Modeling regions of linguistic phenomena, including the inherent spatial dependencies

UFSP SpuR: SynMod+

Merlin Unterfinger, Geography UZH

25 7 2017

Abstract

In this project a new approach to model linguistic regions is presented. To study the spatial distribution of linguistic phenomena concepts from Geography are applied on a linguistic data set, which was collected in the project 'Syntactic Atlas of German-speaking Switzerland' (SADS) during the years 2000 till 2002. Continuous spatial surfaces for the manifestations and intensity of the 'infinitival complementizer' linguistic phenomena in the Swiss-German language are generated by combining techniques from machine learning, like the Support Vector Machine (SVM) and a spatial regularization of the classification map (MRF) with other well-known geographic concepts (Tobler's first law, hiking function and least cost paths) to create linguistic regions, which take the inherent spatial dependencies between the data points into account. Trudgill's gravity index finds application to weight the linguistic influence of two neighbouring points on each other.

The project is supervised by Curdin Derungs, head of the GISLab, a research group of the 'UFSP Language and Space' initiative at the University of Zurich.

Contents

1 Libraries

The following libraries are needed, to execute the code:

- **e1071** – Support vector Machine – (Meyer et al., 2015)
- **FNN** – Fast nearest neighbor search – (Beygelzimer et al., 2013)
- **foreign** – Open foreign data formats (SPSS) – (R Core Team, 2016)
- **gdistance** – Least cost path, walking time – (van Etten, 2017)
- **ggplot2** – Plots – (Wickham, 2009)
- **rasterVis** – Plots of raster – (Perpiñán and Hijmans, 2016)
- **rgdal** – Open Shapefiles – (Bivand et al., 2016)
- **rgeos** – Geoprocessing – (Bivand and Rundel, 2016)
- **sp** – Spatial datatypes – (Pebesma and Bivand, 2005; Bivand et al., 2013)

2 Data

2.1 Linguistic data

The database for this project is the Syntactic Atlas of German-speaking Switzerland (SADS). In the years between 2000 and 2002 about 3'200 participants answered four questionnaires at 383 different survey sites. The questions asked, were about syntactic phenomena of the Swiss German language. Compared to other linguistic surveys, SADS is special, because it has multiple participants respondents per study site and therefore covers the local linguistic diversity. Per study site from 3 to 26 persons (median: 8) were involved in the survey (Bucheli and Glaser, 2002; Jeszenszky and Weibel, 2016).
In this project, four phenomena of the SADS data set are used, which belong to a syntactical construct that is called 'infinitival complementizer'. An infinitival complementizer is a so-called purposive infinitival clause, since the clause expresses a purpose. An example from the English language would be the difference between 'in order' to or 'to'. The original data set has been aggregated to the level of the Swiss communities. This means each community is represented by a point (point data). Due to the strongly varying number of participants per survey site, a majority voting per survey site and phenomena is not suitable. Therfore, a random sampling on survey site is applied to achieve a statistically more stable data base for the study. Each survey site was sampled 8 (median of persons per survey site) times with replacement. Then a majority vote of phenomena's manifestations was applied on the random sampled dataset. The majority vote values (class assignments) serve as input for the further steps.

2.2 Population

The population data of the Swiss communities in the year 2013 is provided by the Federal Office for statistics (BfS). The data set (STATPOP 2013) represents the population of Switzerland aggregated to hectare raster cells. Each raster cell is stored as a polygon. For easier access on the data, the polygons have been rasterized to a GeoTiff (with 100m resolution) using the *to raster* module of ArcGIS.

2.3 Geodata

The digital elevation model DHM200 of Swisstopo delivers the terrain height information used in this study. For visualization purposes the outline of Switzerland is used from the SwissBOUNDARIES3D data set from Swisstopo. The outline of Switzerland was simplified to save processing time, when plotting.

2.4 Overview

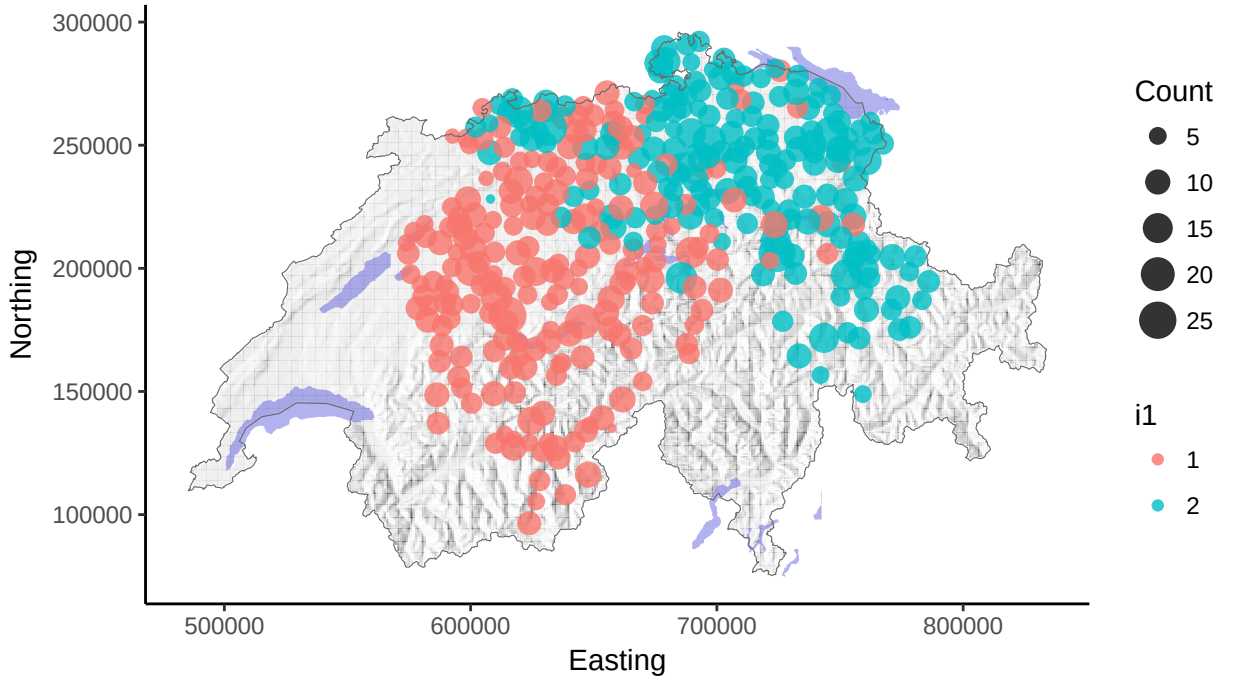

Figure 1: Overview plot of the communities for phenomena i1. The point size represents the number of participants per survey site.

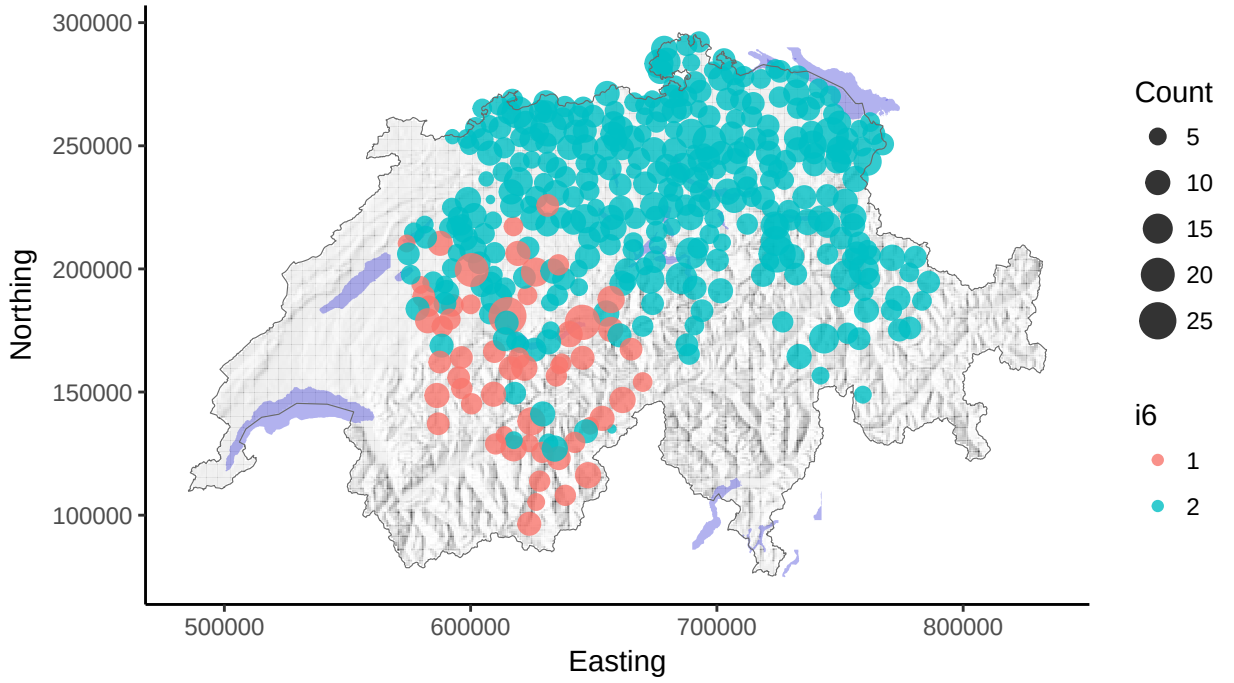

Figure 2: Overview plot of the communities for phenomena iv6. The point size represents the number of participants per survey site.

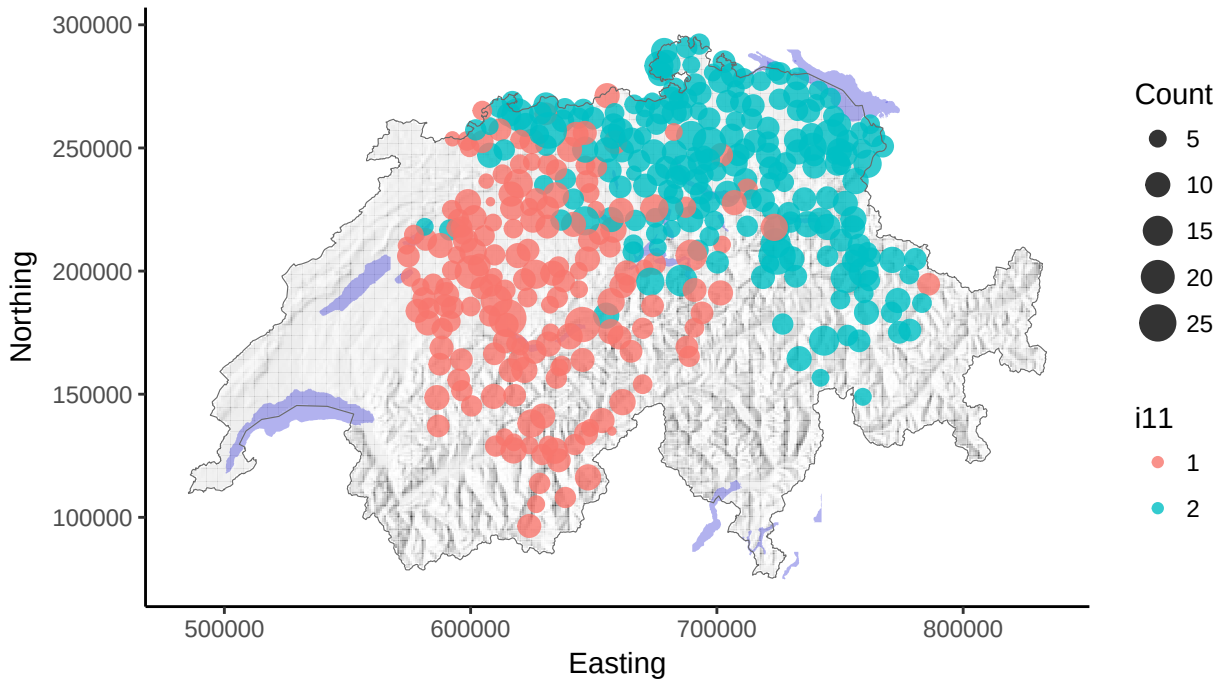

Figure 3: Overview plot of the communities for phenomena i11. The point size represents the number of participants per survey site.

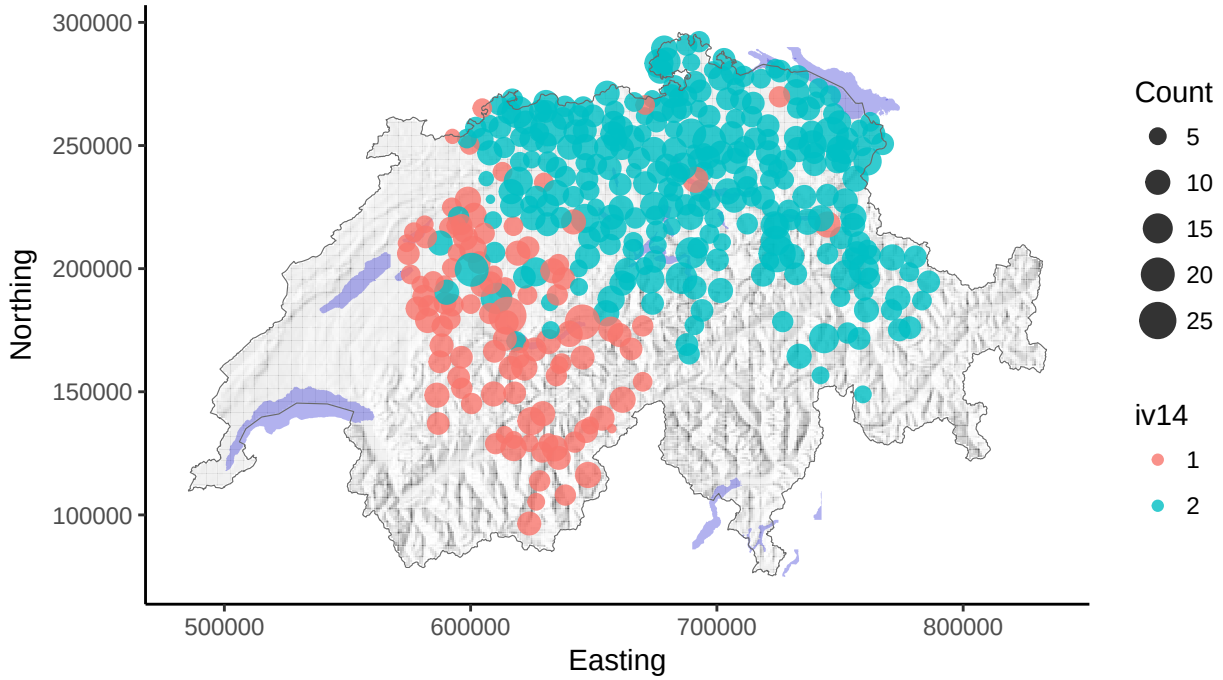

Figure 4: Overview plot of the communities for phenomena iv14. The point size represents the number of participants per survey site.

Table 1: Structure of the processed data set (random sampling & majority vote).

Easting	Northing	BFS	Name	Count	i1	i6	i11	iv14
786007.8	194653.8	3871_1	Klosters-Serneus	8	2	2	1	2
780616.8	204792.1	3893_1	Sankt Antönien	8	2	2	2	2
778152.1	198598.1	3882_1	Küblis	6	2	2	2	2
771204.8	204549.3	3962_1	Schiers	11	2	2	2	2
761946.6	196105.1	3945_1	Trimmis	7	2	2	2	2

Table 1 shows the structure of the processed data set. Each community is represented as a data point, which has a row in the table. The phenomenas are representend by the majority vote (dominant variant) of the random sampling in a particular community.

3 Support Vector Machine

3.1 Parameter Selection

To select the best parameters for the training of the SVM, with a radial basis function (RBF) kernel, a *Grid Search* is performed. The aim is to find the best combination of gamma and cost based on a previously given selection of possible values for each parameter. This step is repeatedly for the four linguistic phenomena and its corresponding data record. The grid search is performed by applying the `tune.svm()` function from the *e1071* package. In the following table, the best parameters for the training of a SVM on this data set, found by a grid search, are listed (Karatzoglou et al., 2006).

Table 2: The best parameters from grid search.

	Gamma	Cost	Error	Dispersion
i1	1.5	10	0.120	0.047
i6	1.5	1	0.089	0.055
i11	2.0	1	0.104	0.037
iv14	2.0	1	0.068	0.031

3.2 Training of the SVM

With the evaluated parameters from the previous section, a SVM is trained for each phenomena. The trained SVM is of type 'RBF', which means it uses a Gaussian kernel and therefore is a non parametric/nonlinear algorithm.

4 Spatial Context

4.1 Setup Grid

A regular grid of data points is expanded over the area of Switzerland. The grid extent is taken from a buffered (8000m) convex hull drawn around all data points.

Table 3: The extent and resolution of the grid.

	Min	Max	Resolution	PixelSize
Easting	566091	794007.4	114.0	2000
Northing	88511	300150.2	105.8	2000

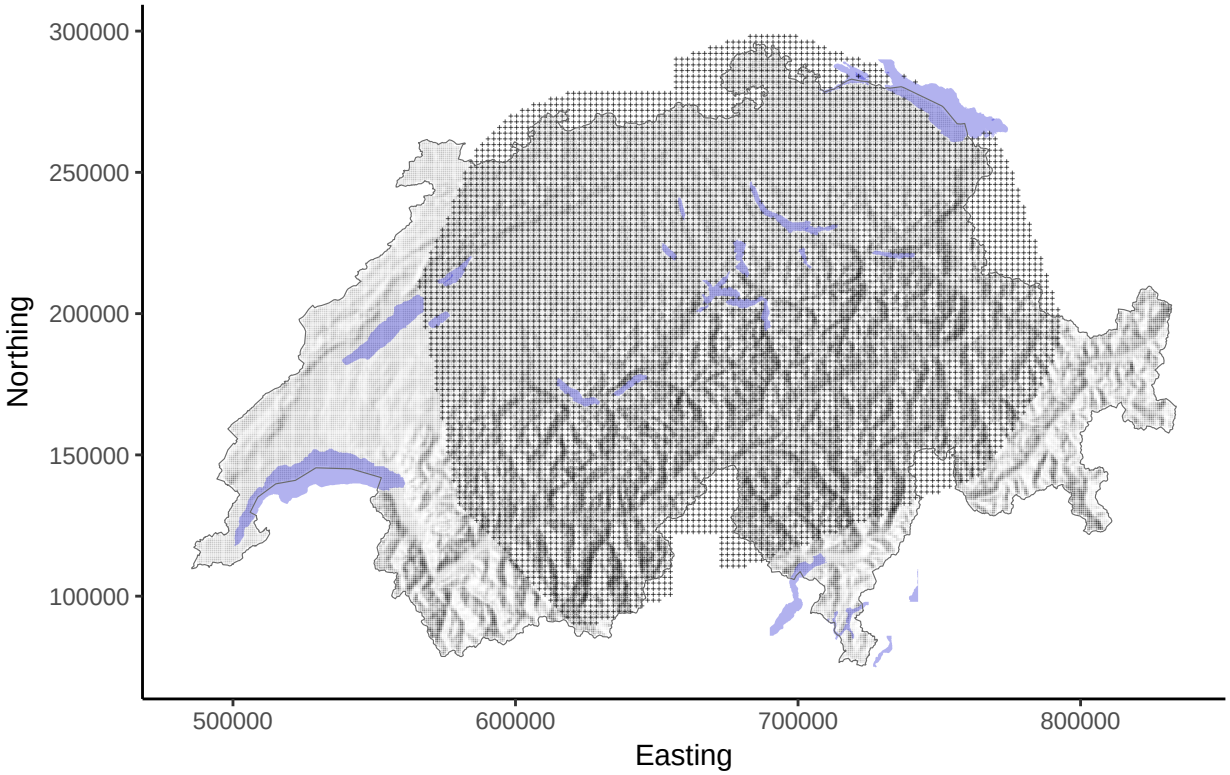

Figure 5: Regular grid over the extent the convex hull.

By applying the `extract()` function from the *raster* package on the grid data points and the data layers of the population *STATPOP13* and the height *DHM25_100m* the values at the grid data points position are extracted.

4.2 Nearest Neighbors

The nearest neighbors of each grid point are extracted by applying the `spDists()` function from the *sp* package on the grid. Then the `knn.index()` function from the *sp* package extracts the indices of the k nearest neighbors.

4.3 Hiking Function

Tobler (1993) defined a hiking function, which is representing the walking velocity W $[\frac{m}{s}]$ dependent on the slope of a path:

$$W = \frac{6e^{-3.5|m+0,05|}}{3.6}$$

Because the units in this study are meters and seconds, the function has to be divided by 3.6 to get from $\frac{km}{h}$ to $\frac{m}{s}$

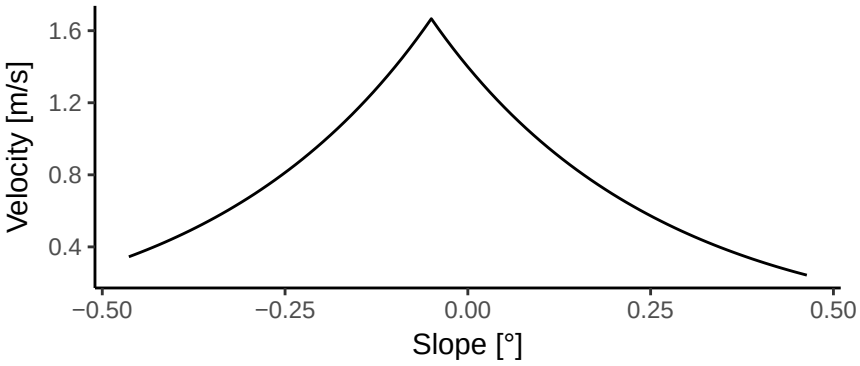

Figure 6: Toblers hiking function.

4.4 Least Cost Path & Walking Time

By using the the *gdistance* package, least cost paths between all neighbors are estimated and the time for walking these paths is extracted with the `costDistance()` function. The least cost path estimation is based on a transition layer (*conductance*), which is generated with the height differences of the pixels in the DHM. In a next step Toblers hiking function is applied on the Slope in 8 directions of each pixel. This produces a transition layer with the walking velocity W in the 8 possible directions at every pixel in the DHM.

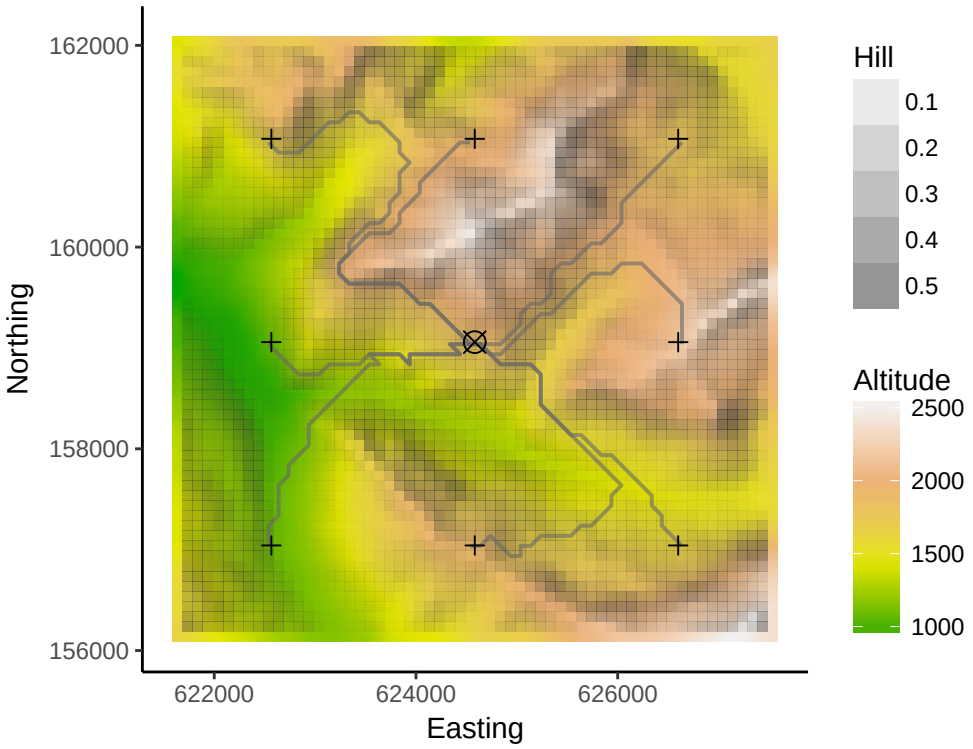

Figure 7: Least cost paths to the k neighbors of an example point.

Because searching least cost paths on such a huge (global) transition layer is computationally expensive, multiple local (smaller) transition layers are generated with a moving window approach that has a variable size. By doing so for every grid point and its k neighbors, a local transition layer is generated and the walking time for the least costs paths to the neighbors is computed locally.

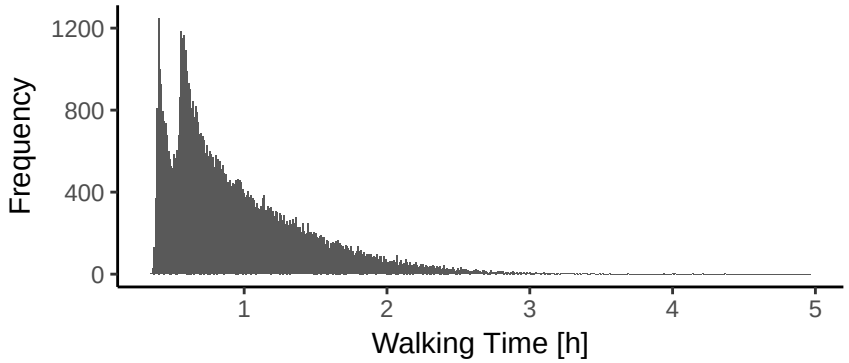

Figure 8: Histogram of the extracted walking times for the grid neighbors.

After the walking time extraction, one big data frame, containing the edges (*ind1*, *ind2*) is created and completed with population values (*p1*, *p2*) for the start and the end point of each edge. Just for completeness the heights (*h1*, *h2*) and the euclidean distance (*dx*) between the points is added. Then the mean slope (m) for each edge is calculated. It is important to state that the edges appear twice in the data frame, this is to represent the two possible directions of the edges.

4.5 Gravity Index

To get a meaningful index of the linguistic influence of two neighbors on each other, a concept from Trudgill (1974) is applied:

$$I_{ij} = \frac{P_i P_j}{d_{ij}^2} \times \frac{P_i}{P_i + P_j}$$

P_i stands for the population size of the data point i (community) and d_{ij} is representing the walking time between the two data points i and j.

Table 4: Some example edges of neighbors in the grid, showing the attribute values.

ind1	ind2	p1	p2	h1	h2	dx	m	t	I
7467	7468	219	188	266.687	255.869	2016.959	0.005	1409.756	130.096
7467	7466	219	1	266.687	273.069	2016.959	-0.003	1446.283	1.176
7467	7391	219	1	266.687	274.250	2015.612	-0.004	1521.379	1.059
7467	7541	219	1	266.687	255.957	2015.612	0.005	1403.339	1.251
7467	7390	219	1	266.687	314.106	2851.459	-0.017	2230.548	0.471
7467	7542	219	3	266.687	253.532	2851.459	0.005	1991.673	1.866

4.6 Weighted Asymmetric Adjacency Matrix

The edge list and the gravity index, representing linguistic influence of two neighbors on each other, is now filled into an weighted asymmetric adjacency matrix. This is done by creating an empty $m \times n$ matrix and then using *ind1* as m and *ind2* as n to fill in the I (gravity index) values.

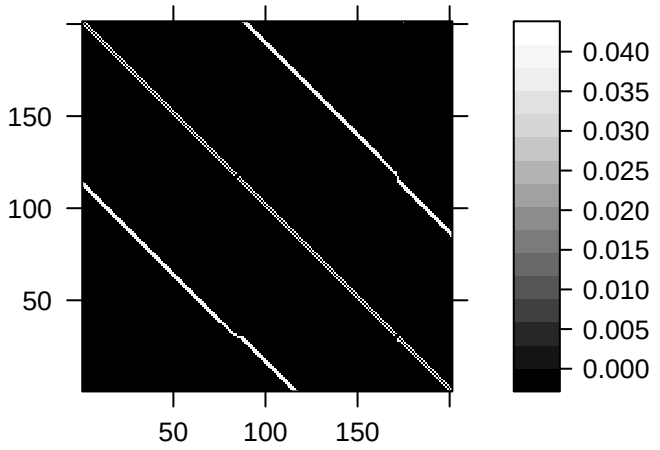

Figure 9: Weighted (linguistic influence) asymmetric adjacency matrix of the grid neighbours. The values in the plot have a logarithmic scale.

5 Spatial Regularization of the Classification Map

5.1 Markov Random Field

The Markov Random Field (MRF) is a statistical method based on a Bayesian approach, which considers spatial dependencies in the decision rule, instead of integrating them in the classification stage. The optimization is done by Iterated Conditional Modes (ICM), which is defined over the minimum energy per data point:

$$y^* = \underset{y \in C}{\operatorname{argmin}} - [\sum_{i=1}^{n} \log(p(y_i|x_i)) + \beta \sum_{i=1}^{n} \sum_{j \in N_i} V(y_i, y_j)]$$

In the MRF the probabilities of the pixel-wise SVM classification $p(y_i|x_i)$ are used as starting point (a priori) for a spatial regularization of the classification map. The spatial dependencies are represented by the term $V(y_i, y_j)$. Because it is not possible to include all the surrounding data points into the decision rule, the influence is limited to local neighborhoods N (also called cliques). In our case the N is given by the nearest grid neighbors. The minima of the global energy is approached by minimizing the local energies in the MRF. A new classification map is obtained by taking the most probable class affiliation y^* (of all possible classes $y \in C$) based on the posterior probabilities of the spatial regularization. (Besag, 1986).

The MRF is constructed using the following components as starting point: (I) probabilities - These are extracted from a prediction done by the SVM trained in the previous part; (II) weighted edges - These are given by the asymmetric adjacency matrix created of the nearest grid neighbors and the index of linguistic influence I_{ij} and (III) the class affiliation of the data points given through the SVM prediction.

5.2 Energy Function

To minimize the global energy of the MRF, there has to be defined an energy function for the local energies of the data points. In this project the energy function definition leans on a paper of Tarabalka et al. (2010). So the the local energy of a data point is defined as:

$$U(x_i) = U_{linguistic}(x_i) + U_{spatial}(x_i)$$

The linguistic energy term only uses the SVM probabilities and is computed as:

$$U_{linguistic}(x_i) = -ln\{P(x_i|L_i)\}$$

And the spatial energy term is calculated using the following equation:

$$U_{spatial}(x_i) = \sum_{x_j \in N_i} \beta I_{ij}(1 - \delta(L_i, L_j))$$

where $\delta(.,.)$ is the Kronecker delta function ($\delta(a,b) = 1$ if $a = b$, and $\delta(a,b) = 0$ otherwise) and β is a parameter that controls the importance of the spatial versus spectral energy terms. I_{ij} is the index of the linguistic influence of between two neighbors, defined in the a previous section.

5.3 Minimizing the Global Energy of the MRF

The global energy of the MRF is defined as the sum of all local energies:

$$E = \sum_{i=1}^{N} U(x_i)$$

The global energy is minimized through the iterative minimization of the local energies. The iterative minimization of the local energies is performed by an adoption of the Metropolis algorithm from Metropolis et al. (1953), based on stochastic relaxation and annealing described by Geman and Geman (1984). Per iteration a randomly chosen data point is assigned to a random class. Based on the observed change $\Delta U = U^{new}(x_i) - U^{old}(x_i)$ of the local energy, the random class assignment is either kept or discarded. If ΔU is negative, the new class assignment is accepted. To avoid the problem of converging to a local minima, also some class changes with a positive ΔU are allowed. But these are only accepted with the probability $p = exp(\frac{-\Delta U}{T})$ (stochastic relaxation), where T (temperature) is a control parameter, which is getting smaller with every iteration (annealing) and thus allows fewer class changes with positive ΔU values (Tarabalka et al., 2010).

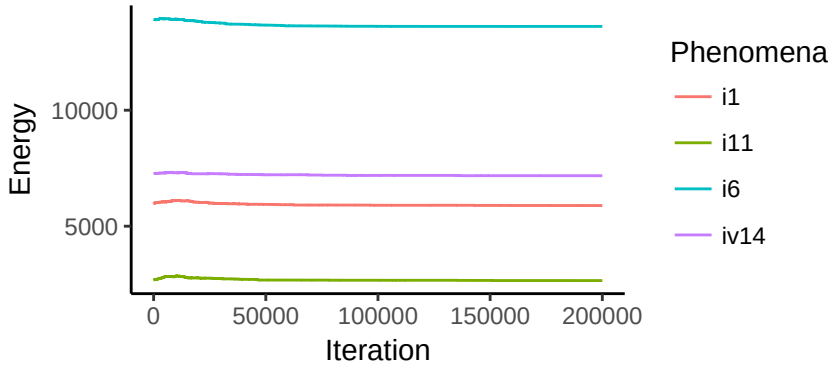

Figure 10: Development of the global energies (in logarithmic scale) during the iterative minimization of the local energies.

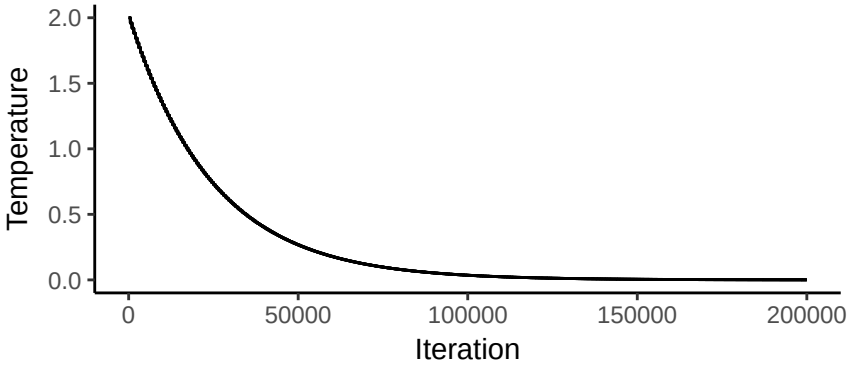

Figure 11: Development of the temperatures during the iterative minimization of the local energies (annealing).

6 Results

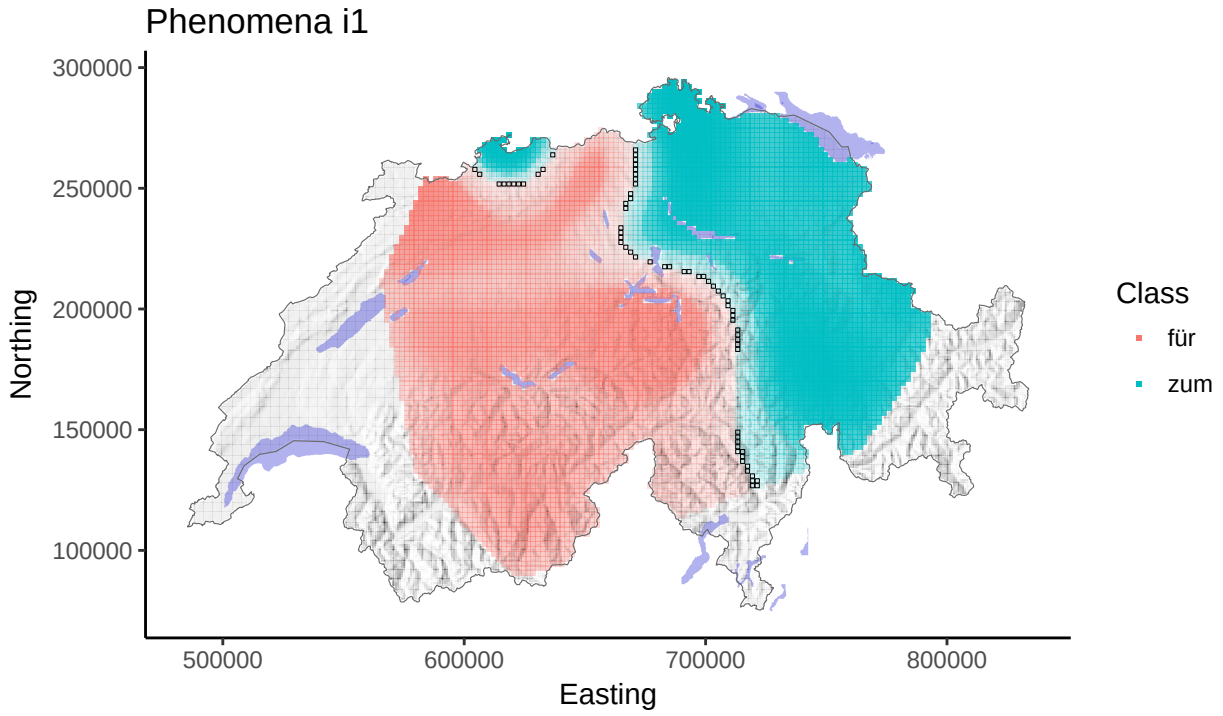

Figure 12: Classification result of the spatial regularization.

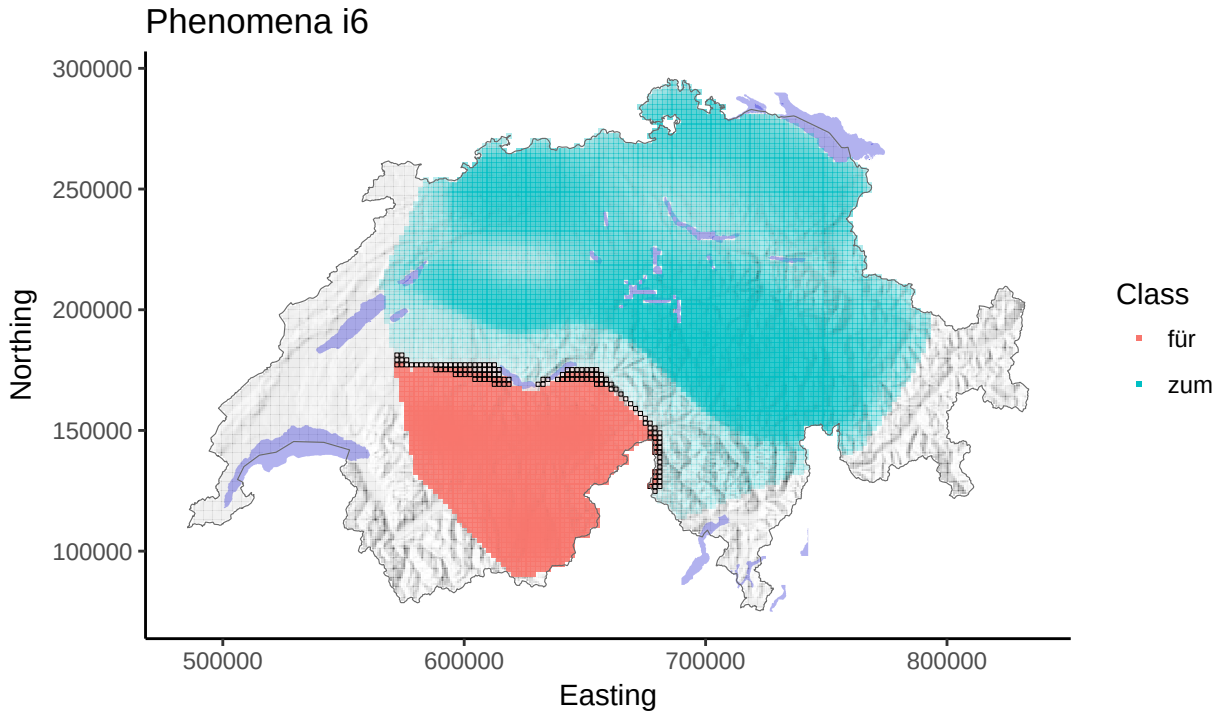

Figure 13: Classification result of the spatial regularization.

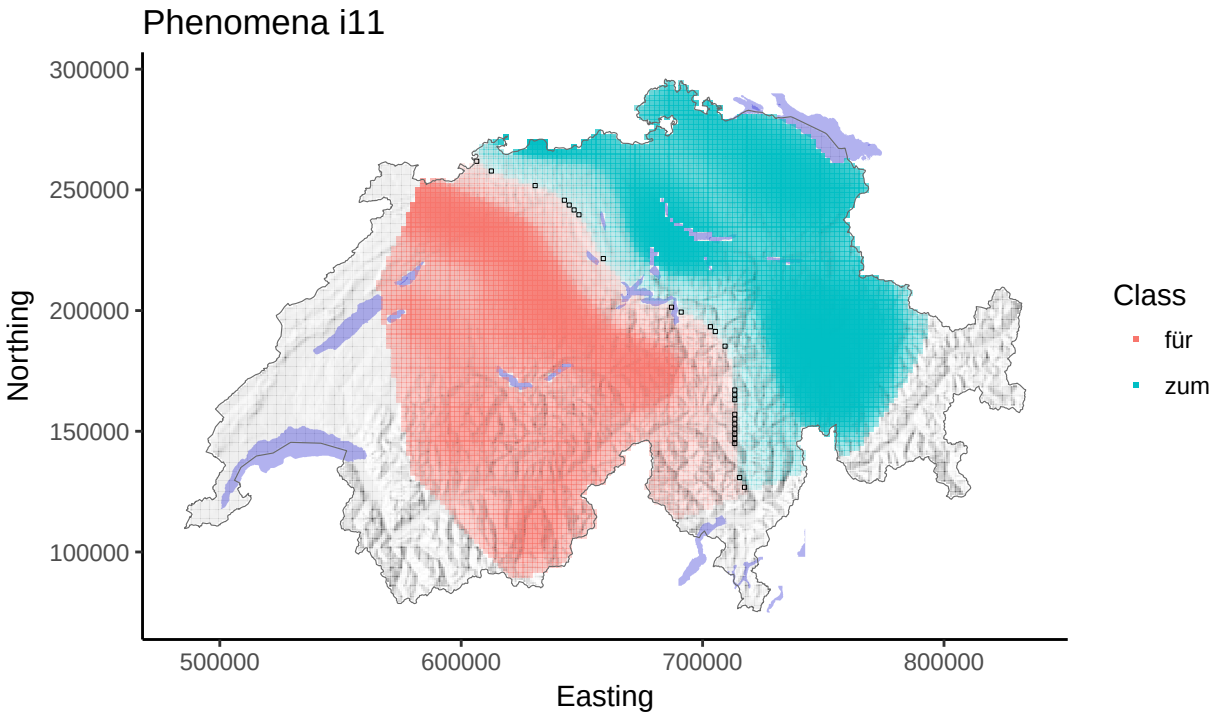

Figure 14: Classification result of the spatial regularization.

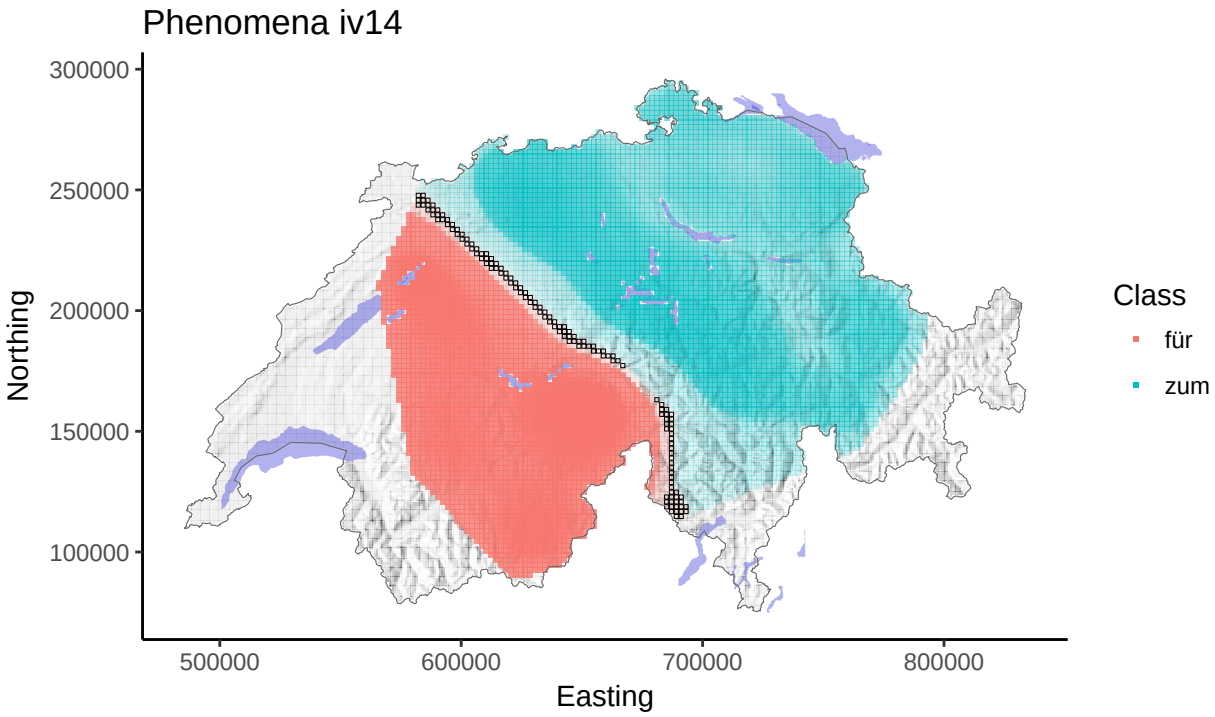

Figure 15: Classification result of the spatial regularization.

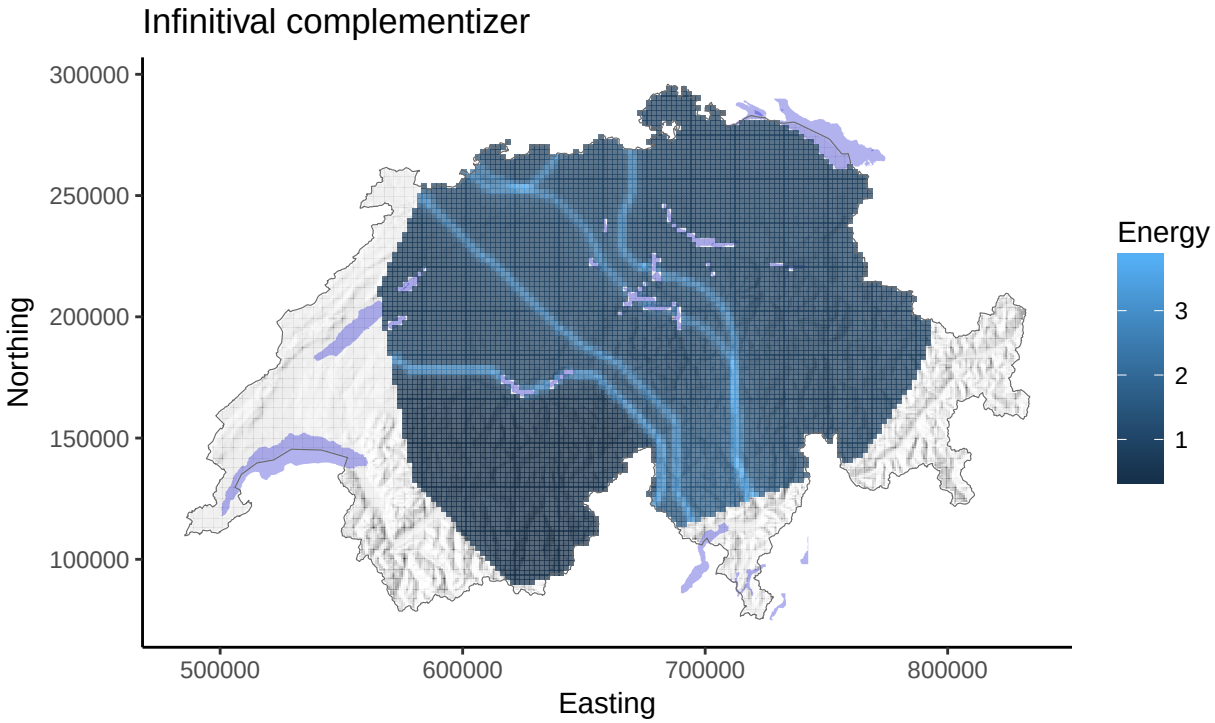

Figure 16: The mean energy of the four phenomena.

References

Besag, J. (1986). On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society*, 48(3):259–302.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2013). *FNN: Fast Nearest Neighbor Search Algorithms and Applications*.

Bivand, R., Keitt, T., and Rowlingson, B. (2016). *rgdal: Bindings for the Geospatial Data Abstraction Library*.

Bivand, R. and Rundel, C. (2016). *rgeos: Interface to Geometry Engine - Open Source (GEOS)*.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with {R}, Second edition*. Springer, NY.

Bucheli, C. and Glaser, E. (2002). The Syntactic Atlas of Swiss German Dialects: Empirical and Methodological Problems. In Barbiers, S., Cornips, L., and van der Kleij, S., editors, *Syntactic Microvariation*, pages 41–73. Meertens Institute Electronic Publications in Linguistics, Amsterdam, 2 edition.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Jeszenszky, P. and Weibel, R. (2016). Modeling transitions between syntactic variants in the dialect continuum. In *The 19th AGILE International Conference on Geographic Information Science, Helsinki (Finnland), 14 June 2016 - 17 June 2016*, number June.

Karatzoglou, A., Meyer, D., and Hornik, K. (2006). Support Vector Algorithm in R. *Journal of Statistical Software*, 15(9):1–28.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal Chemical Physics*, 21(6):1087–1092.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in {R}. *R News*, 5(2):9–13.

Perpiñán, O. and Hijmans, R. (2016). *rasterVis*.

R Core Team (2016). *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...* R package version 0.8-67.

Tarabalka, Y., Fauvel, M., Chanussot, J., and Benediktsson, J. A. (2010). SVM and MRF-Based Method for Accurate Classification of Hyperspectral Images. *IEEE Geoscience and Remote Sensing Letters*, 7(4):736–740.

Tobler, W. (1993). Non-Isotropic Geographic Modeling. *NCGIA Technical Reports*, 93(1):5.

Trudgill, P. (1974). Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society*, 2:215–246.

van Etten, J. (2017). {R} Package {gdistance}: Distances and Routes on Geographical Grids. *Journal of Statistical Software*, 76(13):1–21.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.