Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data

Lüscher, P; Weibel, R; Burghardt, D
Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data

Abstract

This paper presents an ontology-driven approach for spatial database enrichment in support of map generalisation. Ontology-driven spatial database enrichment is a promising means to provide better transparency, flexibility and reusability in comparison to purely algorithmic approaches. Geographic concepts manifested in spatial patterns are formalised by means of ontologies that are used to trigger appropriate low level pattern recognition techniques. The paper focuses on inference in the presence of vagueness, which is common in definitions of spatial phenomena, and on the influence of the complexity of spatial measures on classification accuracy. The concept of the English terraced house serves as an example to demonstrate how geographic concepts can be modelled in an ontology for spatial database enrichment. Owing to their good integration into ontologies, and their ability to deal with vague definitions, supervised Bayesian inference is used for inferring complex concepts. The approach is validated in experiments using large vector datasets representing buildings of four different cities. We compare classification results obtained with the proposed approach to results produced by a more traditional ontology approach. The proposed approach performed considerably better in comparison to the traditional ontology approach. Besides clarifying the benefits of using ontologies in spatial database enrichment, our research demonstrates that Bayesian networks are a suitable method to integrate vague knowledge about conceptualisations in cartography and GIScience.

doi:10.1016/j.compenvurbsys.2009.07.005
Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data

Patrick Lüscher *, Robert Weibel †, Dirk Burghardt °
† Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich (Switzerland)
° Institut für Kartographie, Technische Universität Dresden, Helmholtzstraße 10, 01062 Dresden (Germany)

* Corresponding author. Tel.: +41 44 635 52 17; fax: +41 44 635 68 48.
E-mail addresses: patrick.luescher@geo.uzh.ch (P. Lüscher), robert.weibel@geo.uzh.ch (R. Weibel), dirk.burghardt@tu-dresden.de (D. Burghardt)

Keywords: pattern recognition, spatial database enrichment, ontology, Bayesian network, supervised classification, building types

Abstract

This paper presents an ontology-driven approach for spatial database enrichment in support of map generalisation. Ontology-driven spatial database enrichment is a promising means to provide better transparency, flexibility and reusability in comparison to purely algorithmic approaches. Geographic concepts manifested in spatial patterns are formalised by means of ontologies that are used to trigger appropriate low level pattern recognition techniques. The paper focuses on inference in the presence of vagueness, which is common in definitions of spatial phenomena, and on the influence of the complexity of spatial measures on classification accuracy. The concept of the English terraced house serves as an example to demonstrate how geographic concepts can be modelled in an ontology for spatial database enrichment. Owing to their good integration into ontologies, and their ability to deal with vague definitions, supervised Bayesian inference is used for inferring complex concepts. The approach is validated in experiments using large vector datasets representing buildings of four different cities. We compare classification results obtained with the proposed approach to results produced by a more traditional ontology approach. The proposed approach performed considerably better in comparison to the traditional ontology approach. Besides clarifying the benefits of using ontologies in spatial database enrichment, our research demonstrates that Bayesian networks are a suitable method to integrate vague knowledge about conceptualisations in cartography and GIScience.
1 Introduction

Spatial databases currently produced by national mapping agencies (NMAs) are typically modelled closely after the original map products which they replaced, meaning that they are rich in geometry but poor in semantics, particularly with regards to the representation of higher order geographic concepts that extend beyond the semantics of individual, discrete objects. Examples of geographic concepts that are not coded in current spatial databases include the geomorphological process underlying stretches of a coastline (estuary, fjord, skerry etc.), the extent of an urban settlement, neighbourhood types (residential, industrial etc.), or building types (detached, semi-detached, terrace etc.).

One area that could obviously benefit of richer semantics in spatial databases is map generalisation. Map generalisation aims to derive a model of the geographic reality that is appropriate for portrayal at a certain scale and purpose. It is important to note that this abstraction process is not just a matter of simplification of detailed situations to reduce spatial clutter and therefore guarantee legibility of a map; rather, different phenomena and patterns have to be portrayed at various scale levels (Brassel & Weibel, 1988). Bertin (1967/1999) therefore distinguishes conceptual generalisation and structural generalisation. Conceptual generalisation happens when “a city emerges from a collection of houses and streets”, or a “coal pan from a collection of coal mines”. Structural generalisation simplifies geometry, but conserves conceptualisation. More recently, this dichotomy has been termed model (or model-oriented) generalisation and cartographic generalisation (Grünreich, 1992).

While higher level geographic concepts are not explicitly coded in current spatial databases, they are nevertheless implicitly contained, owing to the fact that there often exists a relationship between the form (i.e. geometry) and function (i.e. semantics) of real-world phenomena, particularly in the built environment. Hence, it is possible – at least to some extent – to “enrich” spatial databases retrospectively, making implicitly contained higher level geographic concepts explicit. This process is termed spatial database enrichment.

In particular, spatial patterns in the urban domain provide the basis for a variety of applications, such as urban planning or pedestrian navigation (Lüscher, Weibel, & Mackaness, 2008). The obvious example, again, is map generalisation. Take the case of a building that is too small to be fully legible on a target map. Here, semantic information is useful in deciding how to proceed: If the building is in a rural area (and hence rather isolated and presumably important), the building may be slightly enlarged; if it is in an urban area, it may be eliminated; and if it happens to be a special type of building such as a hospital, it may be replaced by a special symbol (Steiniger, 2007).

While there are a number of specific algorithms for data enrichment in spatial databases (Lüscher et al., 2008), the goal of the work in the present paper is to provide a modular approach to the overall process. The definition of spatial patterns is formalised through ontologies, which in turn can be used to drive the pattern recognition process.

The general approach was presented in an earlier paper (Lüscher et al., 2008). In the present paper, the following research questions are covered:

1. What methods are suited to classify instances with respect to formal definitions?
2. To what extent is it possible to use only simple measures (such as area and topological relations) to define complex concepts?

The premise is that the pattern recognition process needs to respect uncertainty of spatial data and vagueness of spatial knowledge. To address the first research question, an approach is presented that translates the ontology into a Bayesian network for carrying out fuzzy inference and for including training data. The approach is illustrated step-by-step using a case study that classifies English terraced houses in a topographic dataset. To address the second research question and to put the approach into the context of previous attempts to formalise pattern recognition, an alternative ontology that avoids complex spatial measures is taken as reference. Both ontologies are used to classify four English urban areas.

The remainder of this paper is organised as follows. Section 2 reviews previous approaches to model-based spatial pattern recognition. In Section 3 the need for ontology-driven pattern recognition process is presented, and the approach is outlined. In Section 4 we introduce the case study used in this paper – terraced houses – and define the corresponding ontology. Section 5 argues for an approach of fuzzy inference, based on the translation of the ontology into a Bayesian network. Section 6 presents two sets of experiments, one using a basic ontology not specifically defined for spatial database enrichment, and a second one using the ontology as developed in Section 4. Section 7 presents classification results. Section 8 discusses the ontology-driven approach with particular emphasis on the comparison of the two experiments. Finally, Section 9 rounds off the paper by conclusions and an outlook on future research.

2 Review of relevant literature

2.1 Related work on ontology-based spatial pattern recognition

Klien (2007) presents a framework for annotation of geodata, using Semantic Web technologies (Yu, 2007). She defines semantic annotation as creating links between feature types of a dataset and concepts of an external ontology, and argues that linking based on string-similarity of type/class names alone is too inaccurate. The semantic descriptions in the ontology are therefore used to derive instances of concepts and compare them with actual instances in the database. For example, she defines floodplain as a flat area adjacent to a river and not very much higher in altitude than the river (such that the area is regularly subject to flooding). This definition is translated to the Semantic Web Rule Language (SWRL, 2009). Spatial relations (such as adjacent) are mapped to spatial analysis operations, and regions representing flood plains are inferred through logic deduction. She argues that by following this strategy, instead of implementing a ‘black box’-approach, increased flexibility and transparency to the user is achieved. However, it is further argued that automatic classifications produced by the method are likely error-prone and need to be presented to a human user for final confirmation.

Thomson and Béra (2008) present a methodology for generating urban residential land-use through logic deduction. Increasingly complex spatial aggregates are generated starting from atomic concepts like house, garden, or road. As in the work of Klien (2007), spatial predicates are generated through spatial analysis operations in a GIS and exported to OWL-DL. The Web Ontology Language (OWL, 2008) is a family of
languages to author ontologies. Classification of buildings and plots is then carried out through Description Logic subsumption reasoning (Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2003).

Zhang, Stoter, and Ai (2008) propose a similar approach, although their goal is to improve reusability in cartographic constraint evaluation. During cartographic generalisation, cartographic constraints describe particular spatial settings for which preferred actions exist. For example, there exists a constraint that specifies that ‘roads leading to an isolated building should not be omitted’. Hence, detecting spatial settings corresponds to spatial pattern recognition. The proposed approach works again by decomposing complex spatial settings into simpler measures, and use some kind of predicate logic and/or terminological reasoning to infer instances, although a more detailed account of implementation is not given.

2.2 Uncertainty of geographic objects

Many concepts in the geospatial domain are poorly defined and traditional crisp logic is insufficient in dealing with uncertainty. Klien (2007) points out that “the notion ‘relatively low’ is not expressible in the logic of the representation language” (p. 444), but does not consider uncertainty in her framework further. According to Fisher (1999), there are two kinds of uncertainty associated with poorly defined concepts:

- **Vagueness**, which arises from poor definition of a class or individual object. As a consequence of vagueness, the extent of many spatial phenomena cannot be delimited sharply.
- **Ambiguity**, which arises from differing classification systems. The same road could be denoted as Expressway (by someone with US American background) or as Motorway (by someone with British background).

Dissolving ambiguity for enabling interoperability is one of the main applications of ontologies (Agarwal, 2005). Often concepts do not map one-to-one, but their meaning overlaps partially. Hence, there is increasing research interest in extending conventional reasoning with probabilistic techniques such that not only identical concepts can be deduced, but the most similar ones (Sen, 2008). Translating traditional OWL representations to Bayesian networks (Russell & Norvig, 2003) to carry out probabilistic reasoning is a promising approach (Zheng, Kang, & Kim, 2007). Recently, extensions such as PR-OWL (Costa & Laskey, 2006) or BayesOWL (Ding, Peng, & Pan, 2006) have been introduced to formalise translations from OWL into Bayesian networks.

Vagueness in classification arises because realisations of concepts are often imperfect and come with certain variations. For instance, ponds can be defined as a water body smaller than a lake, but the transition from pond to lake is gradual. As a consequence we are unable to define crisp thresholds for class membership. Fuzzy set theory (Fisher, Wood, & Cheng, 2004; Ladner, Petry, & Cobb, 2003) is an approach to account for this kind of uncertainty by defining fuzzy memberships. An alternative approach is Bayesian decision theory, by which class membership probabilities are estimated.

2.3 Contributions

The key contribution of this paper is the combination of an approach for ontology-based spatial pattern recognition with probabilistic inference to account for vagueness.
A probabilistic Bayesian approach is used for inference. The advantages of Bayesian inference are discussed in Section 5 and can be summarised as follows:

- Good integration into ontologies as shown by previous work on probabilistic OWL;
- sound inference also when multiple decisions are chained; and
- the simplicity of learning conditional probabilities from training data.

A second contribution is the introduction of abstract concepts that are defined algorithmically, but are formulated as simply and generally as possible (so that they can be re-used). A third contribution is the evaluation of the robustness of ontology-driven spatial database enrichment using large extracts of real data.

3 Ontology-driven spatial database enrichment

Lüscher et al. (2008) discussed algorithmic approaches to spatial database enrichment and argued why ontologies should be used to drive the pattern recognition process. Undoubtedly, existing algorithmic methods have been successful in detecting specific spatial patterns, but solutions that solely rely on algorithms also exhibit several important weaknesses:

- They have often been developed and parameterised for specific data models and databases. That limits the reusability of pattern recognition methods across different databases.
- They often make use of bespoke geometric algorithms and/or statistical techniques that do not reveal the ‘mechanics’ of the recognition procedure. Hence, they have limited transparency and explanatory value for the end user.
- They typically cannot be adapted to take into account additional information in the detection procedure, such as topography, which may be important in describing the genesis of certain patterns. That is, they have limited extensibility.

Ontologies have the potential to better inform the pattern recognition process with the aim of improving on some of the limitations of purely algorithmic approaches. Spatial concepts and their (spatial) relationships to other, ‘lower level’ concepts are explicitly modelled in an ontology. While the lowest level concepts are extracted through traditional spatial pattern recognition processes, they can be used to infer the existence of higher level concepts.

This ontology-driven approach proceeds in four steps (Lüscher et al., 2008): We draw on textual descriptions of urban spaces (step 1), then formalise these patterns, their context and hierarchical composition using methods from ontological engineering (Gómez-Pérez, Fernández-López, & Corcho, 2003) (step 2). The ontological definitions of patterns are then used to deductively trigger appropriate pattern recognition algorithms (step 3) in order to detect them in real spatial databases (step 4).

We use the term ‘ontology’ in the sense of the engineering sciences, where it is usually defined as an explicit specification of a shared conceptualisation (Gruber, 1993). It is thus an attempt to capture the knowledge of a certain domain in a systematic way
by breaking it down into the types of entities (concepts) that exist and the relations that hold between them. Therefore, in a first step, knowledge about the domain has to be collected. In this study, knowledge was extracted from the literature on urban development and urban history, complementing this information with the help of dictionaries and thesauri.

4 Ontologies of urban space descriptions

4.1 The case study of English terraced houses

It should be noted that according to the ontology definition given by Gruber (1993), there can be multiple ontologies for the same concept depending on the purpose the ontology is modelled for. The purpose of this research is to model ontologies for the detection of geographical concepts in spatial databases. Such an ontology has been built for the extraction of terraced houses (also called terrace houses or terraces) as they are conceptualised in urban morphology. Relevant concepts of the domain were extracted from a thesaurus of urban morphology (Jones & Larkham, 1991). Several case studies (e.g. Conzen, 1969) and a compendium about “The English Terraced House” (Muthesius, 1982) then gave more insight in the understanding of the concepts. By way of example, Figure 1 shows residential house types identified in the urban morphology literature. Mappings of terraced house settlements are provided in Section 7.

We use terraced houses as a case study for several reasons. First, they represent the most widespread housing type in English cities (Muthesius 1982) and building types such as terraced, semi-detached, and detached houses are commonly used in everyday speech. For instance, they give essential clues to prospective house buyers as to what to expect when reading through real estate advertisements (King, 1994). Second, knowledge about terraces, semi-detached and detached houses is also important in map generalisation. House types are used for typification of residential plots; for example, yards are merged differently in terraced house settlements than in detached and semi-detached settlements. Third, the concept of the terraced house integrates various low level concepts (as will be shown below) that can be re-used in similar concepts (e.g. other residential house types). And finally, it forms in turn a low level concept of other high level concepts, such as ‘residential area’. Hence, it may serve as an exemplar for testing the versatility and reusability of the ontology-driven approach to spatial database enrichment.
A textual description of the English terraced house can be summarised as follows: The construction of terraced houses is closely linked to the Public Health Act of 1875, which was established to improve urban living conditions and resulted in re-housing of population from slum clearance areas (Conzen 1969). The demand for cheap mass housing was met by creating rows of unified buildings sharing sidewalls. Owing to the low social status of the original dwellers, lot sizes and room footprints are small. Terraced houses usually have small front-gardens and possibly attached sculleries and a yard at the rear. Often, multiple rows of houses form an area of a highly regular plot pattern.

A concept map constructed from these descriptions is shown in Figure 2. Relations to simple properties, such as the area of a polygon, were included into the box of the concept itself, while relations that connect two (or more) concepts are drawn as arrows between them. This is for clearer visualisation only.

In the figure, terraced house is defined by its relations to other concepts. Some of those concepts are defined by relating them to even more basic concepts. For instance, the Oxford English Dictionary (Simpson & Weiner 1989) defines a yard as “a comparatively small uncultivated area attached to a house or other building, or enclosed by it”. This means, yard is defined by its area and its relations to uncultivated area and building. The concept map also contains abstract concepts which are to be implemented algorithmically as they constitute general units that are inefficient to break up further. One example is the concept row of houses, which denotes a linear, homogeneous arrangement of adjacent houses.

Having modelled terraced houses as conceptualised by humans, the concept map must be formalised to a pattern recognition process. This consists of two steps: On the one hand, explicit semantics have to be assigned to abstract concepts and relations by mapping them to (often spatial) operations. On the other hand, an algorithm has to carry out the classification process, inferring instances of concepts defined in the ontology.
Through these steps, an ontology is defined. In the remainder of this section, mapping of relations and concepts is discussed. Section 5 presents an approach for fuzzy inference, based on the translation of the ontology into a Bayesian network.

4.2 Mapping of spatial relations and abstract concepts

The meaning of predicates such as `adjacentTo`, `presenceOf`, and `hasArea` has to be interpreted by spatial analysis. `adjacentTo` denotes topological connection (i.e. adjacency) of two areas. The custom of embedding residential houses between front yards and backyards leads to a high proportion of green space in residential settlements. This can be used to establish a contextual measure whether a house lies in a residential neighbourhood or not. `presenceOf(yards)` was therefore mapped to a kernel density measure as it was developed by Chaudhry and Mackaness (2008). Yard density at any location $k$ is given by:

$$
yd_k = \frac{\sum_{i=1}^{n} \sqrt{a_i}}{d_{ki}}
$$

where $a$ is the area of yard $i$, $d_{ki}$ the distance between location $k$ and yard $i$, and $n$ is the number of yards involved in the calculation of density.

It was also mentioned above that some concepts were left abstract because it is inefficient or impossible to define them by relations alone. These involve custom-built algorithms for their instantiation. For the terraced houses ontology, this had to be done for row of houses and areas of parallel rows. The algorithms are discussed in full detail in Lüscher et al. (2008) and are only briefly sketched here. Perceptual alignments were obtained by grouping buildings sharing a common wall and then connecting the centroids of the buildings to a path. The path was broken up at sharp turns, i.e. where the angle between two consecutive segments was larger than 60°. Remaining groups were finally qualified for homogeneity and straightness. The concept areas of parallel rows was derived by identifying the main axes of building groups, clustering these groups using the direction of the axes, and finally qualifying clusters for their homogeneity.

5 Bayesian inference as a technique to derive instances of concepts

5.1 Bayesian inference as a means to integrate probabilistic and crisp decisions

Bayesian inference is a standard approach in pattern classification (Duda, Hart, & Stork, 2001; Rice, 1988; Russel & Norvig, 2003). Assume that we have a categorical variable $C$ that is statistically dependent on a set of evidence variables $F_1, ..., F_n$. For instance, $C$ could be a binary variable that describes the fact whether a building constitutes a terraced house or not, depending on whether it is contained in a homogeneous alignment of houses, the presence of yards, etc.

The Bayesian decision rule tries to minimize the probability of error in a decision by deciding for the most probable outcome. Consider Figure 3, which shows a hypothetical likelihood curve for a building to be a terraced house, if the decision was based exclusively on its area. Let’s assume building $i$ having area $35$ m$^2$ has to be classified. The likelihood of being ‘terraced’ as indicated in the figure is 0.6, while the
likelihood of being ‘not terraced’ is only 0.4. Therefore we decide building \(i\) is ‘terraced’.

Formally, the Bayesian decision rule states that the predicted class \(\hat{C}\) for a given realisation \(F_i = f_1, \ldots, F_n = f_n\) is the class \(c\) which maximises the likelihood \(P(c|f)\). This is mathematically expressed using the operator \(\arg \max\):

\[
\hat{C} = \arg \max_c P(C = c \mid F_i = f_1 \land \ldots \land F_n = f_n)
\]

(2)

\[P(terraced\mid area)\]

\[\text{terraced} \]  
\[\text{terraced} \]

Fig. 3. Hypothetical likelihood curve for terraced house, given the building area.

Any inference can be translated into a conditional probability, including crisp relations with Boolean outcomes, as it happens when an \(is-a\) relation is turned into a Bayesian decision. The likelihood distribution is trivial in these cases, as shown in Table 1.

<table>
<thead>
<tr>
<th>(P(\text{house}))</th>
<th>(is-a(\text{building}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>true</td>
</tr>
<tr>
<td>0.0</td>
<td>false</td>
</tr>
</tbody>
</table>

Table 1. Probability distribution for an \(is-a\) relation.

In the general case, if there are more evidence variables involved than just one, the evidence variables are usually not independent of each other. That is, a joint likelihood distribution has to be created upon which the Bayesian decision is based.

5.2 Chaining Bayesian decisions

The inference process starts with the concepts that can be derived using only concepts that are already in the database, and proceeds incrementally to derived concepts of higher order. In this manner the inference task is translated into a chain of
Bayesian decisions, creating a so-called Bayesian network. Probabilistic inference in Bayesian networks is theoretically well explored (Russel & Norvig, 2003). Consequently, the ontology is turned into a Bayesian network by specifying joint conditional probability distributions for each concept. This can be trivial as in the case of the is-a relation. When fuzzy relations are involved such as in the example of building area, it is easier to learn probability distributions from training samples instead of specifying them manually. In the following section we will show how this can be achieved.

5.3 Learning Bayesian decisions from training data

If the likelihood is to be learned from training data, Equation 2 can be transformed to a more convenient form. The transformation makes use of the Bayes’ theorem:

\[
P(C = c \mid F_1 = f_1 \land \ldots \land F_n = f_n) = \frac{P(F_1 = f_1 \land \ldots \land F_n = f_n \mid C = c)P(C = c)}{P(F_1 = f_1 \land \ldots \land F_n = f_n)}
\]

(3)

The denominator on the right hand side is a scaling factor that guarantees that probabilities sum to one. Recalling the Bayesian decision rule and Equation 2, we are only interested in for which value of \( c \) the term on the right hand side reaches its maximum. The denominator is independent of \( c \) and can therefore be omitted, leading to the following formulation of the Bayesian decision:

\[
\hat{C} = \arg \max_c P(F_1 = f_1 \land \ldots \land F_n = f_n \mid C = c)P(C = c)
\]

(4)

In Equation 4, likelihood has been replaced by the class-conditional joint probability density function.

The advantage of Equation 4 is that density distributions can be estimated using training data. A convenient method to estimate them is to employ kernel density estimation (Silverman, 1986). One can guarantee that the probabilities sum to one if a standard normal distribution function is chosen as kernel. Let \( \tilde{f} = (f_1, \ldots, f_n) \), where \( \tilde{f}_i \) are the training samples with classification \( C = c \). The joint conditional density distribution \( P_c \) is then given by:

\[
P_c(\tilde{f} \mid C = c) = \frac{1}{N} \sum_{i=1}^{N} K(\frac{\tilde{f} - \tilde{f}_i}{\tilde{h}}), \text{ where } K(x) = \frac{1}{(2\pi)^{N/2}} e^{-0.5x^2}
\]

(5)

\( N \) is the number of samples and \( \tilde{h} \) are the bandwidths, which constitute smoothing factors for the density function.

Figure 4 illustrates the calculation of the probability density function. The crosses below the x-axis indicate building area values of terraced houses that were tagged in the training data. Dashed lines are kernels for each sample. The solid line indicates the estimated density function, which is the sum of individual Gaussian curves.
6 Experiments

Two experiments were carried out: The first experiment is based on a definition of terraced house by the Ordnance Survey (Ordnance Survey Ontologies, 2008). This experiment was designed to provide a reference of how much prediction accuracy could be achieved by compact definitions and crisp logic reasoning alone, and where typical problems would arise. In the second experiment, test areas were classified according to the Bayesian inference approach presented above. In Sections 7 and 8, we evaluate classifications by means of their prediction accuracy (compared to human interpretation) and show some typical errors for both experiments.

6.1 Test data

Four urban areas of England were extracted from the Ordnance Survey MasterMap® Topography Layer for the cities of Middlesbrough, Norwich, Portsmouth, and Southampton. The OS MasterMap® Topography Layer models topographic features in urban areas corresponding to a scale of approximately 1:1 250. The extents of the test datasets were chosen such that they include not only residential areas, but a wide variety of urban land-use, i.e. mixed residential with smaller commercial buildings, and large industrial/commercial grounds. Besides traditional, Victorian and Georgian-type residential areas, also more recent settlements were found in the areas. They differ from the traditional type in a less regular arrangement, and a mix of terraced and semi-detached housing types in one block (Marshall, 2005). In the experiments no distinction was made between the two types of settlement periods.

The authors manually attributed buildings in all datasets with ‘terraced’/‘not terraced’ by visual inspection. Besides MasterMap®, aerial photographs provided by Google Earth were used for the manual classification. Table 2 shows some characteristics of the study areas.
Study area | Area covered (East-West / North-South) | # buildings | # terraced houses in manual classification
--- | --- | --- | ---
Middlesbrough | 3.33 km / 3.26 km | 41 667 | 14 138 (33.9%) |
Norwich | 5.63 km / 4.75km | 62 021 | 20 297 (32.7%) |
Portsmouth | 5.80 km / 6.55 km | 80 853 | 37 862 (46.8%) |
Southampton | 3.85 km / 2.80 km | 22 950 | 5 075 (22.1%) |

Table 2. Characteristics of the study areas.

The data enrichment process starts from concepts that are readily available in the database. For instance, for building, an attribute in OS MasterMap® encodes whether a polygon represents open land, transportation or a building. Likewise, instances of uncultivated area can be defined through a combination of two attributes. Therefore, relations were added to the ontologies that define buildings and uncultivated areas as presented in Table 3 (in SWRL Human Readable Syntax).

| ArealPrimitive(?x) ∧ hasAttribute(?x, ?y) ∧ hasName(?y, “DescGroup”) ∧ hasValue(?y, “Building”) ⇒ Building(?x) |
| ArealPrimitive(?x) ∧ hasAttribute(?x, ?y) ∧ hasName(?y, “DescGroup”) ∧ hasValue(?y, “Building”) ∧ hasAttribute(?x, ?z) ∧ hasName(?z, “make”) ∧ hasValue(?y, “multiple”) ⇒ UncultivatedArea(?x) |

Table 3. Rules for asserting building and uncultivated area from OS MasterMap®.

6.2 Experiment based on simple ontology

This experiment was carried out to reveal insights to which extent classification is possible based on very basic spatial operations and crisp inference.

The Ordnance Survey GeoSemantics team provides ontologies of their spatial databases (Ordnance Survey Ontologies, 2008). The aim is to describe the content of OS databases concisely to improve usability and data integration. The first classification experiment was based on the description of terraced house provided in the ‘OS ontology for Buildings and Places’. The natural language description is as follows: “A terrace house is one that is part of a line of connected houses” (Ordnance Survey Ontologies, 2008). The Ordnance Survey GeoSemantics team provides equivalent definitions in Rabbit, a controlled language for authoring ontologies (Hart et al., 2008), and OWL. Table 4 shows the Rabbit definition to ease reading.
House

Every House is a kind of Building. Every House has purpose Housing of People.

End Terrace House

An End Terrace House is anything that:

- is a kind of House;
- is connected to exactly one Terrace House.

Terrace House

A Terrace House is anything that:

- is a kind of House;
- is connected to exactly 2 Terrace Houses; or is connected to exactly one End Terrace House and is connected to exactly one Terrace House.

Table 4. Rabbit definition of terraced house as provided by the Ordnance Survey.

The definition differentiates between houses at the end of a terrace (End Terrace House) and houses within a terrace (Terrace House). We will denote the latter type Mid Terrace House to make a clear distinction. The definition is based on only two types of relations: the functional definition hasPurpose(Housing), and the topological relation isConnectedTo().

In order to carry out the reasoning, the original Ordnance Survey definition was modified in two points. Firstly, there was no information available in OS MasterMap® whether a building serves for dwelling. One possibility would be to integrate data that provide missing information (e.g. from zoning maps or a building register). However, this option was not pursued in this study, as the focus was on pattern recognition from a single topographic database. The hasPurpose(Housing) relation was therefore replaced by a restriction on the area of the building footprint as an approximation. The cut-off values were determined experimentally.

Secondly, the rule for Mid Terrace House contains a reference to itself, which makes reasoning unfeasible by the forward-chaining reasoning mechanism that was employed in the experiment. The rule was therefore simplified to “is connected to exactly 2 Houses”. The modified rules used for classification are given in Table 5.

<table>
<thead>
<tr>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building(?x) ∧ hasArea(?x, ?a) ∧ swrlb:greaterThanOrEqual(?a, 35) ∧ swrlb:smallerThanOrEqual(?a, 160) ⇒ House(?x)</td>
</tr>
<tr>
<td>Building(?x) ∧ isConnectedTo(?x, ?y) ∧ House(?y) ⇒ isConnectedToHouse(?x, ?y)</td>
</tr>
<tr>
<td>House(?x) ∧ (=2 isConnectedToHouse)(?x) ∧ ⇒ MidTerracedHouse(?x)</td>
</tr>
<tr>
<td>Building(?x) ∧ isConnectedTo(?x, ?y) ∧ MidTerracedHouse(?y) ⇒</td>
</tr>
<tr>
<td>isConnectedToMidTerracedHouse(?x, ?y)</td>
</tr>
<tr>
<td>House(?x) ∧ (=1 isConnectedToMidTerracedHouse)(?x) ⇒ EndTerracedHouse(?x)</td>
</tr>
</tbody>
</table>

Table 5. Rules for classifying terraced houses used in the first experiment.
The Jena general purpose reasoning engine (Jena, 2009) was employed to carry out reasoning. Data exchange between the spatial database and Jena happened through OWL as exchange format. For each building, a Java program calculated area and topological connectedness to other buildings, and added this information as OWL properties. As an example, Table 6 shows an OWL extract for one building.

```
<Building rdf:about="http://www.geo.uzh.ch/orus#osgb1000002054799448">
  <hasArea>55.35714999958873</hasArea>
  <isConnectedTo rdf:resource="http://www.geo.uzh.ch/orus#osgb1000002054799449"/>
  <isConnectedTo rdf:resource="http://www.geo.uzh.ch/orus#osgb1000002054949238"/>
  <isConnectedTo rdf:resource="http://www.geo.uzh.ch/orus#osgb1000002054949232"/>
  <isConnectedTo rdf:resource="http://www.geo.uzh.ch/orus#osgb1000002054799458"/>
</Building>
```

Table 6. Exemplary OWL code for a building.

The reasoner thereon classified terraced houses according to the rules presented in Table 5. The classifications were transferred back into the GIS for controlling the results.

### 6.3 Experiment based on Bayesian approach

This experiment used the ontology as presented in Figure 2. The Ordnance Survey MasterMap® datasets used in this study did not have an attribute for the number of floors of buildings. As mentioned previously, there was also no information about building function available. Therefore house and hasHeight(2 floors) were dropped and the relations pointing to house were short cut to building. The authors assessed that the remaining criteria hasArea(small) and presenceOf(yards) provide in most cases enough discriminatory power for classification.

As in the previous experiment the ontology was stated as a set of rules, but inference was carried out using a Bayesian reasoner, which was implemented as a custom-built prototype for ontology-driven database enrichment in Java. Whenever the reasoner has to check if a database object is an instance of a concept, it calls analysis routines for each predicate in the definition of the concept. The routines implement necessary (spatial) analysis functions. Fuzzy predicates are allowed to return any number. For instance, hasArea(small) returns the footprint of the database object (e.g. building); presenceOf(yards) returns the density of yards at the location of the object. The obtained values constitute the evidence variables for Bayesian inference. The Bayesian inference uses training data for estimating a joint probability density distribution as explained in Section 5.3. For each concept definition having fuzzy predicates, a set of positive and negative examples must therefore be given.

For the terraced house concept, all 5 075 buildings of the Southampton area tagged as terraced house were selected as positive samples. A characteristic set of 6 629 buildings from the Southampton area was selected to form samples of non-terraced houses. Figure 5 shows the marginal probability distributions derived from these sample
data. Grey shaded areas denote the acceptance of terraced house, if the decision was based on one criterion only. The is-a(Building) predicate in the definition of terraced house is crisp and has a probability distribution as shown in Table 1.

\[ P(\text{terraced}|\text{area}) \]

\[ P(\text{terraced}|\text{presenceOf(yard)}) \]

\[ P(\text{terraced}|\text{partOf(row of houses)}) \]

\[ P(\text{terraced}|\text{partOf(area of par. rows)}) \]

**Fig. 5.** Marginal probability distributions for uncertain relations of the terraced house concept. Shaded grey areas: Regions with \( P(\text{terraced}) > 0.5 \).

### 7 Evaluation of classification accuracy

In the following, the classification accuracy of the conducted experiments is measured statistically by comparison to human interpretation. Classification accuracy was measured by means of precision, recall, and Cohen’s kappa coefficient \( \kappa \). Precision indicates the probability that a terraced house found by a classification algorithm was also classified as terraced house in manual classification. Recall indicates the probability that a manually classified terraced house is found by the classification algorithm. Cohen’s kappa (Lillesand et al., 2000) is a measure of agreement between classifications; \(-1 \leq \kappa \leq 1\), whereby high values of \( \kappa \) denote a good agreement.

Table 7 presents results produced by the simple ontology approach. The Portsmouth area is classified very well, while results of the other three study areas produce a lower kappa value. This is explained by the fact that the Portsmouth area is a ‘standard’ situation in the sense that highly regular terraced houses dominate. Pure residential
areas were classified generally well, while accuracy in mixed-use and industrial areas was lower.

<table>
<thead>
<tr>
<th>Area</th>
<th># buildings</th>
<th># correct class</th>
<th>% correct class</th>
<th>precision</th>
<th>recall</th>
<th>Cohen's kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southampton</td>
<td>22 950</td>
<td>17 223</td>
<td>75.0%</td>
<td>77.7%</td>
<td>86.4%</td>
<td>0.76</td>
</tr>
<tr>
<td>Middlesbrough</td>
<td>41 667</td>
<td>37 919</td>
<td>91.0%</td>
<td>94.8%</td>
<td>77.8%</td>
<td>0.79</td>
</tr>
<tr>
<td>Norwich</td>
<td>62 021</td>
<td>55 306</td>
<td>89.2%</td>
<td>88.8%</td>
<td>76.2%</td>
<td>0.74</td>
</tr>
<tr>
<td>Portsmouth</td>
<td>80 853</td>
<td>75 559</td>
<td>93.5%</td>
<td>92.5%</td>
<td>93.6%</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 7. Comparison of classification produced by the simple ontology approach and human interpretation.

Table 8 presents the classification accuracies for the experiment using Bayesian inference. It shows that high classification accuracy could be achieved in all four study areas.

<table>
<thead>
<tr>
<th>Area</th>
<th># buildings</th>
<th># correct class</th>
<th>% correct class</th>
<th>precision</th>
<th>recall</th>
<th>Cohen's kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southampton</td>
<td>22 950</td>
<td>22 283</td>
<td>97.1%</td>
<td>90.7%</td>
<td>96.8%</td>
<td>0.92</td>
</tr>
<tr>
<td>Middlesbrough</td>
<td>41 667</td>
<td>40 899</td>
<td>98.2%</td>
<td>98.3%</td>
<td>96.3%</td>
<td>0.96</td>
</tr>
<tr>
<td>Norwich</td>
<td>62 021</td>
<td>59 388</td>
<td>95.8%</td>
<td>93.1%</td>
<td>94.0%</td>
<td>0.90</td>
</tr>
<tr>
<td>Portsmouth</td>
<td>80 853</td>
<td>76 942</td>
<td>95.2%</td>
<td>92.8%</td>
<td>97.3%</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 8. Comparison of classifications produced by Bayesian inference and human interpretation.

*Terraced houses of the Southampton area were part of the training sample.

Figure 6 shows a traditional terraced house neighbourhood as classified by the Bayesian inference approach. Figure 7 depicts a situation in a more ‘modern’ type of settlement, having lower building density and less stringent regularity of the arrangement of rows.
Fig. 6. Traditional terraced house neighbourhood (Middlesbrough). Please note that there are no false positives in this area. OS MasterMap data Ordnance Survey ©Crown Copyright. All rights reserved.

Fig. 7. Modern terraced neighbourhood (Norwich). Colouring as in Figure 6. OS MasterMap data Ordnance Survey ©Crown Copyright. All rights reserved.
8 Discussion

In the following, the Bayesian approach is assessed by comparison to the more traditional simple ontology approach and by making considerations on scalability. The benefits are clarified by means of relating the approach to the case study. Finally, we conclude with perspectives for future research.

8.1 Comparison of common errors produced by the approaches

In the following, we contrast both approaches by discussing common sources of disagreement between the human interpretation and automatic classification as produced by each approach.

8.1.1. Common errors produced by the simple ontology approach

Errors produced by the simple ontology approach can be grouped into two classes.

*Missing linear arrangement:* Figure 8a shows a case where porch roofs classified as house prevent correct classification of a terraced house. The house indicated as ‘MT’ was classified as mid-terraced because it connects to exactly two other houses (one of them being the incorrectly classified porch roof). Buildings indicated as ‘ET’ were classified as end-terraced, because they connect to the mid-terraced house. Most of the terraced houses were not found; they connect to more than two other houses (including the porch roofs). Figure 8b depicts a situation where terraced houses were produced in a heterogeneous, dense built-up block. Even if the buildings in the situation constitute dwellings, the situation would not be perceived as row of terraced houses, but as an assembly of houses randomly built together. The errors in both situations occur because topology alone does not capture the fact of being ‘a line of houses’. A synoptic view is needed to decide on what constitutes an alignment and which houses are parts thereof. In the Bayesian inference experiment this synoptic view is provided by the row of houses concept.

*Special cases in the modelling of features:* Figure 8c shows a situation where small, oblong polygons disconnect otherwise perfectly regular terraces. The polygons are actually small enclosed alleys that connect the street to the backyards. Although they are integrated into the houses (e.g., the first floor above the alley is made up of a room), they are modeled as separate polygons in OS MasterMap®. As a consequence, houses are not topologically connected and are not detected as terraced houses. In the Bayesian inference experiment, the row of houses concept again provides grounds for correct classification. Here, further experiments are needed to establish whether such special cases can be modelled in SWRL rules. However, this would in turn render the description less compact.
Concluding, the simple ontology approach produced reasonable results where situations corresponded to the prototypical conceptualisation. In less clear situations, a synoptic view is missing that cannot be constructed using logic reasoning alone.

8.1.2. Common errors produced by the Bayesian inference approach

A type of false positive produced by the Bayesian inference approach is shown in Figure 9a. There are rows of garages or sheds in the backyards having an area of around 30 m². These were classified as terraced houses by the Bayesian inference approach. The simple ontology experiment, applying a higher area threshold of 35 m², did not reproduce this behaviour, but missed terraced houses in the leftmost vertical row that have an area below 35 m². Obviously, this issue of features with overlapping values cannot be solved without adding more criteria (e.g., detecting backyard sheds in advance).

A second but infrequent type of false positive is shown in Figure 9b. It shows rows of semi-detached houses that are connected with each other through small constructions such as shelter roofs at the entrance. The automatic classification treats them as terraced houses, although they rather correspond to semi-detached houses because effectively
they have three exterior walls that can provide more daylight to inhabitants, whereas terraced houses only have two walls (excluding houses at the end of terraces). In this specific case, the simple ontology approach did not show this misclassification due to the modelling discussed already discussed in Figure 8c (hence also the many false negatives).

**False negatives** were less frequent than false positives. Typically they occurred at boundaries of residential areas, along large streets, and in isolated terraces, where presenceOf(yard) was generally lower.

![Fig. 9. Typical false positives in Bayesian inference. Colouring as in Figure 6. OS MasterMap data Ordnance Survey ©Crown Copyright. All rights reserved.](image)

A general source of disagreement arose in some cases when rows of buildings were not discernable from terraced houses in MasterMap® alone, but from information that was only visible in aerial photographs, such as facades and patterns of access paths and entrances. In other cases, the human operator judged buildings to have a different function than dwelling based on the spatial context visible in the aerial photograph.

### 8.2 Scalability considerations

The expenditure of time is more dependent on the low level algorithms involved than on the inference process itself, and therefore highly dependent on the actual ontology; we therefore constrain to the argument that the approach is practical. The inference of terraced houses took approximately 8 min for the Norwich study area (134 524 database objects) on a 2.66GHz Xeon processor (single task). We therefore argue that the approach is practical, considering that it will typically be run as an off-line process for semantic enrichment of spatial databases rather than in real-time.

The necessity of defining training samples when joint probability distributions cannot be provided by a human operator can be seen as advantage and drawback at the
same time. On the one hand, thresholds or membership functions such as when applying fuzzy set theory (Fisher et al., 2004; Ladner et al., 2003) do not have to be specified, but can be estimated from the training data. This is beneficial when knowledge about the domain is incomplete; for instance, clues in the literature about what ‘small areas’ means for terraced houses are rather vague. The downside is the effort that goes into the selection and tagging of training samples. When using kernel density for estimating probability distributions, density estimation effectively takes place in an \( n \)-dimensional feature space, which is created by the relations to sub-concepts. The more sub-concepts there are for a concept, the more training samples have to be defined to make sure that there are enough characteristic samples in each region of the feature space. This problem is known as the \textit{curse of dimensionality} (Duda et al., 2001).

We also would like to comment on error propagation in the inference process. A concept definition usually relates to other concepts, whose instances are either asserted in the database, or have to be derived first. Poor accuracy in the derivation of related concepts leads to potential errors in the derivation of the composed concept. Since related concepts are derived independently, they should be checked for plausibility before continuing with inferring higher level instances. Therefore, the recognition process has to be supervised and is not fully automatic.

\subsection*{8.3 Benefits of the ontology-driven approach}

The main benefits of this ontology-driven approach can be summarised as follows:

\textbf{Enhanced transparency} is provided since assumptions about the spatial structure of the geographic concepts are explicitly stated. Ontologies can be modelled and validated in collaboration with domain experts (making sure they are consistent with the experts’ conceptualisation of reality), and different conceptualisations of the same terms can be compared, for example to reveal culturally different conceptions.

\textbf{Enhanced flexibility} is provided by being able to align the mapping of ontologies for different databases, or modify parts of an ontology to accommodate locally different settings.

\textbf{Enhanced reusability} is provided since it is a component-oriented approach that allows those parts that have to be implemented in spatial algorithms to be re-used in the derivation of different concepts. For this to happen, basic algorithmic components that provide spatial measures have to be identified and published. They serve as vocabulary that can be used for constructing ontologies. For instance, \texttt{presenceOf} was mapped to a density estimation, which constitutes an algorithmic component. The same component can be re-used to define a variety of patterns, such as the extents of urban areas and woods (Chaudhry & Mackaness, 2008; Mackaness, Pericleous, & Chaudhry, 2008). The concept \texttt{row of houses} can be re-used to define semi-detached houses (containing exactly two instances of \texttt{house} instead of at least three) or so-called perimeter block developments, which are an arrangement of rows of houses along the roads of a roughly square block. The concept \texttt{terraced house} can itself be used to derive even higher level concepts, such as \texttt{residential area}.  

\hfill – 21 –
9 Conclusions and future research

Ontologies of the geographical reality are important because they provide a basis for abstraction of cartographically relevant patterns over large scale changes and for different usages. Hence the automated semantic annotation of spatial databases is a key success factor in support of automated map generalisation. In this paper, a framework for ontology-driven pattern recognition was presented. First, knowledge about the spatial structure of urban concepts is collected in an ontology. Then, the ontology is concretised by mapping it to measurable units. Finally, inference is carried out using Bayesian decision theory, whereas machine learning techniques can be used to learn concept characteristics from examples.

Besides clarifying the benefits of using ontologies in spatial database enrichment, our research has shown that Bayesian networks are a suitable method to integrate vague knowledge about conceptualisations in cartography and GIScience. We have also shown that logic reasoning techniques should best be combined with a set of general algorithmic components in order to achieve satisfying results.

Our future work will focus on the implementation of more concepts (e.g., other residential house types such as semi-detached and detached houses; on residential areas as an aggregation of residential house types) and a further formalisation of the pattern recognition vocabulary; on the evaluation of the choices of algorithms for basic concepts and their influence on extraction results; and on human subject experiments to study where and how people visually detect concepts such as terraces.

Acknowledgments

The research reported in this paper is part of the PhD project of the first author. Funding by the Swiss State Secretariat for Education and Research (SER) through COST Action C21 (project ORUS, Grant no. C05.0081) is gratefully acknowledged. The authors are grateful to the British Ordnance Survey for provision of MasterMap® data.

References


Ruas, & C. Gold (Eds.), *Headway in Spatial Data Handling. Lecture Notes in Geoinformation and Cartography* (pp. 449–466). Berlin / Heidelberg: Springer.


