Methods of Spatial Knowledge Discovery in the Scope of Planning and Development

Martin Behnisch
Philipps-University Marburg, Datenbionik FB 12, Hans-Meerwein-Straße, 35032 Marburg, Germany
Kontakt: Behnisch@urban-data-mining.de

Alfred Ultsch
Philipps-University Marburg, Datenbionik FB 12, Hans-Meerwein-Straße, 35032 Marburg, Germany

Keywords: Spatial Knowledge Discovery, Spatial-Temporal Analysis, Long-term Development, Emergent SOM

Most of the large databases currently available have a strong spatio-temporal component and potentially contain information that might be of value. Cartographic visualizations usually provide information of low dimensional data sets (e.g. distribution, density, correlation, structure or the spatial/temporary change). Data mining in connection with knowledge discovery techniques play an important role for the empirical visualization and examination and of high dimensional spatial data. The increasing discussions about data interoperability require a transparent transfer of spatial semantics (metaphors, abstractions) and a comprehensible syntax (*Shekhar 2003*). However procedures on the basis of knowledge discovery are currently not exactly scrutinised for a meaningful integration into the regional/urban planning and development process (*Demsar, 2006*, *Behnisch, 2009*).

In this contribution spatial knowledge discovery is presented as a cyclical methodological approach to reveal logical or mathematical and partly complex descriptions of patterns and regularities inside a set of high dimensional spatial data. The cyclical procedure is characterized by six main tasks following the initial step of data collection: data inspection, structure visualization, structure definition, structure control, operationalization and knowledge conversion. The practical motivation of this contribution is to brighten the knowledge about the long-term development of (Swiss) communities in terms of patterns and localized properties. The long-term development is usually difficult to quantify and more or less nebulous in context of actual planning and decision processes (*Bätzing/Dickhörner, 2002*). But long-term aspects should be taken into account to avoid dramatic losses of economic, social and cultural capitals in the coming decades. Due to the lack of several long-term data dimensions the initial idea is to start with the development of population between 1850 and 2000 as a kind of overall indicator for the observable situation of communities. The examination of communities (=2896) is realized by 15 time intervals (=15 decades). The key questions are how many long-term patterns will occur and what are the characteristics?

Several indexes are already in use to measure the change in population. Relative difference (ReIDiff) is proposed as an alternative to relative change calculation. It is an appropriate index for classification and in particular for the search of similarity patterns (*Ultsch, 2005*). Data inspection leads to the awareness that there are 15 distributions of population change with difficult (wide) edges. A similarity measure is therefore necessary that allows generating a typology of population dynamics. The idea is to model each of all 15 distributions as a mixture of three characteristic developments: losing communities (e.g. multiplicative process), typical communities (e.g. sum of many unobserved random population is acting independently, CLT theorem), winning communities (e.g. multiplicative process, growth). A mixture model is realized as a composite of a log-normal, normal, log-normal distribution. The expectation maximization algorithm (*Bilmes, 1997*) is used for parameter computation. In addition to other mixture models Pareto density estimation (PDE, *Ultsch, 2003a*) and probability density functions (PDF) are used to ensure the modeling process. The modeled distribution is proofed by Q-Q-plots. Finally each distribution has its own characteristic and observed parameters by decade: Mean, standard deviation and amount of communities.

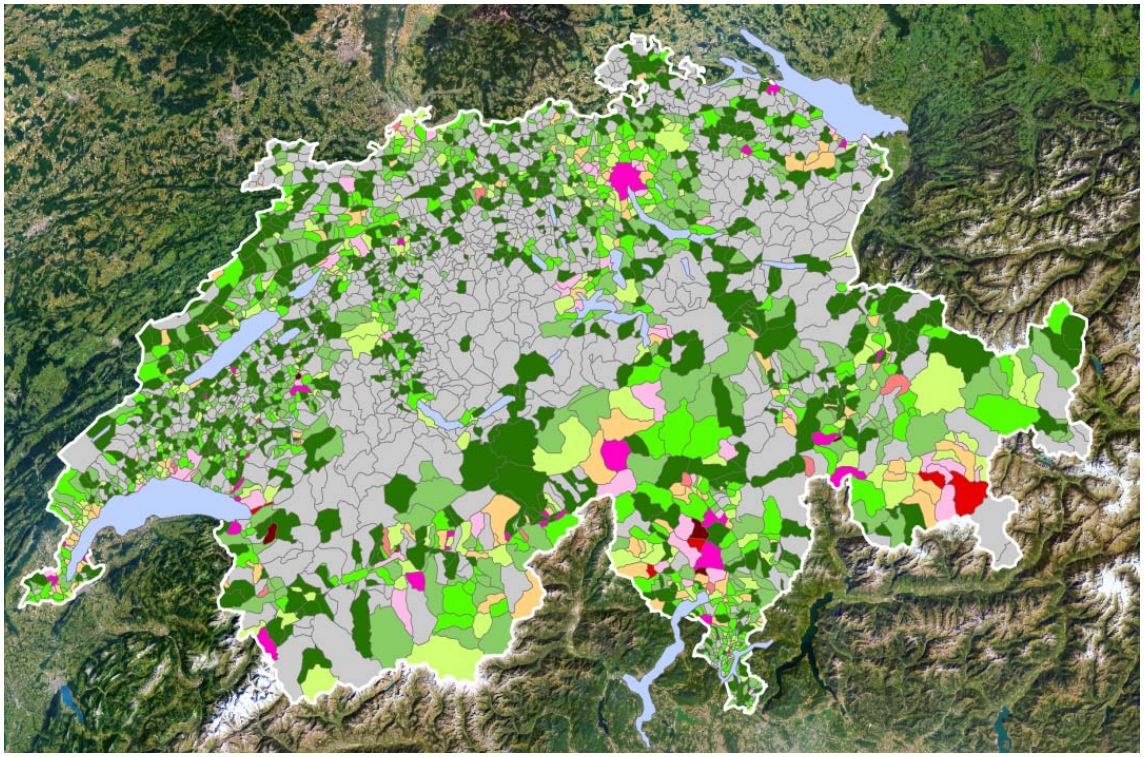
The posterior probability and component probability density function are used to compute the posterior probabilities. The Bayesian theorem offers advantages through its ability to formally incorporate prior knowledge into model specification via prior distributions and allows considering the variability. A specific dynamic class (e.g. winner, typical, and loser) is predominant observed for a given value of population change in a community. The term pattern is used to describe developments consisting of 15 unique dynamic classes (=15 decades). The understanding of each distribution represents a first intermediate result. The visualization of patterns leads to the definition of three periodical subdivisions. Observed patterns are presented in view of the property “Typical” and “Non-Typical” (=Winner or Loser). A procedure of information optimization aims to select relevant patterns for clustering (*Ultsch, 1999*). Spatial localization of probabilities and dynamic classes encourage the discussion with spatial planners to find reasons for the properties of communities. For example cartograms emphasize the cores of agglomerations and other low/high populated regions in visual manner (*Gastner/Newman, 2004*). Cluster maps represent a special choropleth map showing those locations with a significant Local Moran statistics classified by type of spatial correlation (*Anselin, 1995*).

In addition to the understanding of each data distribution the authors aim to understand the structure of several attributes. High dimensional data is projected on a low dimensional grid through projection procedures. As such, the authors believe the Emergent self organizing map (ESOM, *Ultsch 2003b*) complements the investigation of high-dimensional spatial objects. Such method preserves the neighbourhood relationships of high dimensional data. It has the advantage of a nonlinear disentanglement of complex structures. The goal of clustering is to determine the intrinsic grouping in a set of data. But how to decide what constitutes a good clustering? In particular aggregation processes of classes are often necessary to build up a meaningful classification. The corresponding U*-map delivers a geographical landscape of the input data on a projected map (imaginary axis). The cluster boundaries are expressed by mountains, which means the value of height is defining the distance between objects. Data points found in coherent regions are assigned to one cluster.

The confirmation of a clear structure supports the clustering procedure of relevant patterns. It is based on growth indicators summarizing for each period the observed dynamics by decade of a pattern. Eight Swiss population dynamics form the basis for qualitative and quantitative explanations in terms of spatial planning aspects. A finer cluster partition would lead to an elusive and marginal distinguishable amount of clusters. A k-Nearest Neighbor classifier supports the identification of the nearest neighbor that has already class information to allocate an unlabeled community.

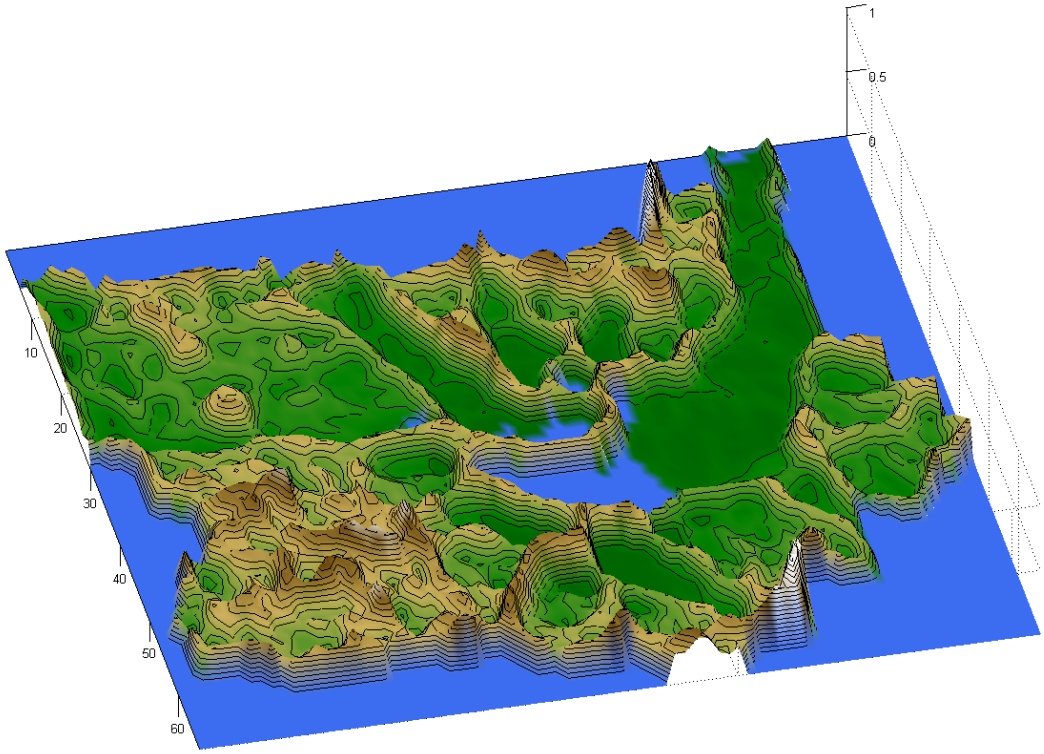
The process of knowledge conversion provides the transition from data to knowledge. Requirements for spatial knowledge discovery are interpretability, novelty and usefulness of results. In context of spatial classifications the knowledge discovery implies a continuous and ongoing search for appropriate spatial abstractions. Results depend on machine generated explanations of classes as well as human interactions or subsequent validations in mind of the involved spatial expert. The aim is to trigger discussions in the application domain. Knowledge should be formulated in two stages: A) Valid and reasonable in terms of statistical tests (significance); B) Useful for planning processes. Against this background the localization of observed long-term dynamics (e.g. spatial semantics) and related attributes is helpful for spatial verification and spatial reasoning. Trees or other symbolic classifier support the understanding of structures respectively class partitions. Furthermore the class partition and other well-known typologies (e.g. community types, urban/rural types, height zones) are compared using contingency tables in order to decide whether or not dependencies are significant. Structure interpretation and validation in mind of the spatial analyst foster the general understanding of the long-term development of Swiss communities.

Figure 1. . Communities and the amount of “Non-Typical” per pattern in 15 decades



non-typical:	0	1	2	3	4	5	6	7	8	9	10
frequency:	852	675	511	359	238	131	70	36	13	6	5 = 2896
cumulative:	29.42	52.73	70.37	82.77	90.99	95.51	97.93	99.17	99.62	99.83	=100%

Figure 2. Structure Detection of high-dimensional data (U*-Map, island view).



Literature

Anselin, L. (1995). Local Indicators of Spatial Association &emdash; LISA. Geographical Analysis 27, pp. 93-115.

Bätzing, W. and Dickhörner, Y.: Die Bevölkerungsentwicklung im Alpenraum 1870-1990 aus der Sicht von Längsschnittanalysen aller Alpengemeinden. Revue de Geographie Alpine. 89, 11--20 (2001)

Behnisch, M. (2009). Urban Data Mining. Karlsruhe: KIT Scientific Press.

Bilmes, J. (1997). A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report (http://crow.ee.washington.edu/people/bulyko/papers/em.pdf, 15.10.2009), University of Berkeley, ICSI-TR-97-021.

Demsar, U. (2006). Data Mining of Geospatial Data: Combining Visual and Automatic Methods, Urban Planning Department, KTH Stockholm.

Gastner, M.T. and Newman, M.E.J. (2004). Diffusion-based method for producing density equalizing maps. In: Proceedings of the NAS, Vol. 101, No. 20, pp. 7499-7504.

Shekhar, S (2003). Spatial Data Mining and Geo-spatial Interoperability. Computer Science Department, University of Minnesota.

Ultsch, A. (2001). Eine Begründung der Pareto 80/20 Regel und Grenzwerte für die ABC Analyse. In: Technical Reports No.30. Department of Mathematics and Computer Science, University of Marburg. Available via DIALOG. http://www.mathematik.uni-marburg.de/databionik/pdf/pubs/2001/ultsch01begruendung Cited 15 Oct 2010.

Ultsch, A. (2003a). Pareto Density Estimation: A Density Estimation for Knowledge Discovery. In D. Baier, K.D. Wernecke (Ed.), Innovations in Classification, Data Science, and Information Systems (pp. 91--100). Berlin: Springer.

Ultsch, A. (2003b) Maps for the visualization of high dimensional data spaces. In T. Yamakawa, editor, Proceedings of the 4th Workshop on Self-Organizing Maps (WSOM'03), pages 225–230.

Ultsch, A. (2005). Pareto Density Estimation: A Density Estimation for Knowledge Discovery, In: Daniel Baier and Klaus Dieter Wernecke: Innovations in Classification, Data Science, and Information Systems (Proceedings of the 27th Annual Conference of the Gesellschaft für Klassifikation e.V. Brandenburg University of Technology, Cottbus, March 12–14, 2003), Berlin, Heidelberg.