Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

Measuring topographic similarity of toponyms

Curdin Derungs
Department of Geography
University of Zürich
Winterthurerstr. 190
8057 Zürich, Switzerland
curdin.derungs@geo.uzh.ch

Ross S. Purves
Department of Geography
University of Zürich
Winterthurerstr. 190
8057 Zürich, Switzerland
ross.purves@geo.uzh.ch

Abstract

We report on research motivated by toponym disambiguation of natural landscape descriptions. Additional to euclidean distance we consider topographic similarity between toponyms as a disambiguation criterion. For this reason we create topographic space, a 2D space derived from multidimensional measurements of geomorphometric characteristics at toponym locations and dimensionally reduced by the use of machine learning algorithms (SOM). Topographic similarity between toponyms in topographic space is a measurement of distance. For toponyms of the same type (e.g. cities or mountains), although distributed all over Switzerland, we observe strong autocorrelation in terms of topography. Comparisons across types of toponyms, for instance between mountains and rivers, indicate that euclidean distance between such toponyms, on a comparable scale, is more proximate than their topographic distance. Additionally the mapping of topographic peculiarities of toponyms to topographic space allows us to visually explore topographic relations. We are in the course of implementing topographic distance in an approach for toponym disambiguation of a large corpus of landscape descriptions.

Keywords: Toponym Disambiguation, Topography, Geomorphometry, Topographic Space, Machine Learning, SOM.

1 Introduction

In this paper we report on research comparing the similarity of *specific* and *generic* descriptions of locations in terms of their toponyms [10]. Specific descriptions take the form of *toponyms* or placenames associated with one or more locations, and form the motivation for our work. Generic descriptions can be related to basic levels or geographic objects, and are terms commonly used to describe types of geographic locale, such as mountain, river or forest [12].

Our original interest stems from a need to disambiguate toponyms associated with a large corpus of articles describing mountaineering activities in the Swiss Alps [1]. Such articles are rich in toponyms, which are often ambiguous, with for example more than 20 instances of the toponym *Schwarzhorn* existing in a gazetteer of Swiss toponyms. Assigning coordinates to such toponyms requires firstly the identification of candidate referents, and secondly toponym resolution, where multiple possible candidate referents are resolved such that a candidate referent is assigned to a single location in space [2]. Disambiguation relies on the use of methods to distinguish between candidate referents, with baselines for instance simply assigning the most common use of a toponym (e.g. London always refers to the capital city of the UK) or relying on textual clues (does a reference to London also refer to Canada, and thus suggest the candidate referent should be resolved to London, Ontario, Canada). In our previous work, we hypothesised that, since the same toponym may be applied to differing landforms, geomorphometric information might be used to resolve toponyms, and demonstrated a considerable improvement over a baseline method for a particular type of toponyms [3].

Since we had established that geomorphometry could distinguish between different instances of toponyms, we set out to explore the similarity between groups of toponyms referring to different generic types of objects (for example, all toponyms classified as mountains or cities in a gazetteer).

Our approach to exploring similarity between groups of toponyms and their types was guided by our work on disambiguation, and we used two basic methods for comparison – firstly, Euclidean distance, as a baseline, reflecting the notion of spatial autocorrelation popularised by Tobler [13] and commonly used in toponym disambiguation (c.f. [6]) and, secondly, what we term *topographic distance*. Topographic distance is calculated by characterising a location in terms of its geomorphometric signature and using machine learning methods to generate a *topographic space* where distance represents the similarity between locations in terms of geomorphometry.

Thus, in our paper, we provide insights into the similarity of toponyms classified as being of the same type in a gazetteer (e.g. toponyms classified as being of the type *mountain*). Furthermore, we show that our implementation of topographic space is useful to facilitate comparisons between different toponyms (e.g. Zürich or Matterhorn) and their generic type (e.g. mountain or river).

Our underlying hypothesis is that toponyms or generics of similar type will be nearer to one another than randomly distributed objects in either Euclidean or topographic space, if these spaces are meaningful in describing the distribution of such objects.

2 Describing topography

Underlying our approach is the assumption that topography is an important attribute in characterising locations related to geographic features identifiable within landscapes [11]. Thus, for example, when shown an image such as that in Figure 1, we recognise mountain peaks and a village in the foreground, and might expect other villages and mountain peaks to have similar topographic characteristics. Such topographic characteristics cannot be described by a single attribute such as elevation or slope (for example the village in the foreground is at a similar elevation to the United Kingdom's highest mountain) and a wide range of geomorphometric classifications have been developed to describe locations on the basis of parameters derived within moving windows.

Thus, for instance, Iwahashi and Pike [5] characterised DEM pixels in terms of 16 classes with different gradient, convexity and texture, while Wood [15] allocated locations to one of six landform types on the basis of the second derivative of local curvature fitted through a quadratic function to moving window.

Figure 1: The village of Sent in the Engadine Valley of Switzerland with the mountain Piz Lischana in the background

To describe the topography at individual locations, we used a SwissTopo Digital Elevation Model (DEM) with a resolution of 25m. Since it is well known that classification of a DEM varies with scale [15], we derived geomorphometric values at a range of scales reflecting different window sizes: in the case of the Iwahashi and Pike [5] characterisation for 75m, 200m, 1km and 10km (Figure 2) and for Wood [15] at 200m and 2km (Figure 3). These window sizes were selected to allow us to capture features of varying sizes, thus for example the village in Figure 1 covers an area of perhaps 1km^2, whilst an individual mountain summit might be found within an area of 200m^2, but a mountain ridge or valley extend for kilometres.

Since both metrics essentially deliver a vector describing membership of classes at each scale, then the result is a vector containing (16 classes x 4 scales) (for Iwahashi and Pike) and (6 classes x 2 scales) (for Wood) classes, resulting in a total of 76 dimensions. We now describe how similarity between vectors describing differing locations can be both visualised and quantified, using machine learning methods to reduce these 76 dimensions to a 2D space.

Figure 2: Geomorphometric classification according to Iwahashi and Pike [5]. Cells of the DEM are classified into 16 classes according to gradient, convexity and texture

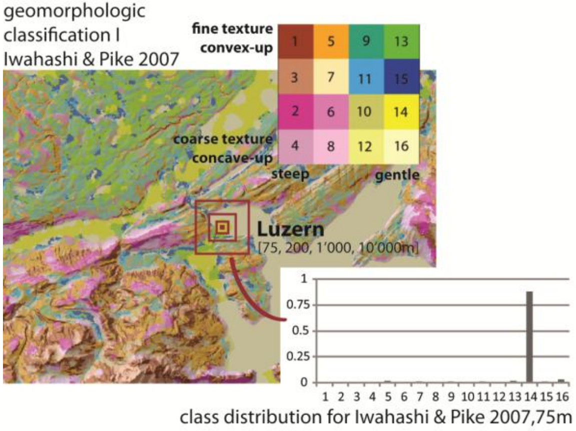
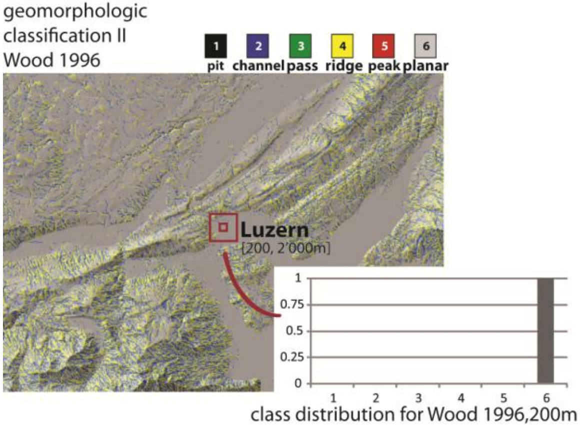


Figure 3: Geomorphometric classification according to Wood [15]. Cells of the DEM are assigned to one of six landform types.



3 From Euclidean to topographic space

Our *Euclidean space* took the form of toponym data extracted from SwissNames, a database of all toponyms shown on Swiss maps of scales ranging from 1:25000-1:500000 for a range of toponym types, and containing more than 150000 entries. Each toponym is assigned coordinates, reflecting the positioning of the toponym on the map. Thus, geographic features are, as is typical in gazetteers, abstracted to points despite the obvious areal nature of, for instance, cities. Distances between toponyms in this Euclidean space are straightforward to calculate on the basis of the coordinates of toponym pairs. Such distances can be related to one another under the assumption that near things are more similar [13].

Topographic space describes each toponym location in terms of the 76 dimensional vectors generated in the previous section. In order to allow both the straightforward calculation of distances between locations in topographic space, and visualisation of this space analogous to our original Euclidean space, we reduced the 76 dimensions to a 2D topographic space. Creation of 2D space from multidimensional vectors is a classical dimension reduction task [7]. We chose to use a

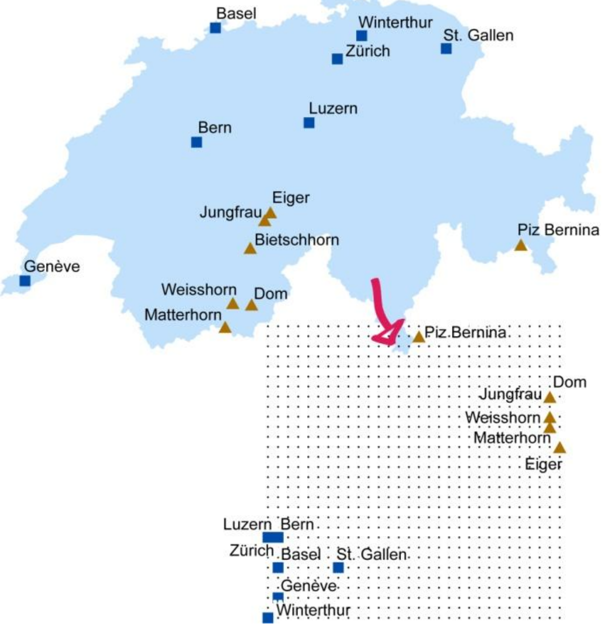
SOM (Self Organising Map) algorithm where the aim is that similarity between locations in vector space is preserved as proximity in 2D space [8]. We assume that this proximity can be measured as a distance within this 2D space, and thus that comparisons can be made between the Euclidean distance between two points in geographic space and their distance in the SOM.

We trained the SOM using the Kohonen package implemented in R [14] using a random sample of 60000 out of 156000 toponym locations, with each location being classified in terms of its vector of 76 topographic dimensions. Training data were presented 5000 times to the 30x30 neuron SOM. Since we assume independence between geomorphometric values at different scales and in the two classifications, a super SOM algorithm was used where independent sets of measurements were passed as separate matrices. Measurements from smaller windows sizes, emphasising the neighbourhood of a location, were given greater weight.

We do not claim that the output SOM is sound in terms of geomorphometric autocorrelation. Although in SOM's distance is a measurement of similarity, the same distance between different neurons in the SOM does not necessarily represent the same similarity value. The U-matrix representation can be used to analyse such variations in terms of dissimilarity between neighbouring neurons (e.g. [8]). However, we chose to use distance within the SOM itself as a proxy for similarity in an initial exploration of the method since the strength of SOM algorithms for the work presented lies in their visual output. Skupin [9] argues that as the number of neurons in the output SOM rises the method starts to function as a spatial layout technique rather than a clustering approach.

After training, any toponym location for which geomorphometric information is known can be mapped to our topographic space. Figure 4 shows some prominent geographic locations in Switzerland represented in Euclidean and topographic space. Note how mountains and cities are clustered in topographic space, reflecting their similar geomorphometric properties.

Figure 4: Representation of locations in Euclidean and topographic space.



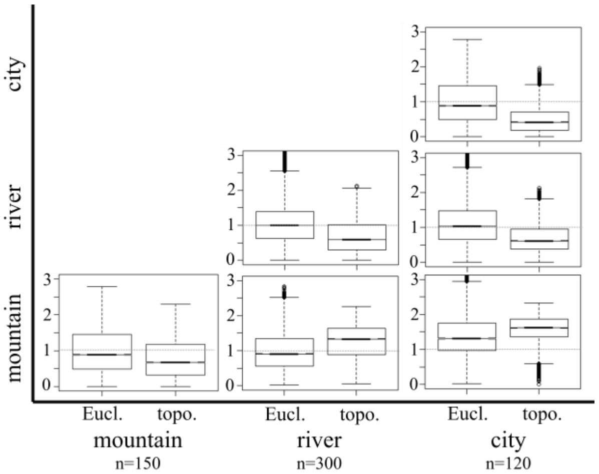
Note that rivers, although incorporated in the following investigations, are not depicted as a cartographic element in Figure 4. A single river, depending on its importance, may be labelled at multiple locations along its length, with for example the Rhine having 28 toponyms in the SwissNames data used here. Thus, although we can characterise rivers in our topographic space, the mapping is n-n rather than 1-1 as is the case for mountains or cities.

4 Comparing toponym types

We illustrate our method by exploring the similarity of toponyms of three different types (cities, mountains and rivers) (Figure 5). These toponym types were selected since we expected mountains and cities to have contrasting geomorphometric characteristics, whilst toponyms referring to rivers might be found in a variety of settings and thus more difficult to classify on the basis of topography.

Distances were calculated between toponym pairs in both Euclidean and topographic space, both within and between toponym types. These distances were then normalised according to the average distance between all toponyms within SwissNames such that a distance of greater than one implies an object is more distant than a randomly selected pair within SwissNames.

Figure 5: Distances between sets of toponyms of mountains, rivers and cities in Euclidean and topographic space.



Each pairwise comparison within a particular toponym type, was found to be significantly different (T-test, p<0.01) in terms of Euclidean and topographic distance, with Euclidean distances being typically nearer 1 (equivalent to the average distance between two randomly selected toponyms from the gazetteer). Thus, in Figure 5, it is clear that all values along the diagonal (i.e. comparison of distances within toponym types) are less than 1 in topographic space. This in turn implies that mountain, river and city toponyms are geomorphometrically more similar to one another than other randomly selected toponyms.

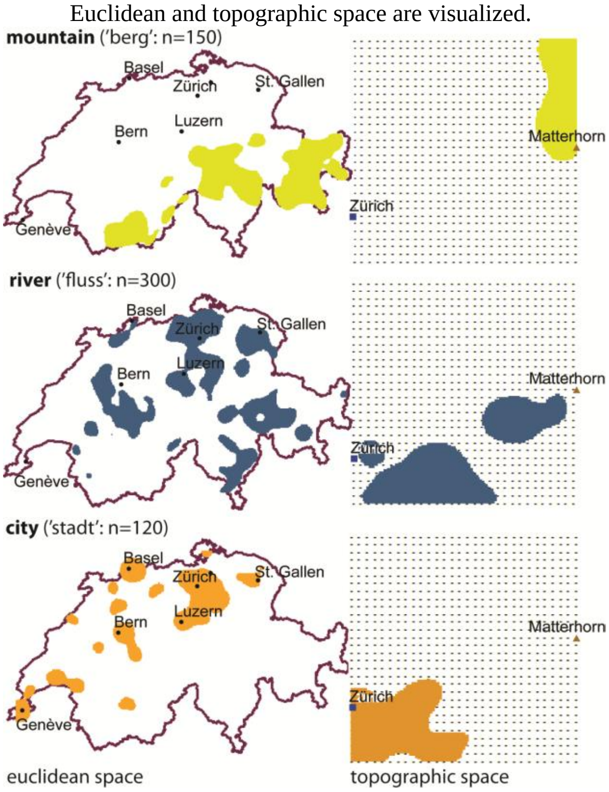
When making comparisons between toponym types, it is clear that mountains are distant from cities in both Euclidean and topographic space. In Euclidean space no difference from a random distance distribution is visible when comparing between rivers and cities or mountains. However, in topographic space rivers are distant from the regions occupied by mountains, but similar to those occupied by cities.

These bulk properties give us confidence that topographic space helps to discriminate between our different toponym types, but does not give insight into the spatial distribution and relationships between toponym types within Switzerland. To explore these properties, we mapped densities of toponym types in both Euclidean and topographic space (Figure 6).

In Euclidean space the distribution of toponym types within Switzerland is reflected, with mountain toponyms associated with the Alpine belt stretching from west to east in the southern half of the country. City toponyms are found in the northern half of Switzerland, roughly corresponding to the region associated with the "Mittelland", a densely populated and highly developed plain. Finally, river toponyms are distributed over large portions of the country in regions associated with both mountains and cities.

All three toponym types are more tightly clustered in topographic than Euclidean space. Furthermore, mountains and cities are found on opposite sides of the SOM, reflecting their topographic distance from one another. The case of rivers is a little different. Here, we see that river toponyms are found in regions within Switzerland associated with both cities and mountains when visualised in Euclidean space. However, in topographic space rivers intersect with the region occupied by cities and appear to occupy different regions from mountains. This accords with the notion that rivers are typically an important part of cities, while in mountainous regions rivers are found in valleys neighbouring, but not part of, the mountains themselves. These visualisations thus may help to understand why, for example, mountains are distant from cities in both Euclidean and topographic space and suggest that the use of topographic space provides an alternative way of exploring the properties of toponym types.

Figure 6: Comparison of spatial distributions of mountain, river and city instances from gazetteer. Top 20% densities in Euclidean and topographic space are visualized.



5 Concluding discussion

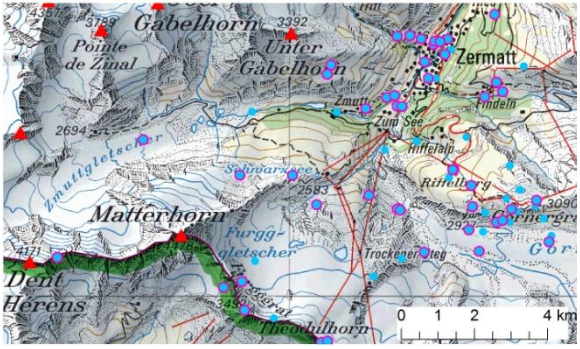
In this paper we extended previous work aimed at disambiguating toponyms using information about their geomorphometry towards the use of topographic space, a generic tool for measuring differences of geomorphometric characteristics between different types of toponyms and geographic generics.

We combined methods to describe the geomorphometry of individual locations associated with point references to toponyms at a variety of scales, and used a machine learning method, SOMs, to project the resulting vector into 2D. This allows on the one hand quantification of topographic differences between toponym types, and on the other hand to visualise regions associated with toponym types. Such information can be used in disambiguation as an additional metric to, for example, assist in the resolution of candidate referents of different types [6]. We are currently implementing and evaluating disambiguation methods for the Text and Berg [1] corpus based on these results, and will extend this implementation to other corpora with containing text with significant landscape related components, for example descriptions used in environmental assessments.

Furthermore, the regionalisation of toponym types in topographic space provides us with a new insight into the distribution of such classes, and suggests new lines of potential research. Thus, for a new toponym of unknown type, it may be possible to suggest likely classifications based on topographic space. One potential application might thus lie in the assigning of types to vernacular toponyms derived from collections such as Flickr [c.f. 4].

A second potential area of research lies in comparing generic descriptions of images in such collections (e.g. references to mountain in Flickr) with toponym data, and may have a role to play in discussions on geographic kinds and their relationships [11, 12]. However, caution is required here. Figure 7 shows a map extract around the iconic Matterhorn, perhaps one of the archetypical mountains referred to in literature and art.

Figure 7: Instances of mountains from SwissNames (red triangles), georeferenced Flickr photographs with the tag *mountain* (blue circles) and those Flickr photographs that reference to *mountain* and *Matterhorn* (blue circle, pink halo).



Source: Background mapping © SwissTopo.

Toponyms referring to mountains within SwissNames are all found, as one would expect, on the summits of individual mountains. However, very few Flickr images are actually situated on mountain peaks, and many are in fact on the valley floor accessible to most tourists. Furthermore, the majority of images tagged with *mountain* also have the tag Matterhorn. Thus, unlike our gazetteer which seeks to describe all spatial features of a particular type, Flickr images concentrate on a single instance of the type and are situated around, rather than on, the mountain. Clearly, varying DEM resolution and the size of moving windows at which geomorphometric properties are captured will resolve some of these issues, but effectively using user generated content such as Flickr to explore such issues is an area of future research that we are currently addressing.

References

[1] N. Bubenhofer, M. Volk, A. Althaus, M. Jitca, M. Bangerter, R. Sennrich. *Text+Berg-Korpus (Release 145.)* Institut für Computerlinguistik, Universität Zürich, 2011.

[2] P. Clough, Extracting Metadata for spatially-aware information retrieval on the internet. In *Proceedings of the ACM Workshop on GIR*, Bremen. 2005.

[3] C. Derungs, RS. Purves, B. Waldvogel. Toponym disambiguation of landscape features using geomorphometric characteristics. In *Proceedings of the 11th International Conference on GeoComputation*, London. 2011.

[4] L. Hollenstein, RS. Purves. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1(1): 21–48, 2010.

[5] J. Iwahashi and RJ. Pike. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*, 15(3):409-440, 2007.

[6] J. Leidner. Toponym resolution in Text: "Which Sheffield is it?" In *Proceedings of the 27th Annual Internation ACM SIGIR Conference*, Sheffield. 2004.

[7] BD. Ripley, editor. *Pattern Recognition and Neural Networks.* Univeristy Press. Cambridge, 1996

[8] A. Skupin. The world of geography: visualizing a knowledge domain with cartographic means. In *Proceedings of the National Academy of Sciences*, Online, 2004.

[9] A. Skupin, A. Esperbé. An Alternative Map of the United States Based on an n-Dimensional Model of Geographic Space. In *Journal of Visual Languages and Computing*, 22(4):290-304, 2011.

[10] S. Shatford. Analyzing the Subject of a Picture: A Theretical Approach. *Cataloging & Classification Quaterly*, 6(3):39-62 1986.

[11] B. Smith, DM. Mark. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B*, 30(3): 411–428, 2003.

[12] B. Smith, DM. Mark. Geographic categories: an ontological investigation. *International Journal of Geographic Information Science*, 15(7):591-612, 2001.

[13] W. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography.* 46(2):234-240, 1970.

[14] R. Wehrens and LMC. Buydens. Self- and Super-organising Maps in R: the kohonen package. *Statistical Software*, 30(6): 1-19, 2007.

[15] J. Wood. *The geomorphological characterisation of digital elevation models.* PhD Thesis, University of Leicester, 1996