Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information

Christoph Mülligann[1], Krzysztof Janowicz[2], Mao Ye[3], and Wang-Chien Lee[3]

[1] Institute for Geoinformatics, University of Münster, Germany
cmuelligann@uni-muenster.de
[2] Department of Geography, University of California, Santa Barbara, USA
jano@geog.ucsb.edu
[3] Department of Computer Science and Engineering,
Pennsylvania State University, USA
{mxy177,wlee}@cse.psu.edu

Abstract. With the increasing success and commercial integration of Volunteered Geographic Information (VGI), the focus shifts away from coverage to data quality and homogeneity. Within the last years, several studies have been published analyzing the positional accuracy of features, completeness of specific attributes, or the topological consistency of line and polygon features. However, most of these studies do not take geographic feature types into account. This is for two reasons. First, and in contrast to street networks, choosing a reference set is difficult. Second, we lack the measures to quantify the degree of feature type miscategorization. In this work, we present a methodology to analyze the spatial-semantic interaction of point features in Volunteered Geographic Information. Feature types in VGI can be considered special in both, the way they are formed and the way they are applied. Given that they reflect community agreement more accurately than top-down approaches, we argue that they should be used as the primary basis for assessing spatial-semantic interaction. We present a case study on a spatial and semantic subset of OpenStreetMap, and introduce a novel semantic similarity measure based on the change history of OpenStreetMap elements. Our results set the stage for systems that assist VGI contributors in suggesting the types of new features, cleaning up existing data, and integrating data from different sources.

1 Introduction

The rise of Volunteered Geographic Information (VGI) as coined by Goodchild[1] is closely tied to projects such as OpenStreetMap (OSM)[1] or Wikimapia[2]. These projects provide open platforms for volunteers to contribute geographic data and make them accessible for others under an open license. Volunteered information

[1] http://www.openstreetmap.org
[2] http://wikimapia.org/

M. Egenhofer et al. (Eds.): COSIT 2011, LNCS 6899, pp. 350–370, 2011.

is acquired and maintained in a different style compared to data provided by professional authorities. Instead of being defined in a top-down manner, geographic feature types in OSM are the result of informal and continuous discussions within the community[3]. Consequently, contributors assign different category tags to features of similar types depending on their local VGI community, previous experience, used software, personal cognition of geographic space and changes of the OSM typing schema. Due to these factors, tags representing feature types change frequently.

With the increasing success of VGI and its integration with projects such as Wikipedia or even commercial products, quality control becomes equally important to mere coverage. Several researchers have studied the quality of Volunteered Geographic Information over the last years [2,3,4]. Tools assisting users in constraint checking, attribute enrichment, or the cleaning of large data sets become more important [5,6,7]. However, most studies do not take geographic feature types into account. In contrast to assessing data quality based on street networks or buildings, choosing reference data for feature types is difficult. Commercial routing data sets can be used to discover missing, dis-placed, or attribute-incomplete streets. Similarly, aerial photography or topographic maps can be used as references for features such as buildings or water bodies. There is no such gold standard for Points Of Interest (POI) feature types such as *Restaurant*, *Pub*, or *Theater*. Arguing that a feature tagged as *Pub* is mis-categorized because it is specified as *Bar* in a commercial data set is troublesome. This problem is not specific to VGI but a long term challenge in harvesting data across multiple gazetteers. Geo-ontologies have been proposed to make the various typing schemata explicit. Besides reference data for comparison, analyzing the usage of feature types in VGI requires measures to quantify the degree of mis-categorization. For instance, confusing pubs with bars is different from tagging a grocery store as a pub. Semantic similarity has been proposed as a measure to determine the difference between feature type definitions [8].

Analyzing the usage and implicit meaning of feature type tags is more than just an academic exercise. Intuitively, we expect a pub to be surrounded by other places that afford drinking alcohol, having a snack, or meeting friends, even though not all of these functions need to be offered by each facility. A waste basket, on the contrary, can be expected to be uniformly distributed within a commercial zoning area. The knowledge which types of features clump together and which most likely do not, can be used to improve VGI. A contributor uploading a fire station POI next to an existing one, may be automatically notified by the user interface that this type of features is not likely to clump together and asked to double check. Using similarity measures, such point patterns can also span the semantic dimension. Pubs are likely to occur next to nightclubs and cafés, but rather unlikely to be grouped around nursing homes. Discovering whether a specific feature, such as a point of interest, is already present before it gets duplicated by another contributor would free the resources of many volunteered editors that constantly work on cleaning up OpenStreetMap

[3] `http://wiki.openstreetmap.org/wiki/Map_Features`

data. Similarly, these editors also change category tags to make them match the latest community agreements or ensure that taxonomies are not confused with partonomies. Such assistant tools and rule systems have been frequently described as the next step in understanding and making use of VGI [4,9,10]. To assist users by suggesting the most likely feature type tags, or notify them if a similar feature already exists in the vicinity, requires the understanding of the spatial as well as the semantic patterns in OSM. In this work, we set the theoretical ground for developing such assistant tools. We present a spatial analysis methodology that identifies spatial-semantic patterns in OSM data and highlight how our approach can be used for tag recommendation and data cleaning. In contrast to existing work on geospatial semantics, we do not require a top-down ontology of geographic feature types, but derive feature type similarity bottom-up from the change history of existing OSM data.

The remainder of this paper is structured as follows. In section 2 we review statistics and measures that underlie our approach. Section 3 describes the development of concept variograms and spatial-semantic point pattern analysis. Both methods require semantic similarity values between feature types. The procedure of deriving these similarities in a bottom-up fashion from OSM data is explained in section 4. After giving an overview of the data set used for the case study (section 5) we describe the results (section 6) and discuss their implications (section 7). Finally, in section 8, we conclude by summarizing our work and point out directions for further work.

2 Related Work

This section introduces the statistical underpinning for our spatial-semantic interaction methodology and points to related work on semantic similarity measurement relevant for the understanding of our research. The geostatistics used in our methodology are well-established within the field of Geographic Information Science. In case of VGI, however, we lack this kind of well-established approaches. Kuhn [11] uses the *hot water* metaphor to picture the urgent need for models specialized on VGI. VGI represents a catalyst, or *hot water* to GI-Science once it is well understood and handled. Both directions of research, the understanding and the handling of VGI need to be integrated. This work is meant to be part of that integration task. In contrast to a social [12] or *producer*-centered [13] view on the topic, we aim at forming a computational basis for the interpretation of VGI datasets.

2.1 Spatial Analysis

Variograms. plot the expected difference between the values measured at two different locations versus their spatial distance. It is applied to describe the spatial dependency of continuous processes. Ahlqvist and Shortridge [14] argue that such a process can also underlie categorical variables like land use classes and introduce *semantic variograms* to analyze landscape heterogeneity. To do

so, they replace the differences between observed numerical variables by a look-up table containing semantic similarities (usually values within the range $[0, 1]$) for each pair of categorical land uses values. The *semantic semivariance* is then defined as a function of distance (or lag) $\mathbf{h}$; see equation 1:

$$\gamma_{SD}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} sd\left[c\left(\mathbf{u}_\alpha\right); c\left(\mathbf{u}_\alpha + \mathbf{h}\right)\right]^2 \tag{1}$$

where N is the number of location pairs separated by spatial lag $\mathbf{h}$, while $sd\left[c\left(\mathbf{u}_\alpha\right); c\left(\mathbf{u}_\alpha + \mathbf{h}\right)\right]$ is the semantic distance between the categorical land use value of points $\mathbf{u}_\alpha$ and $\mathbf{u}_\alpha + \mathbf{h}$ from the look-up table.

Even though point data as well as grids may underlie the computation of variograms, they model fields and not point patterns. Variograms are used to interpolate values between measurement points, e.g. using Kriging [15], whereas the distribution of the measurement points as such is not targeted.

Point Pattern Analysis aims to reveal whether points in a study area are, e.g., clumped, randomly, or regularly distributed. A popular model for a stationary spatial point process is Ripley's K [16]; see equation 2:

$$K\left(s\right) = \lambda^{-1} E \tag{2}$$

where E is the number of occurrences within distance s of an arbitrary point and λ is the the intensity, i.e., the expected number of points, overall density, or average occurrence rate respectively [17]. The function is monotonically non-decreasing [17]. Hence the minimal increase of 0 between two distances s_1 and s_2 means that no additional points are expected when increasing the radius by $s_2 - s_1$ with regard to an arbitrary point. Accordingly a strong increase up to a distance s_x indicates a clustering within this radius. There are superimposed functions that ease visual interpretation of thresholds like s_x (cp. the L index or the linearized version of Ripley's K respectively [18]). However, for a comparison of expected versus observed occurrences this additional step is not necessary.

To our best knowledge, there is no existing methodology to account for semantic aspects in addition to the spatial distribution of point patterns. However, Diggle et al. [19] introduced a second-moment spatio-*temporal* measure for point-processes in which the spatio-temporal occurrence is called an event; see equation 3:

$$K\left(s,t\right) = \lambda^{-1} E \tag{3}$$

where E is the number of events occurring within distance s and time t of an arbitrary event, and λ is the intensity, i.e., the expected number of events per unit space per unit time. Following the spatial definition of Ripley's K, the estimator $\widehat{K}\left(s,t\right)$ [19] can be computed from existing data by equation 4:

$$\widehat{K}\left(s,t\right) = |A|\, T\left(n\left(n-1\right)\right)^{-1} \sum_{j \neq i} w_{ij} v_{ij} I\left(d_{ij} \leq s\right) I\left(u_{ij} \leq t\right) \tag{4}$$

where $|A|$ is the area of a polygon enclosing the spatial domain of interest, T the analogous temporal interval. n is the number of points, I the indicator function and d_{ij} and u_{ij} are the spatial and temporal differences. w_{ij},v_{ij} are weights applied for the correction of edge-effects.

As diagnostic measure for the actual strength of $K(s,t)$, Diggle et al. propose the functions 5 and 6.

$$\widehat{D}(s,t) = \widehat{K}(s,t) - \widehat{K}(s)\,\widehat{K}(t) \tag{5}$$

$$\widehat{D}_0(s,t) = \frac{\widehat{D}(s,t)}{\widehat{K}(s)\,\widehat{K}(t)} \tag{6}$$

$\widehat{K}(s)$ and $\widehat{K}(t)$ are the independent spatial and temporal components of the underlying point process. $\widehat{D}$ describes the absolute difference between the spatio-temporal and an assumed independent spatial and temporal process, $\widehat{D}_0$ its magnitude with regard to an expected number of occurrences in a spatial- and temporal-only process. Space-time interaction is therefore described as a space- and time-dependent factor that measures the influence of the combined point process versus the independent point processes.

2.2 Semantic Similarity

Due to their analogy to spatial proximity functions, semantic similarity measures have been widely studied and applied in GIScience [20,21,14,22,8,23]. Most of these measures are hybrid in a sense that they combine different approaches to similarity, such as features, regions in a multi-dimensional space, or network distances. However, these approaches rely on existing ontologies or scene graphs for comparison. The OSM data set discussed in this work lacks a formal specification of feature types. It also does not support multiple types per feature which excludes classical *bag of words* approaches. In contrast to such set-theoretic approaches, Eck et al. [24] identified probabilistic measures, the association strength in particular, as adequate for normalization purposes because it measures the deviation of observed from expected co-occurrences. Set-theoretic measures like the inclusion index or the Jaccard index return the relative overlap of two sets, still being prone to the absolute number of tags in each of the sets [24].

The association strength, proximity index, or probabilistic affinity index, respectively [25,26,27], is the ratio of the observed number of co-occurrences c_{ij} and the expected number of co-occurrences e_{ij} between tags i and j; see equation 7:

$$e_{ij} = \frac{s_i s_j}{m} \tag{7}$$

s_i and s_j are the total numbers of occurrences for each, i and j, and m is the total number of documents or bag of words respectively. For

$$sim_{AS} = \frac{c_{ij}}{e_{ij}} \tag{8}$$

greater than 1 the number of co-occurrences is higher than expected for assumed statistical independency, lower otherwise.

3 Two Models for Spatial-Semantic Interaction

In this section we explain the changes applied to semantic variograms and the diagnostic measure $\widehat{D}_0$ related to spatio-temporal point processes. While semantic variograms reflect a field view on geographic space where point features are only considered measurement locations for a spatial process such as land cover, Diggle's diagnostics are based on the model of a point process where the distribution of occurrences in two dimensions is the only relevant information.

Semantic variograms have been used to study land cover grids so far. The combination of semantics with variogram is reasonable in this case because land cover is present at any location. Despite the limited amount of classes, the underlying process can be considered continuous. When applying variograms to points of interest, those properties are not given anymore. Given a subset of POI, e.g. amenities, most of the space in between would be void or of no relevance to amenities. Even more, the amenities themselves are not modeled as two-dimensional features. Variograms by nature do not allow for these kinds of situations, because the process to be modeled is considered ubiquitous.

Nevertheless, beyond those theoretical reservations, the computation and careful interpretation of variograms is possible and straightforward compared to the $\widehat{D}_0$ statistic. Investigating both approaches gives us the chance to understand their limitations and which patterns they help to uncover.

Fig. 1 shows the basic steps that precede the computation of our spatial-semantic interaction models. Concept variograms and second-order analysis both require spatial and semantic distances as input data. In our work, the semantic distance is derived from a similarity matrix which also defines the POI to be selected for analysis.

All computations were performed by the statistical language R[4]. In particular we modified functions from the *gstat*[5] and the *splancs*[6] packages.

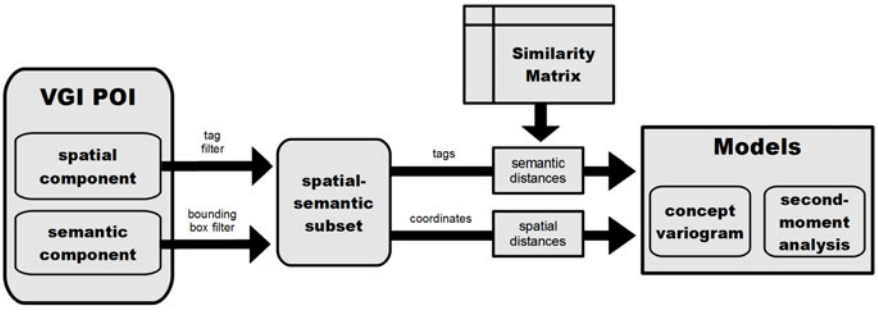

Fig. 1. Workflow of the spatial-semantic interaction analysis

[4] http://www.r-project.org/
[5] http://www.gstat.org/
[6] http://www.maths.lancs.ac.uk/∼rowlings/Splancs/

3.1 Concept Variograms

For the characterization of a certain concept c_k, or categorical value, respectively, we aim at extracting only those spatial-semantic relationships that are relevant for c_k. We achieve this by applying the following change to the semantic variogram definition (cp. section 2):

$$\gamma_{SD}^{c_k}(\mathbf{h}) = \frac{1}{2N(\mathbf{h}, c_k)} \sum_{\alpha=1}^{N(\mathbf{h},c_k)} sd\left[c\left(\mathbf{u}_\alpha\right); c\left(\mathbf{u}_\alpha + \mathbf{h}\right)\right]^2 \tag{9}$$

where $N(\mathbf{h}, c_k)$ is the number of point pairs separated by spatial length $\mathbf{h}$ and fulfilling the condition $c\left(p_i\right) \vee c\left(p_j\right) = c_k$ for each point pair (p_i, p_j). Therefore $\gamma_{SD}^{c_k}(\mathbf{h})$ can be either considered a semantic-enabled version of Goovaerts' indicator variograms [28] or a restricted form of the semantic variogram definition by Ahlqvist et al [14].

3.2 Second-Order Analysis of Spatial-Semantic Clustering

The modifications applied to the second-moment spatio-temporal measure for point-processes by Diggle et al.[19] consist of replacing the temporal by a semantic component as well as restricting the point-pairs contributing to K to those that at least have one point with value s_k. We also changed the notion of E to the more neutral term *occurrence* since *event* only applies to the temporal domain. The altered K-function is then given by:

$$K_{SD}^{c_k}\left(s, sd\right) = \lambda^{-1} E_{c_k} \tag{10}$$

where SD is the semantic distance matrix in use, sd the semantic distance, c_k the selected concept and E_{c_k} the number of further occurrences within spatial distance s and semantic distance sd with respect to occurrences having the categorical value c_k. Accordingly, we propose the following as an estimator for function (10):

$$\begin{aligned}\widehat{K}_{SD}^{c_k}\ \left(s, sd\right) &= |A|\, SD_{\text{range}}\left(n_{c_k}\left(n-1\right)\right)^{-1} \\ &\sum_{j \neq i} w_{ij} I\left(d_{ij} \leq s\right) I\left(sd\left[c\left(p_i\right); c\left(p_j\right)\right] \leq sd\right) I\left(c\left(p_i\right) \vee c\left(p_j\right) = c_k\right)\end{aligned} \tag{11}$$

where SD_{range} is the range of semantic similarity values in the look-up table and n_{c_k} is the number of points with $c\left(p_i\right) = c_k$. Note that no correction of edge-effects (v_{ij} in $\widehat{K}\left(s,t\right)$) is applied to the categorical values due to a lack of metrical spaces for this kind of pair-wise distances. Hence, the notion of an *edge* is not meaningful here. In that regard another modification has to be applied. In contrast to time, semantic distance has a fixed range, usually values from $[0, 1]$. Since we want to examine the whole semantic range of the $\widehat{D}_0$ statistic, the case of POI having a maximum dissimilarity of 1 needs to be addressed separately. Reaching this value, all remaining POI are added to $\widehat{K}_{SD}^{c_k}\left(sd\right)$ and consequently $\widehat{D}_0$ becomes 0. Thereby,

those values within the last similarity interval would be ignored automatically. The traditional approach does not face such issue because the spatial or temporal dimension is seldom captured up to the maximum of the dataset. For semantic distances, however, we define SD_{range} as an open interval $[0, 1)$.

To get a better understanding of how the $\widehat{D}_0$ statistic reflects the underlying spatial-semantic interaction, we created different simulated point patterns based on a simple four-step similarity scale for a test concept c_k. Out of those we show two patterns and their corresponding plots in fig. 2. In the following they will be called pattern *Pattern A* (upper) and pattern *Pattern B* (lower).

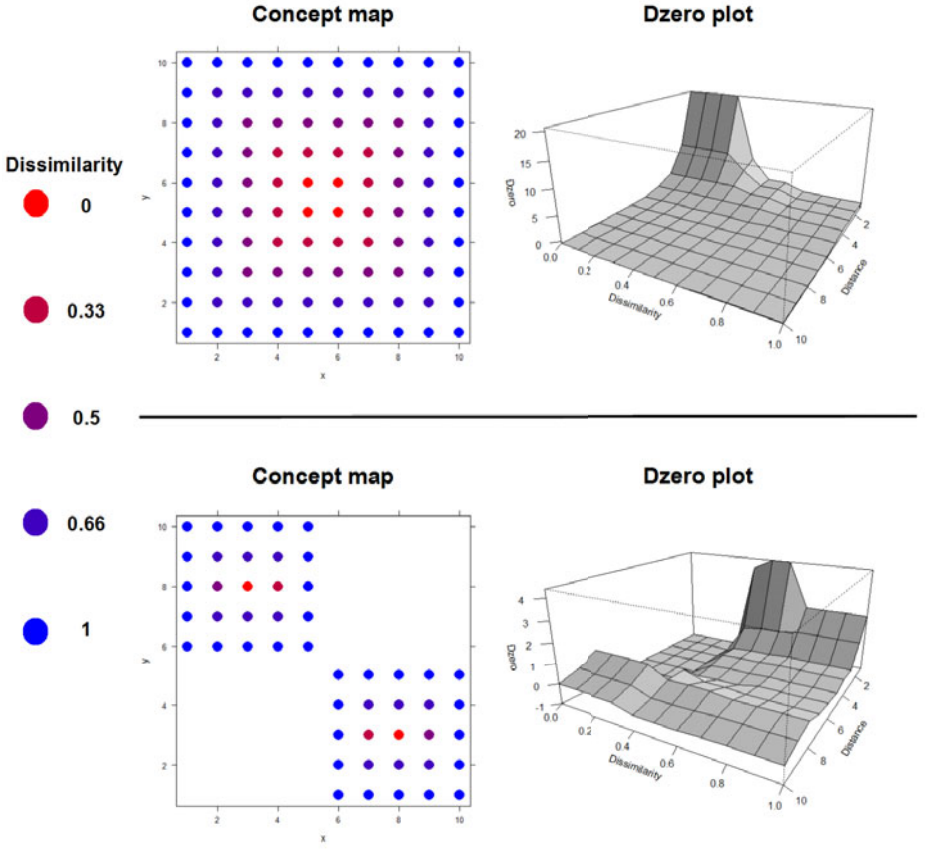

Fig. 2. $\widehat{D}_0$ plots for two different spatial-semantic patterns. Light red dots indicate instances of the concept c_k to be examined.

Pattern A represents perfect spatial-semantic autocorrelation - *perfect* in the sense that spatial distance to instances of c_k (light red dots) is always relative to the semantic distance. Since semantic similarity decreases in both spatial dimensions, outgoing from the four c_k POI, we observe a quadratic drop in the $\widehat{D}_0$ plot. In the following, we focus on the high $\widehat{D}_0$ values in the two-dimensional

(spatial-semantic) interval of $[0, 0], [1, 0.3]$, i.e., the values within spatial distance 0 - 1 and semantic distance 0 - 0.3. The *expected* number of co-occurrences in this interval is low. One the one hand, there is no spatial clustering around c_k POI; the distribution of all POIs is regular. On the other hand, the fraction of POIs of type c_k is only 4%. In other words, in a *random distribution* of the very same set of points it would be quite unlikely that c_k POI would appear right next to each other. However, in the *observed distribution* c_k only occurs as a single cluster. The $\widehat{D}_0$ plots shows that this phenomenon is 20 times more likely to be characteristic than a random distribution. However, as soon as we extent the spatial-semantic interval in either dimension this factor decreases: finding c_k POI within a special distance of 2 is more probable than within distance 1, and finding c_k as well as similar POI (dissimilarity 0.33) is more probable than only c_k POI within distance 1.

Pattern B is a counter-example for spatial autocorrelation. First, c_k POI do not form spatial clusters themselves, second the surrounding POI are mostly dissimilar. Hence, $\widehat{D}_0$ has negative values in the spatial and semantic proximity of c_k POI, i.e., the independent spatial and semantic clustering is stronger than the combined point process. There are two clusters where medium similarity appear only next to a c_k POI. Therefore, $\widehat{D}_0$ is highest for POI with medium similarity.

This behavior is important for our analysis with regard to VGI and Geographic Information in general. The total occurrences of feature types in a geo-dataset may vary for several reasons, e.g., different interests of voluntary mappers, heterogeneous coverage, or ground truth. The interaction signature of a geographic feature type, however, should not be biased by the number of instance occurrences because these do not play a role on the conceptual level. *Pattern B* shows that in the domain of a particular geo-dataset the $\widehat{D}_0$ statistic accounts for the diagnosticity of feature types, i.e., features within a certain similarity range, here 0.3 to 0.6.

Finally, note that the $\widehat{D}_0$ plots would look exactly the same except from the last semantic interval dropping to 0 if POI with dissimilarity 1, i.e., no similarity, would have been incorporated. By not doing so, we are able to visualize the semantic interval $(0.9, 1.0)$ and show that no change occurs there compared to the interval $(0.7, 0.9]$. Points of interest with no similarity still have strong effect on $\widehat{D}_0$ though, because they affect interaction trends of the spatial axis through $\widehat{K}(s)$ (cp. eq. 5 and 6).

4 Deriving Similarity from the OpenStreetMap History

Introducing semantics into geostatistical models is a key contribution of this work. However, we cannot derive similarity values from formal feature types as these do not exist for VGI and, therefore, have to assess pair-wise similarities between each type.

Our case-study is restricted to a particular subset of elements in OSM, namely those that have a key called *amenity*. A key in OSM can be considered a superconcept, its values the subconcepts. Currently the community agrees on 71

different amenity values described in the OSM wiki[7]. By convention these key-value pairs are meant to be applied uniquely, i.e., an OSM node is not supposed to have more than one amenity tag. Therefore, we cannot use *bag-of-words*-based similarity measures between different amenity values. Instead, we obtain the history set (in form of a *bag of words*) for each OSM element. OpenStreetMap offers *diff* files which list all elements that were subjects to change, i.e., creation, modification, or deletion, within a certain time frame.

We use this diff function to compute the history set for all elements. For instance, an element x that was created as a `cafe`, then changed to `restaurant`, then changed back to `cafe,` and finally labeled `bar`, would contain the tags `cafe`, `restaurant`, and `bar` as a bag of words. The number of changes or their sequence is not recorded. For our similarity measure, we assume that such type changes occur due to *semantic confusion* of types by VGI contributors. Based on the history sets, we create a co-occurrence matrix C with c_{ij} containing the number of elements that have been both, tag i and tag j during their history. The diagonal entries of C contain the total number of i / j elements.

Next, we apply the association strength measure to compute the expected number of co-occurrences (cp. eq. 8). In order to get values within $[0, 1]$, sim_{AS} is normalized following equation 12):

$$sim = 1 - \frac{1}{1 + sim_{AS}}. \tag{12}$$

Maximum dissimilarity is marked by a value of 1, while maximum similarity takes the value 0. A value of 0.5 reflects statistical independency.

5 Data and Case Study

This case study examines amenities in London as a spatial-semantic subset of the OpenStreetMap dataset. The semantic similarities of OpenStreetMap amenities are derived from the whole world dataset to achieve a higher degree of significance. In the following, we will describe the spatial and semantic components of amenity points of interest.

5.1 Amenity Points of Interest in OpenStreetMap

Amenity POI may be mapped as *nodes* or *ways* in OpenStreetMap. In the first case they are modeled as point features, as polygons otherwise. For our case study, the bounding box for the London dataset is set to (51.4158,-0.331), (51.6011,0.0796). Data was retrieved from OSM's extended API (see requests[8]). All polygon features were converted to point features after retrieval by a centroid function. Thereby, they can be used in our methodology and we do not loose valuable information. The final dataset contains 20,765 POI with 64 out of 71 different amenity values being present. Table 1 shows their tag counts.

[7] `http://wiki.openstreetmap.org/wiki/Map_Features#Amenity`

[8] `http://xapi.openstreetmap.org/api/0.6/node[amenity=*][bbox=-0.331,51.4158,0.0796,51.6011]`, `http://xapi.openstreetmap.org/api/0.6/way[amenity=*][bbox=-0.331,51.4158,0.0796,51.6011]`

Table 1. Numbers of amenity tags in the London dataset

tag	#	tag	#	tag	#	tag	#
arts_centre	50	atm	330	bank	464	bar	235
bench	219	bicycle_parking	1479	bicycle_rental	343	biergarten	3
bureau_de_change	20	bus_station	20	cafe	1273	car_rental	16
car_sharing	600	car_wash	15	cinema	61	clock	11
college	111	community_centre	62	courthouse	36	crematorium	1
dentist	77	doctors	144	drinking_water	10	embassy	61
fast_food	708	ferry_terminal	19	fire_station	68	fountain	46
fuel	201	grave_yard	16	grit_bin	18	hospital	97
kindergarten	42	library	171	marketplace	32	nightclub	56
parking	1225	pharmacy	275	place_of_worship	1356	police	89
post_box	3086	post_office	284	prison	6	pub	2556
public_building	121	recycling	397	restaurant	1855	sauna	1
school	1358	shelter	17	social_centre	1	social_facility	5
stripclub	1	studio	8	taxi	44	telephone	1274
theatre	111	toilets	214	townhall	19	university	107
vending_machine	3	veterinary	9	waste_basket	225	waste_disposal	3

5.2 Semantic Similarity of Amenities in OpenStreetMap

The total number of features tagged with one of the 71 amenities values is 3,247,409, considering the whole world (state: February 2011). Out of these, 30,538 OSM elements have been subject to tag changes.

For illustration purpose, table 2 shows a subset of the resulting co-occurrence matrix. The diagonal entries show the total number of occurrences of the corresponding element. The similarity values computed based on table 2 are shown in table 3. Due to the low number of overall changes to the amenity dataset, we tested each co-occurrence for statistical significance. The test was carried out as a χ^2 test of the 2x2 contingency table of each tag pair. While the test statistic itself lacks features for semantic similarity, its p-value shows how reliable the raw data is. It turns out that for 25.5% of all co-occurrences the strength of association is not significant on a 95% confidence level. Therefore we assume that our similarity measure has an accuracy of at least 74%.

Table 2. Examples for amenity co-occurrence in OpenStreetMap history sets

	bar	cafe	cinema	community_centre	recycling	theatre	waste_basket
bar	16799	392	2	1	1	2	0
cafe	392	57343	6	1	10	5	3
cinema	2	6	11808	3	0	104	0
community_centre	1	1	3	2306	0	11	0
recycling	1	10	0	0	60309	2	122
theatre	2	5	104	11	2	11569	0
waste_basket	0	3	0	0	122	0	19976

In general, and taking into account that these values have been automatically derived from VGI without any pre-processing, the similarity values are plausible. However, the number of completely dissimilar tags is higher than expected. For instance, the number of tags that have a similarity value lower than 0.1 is 59 for

Table 3. Selected normalized association strengths of OpenStreetMap amenities

	bar	cafe	cinema	community_centre	recycling	theatre	waste_basket
bar	1	0.57	0.03	0.08	0	0.03	0
cafe	0.57	1	0.03	0.02	0.01	0.02	0.01
cinema	0.03	0.03	1	0.26	0	0.71	0
community_centre	0.08	0.02	0.26	1	0	0.57	0
recycling	0	0.01	0	0	1	0.01	0.25
theatre	0.03	0.02	0.71	0.57	0.01	1	0
waste_basket	0	0.01	0	0	0.25	0	1

`bar`, 61 for `cafe`, 62 for `cinema`, 51 for `communtiy_centre`, 67 for `recycling`, 56 for `theatre`, and 68 for `waste_basket` (from an overall number of 71 amenity values). This leads to a coarse semantic granularity with respect to similar category tags.

Additionally, the results are influenced by partonomic (e.g. for `parking`), linguistic (e.g. `bank` and `bench`), and lexical relations (e.g. `watering_place` and `ferry_terminal`). While partonomic relations may be considered a valuable influence on semantic similarity between geographic features, we have to treat the others as errors. Fortunately, in almost all cases, the linguistic or lexical bias comes along with a strong semantic association, (e.g., for `theatre` and `cinema`), or impacts tags that have a very low overall change rate (e.g. `bench`). Nonetheless, this shows that semantic similarities computed out of such history sets should not be applied without prior manual inspection. Finally, concept variograms and point pattern analysis require a dissimilarity values. Therefore, all values were inverted by $dissimilarity = 1 - similarity$.

6 Results

This section presents the results of applying the concept variograms as well as the spatial-semantic point pattern analysis to the OpenStreetMap POI data set for London. Concept variograms and $\widehat{D}_0$ statistics were created for all 64 amenity tags. We selected four amenities, namely `bar`, `cafe`, `post_office` and `theater` to show a spectrum of spatial-semantic interaction and possible interpretations. Fig. 3 shows a map view of all POI in a narrower bounding box, where a warmer color indicates higher similarity to a particular tag.

The data set contains 235 bars, 1273 cafés, 284 post offices, and 111 theaters. Whereas bars and cafés, as well as tags similar to them, are equally prominent in the city center, theaters appear less often and only in a certain region with similar tags. Post offices are regularly spread over the whole area and rarely cluster with similar tags.

Analyzing the spatial autocorrelation with respect to semantic similarities gives a first assessment of the qualitative differences between the four amenities; see fig. 4. `post_office` and `theatre` show nearly no spatial autocorrelation. An increase of semivariance can only be observed for `bar` and `cafe` up to a distance of 300 m. `post_office` shows the lowest overall similarity, followed by `theater`, `cafe` and `bar`. These values (on the y-axis) are called the sill of a variogram. The range of all four variograms, i.e., the maximum distance up to

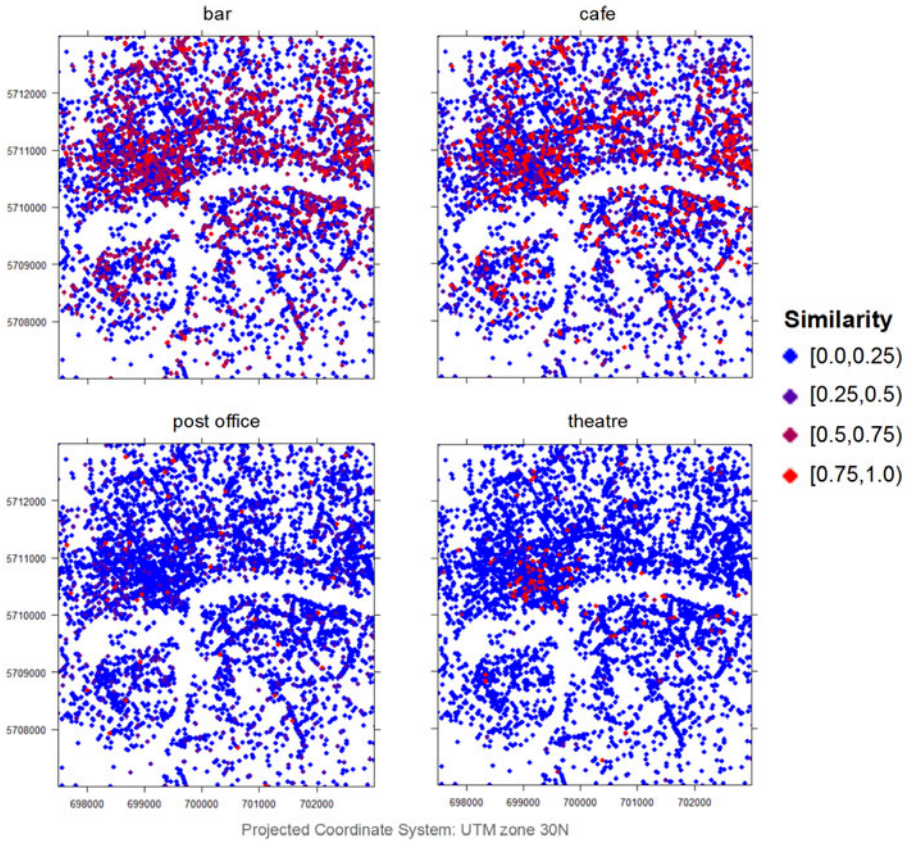

Fig. 3. Similarity values of four amenity tags in a subregion of the London OSM data

which spatial autocorrelation is observed, is roughly between 700 and 1000 m. Therefore, we used the latter value as a threshold for the second-order analysis of spatial-semantic interaction.

Fig. 5 shows $\widehat{D}_0$ plots for the four selected amenities. Bars show spatial-semantic interaction on small spatial and semantic scale, i.e., less than 300 m and below 0.4 dissimilarity. The same applies to cafés regarding their spatial component. The semantic tolerance for interaction appears to be higher than the one for bars here. Theaters show a completely different pattern. The magnitude of interaction highly correlates with spatial and semantic distance. Especially the decrease of interaction in the spatial dimension is smoother for theatres than for cafés and bars. Post offices differ from the other three amenities in showing negative spatial-semantic interaction, i.e., the independent spatial and semantic clustering is stronger than the spatial-semantic one. While dissimilar features have zero interaction with post offices at any distance, negative interaction increases for more similar and closer features. The strength of interaction in general is high for bars and theatres and comparably low for cafes and post offices.

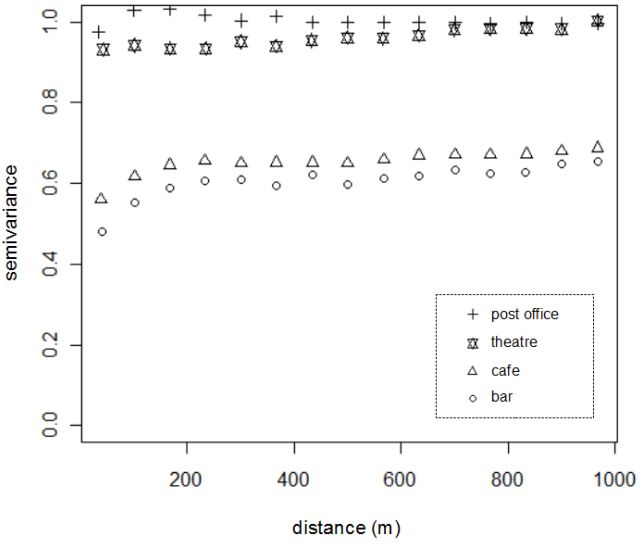

Fig. 4. Concept variograms for four amenity tags in the London dataset

7 Discussion

In this section we discuss the results form the case study and focus on the interpretation of the introduced spatial-semantic interaction models (section 7.1). Subsequently four different application scenarios in the scope of VGI are presented (section 7.2).

7.1 Interpretation of the Results

The examples in section 6 demonstrate that $\widehat{D}_0$ plots have the potential of revealing more information about spatial-semantic interaction than concept variograms. They explicitly plot spatial-semantic interaction on both scales. Therefore, we can observe that, e.g., bars cluster only with very similar amenities, whereas cafés seem to appear in a more diversified environment. Post offices are regularly spaced, primarily with themselves but also with slightly dissimilar features like post boxes. This results in negative spatial-semantic interaction as it occurs in *Pattern B* (cp. section 3). From the $\widehat{D}_0$ plot we can observe that it is more characteristic for post offices to be surrounded by dissimilar than by similar features - due to their public supply function they appear in all kinds of environments.

Theaters seem to be clustered with similar amenities on a much higher spatial and semantic scale that is not captured by the corresponding concept variogram. Nevertheless, theaters can be considered the same interaction type as cafés and bars, if examined on a smaller spatial scale; see fig. 5. Therefore, they correspond

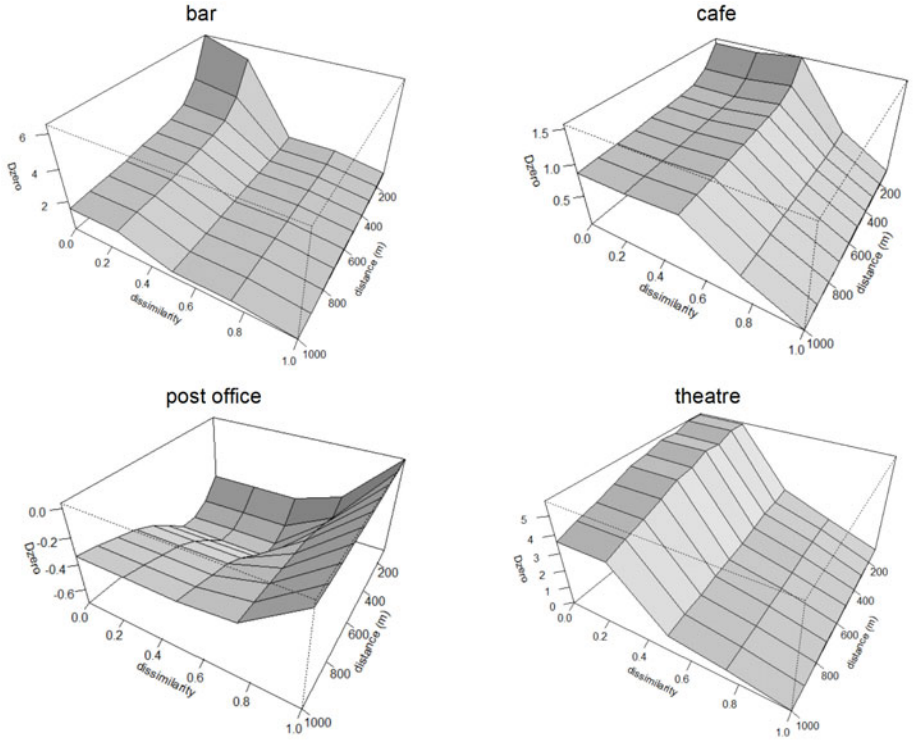

Fig. 5. $\widehat{D}_0$ plots for four amenity tags in the London dataset showing the magnitude of spatial-semantic interaction at different spatial and semantic scales

to the prototypical distribution of *Pattern A* in section 3. In contrast to cafés and bars, theatres only clump in London's city center. Also similar amenities co-occur with theaters under high diagnosticity even for greater distances. When forming a spatial-semantic cluster of certain size, we can assume that a geographic feature has a function that is related to the magnitude of the cluster. A cafe or bar may be important to a block or certain street, whereas a theatre leaves its interaction traces in the whole city center.

The concept variogram of theatres does not reflect the situation described above. There are too many completely dissimilar POI that hide the contribution of similar ones to a possible smaller-scale cluster. This shows the advantage of the point pattern analysis to incorporate the diagnosticity of POI within a certain semantic range. Beyond that, we cannot consider POI to have an underlying continuous spatial process of *theatreness*. It is rather the spatial pattern of theatres, intertwined with the spatial patterns of other amenities, that is characteristic for the geographic feature type *theatre*. By comparing the results of

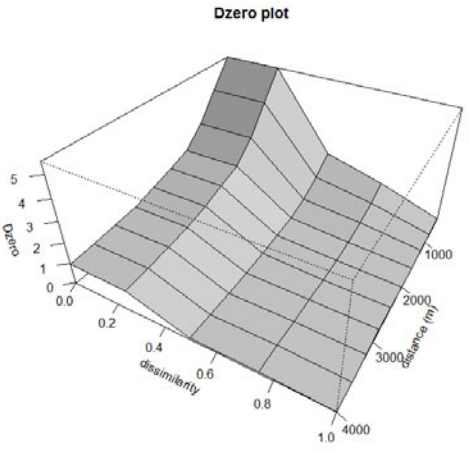

Fig. 6. $\widehat{D}_0$ plot of theatres at a smaller scale

point pattern analysis and concepts variograms we are able to show that the theoretical reservations mentioned in section 3 have practical relevance.

The resolution of the semantic dimension was chosen according to the granularity of the similarity measure (i.e, in 4 intervals). However, the accuracy of $\widehat{D}_0$ is likely to increase if the semantic similarities are distributed over the whole range of possible values. For example, the similarity of `cafe` to `bbq` and `cafe` to `dentist` is 0, and intuitively that could be considered reasonable in terms of their confusion possibility (which is underlying our similarity measure). However, keeping in mind that 0 is the minimal possible similarity we may want to distinguish both cases when assessing semantics in general. So far, the similarity measure applied is rather conservative. Dissimilar features are strongly penalized, which results in coarser semantics. Consequently, $\widehat{D}_0$ plots are not as smooth as in the original spatio-temporal application proposed by Diggle.

7.2 Application of Spatial-Semantic Interaction Models in VGI

Concept variograms and second-order analysis represent the spatial-semantic interaction signature of geographic feature types in a POI dataset. In section 7.1 we argue that the $\widehat{D}_0$ statistics reveal more information than concept variograms. Therefore, we will focus on application scenarios that use the second-order analysis. Their implementation as software remains future work.

The basis for an application of interaction signatures in VGI is a plausibility measure for an individual POI with respect to its feature type. Plausibility in our terminology is different from probability in the sense that we do not aim at predicting feature types in analogy to geostatistical interpolation. It can be computed for an arbitrary location s and feature type c_k based on the comparison between the $\widehat{D}_0$ statistic of a single POI with type c_k at s and the $\widehat{D}_0$ statistic of

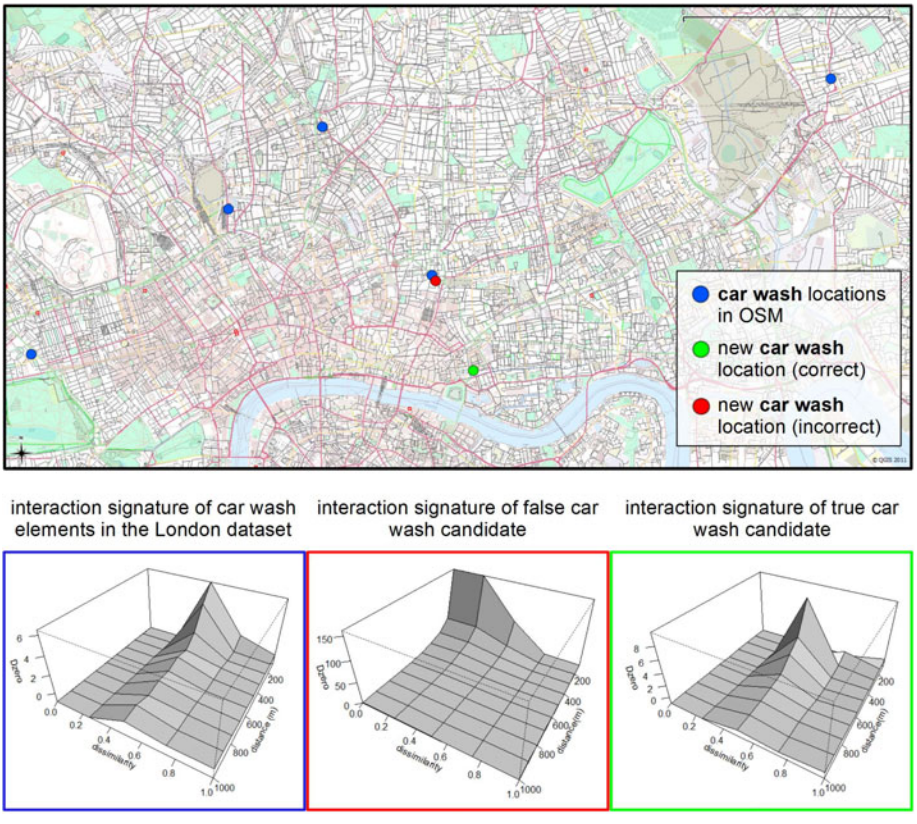

Fig. 7. Second-order analysis in action: two candidates for car wash locations in London are checked for plausibility by comparing their spatial-semantic interaction signature with the existing one. The green dot is a real car wash location extracted from Google Maps. The red dot simulates a duplicate tag. (map rendered by Quantum GIS, bounding box: 51.501,-0.171; 51.557,0.023).

all c_k in the dataset. Fig. 7 shows a real world example of $\widehat{D}_0$ plots. The second plot represents the case of a candidate with low plausibility, the third a case of high plausibility with regard to the first plot. Future work will focus on the numerical comparison between $\widehat{D}_0$ statistics of individuals and feature types as well as their normalization in order to derive a meaningful plausibility measure. Taking the above methodology as a starting point we envision the following application scenarios:

Tag recommendation. Selecting the appropriate tag is a common problem for voluntary mappers. On the one hand, they want to reuse common tags to make sure their POI will be found and rendered. Checking frequency statistics such as taginfo[9] for OSM can be helpful in that regard. On the other hand, contributors

[9] `http://taginfo.openstreetmap.org/`

want to use tags that best describe the corresponding real world entity. This requires browsing and searching the used vocabulary and finally deciding on a tag based on its textual description. With the $\widehat{D}_0$ statistic we can add a criterion such as "which tag is plausible at a certain location?"

Plausibility values for different feature types and arbitrary locations can be ranked to suggest tags by comparing their interaction signature with the local environment. Based on the assumption that the underlying dataset is of reasonable quality, users will more likely select tags from the head of the ranking than from its tail. Hence, the second-order analysis supports mappers by reducing the semantic search space.

Data cleaning. Plausibility can also be applied for cleaning up existing data. Cases of very low plausibility may be forwarded to editors who can check for duplicates or vandalism. For example, a post office tagged next to another post office may be assigned a very low plausibility value, because of its high positive $\widehat{D}_0$ value in the near and similar spectrum (cp. section 7.1). It is more likely that a mapper tagged the very same post office a second time. Fig. 7 depicts such an example of duplicate identification for car wash locations.

However, the second-moment measure should be understood as decision support method for users rather than machine processing. There may still be cases of close post offices that are correct, as well as duplicate bars in a neighborhood of bars and nightclubs. The identification and removal of wrong or redundant data can be guided by our measures but requires manual interaction.

Coverage recommendation. In analogy to the reduction of the semantic search space through tag recommendation, voluntary mappers can also be supported by reducing the search space in the literal sense. A tool that identifies areas in which a certain feature type is likely to occur but not present in the dataset could direct the mapping activities of volunteers to areas where coverage strongly differs by feature type. The possibility of making a valuable contribution can thereby be assessed beforehand. The scenario presented in fig. 7 can be considered the result of coverage recommendation even though such service would rather point to an area in the vicinity of the green dot than its exact location.

Using the $\widehat{D}_0$ statistic for coverage recommendation needs to cope with two problems though. Firstly the influence of a feature type c_k on its own interaction signature must be eliminated. Otherwise high plausibilities can only be expected in the border regions of the area where c_k is actually present. Secondly the second-order analysis models a point process, in contrast to the result of a coverage recommendation, which would be an area. Therefor a sampling of test locations is needed that accounts for the density of instances of c_k itself.

Uncovering implicit partonomy. Given the huge amount of data, it becomes difficult to evaluate how voluntary mappers tag specific locations in comparisons to others, i.e., what users tag in contrast to what they mean, in an aggregated manner. For example, POI tagged as `school` could either represent school

grounds or school buildings. In the latter case it is likely that the regular spacing on city-scale is accompanied by a strong clustering on the city-block-scale (because several buildings jointly form the complex which is commonly considered a school). Whereas tag usage (in this example) could solely be revealed by Ripley's K, a spatial-semantic interaction model is required as soon as different but similar types of schools, e.g., boarding school, public school, or elementary school are present. $\widehat{D}_0$ plots can uncover implicit partonomic assumptions that should be made explicit by either proposing a new tag to the community or providing better descriptions to be considered by mappers in the future. In the car wash example (cp. fig. 7) the $\widehat{D}_0$ plot shows no clustering with similar POI. The red dot can be identified as a duplicate, because car wash facilities are not modeled as building complexes by VGI contributors.

8 Conclusion

In this paper we describe a methodology to characterize the spatial-semantic interaction of points of interest in OpenStreetMap. Inspired by Diggle's [19] second-moment spatio-temporal measure, we combine point pattern analysis as originally proposed by Ripley's [16] with semantic similarity. The resulting spatial-semantic interaction is a measure for the likelihood of features of a certain type to co-occur within a certain semantic and spatial range. The feature type similarities required for our work are not computed from top-down geo-ontologies, but automatically generated bottom-up based on the change history of OpenStreetMap elements. Our methodology sets the theoretical ground for tools to support users in contributing and cleaning up VGI. Users contributing new features may get automatic feature type recommendations based on the location of the new POI and the spatial-semantic interaction within its vicinity. Features that are unlikely to co-occur with other features may be discovered and forwarded to editors.

At the same time, our work has implications on geospatial semantics research in general and geo-ontologies in specific. Instead of aiming at top-down domain ontologies that describe feature types such as pubs by characteristics like having tables, walls, or menus, we argue for a local, bottom-up approach based on their spatial and temporal characteristics. Pubs clump together with other features such as nightclubs or cafés and while they may have different opening hours, they are between those of cafés and nightclubs. Both approaches do not contradict and should be combined. However, it is rather space and time that shape our conceptualization of the world than bags of attributes [29]. As a long-term vision, by examining patterns of spatial-semantic (and temporal [30]) interaction, we aim at extracting prototypical properties of particular feature types, in order to generate unique *semantic signatures*.

Besides integrating the temporal component as well, future work will especially focus on more formal methodologies for validating our results in terms of statistical significance (cp. Diggle's U and residual statistics [19]) and sampling distributions. Our approach can also be improved by resources such as

WordNet[10] to disambiguate and map terms from different repositories containing user-generated *bags of words* and the inclusion of data from location-based social networks like foursquare[11] or whrrl[12].

Acknowledgements. The authors would like to thank Edzer Pebesma, Ashton Shortridge, Ola Ahlqvist, Mike Goodchild, and Peifeng Yin for their fruitful feedback and advice.

References

1. Goodchild, M.F.: Citizens as sensors: the world of volunteered geography. GeoJournal 69(4), 211–221 (2007)
2. Zielstra, D., Zipf, A.: A comparative study of proprietary geodata and volunteered geographic information for Germany. In: 13th AGILE International Conference on Geographic Information Science (2010)
3. Mooney, P., Corcoran, P., Winstanley, A.C.: Towards quality metrics for openstreetmap. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 514–517. ACM, New York (2010)
4. Goodchild, M.F., Glennon, J.A.: Crowdsourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth 3(3), 231–241 (2010)
5. Scheider, S., Possin, J.: Affordance-based algorithms for categorization of road network data. Technical report. University of Münster, Germany (2010)
6. Werder, S., Kieler, B., Sester, M.: Semi-automatic interpretation of buildings and settlement areas in user-generated spatial data. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2010, pp. 330–339. ACM, New York (2010)
7. Trame, J., Keßler, C.: Exploring the lineage of volunteered geographic information with heat maps. In: GeoViz 2011, Hamburg, Germany (2011)
8. Janowicz, K., Keßler, C., Schwarz, M., Wilkes, M., Panov, I., Espeter, M., Bäumer, B.: Algorithm, Implementation and Application of the SIM-DL Similarity Server. In: Fonseca, F.T., Rodríguez, A., Levashkin, S. (eds.) GeoS 2007. LNCS, vol. 4853, pp. 128–145. Springer, Heidelberg (2007)
9. Zook, M., Graham, M., Shelton, T., Gorman, S.: Volunteered geographic information and crowdsourcing disaster relief: A case study of the haitian earthquake. World Medical & Health Policy 2(2), 231–241 (2010)
10. O'Sullivan, D., Unwin, D.: Geographic Information Analysis. Wiley, Chichester (2010)
11. Kuhn, W.: Volunteered geographic information and GIScience. In: NCGIA, UC Santa Barbara, pp. 13–14 (2007)
12. Elwood, S.: Geographic information science: emerging research on the societal implications of the geospatial web. Progress in Human Geography 34(3), 349–357 (2010)
13. Coleman, D., Georgiadou, Y., Labonte, J.: Volunteered Geographic Information: the nature and motivation of produsers. International Journal of Spatial Data Infrastructures Research 4, 332–358 (2009)

[10] `http://wordnet.princeton.edu/`
[11] `http://www.foursquare.com/`
[12] `http://whrrl.com/`

14. Ahlqvist, O., Shortridge, A.: Characterizing land cover structure with semantic variograms. In: Progress in Spatial Data Handling, pp. 401–415 (2006)
15. Cressie, N.: Statistics for Spatial Data (Wiley Series in Probability and Statistics). Wiley-Interscience, Hoboken (1993)
16. Ripley, B.: The second-order analysis of stationary point processes. Journal of Applied Probability 13(2), 255–266 (1976)
17. Daley, D., Vere-Jones, D.: An introduction to the theory of point processes. Springer Series in Statistics (1988)
18. Besag, J.: Contribution to the discussion of Dr. Ripley's paper. JR Stat. Soc. B 39, 193–195 (1977)
19. Diggle, P., Chetwynd, A., Häggkvist, R., Morris, S.: Second-order analysis of space-time clustering. Statistical Methods in Medical Research 4(2), 124 (1995)
20. Rodríguez, A., Egenhofer, M.: Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. International Journal of Geographical Information Science 18(3), 229–256 (2004)
21. Li, B., Fonseca, F.: Tdd - a comprehensive model for qualitative spatial similarity assessment. Spatial Cognition and Computation 6(1), 31–62 (2006)
22. Raubal, M., Adams, A.: The semantic web needs more cognition. Semantic Web Journal 1(1-2), 69–74 (2010)
23. Schwering, A., Raubal, M.: Spatial relations for semantic similarity measurement. In: Akoka, J., Liddle, S.W., Song, I.Y., Bertolotto, M., Comyn-Wattiau, I., van den Heuvel, W.J., Kolp, M., Trujillo, J., Kop, C., Mayr, H. (eds.) ER Workshops 2005. LNCS, vol. 3770, pp. 259–269. Springer, Heidelberg (2005)
24. Van Eck, N.J., Waltman, L.: How to normalize cooccurrence data? An analysis of some well-known similarity measures. Journal of the American Society for Information Science and Technology 60, 1635–1651 (2009)
25. Peters, H.P.F., van Raan, A.F.J.: Co-word-based science maps of chemical engineering. part i: Representations by direct multidimensional scaling. Research Policy 22(1), 23–45 (1993)
26. Rip, A., Courtial, J.: Co-word maps of biotechnology: An example of cognitive scientometrics. Scientometrics 6, 381–400 (1984), doi:10.1007/BF02025827
27. Zitt, M., Bassecoulard, E., Okubo, Y.: Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. Scientometrics 47, 627–657 (2000), doi:10.1023/A:1005632319799
28. Goovaerts, P.: Geostatistics for natural resources evaluation. Oxford University Press, USA (1997)
29. Janowicz, K.: The role of space and time for knowledge organization on the semantic web. Semantic Web Journal 1(1-2), 25–32 (2010)
30. Ye, M., Shou, D., Lee, W.C., Yin, P., Janowicz, K.: On the semantic annotation of places in location-based social networks. In: ACM SIGKDD (forthcoming, 2011)