J Geogr Syst (2015) 17:137–156
DOI 10.1007/s10109-015-0209-3

ORIGINAL ARTICLE

# Finding similar places using the observation-to-generalization place model

Benjamin Adams[1]

Received: 10 April 2014 / Accepted: 2 March 2015 / Published online: 17 March 2015

**Abstract** In this article, a novel observation-to-generalization place model is proposed. It is shown how this model can be used to formally define the problem of finding geographically similar places. The observation-to-generalization model differentiates between observations of phenomena in the environment at a specific location and time, and generalizations about places that are inferred from these observations. A suite of operations is defined to find similar places based on the invariance of generalized place properties, and it is demonstrated how these functions can be applied to the problem of finding similar places based on the topics that people write about in place descriptions. One use for similar-place search is for exploratory research that will enable investigators to perform case–control studies on place data.

**Keywords** Place · Gazetteer · Similarity · Geographic information retrieval

**JEL Classification** C19 · C65

## 1 Introduction

In the first chapter of his book, *Changes in the Land*, William Cronon writes that as an ecological historian he is "always faced with the problem of generalizing from a *local* description to a *regional* landscape..." (Cronon 2003, p. 6). This process of generalizing a property of a place from individual observations is a common one found in nearly every discipline where there is need to operationalize information about places or regions (e.g., ecology, climate science, and geopolitics). However,

✉ Benjamin Adams
b.adams@auckland.ac.nz

[1] Centre for eResearch, The University of Auckland, Auckland, New Zealand

place knowledge bases, such as digital gazetteers, rarely if ever explicitly model the distinction between individual observations of the environment and the generalizations about places that are derived from these observations. A *generalization* is here defined as *a property of a place, represented in a gazetteer or geographic information system, which has been generated from a set of individual environmental observations via an inferential mechanism*. An example of a generalization is an aggregate attribute, such as a geodemographic or an environmental quality index, associated with an administrative region that is represented geometrically by a polygon in a geographic information system (Montero et al. 2010; Petersen et al. 2011).

In this article, a new observation-to-place model is proposed to organize geographic information in a manner that records the relationship between observations and generalized properties of places. The proposed model serves as a foundation to formally define the problem of similar-place search. Two places are similar if they are invariant (or similar) with respect to a set of properties. An example in the climate domain relates to two temperate places with wet winters and dry summers. Often geographic categories are defined as sets of similar places, e.g., in the above example, those two places are instances of class C places in Köppen climate classification system (Peel et al. 2007). Geographic analogs are not restricted to the physical sciences, however. For example, social scientists commonly differentiate between countries with ‘developed’ and ‘developing’ economies based on various measures. Likewise, laypeople will find places that are analogous based on factors.

The search for similar places is an underdeveloped issue in geographic information science. Recently, there has been progress in developing some systems that search for similar places for specific applications (see, e.g., JournalMap).[1] However, identification of similar places from an information science perspective is a nascent field. One step in this direction is made by a recent dissertation on the development of a geographic analog engine where “place-analog search [is] regarded as an application of entity similarity measurement” (Banchuen 2008). Although systems such as JournalMap and the geographic analog engine attend to the idea of finding similar places for a given task, there is little that is systematic in how they approach the problem of figuring out which properties are relevant for comparing places. An inherent problem is the conflation of spatially referenced *observation* data and *place-based* properties—a problem that also harkens back to the traditional space and place dichotomy. A proposal is made here to more clearly model the distinction between observations of phenomena and place properties that are generalized from these observations. As a demonstration of this model, a suite of similar-place reasoning functions is presented, and it is shown how these operations can be used to find similar places based on written descriptions.

Organizing geographic information in a tiered ontology that separates geographic data that are observations from generalized objects is a key requirement for building geographic information systems that interoperate with human users (Frank 2001; Couclelis 2010). Clearly, delineating the difference between observation-based data and generalized place-based data is important not just as an ontological question. It is important because the conflation of these two types of data is easily done when

[1] http://journalmap.org/.

geographic data (e.g., on the Linked Open Data web) about places and sensor observations are treated as the same thing. This is a relevant problem today as we get more and more location-based sensor data while also wanting to find patterns and find common causes for effects seen in geographic regions. Equating location data with place data, without describing the inference mechanism that leads from one to the other, leads to erroneous claims about the equalities between places. This is not a new concept in spatial analysis—the modifiable areal unit problem and other ecological fallacies that can arise from spatially aggregated data are well studied (Selvin 1958; Fotheringham and Wong 1991).

Thus, this model is first and foremost a practical guide for *organizing* data to clarify what types of inferences we can make based on the comparison of place properties. Furthermore, we make the claim that places can only be thought of as similar in as much as they are invariant based on generalized properties. Otherwise, we are led to the idiographic perspective that all places are unique. We do not claim that it will not be useful in some cases to compare two locations based on single observation values, but seeking evidence for hypotheses about why places are the way they are rarely benefit from such comparisons.

The rest of this paper is organized as follows. The next section provides background material on place and observational models. Section 3 introduces the observation-to-generalization place model. Section 4 defines similar-place search operation templates modeled with the observation-to-generalization model, and the use of similar-place search for case–control studies is discussed. In Sect. 5, the application of these functions is demonstrated with a use case, and finally, we conclude with discussing future work.

## 2 Background

In this section, an overview of previous work on the representation of places in computer systems is provided.

### 2.1 Computing place

Computational models of place tend to be simple when compared to the ways in which place has been conceptualized in geography, environmental psychology, and related disciplines. Part of the reason is simply that places are not crisp canonically defined entities, and instead, they are vague with ill-defined spatial footprints (Montello et al. 2003). Perhaps more important, however, is that place is an experience-based, dynamic construct that is socially mediated, and therefore, it is highly contextual (Tuan 1977; Relph 1976; Cresswell 2004).

A recent special issue of the journal *Spatial Cognition and Computation* was dedicated to the problem of modeling place in computational systems. In the introductory editorial of the issue, the editors state:

> Modeling place involves, among other issues, finding computational models to capture and express the meaning of a place name.... Research on computational place modeling will have a substantial impact on several application areas, such

as spatial recommender systems, urban planning, marketing, and on information retrieval in general (Winter et al. 2009, pp. 1–2).

The prototypical example of a place model in information systems is the digital gazetteer, a dictionary of named places formally defined as a set of place name, feature type, and geographic footprint triples (Hill 2006). Each triple defines a relation that maps a place name to a feature type and geographic footprint. Additional relations can be defined between place names and other attributes (e.g., a population field). Although research on digital gazetteers began as an academic endeavor (e.g., the Alexandria Digital Library), a plethora of commercial, governmental, and open-source digital gazetteer variants have been developed, such as Geographic Names Information System (GNIS) from the United States Geological Survey, OpenStreetMap, Geonames.org, Google Maps, and Bing Maps (Hill 2006). These gazetteers are populated from authoritative data sources as well as, more recently, volunteered information crowdsourced from a large group of users. Keßler et al. (2009) have proposed a new generation of digital gazetteer that takes the next step by combining formal semantics of geographic types using description logics with user-generated content from non-authoritative sources.

Much of the work on modeling place in gazetteers has focused on the appropriate way to represent the spatial profile of a place. This is a natural outcome of the existing geographic information system emphasis on vector/raster representations (Couclelis 1992). Representing the spatial footprint of places as crisp points, polylines, and polygons has the advantage that well-defined and efficient spatial operations can be used to query the gazetteer. Raster (and other non-vector) data formats can be used to represent field-based attribute data and uncertain/fuzzy regions in a gazetteer, although these data are used less often in online gazetteers due to the added computational overhead (Goodchild et al. 1998).

While gazetteers are traditionally built from structured knowledge about places, knowledge can also be discovered from unstructured data. This semantic enrichment of place knowledge can be done by inferring knowledge about places from web pages, social media, and mobile systems that refer to them (Alves et al. 2009). Attributes of locations can be identified from examining photos tagged to a proximate location (Leung and Newsam 2010). In line with research on place identity and the role of social relations in the formation of place, one can use information about the people (e.g., their user profiles in online social networks) who visit a place to develop a model of that place (Graham and Gosling 2011). Location context is often used as a proxy for place to personalize results in information retrieval and recommender systems, and this location information is easily acquired on mobile devices.

## 2.2 Observation and measurement models

To aid semantic enablement of geospatial services and sensor interoperability, a number of formal models for observations and measurements have been developed in recent years. The OGC Observations and Measurements (O & M) ISO standard defines an observation schema when an observation is an action that results in a measurement value of a property of a feature of interest through a procedure.[2] OGC O & M has been formally mapped to the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) upper ontology (Probst 2008; Gangemi et al. 2002). The Science Environment for Ecological Knowledge (SEEK) Extensible Observation Ontology (OBOE) has many commonalities with OGC O & M, but it adds a context relationship that enables modeling how observations are related to one another (Madin et al. 2007). With the advent of the notion of human sensors providing observation data in the form of volunteered geographic information, a stimulus-centric approach to observation modeling has also been developed (Stasch et al. 2009; Goodchild 2007).

## 3 Observation-to-generalization place model

In this section, an abstract model is defined for representing place information that extends the standard gazetteer entry model with an observation-to-generalization model for representing place properties (Hill 2006). A distinction is made between *generalized place attributes*, assigned holistically to a place, and *observed attributes* of the environment, tied to specific observation events located in space. This distinction helps to delineate between properties of places and properties of spatially referenced phenomena, so that geographic knowledge reasoning systems can handle these two types of information in distinct manners. We use this distinction to help define a set of operations for finding geographical place analogs in Sect. 4.

At minimum, a gazetteer entry will have three elements, $<N, t, g>$, where $N$ is the name of the feature (i.e., place name or toponym), $t$ is the type or class of the feature in typology, and $g$ is the spatial footprint, often represented as a point or polygon (Hill 2006). There are several ways in which one can extend this model, including with temporal information and linked data (Keßler et al. 2009). Here an observation-to-generalization approach is proposed for modeling the properties of places. The observation component of this model is similar to the Observations and Measurements ISO standard schema for encoding observations of a feature of interest, but in contrast to that model it differentiates between observations and generalizations (i.e., the interpretations of the observations).

One advantage of this observation-to-generalization model for places is that it sidesteps the objective versus subjective debate on place representation. Because general properties of places can be linked via inferential mechanisms to observations, including those performed by individuals, it is possible to generate heterogenous and even individualized representations of a single place from different sets of observation data. Implicitly, this model makes the claim that the characteristics of places are derived from observations of the environment, which tallies with phenomenological theories of place (Tuan 1977; Cresswell 2004). The model does not have an explicit framing of the subjective nature of observations and generalizations, but does preclude reification statements being asserted over the process used to make the observations or the inference mechanism used for generalization.

[2] http://www.iso.org/iso/catalogue_detail.htm?csnumber=32574.

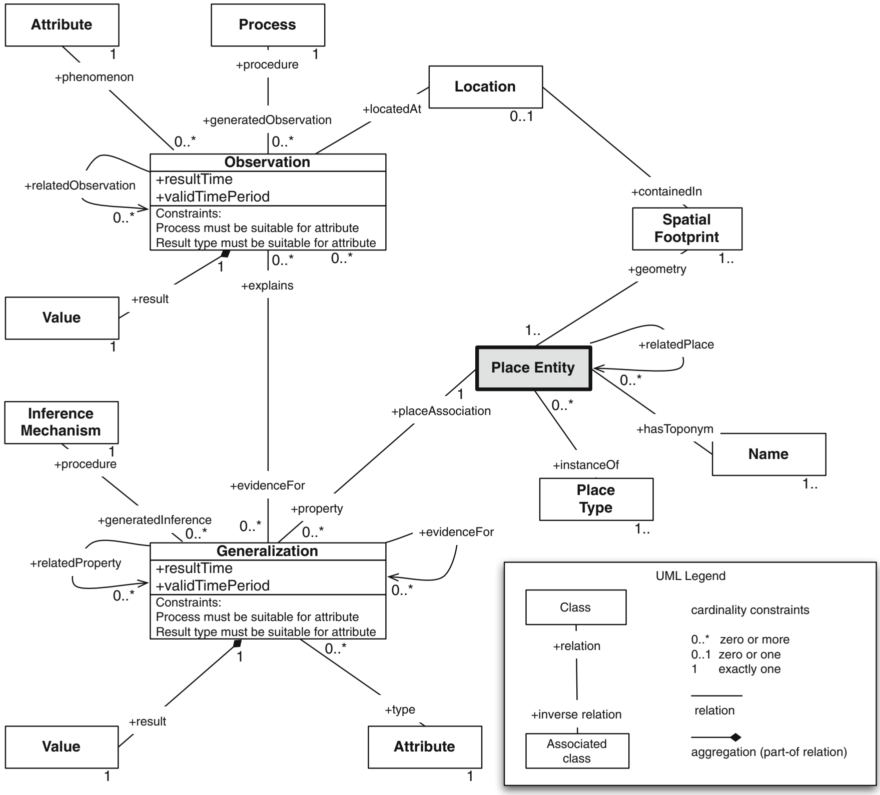

**Fig. 1** UML schematic of the observation-to-generalization place model

## 3.1 Place attributes

In the observation-to-generalization place model shown in Fig. 1, a place entity is defined by a 6-tuple: a set of toponyms, type, a set of spatial footprints, a set of associated observations, a set of generalizations, and a set of relations to other places. The particular choice of interplace relationships, typology of places, and geometric representations of the spatial footprints remains unspecified in the model to allow for different application-dependent solutions. In the latter case, it is indeed possible to represent a "spatial" footprint in terms of networked relations, leading to alternative models of place, such as contrast sets Winter and Freksa (2012).

Observations result in information entities that record properties of environmental phenomena, and generalizations describe properties of a place (Probst 2006).[3] To avoid confusion, in the following text, the term *attribute* refers to the type of property and *property* to the value assigned to the attribute. Thus, *attribute value* and *property* are synonyms. For example, *color* is an *attribute* and *red* is a *property*

[3] The meaning of the term 'generalization' here should not be confused with the meanings used in either cartography or logic.

or *attribute value*. As described in more detail below, the data structure associated with a property need not be restricted to a scalar. That is, the attribute data structure can be complex, e.g., an attribute space with semantically defined quality dimensions. Using the previous example, *color* can be modeled as an attribute space made up of three quality dimensions: *hue*, *value*, and *saturation*. The *hue* attribute describes the direction on the color wheel, *value* characterizes the brightness of the color, and the saturation attribute captures the degree to which the color is pure or more ‘washed out’ (i.e., mixed with white). The color property *red* can be modeled as a vector or region within that attribute space depending on the application need.

### 3.2 Observed property

An observed property is the result of a measurement of the environment by a sensor, whether it be a mechanical sensor such as a temperature gauge or a human sensor who records a written description of a place, at a specific location and time. A liberal reading of the term *measurement* is used here to involve not only observation methods that generate signals in the form of quantitative results but also other kinds of annotations about places as well as complex data structures. While there exist philosophical debates about whether the observation results from “human sensors” should be equated with those from instrument sensors, the main purpose of the observation-to-generalization model is to act as a practical guide for organizing place-based geographic information systems that deal with both space- and place-based data. It is not intended to fully address the problem of semantic interoperability, which may require a more nuanced understanding of the ontology of observation and measurement (Kuhn 2003; Probst 2006; Schade et al. 2012).

Examples of observations include temperature readings, ozone measurements, crime reports, photographs, and georeferenced tweets. As in the observation and measurement standard, multiple observations can be made of the same attribute, and the location of the observation can have a different representation than the spatial footprint of the feature of interest (i.e., the place).

An observation has multiple components: the observation location, time of result, valid time period, the phenomenon being observed, the process used to produce the result, and the result of the observation (i.e., the attribute value). In accordance with the observation and measurement standard, an observation which has a location that extends in one or more dimensions and varies in value along that spatial extent can be modeled as a *coverage* attribute. An example of such a coverage observation is a satellite image. An observation may have zero or more *evidenceFor* associations with place generalizations if it is used as evidence in the generalization inference. An observation can also be related to other observations. As with places, these types of relations remain undefined in the model.

### 3.3 Generalized attribute values

A generalized property is a place property that is assigned to the place as a whole. Generalized attribute values result from some kind of inferential process, such as a statistical inference on a sample, simulation, or algorithm, performed on place observations or other interpreted data. In many cases, this inferential process will be opaque, as in the case when place attributes are incorporated into the knowledge base without any provenance information (e.g., linked data from government sources).

The key distinction between *generalized* and *observed* attributes is that *generalizations* are directly associated only with places, not with locations. These generalized attributes are indirectly associated with locations in two manners: 1) by the relation of the place to its spatial footprint and 2) when relations are defined between a generalized attribute value and the observations (with locations) used to derive it. However, spatial statistical analyses of place generalizations based on these indirect associations to spatial location should be interpreted as more uncertain than similar analyses on direct observation data. Although we do not focus on data quality here, this last point is related to the issue of quality standards and error propagation in GISs (Heuvelink 1998).

Examples of generalized attributes for a place are population counts, temperature seasonality, and median income. A spatial footprint for a place can be a special case of a generalized property. In most cases in gazetteers, the spatial footprints are assigned to places independently of other attributes, but deriving the spatial footprint from observations has been explored in the literature (Montello et al. 2003). Spatial relationships with other places can also be thought of as generalized attributes, but they can also be directly inferred from spatial footprints, so modeling them in this way is most likely an overkill.

A generalization is similar to an observation with the following exceptions. A generalization is associated with one and only one place, unlike observations which can have multiple place associations. Instead of an observation procedure, a generalization is generated by an inference mechanism. Generalizations do not have a location, only a place association.

Although an argument can be made to model all measured attributes as generalized attributes due to implied inference mechanisms built into the use of scientific instruments as well as human cognition, from a pragmatic perspective, it seems reasonable to consider georeferenced data that come directly as output from sensors (including human sensors) as non-generalized data [(a similar distinction is made by Frank (2001) and Couclelis (2010)]. Both observed and generalized attributes can be used for comparing places, but the main utility for making this distinction in the model derives from the claim that when finding similar geographic places, similarity based on keeping generalized attribute values invariant will be of primary interest. For example, finding similar places by comparing individual observations such as temperature readings at a specific times and locations is much less useful than comparing the places based on generalized climactic variables. Furthermore, directly comparing two sets of observation data for similarity, e.g., in point pattern analysis, might be better characterized as a form of spatial analog (based on spatial similarity measurements) rather than place-based geographic analog (Gatrell et al. 1996).

### 3.4 Similarity across multiple properties

There exists a wealth of heterogeneous knowledge about places that is readily available, including structured data about population and climate, physical structure, affordance properties, spatial relations with other places (e.g., *contained inside*), and distance relationships. All of these properties become different dimensions on which the similarity of two places can be judged. A system that enables to explore geographic knowledge should be able to integrate these different kinds of knowledge into a common, flexible framework.

Using the terminology from conceptual space theory, each of these attributes constitutes a separable domain on which places can be compared (Gärdenfors 2000). In some cases, these domains can be defined by a set of quantitative quality dimensions, but in contrast to conceptual space theory no restriction is made here that the structure of these domains *must* be represented geometrically. Rather, the only requirement is that for each attribute domain, the similarity function takes the form $P \times P \rightarrow [0, \ldots, 1]$, where $P$ is the set of all property values in the domain. This restriction that the result is a value in $[0, \ldots, 1]$ ensures that the different similarity measures can be combined (Janowicz et al. 2011). A similarity value of one means maximally similar, and zero means maximally dissimilar. The similarity function needs not be symmetric.

The similarity of two places is defined as a multiparameter weighted measure of the similarities of different attributes. A weighted product measure is defined in Eq. (1).

$$\prod_{i=1}^{n} (\mathrm{Sim}_i)^{\frac{w_i}{n}} \tag{1}$$

where $n$ is the number of attributes being compared, $\mathrm{Sim}_i$ is the similarity result for the $i$th attribute, $w_i$ is the weight on the attribute. The product measure has the property that if one attribute similarity is zero, then the whole product goes to zero. A weighted sum measure is an alternative approach that does not have this property [see Eq. (2)].

$$\sum_{i=1}^{n} (\mathrm{Sim}_i) w_i. \tag{2}$$

The resulting compound similarity values are only commensurable with other similarity values given the same weighting. Therefore, there is no need to normalize the compound value.

The weights provide a means to introduce context into the similarity judgment (Janowicz et al. 2011). For many cases, the weights on the dimensions will be standardized based on a theoretical model with well-defined weights on the variables. For example, such models include socioeconomic indices, such as the Human Development Index and the Ocean Health Index (United Nations Development Programme 1990; Halpern et al. 2012). The weights can also generate personalized similarity results, where the weights are highly context-dependent. For example, the weights can be inferred based on previous knowledge about user preferences (Janowicz et al. 2010).

## 4 Operationalizing place similarity for geographic research

Place information systems that can identify potential target places which are similar to a source would be a valuable resource for many different kinds of studies that aim to make inferences and predictions about places via similarity reasoning. One key aspect of similarity search is that it is context-dependent, and while the target places should be similar with respect to a small set of properties, they will undoubtedly differ with respect to other properties. Furthermore, it might be that those differences can be a useful constraint in the search. For example, someone might be interested in finding climatic analogs to Santa Barbara that differ with respect to a specific ecological variable in order to test a hypothesis.

This example illustrates that scientists can use the idea of similar places to identify a set of "case" places for use in case–control studies. A case–control study is a retrospective study based on existing data. Case–control studies have been used effectively by epidemiologists, most notably in showing the link between smoking and lung cancer (Schulz and Grimes 2002). Figure 2 shows a schematic of how case–control study design can be adopted for research on a population of places rather than human subjects. A common type of research problem involves hypothesizing a cause for a place property (a *consequent property*). Another place property can be hypothesized to be a *precursory property*, i.e., its presence at an earlier time is indicative of causal process leading to the consequent property. One way this hypothesis can be tested is by examining two sets of other places. One set shares the property with the source place and the other set does not. By examining the presence or lack of the presence of the precursory property (the "exposure" in

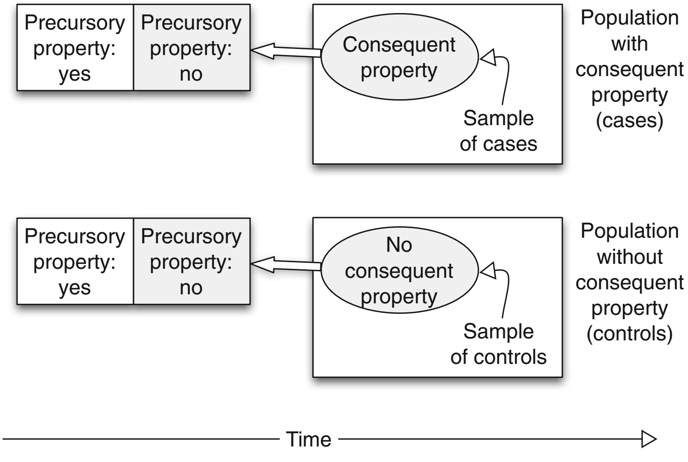

**Fig. 2** Case–control study design for place-based research, derived from Schulz and Grimes (2002)

an epidemiological context) in the case and control places, the hypothesis can be supported or rejected. As in epidemiology, when selecting control places, it is important that “controls should represent the population at risk of becoming cases” (Schulz and Grimes 2002, p. 432). This need for representativeness in the control places can be specified in terms of constraints on the types of places being investigated. For example, we might limit the control places to only *cities* if appropriate for the study. Rather than matching based on type, another method for finding appropriate control places is to find places that are similar with respect to a set of background properties but which differ in terms of the consequent property being investigated.

In geographic research, a case–control study can be particularly efficacious because it is based on available observation data and does not require an investigator to obtain data for a fully randomized controlled study, which can be difficult or even impossible for geographic-scale problems. Use of spatial analytic techniques in epidemiological case–control studies has been successful due to supporting tools and methods (Jacquez 2000; Bivand and Gebhardt 2000), and place-based case–control studies hold promise given appropriate infrastructure. This kind of data-driven scientific research has gained attraction in other domains as well in recent years (Kell and Oliver 2004; Hey et al. 2009).

### 4.1 Finding similar places

In this section, a series of operations is presented for finding geographically similar places. The problem of similar-place search is framed generally, so that it can be applied to many types of place properties. The operations operate on a set of places defined using the observation-to-generalization place model. The following are some examples of the kinds of similar-place searches that this framework aims to facilitate:[4]

- Places similar to Vienna based on Wikipedia text topics,
- The European analog of the Grand Canyon,
- The warm version of Dublin,
- What place best matches New York City with beaches,
- What place best matches Santa Barbara with civil war history.

The general approach for finding similar places is to specify the source place and the set the invariant properties to get a candidate set of target places. The similar-place search problem is defined as follows. Let $X$ be a set of places modeled using the observation-to-generalization place model and let $s \in X$ be a source place. Let $A^I$ be the set of attributes that should have invariant values. The invariant properties comprise a set of attributes, $A^I$, over which the source place and each target place should have generalization results of those attributes that are similar. The template

[4] The search functions are based on place entities in a knowledge base. These examples are illustrative shorthand. To perform natural language searches like these, one would also need to disambiguate places with the same toponym and integrate exonyms.

of a similar-place search operation has $s$ and $A^I$ as parameters. Additional parameters are added for specific sub-types of similar-place search operations. Every operation returns a finite ordered set (in terms of a greater than similarity relation) of target places, $T = \{t_1, t_2, \ldots, t_n\}$, such that $T \subseteq X$.

Following is a set of two general place similar-place search operations. The first function is a weighted similarity measure based on the attributes of the places. For the second function (of which there are two variations), the notion of a *contrast property* is introduced. The idea of a *contrast property* is derived from the idea of the contrast class in conceptual space theory (Gärdenfors 2000). In conceptual spaces, it is proposed that non-monotonic property–concept combinations such as *white wine* can be modeled using geometric operations on quality dimensions as opposed to set theoretic intersection or union operations on classes defined in terms of necessary and sufficient features. The combination *white wine* is non-monotonic in the case that the *wine* category is defined as having a range of colors from reddish to yellow, and therefore, the intersection of all *white* things with all *wine* things is an empty set. Instead, using geometric structures, an instance of *white wine* can be classified as falling within a modified "yellowish" region in the attribute space. The basic approach is to modify the representation of the property (e.g., *white*) by stretching it over the region representing the color property for *wine*. The notion of contrast class was formalized in (Adams and Raubal 2009; Adams and Janowicz 2011). Herein, a *contrast property* takes a similar approach; however, the implementation is more flexible given different kinds of attribute representations.

*SF1: Similar places to source based on set of properties* Let $A^I$ be a set of $n$ generalized place attributes, $A = \{a_1, a_2, \ldots, a_n\}$, where $a_1$ is the first attribute and so on. Let $W$ be a set of weights, one for each attribute. The similarity function for each attribute, $\mathrm{sim}_i$, will be determined by its type. The overall similarity measure for the places [see Eq. (3)] is a weighted combination of these similarity measures $(\mathrm{sim}_1, \ldots, \mathrm{sim}_n)$ according to the weights in $W$ [see Eqs. (1)–(2)]. The resulting set of target places are the top-K most similar based on this overall similarity measure.

$$SF1(X, s, A, W) \rightarrow T \tag{3}$$

An example of this search function is to find the best match to *New York City* based on a combination of *climate* and *demographic* variables.

*SF2a: Similar places to source, modified by quantitative scalar contrast property* The goal of $SF1$ is to find a set of target places that are similar to a source place. In contrast, function $SF2a$ [see Eq. (4)] finds a set of target places that are similar overall to the source place in terms of a set of relevant properties, but differ in terms of a contrast property, $a^*$. The manner in which it differs is determined by the modifier parameter $m \in \{+, \log +, -, \log -, \backslash\}$. A "$+$" or "$\log +$" modifier means that targets are desired that have a higher value for property $a^*$, and a "$-$" or "$\log -$" modifier means that targets are desired that have a lower value for topic $a^*$. The log modifiers will often be preferable when the property value is on a ratio scale. The "$\backslash$" modifier means that the target has a dissimilar value for topic $a^*$. Let

$y$ be a threshold parameter with value in $[0, \ldots, 1]$ that is used in conjunction with the modifier.

$$SF2a(X, s, A, W, a^*, m, y) \rightarrow T \tag{4}$$

As in the case of *SF1*, a candidate set $T'$ of target places is found based on $A$ similarity. $T'$ is then filtered to a subset $T$ based on the given contrast property. Let $sim^*$ be the similarity function for attribute $a^*$. Let $r^s$ be the value of $a^*$ for the source, and $r^t$ be the value of $a^*$ for the target.

If $m$ is "+", then let $r^{\mathrm{MAX}}$ be equal to the maximum value for $a^*$ in $T'$. Using this let $y' = y(r^{\mathrm{MAX}} - r^s) + r^s$. The target place is included in $T$ only if $r^t > r^s \wedge r^t > y'$. If $m$ is "*log+*", then these conditions are the same but $r^{\mathrm{MAX}}$, $r^s$, and $r^t$ are log-scaled: $y' = y(\log r^{\mathrm{MAX}} + \log r^s) + \log r^s$.

If $m$ is "−", then let $r^{\mathrm{MIN}}$ be equal to the minimum value for $a^*$ in $T'$. Using this let $y' = r^s - y(r^s - r^{\mathrm{MIN}})$. The target place is included in $T$ only if $r^t < r^s \wedge r^s < y'$. If $m$ is "*log−*", then these conditions are the same but $r^{\mathrm{MIN}}$, $r^s$, and $r^t$ are log-scaled: $y' = y(\log r^{\mathrm{MAX}} - \log r^s) + \log r^s$.

An example of this search function is to find the best match to *warm Anchorage*, where *warm* is a contrast property defined using the "+" modifier for the mean annual temperature attribute.

*SF2b: Similar places to source based on description, modified by categorical contrast property* This function given by Eq. (5) is similar to *SF*2*a*, except that the contrast property is a categorical property, such as place type or spatial relation. Let $a^C$ be a place attribute, $v$ a valid attribute value for $a^C$, and $m$ be the modifier parameter, such that $m \in \{+, -\}$.

$$SF2b(X, s, A, W, a^C, m, v) \rightarrow T \tag{5}$$

The candidate target set, $T'$, is found as in *SF*1. If $m$ is "+", then a target place in $T'$ is only included in $T$ if the value of $a^C$ is $v$. If $m$ is "−", it is only included if $a^C$ is not equal to $v$. Rather than a Boolean check, this function can easily be extended to use a threshold value based on semantic similarity measurements between attribute values (as in *SF*2*a*) (Janowicz and Wilkes 2009).

An example of this search function is to find the best match to *New York City in Europe*, where *in Europe* is defined using the '+' modifier and 'Europe' value for a spatial inclusion relation.

## 4.2 Using similar-place search for a case–control study

The similar-place search functions described above can be applied to the task of finding sets of case and control places for a case–control study. The "effect" property in a case–control study, $p_e$, is a generalized attribute value for the attribute $a_e$. Thus, a set of case places are places that have the "effect" property (i.e., they are invariant with respect to $a_e$). In the case of quantitative attributes, this invariance can be defined in terms of being within a specified distance threshold, rather than having an exact value, whereas for categorical attributes it is a Boolean measure.

A set of control places can be found by using $p_e$ as a negative contrast property (see functions *SF*2*a* and *SF*2*b*) and finding a set of target places to a source place that have the property $p_e$. In order to reduce sampling bias, once we have candidate sets of case and control places, we want to choose control places such that they are most representative of the types of places in the case set. One potential approach is to use propensity score matching, so that the distribution of baseline properties in the control places matches the distribution of those properties in the case places (Rosenbaum and Rubin 1983). In this way, a more representative subset of the target places is used as control places. However, this approach will require that the baseline properties be categorical; otherwise, an appropriate implementation of the property matching criteria will need to be developed. Once an investigator has sets of case and control places, then the presence or absence of the consequent property in the case and controls can be used to support or refute the hypothesis.

## 5 Application to natural language observations and generalizations

In this section, we demonstrate how the similar-place search operations described in the previous section can be applied to the case of finding similar places based on what people write about them. Natural language descriptions represent a large amount of the crowdsourced volunteered geographic information on the web. For example, the English version of Wikipedia contains over 600,000 *place* articles each with natural language text describing a place on the earth (Lehmann et al. 2009). Recently, a number of text mining techniques have been developed that allow us to operationalize these data through statistical means.

One particularly popular approach to discovering the latent topics in a corpus of documents is the latent Dirichlet allocation (LDA) model, which is a generative statistical model that describes the creation of a text document as a kind of random process (Blei et al. 2003). Each word in a document is selected by first picking a topic and then selecting a word from that topic. The data mining inference that we make using the model is to go backward from the observed words that we see in the corpus to the most likely topics to have generated those words. In order to do this inference a number of techniques have been developed including expectation maximization, variational Bayes, and Gibbs sampling Markov chain Monte Carlo (Blei et al. 2003; Griffiths and Steyvers 2004). As LDA is modular, it is easy to extend the model, and several extensions to LDA have been developed to characterize the mixture of topics associated with a place (Wang et al. 2007; Ramage et al. 2009; Eisenstein et al. 2010; Hao et al. 2010; Sizov 2010; Yin et al. 2011; Hong et al. 2012; Adams and Janowicz 2012). Figure 3 illustrates sample topics that are automatically discovered in a corpus of 275,000 travel blog entries (Adams and McKenzie 2013).

The operationalized knowledge about a place that is extracted using these topic modeling approaches has the data representation of a vector in an n-dimensional topic space. Each topic value is a probability of that topic $[0, \ldots, 1]$, where all the

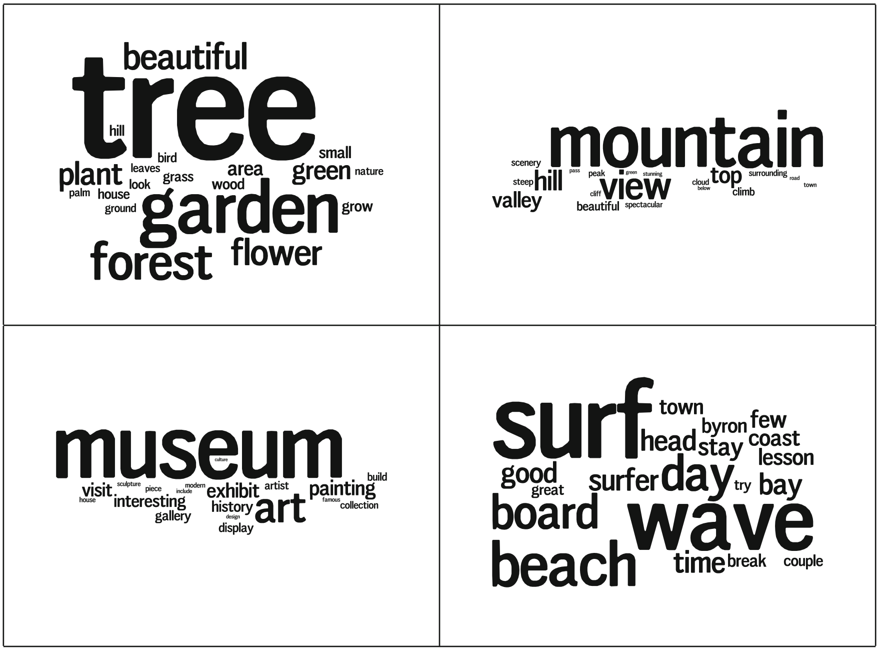

**Fig. 3** Sample latent topics discovered from running the latent Dirichlet allocation text mining technique on 275,000 travel blog entries. The size of each word in a topic is indicative of its relative probabilistic weight of being generated when that topic is selected. Each document in the corpus (and in turn place) is characterized as a probability vector over all 200 topics

values sum to one. Once we identify a vector of topic probabilities for each place, we can calculate their similarity using a variety of techniques, such as Euclidean distance, relative entropy, and Jensen Shannon (JS) divergence. The JS divergence is a symmetric measure of the distance between two multinomial probability distributions, derived from the Kullback–Leibler divergence $D_{\mathrm{KL}}$ (or relative entropy measure) shown in Eq. (6). Let $P$ and $Q$ be discrete probability vectors over $i \in \{1, \ldots, n\}$ values (e.g., topic vectors for two places) and $M = \frac{1}{2}(P + Q)$. The JS divergence measure is given in Eq. (7).

$$D_{\mathrm{KL}}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{6}$$

$$JS(P|Q) = \frac{1}{2} D_{\mathrm{KL}}(P|M) + \frac{1}{2} D_{\mathrm{KL}}(Q|M). \tag{7}$$

Because the dimensions of this topic space are semantically interpretable, it means that reasoning operations are not limited solely to similarity measurement between places; we can use the term distributions associated with each dimension (i.e., topic) to do automated reasoning to find real-world entities that best match novel combinations of terms with other places, such as *London* + *beach*, using contrast properties. Furthermore, because different text corpora generate different topic spaces and vectorial representations of places, each corpus presents a different domain by which the similarity of places can be judged. For example, Los Angeles as described in travel blog entries is very different from Los Angeles as described in Wikipedia or in historical literature.

### 5.1 Topic space-generalized attributes

The abstract observation-to-generalization place model does not predefine the types of attributes associated with observations and general properties, and it is possible to have complex-valued attributes that are represented with compound data structures, such as vectors, records, and objects. One example of a complex attribute is a topic space that is generated from a corpus of place descriptions using the techniques described previously.

A georeferenced natural language document can be modeled as an observation of a place by a human sensor. In particular, the attribute type is a textual description with the data-type structure of a term vector. Topic modeling is one inference mechanism for taking a set of these observations and producing a generalized attribute value, which can be assigned to a whole place. In this case, the generalized attribute is the place's topic mixture with the complex data-type structure of a topic space and result as topic vector in that space.

### 5.2 Thematically similar-place search operations

Using the templates defined in Sect. 4.2, we can define a set of specific similar-place search functions based on the topics derived from LDA.

*TSF1: Similar places based on descriptions from one corpus* Let $\Theta^I$ be a singleton set containing a topic space attribute, $a^{TS}$, from a set of place descriptions [see Eq. (8)].

$$TSF1(X, s, a^{TS}) \rightarrow T \tag{8}$$

Let $JS_2$ be the JS divergence function with base 2 logarithm, which is used because it produces a result in $[0, \ldots, 1]$. The similarity function, $sim^{TS}$, for $a^{TS}$ is defined as $1 - JS_2$. The set of target places are the top places based on $sim^{TS}$.

An example of this search function is to find the most similar places to *Sydney, Australia based on travel blog topics*.

The exploratory potential of applying this and the subsequent functions to discover relations between places is demonstrated by the following result. Using the travel blog corpus described in Fig. 3 as the input data, $TSF1$ was tested to find the most similar locations to Baghdad, Iraq, in terms of travel blog topics. An unexpected result was that Potsdam, Germany, was returned as the fourth most similar location by this measure. Although Al Asad Airbase in Iraq and Kandahar in Afghanistan (the top two most similar places) intuitively make sense as similar places, the high ranking of Potsdam is perhaps surprising. However, upon investigation of the topics in the relevant entries, it can be seen that both Potsdam and Baghdad share topics related to war, war-related conferences (Potsdam was a meeting point for the Allied powers soon after the end of World War II), and palaces.

*TSF2: Similar places with respect to a single topic* In the previous example, target analogs are found based on invariance with respect to a whole mixture of topics, but the topic space attribute is itself a complex attribute made up of individual topics. Here a function [see Eq. (9)] is described for finding places that are similar to a source place based on invariance of one topic. Let $a^{TS}$ be a topic space attribute where $q$ denotes the $q$th dimension of the topic space.

$$TSF2(X, s, a^{TS}, q) \rightarrow T \tag{9}$$

Let $r$ be the result along the $q$th dimension in the $a^{TS}$ value for a place; thus, $r$ is the probability that the topic will generate a particular word in a description of that place. From this, we can define a binomial distribution $\mathbf{r} = [r, 1 - r]$ for this place; that is the probability of $q$ and $\neg q$ for a place. Let $\mathbf{r^s}$ be $\mathbf{r}$ for the source place $s$ and $\mathbf{r^t}$ be $\mathbf{r}$ for a candidate target place $t$. The similarity, $\mathrm{sim}^q$, of $s$ and $t$ is defined as $1 - JS_2(\mathbf{r^s}, \mathbf{r^t})$. The set of target analogs are the top places based on $\mathrm{sim}^q$.

An example of this search function is to find the most similar places to *Santa Barbara with respect to 'beaches, sand, sun' topic.*

*TSF3: Similar places based on descriptions from multiple corpora* Let $\Pi^I$ be a set containing $n$ topic space attributes, $A^{TS} = \{a_1^{TS}, a_2^{TS}, \ldots, a_n^{TS}\}$, such that $n > 1$. Let $W$ be a set of weights, one for each topic space attribute. The similarity function for each topic space attribute is $sim^{TS} = 1 - JS_2$ as above. The overall similarity measure [see Eq. (10)] uses Eq. (1) to do a weighted combination of these similarity measures based on the weights in $W$, and the set of target places are the topmost similar based on this measure.

$$TSF3(X, s, A^{TS}, W) \rightarrow T \tag{10}$$

An example of this search function is to find the topmost similar places to *Seattle in terms of Wikipedia and travel blog entry topics.*

*TSF4: Similar places to source based on description, modified by contrast topic* In the previous examples, the goal was to find a set of target places that are similar to a source place based on descriptions or other properties. This function [see Eq. (11)] finds a set of target places that are similar overall to the source place in topic space, but which differ in terms of a *contrast topic*, $q^{\mathrm{CON}}$. The manner in which it differs is determined by the modifier parameter $m \in \{+, -, \backslash\}$. A "+" modifier means that targets are desired that have a higher value for topic $q^{\mathrm{CON}}$ and a "$-$" modifier means that targets are desired that have a lower value for topic $q^{\mathrm{CON}}$. The "\" modifier means that the target has a dissimilar value for topic $q^{\mathrm{CON}}$. Let $y$ be a threshold parameter $[0, \ldots, 1]$ used in conjunction with the modifier.

$$TSF4(X, s, a^{TS}, q^{\mathrm{CON}}, m, y) \rightarrow T \tag{11}$$

A candidate target set $T'$ is first found using the technique in $TSF1$, but these results are then filtered based on the contrast topic. Let $r^s$, $r^t$, $\mathbf{r^s}$, and $\mathbf{r^t}$ be defined for $q^{\mathrm{CON}}$ as in function $TSF2$.

If $m$ is "+", then let $r^{\mathrm{MAX}}$ be equal to the maximum value for $q^{\mathrm{CON}}$ in $T'$. Using this, let $y' = y(r^{\mathrm{MAX}} - r^s) + r^s$. The target place is included only if $r^t > r^s \wedge r^t > y'$.

If $m$ is "−", then let $r^{\mathrm{MIN}}$ be equal to the minimum value for $q^{\mathrm{CON}}$ in $T'$. Using this, let $y' = r^s - y(r^s - r^{\mathrm{MIN}})$. The target place is included only if $r^t < r^s \wedge r^s < y'$.

If $m$ is "\", then the target place is included only if $JS_2(\mathbf{r^s}, \mathbf{r^t}) > \mathbf{y}$.

Examples of this search function include to find the most similar places to *Denver + 'beaches, sand, surf'* or *Anchorage − 'snow, cold, wind'*.

## 6 Closing remarks

Future work will explore the efficacy of the observation-to-generalization place model for place-based analog search engines. This includes evaluating its value for several domains of place knowledge, using socioeconomic, environmental, and historical data. Finding places similar to arbitrarily shaped regions on the earth is another interesting extension to the similar-place search functions that is worth exploring. For example, rather than finding a place similar to a named place, a user might wish to find places similar to a region that is defined by an arbitrarily drawn polygon. In addition, the target places might also be arbitrarily shaped regions. Solving this problem is difficult because it would require repeated re-calculation of generalized attribute values from observation data. Furthermore, understanding the search space for target places of this kind is difficult because there are infinitely many ways to subdivide the space into candidate regions.

## References

Adams B, Janowicz K (2011) Constructing geo-ontologies by reification of observation data. In: Cruz IF, Agrawal D, Jensen CS, Ofek E, Tanin E (eds) GIS. ACM, Chicago, pp 309–318

Adams B, Janowicz K (2012) On the geo-indicativeness of non-georeferenced text. In: Breslin JG, Ellison NB, Shanahan JG, Tufekci Z (eds) ICWSM. The AAAI Press, Dublin, pp 375–378

Adams B, McKenzie G (2013) Inferring thematic places from spatially referenced natural language descriptions. In: Sui D, Elwood S, Goodchild M (eds) Crowdsourcing geographic knowledge. Springer, Berlin, pp 201–221

Adams B, Raubal M (2009) A metric conceptual space algebra. In: Hornsby K, Claramunt C, Denis M, Ligozat G (eds) Spatial information theory, lecture notes in computer science, vol 5756. Springer, Berlin, pp 51–68

Alves AO, Pereira FC, Biderman A, Ratti C (2009) Place enrichment by mining the web. In: Proceedings of the European conference on ambient intelligence (Am I '09). Springer, Berlin, pp 66–77

Banchuen T (2008) The geographical analog engine: hybrid numeric and semantic similarity measures for U.S. cities. Ph.D. thesis, The Pennsylvania State University

Bivand R, Gebhardt A (2000) Implementing functions for spatial statistical analysis using the language. J Geogr Syst 2(3):307–317

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(4/5):993–1022

Couclelis H (1992) People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. In: Frank AU, Campari I, Formentini U (eds) Spatio-temporal reasoning, lecture notes in computer science, vol 639. Springer, Berlin, pp 65–77

Couclelis H (2010) Ontologies of geographic information. Int J Geogr Inf Sci 24(12):1785–1809

Cresswell T (2004) Place: a short introduction. Blackwell Publishing Ltd, Oxford

Cronon W (2003) Changes in the land: Indians, colonists, and the ecology of New England. 20th, anniversary edn. Hill and Wang, New York

Eisenstein J, O'Connor B, Smith NA, Xing EP (2010) A latent variable model for geographic lexical variation. In: Li H, Màrquez L (eds) EMNLP. ACL, Cambridge, pp 1277–1287

Fotheringham AS, Wong DW (1991) The modifiable areal unit problem in multivariate statistical analysis. Environ Plan A 23(7):1025–1044

Frank AU (2001) Tiers of ontology and consistency constraints in geographical information systems. Int J Geogr Inf Sci 15(7):667–678

Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L (2002) Sweetening ontologies with DOLCE. In: Gómez-Pérez A, Benjamins VR (eds) EKAW, Lecture notes in computer science, vol 2473. Springer, Berlin, pp 166–181

Gärdenfors P (2000) Conceptual spaces: the geometry of thought. A Bradford book. MIT Press, Cambridge

Gatrell AC, Bailey TC, Diggle PJ, Rowlingson BS (1996) Spatial point pattern analysis and its application in geographical epidemiology. Trans Inst Br Geogr 21(1):256–274

Goodchild M (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221

Goodchild M, Montello D, Fohl P, Gottsegen J (1998) Fuzzy spatial queries in digital spatial data libraries. In: Simpson PK (ed) Proceedings of the IEEE international conference on fuzzy sysem, vol 1. IEEE, Anchorage, Alaska, pp 205–210

Graham LT, Gosling SD (2011) Can the ambiance of a place be determined by the user profiles of the people who visit it? In: Adamic LA, Baeza-Yates RA, Counts S (eds) ICWSM. The AAAI Press, Barcelona, pp 145–152

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(Suppl. 1):5228–5235

Halpern B, Longo C, Hardy D, McLeod K, Samhouri J, Katona S, Kleisner K, Lester S, O'Leary J, Ranelletti M, Rosenberg A, Scarborough C, Selig E, Best B, Brumbaugh D, Chapin F, Crowder L, Daly K, Doney S, Elfes C, Fogarty M, Gaines S, Jacobsen K, Karrer L, Leslie H, Neeley E, Pauly D, Polasky S, Ris B, St Martin K, Stone G, Sumaila U, Zeller D (2012) An index to assess the health and benefits of the global ocean. Nature 488(7413):615–620

Hao Q, Cai R, Wang C, Xiao R, Yang JM, Pang Y, Zhang L (2010) Equip tourists with knowledge mined from travelogues. In: Rappa M, Jones P, Freire J, Chakrabarti S (eds) WWW. ACM, Raleigh, pp 401–410

Heuvelink GBM (1998) Error propagation in environmental modelling With GIS. Taylor & Francis, New York

Hey T, Tansley S, Tolle KM (eds) (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond

Hill LL (2006) Georeferencing: the geographic associations of information. MIT Press, Cambridge

Hong L, Ahmed A, Gurumurthy S, Smola AJ, Tsioutsiouliklis K (2012) Discovering geographical topics in the twitter stream. In: Mille A, Gandon FL, Misselis J, Rabinovich M, Staab S (eds) WWW. ACM, Lyon, pp 769–778

Jacquez GM (2000) Spatial analysis in epidemiology: nascent science or a failure of GIS? J Geogr Syst 2(1):91–97

Janowicz K, Adams B, Raubal M (2010) Semantic referencing—determining context weights for similarity measurement. In: Fabrikant SI, Reichenbacher T, van Kreveld MJ, Schlieder C (eds) GIScience, Lecture notes in computer science, vol 6292. Springer, Berlin, pp 70–84

Janowicz K, Raubal M, Kuhn W (2011) The semantics of similarity in geographic information retrieval. J Spat Inf Sci 2(1):29–57

Janowicz K, Wilkes M (2009) SIM-DL_A: a novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity. ESWC, Heraklion

Kell DB, Oliver SG (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. BioEssays 26(1):99–105

Keßler C, Janowicz K, Bishr M (2009) An agenda for the next generation gazetteer: geographic information contribution and retrieval. In: Wolfson O, Agrawal D, Lu CT (eds) GIS. ACM, Seattle, pp 91–100

Kuhn W (2003) Semantic reference systems. Int J Geogr Inf Sci 17(5):405–409

Lehmann J, Bizer C, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) DBpedia—a crystallization point for the web of data. J Web Semant 7(3):154–165

Leung D, Newsam S (2010) Proximate sensing: inferring what-is-where from georeferenced photo collections. In: Davis L, Malik J (eds) CVPR. IEEE, San Francisco, pp 2955–2962

Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D, Villa F (2007) An ontology for describing and synthesizing ecological observation data. Ecol Inform 2(3):279–296. doi:10.1016/j.ecoinf.2007.05.004

Montello DR, Goodchild MF, Gottsegen J, Fohl P (2003) Where's downtown? Behavioral methods for determining referents of vague spatial queries. Spat Cogn Comput 3(2):185–204

Montero JM, Chasco C, Larraz B (2010) Building an environmental quality index for a big city: a spatial interpolation approach combined with a distance indicator. J Geogr Syst 12(4):435–459

Peel MC, Finlayson BL, McMahon TA (2007) Updated world map of the Köppen-Geiger climate classification. Hydrol Earth Syst Sci Discuss 4(2):439–473

Petersen J, Gibin M, Longley P, Mateos P, Atkinson P, Ashby D (2011) Geodemographics as a tool for targeting neighbourhoods in public health campaigns. J Geogr Syst 13(2):173–192. doi:10.1007/s10109-010-0113-9

Probst F (2006) Ontological analysis of observations and measurements. In: Raubal M, Miller HJ, Frank AU, Goodchild MF (eds) GIScience, Lecture notes in computer science, vol 4197. Springer, Berlin, pp 304–320

Probst F (2008) Observations, measurements and semantic reference spaces. Appl Ontol 3(1–2):63–89

Ramage D, Hall D, Nallapati R, Manning CD (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Koehn P, Mihalcea R (eds) EMNLP. ACL, Cambridge, pp 248–256

Relph E (1976) Place and placelessness. Pion, London

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Schade S, Ostermann F, Spinsanti L, Kuhn W (2012) Semantic observation integration. Future Internet 4(3):807–829

Schulz KF, Grimes DA (2002) Case–control studies: research in reverse. Lancet 359(9304):431–434

Selvin HC (1958) Durkheim's suicide and problems of empirical research. Am J Sociol 63(6):607–619

Sizov S (2010) GeoFolk: latent spatial semantics in web 2.0 social media. In: Suel T, Davison BD (eds) WSDM. ACM, New York, pp 281–290

Stasch C, Janowicz K, Bröring A, Reis I, Kuhn W (2009) A stimulus-centric algebraic approach to sensors and observations. In: Trigoni N, Markham A, Nawaz S (eds) GSN, Lecture notes in computer science, vol 5659. Springer, Berlin, pp 169–179

Tuan YF (1977) Space and Place: the Perspective of Experience. The Regents of the University of Minnesota, Saint Paul

United Nations Development Programme (1990) Human development report 1990. Oxford University Press, New York

Wang C, Wang J, Xie X, Ma WY (2007) Mining geographic knowledge using location aware topic model. In: Purves R, Jones C (eds) GIR. ACM, Lisbon, pp 65–70

Winter S, Freksa C (2012) Approaching the notion of place by contrast. J Spat Inf Sci 5:31–50

Winter S, Kuhn W, Krüger A (2009) Guest editorial: does place have a place in geographic information science? Spat Cogn Comput 9(3):171–173

Yin Z, Cao L, Han J, Zhai C, Huang TS (2011) Geographical topic discovery and comparison. In: Srinivasan S, Ramamritham K, Kumar A, Ravindra MP, Bertino E, Kumar R (eds) WWW. ACM, Hyderabad, pp 247–256