Semi-Automatic Interpretation of Buildings and Settlement Areas in User-Generated Spatial Data

Stefan Werder, Birgit Kieler, Monika Sester
Institute of Cartography and Geoinformatics
Leibniz Universität Hannover
Appelstr. 9a, 30167 Hannover

{Stefan.Werder, Birgit.Kieler, Monika.Sester}@ikg.uni-hannover.de

ABSTRACT

In recent times the amount of spatial data being collected by voluntary users, e.g. as part of the OpenStreetMap project, is rapidly increasing. Due to the fact, that everyone can participate in this social collaboration, the completeness and accuracy of the data is very heterogeneous. Although a object catalogue exists as part of the OSM project, users are not restricted which attributes they set and to which detail. Therefore the geometry of a feature is more reliable than its attributes. However, in order to use the data for analysis purposes, knowledge about the semantic contents is of importance.

In our work, we propose an approach to classify spatial data solely based on geometric and topologic characteristics. We use both building outlines and road network information. In the first step, topology errors are fixed in order to create a consistent dataset. In the second step, we use unsupervised classification to separate buildings into clusters sharing the same characteristics. Including expert knowledge by visual inspection and interaction, some of these clusters are grouped together and semantically enriched. In the third step, we transfer the derived information from individual buildings to city blocks that are enclosed by edges of the road network. We evaluate our approach with test datasets from OSM and available authoritative datasets. Our results show, that enrichment of user-generated data is possible based on geometric and topologic feature characteristics.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data Mining*; I.5.1 [**Pattern Recognition**]: Models – *Geometric*; I.5.3 [**Pattern Recognition**]: Clustering – *Algorithms*.

General Terms

Algorithms, Measurement, Verification.

Keywords

Spatial Data Mining, Settlement Types, Generalization.

GIS '10, 03-NOV-2010, San Jose CA, USA

1. INTRODUCTION

In the past, the production of topographic maps was solely the task of national mapping agencies. Today, however, maps are excessively used and produced not merely by experts, like cartographers, but rather by non-experts, reinforced by numerous web applications and mobile devices (e.g. Google Maps and various location based services). For this reason, the OpenStreetMap (OSM) project was established in 2004 with the aim to create a free digital map of the world using data collected by volunteers. Due to the fact that everyone can participate in this social collaboration, the amount of spatial data is rapidly increasing. The collected data is also very heterogeneous compared to proprietary data.

Several studies investigated OSM data regarding its completeness and accuracy. They conclude that data density and level of detail are much higher in larger cities than in rural areas, explained by the presence of more active project members [1, 2]. Though geometric uncertainties of the data can be explained by the use of various acquisition devices with different accuracies, the reason for the heterogeneity concerning semantic attributes are manifold. Although a map object catalogue exists, the users are not obliged to use the recommended tags for a precise description of the captured features. Often, attributes are assigned rather arbitrarily or are completely missing. Still, however, reliable semantics are very important for a correct interpretation of spatial data, and also a prerequisite for successful data integration, data analysis and map generalization, just to name a few. The additional information that a building represents, e.g. a church, can prevent its elimination during generalization. Equally important are semantic information at a higher level, such as the membership to a building alignment in order to preserve these implicit given patterns during a generalization process.

This leads to two primary questions addressed by current cartographic research. First, how can we exploit the potentially rich content of freely available and frequently updated user-generated data for spatial applications? Secondly, how can we automatically extract semantics of different detail by analyzing the geometry and topology of individual spatial features or by recognizing spatial patterns?

In this paper, we present an approach which combines both objectives, that is, we use OSM data for the detection of different types of higher level settlement areas based on previous building classification, in order to distinguish e.g. residential areas from industrial areas. We analyze individual feature characteristics

based on several measures, which represent their geometrical properties and topological relations to neighboring features.

Several works in literature have proposed solutions in recognizing patterns in spatial data. Heinzle [3] works on the detection of typical patterns in road networks. Steiniger et al. [4] propose an approach for recognizing island structures. However, in particular, the recognition of building patterns for generalization purposes are comprehensively investigated. While Christophe & Ruas [5] present a method for the detection of linear building alignments, Zhang et al. [6] extend the focus also on nonlinear structures, including grid-like and unstructured building patterns. All mentioned approaches use either geometric algorithms or statistical techniques for the detection of patterns.

On the contrary Lüscher et al. [7, 8] use an ontology-driven approach for pattern recognition, by formalizing semantics of e.g. terraced houses in terms of geometrical and topological conditions in an ontology. The success in detecting features using this manually built model is highly dependent on the quality and completeness of the ontology. Furthermore, for each concept a complete ontology is necessary.

To overcome this disadvantage we propose a data-driven approach. Also Burghardt & Steiniger [9] describe types of settlements with respect to buildings. Based on the homogeneity evaluation of several geometrical and topological measures in building alignments, the settlement type regions (urban, sub-urban, rural, inner city, commercial/industrial area) are defined manually. A similar approach has been taken by Meinel [10] who targets at the automatic delineation and interpretation of settlement classes from rasterized topographic maps. Sester [11] used a supervised Machine Learning approach to create decision trees for the description of topographic map features.

In contrast, the approach that we propose uses unsupervised classification for the derivation of implicitly given knowledge in the data. Thereby, the fundamental problem is the definition of relevant feature characteristics, which allows a clear grouping of features with similar geometrical attributes and also semantic meaning. To what extent geometric similarity of data instances corresponds to semantic similarity was previously studied by Kieler [12].

This paper is structured as follows. First the interpretation scheme based on an unsupervised classification is presented (Section 2). Then we describe the datasets that are used in this study and the preprocessing and enrichment steps that are necessary (Section 3). Subsequently the presentation and evaluation of the results obtained in our experimental tests are given in Section 4 and 5. Furthermore we propose a solution for integrating our results back into the original OSM data. Section 6 summarizes the achievements and gives an outlook on future work.

2. WORKFLOW

The overall workflow of our research is shown in Figure 1. Basically it consists of two parts. In the first part, buildings and city blocks from a training dataset are automatically separated into different clusters, based on their geometric and topologic characteristics. The results are judged by an expert who assigns the clusters to visually meaningful settlement classes. This visual inspection process is in the spirit of the discipline of Visual Analytics, which aims at integrating humans and machines and exploiting their relative benefits in cognition (human) and calculation (machine). After the training, in the second part, buildings and city blocks of test datasets can be automatically assigned to the respective clusters.

The parts are described in more detail in the following. The buildings of the OSM training dataset have to be pre-processed in order to create a consistent dataset, especially concerning topology. Based on the shape, i.e. the geometry, and the relations, i.e. the topology, several characteristic measures are computed for each building. The enriched data is then processed by an unsupervised classification, which automatically determines clusters of buildings that are similar based on the calculated measures. Some of the clusters are then grouped together using expert knowledge to form similar building types.

In our approach, we also seek for an interpretation of city blocks, which represent aggregations of individual buildings, again based on unsupervised clustering. City blocks are geometrically derived based on the road lines that enclose the blocks. Road lines refer here to any type of transport network, including e.g. motorways as well as pedestrian paths. The derived city blocks are used to delineate settlement areas and are also enriched by geometric measures. In addition, the cluster types of the buildings (from the first step) located inside a block are used as additional characteristics in the unsupervised classification of the city blocks.

In the test phase, the learned model is subsequently applied to another OSM and a cadastral test dataset. The result of the clustering are buildings and city blocks with similar visual (primarily geometric and topologic) characteristics. As such, they do not have a dedicated semantics. In order evaluate the results, we compare them to settlement types given in a topographic dataset.

For the presented approach we use Data Mining techniques. The term Data Mining refers to the process of discovering interesting, implicit, and previously unknown knowledge from large databases [13]. We apply a clustering process, in order to group objects with similar attributes into clusters, which are dissimilar to objects in other clusters. In general, four categories of clustering algorithms exist: hierarchical, density-based, grid-based, and partitioning methods. For our approach we chose a partitioning algorithm. The K-Means-Algorithm would indeed be the simplest and fastest algorithm but a negative aspect is that the number of cluster centers k must be known in advance and the algorithm does not yield reliably the best solution. Also the results are not the same in each run, because it does depend on a great deal on the initial random assignments. There would have been many test runs needed to train the classifier for our specific task. For this reason we chose the Expectation Maximization algorithm (EM), because this algorithm organizes objects into k clusters, without a priori information neither on the number of clusters nor their composition, so that the total deviation of each object from a cluster distribution is minimized. That implies, that the EM algorithm yields for each run the same result, however the success of the algorithm depends strongly on the assumed probability distribution. For more details see [14, 15].

For our investigation we specially use the EM algorithm implementation of the WEKA Data Mining Software (http://www.cs.waikato.ac.nz/ml/weka).

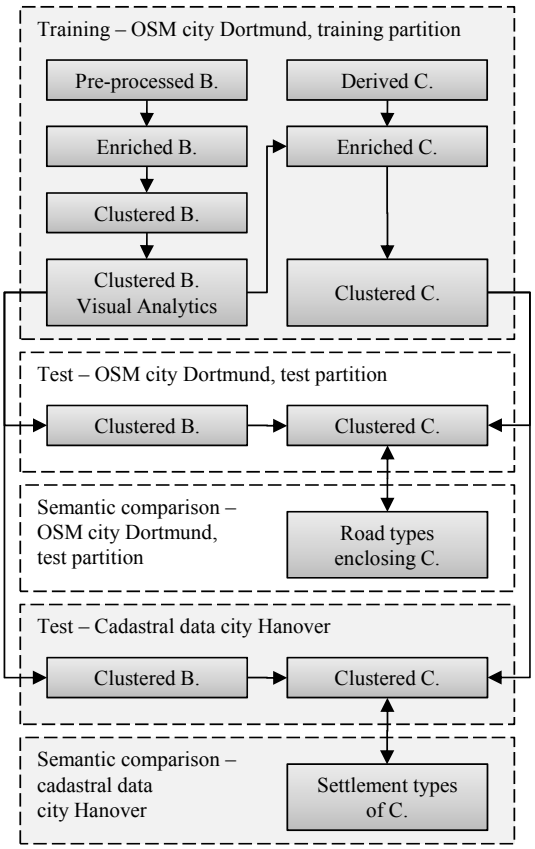


Figure 1. Workflow; (B)uildings, (C)ity blocks

3. DATASETS AND DATA ENRICHMENT

3.1 Datasets and Pre-Processing

We use datasets of two German cities for automatic interpretation of settlement areas. For both cities detailed building outlines and detailed road lines are used in the workflow. The dataset for the city of Dortmund covers about 38 km² and is taken solely from OSM (Figure 2).

Although the OSM project started with mapping streets, it aims at creating a digital world map of everything. Its detail regarding transport networks is still one of its strengths. The dataset includes information about the complete road hierarchy, ranging from motorways to residential streets, cycle ways, and foot paths. But OSM also includes other feature types in some areas, such as buildings or land use. Detailed building polygons exist e.g. for Vienna (Austria), Grenoble (France), Prague (Czech Republic), and Cambridge (USA). Such detailed data is based on digitization of analogue maps or aerial photographs, donations by local authorities, or extensive mapping activities of volunteers.

As already mentioned in the introduction, OSM data is captured mostly by non-experts. Therefore some pre-processing steps are necessary in order to gain a consistent dataset. Topological errors have to be fixed in the derived output. These errors are common to OSM data for two reasons. Firstly, OSM is aimed at creating maps, therefore duplicate features and small gaps or overlays of buildings do not change the map output significantly. Secondly, topology is often not common to non-experts. In the used dataset for the city of Dortmund, e.g. about 16% of all buildings do not comply to the rule that polygons must not overlap. Therefore all topological errors were fixed by using both automatic and manual tools.

The second city we use in our work is Hanover. In contrast to the city of Dortmund, the building outlines and road lines for about 21 km² of Hanover are taken from data captured by the national mapping agency. Detailed building outlines are provided by real estate cadastral data (ALK), whereas road lines and settlement area types are provided by topographic data (ATKIS).

Unfortunately there are no further detailed attributes given, such as height information of the buildings. This information would improve the classification e.g. by distinguishing different building types with a similar outline, like detached houses from standalone high rise buildings.

In Table 1, the total number of buildings and city blocks for the datasets are listed. In order to create a training and test dataset for unsupervised classification, OSM data of Dortmund is split into two partitions as shown in Figure 2.

Table 1. Number of buildings and city blocks per dataset

City	# Buildings	# City blocks
Dortmund (training partition)	8413	626
Dortmund (test partition)	10294	651
Hanover	14520	556

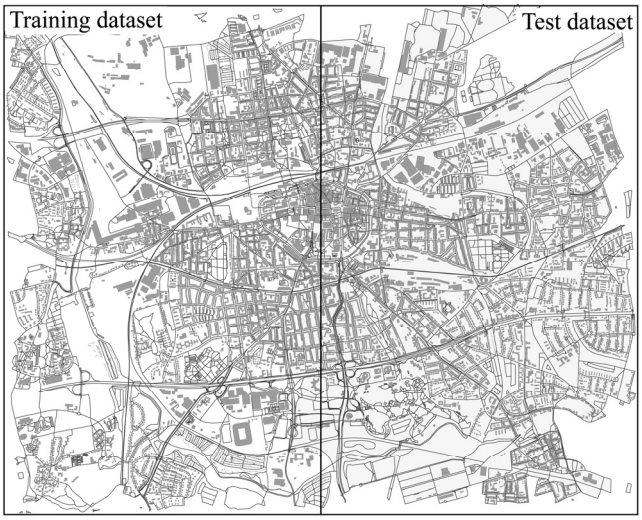


Figure 2. Partitions of the city of Dortmund

3.2 Data Enrichment

Before starting with the clustering process, several geometric and topologic measures have to be determined, that allow grouping of features in different classes. We first present the measures for the enrichment of buildings (Section 3.2.1) and then for city blocks (Section 3.2.2).

3.2.1 Buildings

The measures we chose for the enrichment of the building features are listed in Table 2. Beside the basic geometric attributes of a building like area and perimeter *(BuildArea, BuildP)*, we also consider the minimum enclosing rectangle *(Mer)* with its parameters area *(MerArea)*, length *(Mer L)* and width *(Mer W)*.

The complexity of a building can be expressed by the ratio of the building area to the area of its Mer *(BuildMerRa)*. That is the more complex a building is, the lower the ratio becomes. The ratio *(BuildMerEl)* between length and width characterizes the elongation of the feature. The rectangularity *(BuildRect)* is based on the assumption that buildings generally have parallel borders and right angles. Here, we disregard all vertices of the building outline leading to an angle difference smaller than 10° to the previous vertex. By this approach errors caused by inaccurate digitization are ignored. The measure compactness relative to a square *(BuildComp)* also gives an indication for the shape complexity of the building. Furthermore, topological relations to neighboring features allow to draw inferences about the building type. That is, e.g. a terraced house has at least one but no more than two neighbors, whereas a semi-detached house has exactly one neighbor *(BuildNeigh)*. To minimize the number of neighbors, we exclude building neighbor relations, which meet only in one point. If no adjacent neighbor exists, we compute the shortest distance to the next building *(BuildDist)*.

Based on a correlation analysis we identified the measures, that are not correlated. In Table 3 the correlation matrix is shown. The matrix is symmetric and rows and columns contain the measures of Table 2 in the same order. There is a strong correlation (values higher than 0.75 marked by gray fields) between *BuildArea, BuildP, MerArea*, *Mer L* and *Mer W*. Therefore we only consider the bold marked measures in Table 2 for the building clustering process.

Most of these measures are part of a software component, which is used for checking the data integrity of spatial datasets. Werder [17] gives a detailed overview about spatial constraints, as well as their formalization.

Table 2. Building measures

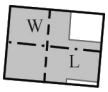
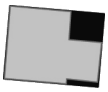
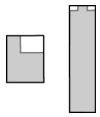
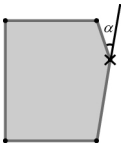
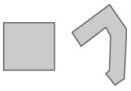
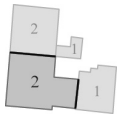
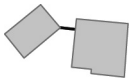
Measure	Description / Formula	Figure
BuildArea *BuildP*	Area of the building outline Perimeter of the building	
MerArea *Mer L* *Mer W*	Area of the minimum enclosing rectangle (*Mer*) Length (*L*) and Width (*W*) of Mer	
BuildMerRa	$BuildMerRa = \frac{BuildArea}{MerArea}$	
BuildMerEl	$BuildMerEl = \frac{Mer\,L}{Mer\,W}$	 1.3 4.2
BuildRect	Rectangularity of the building (threshold $\alpha < 10°$) $BuildRect = 1 - \frac{\sum_{i=1}^{n} \left\lvert \lvert\Delta\alpha_i\rvert - \frac{\pi}{2} \right\rvert}{n\frac{\pi}{2}}$ n = number of polygon points α = angle	
BuildComp	Compactness relative to a square $BuildComp = \frac{BuildP^2}{BuildArea \cdot 4^2}$	 1 2.2
BuildNeigh	Number of neighbors except single point contacts	
BuildDist	Shortest distance between neighboring buildings	

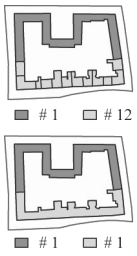
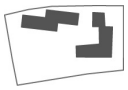
Table 3. Correlation matrix; high correlated values are marked by gray fields

BuildArea	1										
BuildP	0.82	1									
MerArea	0.90	0.80	1								
Mer L	0.79	0.95	0.75	1							
Mer W	0.82	0.91	0.78	0.82	1						
BuildMerRa	-0.26	-0.53	-0.31	-0.49	-0.56	1					
BuildMerEl	0.10	0.27	0.07	0.46	-0.01	-0.05	1				
BuildRect	-0.13	-0.18	-0.11	-0.19	-0.22	0.42	-0.01	1			
BuildComp	0.33	0.73	0.41	0.68	0.53	-0.67	0.46	-0.13	1		
BuildNeigh	-0.07	-0.05	-0.07	-0.05	-0.02	-0.08	-0.02	-0.13	-0.03	1	
BuildDist	0.12	0.18	0.10	0.18	0.17	-0.08	0.07	-0.01	0.13	-0.43	1

3.2.2 City Blocks

In order to enrich a city block feature we calculate the measures listed in Table 4. For this purpose some of the buildings measures can be reused, e.g. the area *(BlockArea)* and compactness *(BlockComp)*. Additionally we introduce the spatial relation *containment* by determining the number of buildings lying within one city block *(BlockNbrBuildG_i)*, though differentiated by the grouped building clusters from the first step of the process. Based on the mixture of different building types in a city block, the land use type can be distinguished. Due to strong varieties during the building modeling process, we aggregate adjacent buildings with the same type to a single feature (see third row in Table 4). Furthermore, the rate of the built-over area is a measure for the density in a city block and consequently for the settlement type *(BlockDensBuildG_i)*.

Table 4. City block measures

Measure	Description / Formula	Figure
BlockArea	Area of the city block	See Table 2.
BlockComp	Compactness relative to a square	
BlockNbrBuildG$_i$	Number of buildings for each type G_i within a city block (bottom: after aggregation)	
ToAreaG$_i$	Total area for each building type within a city block	
BlockDensBuildG$_i$	$BlockDensBuildG_i = \frac{ToAreaG_i}{BlockArea}$	

4. EXPERIMENTAL RESULTS

4.1 Building Classification

We use a two step approach for building classification. In the first step, clusters are automatically determined by an unsupervised classification. It is based on the measures presented in the previous section. In the second step, clusters are grouped together in order to reduce the number of building clusters. This is done by visual inspection and is a manual step based on expert knowledge about settlement patterns. If buildings from two or more clusters dominantly fall into the same settlement pattern, these clusters are grouped together to form a single, more homogeneous, cluster.

4.1.1 Unsupervised Classification

Using the uncorrelated measures from Section 3.2.1, unsupervised classification of the training partition of Dortmund results in ten building clusters. The relative distribution of total 8413 buildings among these clusters (C1 ... C10) is shown in Figure 3.

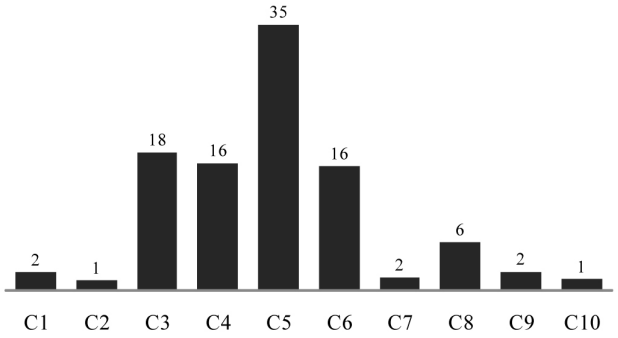


Figure 3. Relative distribution of buildings among clusters [%]

Due to clarity reasons, the individual values for the defined building measures of each of the ten clusters are not listed. Instead, the minimum and maximum values calculated over all clusters are shown in Table 5. The table also includes the minimum and maximum standard deviations.

Table 5. Min. and max. values for building measures

Measure	Minimum		Maximum	
	mean	std.dev.	mean	std.dev.
BuildArea	94.17	57.64	4305.50	6112.34
BuildMerRa	0.44	0.00	1.00	0.14
BuildMerEl	1.36	0.27	3.47	2.56
BuildRect	0.75	0.00	1.00	0.22
BuildComp	1.04	0.04	3.89	1.76
BuildNeigh	0.01	0.11	2.65	1.58
BuildDist	0.00	0.01	21.75	77.83

The partly high standard deviations, as well as the differences between minimum and maximum values, can be traced back to the variety of building geometries in the training dataset. It ranges from small, strictly rectangular building geometries in plot gardens to large and geometrically complex industrial buildings.

4.1.2 Visual Inspection and Grouping

Although the unsupervised classification in exactly ten clusters is statistically sound, neighboring buildings of roughly the same shape may fall into different clusters. As shown in Figure 4, the individual buildings forming a building block are assigned to different clusters. Nevertheless, from a semantic viewpoint, the small deviations in the shape of buildings in a single row of the block are negligible.

From the perspective of interpretation of settlement areas, these buildings share the same settlement type, namely block structure. Therefore it is reasonable to group the building clusters, in this case C4, C5, and C6, into a single cluster G2. This process is guided by expert knowledge as well as by comparing the values of the defined building measures. In order to illustrate this decision, the calculated measures for this largest building group G2 are summarized in Table 6. The measures that characterize

this group are *BuildArea*, *BuildNeigh*, and *BuildDist*. As can be also seen in Figure 4, clusters C4 and C5 cover smaller buildings, whereas C6 covers slightly larger buildings. Buildings of this group have a mean number of neighbors that is between one and three, which is also reflected in the value for the measure *BuildDist*. The values of all other measures differ between the three clusters, e.g. buildings in C6 are often elongated, whereas buildings in C4 and C5 correspond more to the form of a square.

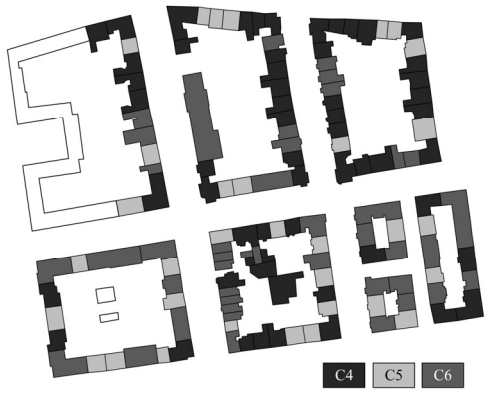


Figure 4. Block structure

Table 6. Grouping of clusters C4, C5, and C6 to G2

Measure		Cluster		
		C4	**C5**	**C6**
BuildArea	mean	189.86	122.69	590.53
	std.dev.	106.95	57.64	477.47
BuildMerRa	mean	0.83	0.99	0.93
	std.dev.	0.09	0.02	0.06
BuildMerEl	mean	1.42	1.36	3.18
	std.dev.	0.37	0.27	1.27
BuildRect	mean	0.79	0.99	0.97
	std.dev.	0.09	0.01	0.04
BuildComp	mean	1.16	1.04	1.5
	std.dev.	0.16	0.04	0.32
BuildNeigh	mean	1.89	1.46	2.65
	std.dev.	0.69	0.75	1.44
BuildDist	mean	0.01	0.29	0.00
	std.dev.	0.09	1.05	0.01

In this way, starting with ten clusters from unsupervised classification, these are grouped into a total of five clusters being more closely related to settlement types. The following descriptions for the grouped clusters cover the majority of the buildings, however also individual buildings fall into these groups, that differ from the group characteristics. The grouped cluster G1 (C1, C2, C3) includes smaller, standalone, and rectangular buildings. G2 (C4, C5, C6) combines homogenous building blocks or rows. G3 actually corresponds to the single building cluster C7, because this cluster is already homogeneous. Its buildings are in most cases large and complex, i.e. having a small value for the *BuildRect* measure. The grouped cluster G4 (C8, C9) consists of buildings that mainly have the shape of the letters I, L, T, or U. It includes buildings with a wide range of area sizes, starting from single family detached houses to large buildings. The last building group, G5, corresponds to C10 and is characterized by maximum values for the measures *BuildComp* and *BuildMerRa*, which are based on the complexity of buildings in this cluster. In Figure 5 exemplary buildings are shown for each of the grouped clusters.

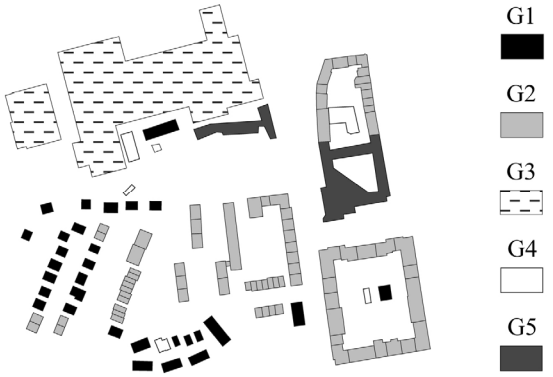


Figure 5. Building examples for grouped clusters

4.2 City Block Classification

City blocks are classified based on two measure types. The first measure type describes the shape of the block and includes area and compactness. The second measure type is based on the building classification from Section 4.1. Both the number of aggregated buildings as well as their occupancy of the city block are used as input for the classification. By this approach, the classification result is transferred from individual aggregated buildings to the enclosing city blocks, which represent a coverage. Therefore, this step can also be seen as a generalization to a smaller map scale.

Unsupervised classification of the city blocks, using the previously presented measures, reveals six clusters. For the 626 city blocks of the training area the relative distribution into these clusters is shown in Figure 6. The values for the other datasets are discussed in detail in Section 4.3.

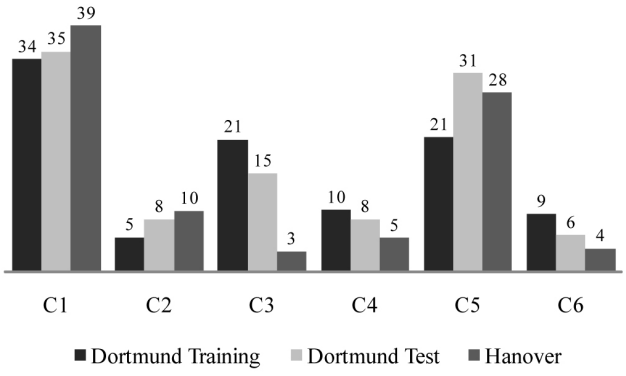


Figure 6. City block clusters for all datasets [%]

In contrast to the two step approach of building clustering and cluster grouping, the automatically determined city block clusters show already distinctive characteristics, thus no subsequent grouping step is needed. The values of the three important measures *BlockArea*, *BlockComp*, and the summarized *BlockDensBuild* for all building clusters, are shown in Figure 7. All values are normalized to [0,1] based on minimum and maximum value for a measure among all clusters, e.g. *BlockComp* has a minimum value of 1.29 (C1) and a maximum value of 2.45 (C2).

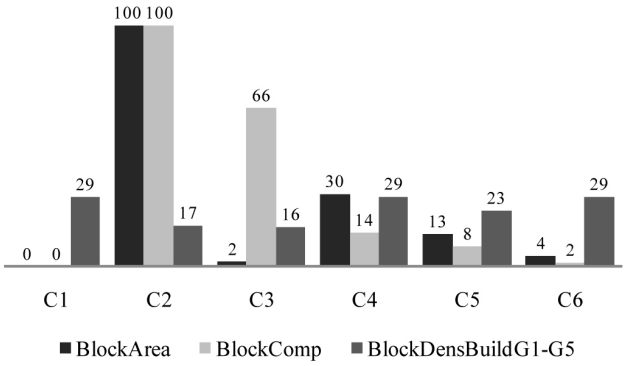


Figure 7. Selected measures for city block clusters [%]

City block cluster C1 is characterized by minimum values for the measures *BlockArea* and *BlockComp*. From the 29% summarized building density over all grouped clusters, about 25% is solely contributed by building group G2, which represents building block structure. *BlockComp* has the minimum value of 1.29, leading to compact polygons for the city blocks. Therefore C1 can be labeled as block structure, which may be closed but also consist of building rows.

Cluster C2 has the maximum values for both *BlockArea* and *BlockComp*. It also has a low value for the building density. However, the total density of 17% is distributed between all building groups leading to an heterogeneous building structure inside the city block.

An open structure is also characteristic for cluster C3. It is made up of a combination of building clusters G1 and G4, which include smaller and standalone as well as letter-shaped buildings.

The value of 29% for the overall building density of cluster C4 can be traced back to about 24% solely from building group G3. This group combines both complex and large buildings.

City block cluster C5 is dominated by buildings in block structure (G2), that can be also found in C1. In contrast, the block structure shows more gaps between individual buildings and also has in most cases a regular structure.

For the overall building density of 29% for cluster C6, building group G5 contributes about 23%. Having maximum values for *BuildComp* and *BuildMerRa*, C6 is made up of complex shaped buildings. In contrast, the measure *BlockComp* for the city block cluster has a rather low value. Therefore C6 can be characterized as non-compact buildings in compact city blocks.

Figure 8 presents a part of the training data showing both building groups and city block clusters.

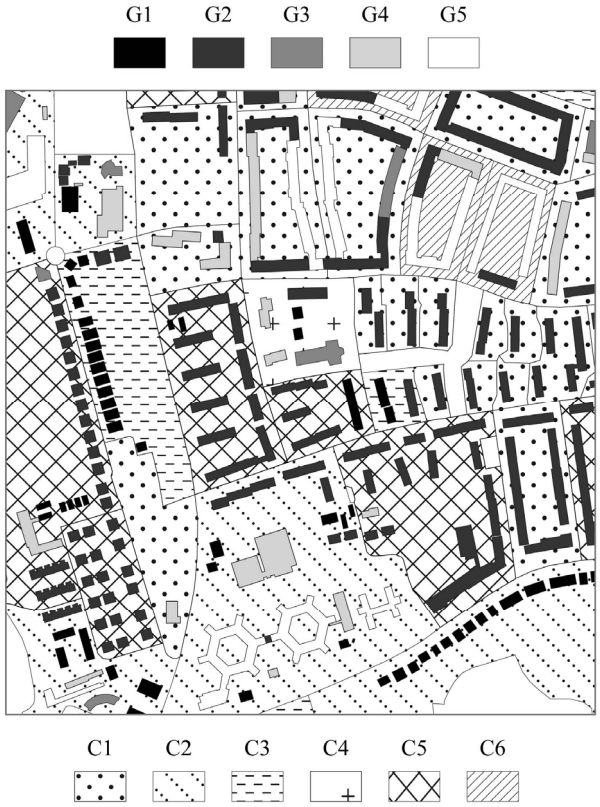


Figure 8. Area from training dataset with building groups (G1 ... G5) and city block clusters (C1 ... C6)

4.3 Evaluation of Clustering Results

In the approach, several configurations with different geometric and topologic measures have been tested. The result presented in this paper reflects the solution after an appropriate filtering of correlated measures. However, it was possible to extract and combine clusters that geometrically and visually correspond to different building and city block types also for configurations having correlated or insignificant measures.

As most important part of the evaluation, the transferability of the obtained clusters from training to test data was investigated. Firstly, the results were applied to training data of the same city and same data provider, namely the OSM project. Secondly, they were applied to data of a different city from a different data provider, namely data obtained from cadastre and topography.

In Figure 9 the distribution of the five grouped building clusters is shown for each scenario. Using the model from the training dataset for clustering test datasets reveals their individual characteristics but also differences to the training dataset.

The OSM test dataset of the city of Dortmund reflects the actual settlement area types pretty well. East of the city center, the dominance of building blocks (G2) is partly exchanged for single family detached houses (G1). Also more large buildings,

representing industrial sites (G3), are located in the test dataset. Finally, covering more the geometric aspect of building outlines, building group G4 brings in additional letter-shaped buildings in the Dortmund test dataset.

The second test dataset, cadastral data of the city of Hanover, proves the transferability of the derived clusters also to a more dense city center area. Little standalone buildings exist in the dataset (G1). Instead, the settlement structure is clearly dominated by building group G2, representing homogeneous block or row structure, as the value of 77% indicates in Figure 9.

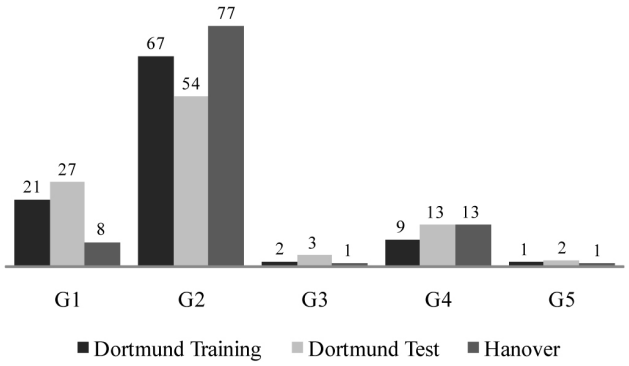


Figure 9. Building groups for all datasets [%]

These changes in settlement types are also reflected in the distribution of the city blocks for the test dataset, which can be seen in Figure 6.

For the OSM test dataset the differences are most significant for city block clusters C3 and C5. Blocks with smaller and standalone buildings (C3) decreased by 6%, whereas open block structures (C5) increased by 10%. This is due to the fact that, upon visual inspection, the test area covers more suburbs with multifamily houses instead of single family houses.

The high density of buildings in the cadastral dataset is also reflected in the distribution of its city blocks, as can be seen in Figure 6. Block structures, mainly summed up in city block clusters C1 and C5, dominate the visual perception of the city center area. This in turn reduces drastically the number of blocks with an open structure amalgamated in class C3, which drops from 21% for the training dataset to only 3% in the cadastral test dataset.

The evaluation of the transferability of both building and city block clusters to training datasets performed well. As the identified clusters cover settlement characteristics from dense city centers to the open structure of suburbs, they are also likely transferable to other settlement areas.

5. COMPARISON WITH KNOWN SEMANTIC CLASSES

Up to know, clusters have been derived, which are visually plausible, but do not have a clear semantics attached to it. In the following, these clusters are linked and compared to existing semantic classes using two different data sources.

5.1 OpenStreetMap Road Lines

The road network in OSM data is modeled in high detail. The map object catalogue [16] lists more than 25 values for the tag "highway". Because the city blocks we use in our approach are directly derived from the OSM road network, we are able to compare the cluster of a city block with the type of roads that enclose the block.

We group the road types from OSM into the following five categories: Highway (Motorway, Primary, Trunk), DistrictRoad (Secondary, Tertiary, Road, Unclassified), Residential (Living street, Residential), Service and Pedestrian (Footway, Path, Pedestrian, Steps, Track, Cycle way). For the comparison, we summarize the line lengths of all roads enclosing the city blocks and normalize the results to [0,1] by dividing the obtained value through the sum of all city block perimeters. The resulting statistic is shown in Table 7.

Table 7. City block clusters to OSM road lines [%]

Road type	City Block Cluster					
	C1	C2	C3	C4	C5	C6
Highway	2	2	0	4	1	2
DistrictRoad	12	20	6	16	10	12
Residential	64	28	22	15	56	47
Service	4	7	7	20	6	13
Pedestrian	17	21	50	22	20	22

The semantic of the road types can then be compared to the city blocks they enclose. Cluster C1 and C5 are dominated by Residential roads, which indicate residential settlement areas. Both clusters are characterized by a block structure, either closed or in row, as shown in Figure 8. This actually shows a high conformance to the structure of residential areas in German city centers.

City block clusters C2 and C4 do not have clear dominant road types. This complies to the already described fact, that C2 describes a heterogeneous building structure. C4 is composed of large and complex buildings. When inspected visually, these two clusters include both residential and industrial areas. Especially cluster C4 covers mainly industrial sites, which is also confirmed by high values for the road types Highway and Service.

Pedestrian areas make up more than half of the road types of city block cluster C3. When compared to the OSM data, mainly plot gardens fall into this cluster, which shows a fairly good separation of this settlement area type.

From the values of the road types for cluster C6, it can be also characterized as mainly residential. However, this cluster cannot be clearly assigned to a single settlement area type, because it is solely defined by the compactness of both city block and building polygons.

At this point, we also want to answer the question of how to integrate our results back into OSM in order to improve and enrich the data. The most obvious way is to use the calculated measures from Section 3.2 to detect geometric and topologic

errors or outliers, e.g. buildings having an area smaller than 10m² or a distance to neighboring buildings smaller than 1m. For the semantic enrichment, the cluster type can be introduced to the OSM map object catalogue [16] as a new tag for buildings or instead of the "yes" value for the tag "building". As city blocks are not yet considered in the map object catalogue, they can be added to the schema as a new polygon type with respective attribute-value pairs. However, for each change in the enclosing road lines also the update of the respective city blocks has to be triggered automatically.

5.2 ATKIS Settlement Areas

For the comparison of the derived city block classification of Section 4.2 with the settlement types of ATKIS we first present the existing settlement types along with short textual object catalogue descriptions. In order to avoid the complete presentation of all available object classes [18], we only consider the following five main classes:

1) Residential Area: Area with buildings, predominantly or solely used for residential purposes. Besides these residential buildings also shops to supply this area, non-disturbing craft producers, facilities for religious, cultural, social and sanitary purposes are permitted.
2) Industrial Area: Area with buildings, predominantly or solely used for industrial or craft producing purposes. This includes, e.g. shopping malls, warehouses/depots, large-scale commercial farms, processing and disposal plants and trade fair facilities.
3) Mixed Area: Area with buildings without a typical purpose. This includes especially areas with a rural character, e.g. agricultural or forestry companies, residential buildings and central areas in a city with commercial buildings and vital economic and administrative facilities.
4) Area of Special Usage: Area with buildings of certain purposes. This includes purposes of administration, health and social affairs, education, research, culture, safety and order, vacation or weekend homes and national defense.
5) Green Area: Area with green spaces and sport grounds.

These class descriptions are very detailed, but in our opinion this strict distinction cannot be determined solely from visual characteristics, e.g. the dominant functional usage of a settlement area cannot be solely seen from an aerial image. For this reason we expect that our derived city block classification differs from the semantics used in the ATKIS object catalogue.

For the evaluation we calculated the intersection between both classifications. In Figure 10 the distribution of the settlement areas of ATKIS in the city block clusters of our approach is shown.

By analyzing these values we first reveal that the class Mixed Area has the largest proportion in four of the six city block clusters. One reason could be, like the name implies, that the settlement type comprises a large variety of buildings and therefore a clear assignment to one specific cluster is not possible. Furthermore, it is obvious that Residential Area and Mixed Area predominantly occur together in the same range in city block cluster C1 and C5. In addition, C5 is dominated by Areas of Special Usage, noting that the settlement type has the highest ratio related to the overall area of the clusters. However, that confirms our assumption, that both clusters C1 and C5 represent residential areas, and therefore semantic similarities between the different class definitions exist.

Green Areas are also present in each city block cluster, e.g. parks for recreation purposes located in each residential area. In cluster C2 all ATKIS settlement types are present at a similar level, backing our interpretation of a very heterogeneous building structure in this cluster. Furthermore, the ratio of the industrial areas related to the overall size of this cluster (16%) is higher compared to the other five clusters.

In conclusion, we observe, that the derived semantic annotations for our city block clusters fit more to the OSM data (Section 5.1) than to the ATKIS settlement areas (Section 5.2). This result can be explained by the classification of the OSM road line data, finally forming our city blocks, that is more detailed and strongly influenced by the visual characteristics and structure of the individual buildings. For example for volunteers it is easier to recognize road lines which are only used by pedestrians or road lines in a residential area. Thus our building blocks form a settlement characteristics; in contrast, ATKIS more refers to a general land use characteristics, which also includes functional parameters not visible in geometric structures (like area of special usage).

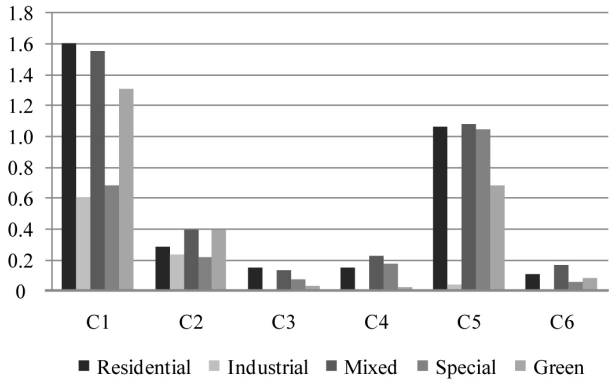


Figure 10. Area distribution of ATKIS settlement types into the derived city block clusters [km²]

6. CONCLUSION AND OUTLOOK

The volunteered geographic information of the OpenStreetMap project offer a large amount of data, but the data may be incomplete or unequally distributed and therefore cannot replace systematically acquired official data, especially if a consistent coverage is required. Semantic annotation of this data, however, makes it more valuable and also more usable for different purposes. Examples are data integration, data analysis and map generalization.

We have presented an approach for deriving semantic annotations for building and city block features based on the analyses of geometric and topologic characteristics. The advantage of the approach is that it is based on an unsupervised method, which avoid the often times-consuming and difficult generation of appropriate training data. Our results are visually convincing and

provide additional information, which can be exploited in various ways, especially for applications where the individual classification is not of highest relevance and a more general summative information is needed. One such application is the determination of a settlement typology which can be used for spatial disaggregation of statistical parameters regarding population, housing, economics and infrastructure. Examples for its usage are a better estimation of the spatial distribution of inhabitants over an area or the demand for infrastructure (e.g. heating [19]), or better monitoring of settlement and open space development [10]. The information about building structures can be used for cartographic generalization and visualization.

There are several issues that give rise to further research. On the one hand, additional measures could be determined, e.g. the maximum width of a building or shape complexity based on Fourier analysis. Also, an Bayesian interpretation scheme as proposed by Lüscher et al. [9] could be beneficial, as it also takes the probabilities of the characteristics into account and thus is able to model uncertainties and also interdependencies.

ACKNOWLEDGMENTS

The authors would like to thank C. Brenner for valuable discussions.

The research described in this paper is funded by the German Science Foundation (DFG) and the German Federal Ministry of Education and Research (BMBF). The support is gratefully acknowledged.

REFERENCES

[1] Haklay, M. 2008. How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. Under review in Environment & Planing B: Planning and Design.

[2] Zielstra, D., and Zipf, A. 2010. A Comparative Study of proprietary Geodata and Volunteered Geographic Information for Germany. In Proceedings of the 13th AGILE International Conference on Geographic Information Science (Guimarães, Portugal).

[3] Heinzle, F., Anders, K.-H., and Sester, M. 2006. Pattern Recognition in Road Networks on the Example of Circular Road Detection. In *Geographic Information Science, LNCS*. Berlin: Springer, 4197, 153-167.

[4] Steiniger, S., Burghardt, D., and Weibel, R. 2006. Recognition of Island Structures for Map Generalization. ACM-GIS 2006 - International Symposium on Advances in Geographic Information Systems, 67-74.

[5] Christophe, S., and Ruas, A. 2002. Detecting Building Alignments for Generalization Purposes. Advances in Spatial Data Handling. Berlin: Springer, 419-432.

[6] Zhang, X., Ai, T., and Stoter, J. 2010. Characterization and Detection of Building Patterns in Cartographic Data. In *Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*. 38, 2.

[7] Lüscher, P., Weibel, R., and Mackaness, W. 2008. Where is the Terraced House? On The Use of Ontologies for Recognition of Urban Concepts in Cartographic Databases. In Headway in Spatial Data Handling*, LNGC*. Berlin: Springer, 449-466.

[8] Lüscher, P., Weibel, R., and Burghardt, D. 2009. Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data. Computers, Enviroment and Urban Systems. 33, 363-374.

[9] Burghardt, D., and Steiniger, S. 2005. Usage of Principal Component Analysis in the Process of Automated Generalisation. In Proceedings of 22nd International Cartographic Coference (A Coruña, Spain).

[10] Meinel, G. 2008. High Resolution Analysis of Settlement Structure on Base of Topographic Raster Maps - Method and Implementation. In *Computational Science and Its Applications - ICCSA 2008, LNCS*. Berlin: Springer, 5072, 16-25.

[11] Sester, M. 2000. Knowledge acquisition for the automatic interpretation of spatial data. International Journal of Geographic Information Science, 14, 1-24.

[12] Kieler, B., Sester, M., Wang, H., and Jiang, J. 2007. Semantic Data Integration: Data of Similar and Different Scales, *Photogrammetrie Fernerkundung Geoinformation (PFG)*. 6 , 447-457.

[13] Fayad, U., Piatetsky-Shapiro, G., Smith, P., and Uthurusamy, R. 1996. Advances in Knowledge Discovery and Data Mining. Menlo Park CA: American Association for Artificial Intelligence.

[14] Bradley, P., Fayyad, U.M., and Reima, C.A. 1998. Scaling EM (Expectation-Maximization) clustering to large databases. *Microsoft Research Technical Report 98-35*.

[15] Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*. B 39, 1-38.

[16] OpenStreetMap Foundation. 2010. Map Features. http://wiki.openstreetmap.org/w/index.php?title=Map_Features, retrieved on 2010-06-14.

[17] Werder, S. 2009. Formalization of Spatial Constraints. In *Proceedings of 12th AGILE Conference on GIScience*. CD.

[18] AdV. 2008. ATKIS-Objektartenkatalog. http://www.atkis.de, retrieved on 2010-06-30.

[19] Neidhart, H., and Sester, M. 2006. Creating a digital thermal map using laser scanning and GIS. In *Proceedings of the District Heat and Cooling Symposium* (Hanover, Germany).