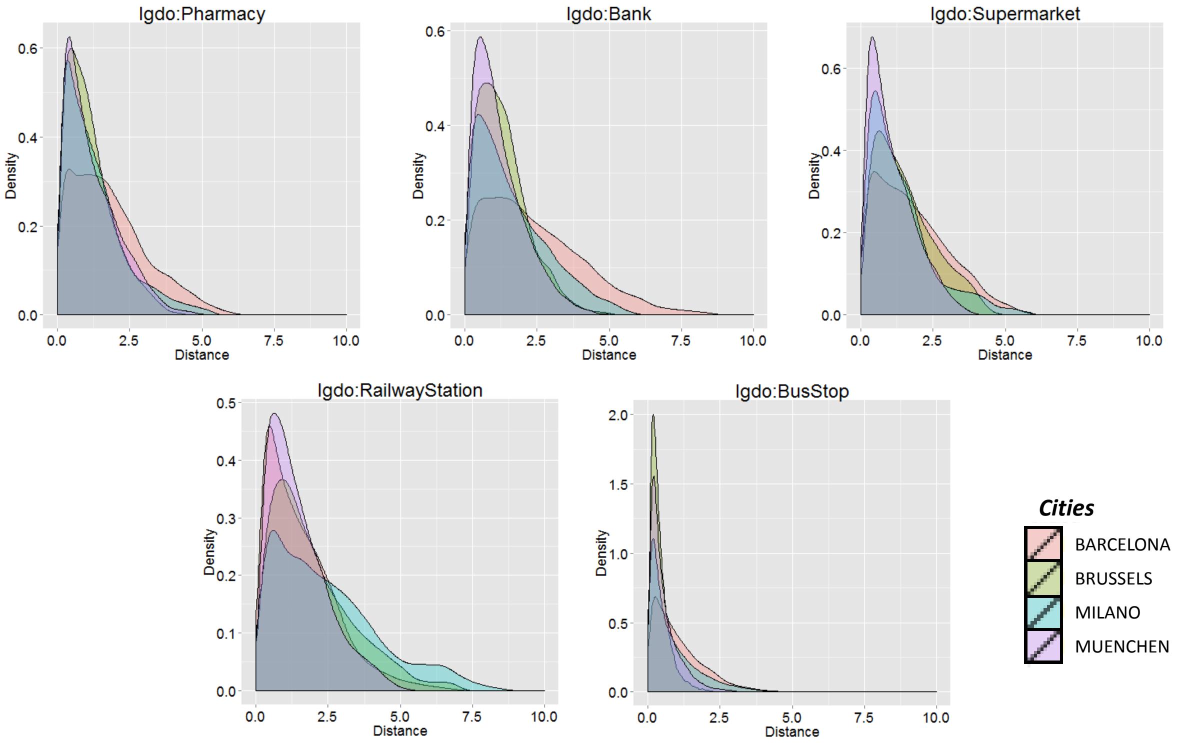

Figure 3. Distribution in the four cities of the five top-ranked predictors, according to their information gain.

Table 5. Overall accuracy obtained by multiple models trained with different numbers of predictors and observations.

Overall Accuracy	***All***	***BCi.BCl***			***BCi.SCl***			***SCi.SCl***	
	50	***50***	***11***	***5***	***50***	***11***	***5***	***11***	***5***
Milano-Barcelona-Brussels → München	0.37	0.36	0.36	0.29	0.26	0.39	0.48	0.41	**0.50**
München-Barcelona-Brussels → Milano	0.25	0.24	0.46	0.34	0.29	**0.49**	0.44	0.25	0.41
Milano-München-Brussels → Barcelona	0.14	0.21	0.24	**0.39**	0.19	0.21	0.23	0.22	0.22
Milano-München-Barcelona → Brussels	0.40	0.38	0.38	0.38	0.42	0.45	0.48	0.42	**0.49**

All of the performed tests, varying both the number of predictors and the sampling methods, are shown in Table 5. Generally, the overall accuracy values are not so exciting (no values higher than 50%), and the general tendency is that the lower the number of predictors, the higher the overall accuracy. Cohen's kappa coefficients follow the same tendency.

The best results are obtained, on average, using the stratified sampling methods (BCi.SCl and SCi.SCl) with five predictors. By looking at the table from an inter-city perspective, we can observe that, once again, the city with the best results is München (50% accuracy with five predictors and SCi.SCl sampling), and the city with the worst results is Barcelona, which reaches a maximum accuracy of 39%.

The confusion matrix for München (see Table 6a) resulting from the classifier trained with five predictors and with the stratified sampling on both classes and cities (SCi.SCl) strengthens the considerations already discussed in the previous experiments. The misclassification error between the "Dense residential" class and the "Sparse residential" class is even clearer (20% cells correctly classified as the "Dense residential" class and 64% incorrectly labeled as the "Sparse residential" class), as well as the difficulties in predicting the "Industrial residential" class (higher errors spread between the "Sparse residential", "Agricultural" and "Nature" classes). The confusion matrix also highlights relevant misclassification errors between the "Agricultural" and "Nature" classes; however, this is reasonable.

As regards Barcelona (see Table 6b), "Dense residential" and "Nature" are the best predicted classes, whereas the distinction between the other three classes is not correctly modeled, showing the difficulties in detecting these kinds of land use.

From these results, it is evident that it is hard to train a classifier to predict a city land use, without any previous knowledge about the city itself. Indeed, the results shown in Section 5 (model trained with data coming from all of the cities) are much better, even if only little background information was given to the training phase. This could strengthen the hypothesis that each city has its own peculiar characteristics and intrinsic patterns. Therefore, a single model could fail to be general enough to predict other unknown urban environments. We further investigate this idea by conducting a hypothesis test ("two-proportion Z-test") to evaluate if the four selected cities are likely to get similar metrics on the five classes; our aim is to verify if the difference in sensitivity and specificity values between the cities is statistically significant.

Table 6. Confusion matrix of München and Barcelona with the five classes (dense residential, sparse residential, industrial, agricultural, nature); result obtained with a cross-city model with background knowledge sampled with SCi.SCl and 50 predictors. (**a**) Best city: München; (**b**) Worst city: Barcelona.

(a)

		Real				
		Dense	**Sparse**	**Ind.**	**Agric.**	**Nature**
Predicted	**Dense**	**0.20**	0.64	0.10	0.03	0.03
	Sparse	0.01	**0.59**	0.12	0.14	0.14
	Ind.	0.00	0.18	**0.23**	0.41	0.18
	Agric.	0.00	0.05	0.05	**0.54**	0.36
	Nature	0.00	0.02	0.04	0.45	**0.48**

(b)

		Real				
		Dense	**Sparse**	**Ind.**	**Agric.**	**Nature**
Predicted	**Dense**	**0.93**	0.01	0.03	0.01	0.02
	Sparse	0.37	**0.12**	0.20	0.08	0.23
	Ind.	0.20	0.18	**0.20**	0.15	0.27
	Agric.	0.04	0.11	0.15	**0.19**	0.51
	Nature	0.03	0.09	0.05	0.07	**0.76**

Table 7 presents the hypothesis tests' *p*-values, computed for both the sensitivity and specificity metrics (*cf.* Table 1), for each couple of cities and for each class; colored cells highlight the results that are not statistically significant at the 5% significance level, therefore cases in which the null hypothesis cannot be rejected.

The results in Table 7 prove that in most cases, the difference between cities is statistically significant (the majority of white cells in the table), except for the "Dense residential" class in which the difference is almost always not statistically significant. Since the "Dense residential" class corresponds to CORINE Category 111 "Continuous urban fabric", also, this result seems reasonable: our experiments are focused on urban areas, and indeed, that land use type is more typical for cities.

To sum up, the reason for the limitation of the model trained without any previous knowledge probably lies in the intrinsic peculiarities of each city, which get so strongly reflected in the classification model that they cannot be applied to another place, and in the available predictors, which have different reliability from place to place.

Table 7. *p*-values of the hypothesis tests on the difference in sensitivity and specificity between cities ("two-proportion Z-test"); five-classes classification (dense residential, sparse residential, industrial, agricultural, nature); dark gray: not significant at the 5% significance level; light gray: significant at the 10%, but not at the 5% significance level.

Cities Comparison	**Sensitivity**					**Specificity**				
	Dense	**Sparse**	**Indust.**	**Agric.**	**Nature**	**Dense**	**Sparse**	**Indust.**	**Agric.**	**Nature**
Barcelona-Brussels	0.30	0.00	0.07	0.17	0.01	0.06	0.00	0.05	0.00	0.00
Barcelona-Milano	0.00	0.00	0.02	0.00	0.00	0.07	0.00	0.00	0.00	0.00
Barcelona-München	0.51	0.00	0.00	0.00	0.67	0.06	0.00	0.13	0.00	0.00
Brussels-Milano	0.37	0.10	0.88	0.00	0.00	0.36	0.00	0.00	0.01	0.07
Brussels-München	0.77	0.00	0.06	0.00	0.05	0.92	0.00	0.72	0.00	0.01
Milano-München	0.20	0.00	0.06	0.00	0.00	0.38	0.45	0.00	0.00	0.00

6.2. Classification with Two Levels

Bearing in mind the results obtained in the classification with five classes and in the hypothesis test of Table 7, we finally investigate whether we can obtain better results by building a single cross-city classification model only for residential land use typology. In other words, we test whether a stronger similarity between cities can be found by changing the granularity level of the land use classification. That is why we create a binomial classification, by merging, on the one hand, the "Dense residential" and "Sparse residential" classes and, on the other hand, the remaining classes (industrial/commercial, agricultural and natural areas).

We train a classifier using the SCi.SCl sampling strategy shown in the previous section and the top five predictors. We obtain high values of the overall accuracy for all of the cities (ranging from 71% of Brussels to 83% of München). Table 8 illustrates the confusion matrix obtained for München. The percentage of cells correctly classified is very high on both classes (76% for residential and 88% for non-residential).

Table 8. Confusion matrix of München obtained with a binomial model sampled with SCi.SCl and 5 predictors.

		Real	
		Residential	**Other**
Predicted	**Residential**	**0.76**	0.24
	Other	0.12	**0.88**

If we also look at Figure 4, which shows the spatial distribution of the misclassified cells, we can notice that most of the errors lie again on the "boundaries" between residential (colored in dark grey) and natural areas (colored as white). The blue cells represents the cells incorrectly classified as residential areas (24% in Table 8), whereas the green cells are the mispredicted non-residential ones.

To sum up this third and last set of experiments, with two classes, we obtain a great improvement with respect to the five-class classification. These results therefore confirm the hints given by the hypothesis test and our previous considerations on those results.

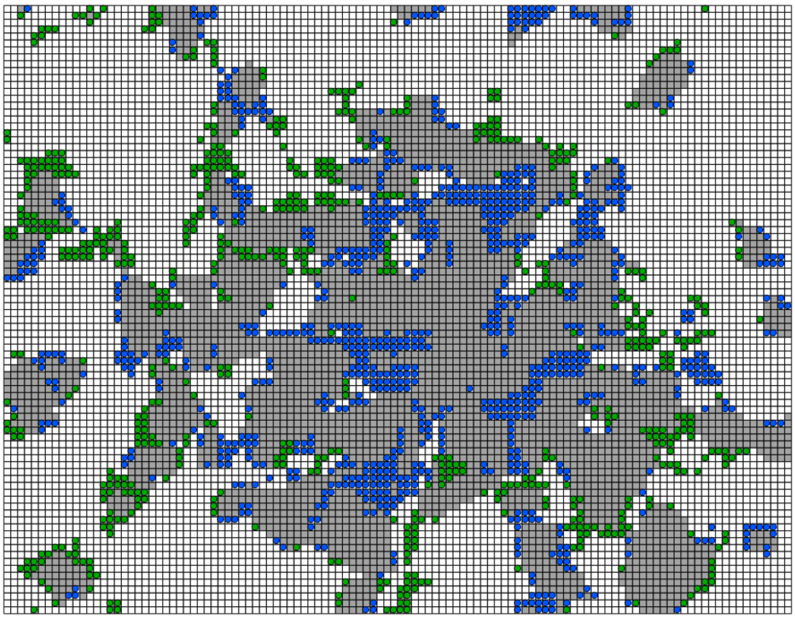

Figure 4. Errors of binomial classification on München (sampling with SCi.SCl and five predictors). Dark gray represents residential areas. Blue dots are the cells erroneously classified as residential. Green dots are cells erroneously classified as not-residential.

7. Discussion

Our experiments showed very good results (overall accuracy always higher than 75%) in predicting the land use of a city when using a classifier that takes into account some background knowledge of the city itself during its training phase. The quantitative and qualitative evaluations presented in the paper therefore support our hypothesis that linked open geospatial data can be successfully used in support of producing or updating other expensive spatial data sources, like the CORINE land use dataset [13]. The application to four different European cities demonstrated the repeatability and general validity of our methodology, provided that some background information of the urban environment is given. Our solution can be proven to be useful for monitoring and detecting the land use changes: for example, if our predictive models, applied to the same urban area after a period of time, show different land use predictions in limited zones, the CORINE maps' revision can focus on those regions only. In this way, the update process of land use information could be done in a less expensive and more tailored way.

The inferior results obtained in predicting the urban land use without any background knowledge of the city itself show the limits of our current methodology, at least when using the five-level classification with the illustrated experimental settings. Indeed, we obtained better results when using a simpler two-level land use classification, focusing the prediction on residential *vs.* non-residential areas. Furthermore, moving from a classification with a handful of levels to the full CORINE taxonomy (which includes a hierarchy of more that 40 classes) will require further investigation. Our first experiment in training a model of all CORINE classes in the Milano area (not reported in this paper) resulted in sensitivities higher than 70% in only two classes corresponding to the categories with the highest cardinality and in sensitivities close to 0% in the minority classes, which were predominant in a very limited number of grid cells.

This means that, to draw some conclusions on the feasibility to predict the full CORINE classification, future studies will need to include larger and more comprehensive samples.

In our opinion, the reason for the outlined limitations of models without background knowledge lies, on the one hand, in the intrinsic peculiarities of each city: the specific characteristics of the cities used in the training phase get so strongly reflected in the resulting classifier, that the predictive model can become unsuitable to be applied to another place. On the other hand, the reason for the models' difference can be traced back to the geo-information we used as predictors, *i.e.*, the distances to the closest POI of specific categories. Those distances were computed on the linked open geospatial data that ultimately come from OpenStreetMap; since the latter is a VGI [14] initiative, the diversity in our results could also be caused by the inherent inhomogeneity of the mapping efforts of OpenStreetMap volunteers in the selected cities. In other words, the varying levels of data completeness and reliability from place to place in OpenStreetMap can indeed be the root cause of the strong differences between the cities as reflected in the trained classifiers. Several metrics are being proposed in literature to analyze the "quality" of OpenStreetMap [27,28], and we intend to take those measures into account in the future extensions of the presented approach.

On the positive side, regarding the evolution of the land use information over time, OpenStreetMap has proven effective at reflecting the actual dynamics of a place, both in generic urban places [29] and especially in response to emergency situations [30]. Indeed, the motivations of mappers for a long-term contribution show that OpenStreetMap is destined to stay as a reference geographic information source [31]. Moreover, in this study, we manually selected only a subset of the available POI categories to build our classifier predictor variables; this choice can be questionable, and we will work on improving the POI selection to get the best possible coverage of all of the land use types. For example, we will probably need to select additional POIs that better characterize industrial/commercial, agricultural and natural areas or other more specific land uses.

Finally, our future work will focus on extending our experiments to make the resulting solution more effective and robust; one option on our research agenda is to combine and complement the geo-information from OpenStreetMap with other heterogeneous sources, like phone activity data, which have often been used as land use predictors.

8. Comparison with State-of-the-Art Approaches

Land use classification is an extensively explored topic in urban planning, environmental engineering and geo-science in general, because of its implications on our everyday life, especially in the context of the design of urban environments. It is well known, for example, that several different aspects influence land use and its change over time, like socio-economic factors [32], travel behaviors [33] or the interactions with urban stakeholders [34].

When dealing with the challenges of land use change analysis [35], it is also worth distinguishing between land cover, the physical cover of the Earth's surface, and land use, what the land is used for by humans. As a matter of fact, most automatic and semi-automatic approaches operating on remote sensing information, like satellite imagery, are aimed at determining landscape metrics [36] and land cover classification [37], because those data sources can help with detecting only visually-observable changes. This is also the case for the CORINE classification of Europe [13], which undergoes a long, complex and only partially automated processing of remote sensing data.

On the other hand, people play an increasingly relevant role with respect to geo-information in general [14,15], because of the emergence of the so-called Geospatial Web in the last decade: a large user base is involved in the collection and curation of geospatial metadata, especially road networks and POIs, including information that is relevant for urban planning. For example, also in the context of land cover, initiatives like Geo-Wiki [38] leverage crowdsourcing and/or citizen science efforts to improve the environment knowledge.

This is also the reason why a number of methods, approaches and experiments have been focusing on exploiting heterogeneous information sources to classify land use; in most cases, those sources are either produced by humans or derived from their everyday activities. Notable examples are land use information derived from social media (e.g., geo-tagged tweets as in [7]), mobile phone activity (like clustered calling patterns in [39], location clusters derived from activity patterns in [8] and aggregated time series considering both temporal patterns and call volumes in [9]), transport and mobility data (e.g., GPS trajectory datasets, as in [10], or bus smart card data, as in [40]), as well as a mix of different open and enterprise data sources (as in our previous work [41]).

Last, but not least, the approach presented in this paper is not the first one to exploit POI information from OpenStreetMap, which is also used by [11] to derive land use patterns and by [12] to identify urban parcels. With respect to those works, our original contribution lies in the characterization of the space with the distance to the closest POI of a given type, the repeatability of the experiments and the quite moderate spatial resolution of the urban environments (in the order of 250 meters), including the areas surrounding the cities, which exhibit mixed land uses.

9. Conclusions

For the last few years, the increasing availability of linked open geospatial data has paved the way for innovative solutions in the smart city domain. In the context of urban planning and monitoring, the wealth of data available today represents an opportunity to introduce incremental or disruptive innovations in the urban data management processes.

The study presented in this paper is framed in this context. Our research focuses on the exploitation of diverse and heterogeneous data sources for land use monitoring. More specifically, we demonstrated that open geospatial data indeed reflects the territory usage and, thus, can prove to be additional and relevant input information to be leveraged in urban planning. In particular, in this paper, we proposed a knowledge discovery approach to extract urban land use from linked open data of a geospatial nature: predictive models trained with geo-information related to urban POIs were used to classify the city land use according to a five-level classification derived from the European CORINE taxonomy.

The uniqueness of our proposed method lies in replicating our experiments on four different European cities, thus ensuring the repeatability and generality of the proposed solution. Furthermore, our study also ensures reproducibility, since we described in detail the experiments performed, as well as the employed datasets, which are also made available on-line.

An original contribution is the investigation of the level of background knowledge required as input. We demonstrated that, to get better classification results, it is essential to include in the model some information about the city to be predicted. Actually we obtained the best results with a city-specific model that reached 87% of overall accuracy (training the algorithm on the city to predict) and with a cross-city model with background knowledge resulting in overall accuracy up to 80%. In general, we obtained balanced sensitivity values across classes that are comparable and, in most cases, better than in pre-existing literature. Still, further investigation is required to make our methodology applicable to any urban environment, and additional experiments are needed to discover more specific land uses.

While we do not claim that the proposed predictive methods can be used to generate detailed land use maps, we believe that the introduction of openly available information can provide valid support to the monitoring and updating of urban planning information, like the European CORINE maps. Indeed, land monitoring in Europe is experiencing a radical revision of processes in terms of methods and classifications with the EAGLE initiative [6]; in this context, innovative solutions like ours, based on the employment of diverse geo-information sources, can provide a valuable contribution.

Acknowledgments

This work was supported by the PROACTIVE project (ID 40723101) co-funded by Regione Lombardia (POR-FESR 2007-2013).

Author Contributions

All authors contributed equally to this work.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Celino, I.; Kotoulas, S. Smart cities. *IEEE Internet Compu.* **2013**, *17*, 8–11.
2. Janowicz, K.; Scheider, S.; Pehle, T.; Hart, G. Geospatial semantics and linked spatiotemporal data—Past, present, and future. *Semant. Web* **2012**, *3*, 321–332.
3. Hemerly, J. Public policy considerations for data-driven innovation. *Computer* **2013**, *46*, 25–31.
4. Talbot, D. Big data from cheap phones. *MIT Technol. Rev.* **2013**, *116*, 50–54.
5. Mandel, M. Beyond goods and services: The (unmeasured) rise of the data-driven economy. *Progress. Policy Inst.* **2012**, *10*, 1–14.
6. Arnold, S.; Kosztra, B.; Banko, G.; Smith, G.; Hazeu, G.; Bock, M.; Valcarcel Sanz, N. The EAGLE concept—A vision of a future European Land Monitoring Framework. In Proceedings of the 33rd EARSeL Symposium "Towards Horizon 2020", Matera, Italy, 3–6 June 2013.
7. Frias-Martinez, V.; Soto, V.; Hohwald, H.; Frias-Martinez, E. Characterizing urban landscapes using geolocated tweets. In Proceedings of 2012 International Conference on Social Computing (SocialCom) (Privacy, Security, Risk and Trust (PASSAT)), Amsterdam, Netherland, 3–5 September 2012.
8. Toole, J.L.; Ulm, M.; González, M.C.; Bauer, D. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing (UrbComp 2012), Beijing, China, 12 August 2012.
9. 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. *Int. J. Geogr. Inform. Sci.* **2014**, *28*, 1988–2007.
10. Yuan, J.; Zheng, Y.; Xie, X. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012.
11. Jokar Arsanjani, J.; Helbich, M.; Bakillah, M.; Hagenauer, J.; Zipf, A. Toward mapping land-use patterns from volunteered geographic information. *Int. J. Geogr. Inform. Sci.* **2013**, *27*, 2264–2278.
12. Long, Y.; Liu, X. Automated Identification and Characterization of Parcels (AICP) with OpenStreetMap and Points of Interest. 2013. Available online: http://arxiv.org/abs/1311.6165 (accessed on 25 June 2015).
13. Büttner, G.; Kosztra, B.; Maucha, G.; Pataki, R. *Implementation and Achievements of CLC2006*; Technical report for European Environment Agency (EEA): Copenhagen, Denmark, 2012.
14. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221.
15. Goodchild, M.F. Spatial accuracy 2.0. In Proceedings of the Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, China, 25–27 June 2008.
16. Stadler, C.; Lehmann, J.; Höffner, K.; Auer, S. LinkedGeoData: A core for a web of spatial open data. *Semant. Web J.* **2012**, *3*, 333–354.

17. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data—The story so far. *Int. J. Semant. Web Inf. Syst.* **2009**, *5*, 1–22.
18. Duan, K.B.; Keerthi, S. Which is the best multiclass SVM method? An empirical study. In *Multiple Classifier Systems*; Oza, N., Polikar, R., Kittler, J., Roli, F., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2005; Volume 3541, pp. 278–285.
19. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA, USA, 21 June–18 July 1965.
20. Perry, M.; Herring, J. *OGC GeoSPARQL—A Geographic Query Language for RDF Data*; Technical report for Open Geospatial Consortium: Wayland, MA, USA, 2011.
21. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2009.
22. Kotsiantis, S.B.; Zaharakis, I.; Pintelas, P. Supervised machine learning: A review of classification techniques. In *Emerging Artificial Intelligence Applications in Computer Engineering*; Maglogiannis, I., Karpouzis, K., Wallace, M., Soldatos, J., Eds; IOS Press: Amsterdam, Netherlands, 2007; pp. 3–24.
23. Vatsavai, R.R.; Bright, E.; Varun, C.; Budhendra, B.; Cheriyadat, A.; Grasser, J. Machine learning approaches for high-resolution urban land cover classification: A comparative study. In Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications, Washington, DC, USA, 23–25 May 2011.
24. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, USA, 27–29 July 1992.
25. Hsu, C.W.; Chang, C.C.; Lin, C.J. *A Practical Guide to Support Vector Classification*; National Taiwan University: Taipei, Taiwan, 2010.
26. Shannon, C.E. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.* **2001**, *5*, 3–55.
27. Mooney, P.; Corcoran, P.; Winstanley, A.C. Towards quality metrics for OpenStreetMap. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010.
28. Keßler, C.; de Groot, R.T.A. Trust as a proxy measure for the quality of Volunteered Geographic Information in the case of OpenStreetMap. In *Geographic Information Science at the Heart of Europe*; Springer International Publishing: Cham, Switzerland, 2013; pp. 21–37.
29. Hristova, D.; Quattrone, G.; Mashhadi, A.J.; Capra, L. The life of the party: Impact of social mapping in OpenStreetMap. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Cambridge, MA, USA, 8–11 July 2013.
30. Palen, L.; Soden, R.; Anderson, T.J.; Barrenechea, M. Success & scale in a data-producing organization: The socio-technical evolution of OpenStreetMap in response to humanitarian events. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea, 18–23 April 2015.

31. Budhathoki, N.R.; Haythornthwaite, C. Motivation for open collaboration crowd and community models and the case of OpenStreetMap. *Am. Behav. Sci.* **2013**, *57*, 548–575.
32. Stead, D. Relationships between land use, socioeconomic factors, and travel patterns in Britain. *Environ. plan. B* **2001**, *28*, 499–528.
33. Maat, K.; Van Wee, B.; Stead, D. Land use and travel behavior: Expected effects from the perspective of utility theory and activity-based theories. *Environ. Plan. B: Plan. Des.* **2005**, *32*, 33–46.
34. Jjumba, A.; Dragićević, S. High resolution urban land-use change modeling: Agent iCity approach. *Appl. Spatial Anal. Policy* **2012**, *5*, 291–315.
35. Erickson, A.; Rogers, L.; Hurvitz, P.; Harris, J. Challenges and solutions for a regional land use change analysis. In Proceedings of ESRI International User Conference, San Diego, CA, USA, 8–12 July 2013.
36. Herold, M.; Scepan, J.; Clarke, K.C. The use of remote sensing and landscape metrics to describe structures and changes in urban land uses. *Environ. Plan. A* **2002**, *34*, 1443–1458.
37. Jiang, D.; Huang, Y.; Zhuang, D.; Zhu, Y.; Xu, X.; Ren, H. A simple semi-automatic approach for land cover classification from multispectral remote sensing imagery. *PloS one* **2012**, *7*, e45889.
38. Fritz, S.; McCallum, I.; Schill, C.; Perger, C.; See, L.; Schepaschenko, D.; van der Velde, M.; Kraxner, F.; Obersteiner, M. Geo-Wiki: An online platform for improving global land cover. *Environ. Modell. Software* **2012**, *31*, 110–123.
39. Soto, V.; Frías-Martínez, E. Automated land use identification using cell-phone records. In Proceedings of the 3rd ACM International Workshop on MobiArch, Bethesda, MD, USA, 28 June–1 July 2011.
40. Han, H.; Yu, X.; Long, Y. Discovering Functional Zones Using Bus Smart Card Data and Points of Interest in Beijing. 2015. Available online: http://arxiv.org/abs/1503.03131 (accessed on 25 June 2015).
41. Re Calegari, G.; Celino, I. Smart urban planning support through web data science on open and enterprise data. In Proceedings of the 24th International Conference on World Wide Web Companion, Florence, Italy, 18–22 May 2015.