Transactions in GIS

Research Article

Transactions in GIS, 2014, 18(6): 852–876

Exploring and Visualizing Differences in Geographic and Linguistic Web Coverage

Ramya Venkateswaran, Robert Weibel and Ross S. Purves

Department of Geography, University of Zurich

Abstract

This article reports on a study performed to understand the geographic and linguistic coverage of web resources, focusing on the example of tourism-related themes in Switzerland. Search engine queries of web documents were used to gather counts for phrases in four different languages. The study focused on selected populated places and tourist attractions in Switzerland from three gazetteer datasets: topographic gazetteer data from the Swiss national mapping agency (SwissTopo); POI data from a commercial data provider (Tele Atlas) and user generated geographic content (geonames.org). The web counts illustrate the geographic extent and trends of web coverage of tourism for different languages. Results show that coverage for local languages, i.e. German, French and Italian, is more strongly related to the region of the spoken language. Correlation of the web counts to typical tourism indicators, e.g. population and number of hotel nights rented per year, are also computed and compared.

1 Introduction

The use of web content, both in the form of unstructured text and objects with explicit georeferencing, is an increasingly popular way of exploring a wide range of geographic questions (Egenhofer 2002; Leidner and Lieberman 2011; Jones and Purves 2008; Purves 2011). However, it is very unlikely that web content is evenly distributed in space, and studies which seek to draw conclusions based on, for example, variations in density, must first estimate the underlying density of the collection of interest. Implicit assumptions about homogeneity of coverage can be misleading, and Pasley et al. (2008) set out to explore how web coverage varied for different forms of social media in the UK, correlating coverage of a variety of sites with overall web coverage and population.

The main contribution of this article is use of the web counting method (Kilgarriff and Grefenstette 2003) to create a reproducible method for examining the web coverage and its variations due to language. We do this by counting the individual number of web pages returned by a search engine based on tourism search queries containing toponyms (place names) obtained from gazetteers. This is a continuation of previous work (Venkateswaran 2010) and we discuss web coverage in detail, exploring how it is affected by different influencing factors. Importantly, this article focuses on examining the coverage in unstructured textual information such as web documents, rather than in structured databases. For the remainder of the article, the term web count is used to refer to the number of web documents that are matched to an input query by a web search engine. As an application domain and study area,

Address for correspondence: Ramya Venkateswaran, Department of Geography, University of Zurich, Winterthurerstr. 190 Zurich, Switzerland. E-mail: ramya.venkateswaran@geo.uzh.ch

Acknowledgements: The research reported in this article represents a part of the PhD project of the first author. Funding by the Swiss National Science Foundation through project Generalization for Portrayal in Web and Wireless Mapping (GenW2 +) (SNF No. 200020–138109) is gratefully acknowledged.

 doi: 10.1111/tgis.12071

we use tourism in Switzerland, since tourism is an important factor in the Swiss economy, and is often used as a prototypical application for the utilization of web content. Switzerland is a multilingual country and is visited by tourists speaking many different languages. Therefore, to complete the picture, both geographic web coverage and variation as a function of language are examined. Ad hoc tourist information is readily available on the web in the form of pages that contain news, lists, catalogues, reviews, blogs and multimedia content related to activities targeted to particular regions. Although previous work has shown that web coverage is, unsurprisingly, not homogeneous (Pasley et al. 2008; Venkateswaran 2010), little work has addressed the issue of *how* it varies, beyond obvious relationships to population. For example, Crandall et al. (2009) plot a density map of geotagged Flickr images from all over the world. From this map, one might hypothesize that density of Flickr images correlates with population, Internet connectivity, popularity of Flickr as a social media service, popularity of a given place due to tourism, or some other explanatory variables. With such hypotheses as a starting point, we examine the correlation between web coverage and possible predictor variables that could be used to explain it. We choose to investigate these questions through a case study, since we assert that detailed local knowledge (in our case of Switzerland) is necessary to analyze and discuss the spatial patterns and geographic relationships identified in work of this nature. The key questions driving this research are therefore:

1. How does the geographic distribution of web coverage for tourism-related themes vary across Switzerland?
2. Are there any differences in web coverage distribution for different languages and gazetteer datasets?
3. How do factors such as population and touristic popularity of a place affect web coverage?

Our intention is to develop a simple, repeatable method which allows us to explore web coverage. Such maps of coverage can then be used as baselines to explore variation in coverage (either in time or space), rather than simply assuming either homogeneity of coverage or that coverage simply varies as a function of population.

The article is structured as follows. Section 2 reviews related work, while Section 3 describes the methods that were used to gather web counts and provides details on the different datasets used. Sections 4 to 6 then present the results of several analyzes exploring the above three research questions. In Section 7 we discuss our results in light of these research questions, and in Section 8 we conclude with some possible extensions for future work.

2 Related Work

Geographic information is widely available on the web both in the form of unstructured text and georeferenced multimedia content with associated descriptive information (e.g. Flickr images with image tags). This information can be mined and analyzed by a wide variety of techniques, with a crucial difference being whether references to location are explicitly linked to a unique location (i.e. through coordinates and a reference system), or contain potential ambiguity in the form of a toponym (Hill 2006; Leidner and Lieberman 2011).

2.1 Toponym Recognition and Resolution

The web is an important source of geographical information. For example, Hill (2006) estimates that up to 70% of text documents contain placename references, while Sanderson and

Kohler (2004) suggested that 13–15% of all search engine queries contained place names or some kind of geographic term. Thus, the unambiguous identification of toponyms in text and the assignment of a unique set of coordinates is a key task (Leidner and Lieberman 2011).

Geoparsing involves identifying and disambiguating place names or toponyms in a corpus of text that is part of unstructured content (Leidner and Lieberman 2011). Geoparsing, can be achieved through simple gazetteer lookups (Hill 2006), rule-based methods applied in natural language (Cunningham et al. 2002), and/or machine learning (Leidner 2007). Geocoding, on the other hand, is the process of assigning unique geographic identifiers, usually coordinates, to toponyms that have been extracted from unstructured content in the geoparsing step. One of the main issues with the process of geoparsing and geocoding is performing this process automatically and unambiguously, as all toponyms are not uniquely named. Toponym ambiguity is a special case of word sense ambiguity, a term commonly used in computational linguistics, for a word with more than one meaning. In the case of toponyms, this must be resolved in order to identify and ground toponyms uniquely. Amitay et al. (2004), explains that a geo/non-geo ambiguity arises if the place name has a non-geographic meaning, such as Washington as a place vs. Washington as the name of a person, while a geo/geo ambiguity arises if there exist two distinct places with the same name (e.g. London UK vs. London Ontario). A wide variety of methods are used in dealing with toponym ambiguity (Buscaldi 2011), ranging from simple default rule-based methods based on, for example, population (Rauch et al. 2003; Zong et al. 2005), through methods based on exploiting toponym hierarchies (e.g. Buscaldi and Rosso 2008) to context-based disambiguation (e.g. Overell and Rüger 2008).

2.2 Web Counting and Web Coverage

In this research, the web counting process includes the formation of a search phrase, which is passed to a search engine and the result, the count, is gathered. This is done for four different languages, thereby generating four sets of web counts, one for each language. Some examples of these phrases are discussed in the section that discusses the approach (Section 3.1). As discussed earlier, this count is the number of documents that the search engine indicates as a match for the search phrase. Pasley et al. (2008) used counts as a proxy for coverage, and thus density, by retrieving the total number of occurrences of documents with a given toponym from a search engine index via an API (Application Programming Interface). Web counts have been used in a variety of other studies, typically in information retrieval and search, as well as for web statistics. Keller and Lapata (2003) and Lapata and Keller (2005) used web counts to investigate the performance of web-based models for several natural language processing (NLP) tasks and to approximate bigram counts. Web counts were also used to estimate the size of the web through English search queries (Kilgarriff and Grefenstette 2003). However, web counts are not a perfect estimate of what really exists. The coverage of different search engine collections and their individual methods for approximating the number of web pages matched, often introduce biases in the results. These issues are discussed in the next section.

There are other approaches to analyzing a text corpus, in addition to web counts. For example, Hecht and Gergle (2010a, b) measured the diversity of the Wikipedia corpus in 25 different languages by counting the concepts that were included and the ways in which these concepts were described. In Volk's (2009) work on the Text + Berg project, he accumulated counts for occurrences of mountain names from the yearbooks of the Swiss Alpine Club over a period of time to mine mountain names.

In more specifically geographic applications, Tezuka et al. (2004) calculated the cognitive significance of landmarks using the number of documents collected from the web. By using trigger phrases, such as "hotels in XX" to retrieve web counts, it is possible to detect and identify candidate place names in web documents, and thus identify instances in which a named entity refers to a place (Twaroch et al. 2008).

Many researchers realized the problems of uneven data coverage. For example Graham et al. (2012) report through their cartograms the digital divide in the geography of the Internet by examining the raw number of Internet users in each country as well as the percentage of the population with Internet access. They later examine georeferenced tweets produced by Twitter users all over the world and plot a spatial tree map (Graham et al. 2013). This map clearly shows the inequality in the geography of content. Li et al. (2013) use georeferenced Twitter and Flickr data to derive patterns rather than using only one of them, as they acknowledge that there is uneven distribution of the data generated in social media and the nature of such data has to be understood and used appropriately. All the above work suggests that web content is not homogeneous and varies due to a variety of reasons.

3 Methods

3.1 Approach

This section introduces methods that were used to establish the geographic and linguistic web coverage for tourism in Switzerland. From previous work (Venkateswaran 2010) we have first results suggesting that web content is linked to and varies with language. In this article, however, we go deeper and study other factors that may affect web coverage. Since the coverage problem focuses on tourism-related themes in Switzerland, it is essential first to understand the linguistic background of Switzerland. Switzerland has four official languages: German, French, Italian and Romansh. According to the Swiss Federal Office of Statistics (2000 Swiss census) the number of native speakers is approximately 64% for German (all dialects), 20% for French, 6.5% for Italian and 0.5% for Romansh. As discussed earlier, our aim was to examine tourism-related phrases in different languages, especially those important in the context of Switzerland. Although English is not one of the national languages of Switzerland, it is an important language with respect to tourism in Switzerland. Therefore, English was also selected as one of the languages. Given the proportionally low number of Romansh speakers, Romansh was not selected for this study. Web counts were therefore examined in German, French, Italian, and English.

In order to gather web counts, the following trigger phrases: <"*Toponym*" Schweiz tourismus>, <"*Toponym*" Suisse tourisme>, <"*Toponym*" Svizzera turismo> and <"*Toponym*" Switzerland tourism> were used. These phrases were made up of a toponym, followed by translations of Switzerland and tourism into the four different languages. The toponyms were selected from three datasets, two of which contained names of populated places and points of interest from the third. An example of a search query is thus <"La Chaux-de-Fonds" Switzerland tourism>. Toponyms were placed in quotes so that only exact matches were found, in particular for toponyms which were made up of multiple words. The phrases were selected from initial testing with a combination of the toponyms with canton names (Switzerland is a federal state made up of 26 cantons), country- and tourism- related terms such as "attractions", "places to visit", etc. The country along with the keyword "tourism" seemed to yield the highest web counts. Furthermore, work by Hollenstein and Purves (2010), reports that tourists are more likely to tag photographs on Flickr as a combination of a town or city name and country rather than higher level administrative units such as state or canton. We assume that this behavior might be replicated in other web content.

In the case of the point of interest (POI) data, specifically related to tourism, the word "tourism" and its translations in the three other languages were omitted from the search phrase. This is because most POIs were typical tourist locations, hence it could be assumed that the toponym was directly related to tourism.

To determine the number of hits (denoted as web counts in the remainder of the article) we used the Yahoo! Search BOSS API (http://developer.yahoo.com/search/boss) for the above sets of queries. The API has a wide variety of parameters that can be supplied, thereby influencing results. For instance, the type of the web content can be specified using the *type* parameter, including specifying the format of the documents that match the search query, for instance html, text, pdf, doc, etc. Since our main aim was to study the aggregate coverage, we concluded that the format of the web content did not matter and that any web page that contained these terms was a candidate contributing to the web count. Another instance is the query operator. Boolean operators like "AND" and "OR" can be used to combine query words, and hence could be used in the search phrases (discussed above). However, we found out that there was no significant change in the web count with or without the AND operator, hence we did not make use of it. Furthermore, the search was not restricted to the top level domain ".ch", since many tourism websites are hosted under ".com". Finally, as locale we used the default "en-us", since preliminary experiments had shown that many tourism websites use this locale rather than the local locale (e.g. "de-ch", "fr-ch", etc.).

The counts were extracted in February 2010 and this cache of counts has been analyzed further in the research. The API returns *totalhits* and *deephits*. Both these values are approximate counts of the number of web documents that exist as, firstly, the Yahoo! Search BOSS API returns only a smaller proportion or a snapshot of the web, instead of all web documents and, secondly, the number of hits returned is an approximation based on proprietary code. *Totalhits* does not contain duplicates while *deephits* reflects duplicate documents and all documents from a host. Hence, we selected *totalhits* as the web count for our study.

3.2 Toponym Data

The toponyms for the search phrase were taken from the following three datasets: SwissNames, the Tele Atlas Points of Interest (POI) dataset and the GeoNames gazetteer dataset. SwissNames is provided by the Swiss Federal Office of Topography (swisstopo; http://www.swisstopo.admin.ch/). The dataset contains 155,571 place names in 62 categories shown on the swisstopo 1:25,000 map, and contains other essential pieces of information such as coordinates, altitude, "Gemeinde" (commune) name and canton name. The POI dataset was provided by Tele Atlas BV 2010 (see http://www.teleatlas.com/index.htm for additional details). It contains 54,912 points of interest in 50 categories. The POIs are attributed with important information including coordinates, name, address and other details. The GeoNames gazetteer is provided online by www.geonames.org. The data for Switzerland contained, at the time of our experiments, 20,726 place names in 107 categories, also known as feature classes. One of the highlights of the dataset is that along with *placenames*, it also lists *asciinames* and *alternatenames*. The *asciinames* restrict spellings to only ASCII letters, while *alternatenames* spell out the place name in a number of other languages. In Section 3.6 we discuss how we made use of this additional information.

The above selection of datasets covers three different types of data sources: Topographic data by a national mapping agency, POI data from a commercial data provider, and, to some extent in the case of GeoNames, user generated geographic content. In the following, we will describe the analyses that make use of the above datasets.

Table 1 SwissNames list of populated places that were selected for the experiment

SwissNames code	Explanation
HGemeinde	city > 50,000 inhabitants
GGemeinde	city 10,000–50,000 inhabitants
MGemeinde	town 2,000–10,000 inhabitants
KGemeinde	village < 2,000 inhabitants
GOrtschaft	large settlement > 2,000 inhabitants
MOrtschaft	middle settlement < 2,000 inhabitants
KOrtschaft	small settlement 50–100 inhabitants

3.3 Settlements from SwissNames

From SwissNames, we selected all the toponyms of populated places: cities, towns, villages and settlements as shown in Table 1. This toponym set contained 7,949 populated places in Switzerland. Out of the 7,949 records, 1,704 places were eliminated because of geo/non-geo ambiguities that caused the counts to be artificially high (see Section 3.7 for more detail on ambiguities). Following this, web counting was achieved using the approach described above.

3.4 Tourist Destinations from Tele Atlas POI

From this dataset, a list of 787 tourist destinations were extracted from the Tele Atlas database, by filtering towns or points of interest that were explicitly marked "Important Tourist Attraction". The web counting approach described above was then performed. No ambiguities were identified, in contrast to the previous analysis with SwissNames.

3.5 Populated places from GeoNames

For the third analysis, names of populated places were extracted from the GeoNames gazetteer. As discussed above, the gazetteer provided information on toponyms, their corresponding feature codes and population. Among all the toponyms, only toponyms with feature code 'PPL' and 'PPLA' were chosen. In GeoNames, PPL is a populated place, defined as "a city, town, village, or other agglomeration of buildings where people live and work", while PPLA is a seat of a first-order administrative division. Other populated toponym categories like PPLA2, PPLC, PPLL exist but were not considered for the study, as they were too small (population-wise) or already included in PPL and PPLA. Initially 4,337 entries were selected, of which 412 were deleted due to geo/geo and geo/non-geo ambiguities and another 277 were aggregated (cf. Section 3.7 for the method) and then deleted due to repetitions.

3.6 Modifications to Toponyms in SwissNames and GeoNames

After the three sets of toponyms were selected from the three different datasets, some translations and changes in the spelling were made. These modifications were performed on the first two sets of toponyms; settlements (SwissNames) and populated places (GeoNames) (Table 4), as the web counts could be slightly skewed or biased for several reasons, such as:

- The datasets contain toponyms that are in the local language. For instance, the towns in the French-speaking part of Switzerland are in French (e.g. Genève) and towns in the German speaking part of Switzerland are in German (e.g. Zürich). This skews the search results and in turn the web counts, as the local name may not be used in a website of a different language, hence not reflecting the real nature of the coverage. Therefore, these web counts were also examined after translating the toponyms to the particular language of examination (e.g. Geneva for English and Zurigo for Italian). All the translated names of the toponyms were extracted from Wikipedia using WikAPIdia (http://collablab.northwestern.edu/wikapidia_api/Wikapidia/Home.html).
- Occurrences of diacritics such as "ö", "é", "è", etc. in a toponym are highly language-specific. The content on the web in a particular language often does not contain toponyms with the special characters of another language. For instance, 'Zürich' is spelt as 'Zurich' in English and French, causing the counts to be skewed, as the search phrase <"Zürich" Switzerland tourism> does not appear as frequently as <"Zurich" Switzerland tourism> in English and French web pages. A preliminary examination caused the number of counts to drastically increase to 111,139 for "Zurich", as compared with 28,227 for "Zürich" with English search terms. Hence, on the basis of this observation, web counts were also examined by considering the toponyms in the ASCII form, after replacing any non-ASCII character with its respective ASCII character (for example "ü" with "u", "è" with "e", etc.).
- Some toponyms in Switzerland are spelt with another "e" to replace the *umlaut* diacritic (¨) in the German spelling (Table 2). Thus "ü" is replaced by "ue", and "ä" is replaced by "ae". This was also applied to the toponym set and counts were examined again.

If a toponym did not have a translation or diacritic, then the (unchanged) web count for the original toponym name was considered. Table 3 shows a case-by-case example of how web counts change depending on whether the toponym is in the local language, taken in ASCII format, spelling changed, or translated.

Table 2 Examples of spelling changed toponyms (only for English)

Name	Changed spelling
Zürich	Zuerich
Graubünden	Graubuenden
Grächen	Graechen

Table 3 Comparison between original, ASCII-converted, spelling changed and translated toponyms

Original counts	Higher counts	What was higher?
Glarus Suisse tourisme **~3,000**	Glaris Suisse tourisme **~20,000**	Translated toponym into French (from German)
Neuchâtel Switzerland tourism **~20,000**	Neuchatel Switzerland tourism **~30,000**	ASCII toponyms (instead of French spelling)
Zürich Switzerland tourism **~30,000**	Zuerich Switzerland tourism **~35,000**	Toponym without diacritic (from German)

Table 4 Final list of toponym sets

Analyzes	Sets containing
Settlements from SwissNames	(1) Original toponyms; (2) translated toponyms; (3) ASCII toponyms; (4) toponyms without diacritic
Tourist destinations from Tele Atlas POI	(5) Original toponyms
Populated places from GeoNames	(6) Original toponyms; (7) translated toponyms; (8) ASCII toponyms; (9) toponyms without diacritic

Table 5 Ambiguities in toponyms. Numbers in bold typeface denote final values chosen as explained above

Toponym	Coordinates	Population	German web counts
Aesch	47.47104, 7.5973	**10,138**	3,791
Aesch	47.26667, 8.25	911	3,791
Aesch	46.88333, 8.8	0	**3,988**

Having carried out all of these operations, a final set of nine toponyms datasets was generated (Table 4). For each of these nine toponym sets, the web counts were generated in the four languages, amounting to a total of 36 individual runs.

3.7 Toponym Ambiguities

We chose to disambiguate geo/geo ambiguous toponyms using a simple, but effective metric, i.e. population, which results in a one sense per discourse representation (Rauch et al. 2003). This strategy should work in most cases, although if the ambiguous toponym is a tourist destination with few permanent inhabitants it may fail. Table 5 shows an example, where "Aesch" is treated as a geo/geo ambiguity and the duplicate entries were deleted by the procedure explained above. Table 5 also shows another effect, visible in the last column. Occasionally, toponyms shared the same name but had different web counts. While this may seem surprising, the difference can be attributed to cache updates that might have happened on Yahoo! at any point during a processing run. Hence, as the final count, the highest web count was selected for the set of toponyms that had a common name.

In the case of geo/non-geo ambiguities, we prepared stop word lists of common words and commonly used geographic terms such as "berg" (mountain in German), "stein" (stone in German), etc., in four languages. Toponyms with these names were automatically deleted and not examined. We also used simple methods such as comparing the web counts with population and found several toponyms with an extremely high web count but a very low population count. With the help of local knowledge, we found many of these toponyms were geo/non-geo ambiguities and, as for the previous cases, they were deleted from our list. Finally, we manually went through the list of the top 100 web counts and deleted all the geo/non-geo ambiguities identified for all four languages. Table 6 shows some examples of typically occurring toponym ambiguities, along with the number of times they appeared in the SwissNames

Table 6 Ambiguities in toponyms. Language wise typically occurring top five toponym ambiguities

Toponym (German)	No. of times	Meaning in English	Toponym (French)	No. of times	Meaning in English	Toponym (Italian)	No. of times	Meaning in English
Alle	1	All	Au	9	To	Del	1	The
Platz	2	Place	Nord	2	North	Alle	1	To
Markt	1	Market	Plan	2	Map	Stampa	1	Print
Bild	2	Picture	Mon	1	Mine	Nord	2	North
Berg	11	Mountain	Premier	1	First	Valle	1	Valley

Table 7 Top 10 language wise Web counts pre-filtered for toponym ambiguities

Toponym (German)	Web count	Toponym (English)	Web count	Toponym (French)	Web count	Toponym (Italian)	Web count
Alle	404,649	First	1,252,161	Au	1,219,563	Del	392,964
Platz	241,955	Costa	1,126,839	Nord	682,306	Alle	214,242
Markt	229,388	Full	821,796	Plan	635,139	Stampa	149,994
Bild	211,901	Sales	582,168	Mon	454,508	Nord	120,429
Berg	210,786	Plan	548,689	Premier	381,920	Valle	97,877
Buch	183,385	Far	413,484	Provence	329,040	Costa	93,269
Ins	154,673	Bissau	375,120	Rue	286,517	Strada	79,388
Schutz	151,891	Seen	314,905	Font	232,269	Far	71,711
Plan	140,621	Play	308,726	Champagne	209,970	Piazza	70,948
Bad	131,255	Says	291,378	Tavers	195,906	Isola	67,997

dataset. Table 7 shows the pre-filtered top 10 web counts in four languages. It is clear that high web counts are dominated by ambiguous uses of toponyms, which typically do not refer to locations, and thus filtering the web counts is important.

4 Results

4.1 Geographic Web Coverage

In our presentation of results, we start with the outcome of analyzes related to Research Question 1, seeking to establish geographic web coverage related to tourism-related themes in Switzerland. Figure 1 shows the resulting web counts for Swiss toponyms. Web counts were sorted individually for each language in decreasing order. That is, the sorting order differs between languages, and thus the graphs suggest the trends and the frequency distribution of the web counts over all toponyms, rather than the specific web counts per individual toponym.

Looking at Figure 1 the general trend seems to be that German has the highest counts overall as well as for many individual locations. This reflects the dominance of German as the most widely spoken language in Switzerland. The counts for Italian, on the other hand, are lowest, again in line with the observation that Italian is less frequently spoken in Switzerland than German and French.

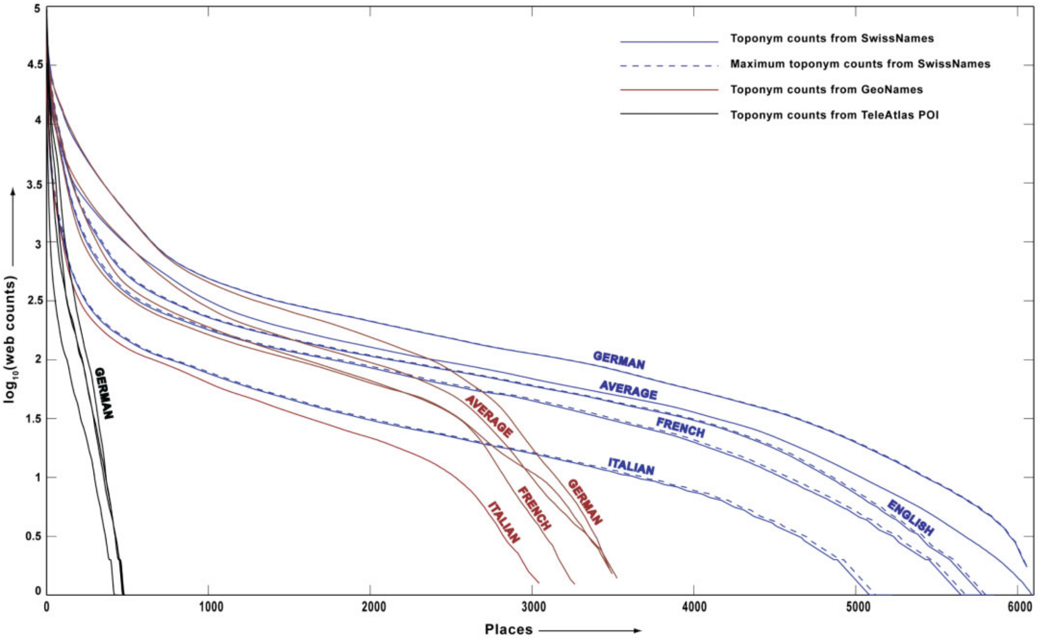

Figure 1 Plot of the web counts vs. places (tourist attractions), plotted on a logarithmic scale with colours reflecting different gazetteer data sources: SwissNames (blue), Tele Atlas POI (black), and GeoNames (red). Dashed line denotes maximum toponym counts resulting in spelling modifications made as explained in Section 3.6. In each colour, the four different lines indicate the four different languages that were chosen for this study

Counts for Tele Atlas POI data were lowest. This is despite the fact that the word "tourism" (and its translations) was omitted from the search phrase used for this gazetteer data set (cf. Section 3.4 for more details on the Tele Atlas dataset), resulting in a less restricted search. This result suggests that the cumulative tourism web content in Switzerland is greater for individual cities than for specific tourism attractions. In other words, city names are typically used when referring to tourism rather than more specific names related to individual attractions. Figure 2 represents a selection from Figure 1, focusing on the results for the SwissNames gazetteer in order to highlight some individual counts. The tags on this graph, through their position, symbolize the approximate value of the web count for selected places. Bern, the capital of Switzerland and also an important tourist destination, is the top ranked place name. The two bar charts in Figure 3 show the resulting web counts from the SwissNames and GeoNames datasets for toponyms. In terms of the web counts, both datasets exhibit similar characteristics, with few toponyms that yielded zero counts for German and English and many zero valued web counts for Italian.

Figures 4 and 5 show coverage maps of Switzerland, using the web counts generated from the SwissNames and GeoNames datasets. In both datasets, we eliminated toponyms that had cumulative web counts equal to 0, though this was rare (cf. Figure 3). Hence, the lowest value for the average web count was 0.25. We also noticed that there were many toponyms for which the web counts were 0 for three languages and high for the fourth language, which happened to be the language spoken in that area. This is due to the fact that many tourism-related toponyms are in the local language (i.e. in German, French and Italian, as opposed to English)

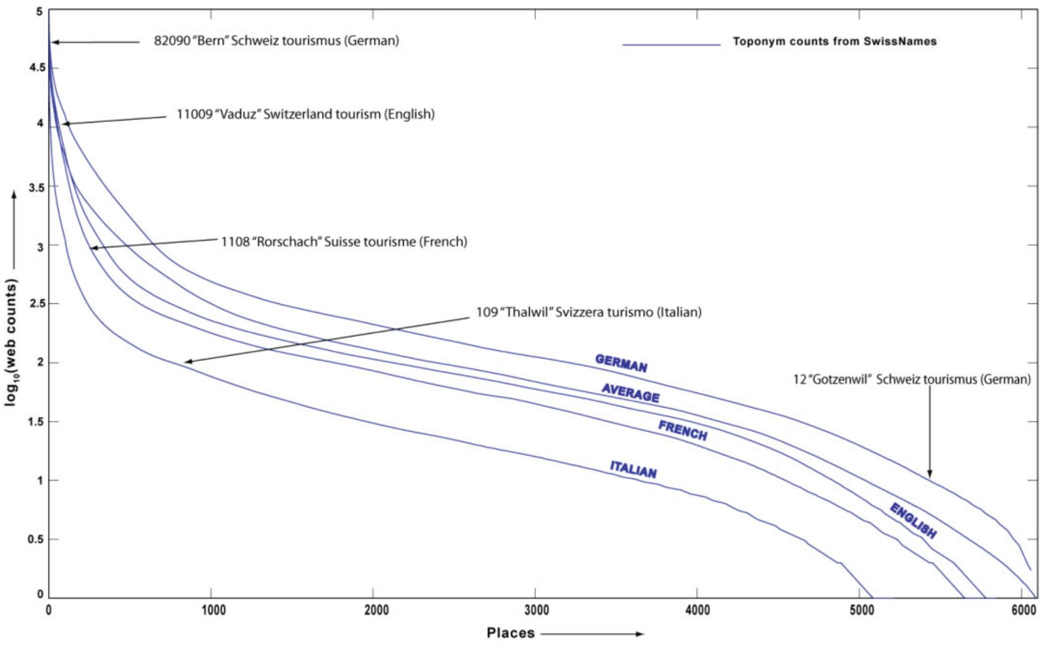

Figure 2 Selection of SwissNames from Figure 1 with place names and its approximate value of the web count for selected places

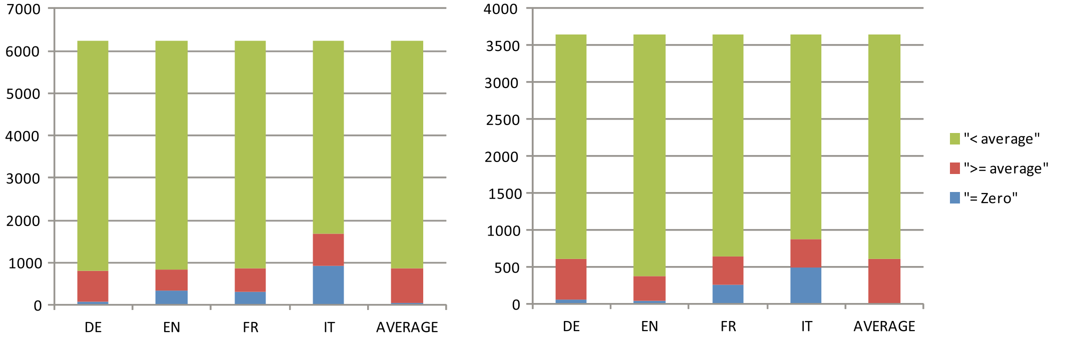

Figure 3 Bar charts showing the web count summary by language for SwissNames (left) and GeoNames (right)

and some have complicated names. Also, many of the entries in the POI dataset relate to transportation tourist attractions such as "Luftseilbahn" (German word for cable car), "Gondelbahn" (German word for gondola lift), "télésiège" (French word for a chair-lift), etc. These names are given in the local language and yielded 0 or very low counts for other languages. A typical example is the "Felsenegg Luftseilbahn", which is an important tourist attraction near Zurich. "Luftseilbahn" is the German word for the cable car to a place called Felsenegg. But English, French and Italian web pages do not use the word "Luftseilbahn", instead they use the corresponding translated word for "Luftseilbahn" (cable car). Since "Luftseilbahn Felsenegg" is the official name, it is rarely found in English, French or Italian web pages but yields a high count in German.

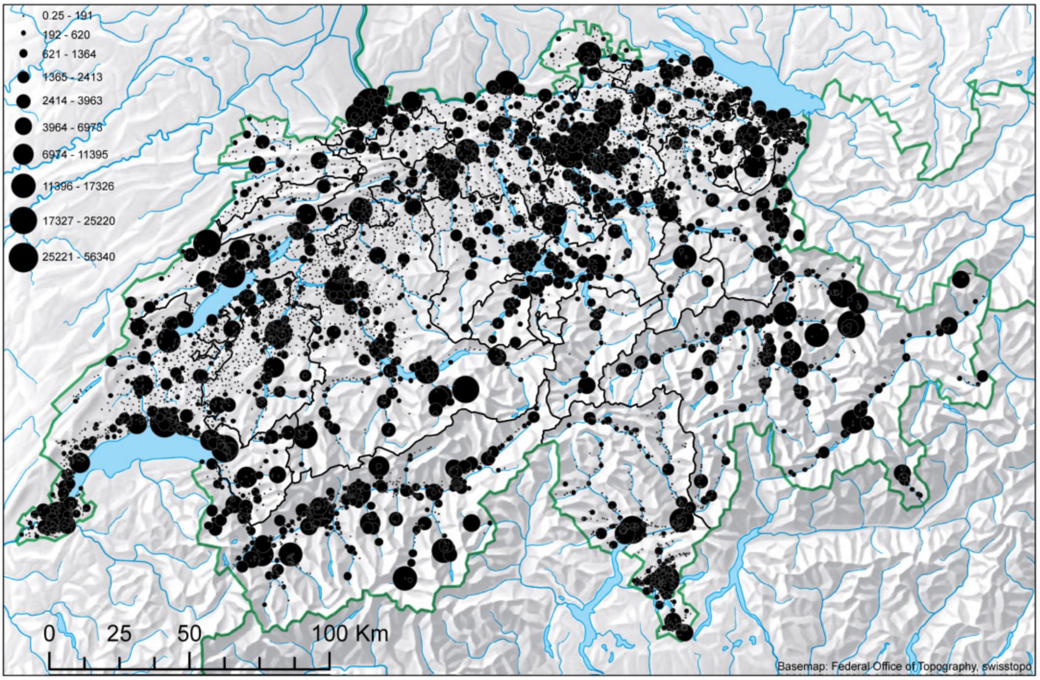

Figure 4 Map showing the geographic coverage by graduated circles of web counts for the toponyms from SwissNames. Size of the circle depends on the average web count of all the four languages, for a given toponym, with legend values decided by the Jenks classification method

Table 8 List of places in SwissNames and Tele Atlas with the counts of highest frequency

SwissNames		Tele Atlas POI	
Frequency of highest counts	Keywords in different languages	Frequency of highest counts	Keywords in different languages
555	German	290	German
144	French	112	French
16	English	78	English
14	Italian	22	Italian
		185	0 counts
729	**Total**	**787**	**Total**

Figure 3 and Table 8 show a comparison between the two datasets. In Figure 3, we attempt to compare the SwissNames and GeoNames datasets. With respect to their toponym content the two lists are quite similar as they have similar 0 values and values that are above and below the average web count (Table 8). To compare the SwissNames and Tele Atlas POI dataset, toponyms of places with more than 2,000 inhabitants were selected. This gave us 729 toponyms that we compared with 787 tourism POIs.

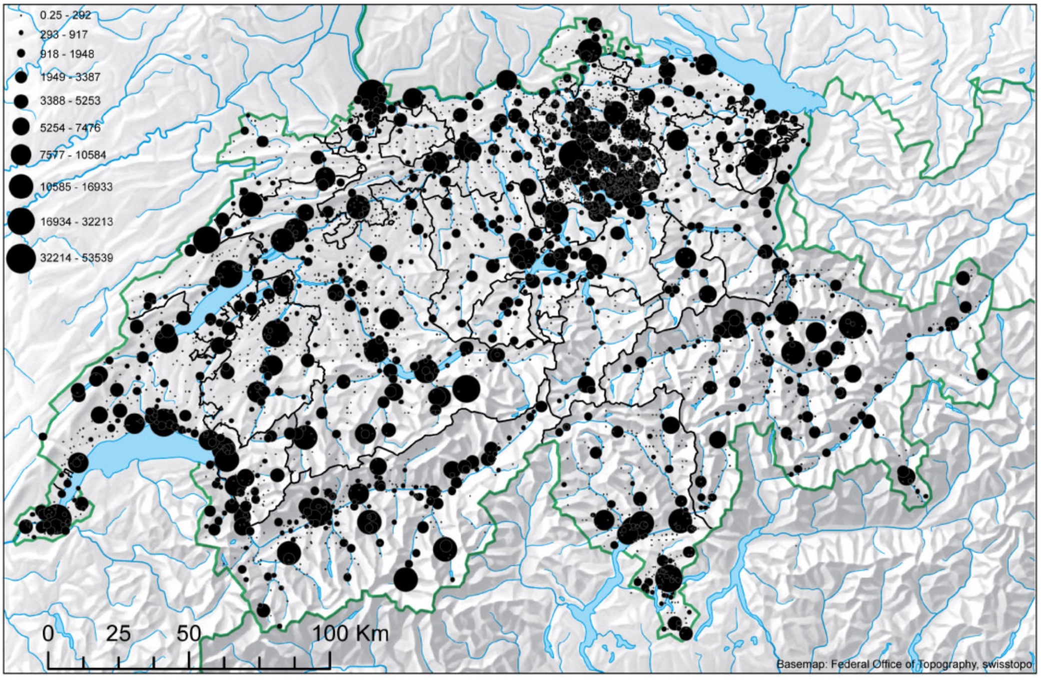

Figure 5 Map showing the geographic coverage by graduated circles of web counts for the toponyms from GeoNames. Size of the circle depends on the average web count of all the four languages, for a given toponym, with legend values decided by the Jenks classification method

Table 8 shows the number of times a count in a certain language was the highest among the three other languages. Out of the subset of records selected, for both datasets German web counts were highest and once again lowest for Italian web counts. There were no web counts that yielded 0 in the SwissNames dataset. Therefore, we can conclude that the two sets are quite similar in the trends of the web counts presented in Figure 3 and Table 8, respectively. The number of overlapping toponyms was 2,621; hence, half of the GeoNames dataset was part of the SwissNames dataset. Thus, for the remainder of the experiments and analyses, only the SwissNames dataset was used, as other datasets seemed to be similar in content or coverage.

4.2 Linguistic Coverage

In this section, we will explore more closely Research Question 2, relating to linguistic differences in web coverage. From the dashed line in Figure 1, it can be seen that changes in spelling and translations do not make a difference in the trends. While this is the case for the overall trends, in the case of toponyms with high web counts, the order changes considerably. Table 9 shows the top 10 web counts with toponyms in original names, along with search phrases in different languages. Table 10, on the other hand, shows maximum web counts selected from search phrases using the original names, ASCII spelled names and translated names. The number of toponyms whose web counts increased is highest for Italian and lowest for German. This is most likely because the German-speaking region of Switzerland is comparatively the

Table 9 Top 10 web counts in four different languages

Toponyms and web counts used with **German** phrases		Toponyms and web counts used with **English** phrases		Toponyms and web counts used with **French** phrases		Toponyms and web counts used with **Italian** phrases	
Zürich	112,074	Basel	64,517	Genève	182,006	Lugano	31,721
Bern	82,090	Bern	53,620	Lausanne	65,637	Locarno	28,749
Basel	72,679	Lausanne	46,037	La Chaux-de-Fonds	53,945	Bellinzona	22,622
Freiburg	50,344	Grindelwald	42,305	Yverdon-les-Bains	38,541	Chiasso	13,736
Luzern	49,000	Zürich	33,534	Neuchâtel	38,084	Mendrisio	13,319
Grindelwald	43,232	Davos	28,232	Montreux	34,444	St. Moritz	11,961
Glarus	42,463	Locarno	26,352	Sion	33,327	Zermatt	11,825
St. Gallen	41,291	Sion	26,213	Morges	25,242	Zürich	10,966
Aarau	40,025	Lugano	25,870	Davos	23,031	Ascona	9,903
La Chaux-de-Fonds	35,223	Genève	24,394	Zürich	20,051	Basel	9,353

Table 10 Top 10 web counts with toponyms showing maximum web counts selected from search phrases using the original names, ASCII spelled names and translated names. Toponyms whose counts changed because of the modified spelling, are in bold typeface

Changes for toponyms and web counts with **German** phrases		Changes for toponyms and web counts with **English** phrases		Changes for toponyms and web counts with **French** phrases		Changes for toponyms and web counts with **Italian** phrases	
Zürich	112,074	**Geneva**	185,819	Genève	182,006	Lugano	31,721
Bern	82,090	**Zurich**	123,715	Lausanne	65,637	Locarno	28,749
Basel	72,679	Basel	64,517	La Chaux-de-Fonds	53,945	**Ginevra**	26,650
Freiburg	50,344	Bern	53,620	**Berne**	49,938	**Zurigo**	26,544
Luzern	49,000	Lausanne	46,037	Zurich	46,146	Bellinzona	22,622
Genf	44,108	Grindelwald	42,305	**Fribourg**	43,603	**Berna**	16,051
Grindelwald	43,232	**Neuchatel**	31,712	**Bâle**	38,575	**Losanna**	15,528
Glarus	42,463	Davos	28,232	Yverdon-les-Bains	38,541	**Basilea**	14,764
St. Gallen	41,291	Locarno	26,352	Neuchâtel	38,084	Chiasso	13,736
Aarau	40,025	Sion	26,213	Montreux	34,444	Mendrisio	13,319

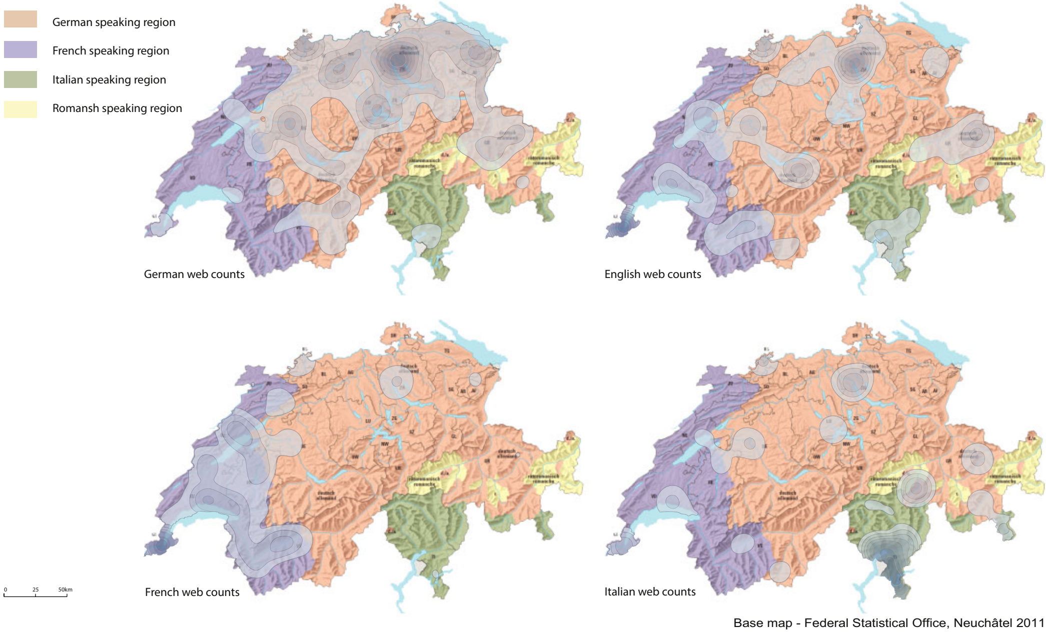

Figure 6 Kernel density map using web counts from GeoNames dataset shown on language region, after toponym spellings were changed

largest and thus has more toponyms than the other language regions, therefore the probability of a toponym occurring in the German-speaking region is high. In turn, many of Switzerland's important places in terms of tourism and population are situated in the German-speaking region. On the other hand, the Italian-speaking region is the smallest and thus has the least toponyms. We see that out of the 10 toponyms in Table 9, five are pushed down the list when translated into Italian (Table 10). Also the toponyms whose counts increased for Italian are not places located in the Italian-speaking regions, but are important populated places in Switzerland, such as Zurich. The above observations suggest that web content seems to have toponyms translated into the language being used in the web page, rather than using the toponym in its local language. For example, Geneva when spelt as Genève yields an English web count of 24,394, which increases to 185,819 when Geneva is used.

To examine the language bias we plotted the kernel density estimate (KDE surfaces) of the web counts in four languages on the language region map of Switzerland. KDE is a useful method for highlighting patterns of overall density distribution in point data. One key parameter that must be chosen for KDE is the bandwidth, or smoothing parameter. The average nearest neighbor distance for a set of points is one indicator for local bandwidth selection (Silverman 1986). For our study, we used the places with the top 50 web counts, as they have the strongest influence on the density distribution because they correspond to the major cities and tourist resorts. For these toponyms, the average nearest neighbor distance – and hence the bandwidth – is approximately 15 km.

In the map (Figure 6) a clear bias of language to region can be seen through web counts that cluster on the German, French and Italian language regions. English web counts, on the other hand, show a dispersed behavior with similar coverage across different regions. For example, in the case of the German web coverage, it is biased to the extent that important

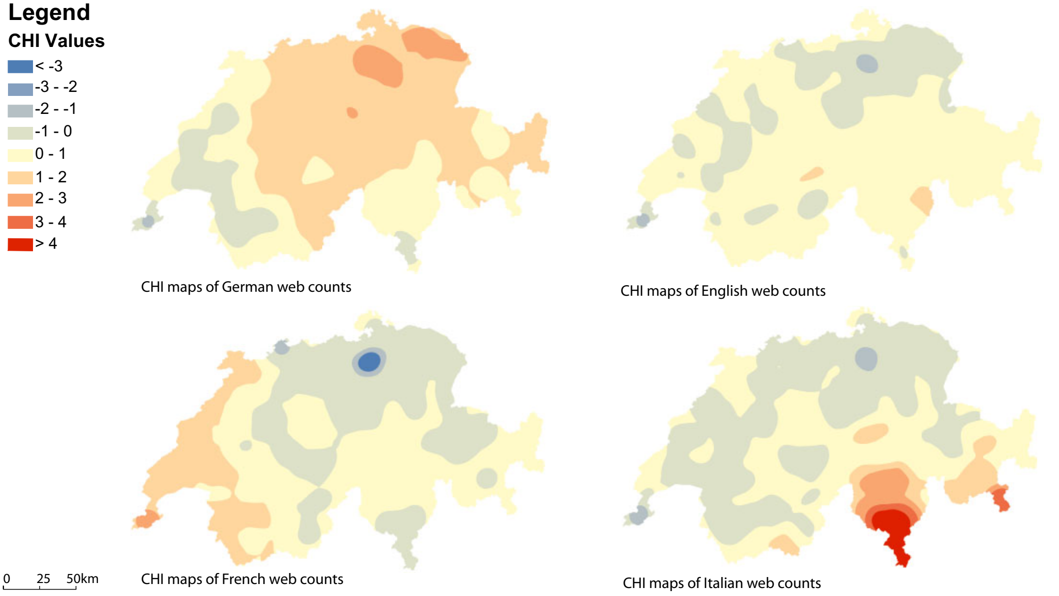

Figure 7 Map showing χ values comparing kernel densities of average and language web counts. Positive χ values (red colours) denote areas where language web counts were higher than average web counts, while negative χ values (blue colours) denote areas where language web counts were lower than average web counts

cities such as Geneva (situated in the French speaking part of Switzerland) and Lugano (situated in the Italian speaking part of Switzerland) are hardly visible. Web counts in the French language also show similar behavior. To better examine this language bias with the language regions, we generated kernel density estimates, visualizing the χ values, calculated by comparing the observed web counts with the average web counts (i.e. the expected number), where the observed web count was the actual web count for any of the languages, and the expected web count was the average web count, again in each of the four languages. This allows us to study differences between the web counts in the four languages, as we can see in Figure 7. χ values were computed as follows:

$$\chi = \frac{(obs - \exp)}{\sqrt{\exp}} \tag{1}$$

Red shaded areas denote positive χ values, which means that the observed value was higher than the expected value, while the blue shaded areas denote negative χ values, meaning that the observed value was lower than the expected value. For French and Italian web counts the red shaded areas coincide with the language regions. For German web counts, a similar though less pronounced pattern can be seen. English, on the other hand, is not only almost uniform throughout Switzerland, but also seems to show lesser contrast in the χ values. This confirms our observation that coverage in English is more evenly distributed than the other languages.

To measure spatial autocorrelation, we computed the Moran's I (Table 11). Moran's I always ranges from −1 to 1 and a value near +1 indicates clustering, while a value near −1 indicates dispersion in the values of a variable. To test for the null hypothesis (no spatial

Table 11 Spatial autocorrelation of language with place (clustered patterns in bold)

	Computed with all points		Computed with points only in the corresponding language region	
Measure	Moran's Index	Z-score	Moran's Index	Z-score
Average	0.005972	**2.402686**	–	–
German	0.0020140	**7.954030**	0.011856	**5.4978**
English	0.003957	1.615825	–	–
French	0.022078	**9.983075**	0.009924	**2.095638**
Italian	0.034569	**14.236829**	–0.002434	–0.074634

autocorrelation), we also calculated a Z-score. A Z-score between 1.96 and –1.96 indicates no statistical significance. Looking at the first part of Table 11 we note that all the points show a clustered pattern except for English. Since the language areas for Italian and French are smaller their Z-scores are very high. To study the spatial autocorrelation in the individual language regions, we extracted three sets of points, by intersecting the toponym points with each of the three language regions. That is, one point set was generated for the German speaking region, a second point set in the French speaking region and a third set in the Italian speaking area of Switzerland. For all points except Italian we see a high Z-score and no pattern of dispersed points.

4.3 Influencing Factors

This section is devoted to Research Question 3, thus establishing the correlation of web coverage with independent variables. We start with an analysis of clusters in the web counts data, in order to gain a better impression of the geographic distribution of web coverage. While Moran's I can give an impression of the global degree of concentration and spatial autocorrelation in a spatial variable, it does not allow local patterns of spatial autocorrelation to be revealed. We therefore used a measure of local spatial autocorrelation, the Getis-Ord G_i^* statistic (Ord and Getis 1995) on the average web counts across all languages. The output of the G_i^* statistic is a Z-score for each point, representing the statistical significance of clustering for a specified distance. Highly positive values denote so-called hot spots, while clusters of highly negative values are termed cold spots. In the maps in Figures 8a and b we can see that there are a several hot spots, but no cold spots. Figure 8a shows the hotspots for the average of all web counts for Switzerland. Figure 8b on the other hand shows the hotspots per individual language, along with the language regions of Switzerland. The hot spots correspond to places such as Zurich, Basel, Bern, Geneva, La Chaux-de-Fonds, Lausanne, Grindelwald, Zermatt, Davos and Lucerne, in effect the top 10 counts when all four languages are considered.

From the kernel density estimation (Figure 6), it was clear that there is a bias towards big cities, irrespective of the language. One potential reason may be that cities have higher populations; hence, becoming centres for hotels, transport, and services related to tourism. To examine this we compared the kernel densities of population and the average web counts in all of the languages, thereby once again generating χ values where we used population density to

(a)

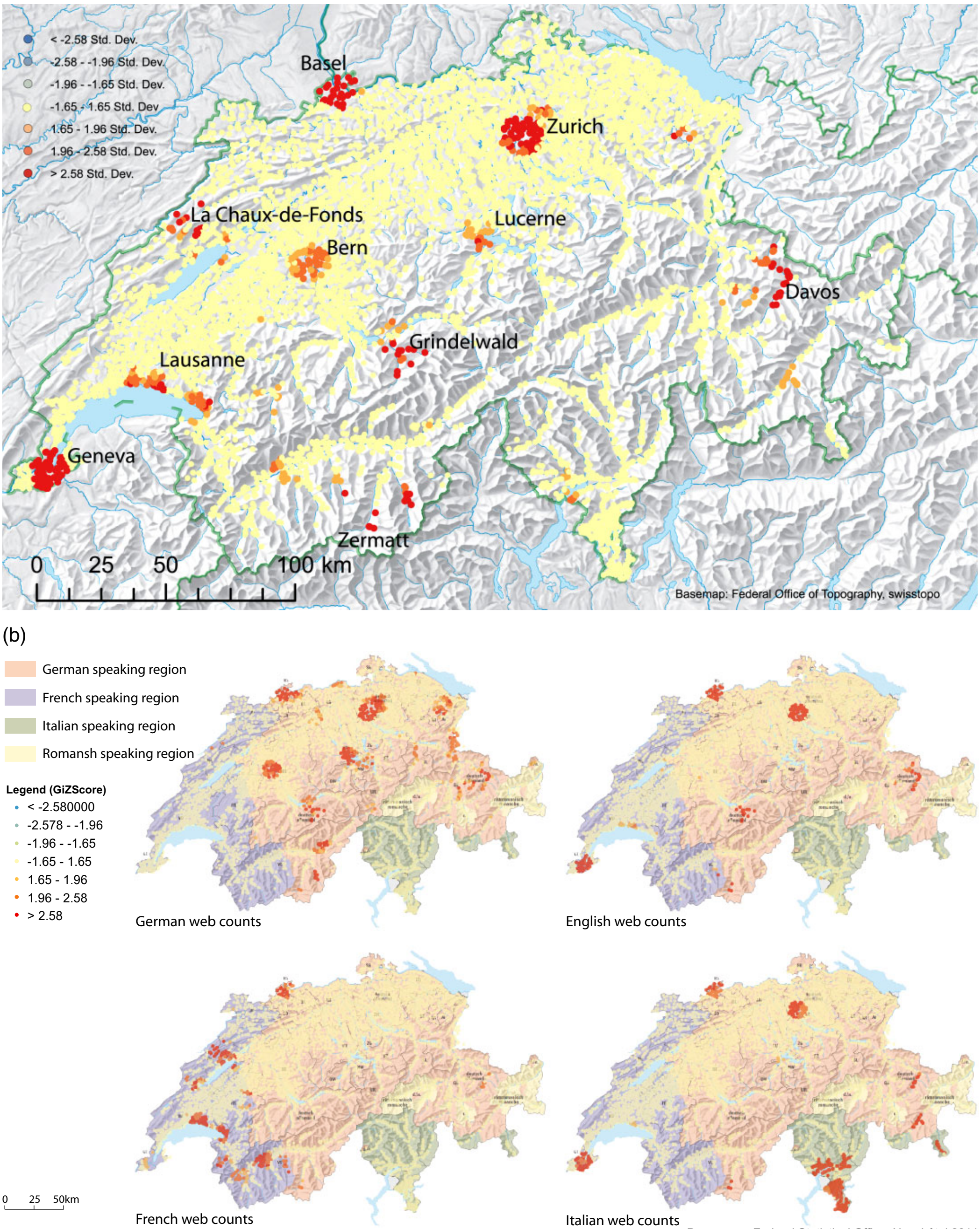

Figure 8 (a) Hotspot analysis of average web counts over all languages. Several hotspots but no cold spots can be seen. Top 10 places are labelled (approximate); and (b) Hotspot analysis of web counts for individual languages. Several hotspots but no cold spots can be seen

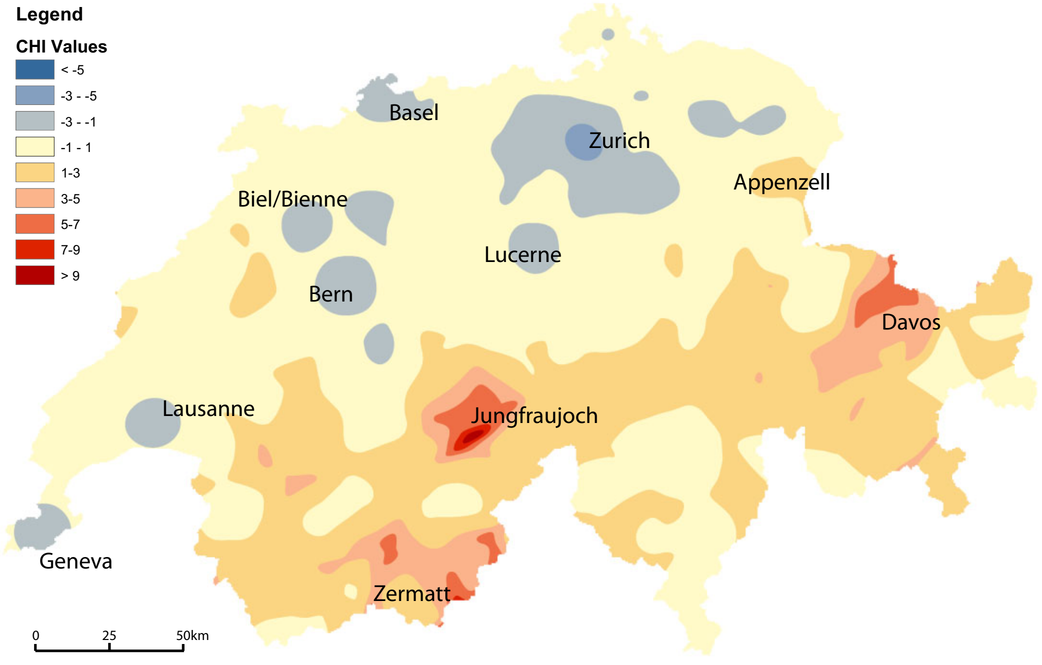

Figure 9 CHI map comparing kernel densities of population and average web counts of Switzerland. Positive χ values (red colours) denote web counts higher than population, while negative χ values (blue colours) denote population higher than web counts. Labelling is approximate

derive expected values. The kernel density estimation was plotted for a 20 km radius at a resolution of 1 km (Figure 9). The most obvious effect is that many of the χ values are between –1 and 1 (yellow parts in the map), showing that expected (population) and observed (web counts) values are similar. Some highly populated cities such as Zurich, Bern, Biel/Bienne, Basel, Lausanne and Geneva (shades of blue), have negative χ values i.e. higher population than web counts. Some typical tourist destinations such as the areas around Jungfraujoch, Zermatt, Davos and Appenzell (shades of red and orange), have positive χ values, i.e. lower population than web counts. The entire area of the Alps also has low population and higher web counts than expected; therefore the entire region has an orangey shade.

Finally, we measured the correlation of these variables with the web counts. The second column of Table 12 first shows the correlation between the populations of places with the web counts; the third column then presents the correlation between the populations of each canton with the web counts per canton. As a proxy for how significant a tourist destination a place is, we considered the number of rented hotel nights per year for that canton. It is possible that many of these rented hotel nights were used for business purposes. In our current work we follow the definition of tourists given by the United Nations' Conference on International Travel and Tourism, 1963 (Leiper, 1979): tourists are "temporary visitors staying at least twenty-four hours in the country visited and the purpose of whose journey can be classified under one of the following headings: (a) leisure (recreation, holiday, health, study, religion, and sport), (b) business, family, mission, meeting." The corresponding correlation coefficients are presented in the last column of Table 12. Note that data was available only until 2003, hence the contents of Table 12 is for the

Table 12 Correlation (r) of web counts with population and hotel nights (highest correlation per language highlighted in bold)

Language	Correlation with population (all places with information)	Correlation with population (cantons only)	Correlation with hotel nights rented per year (cantons only)
German	0.3817	**0.6676**	0.4508
English	0.1811	0.2793	**0.5079**
French	0.2056	0.2159	**0.2360**
Italian	0.0612	0.5496	**0.6023**

year of 2003. The statistical data are published by the Swiss Federal Office of Statistics, Neuchâtel (see http://www.bfs.admin.ch/bfs/portal/en/index.html for additional details).

5 Discussion

In this section we look back at the research questions we asked in the introduction and discuss them individually.

1. *How does the geographic distribution of web coverage for tourism-related themes vary across Switzerland?*
   In the current article we measured the web coverage through the number of web documents that exist for a given location known as the web count (Pasley et al. 2008). The web counts are only approximate values for measuring the coverage and this method works only for relative numbers and does not account for artificially high occurrences of a toponym due to ambiguities or other reasons. They convey aggregate coverage rather than individual trends, as some toponyms were removed due to semantic ambiguities. Also, the search engine may have limited coverage and this also might introduce a bias in our results. Using web counts is a relatively straightforward method of measuring the background coverage of a particular collection and can be quickly carried out. Such an approach then allows the exploration of values which differ from the underlying distribution. The web counts are not merely artefacts of overall coverage but we would argue that generating web counts is a fundamental first step before drawing conclusions based on the coverage of some specialized collection. Furthermore, the web counts are only a proxy of what really exists in terms of content regarding a particular theme. Nevertheless, inspection of the top 20 web pages in our case revealed that these pages most often contained web pages from the official website of the city, Wikipedia, Wikitravel, TripAdvisor, Qype, Yelp, MySwitzerland, Viator, Yahoo! Travel, etc., which clearly relate to tourism.

   The geographic distribution of these web counts seems to be most affected by language (Figure 6) and population (Figure 9). This can also be seen in Table 9 showing the top 10 web counts. These toponyms are often major cities and they can be seen clearly in the map showing the hotspots (Figure 8a), which also suggests that there is a correlation of higher web counts to these cities and their neighboring places. Roundish clusters of hotspots can also be seen for cities such as Zurich, Geneva and Basel; hence, proximity of a place to a big city also seems to play a role in higher coverage. On the other hand the hotspots for Grindelwald, La Chaux-de-Fonds and Davos are more linear. This behavior suggests that there are several distinct points of interest, rather than a cluster of points around a larger place (e.g. in the Grindelwald area), or that the coverage depends on the terrain, e.g. for linear patterned hotspots in the valley surrounding Davos. The different coverage maps (Figures 4 and 5) also show that the coverage is affected by the datasets used. The valleys are better covered than the mountainous areas. Big cities have larger circles and the area around the Alps in general has sparse coverage, but comparatively it is higher in the SwissNames dataset.

2. *Are there any differences in web coverage distribution for different languages and gazetteer datasets?*
The web counts differ for different languages and this is seen very clearly in the graphs (Figures 1 and 2) and coverage diagrams (Figures 4 and 5). German is very well covered but Italian is not, corresponding to the linguistic distribution of the Swiss population. On the other hand we see that the spatial autocorrelation is the least for English, translating into wider coverage area and the tendency towards the coverage being dispersed as compared with the other languages. French web counts, on the other hand, seems to have moderate coverage but are spatially highly correlated with the French-speaking region.

From the two bar charts (Figure 3), English and French show similar behavior in both SwissNames and Geonames gazetteer datasets in terms of the cardinality of their web counts being similar. However, after looking at the kernel density map (Figure 6), we can see a clear bias of French web counts to the French speaking part. This is also the case for toponyms in the German and Italian speaking part of Switzerland; they are better covered in the German and Italian languages, respectively.

The maps (Figure 7) of χ values show the comparison between the average web counts and the web counts in four languages, thus comparing the difference between expected and observed counts. Assuming that calculating the average web count is a way to reduce the bias caused by language, we are able to examine how much each language differs from average web counts. English, as mentioned earlier, seems to converge (lighter colours) more than the other languages, hinting that coverage is more homogeneous than in other languages.

3. *How do factors such as population and touristic popularity of a place affect the coverage?*
One might guess that the population of a place has a positive effect on the amount of web content for a given place. Highly populated places tend to have better transport infrastructure and more information that is important in the context of tourism is potentially available. Considering Table 10, Zurich, Geneva, Bern and Basel are present in the top 10 web counts across all the languages and they are also the four most populated cities in Switzerland. However, when we computed the correlation between population and counts for places and cantons the results were not what we expected. On further examination we noticed this behavior could be because of a very large number of geo/non-geo ambiguities. These ambiguities cause the web count to be artificially high for many tiny villages, not of interest to most tourists, e.g. Wald (forest in German) and Burg (castle in German). Hence, for a more meaningful result, we computed the correlation (r) between places with the top 100 average web counts and their corresponding population. The result was 0.73, which hinted at a positive and somewhat strong correlation. The places with top 100 average web counts were chosen simply because for all languages, we performed a manual disambiguation.

To measure the popularity of a tourist destination is not straightforward. The web counts themselves do convey some information about the popularity of a place, but not explicitly. Hence, we selected the number of hotel nights rented per year per canton as a better indicator of touristic popularity and compared them to the web counts of the corresponding cantons through the method of correlation. We found that for French, English and Italian the correlations of web counts to hotel nights per year are higher than web counts to population (Table 12). With the factors that we have examined above it is difficult to point that the coverage is affected by a list of deterministic factors and tag the coverage with individual correlations. We only have hints in the form of correlations from the big players such as population and language. For a given place, spatial factors such as its terrain, daily flow of people in and out, public transport connections (especially in the case of Switzerland) and its vicinity to a big city or important landmark could affect the coverage. We have also not directly examined any temporal factors such as the season or time proximity to a big festival or event. It is possible that a certain toponym may have high counts because of the above reasons.

Studies in geographic information retrieval often use quantitative web information about places for various decisions and assume homogeneity (Jones et al. 2008). Our main point in the article is to emphasize that web coverage varies geographically and linguistically and is not homogeneous. This also means that a method of normalization is needed when dealing with quantitative analysis of web resources, as results could be biased by the amount of unequal data that exists for different places. Web counts, population and touristic popularity are parameters that could be used for normalization. Not only that, but there are big differences attributed to language, and we try to show this in the graduated circle maps (Figures 4 and 5) and by visualizing the differences between the web coverage for four different languages (Figure 6). We are able to visually show how language causes bias to the extent that, for a given place, the amount of web content is sometimes very low for a particular language and very high for another (Figure 7). This means that while conducting research, the language in which it is conducted needs to be selected carefully, as the results could greatly vary depending on the language they use. This is true especially for places that are multilingual.

6 Conclusions

In this article we examined the web coverage through a simple method of web counts, with a focus on the variation in different languages. As well as measuring coverage we examine how various toponym spellings affect the coverage. While the basic method is not a new one, our main contribution lies in examining the geographic and linguistic coverage across Switzerland and exploiting various methods to visualize and analyze the differences between them. We also focused on using unbiased data by looking at toponyms from three different datasets and not only for just populated places, but also examined toponyms in the form of explicit tourist attractions.

However, there are a series of issues that remain unsolved with respect to our work. One of our main challenges was toponym ambiguity. Firstly, we solved this using a simple approach and removed geo/non-geo toponym ambiguities. Ideally, it would have been more useful to examine them and apply disambiguation methods such as the ones discussed in the background work in Section 2.1. Secondly, we did not make use of a timeline. We harvested counts for a whole month, but one can imagine that the effects of seasons and the time of the year play a very important role on tourism web content. In winter, the probability that most pages talk about winter-related activities and associated places is higher. Hence, our research suggests only a trend of a snapshot, rather than the exact picture. Thirdly, we are bound by the coverage of the toponym dataset itself and its lack of inclusion of vernacular place names. Lastly, we have not thoroughly researched the difference in web counts arising due to the use of different locales while sending the query to the search engine, something that should be addressed in future research.

Studying the web coverage for tourism in Switzerland is part of a plan to explore tourism information from the web for mobile location-based services. In the process of our web counting experiment, we have gathered plenty of georeferenced User Generated Content (UGC) information mainly in the form of text. In the next step we will gather image data (from the Flickr image sharing platform) and their tags. Together with the counts data, there is much information that can be obtained from these images and their tags (Jain et al. 2010; Popescu and Grefenstette 2009). From these tags it is possible to extract place-based semantics (Rattenbury et al. 2007), such as activities performed in a place, along with their popularity with respect to a certain toponym. It will then be possible to make inferences on how place can be described by these activities and to also automatically extract activity locations. It is further possible to record the above extracted web counts, along with tourism indicators such as population and hotel rents per night, in an auxiliary data structure, which can be linked to a spatial database (such as a multiple representation database, or MRDB) via a gazetteer. This provides a way of enriching the spatial data with non-topographic, semantic information that in a later stage may inform processes of portrayal in web and mobile services (e.g. tourism-related location-based services), such as real-time map generalization (Bereuter and Weibel 2013).

References

Amitay E, Har'El N, Sivan R, and Soffer A 2004 Web-a-where: Geotagging web content. In *Proceedings of the Twenty-seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, United Kingdom: 273–80

Bereuter P and Weibel R 2013 Real-time generalization of point data in mobile and web mapping using quadtrees. *Cartography and Geographic Information Science* 40: 273–80

Buscaldi D and Rosso P 2008 A conceptual density based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science* 22: 301–13

Buscaldi D 2011 Approaches to disambiguating toponyms. *SIGSPATIAL Special* 3(2): 16–9

Crandall D J, Backstrom L, Huttenlocher D, and Kleinberg J 2009 Mapping the world's photos. In *Proceedings of the Eighteenth International Conference on the World Wide Web*, Madrid, Spain: 761–70

Cunningham H, Maynard D, Bontcheva K, and Tablan V 2002 GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania

Egenhofer M J 2002 Toward the semantic geospatial web. In *Proceedings of the Tenth ACM International Symposium on Advances in Geographic Information Systems*, McLean, Virginia: 1–4

Graham M, Hale S, and Stephens M 2012 Digital divide: The geography of internet access. *Environment and Planning A* 44: 1009–10

Graham M, Stephens M, and Hale S 2013 Mapping the geoweb: A geography of Twitter. *Environment and Planning A* 45: 100–2

Hecht B J and Gergle D 2010a On the "localness" of user-generated content. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '10)*, Savannah, Georgia: 229–32

Hecht B J and Gergle D 2010b The tower of Babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the Twenty-eighth International Conference on Human Factors in Computing Systems*, Atlanta, Georgia: 291–300

Hill L L 2006 *Georeferencing: The Geographic Associations of Information*. Cambridge, MA, MIT Press

Hollenstein L and Purves R S 2010 Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science* 1(1): 21–48

Jain S, Seufert S, and Bedathur S 2010 Antourage: mining distance-constrained trips from Flickr. In *Proceedings of the Nineteenth International Conference on World Wide Web*, Raleigh, North Carolina: 1121–22

Jones C B, Purves R S, Clough P D, and Joho H 2008 Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science* 22: 1045–65

Jones C B and Purves R S 2008 Geographical information retrieval. *International Journal of Geographical Information Science* 22: 219–28

Keller F and Lapata M 2003 Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29: 459–84

Kilgarriff A and Grefenstette G 2003 Introduction to the special issue on the web as corpus. *Computational Linguistics* 29: 333–47

Lapata M and Keller F 2005 Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing* 2(1): 1–30

Leidner J L 2007 Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Unpublished PhD Dissertation, School of Informatics, University of Edinburgh

Leidner J L and Lieberman M D 2011 Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special* 3(2): 5–11

Leiper N 1979 The framework of tourism: Towards a definition of tourism, tourist, and the tourist industry. *Annals of Tourism Research* 6: 390–407

Li L, Goodchild M F, and Xu B 2013 Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science* 40: 61–77

Ord J K and Getis A 1995 Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27: 286–306

Overell S and Rüger S 2008 Using co-occurrence models for placename disambiguation. *International Journal of Geographic Information Science* 22: 265–87

Pasley R, Clough P, Purves R S, and Twaroch F A 2008 Mapping geographic coverage of the web. In *Proceedings of the Sixteenth ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Irvine, California: 1–9

Popescu A and Grefenstette G 2009 Deducing trip related information from Flickr. In *Proceedings of 18th International World Wide Web Conference*, Madrid, Spain: 1183–84

Purves R S 2011 Methods, Examples and Pitfalls in the Exploitation of the Geospatial Web. In Hesse-Biber S N (ed) *The Handbook of Emergent Technologies in Social Research*. New York, Oxford University Press: 592–622

Rattenbury T, Good N, and Naaman M 2007 Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands: 103–10

Rauch E, Bukatin M, and Baker K A 2003 A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Edmonton, Alberta: 50–4

Sanderson M and Kohler J 2004 Analyzing geographic queries. In *Proceedings of the 2004 Workshop on Geographic Information Retrieval*, Sheffield, United Kingdom

Silverman B W 1986 *Density Estimation for Statistics and Data Analysis*. Boca Raton, FL, Chapman and Hall

Tezuka T, Yokota Y, Iwaihara M, Tanaka K, and Zhou X 2004 Extraction of cognitively-significant place names and regions from Web-based physical proximity co-occurrences. In Zhou X, Su S, Papazoglou M P, Orlowska M E, and Jeffery K G (eds) *Web Information Systems: WISE 2004*. Berlin, Springer-Verlag Lecture Notes in Computer Science Vol. 3306: 113–24

Twaroch F A, Smart P D, and Jones C B 2008 Mining the web to detect place names. In *Proceedings of the Second International Workshop on Geographic Information Retrieval*, Napa Valley, California: 43–4

Venkateswaran R 2010 A study of the tourism Web coverage in Switzerland. In *Proceedings of the Sixth International Conference on Geographic Information Science (GIScience 2010)*, Zurich, Switzerland

Volk M 2009 How many mountains are there in Switzerland? Explorations of the SwissTopo Name List. In Clematide S, Klenner M, and Volk M (eds) *Searching Answers: A Festschrift for Michael Hess on the Occasion of his 60th Birthday*. Cottbus, Germany, MV-Verlag: 127–40

Zong W, Wu D, Sun A, Lim E, and Goh D 2005 On assigning place names to geography related web pages. In *Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, Colorado: 354–62

