Challenges and prospects of spatial machine learning

Habilitationsschrift
zur
Erlangung des akademischen Grades
Doctor rerum naturalium habilitatus (Dr. rer. nat. habil.)
der Agrar- und Umweltwissenschaftlichen Fakultät
der Universität Rostock

Vorgelegt von
Dr. rer. nat. Julian Christian Hagenauer, geb. am 17.10.1980 in Eschwege
aus Stuttgart

Stuttgart, 2022

https://doi.org/10.18453/rosdok_id00004228

Gutachter:
Prof. Dr.Philip Marzahn, Universität Rostock, Agrar- und Umweltwissenschaftliche Fakultät
Prof. Dr. Nguyen Xuan Thinh, Technische Universität Dortmund, Fakultät Raumplanung
Assoc. Prof. Dr. Johannes Scholz, Technische Universität Graz, Geodätisches Institut

Jahr der Einreichung: 2022
Jahr der Verteidigung und Probevorlesung: 2023

Abstract

With the increasing digitalization of our society and the accompanying development of technologies that facilitate the collection, sharing and storage of spatial data, larger amounts of spatial data are available than ever before. However, not only is the sheer amount of spatial data available unprecedented, but also its complexity, dynamics and diversity. This raises the question of the applicability of traditional statistical models, which often cannot be scaled to handle large amounts of data and depend on strict assumptions. In contrast, spatial machine learning is able to model the relationships within large and complex datasets with basically no assumptions and within a reasonable computational time. However, spatial machine learning is also subject to several challenges.

The main objective of this thesis is to improve the usefulness of spatial machine learning for the spatial sciences and to allow its unused potential to be exploited. To achieve this objective, this thesis addresses several important but distinct challenges which spatial machine learning is facing. These are the modeling of spatial autocorrelation and spatial heterogeneity, the selection of an appropriate model for a given spatial problem, and the understanding of complex spatial machine learning models.

First, this these develops spatial machine learning algorithms for spatial prediction and clustering tasks. By accounting for either spatial autocorrelation or spatial heterogeneity, these algorithms allow for more accurate modeling of the data.

Second, this thesis compares the usefulness of different machine learning algorithms for different spatial problems. This comparison not only reveals promising models for the problems at hand and provides further insight into them, but also suggests suitable models for related problems.

Finally, this thesis investigates different approaches that support the understanding of spatial machine learning models. Such approaches are needed because they can suggest potential hypotheses for exploration and thus support the discovery of new knowledge. To provide a broad perspective on different spatial problems, all of the research presented is based on applications from such diverse fields as health, real estate pricing, land-use change, transportation, and spatial planning.

The results of this thesis underpin the usefulness of spatial machine learning for spatial sciences. In particular, they show that the flexibility of spatial machine learning allows the challenges mentioned above to be successfully addressed. Despite these promising results, the contributions of this thesis represent only a portion of a mostly unexplored research area. The complexity of most spatial problems as well as the flexibility of spatial machine learning offer great potential for further research.

Contents

List of figures ix

List of tables xiii

I. Synopsis 1

1. Introduction 3
  1.1. Background 3
    1.1.1. Machine learning 4
    1.1.2. Spatial machine learning 6
    1.1.3. Spatial autocorrelation and spatial heterogeneity 8
    1.1.4. Selecting a model 11
    1.1.5. Understanding a model 14
  1.2. Problem 16
  1.3. Objective 16
  1.4. Research questions 17
  1.5. Structure of the thesis 17

2. Summary of publications 21
  2.1. Clustering contextual neural gas: A new approach for spatial planning and analysis tasks 21
    2.1.1. Motivation 21
    2.1.2. Methods 21
    2.1.3. Results 21
    2.1.4. Conclusion 22
  2.2. Weighted merge context for clustering spatial data 22
    2.2.1. Motivation 22
    2.2.2. Methods 22
    2.2.3. Results 22
    2.2.4. Conclusion 23

2.3. SPAWNN: A toolkit for spatial clustering with artificial neural networks 23
2.3.1. Motivation . . . 23
2.3.2. Methods . . . 23
2.3.3. Results . . . 23
2.3.4. Conclusion . . . 24
2.4. A comparison of machine learning models for predicting travel mode choice . . . 24
2.4.1. Motivation . . . 24
2.4.2. Methods . . . 24
2.4.3. Results . . . 24
2.4.4. Conclusion . . . 25
2.5. Local modelling of land consumption . . . 25
2.5.1. Motivation . . . 25
2.5.2. Methods . . . 25
2.5.3. Results . . . 26
2.5.4. Conclusion . . . 26
2.6. A comparison of machine learning models for predicting land consumption 26
2.6.1. Motivation . . . 26
2.6.2. Methods . . . 26
2.6.3. Results . . . 27
2.6.4. Conclusion . . . 27
2.7. Relative importance of perceived neighborhood characteristics for depression . . . 27
2.7.1. Motivation . . . 27
2.7.2. Methods . . . 27
2.7.3. Results . . . 28
2.7.4. Conclusion . . . 28
2.8. A geographically weighted artificial neural network (GWANN) . . . 28
2.8.1. Motivation . . . 28
2.8.2. Methods . . . 29
2.8.3. Results . . . 29
2.8.4. Conclusion . . . 29

3. Discussion **31**

4. Conclusion **35**

Bibliography **37**

II. Publications **45**

5. Clustering contextual neural gas: A new approach for spatial planning and analysis tasks **47**
Abstract . . . 49
5.1. Introduction . . . 49
5.2. Methodical background . . . 52
5.2.1. Contextual neural gas . . . 52
5.2.2. Competitive Hebbian learning . . . 53
5.2.3. Multi-level modularity optimization . . . 54
5.3. Workflow . . . 55
5.4. Experiments . . . 55
5.4.1. Synthetic data . . . 56
5.4.2. Practical application . . . 57
5.5. Conclusion and further work . . . 62

Bibliography **67**

6. Weighted merge context for clustering and quantizing spatial data with self-organizing neural networks **71**
Abstract . . . 73
6.1. Introduction . . . 73
6.2. Background . . . 75
6.2.1. Neural gas . . . 75
6.2.2. Merge context . . . 76
6.3. Weighted merge context . . . 77
6.4. Case studies . . . 79
6.4.1. Quantization of binary grid data . . . 79
6.4.2. Quantization and clustering of (continuous) real-world vector data 84
6.5. Closing remarks . . . 87

Bibliography **89**

7. SPAWNN: A toolkit for *SP*atial *A*nalysis *W*ith self-organizing *N*eural *N*etworks **91**
Abstract . . . 93
7.1. Introduction . . . 93
7.2. Related work . . . 95

7.3. Self-organizing neural networks 97
    7.3.1. Self-organizing map 98
    7.3.2. Neural gas 98
7.4. Spatial context models 99
    7.4.1. Augmented input vectors 99
    7.4.2. Weighted distance 100
    7.4.3. GeoSOM 100
    7.4.4. Contextual neural gas 100
    7.4.5. Weighted merge context 101
7.5. The SPAWNN toolkit 101
7.6. Performance study 106
7.7. Case study 107
    7.7.1. Outlier analysis 108
    7.7.2. Correlation analysis 110
    7.7.3. Cluster analysis 111
7.8. Conclusion 114

Bibliography **117**

8. A comparative study of machine learning classifiers for modeling travel mode choice **123**
    Abstract 125
8.1. Introduction 125
8.2. Materials and methods 128
    8.2.1. Data 128
    8.2.2. Classifiers 131
    8.2.3. Model comparison 133
    8.2.4. Variable importance 134
8.3. Results 134
    8.3.1. Classification performance 134
    8.3.2. Variable importance 136
8.4. Discussion 137
    8.4.1. Classification performance 137
    8.4.2. Variable importance 141
8.5. Conclusion 142

Bibliography **145**

9. Local modelling of land consumption in Germany with RegioClust 151
Abstract 153
9.1. Introduction 155
9.2. Materials and methods 157
9.2.1. Study area 157
9.2.2. Data 158
9.2.3. Methods 158
9.3. Results 161
9.3.1. Descriptive statistics 161
9.3.2. Model fit 162
9.3.3. Coefficients of RegioClust and GWR 165
9.3.4. Predictions of RegioClust and GWR 168
9.4. Discussion 170
9.5. Conclusion 172

Bibliography 173

10. Assessing the performance of 38 machine learning models: the case of land consumption rates in Bavaria, Germany 181
Abstract 183
10.1. Introduction 185
10.2. Materials 187
10.2.1. Study area 187
10.2.2. Data 187
10.3. Methods 190
10.4. Results 198
10.5. Discussion 203
10.5.1. Strengths and limitations 206
10.6. Conclusion 207

Bibliography 209

11. Relative importance of perceived physical and social neighborhood characteristics for depression: a machine learning approach 217
Abstract 219
11.1. Introduction 221
11.2. Materials and methods 222
11.2.1. Study setting and participants 222

11.2.2. Data 223
11.2.3. Statistical analyses 225
11.3. Results 226
11.3.1. Sample description 226
11.3.2. Model fits 226
11.3.3. Variable importance 226
11.3.4. Correlation assessments 228
11.3.5. Variable interactions 228
11.4. Discussion 232
11.4.1. Main findings in the context of available evidence 232
11.4.2. Strengths and limitations 235
11.5. Conclusions 236

Bibliography **239**

12. A geographically weighted artificial neural network **245**
Abstract 247
12.1. Introduction 247
12.2. Methods 249
12.2.1. Artificial neural network 249
12.2.2. Geographically weighted regression 251
12.2.3. Geographically weighted artificial neural network 252
12.3. Experiments 253
12.3.1. Experiment 1: Synthetic data 255
12.3.2. Experiment 2: House prices in Austria 261
12.4. Conclusion 265

Bibliography **269**

Appendix **275**
Thesen der Habilitationsschrift 275

List of figures

5.1. Synthetic data set . . . 57
5.2. Mean modularity score for the synthetic data set . . . 58
5.3. Clustering results for $l = 7$ . . . 59
5.4. Philadelphia analysis sections . . . 59
5.5. Variables used for the experiment . . . 61
5.6. Mean modularity score for the real-world data set . . . 62
5.7. Clustering results for $l = 2$, $l = 10$, and $l = 13$ . . . 63

6.1. Artificially created binary grid data set . . . 80
6.2. Exemplary probability fields of a prototype k for different distances l . . . 81
6.3. Uncertainty of WMNG for the binary grid data set . . . 83
6.4. Votes for Bush in the 2004 US presidential election . . . 84
6.5. Cartographic maps of the clusters outlined by WMNG . . . 86

7.1. Architecture of the SPAWNN toolkit . . . 102
7.2. View for selecting the self-organizing neural network and the spatial context model . . . 104
7.3. Linked view of a self-organizing neural network and a geographic map of the data . . . 104
7.4. Mean computation times for different numbers of regular attributes . . . 106
7.5. Mean computation times for different numbers of location attributes . . . 107
7.6. Distance matrix and cartographic map of the GeoSOM . . . 109
7.7. Distance-based representation and geographic map of the CNG . . . 110
7.8. GeoSOM component planes for the rates of African Americans and white population . . . 111
7.9. Neurons of the CNG for the rates of African Americans and white population 112
7.10. Clustering results for CNG . . . 113
7.11. Clustering results for the GeoSOM . . . 113

8.1. Accuracy for each classifier . . . 135
8.2. Sensitivity for each classifier . . . 136

8.3. Overall variable importance 138
8.4. Variable importance for MNL 139
8.5. Variable importance for BOOST 139
8.6. Variable importance for RF 140

9.1. Drivers of land consumption for the year 2000 159
9.2. The hierarchical clustering steps of RegioClust 160
9.3. LCRs for 2000–10 162
9.4. AICc statistics for RegioClust 163
9.5. Number of outlying coefficients for RegioClust 164
9.6. LCRs for 2000–10 165
9.7. Estimated local coefficients of RegioClust 166
9.8. Estimated local coefficients of GWR 167
9.9. Predicted LCRs for 2010–20 169

10.1. Workflow of parameter selection, model building, and prediction 197
10.2. Performance boxplots of the cross-validation results 199
10.3. Residuals of the *xgbTree* model and *nnet* model 200
10.4. Predicted LCRs for 2015–2021 201
10.5. Observed LCRS for 2009–2015 and predicted LCRs for 2015–2021 when using *xgbTree* 202
10.6. Variable importance for *xgbTree* 203
10.7. Partial dependence plots for the most relevant variables 204
10.8. Computational times to build the final models 205

11.1. Summary of the crossvalidated model fits 228
11.2. Clustered heat map of variable importance 229
11.3. Partial dependence plots for PHQ-9 230
11.4. Overall variable interaction 231
11.5. Two-dimensional partial dependence plots for PHQ-9 233

12.1. ANN with three layers 250
12.2. GWANN with three layers 252
12.3. Coefficients' surfaces of synthetic datasets 256
12.4. Number of iterations until convergence 258
12.5. Determined bandwidths (fixed) 259
12.6. Estimated prediction error 259
12.7. Estimated coefficient surfaces of GWR 260

12.8. Connection weights of GWANN as surfaces 261
12.9. Number of iterations until convergence 263
12.10. Determined bandwidths (adaptive) 264
12.11. Estimated prediction error 264
12.12. Difference in predicted house prices for GWR and GWANN 266
12.13. Moran's *I* statistics of the residuals 266

List of tables

1.1. Contributions of each publication to the research questions 18

5.1. Mean homogeneity of the attributes for different clusterings 64

6.1. Selected relevant statistics for the trained WMNG 85

8.1. Description of the variables 129
8.2. Descriptive statistics of the trip data set 130
8.3. Results of Wilcoxon rank-sum tests for differences in accuracy 135

10.1. Descriptive statistics for 2009 and 2015 189
10.2. Summary of the tested models 191

11.1. Sample characteristics 227

Part I.

Synopsis

1. Introduction

1.1. Background

The amount of data that is available has increased dramatically in recent years. This trend can be attributed to the increasing digitization of our society. More and more of our activities are now digitally recorded (Goodfellow et al., 2016) and since most of these activities take place in a spatial context, these recorded data are increasingly spatial in form. This has been made possible with the advent of location-aware technologies such as Global Positioning Systems (GPS) (e.g., Zumberge et al., 1997), sensors carried by individuals in mobile phones, attached to vehicles, or embedded in infrastructures (e.g., Johnson and Trivedi, 2011), remote sensors carried by airborne and satellite platforms (e.g., Asner et al., 2012), and georeferenced social networks (e.g., Sui and Goodchild, 2011). The availability of large amounts of spatial data has led to a change in the spatial sciences from a data-scarce to a data-rich environment (Miller and Goodchild, 2015). Making use of this data can offer unprecedented possibilities to find solutions for important problems in relevant fields and to improve our understanding of the world.

While traditional statistical methods have long been the standard tool for solving spatial problems, their applicability is becoming limited for two main reasons. Firstly, most of these methods were developed at a time when there was no need for methods that could be scaled to handle large datasets; that is, large not only in terms of the records represented, but also in terms of the dimensions over which the data were gathered. When traditional statistical methods are applied to these datasets, the computing time required is a hurdle that is difficult to overcome (e.g., Gahegan, 2000; Miller and Han, 2009). Secondly, most of these methods rely on strict assumptions about the data. Spatial data, however, have inherent special properties which hardly met the assumptions of most traditional statistical methods (e.g., Anselin, 1989; Gahegan, 2000). For instance, relationships often vary by location (e.g., Fotheringham, 2009) and are nonlinear (Openshaw, 1999); non-normal and complex data distributions can be expected (Openshaw, 1999); and observations tend not to be independent of each other (e.g., Getis, 2010). In addition, since spatial datasets often originate from different sources, these properties may vary in terms of the dimensions over which the data are gathered.

One promising approach to overcoming these limitations is the use of machine learning. Machine learning is a computational approach to modeling relationships within large and complex datasets with essentially no assumptions and within a reasonable computational time. It allows researchers to tackle problems involving knowledge of the real world and to make decisions that appear subjective (Goodfellow et al., 2016). Subjective in this context means that the calculations that led to these decisions are not immediately apparent to the analyst. The spatial extension of machine learning is referred to as spatial machine learning. Spatial machine leaning is particularly promising for the spatial sciences, as it is able to respond specifically to the nuances of different locations, given that information about the spatial structure of the data is provided in an adequate way (Gahegan, 2003).

Despite the advantages of spatial machine learning, it is not able to overcome all of the challenges faced by traditional statistical models. For instance, the issue of how to effectively model spatial autocorrelation and spatial heterogeneity with spatial machine learning is still an open question. In addition, spatial machine learning also introduces certain new challenges to spatial sciences. For instance, due to the plethora of available machine learning models, it is usually unclear which one is most useful for solving a certain spatial problem. Developing an understanding of most machine learning models is also challenging, because their inner workings can rarely be traced by human beings. Overcoming these challenges is likely to improve the usefulness of spatial machine learning for the spatial sciences and to allow its unused potential to be exploited.

The rest of this section is structured as follows. A formal introduction to machine learning is given in subsection 1.1.1, and building on this, spatial machine learning is described in subsection 1.1.2. Subsection 1.1.3 discusses the challenges of modeling spatial autocorrelation and spatial heterogeneity, subsection 1.1.4 addresses the issue of model selection, and finally subsection 1.1.5 focuses on the difficulty of understanding a machine learning model.

1.1.1. Machine learning

The term machine learning was introduced by Samuel (1959) in an article in which he proposed an algorithm for playing checkers. A key feature of this algorithm was that it improved itself by experience, based on games played previously. A formal definition of machine learning was provided by Mitchell (1997): "A computer program is said to learn from experience E with respect to some task T and some performance measure

P, if its performance on T, as measured by P, improves with experience E". While exact definitions of T, P, and E are still lacking, informal descriptions and examples are provided in the following.

Task T

The overall aim of machine learning is to solve a task T that is usually too difficult to solve with fixed algorithms written by human beings (Goodfellow et al., 2016). There are numerous different kinds of tasks that can be solved by machine learning; however, learning does not in itself represent a task of machine learning but a means of solving a task.

The most common machine learning tasks are regression, classification, and clustering. In a regression task, the machine learning algorithm is asked to predict a real-valued number for some given observation. Examples of regression tasks include forecasting the energy consumption of buildings (e.g., Neto and Fiorelli, 2008) and modeling the individual credit risk of customers (e.g., Kruppa et al., 2013).

In a classification task, the machine learning algorithm is asked to determine one of several pre-specified categories to which an observation belongs. Classification is closely related to regression, except that the format of the output is categorical rather than numerical. Examples of classification tasks include sentiment analysis of texts (Ye et al., 2009) and object detection in images (Zhao et al., 2019).

Finally, in a clustering task, the machine learning algorithm is asked to organize observations into groups based on their similarity (Jain et al., 1999). The aim is that these groups should represent the data as closely as possible. Examples of clustering tasks include the identification of similar protein sequences (e.g., Steinegger and Söding, 2018) and social network analysis (e.g., Cai et al., 2015).

Performance measure P

A performance measures P quantitatively evaluates the ability of a machine learning algorithm to solve a given task T. For regression tasks, the mean square error is often used for evaluation, while the classification performance is often measured by accuracy. The latter evaluates the proportion of correct predictions among the total number of observations examined. For clustering tasks, a metric that is often used is the within-cluster variance. This metric evaluates the distances of the observations from the mean of the clusters to which they are assigned.

Experience E

Machine learning algorithms are allowed to experience some data which enables them to learn to solve a given task. In order to support the experience, it is important that the data is appropriately encoded and presented to the algorithm. Commonly, the data is organized into a dataset consisting of a collection of observations, each of which consists of a set of variables.

Depending on the kind of experience E they are allowed to have during the training, machine learning algorithms can be broadly categorized into supervised and unsupervised approaches (Goodfellow et al., 2016). In supervised learning, the algorithm tries to learn to predict a target value from the observations when experiencing the dataset. To allow this process to take place, each observation in the dataset must be assigned a target value. The term "supervised learning" originates from the view that the target value is provided by a teacher who also gives feedback about the learning progress of the machine learning algorithm (Goodfellow et al., 2016).

In unsupervised learning, the machine learning algorithm tries to learn the useful properties of the structure of the dataset when experiencing a dataset; in this case, there is no teacher and the machine learning algorithm has to make sense of the data on its own.

Regression and classification tasks are usually solved using supervised learning, whereas clustering tasks are tackled using unsupervised learning. The distinction between supervised and unsupervised learning, though, can be blurred and other learning paradigms exist. For instance, in reinforcement learning, a machine learning algorithm does not experience a fixed dataset but interacts with an environment, so that there is a feedback loop between the machine learning algorithm and its experience (Goodfellow et al., 2016); however, these algorithms are outside the scope of this thesis.

1.1.2. Spatial machine learning

The current literature does not provide a common definition of spatial machine learning. In fact, the terms spatial machine learning, geocomputation, and data-driven geography are often used rather loosely and even interchangeably. In general, spatial machine learning can be considered a spatial extension of aspatial machine learning. By adapting Goodchild's (2017) definition of spatial data analysis, it can be said that spatial machine learning differs strictly from aspatial machine learning in that its results are not invariant under relocation of the observations. Or in other words, space matters in spatial machine learning. More formally, the definition of machine learning put forward by Mitchell (1997) can be extended to spatial machine learning in the sense that it is additionally required that the task T and the experience E must be spatial. Whether the performance measure P must also be spatial depends on T and E. An informal description and examples of what is meant by spatial T, spatial P, and spatial E is provided in the following.

Spatial task T

In a spatial machine learning task T, the algorithm takes the spatial structure of the observations into account when processing them. In an analogous way to aspatial tasks, the most common spatial tasks are spatial regression, spatial classification, and spatial clustering.

In a spatial regression task, the machine learning algorithm is asked to predict a real-valued number for a given observation, taking into account the observation's spatial location. Examples of spatial regression tasks include hedonic house price modeling (e.g., Helbich et al., 2014) and landslide susceptibility mapping (e.g., Kavzoglu et al., 2019).

A special case of spatial prediction is spatial interpolation, which is the task of predicting numeric values at arbitrary locations within a region defined by given observations. It is commonly used to create surface maps of some quantities. Examples of spatial interpolation include the estimation of particulate matter concentrations (e.g., Sampson et al., 2013) and the mapping of organic carbon in soil (Kerry et al., 2012).

Spatial classification is similar to spatial regression except that the format of the output is categorical. Examples of spatial classification tasks include the prediction of travel mode choices (e.g., Hagenauer and Helbich, 2017) and the modeling of changes in land use (e.g., Pijanowski et al., 2002).

Finally, in a spatial clustering task, the machine learning algorithm is asked to organize observations into groups based on their similarity, taking into account the spatial location of each observation. Analogously to aspatial clustering, the aim is that the groups should represent the data as closely as possible, but also with respect to space. Examples of spatial clustering tasks are the identification of counties with similar socio-economic characteristics (e.g., Hagenauer, 2015) and the detection of crime hotspots (e.g., Nakaya and Yano, 2010).

A special case of spatial clustering is regionalization. In this case, it is additionally required that the resulting groups of observations form contiguous regions that partition the space (Guo, 2008; Haining, 2003). The resulting regions typically allow for a more intuitive perception and thus are particularly useful in supporting a decision making process. Examples of regionalization include the identification of housing markets (e.g., Helbich et al., 2014) and the outlining of regions for the analysis of cancer risk (e.g., Wang et al., 2012).

Spatial performance measure P

A spatial performance measure P quantitatively evaluates the ability of a machine learning algorithm to solve a given spatial task T. For this purpose, it is often useful to evaluate the degree to which the machine learning algorithm has been successful in taking into account the spatial properties of the data.

For regression tasks, it is often useful to evaluate the spatial autocorrelation of the residuals in addition to the mean square error. If the degree of spatial autocorrelation between the residuals is high, this indicates that the machine learning algorithm failed to capture the spatial properties of the dataset (Fotheringham, 2009; Getis, 2010).

For clustering tasks, in addition to the within-cluster variance, it is often useful to measure the degree of spatial closeness between the observations in each group (cluster). If they are not spatially close, this indicates that the clustering model is wrong, as distant observations usually represent measurements of different spatial processes (spatial heterogeneity).

Spatial experience E

A spatial experience E enables a machine learning algorithm to learn information to solve a spatial task T. For this purpose, it is important to consider how the machine learning algorithm experiences space, or, in other words, how spatial information is provided to the algorithm (Gahegan, 2000). If space is not adequately represented, the machine learning algorithm might not be able to learn the information necessary to solve the spatial task at hand. This can be done, for instance, by using a weight matrix or by encoding the location of the observations using one or more additional variables. More details about this issue are provided in subsection 1.1.3. Apart from this property, there are no substantial formal differences between spatial and aspatial experience.

1.1.3. Spatial autocorrelation and spatial heterogeneity

In practice, observations that are spatially close tend to be more similar (i.e., less independent of each other) than observations that are far apart. This property is commonly referred to as (positive) spatial autocorrelation. [1]

Spatial autocorrelation is important to geography because it governs many spatial processes. Its importance led Tobler (1970) to phrase this principle in terms what he referred to as the first law of geography: "Everything is related to everything else, but near things are more related than distant things". Spatial autocorrelation is similar to temporal autocorrelation, as present observations depend on past observations, but can be considered more complex since it affects more than one dimension (Anselin, 1989; Griffith, 1993).

In general, when observations are spatially autocorrelated, they share a certain amount of information and thus the entire dataset is less informative; in other words, a larger dataset of spatially dependent observations is required in order to obtain the same degree of information as if the observations where spatially independent (Anselin, 1989). Also, spatial autocorrelation violates the assumption of independence of observations upon which most traditional statistical methods are predicated. As a consequence, significance tests and measures of model fit may be misleading if spatial autocorrelation is not taken into account (Anselin and Griffith, 1988). A number of spatial regression methods have been proposed that consider this aspect, with the most common being spatial lag, spatial error, and spatial Durbin models (e.g., LeSage, 2008). All of these models use a spatial weight matrix to represent the spatial dependence between the observations. The difference between them is that spatial lag models assume that spatial autocorrelation is present in the dependent variable, spatial error models assume that spatial autocorrelation is present in the error term, and spatial Durbin models assume that spatial autocorrelation is present in both the dependent variable and the independent variables.

Nevertheless, spatial autocorrelation is not only a nuisance, but can also serve as a valuable source of information about a spatial process. While it does not necessarily imply causality, it provides evidence for causality, which can be assessed in the light of theory or other forms of evidence (Miller, 2004). Spatial autocorrelation can also be exploited for spatial interpolation tasks, where the value of an arbitrary location within a region is estimated using observations that are spatially close, based on the assumption that they are subject to the same spatial process.

Measures of spatial autocorrelation are particularly useful in terms of describing the spatial structure of the data and verifying and validating the assumptions of a model. A

---

[1] In theory, nearby observations can also tend to be less similar than observations that are far apart (negative spatial autocorrelation), although this kind of relationship is very rare in practice and is often the result of measurement at a scale that is much greater than that at which the process operates (Getis and Ord, 1992). Henceforth in this thesis, all references to spatial autocorrelation refer to positive autocorrelation.

variety of measures of spatial autocorrelation have been developed in the past. Of these, the most commonly used are Moran's I (Moran, 1950), Geary's C (Geary, 1954), and variograms (Cressie, 1993), although the usefulness of these measures is limited when spatial autocorrelation shows instability in the form of spatial heterogeneity, spatial drift, or spatial regimes (Anselin, 1996). This limitation motivated the development of local measures of spatial autocorrelation that decompose the underlying global measures by location, the most common of which are the local Moran's I (Anselin, 1995) and the Getis-Ord statistic (Getis and Ord, 1992).

Spatial heterogeneity (or spatial non-stationarity) is closely related to spatial autocorrelation. It refers to the property of spatial processes to vary by location due to the intrinsic degree of uniqueness exhibited by every location. As a consequence, global parameters are not appropriate to describe spatial heterogeneous processes. More formally, Anselin and Griffith (1988) define spatial heterogeneity as structural instability in the form of systematically varying model parameters or different response functions.

Ignoring spatial heterogeneity can cause biased parameter estimation, misleading significance levels and suboptimal predictions (Anselin and Griffith, 1988). A number of spatial regression methods have been proposed to account for spatial heterogeneity. Notable examples include the expansion method (Casetti, 1972), weighted spatial adaptive filtering (Gorr and Olligschlaeger, 1994), Eigenvector spatial filtering (Griffith, 2003), and geographically weighted regression (GWR) (Brunsdon et al., 1996).

In general, spatial machine learning does not assume spatial independence of observations; however, the redundancy of information due to spatial autocorrelation may make the learning of patterns more difficult if spatial autocorrelation is not appropriately taken into account. Spatial machine learning also does not typically assume spatial stationarity of the relationships between observations. However, a machine learning algorithm might not be able to learn significant patterns or may produce suboptimal predictions if spatial heterogeneity is not considered. The question of how to account for spatial autocorrelation and spatial heterogeneity is one of the main challenges associated with spatial machine learning.

In order to take into consideration the spatial autocorrelation and/or spatial heterogeneity, information about the spatial structure of the observations must be provided to the machine learning algorithm. A straightforward approach is to provide such information in the form of additional independent variables; these variables do not directly reflect the complete spatial structure of the observations, but allow the model to learn the spatial relationships that are needed to perform its task. For instance, Credit (2021) uses a random forest model with a spatial lag term in order to account for spatial autocorrelation, whereas Georganos et al. (2019) use a random forest model which includes the coordinates of the observations as additional independent variables in order to account for spatial heterogeneity. While this approach allows to use aspatial machine learning algorithms in order to solve spatial problems, it is then often unclear and one has little control over exactly how the algorithms take spatial autocorrelation or spatial heterogeneity into account. In fact, for complex machine learning algorithms like random forests or artificial neural networks, which are essentially black box models, it is not immediately clear that using the coordinates of the observations as additional independent variables accounts for spatial autocorrelation rather than spatial heterogeneity or even both.

A more elaborated approach is to provide information about the spatial structure of the observations in the form of a weight matrix (Getis, 2009). For instance, Fotheringham (1998) and Hagenauer and Helbich (2022) utilize a weight matrix in order to consider spatial heterogeneity when building a regression model, whereas Hagenauer (2016) utilizes a weight matrix in order to consider spatial autocorrelation when performing spatial cluster analysis. The usage of a weight matrix enables detailed control on how the spatial structure of the observations is represented. In particular, it enables the representations of very different and complex spatial structures (e.g., non-Euclidean distance relationships). However, in order for spatial machine learning algorithms to be able to use the information provided in the form of weight matrices, they must be specifically designed for this purpose; aspatial machine learning algorithms cannot be readily used. Also, it must be noted that using weight matrices can be computationally expensive, in particular when the number of observations is large (Miller and Wentz, 2003).

1.1.4. Selecting a model

The number of available machine learning algorithms is huge and ever increasing (Fernández-Delgado et al., 2014), which raises the question of how to select an algorithm. The *no free lunch theorem* states that for optimization problems, the average performance of any pair of algorithms across all possible problems is identical (Wolpert and Macready, 1997). Since most machine learning problems can be reduced to optimization problems (i.e., they seek to minimize some cost function) (Bennett and Parrado-Hernández, 2006), this theorem also applies to machine learning. It is important to note, though, that the theorem only holds on average across all possible problems; it does not exclude the possibility that for certain classes of problems there exist algorithms that are better than others. Hence, when selecting a machine learning algorithm, it is not the goal to find the best algorithm for all kind of tasks but one that performs well for carrying out a particular task at hand.

For regression and classification tasks, one is usually interested in how well a machine learning algorithm performs on data it has not seen before, since this determines how well it will work when it is deployed in the real world (Goodfellow et al., 2016). In other words, one is usually interested in the ability of the algorithm to generalize to previously unseen observations.

In order to evaluate the generalization performance of a machine learning algorithm, it is useful to distinguish between the dataset that was used for learning and an independent test dataset, which is independent from the training dataset but identically distributed (i.e., both are subject to the same probability distribution) (Goodfellow et al., 2016). Using the performance measures presented in subsections 1.1.1 and 1.1.2, one can then calculate the error for a machine learning algorithm using either the training dataset (training error) or test dataset (test error). The test error is an estimate of the generalization performance of the machine learning algorithm. The process of partitioning the available dataset into training and test datasets and then subsequently measuring the generalization performance is commonly referred to as cross-validation (Hastie et al., 2009).

When training a machine learning algorithm, one is interested in minimizing both the training and test errors. These errors, however, are not independent of each other; if the training error is small, it often results in increase in the test error, meaning that the machine learning algorithm is *overfitting* the training data. If machine learning algorithm is not able to reduce the training error, it is *underfitting* the training data. Under- and overfitting are two central challenges of machine learning (Goodfellow et al., 2016).

If few data are available and hence the size of the test dataset is small, cross-validation tends to produce unreliable estimates of an algorithm's generalization performance. As a way around this problem, k-fold cross-validation was proposed, in which the data are randomly partitioned into k disjoint subsets. One subset at a time is used to test the performance of the algorithm, while the others are used for training. Then, the mean performance over all folds is reported. One problem with this procedure, however, is that for $k > 2$ the datasets used for training are not independent of each other (i.e., they overlap), resulting in a biased estimate of the performance (Dietterich, 1998).

When using cross-validation with spatial data, there is a risk that spatially close and thus spatially autocorrelated observations will be assigned to the test and training datasets. As a consequence, the test and training datasets are not independent of each other and hence the resulting estimate is too optimistic. Two main approaches for dealing with this issue can be distinguished. In the first approach, observations are assigned to the training and test datasets in such a way that the distance between each observation in the test dataset and each observation in the training dataset exceeds a certain threshold (e.g., Brenning, 2005; Le Rest et al., 2014). This threshold is determined so that observations whose distance from each other exceeds the threshold can be considered spatially independent. One drawback of this approach is that due to the distance constraint, it is often impossible to assign all available observations to either the test or training dataset, which affects the resulting estimate of generalization performance. This is a particular problem when dealing with spatial datasets that are small and in which the observations are highly spatially autocorrelated.

In the second approach, the study area is partitioned into a number of spatial regions, for example using k-means clustering (e.g., Brenning, 2012) or a predefined partitioning scheme (e.g., Bahn and McGill, 2013; Wenger and Olden, 2012). The observations within each of the k regions of the partitioning represent disjoint subsets of observations. These subsets are then used with the basic k-fold cross-validation procedure, rather than randomly chosen subsets, for performance estimation. Since observations in different regions tend to be distant from each other, they are also mostly spatially independent from each other. However, in contrast to the first approach, this approach does not guarantee spatial independence. For instance, if two regions are close to each other (e.g., they share a border), some observations within these regions are necessarily also close to each other and are therefore likely to be spatially dependent.

Another risk which is often not considered when using cross-validation with spatial data is that the data generating process may vary by location (i.e., it may be spatially heterogeneous). If this is the case, a global estimate of the generalization performance is not appropriate, as it obscures spatial variations which might be of interest. Instead, it would be more useful to spatially disaggregate the cross-validation procedure to obtain local estimates of the generalization performance. How exactly this could be done, though, is still open to research.

When aiming to solve a task using machine learning, one is usually interested in selecting the model with the best generalization performance from a set of models as well as in obtaining an unbiased estimate of the generalization performance of the finally selected model. When cross-validation is used naïvely for both purposes, there is the risk that the estimate of the generalization performance of the finally selected model will be overly optimistic (i.e. biased), as the same test dataset is used twice: once for selecting the best model and once for estimating its generalization performance. To avoid this problem, a common approach is to split the original training dataset once more into a smaller training dataset and a validation dataset. This approach is termed nested cross-validation. The resulting training dataset is used to build the models, whereas the validation dataset is used to estimate the generalization performance of the models when comparing them for the purpose of model selection, and the test dataset is used to obtain an unbiased estimate of generalization performance of the finally selected model. It should be mentioned that nested cross-validation can also be used with folded cross-validation and spatial cross-validation.

Most machine learning algorithms have several settings that control the behavior of the learning algorithm (Goodfellow et al., 2016). These are called hyperparameters and are usually not learned by the algorithm but are chosen prior to training. Two models built using the same learning algorithm and the same training data but with different hyperparameters are considered to be different models. Hence, the problem of selecting an appropriate model arises not only when choosing a machine learning algorithm but also when determining hyperparameters.

1.1.5. Understanding a model

In general terms, understanding a model refers to comprehending the relationships of a model, i.e., why it produced a certain output for a given input. The internal logic and inner workings of most machine learning models are too complex for human beings to understand (Carvalho et al., 2019); they are black box models that are easier to experiment with than to understand (Golovin et al., 2017). However, understanding a machine learning model is important for several reasons. Firstly, understanding a machine learning model increases the trust of a user in the model's predictions and hence the willingness to follow the recommendations associated with those predictions (Freitas, 2014). This is a particular concern in applications where the predictions of the model can have a critical effect on high-stakes decision making, such as in medicine (e.g., Holzinger et al., 2019) or criminal justice (Fisher et al., 2019). Secondly, understanding a machine learning model makes it easier to find and correct errors, since the causes of these errors can be directly traced by a human. This ultimately helps to develop more accurate and truthful machine learning models. Finally, understanding a machine learning model supports the discovery of new knowledge, as it gives insights into the learned relationships. While these relationships do not imply causality, they may suggest potential hypotheses and theories for a researcher to explore and hence serve as building blocks for knowledge discovery.

Numerous methods have been proposed that can enhance the understanding of machine learning models by providing explanations of their reasoning. These methods can be broadly categorized based on the kind of explanation they provide (Molnar, 2019). Firstly, some methods explain a model by providing summaries of the model's relationships; examples include variable importance statistics (e.g., Wei et al., 2015) and partial dependence plots (Friedman and Meulman, 2003). Secondly, others explain a model by providing summaries of the model's learned parameters. Examples include the parameters of a linear model and the connection weights of a self-organizing map (e.g., Vesanto, 1999). Thirdly, some methods explain a model by providing examples or counterfactual examples that are meaningful and can be interpreted to explain the model's relationships (e.g., Kim et al., 2016). These kind of methods are particularly useful if the input consists of structured data, such as images of text. Finally, some methods utilize an intrinsically interpretable surrogate model to approximate the relationships of a more complex model, at either a local or a global level (e.g., Ribeiro et al., 2016). The surrogate model then provides an explanation of the more complex model.

The learned relationships of a spatial machine learning model also reflect the spatial structure of the observations in some form. For instance, they may vary by location (spatial heterogeneity) or may refer to the similarity of nearby observations (spatial autocorrelation). It is essential to account for these kinds of relationships when explaining a spatial machine learning model, as they provide evidence for how the output of the model is affected by the spatial structure of the observations. In practice, due to the complexity of the spatial structure of the observations, the relationships of a spatial machine learning model also tend to be complex, which makes an explanation difficult. In some cases, it is possible to explain the relationships of the model using maps, which are considered to be particularly effective for communicating spatial information (MacEachren, 2004). For instance, Brunsdon et al. (1996) use maps to explain the spatial variation in learned model parameters.

It should be noted that explanation methods generally provide summaries or show trends in the relationships of a model that are not completely faithful to the actual relationships (Rudin, 2019). If they were, the explanation would be equal to the model, which would then not be needed in the first place. In other words, the explanations reduce an original model to a simplified form for presentation and examination. As a consequence, there is a risk that the provided explanations will be misleading and that the understanding of the model will be wrong. However, it can be argued that these explanation methods at least give some evidence of the actual relationships of a model and thus can still be of use in terms of understanding the model. One way to avoid this issue completely is to use models that do not require an explanation because they are simple enough to be understood by a human. This kind of model are commonly referred to as interpretable machine learning models (Rudin, 2019). Interpretable machine learning models are usually constrained in terms of model form, such that they are either useful to someone or obey structural knowledge of the domain, such as monotonicity, causality, structural constraints, or physical constraints that arise from the domain (Rudin, 2019). Examples of interpretable machine learning models include linear regression, logistic regression, and decision trees (Molnar, 2019).

1.2. Problem

Machine learning algorithms have already attracted significant attention in terms of applications in the spatial sciences and some progress has been made (e.g., Gahegan, 2003; Miller and Han, 2009). In particular, continuous efforts have been made to make spatial machine learning models more accurate (Gahegan, 2017) and to make the learned relationships more explicit in order to provide insights into the data generating process (e.g., Georganos et al., 2019). The limitations and potentials of spatial machine learning, though, are still unknown and have yet to be defined.

1.3. Objective

The main objective of this thesis is to improve the usefulness of spatial machine learning for the spatial sciences and to allow its unused potential to be exploited. With this in mind, this thesis focuses on three important challenges which spatial machine learning is facing: modeling of spatial autocorrelation and spatial heterogeneity, model selection for certain spatial tasks, and effective means of understanding models.

More specifically, while aspatial problems can often be solved by basic machine learning algorithms, solving spatial problems typically requires consideration of the special properties of spatial data, so that the algorithm is able to learn the relevant relationships accurately. How to effectively account for these special properties is often unclear. Therefore, this thesis introduces new spatial machine algorithms for spatial prediction and spatial clustering tasks which take into account the special properties of spatial data. In addition, although the literature already proposed numerous machine learning algorithms with different properties and capabilities (Fernández-Delgado et al., 2014), it is often not clear which algorithm is most suitable for solving a particular spatial task. This thesis therefore explores the usefulness of a broad set of different machine learning algorithms for certain spatial tasks. Finally, there is a need for spatial machine learning to move from prediction to understanding (Gahegan, 2020). Therefore, this thesis investigates different means of explaining the learned relationships of spatial machine learning models.

1.4. Research questions

To achieve the objectives of this thesis, the following specific research questions will be addressed:

1. How to account for spatial heterogeneity in spatial prediction tasks?
2. How to account for spatial autocorrelation in spatial clustering tasks?
3. How to account for spatial heterogeneity in spatial clustering tasks?
4. How can spatial clusters with complex structures be effectively outlined?
5. Which machine learning algorithms are useful for spatial prediction tasks?
6. Which machine learning algorithms are useful for spatial clustering tasks?
7. Which approaches are useful to explain spatial machine learning models?
8. How can location be used to support the explanation of spatial machine learning models?

1.5. Structure of the thesis

This thesis draws on eight publications in order to address the aforementioned research questions. Of these, seven are research articles that have been published in peer-reviewed scientific journals and one is a peer-reviewed book chapter. These publications are listed below.

1. J. Hagenauer (2015). Clustering contextual neural gas: A new approach for spatial planning and analysis tasks. In: M. Helbich, J. J. Arsanjani, and M. Leitner (eds.). *Computational Approaches for Urban Environments.* Springer, 77–94.
2. J. Hagenauer (2016). Weighted merge context for clustering and quantizing spatial data with self-organizing neural networks. *Journal of Geographical Systems*, 18(1), 1–15.
3. J. Hagenauer and M. Helbich (2016). SPAWNN: A toolkit for spatial analysis with self-organizing neural networks. *Transactions in GIS*, 20(5), 755–774.

Table 1.1.: Contributions of the publications to the research questions.

Publication	Research question							
	1	2	3	4	5	6	7	8
1			X	X		X		
2		X				X		
3		X	X	X		X	X	X
4					X		X	
5	X			X	X		X	X
6	X				X		X	X
7					X		X	
8	X				X		X	X

4. J. Hagenauer and M. Helbich (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, 273–282.

5. J. Hagenauer and M. Helbich (2018). Local modelling of land consumption in Germany with RegioClust. *International Journal of Applied Earth Observation and Geoinformation*, 65, 46–56.

6. J. Hagenauer, H. Omrani, and M. Helbich (2019). Assessing the performance of 38 machine learning models: The case of land consumption rates in Bavaria, Germany. *International Journal of Geographical Information Science*, 33(7), 1399–1419.

7. M. Helbich, J. Hagenauer, and H. Roberts (2020). Relative importance of perceived physical and social neighborhood characteristics for depression: A machine learning approach. *Social Psychiatry and Psychiatric Epidemiology*, 55, 599–610.

8. J. Hagenauer and M. Helbich (2022). A geographically weighted aritifical neural network. *International Journal of Geographical Information Science*, 36 (2), 215–235.

Each publication contributes to answering more than one research question. Table 1.1 summarizes which publication addresses which research question and how these publications relate to each other with respect to the research questions.

This thesis is divided into two distinct parts. The first part (synopsis) provides a summary that places the perspectives and contributions of each publication into a scientific context. The second part (publications) presents the publications as they appeared in the original sources but in a consistent format.

This chapter has presented the scientific background and motivation for this thesis. The research objective of this thesis has been determined and a set of research questions has been identified. The rest of the synopsis is organized as follows. Chapter 2 presents an executive summary of the eight publications which form the basis of the research conducted. Chapter 3 discusses the results of the publications with respect to the challenges addressed, identifies limitations, and lists the problems that remain open. Finally, Chapter 4 summarizes the contributions of this thesis and draws final conclusions.

2. Summary of publications

The main body of this thesis consists of the eight publications. This chapter summarizes the motivations, methods, results, and conclusions of each of these publications. It does not replace the publications themselves, but provides structured access to the research presented in each.

2.1. Clustering contextual neural gas: A new approach for spatial planning and analysis tasks

J. Hagenauer (2015). Clustering contextual neural gas: A new approach for spatial planning and analysis tasks. In: M. Helbich, J. J. Arsanjani, and M. Leitner (eds.). *Computational Approaches for Urban Environments.* Springer, 77–94.

2.1.1. Motivation

Many tasks, and particularly those in the area of spatial planning, require homogeneous regions that aggregate spatial observations in a meaningful way. Effectively determining such regions is a difficult task and the appropriate number of regions is typically not known a priori.

2.1.2. Methods

This publication proposed a method is proposed that combines CNG with topology learning and graph clustering to outline homogeneous regions. The method accounts for spatial heterogeneity by utilizing a CNG. Moreover, this method does not require prior knowledge about the actual number of clusters in the data, as it uses a modularity score when clustering the learned topology. Two experiments based on a synthetic and a real-world dataset were used to evaluate the method.

2.1.3. Results

The results for the synthetic dataset showed that the proposed method could correctly identify clusters in a predefined setting. The results for the real-world dataset demonstrated that the method was also able to outline meaningful and theoretically sound clusters in a practical setting.

2.1.4. Conclusion

The proposed method combines CNG, topology learning, and graph clustering. By utilizing the specific properties of the individual components, the method enables the identification of homogeneous regions without prior knowledge about their total number. This makes the method particular well-suited for spatial analysis and planning tasks.

2.2. Weighted merge context for clustering spatial data

J. Hagenauer (2016). Weighted merge context for clustering and quantizing spatial data with self-organizing neural networks. *Journal of Geographical Systems*, 18(1), 1–15.

2.2.1. Motivation

Merge context has already been shown to be useful for quantizing and clustering sequential data (Strickert and Hammer, 2005). A generalization of this approach to the spatial domain has the potential to be useful for spatial data analysis.

2.2.2. Methods

This publication introduced weighted merge context (WMC), a generalization of merge context, which recursively takes into account neighboring observations by using a weight matrix. WMC was combined with a neural gas (NG) algorithm to obtain weighted merge neural gas (WMNG). The usefulness of WMNG for quantizing and clustering spatial data was investigated using a simulated binary grid dataset and a real-world continuous dataset.

2.2.3. Results

The results for the simulated dataset showed that WMNG was able to effectively quantize the spatial context of binary grid data for a wide range of parameter settings. However, although its effectiveness was high for low distances, it decreased rapidly with distance. The results for the real-world dataset demonstrated that WMNG was able to give coherent clusters in a practical setting.

2.2.4. Conclusion

WMC's rich representation of the spatial context of the observations enabled WMNG to outline meaningful clusters and to effectively quantize spatial data. The use of a weight matrix also made WMC useful when the distance relationships between the observations were non-Euclidean, for example when observations represent complex areas. Furthermore, WMC can be combined with almost any quantization algorithm and therefore has the potential to be useful for further applications.

2.3. SPAWNN: A toolkit for spatial clustering with artificial neural networks

J. Hagenauer and M. Helbich (2016). SPAWNN: A toolkit for spatial analysis with self-organizing neural networks. *Transactions in GIS*, 20(5), 755–774.

2.3.1. Motivation

The use of self-organizing neural networks for spatial cluster analysis is a complex task that typically comprises multiple and often iteratively repeating steps. A combined toolkit that provides interactive and visual means to support the analyst in this task has the potential to provide new insights, particularly when used with complex and high-dimensional spatial datasets.

2.3.2. Methods

This publication introduced a toolkit for spatial analysis with self-organizing neural networks. The toolkit distinguishes between self-organizing neural networks and spatial context models, which can be combined with networks to account for spatial autocorrelation or spatial heterogeneity. In addition, it interactively links different self-organizing networks and data visualizations in an intuitive manner, to facilitate explorative data analysis. The computational performance of the implemented algorithms was investigated using high-dimensional synthetic datasets, while the distinctive features of the toolkit were explored using a real-world case study involving socio-economic data on Philadelphia, Pennsylvania.

2.3.3. Results

The results showed that a better understanding of the data could be obtained by utilizing self-organizing networks with different properties (i.e., CNG and GeoSOM) and linking them to geographic maps for visualization purposes. The level of computational performance of the context models was reasonable, even for high-dimensional spatial datasets.

2.3.4. Conclusion

The distinction made in the toolkit between self-organizing neural networks and spatial context models is useful, as it maintained modularity and gave rise to a multitude of useful combinations for analyzing spatial data with self-organizing neural networks. The linkage between the different networks and data visualizations allows for interaction between the analyst, the data, and the trained networks, thus enabling an improved understanding of the results.

2.4. A comparison of machine learning models for predicting travel mode choice

J. Hagenauer and M. Helbich (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, 273–282.

2.4.1. Motivation

The analysis of travel mode choice is an important task in the field of transportation planning and policy making, in order to be able to understand and predict travel demand. However, the usefulness of machine learning algorithms as well as the importance of environmental factors in modeling travel mode choice are still largely unexplored.

2.4.2. Methods

The publication compared seven machine learning classifiers for travel mode choice prediction using data from a Dutch travel diary of the years 2010 to 2012. To evaluate the performance of the classifiers, their accuracy and sensitivity was estimated using cross-validation. The class imbalance problem was accounted for by random resampling. In addition, a permutation-based approach was used to investigate the importance of different independent variables and how they relate to different modes of travel.

2.4.3. Results

The results showed that random forest performed significantly better than the other classifiers, including the commonly used multinomial logit model. Sensitivity analysis revealed that public transport and car trips could be predicted with the highest sensitivity by all classifiers, while walking and bicycle trips were predicted with the lowest sensitivity. Analysis of variable importance showed that while the trip distance was found to be the most important variable, the importance of the other variables varied with the classifier and travel mode. The importance of the meteorological variables was highest for the support vector machine, while temperature was particularly important when predicting bicycle and public transport trips.

2.4.4. Conclusion

The comparison of different classifiers and the analysis of variable importance is useful in gaining a better understanding of the relationships within the data and allows for effective modeling of travel mode choice. The superior performance of random forest and the poor performance of the multinomial logit model suggest that the relationships that determine travel mode choice are highly complex. Trip distance is a main driver for travel mode choice, whereas meteorological variables play a minor role.

2.5. Local modelling of land consumption

J. Hagenauer and M. Helbich (2018). Local modelling of land consumption in Germany with RegioClust. *International Journal of Applied Earth Observation and Geoinformation*, 65, 46–56.

2.5.1. Motivation

To prevent a further increase in land consumption in Germany, it is necessary to gain an understanding of actual and future land consumption patterns. In this context, a modeling approach that outlines clearly defined regions with similar relationships between land consumption and its drivers seems promising.

2.5.2. Methods

This publication introduced RegioClust, an algorithm that combines hierarchical clustering with regression analysis. The performance of RegioClust was compared to geographically weighted regression (GWR) using AICc scores and Moran's I.

2.5.3. Results

RegioClust provided better model fits than GWR with respect to AICc, but tended towards local overfitting when its hyperparameters were not chosen appropriately. The values of Moran's I for the residuals of RegioClust and GWR were comparable. RegioClust and GWR predicted an increase in land consumption rates in eastern Germany for 2010–20, while only GWR forecast an increase for western Germany. Of all the variables, population density had the highest importance for both models.

2.5.4. Conclusion

RegioClust and GWR provide evidence that LCR drivers vary substantially across Germany and that the most important driver of land consumption is population density. The predictions of both approaches indicated that the policy target of reducing the land consumption rate to 30 ha per day in 2020 will not be achieved.

2.6. A comparison of machine learning models for predicting land consumption

J. Hagenauer, H. Omrani, and M. Helbich (2019). Assessing the performance of 38 machine learning models: The case of land consumption rates in Bavaria, Germany. *International Journal of Geographical Information Science*, 33(7), 1399–1419.

2.6.1. Motivation

Due to irreversible affects on the environment, it is important for policymakers to use the most accurate models of land consumption available. While machine learning algorithms seem promising for this purpose, it is still unclear which algorithms perform well and how they can support an understanding of land consumption.

2.6.2. Methods

This publication compared 38 machine learning models of land consumption rates (LCRs) in Bavaria, Germany, using publicly available data. Unexplained locational effects were adjusted for by considering the longitude and latitude of each municipality. To assess the performance of each model, the mean absolute error (MAE), the root-mean-square error (RMSE), and the coefficient of determination (R^2) were estimated using cross-validation. Partial dependence plots, variable importance statistics, and residual maps were used to analyze the models and explain the learned relationships.

2.6.3. Results

All models consistently predicted that LCRs for Bavaria will increase. The best performance was obtained from eXtreme gradient boosting decision trees (xgbTree) performed best with respect to the RMSE (0.500) and R^2 (0.183), while the support vector machine with polynomial kernel gave the lowest MAE (0.288). The generalized additive model and the random forest models also performed well. The most important variables for xgbTree were the built-up area, population density, and terrain ruggedness. In addition, for xgbTree, all variables showed a nonlinear association with LCRs and municipalities in the northwest and south of Bavaria were associated with higher LCRs.

2.6.4. Conclusion

The comparison and analysis of different machine learning algorithms is useful in terms of gaining a better understanding of the relationships within the data and allow effective modeling of LCRs. The superior performance of models with high modeling capacities (e.g., xgbTree and random forest) indicates that the relationships that determine land consumption rates are highly complex. Population density is a main driver of land consumption. The predictions provide empirical evidence that the LCRs for Bavaria will increase.

2.7. Relative importance of perceived neighborhood characteristics for depression

M. Helbich, J. Hagenauer, and H. Roberts (2020). Relative importance of perceived physical and social neighborhood characteristics for depression: A machine learning approach. *Social Psychiatry and Psychiatric Epidemiology*, 55, 599–610.

2.7.1. Motivation

Recent research suggests that physical and social neighborhood environments are determinants for depression, although how and to what extent different combinations of neighborhood characteristics affect the severity of depression is currently unknown.

2.7.2. Methods

This publication compared supervised machine learning models with the aim of investigating the relationship between the perceived neighborhood environment and depression severity. For this purpose, cross-sectional data drawn from a population-representative sample of the Netherlands were used. The severity of depression was measured with the standardized Patient Health Questionnaire, while perceptions of the neighborhood were assessed with a separate questionnaire.

2.7.3. Results

The results indicated that neighborhood social cohesion, pleasantness, and safety were negatively correlated with the risk of depression, whereas perceived distance from green space and traffic were positively correlated. No correlation with depression risk was found for the perceived distance from blue space and urbanicity. A high risk of depression was found for young adults, low income earners, low-educated, unemployed, and divorced people. The risk of depression was more strongly determined by personal attributes (e.g., age, marital and employment status) than neighborhood characteristics. The results were robust across different models.

2.7.4. Conclusion

Depression severity is, independent of socio-demographic characteristics, affected by the perceived social environment. In contrast to person-level and social neighborhood characteristics, the importance of the physical neighborhood environment for depression risk is low.

The severity of depression is affected by the perceived social environment, independently of socio-demographic characteristics. Unlike person-level and social neighborhood characteristics, the importance of the physical neighborhood environment is low in terms of the risk of depression.

2.8. A geographically weighted artificial neural network (GWANN)

J. Hagenauer and M. Helbich (2022). A geographically weighted aritifical neural network. *International Journal of Geographical Information Science*, 36 (2), 215–235.

2.8.1. Motivation

Geographically weighted regression (GWR) assumes that the relationships between dependent and the independent variables are linear; in practice, however, it is often the case that these variables are nonlinearly associated.

2.8.2. Methods

The publication proposed a geographically weighted artificial neural network (GWANN). GWANN combines geographical weighting with artificial neural networks, which are able to learn complex nonlinear relationships in a data-driven manner without assumptions. GWANN was compared to GWR using synthetic data with known spatial characteristics and a real-world case study. To assess the performance of GWANN and GWR, the root-mean-square errors (RMSEs) were estimated using cross-validation.

2.8.3. Results

The results for the synthetic data showed that GWANN performed better than GWR when the relationships within the data were nonlinear and their spatial variance was high. The results for the real-world dataset demonstrated that the performance of GWANN could also be superior in a practical setting.

2.8.4. Conclusion

GWANN is able to model spatially varying nonlinear relationships without assumptions, a useful aspect for many practical applications. However, the computations performed within the network are complex, which makes an understanding of the learned relationships difficult.

3. Discussion

This thesis developed spatial machine learning algorithms for spatial prediction and clustering tasks. When solving these tasks, the algorithms take into account either spatial autocorrelation or spatial heterogeneity. Due to their inherent flexibility and adaptability, most of the algorithms developed here are based on artificial neural networks. Although the experimental results show that these algorithms can be more accurate than existing ones, certain limitations should be noted.

Spatial autocorrelation and spatial heterogeneity often occur together and may be observationally equivalent (Anselin, 2001). For instance, if residuals are spatially clustered, it is unclear whether this is due to spatial heterogeneity, spatial autocorrelation, or both. As a consequence, a spatial machine learning algorithm that takes into consideration either spatial autocorrelation or spatial heterogeneity may not be able to appropriately reflect the relationships within the spatial data.

In addition, the performance of most machine learning algorithms depends strictly on the choice of the hyperparameters. The spatial machine learning algorithms proposed here introduce one or more additional hyperparameters that control the degree to which either spatial autocorrelation or spatial autocorrelation is accounted for. These additional hyperparameters complicate the process of determining appropriate hyperparameters. This can be in particular a concern when the hyperparameters are not independent of each other, many settings of hyperparameters must be empirically evaluated, or the computational time needed to train the machine learning model is high.

In general, the usefulness of a spatial machine algorithm for a given application depends strongly on whether the task performed by the algorithm is appropriate in terms of solving the underlying problem and how the resulting solution is evaluated. This thesis comprehensively compared the usefulness of different machine learning models for three applications, namely the prediction of travel mode choice, land consumption, and depression risk. This comparison revealed that tree-based methods such as random forest and gradient boosting can be particularly useful for prediction tasks in a spatial context. Nevertheless, there are some limitations that need to be taken into account.

For instance, the machine learning models were primarily evaluated based on their prediction performance, computational performance, or the spatial distribution of the residuals. However, the suitability of an algorithm for a particular task can only be evaluated to a limited extent if a small number of performance measures is considered. Depending on the task the algorithm is intended to solve, other measures such as understandability, number of parameters, or robustness to data errors and outliers may also be useful.

Another issue that is closely related to model selection is how to determine the appropriate hyperparameters when comparing machine learning models. Although the choice of hyperparameters critically affects the results of a model, there is no definitive method for determining the most appropriate parameters. In this thesis, a grid search within a manually specified subspace was performed in order to search for appropriate hyperparameters and the performance for each setting of hyperparameters was evaluated using cross-validation. While this approach produces good results in practice, it cannot be guaranteed to find the most appropriate hyperparameters.

In addition, the number of available machine learning models is vast. For practical reasons, it is therefore necessary to limit a comparative study to a subset of all the available models that seem to be promising from an a priori point of view. The decision on which machine learning models seem promising should be based on the domain knowledge of the analyst; however, this decision is also necessarily subjective to some extent and therefore bears the risk of introducing selection bias into the comparison.

The number of available machine learning models is constantly increasing due to the scientific progress that is being made in this domain. For instance, progress in representational learning has led to the development of numerous variants of artificial neural networks (e.g., Schmidhuber, 2015) which could potentially also play an important role in the spatial sciences. Hence, any comparative study can at most provide a snapshot of the current state of research and should be frequently updated and repeated in order to ensure significance in its field. To make this feasible in practice, it is necessary that the reproducibility of comparative studies is maintained (e.g., Nüst and Pebesma, 2020).

A great deal of the research on spatial data machine learning has focused on improving predictive performance, but in order to further exploit the large amounts of spatial data that are now available, it is necessary to move beyond prediction towards explanation (Gahegan, 2017). This thesis has investigated several different approaches to support the understanding of the models and their learned relationships. These approaches included the visualization of the results and model parameters using maps, the investigation of variable importance, and (spatial) partial dependence plots. In particular, the ability to visualize the model parameters with maps has proven useful, as maps are able to effectively communicate spatial distance relationships (MacEachren, 2004). One limitation of these approaches, however, is that they require that the model parameters can be associated with some location in space, which is not always feasible.

It should also be noted that these approaches provide explanations which are, for the sake of comprehensibility, not completely faithful to the original model (Rudin, 2019). For instance, partial dependence plots only provide insights into the joint relationships of a subset of independent variables (for practical reasons, usually one or two independent variables) and parameter maps do not show the interactions between parameters at all. In order to avoid the use of these unfaithful explanatory approaches, machine learning models are needed that are both true (i.e., highly accurate) and understandable at the same time.

Moreover, the approaches investigated here for explaining machine learning models are only useful in so far as they support an understanding of how the model works, or in other words why the model produces an output for a certain input; they do not explain how the real world works. It is still the task of the analyst to draw conclusions from the provided explanations, to build new hypotheses using domain knowledge, and to test whether these hypotheses hold true in the real world. To facilitate this process, a single integrated system must be developed to replace the many fragmented systems for representation and analysis that are currently available (Gahegan, 2020).

4. Conclusion

This thesis contributed to increasing the usefulness of spatial machine learning for the spatial sciences by developing applications and methods that facilitate the solution of complex spatial problems. To achieve this, it focused on addressing particularly important challenges. These challenges were the modeling of spatial autocorrelation and spatial heterogeneity, the selection of an appropriate model for a given spatial problem, and the understanding of complex spatial machine learning models.

Despite the progress that have been achieved, spatial machine learning is not and probably never will be a panacea that can solve all the problems of spatial science. In particular, spatial machine learning models are only as good as the data used to train the model are representative, the models are capable of learning important relationships within the data, and the analyst is able to make the appropriate decisions when building the model. The same limitations also apply to traditional statistical models. The fundamental difference, though, is the different degree of reliance on assumptions. In contrast to traditional statistical models, spatial machine learning models do rely on no or only few assumptions. On the one hand, this is a strength, because this allows the model to learn new and unexpected relationships within the data, and hence can support the generation of new hypotheses and the discovery of knowledge. On the other hand, this is a weakness, because the resulting models are often overly complex and the process of building a truthful model places additional demands on the spatial scientist. For this reason, spatial machine learning does not replace more traditional statistical methods but instead enhances the toolbox of spatial scientists through methods which are particularly useful in our current data-rich environment.

As discussed in the previous section, the proposed approaches for addressing the challenges that are the focus of this thesis are not without limitations. The challenges addressed here also represent only a small portion of the challenges which spatial machine learning is facing. This means that the potential of machine learning in the context of spatial sciences is still far from being realized. Further research in the area of spatial machine learning can be expected to further increase its usefulness for spatial sciences, which in turn will lead to new insights into spatial phenomena and ultimately to a better understanding of our world.

Bibliography

Anselin, L. (1989). *What is special about spatial data? Alternative perspectives on spatial data analysis.* Tech. rep. 89-4. UC Santa Barbara: National Center for Geographic Information and Analysis.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, 27(2), 93–115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: M. Fischer, H. J. Scholten, and D. Unwin (eds.). *Spatial Analytical Perspectives on GIS.* Taylor & Francis, London, United Kingdom, 111–125.

Anselin, L. (2001). Spatial econometrics. *A Companion to Theoretical Econometrics*, 310330.

Anselin, L. and Griffith, D. A. (1988). Do spatial effects really matter in regression analysis? *Papers in Regional Science*, 65(1), 11–34.

Asner, G. P., Knapp, D. E., Boardman, J., Green, R. O., Kennedy-Bowdoin, T., Eastwood, M., Martin, R. E., Anderson, C., and Field, C. B. (2012). Carnegie Airborne Observatory-2: Increasing science data dimensionality via high-fidelity multi-sensor fusion. *Remote Sensing of Environment*, 124, 454–465.

Bahn, V. and McGill, B. J. (2013). Testing the predictive performance of distribution models. *Oikos*, 122(3), 321–331.

Bennett, K. P. and Parrado-Hernández, E. (2006). The interplay of optimization and machine learning research. *The Journal of Machine Learning Research*, 7, 1265–1281.

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, 5(6), 853–862.

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: *2012 IEEE international geoscience and remote sensing symposium.* IEEE, 5372–5375.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4), 281–298.

Cai, Q., Gong, M., Ma, L., Ruan, S., Yuan, F., and Jiao, L. (2015). Greedy discrete particle swarm optimization for large-scale social network clustering. *Information Sciences*, 316, 503–516.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832.

Casetti, E. (1972). Generating models by the expansion method: Applications to geographical research. *Geographical Analysis*, 4(1), 81–91.

Credit, K. (2021). Spatial models or random forest? Evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in Los Angeles. *Geographical Analysis*.

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895–1923.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1), 3133–3181.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.

Fotheringham, A. S. (1998). Trends in quantitative methods II: Stressing the computational. *Progress in Human Geography*, 22(2), 283–292.

Fotheringham, A. S. (2009). "The problem of spatial autocorrelation" and local spatial statistics. *Geographical Analysis*, 41(4), 398–403.

Freitas, A. A. (2014). Comprehensible classification models: A position paper. *ACM SIGKDD Explorations Newsletter*, 15(1), 1–10.

Friedman, J. H. and Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22(9), 1365–1381.

Gahegan, M. (2000). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, 32(2), 113–139.

Gahegan, M. (2003). Is inductive machine learning just another wild goose (or might it lay the golden egg)? *International Journal of Geographical Information Science*, 17(1), 69–92.

Gahegan, M. (2017). International encyclopedia of geography: People, the earth, environment and technology. In: *International Encyclopedia of Geography*. Ed. by D. Richardson, N. Castree, M. F. Goodchild, A. Kobayashi, W. Liu, and R. A. Marston. American Cancer Society. Chap. Geocomputation, 1–5.

Gahegan, M. (2020). Fourth paradigm GIScience? Prospects for automated discovery and explanation from data. *International Journal of Geographical Information Science*, 34(1), 1–21.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated Statistician*, 5(3), 115–146.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S. (2019). Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 1–16.

Getis, A. (2009). Spatial weights matrices. *Geographical Analysis*, 41(4), 404–410.

Getis, A. (2010). Spatial autocorrelation. In: *Handbook of Applied Spatial Analysis*. Springer, 255–278.

Getis, A. and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3), 189–206.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D. (2017). Google vizier: A service for black-box optimization. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1487–1495.

Goodchild, M. F. (2017). Data analysis, spatial. In: X. Z. Shashi Shekhar Hui Xiong and (ed.). *Encyclopedia of GIS*. Springer US, 405–409.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep learning*. `http://www.deeplearningbook.org`. Cambridge, MA, USA: MIT Press.

Gorr, W. L. and Olligschlaeger, A. M. (1994). Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market modeling. *Geographical Analysis*, 26(1), 67–87.

Griffith, D. A. (2003). *Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization*. Springer Science & Business Media.

Griffith, D. A. (1993). Which spatial statistics techniques should be converted to GIS functions? In: *Geographic Information Systems, Spatial Modelling and Policy Evaluation*. Springer, 103–114.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22(7), 801–823.

Hagenauer, J. (2015). Clustering contextual neural gas: A new approach for spatial planning and analysis tasks. In: M. Helbich, J. J. Arsanjani, and M. Leitner (eds.). *Computational Approaches for Urban Environments*. Springer, 77–94.

Hagenauer, J. (2016). Weighted merge context for clustering and quantizing spatial data with self-organizing neural networks. *Journal of Geographical Systems*, 18(1), 1–15.

Hagenauer, J. and Helbich, M. (2016). SPAWNN: A toolkit for spatial analysis with self-organizing neural networks. *Transactions in GIS*, 20(5), 755–774.

Hagenauer, J. and Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, 273–282.

Hagenauer, J. and Helbich, M. (2018). Local modelling of land consumption in Germany with RegioClust. *International Journal of Applied Earth Observation and Geoinformation*, 65, 46–56.

Hagenauer, J. and Helbich, M. (2022). A geographically weighted aritifical neural network. *International Journal of Geographical Information Science*, 36 (2), 215–235.

Hagenauer, J., Omrani, H., and Helbich, M. (2019). Assessing the performance of 38 machine learning models: The case of land consumption rates in Bavaria, Germany. *International Journal of Geographical Information Science*, 33(7), 1399–1419.

Haining, R. (2003). *Spatial data analysis: theory and practice*. Cambridge university press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction*. 2nd ed. Springer.

Helbich, M., Brunauer, W., Vaz, E., and Nijkamp, P. (2014). Spatial heterogeneity in hedonic house price models: The case of Austria. *Urban Studies*, 51(2), 390–411.

Helbich, M., Hagenauer, J., and Roberts, H. (2020). Relative importance of perceived physical and social neighborhood characteristics for depression: A machine learning approach. *Social Psychiatry and Psychiatric Epidemiology*, 55, 599–610.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, 31(3), 264–323.

Johnson, D. A. and Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. In: *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1609–1615.

Kavzoglu, T., Colkesen, I., and Sahin, E. K. (2019). Machine learning techniques in landslide susceptibility mapping: A survey and a case study. *Landslides: Theory, Practice and Modelling*, 283–301.

Kerry, R., Goovaerts, P., Rawlins, B. G., and Marchant, B. P. (2012). Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. *Geoderma*, 170, 347–358.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. In: *NIPS*, 2280–2288.

Kruppa, J., Schwarz, A., Arminger, G., and Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13), 5125–5131.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., and Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, 23(7), 811–820.

LeSage, J. P. (2008). An introduction to spatial econometrics. *Revue d'économie industrielle*, (123), 19–44.

MacEachren, A. M. (2004). *How maps work: Representation, visualization, and design.* Guilford Press.

Miller, H. J. and Wentz, E. A. (2003). Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers*, 93(3), 574–594.

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2), 284–289.

Miller, H. J. and Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4), 449–461.

Miller, H. J. and Han, J. (2009). *Geographic data mining and knowledge discovery.* CRC press.

Mitchell, T. M. (1997). *Machine learning.* New York: McGraw-Hill.

Molnar, C. (2019). *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* `https://christophm.github.io/interpretable-ml-book/`.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17–23.

Nakaya, T. and Yano, K. (2010). Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3), 223–239.

Neto, A. H. and Fiorelli, F. A. S. (2008). Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and Buildings*, 40(12), 2169–2176.

Nüst, D. and Pebesma, E. (2020). Practical reproducibility in geography and geosciences. *Annals of the American Association of Geographers*, 1–11.

Openshaw, S. (1999). Geographical data mining: Key design issues. In: *Proceedings of the 4th International Conference on GeoComputation.*

Pijanowski, B. C., Brown, D. G., Shellito, B. A., and Manik, G. A. (2002). Using neural networks and GIS to forecast land use changes: A land transformation model. *Computers, Environment and Urban Systems*, 26(6), 553–575.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., Sheppard, L., Larson, T. V., and Kaufman, J. D. (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2. 5 concentrations in epidemiology. *Atmospheric Environment*, 75, 383–392.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210–229.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117.

Steinegger, M. and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1), 1–8.

Strickert, M. and Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing*, 64, 39–71.

Sui, D. and Goodchild, M. F. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographical Information Science*, 25(11), 1737–1748.

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

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2), 111–126.

Wang, F., Guo, D., and McLafferty, S. (2012). Constructing geographic areas for cancer data analysis: A case study on late-stage breast cancer risk in Illinois. *Applied Geography*, 35(1-2), 1–11.

Wei, P., Lu, Z., and Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, 142, 399–432.

Wenger, S. J. and Olden, J. D. (2012). Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3(2), 260–267.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82.

Zhao, Z.-Q., Zheng, P., Xu, S.-t., and Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), 3212–3232.

Zumberge, J. F., Heflin, M. B., Jefferson, D. C., Watkins, M. M., and Webb, F. H. (1997). Precise point positioning for the efficient and robust analysis of GPS data from large networks. *Journal of Geophysical Research: Solid Earth*, 102(B3), 5005–5017.

Part II.

Publications

5. Clustering contextual neural gas: A new approach for spatial planning and analysis tasks

Authors

Julian Hagenauer

Book

Computational approaches for urban environments

Status

Published online 11 November 2014

Contribution statement

Julian Hagenauer has developed the methods, designed the experiments, and has written the manuscript for the study.

Abstract

Spatial clustering is a method that can reveal structures and identify groupings in large spatial data sets, which is in particular useful for spatial planning and analysis tasks. A recent and powerful clustering algorithm for spatial data is contextual neural gas (CNG). The CNG algorithm is closely related to the basic self-organizing map algorithm, but additionally takes spatial dependence into account. However, like most clustering algorithms, the CNG requires the analyst to specify the number of clusters beforehand. Even though the chosen number of cluster critically affects the results of the clustering, it is unclear how to determine it. This study introduces a new method which combines CNG, the learning of the CNG's topology, and graph clustering. It can be used to cluster spatial data without any prior knowledge of present clusters in the data. The proposed method is in particular useful for spatial planning and analysis tasks, because it provides means to find groupings in the data and identify homogeneous regions. To evaluate the method this study draws from two experiments which are based on an synthetic and a real-world data set. The results of the synthetic data set show that it can correctly identify clusters in a predefined setting. The results of the real-world data set demonstrate that the proposed method outlines meaningful and theoretically sound regions.

Keywords: Artificial neural networks; cluster analysis; spatial planning

5.1. Introduction

Clustering is the task of organizing observations into clusters such that the similarity of observations within a cluster is minimized and the dissimilarity between the clusters is maximized. It is particularly useful if no categorization or labeling of the observations is available, but some structural organization is needed. Many different clustering algorithms have been developed, mainly in the fields of statistics and machine learning. These clustering algorithms can be broadly classified by their used paradigms. One of the most prominent and widely used clustering paradigms is partitioning clustering. Partitioning clustering algorithms, such as k-Means or neural gas (NG; T. Martinetz and Schulten, 1991), divide a set of observations into a non-overlapping set of clusters. Each observation is assigned to the cluster which it is closest to. For large data sets, partitioning clustering algorithms are typically more computationally effective than, e.g., hierarchical clustering algorithms (Jain et al., 1999). However, a severe disadvantage of them is that they require the analyst to choose the number of desired clusters beforehand.

There are several important special cases of clustering. One such case is spatial clustering, which deals with the clustering of spatially located observations. A basic property of such observations is that they are likely to be spatially dependent. Spatial dependence states that observations that are spatially located close to each other tend to have similar characteristics. This property is essential to spatial sciences because without it variation of phenomena would be independent of location, and thus the notion of region would be totally meaningless (Goodchild, 1986). The presence of spatial dependence has been traditionally regarded as problematic for statistical analysis, which typically requires sample independence (Bailey and Gatrell, 1995). However, it can also serve as a valuable source of information about spatial processes, because it provides evidence of causality (Miller, 2004). Therefore, it is important for spatial clustering algorithms to take spatial dependence into account in order to utilize the full range of available information for discovering spatial patterns.

Spatial clustering is of special importance for spatial planning tasks: Administrative areas typically have their roots in historic administrative divisions of space, which disregard the non-spatial characteristics of place. As a consequence, they often intersect contiguous regions and are often inhomogeneous. Decisions made concerning the planning, distribution, and allocation of resources among such administrative areas are likely to be ineffective and meaningless (Amedeo, 1969). In fact, it has been shown by Van Der Laan and Schalke (2001) that local policies are more effective for homogeneous regions. These concerns are very closely related to the modifiable areal unit problem (MAUP; Openshaw, 1984). Spatial analysis typically requires manageable discrete descriptions of spatial processes, which are continuous. For this purpose, it is necessary to aggregate observations over areal units. The MAUP states that the outline of these units and the scale of aggregation critically affect the results of any spatial analysis. In general, it is useful if the observations that are aggregated over the same areal unit are similar to each other. Consequently, since spatial clustering outlines mostly coherent and homogeneous areas, it has potential to serve as a valuable tool for spatial planning and analysis tasks (e.g., Helbich et al., 2013).

Various spatial clustering algorithms have been developed in the past (see e.g., Han et al., 2001). Most of these methods are based on general-purpose clustering algorithms that have limited capabilities in recognizing spatial patterns that involve neighbors or cannot deal with high-dimensional data sets (Guo et al., 2003). Contextual neural gas (CNG; Hagenauer and Helbich, 2013) is a recently developed algorithm for clustering spatial data that is specially designed for spatial data mining. The CNG algorithm combines the concepts of the NG algorithm with the GeoSOM (Bação et al., 2005), a variant of the famous self-organizing map algorithm (SOM; Kohonen, 1982, 2001) that takes spatial dependence into account. A particular advantage of the CNG is that it quantizes the data space better than the GeoSOM, because the adaptation of the CNG's neurons, in contrast to the SOM, does not depend on some predefined and fixed topology (Hagenauer and Helbich, 2013). However, the topology of the SOM facilitates the analysis of the SOM and hence supports understanding of the properties of the data (e.g., Arribas-Bel and Schmidt, 2013; Hagenauer et al., 2011; Skupin and Esperbé, 2011), In particular, it is useful for determining the actual number of clusters in the data, either computationally (e.g., J. Costa and De Andrade Netto, 1999; Murtagh, 1995) or by visualizing it (see Flexer, 2001).

Another important special case of clustering is graph clustering. A graph is a set of vertices and a set of edges that are connections between pairs of vertices. The edges can have a weight assigned which indicates the strength of the connection and can be directed or undirected. The task of spatial clustering is organizing vertices of a graph into clusters such that vertices within a cluster are better connected than vertices within different clusters. The ability to find and analyze clusters is useful for understanding and visualizing the structure of networks, which is of great importance in many research areas that deal with social, technological, or information systems. A lot of different algorithms have been developed in the past for this purpose (see Schaeffer, 2007). From the large set of available algorithms, the heuristic multi-level modularity optimization algorithm (MLMO; Blondel et al., 2008) is particularly promising, because it is exceptionally fast even for very large graphs and automatically determines the number of clusters in the graph by optimizing its quality score.

This study introduces a new method that combines CNG, topology learning, and graph clustering algorithms to outline clusters in CNG: The method consists of the following steps: First, a CNG consisting of a sufficient number of neurons is trained. Second, a topology of the neurons is learned and the resulting topology is considered a weighted graph. Finally, this graph is clustered using advanced graph clustering algorithms, which does not require to specify the desired number of clusters. The resulting clusters represent homogeneous regions in the input data. Since the number of clusters is automatically determined depending on the topological patterns of the graph, the method is especially useful for outlining spatial clusters when no prior knowledge about the actual number of clusters available.

This workflow is closely related to the clustering approach using the GeoSOM algorithm. In this approach, a GeoSOM consisting of a sufficiently large number is trained to project the input data onto a two-dimensional map. Subsequently, the map is visualized, usually by means of a U-matrix (Ultsch and Siemon, 1990). Clusters appear on the U-matrix as valleys, cluster boundaries as ridges. However, for complex and high-dimensional data sets U-matrices often shown no clear patterns so that it is difficult or even impossible to determine clusters, in particular for computational methods.

The proposed method is also closely related to the approach by J. A. F. Costa and Oliveira (2007). In their approach, they train a growing neural gas (GNG; Fritzke, 1995) to obtain a topology. The main differences between the basic NG and the GNG algorithm are, that the GNG does not require to specify the number of neurons beforehand and that it forms a topology in the process of training the network. However, the GNG also introduces numerous additional parameters, which must be set appropriately to obtain reasonable results. Then, in a post-processing step, they modify the topology of the GNG by heuristically removing connections between neurons; disjunctive sections of the topology are considered clusters. However, which connections are removed depends on arbitrary chosen threshold levels and critically affects the results. Additionally, complex structural properties of the topology are totally disregarded. Moreover, their approach is not appropriate for clustering spatial data, because it merely uses a basic NG algorithm, which does not take into account spatial dependency.

This study is structured as follows: Section 5.2 introduces the algorithm which this study utilizes, while Section 5.3 briefly explains the consecutive steps of which the proposed method consists of. The usefulness of the method is demonstrated in two different experiments in Section 5.4. Finally, Section 5.5 concludes with some remarks and identifies future work.

5.2. Methodical background

5.2.1. Contextual neural gas

Contextual Neural Gas (CNG; Hagenauer and Helbich, 2013) is a spatial clustering algorithm that combines the concepts of the GeoSOM with the NG algorithm. Like basic NG, it consists of an arbitrary number of neurons, which are not subject to any topological restrictions, and provides a non-linear mapping in high-dimensional data space. In each step of the training process, an input vector is selected from the input data and each neuron is moved into its direction. Thereby, the strength of the movement depends on the neurons' ranking order with respect to the distance to the input vector, the adaptation rate, and the neighborhood range. Both, the neighborhood range and adaptation rate are typically chosen to decrease with time.

CNG differs from basic NG in the determination of the neurons' ranking order, which

CNG does in a two-phase procedure to incorporate spatial dependence. In the first step, the neurons are ordered by spatial closeness. In the second step, the first l neurons of the resulting spatial ordering are reordered within their ranks with respect to the similarity of attributes.

The parameter l determines the strength of spatial dependence which is incorporated into the mapping. If $l = 1$, the ordering in the second step has no effect on the final ordering at all. As a consequence, the adaption of the neurons depends solely on spatial closeness. The attributes of the input data are ignored. If l is increased, the ordering of the l spatially closest neurons depends on attribute similarity. Hence, spatial closeness is less important for the final ordering and less spatial dependence is being incorporated. If l equals the total number of neurons, the spatial ordering does not matter for the final ordering, because all neurons are totally reordered in the second step by similarity of attributes. Consequently, no spatial dependence is incorporated at all.

CNG has several advantages over other spatial clustering algorithms: Like the GeoSOM, CNG enforces spatial proximity between observations and neurons by means of neural distance, defined by either the maps topology or the rank ordering or neurons. Consequently, it is not necessary to weight or scale spatial proximity and attribute similarity in the data space. Furthermore, the neurons are basically local averages. Thus, the process of incorporating spatial dependence is less sensitive to random variations in the input data. Finally, the parameter l restricts the mapping of observations; all observations are always mapped to one of its l spatially closest neurons. Hence, the mapping maintains a certain degree of spatial closeness, even for observations whose attributes are very different from those of their spatial neighbors (spatial outliers).

5.2.2. Competitive Hebbian learning

Competitive Hebbian learning (CHL; T. Martinetz and Schulten, 1991; T. Martinetz, 1993) forms a topology on a set of neurons by creating a number of connections between neighboring neurons. In more detail, the learning algorithm can be described as follows: For each input vector, the two closest neurons are determined and a connection between these is added to the total set of connections. Thereby, closeness is typically measured by Euclidean distance. After all input vectors have been presented, the set of connections represents the topology of the underlying data.

The resulting graph optimally preserves the topology in a very general sense (T. Martinetz, 1993). In particular, each connection between two neurons belongs to the Delaunay triangulation corresponding to the neurons in data space. The theoretical foundations of CHL in terms topology preservation have been provided by (Edelsbrunner and Shah, 1997).

CHL is especially useful for NG and other vector quantization algorithms, which do not define a topological structure. It can be applied concurrently to the training of NG or as a post-processing step. However, in the first case the movement of neurons during the training may make previously learned connections invalid. Therefore, it is necessary to constantly adapt the topology to these movements, e.g., by removing outdated connections (T. Martinetz and Schulten, 1991). In the latter case, NG is trained before CHL is applied, and hence the topology is not affected by the movement of the neurons. For simplicity, this study applies CHL as a post-processing step.

Since its introduction, numerous extensions and variants of the CHL algorithm has been proposed. e.g., De Silva and Carlsson (2004) presented a generalization of CHL which produces a simplicial complex instead of a graph. An alternative to CHL was presented by Aupetit (2005). In this approach, each edge and vertex of the Delaunay triangulation is the basis of a generative model, so that the triangulation generates a mixture of Gaussian density functions; the likelihood of the set of model parameters is maximized using the Expectation-Maximization algorithm.

5.2.3. Multi-level modularity optimization

The multi-level modularity optimization algorithm (MLMO; Blondel et al., 2008) is a heuristic method which seeks to find a clustering of a graph with maximum modularity. Modularity is a quality measure that evaluates the density of connections inside a cluster as compared with the connection between different clusters (Newman and Girvan, 2004). Because optimizing modularity is a problem that is computationally hard (Brandes et al., 2008), heuristic algorithms are inevitable for practical applications.

The MLMO algorithm consists of two phases: Initially, each vertex of a graph is assigned to a single cluster. Then, in the first phase, each vertex is assigned to the cluster of the neighboring vertex which yields the largest increase of modularity, as long as it is positive. In the second phase, the original graph is replaced by a newly built graph whose vertices are the clusters found during the first phase. Connections between the new graphs' vertices exist if there is at least one connection between vertices of the corresponding clusters in the original graph. The two phases are iteratively repeated until there are no more changes to the graph and a maximum of modularity is reached.

The MLMO algorithm is computationally efficient and scales very well, because the number of clusters dramatically reduces with each pass. In particular, computer simulations on large graphs indicated that its complexity is linear on typical and sparse data (Blondel et al., 2008). A limitation of most modularity optimizing clustering algorithms is that they fail to detect small clusters in very large graphs (Fortunato and Barthelemy, 2006). However, the MLMO algorithm seems to be unaffected by this limitation because of its multi-level nature (Blondel et al., 2008). In fact, it has been shown by Fortunato (2010) and Lancichinetti and Fortunato (2009) that the quality of the MLMO's results is superior to that of many other graph clustering algorithms

5.3. Workflow

The proposed method consists of three major steps that are typically executed in sequential order:

1. *Contextual Neural Gas:* The CNG algorithm clusters the data set into n spatial clusters, where n is the number of neurons. The actual number of clusters in the data is typically unknown; n must be chosen large enough so that a reasonable cluster structure can be detected in the following steps. However, if n is too large, some of the CNG's neurons may not map any data at all. These neurons must not be removed, because the rank ordering of CNG depends on number of neurons (Hagenauer and Helbich, 2013).

2. *Topology learning:* A topology of the CNG's neurons is learned with a modification of the CHL algorithm. The algorithm can be described as follows: For each input vector, the ranking order of the neurons is determined according to the two-phase procedure of the CNG and a connection between the two highest ranked neurons is added to the connection set. Additionally, the number of times a connection has been added to the set is stored for each connection. This number finally indicates the strength of a connection, and is of use in the clustering step.

3. *Graph clustering:* Before clustering the resulting graph, single vertices that are not connected to any other vertex are removed because the neurons that these vertices represent do not map any data and bear no valuable topological information. Then the graph is clustered based on its structural properties using the MLMO algorithm.

5.4. Experiments

To evaluate the proposed method, two experiments on different data sets are conducted. In both experiments a CNG with 25 neurons is applied. The neurons are randomly initialized, and the training time is set to $100,000$ iterations. The neighborhood range and adaptation rate are chosen as proposed by T. M. Martinetz et al. (1993).

5.4.1. Synthetic data

In this experiment an synthetic data set is constructed whose properties are clearly determined. Consequently, the results of the proposed method can be easily evaluated. The data set consists of five clusters: One large cluster in the middle with low point density and four smaller clusters in the corners with higher point density (see Figure 5.1). Each cluster contains 200 random data points and each point has three attributes: Their x and y coordinates and an synthetic attribute, whose value is one for the middle cluster and otherwise zero.

The main challenge when clustering this data set is to differentiate between the spatial clusters in the corners of the data set, because their borders are defined by spatial point density. Spatial clustering algorithms which solely consider the spatial distances between points and/or the similarity of the points' attribute value are likely to fail to correctly identify the spatial clusters. Additionally, the clustering of the data set becomes much more difficult, if the actual number of clusters is not known beforehand.

The results for the synthetic data set depend on the parameter l. If l is set too low, differences in the observations' attribute values are neglected and the resulting graph therefore exhibits no distinct clusters. Otherwise, if l is set too high, the clustering does not consider the spatial configuration of the data set, resulting in a graph with only two clusters, representing the data points with the values 0 and 1. However, because of the clearly defined cluster structure of the data set, it can be assumed that l is chosen correctly, if the modularity of the resulting graph is maximal.

Figure 5.2 plots the mean modularity scores of 100 runs for different settings of l. For $l > 16$ the mean modularity score is basically constant and at its minimum. Hence, for large l-values the spatial configuration of the data set has no significant effect on the clustering of the resulting graph. Furthermore, the plot shows multiple local maxima; the highest mean modularity score (0.724) is achieved with $l = 7$.

Figure 5.3 exemplarily shows the graph resulting from a CNG that has been trained with $l = 7$ and its clustering, indicated by the colored vertices. At first, it is notable that the depicted graph consists of only 17 vertices, although CNG consists of 25 neurons. Generally, the number of neurons which map no data at all increases rapidly if l is increased, because of the simple clustering structure of the data set. The large number of vertices present indicates that the incorporated degree of spatial dependence

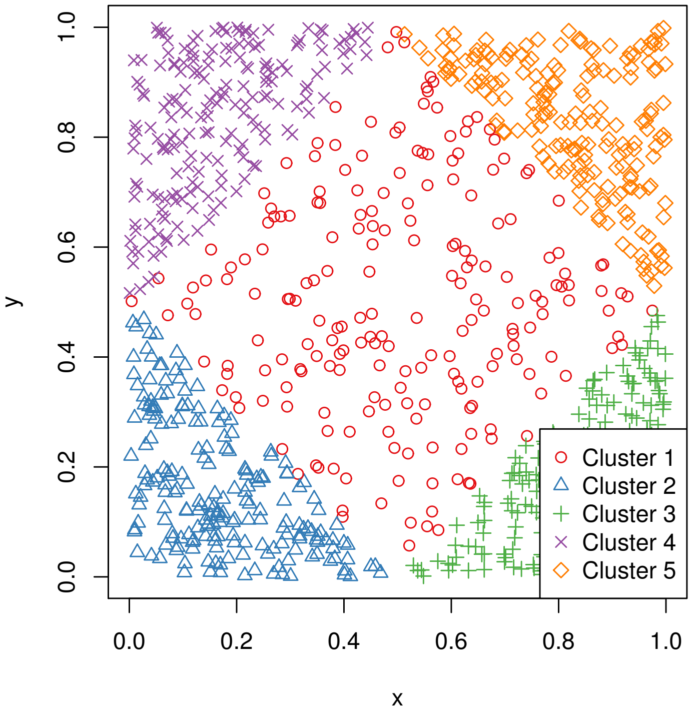


Figure 5.1.: Synthetic data set. The attribute values of the data points of cluster 1 are 0, the attribute values of the other clusters 1.

is significant. Moreover, the figure reveals that the MLMO algorithm identified the five clusters of the data set. It is notable that the middle cluster consists of more vertices than the other clusters, even though the middle cluster consists of the same number of data points as the clusters in the corner. The reason for this is that, because of the low value of the l-parameter, the distribution of the CNG's neurons depends heavily on the spatial distribution of observations. Hence, since the spatial extent of the middle cluster is four times that of the clusters in the corners, it is mapped by more neurons.

Finally, it can be seen that the three of the graph's clusters in the corners are connected, which is likely due to the small distance between them. Because the MLMO algorithm has taken the weighting of the connections into account, the corner clusters are correctly distinguished.

5.4.2. Practical application

To evaluate the practical applicability of the proposed method, it is used for delineating homogeneous regions in the city of Philadelphia, Pennsylvania. The city is situated in the Northeastern United States along the Delaware and Schuylkill rivers and consists

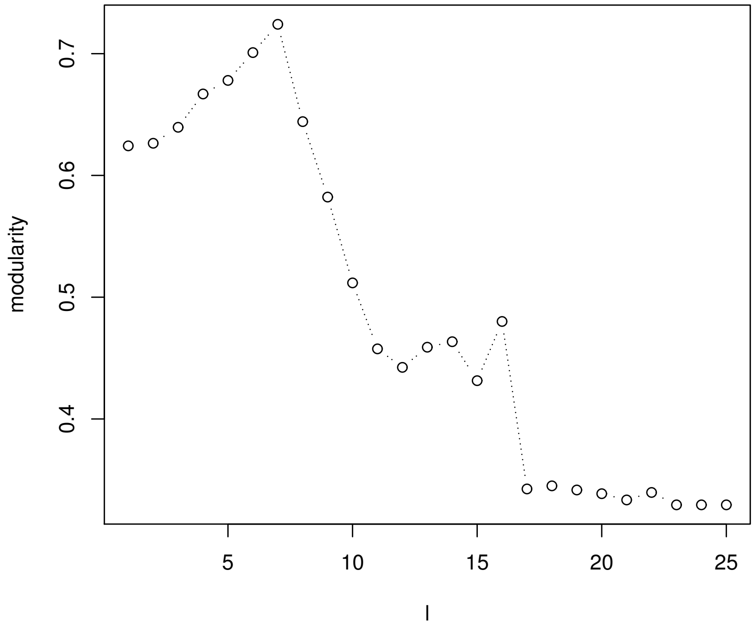


Figure 5.2.: Mean modularity score of 100 runs for different settings of l for the synthetic data set.

of an area of approximately 369 square kilometers. Philadelphia is currently the fifth largest city in the United States with an estimated population in 2012 of 1.5 million people. Philadelphia is the economic and cultural center of the Delaware Valley, the sixth-largest metropolitan area of the United States. The city is of particular interest because it has experienced dramatic changes in its ethnic and racial makeup in the last two decades (The Philadelphia Research Initiative, 2011). Hence, dynamic approaches to outline homogeneous regions are essential in this context. However, the validation of the results is difficult because there is no correct solution to the problem in a formal sense. The results of the proposed method are evaluated in this experiment by comparing them with the planning analysis sections (PAS) of the Philadelphia City Planning Commission (PCPC; Philadelphia City Planning Commission, 2004) and linking them to existing demographic knowledge. Each section of the PAS contains a number of census tracts that roughly correspond to general socio-economic divisions existing within the city (Wolfgang et al., 1987). Even though the PAS where designed for administrative purposes decades ago, they are still currently used for planning and analysis tasks (e.g., Pearsall and Christman, 2012). Figure 5.4 shows the 12 regions of the PAS.

The experiment uses tract-level data about ethnicity, race, age, housing, and households in Philadelphia from the 2010 US Census (see Figure 5.5). Tracts without

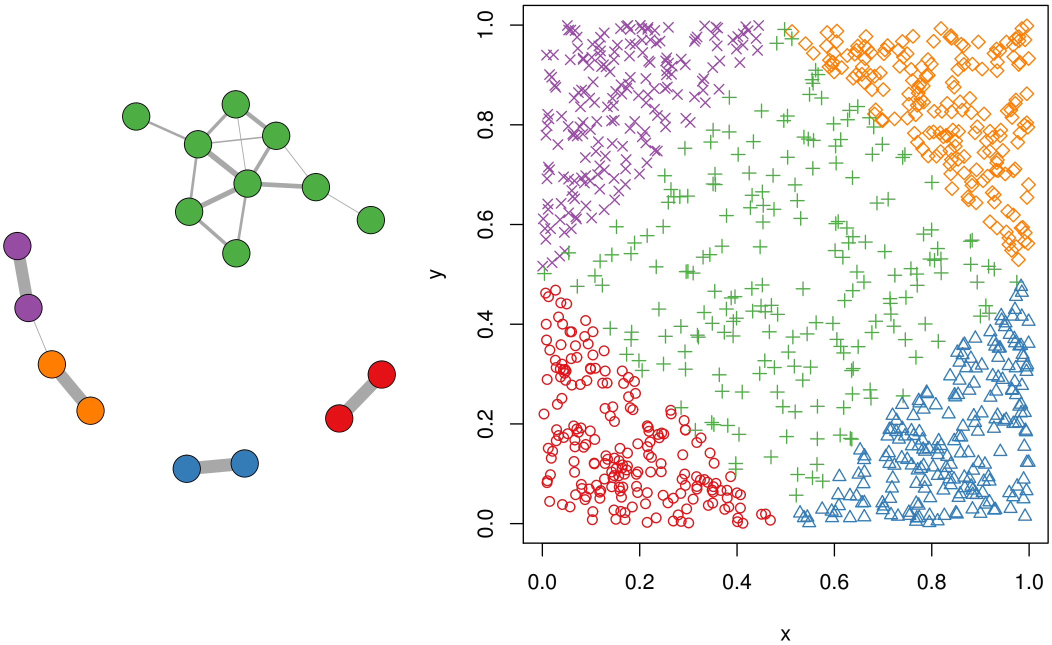


Figure 5.3.: Clustering results for $l = 7$. The data points (right) are colored according to the colors of the detected clusters of the graph (left). The thickness of the graph's edges corresponds to the weights of the connections.

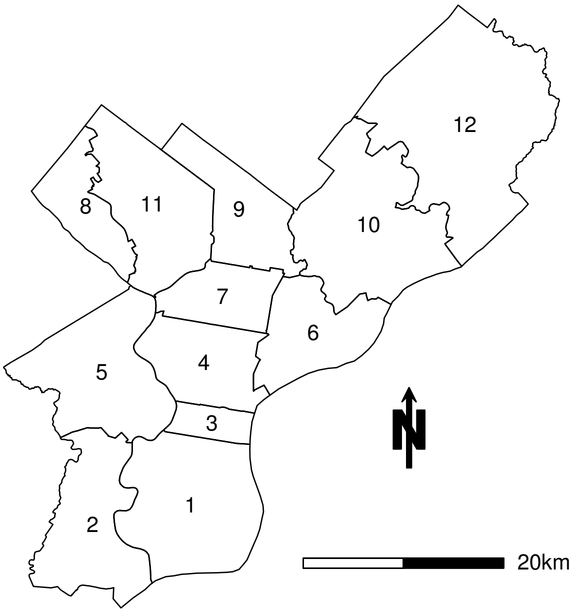


Figure 5.4.: Philadelphia analysis sections (PAS).

population are removed from the data set, and all attributes are standardized to zero mean and unit variance to make them comparable. Overall, the study site consists of 380 census tracts.

Similar to the previous experiment, it is unclear how much spatial dependence should be incorporated into the CNG's learning process to obtain reasonable results. Figure 5.6 shows the mean modularity scores of 100 runs for different settings of l.

The highest mean modularity score (0.646) is achieved with $l = 2$ and $l = 3$. Notable local maxima can be observed for $l = 10$ (0.598) and $l = 13$ (0.600). For $l > 19$, the modularity score is basically at its minimum value.

In contrast to the previous experiment, no prior knowledge about clusters in the data set is available; any parameter l might be as reasonable as any other one. However, based on the objective of this experiment three demands on l should be met: First, parameter l should be chosen so that the modularity score is high, because a high modularity score is a strong indicator of a clear-cut clustering structure. Second, the parameter l should be high enough so that a fair portion of the tracts' social and demographic characteristics is taken into account in the process of clustering. Third, parameter l should be low enough so that the resulting clusters tend to be spatially contiguous. Spatial contiguity is in particular a useful property for spatial planning and policy making, because spatially contiguous clusters can typically be described by a single spatial outline, which eases perception and understanding of the clusters.

Figure 5.3 exemplarily shows the graph resulting from a CNG that has been trained with $l = 2$, $l = 10$, and $l = 13$ alongside the regions that result from clusters of the graphs. While the clusters for $l = 2$ are the most spatially contiguous, there is no notable difference between $l = 10$ and $l = 13$ observable. Additionally, the graph for $l = 2$ seems much more clearly structured than the graphs for $l = 10$ and $l = 13$.

Furthermore, comparing Figure 5.7 with Figure 5.4 it can be seen that the PAS, which were designed for planning purposes, do not correspond well to the obtained clusters. The PAS consist of 12 different regions, whereas the proposed method determined only six clusters.

In order to compare the non-spatial characteristics of the clustering results and of the PAS, their mean homogeneity with respect to the different attributes is compared. The homogeneity of a cluster is calculated as the sample standard deviation of the differences between the cluster's center and the data that is assigned to it. Table 5.1 shows the mean homogeneity values of the attributes for the different clusterings. Notably, the homogeneity for all attributes decreases with increasing l. Furthermore, even though the PAS consists of double as much clusters as the clustering for $l = 13$, its mean homogeneity is mostly equal or worse. However, in contrast to the other

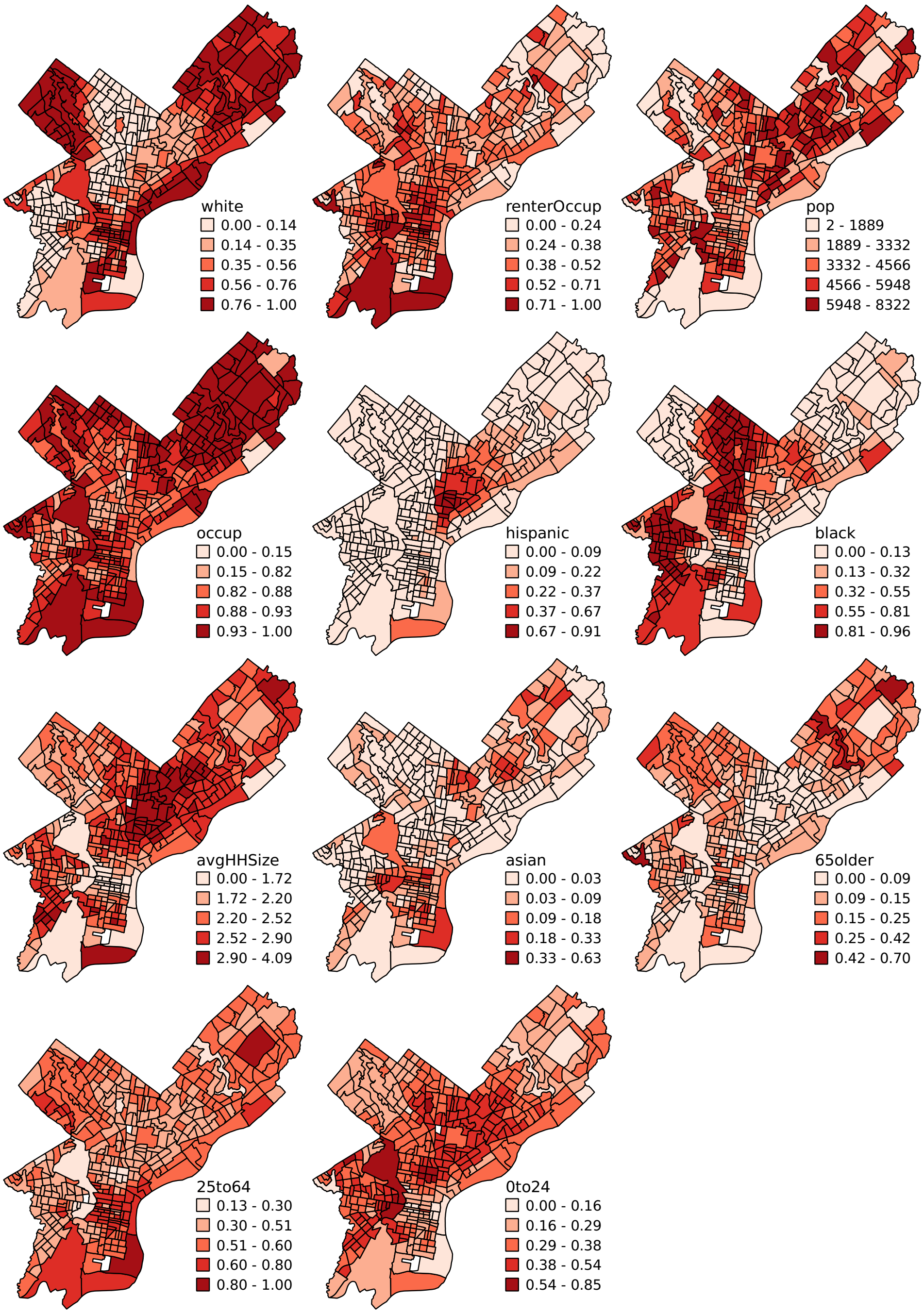


Figure 5.5.: Variables used for the experiment: Rate of Whites (white), Blacks (black), Asians (asian), Hispanics (hispanics), renter-occupied houses (renterOccup), occupied houses (occup), population younger than 25 years (0to24), population between 25 and 64 years (25to64), and population older than 64 years (65older); average size of households (avgHHSize); total population (pop).

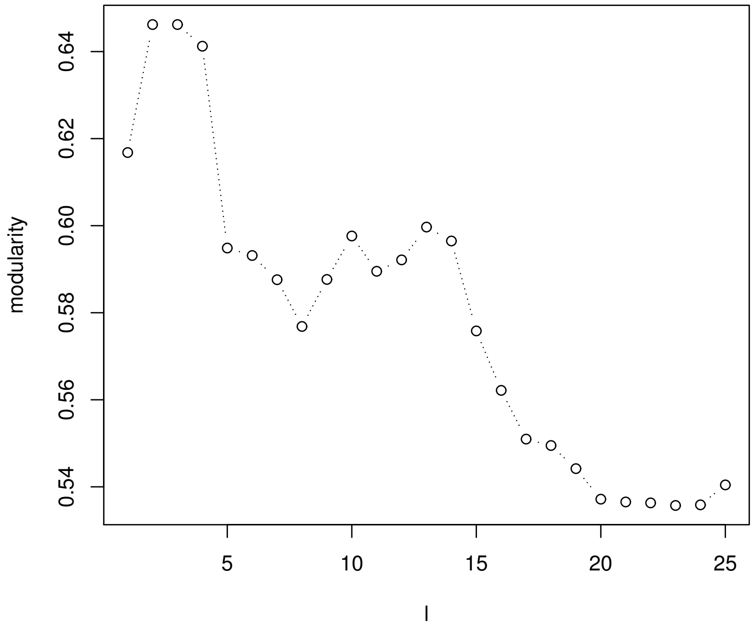


Figure 5.6.: Mean modularity scores of 100 runs for different settings of l for the real-world data set.

clusterings, the PAS is perfectly spatially contiguous. The clustering for $l = 2$ is nearly as spatially contiguous as the PAS, but it is less homogeneous than the PAS with respect to the attributes 65older, black, hispanic, and occup. However, for the majority of attributes the clustering for $l = 2$ is still more homogeneous than the PAS.

Philadelphia is one of the most segregated cities in the US; even the most affluent Blacks live in neighborhoods that are close to majority black (Logan, 2011). Hence, it can be expected that these neighborhood emerge as distinct clusters in the clustering results. Comparing Figure 5.4 with Figure 5.5 reveals that the predominantly black neighborhoods, especially in North Philadelphia, are mixed up with non-black neighborhoods or are separated by the outlines of the PAS (e.g., sections 7, 9, and 11). Also the clustering for $l = 2$ (compare Figure 5.7 with Figure 5.5) does not clearly identify the predominantly black neighborhoods. However, these neighborhoods are clearly outlined by cluster 3 of the clustering for $l = 10$ and cluster 5 of the clustering for $l = 13$.

5.5. Conclusion and further work

This study presented a new method which combines CNG, topology learning, and graph clustering to outline homogeneous regions, taking into account spatial dependence.

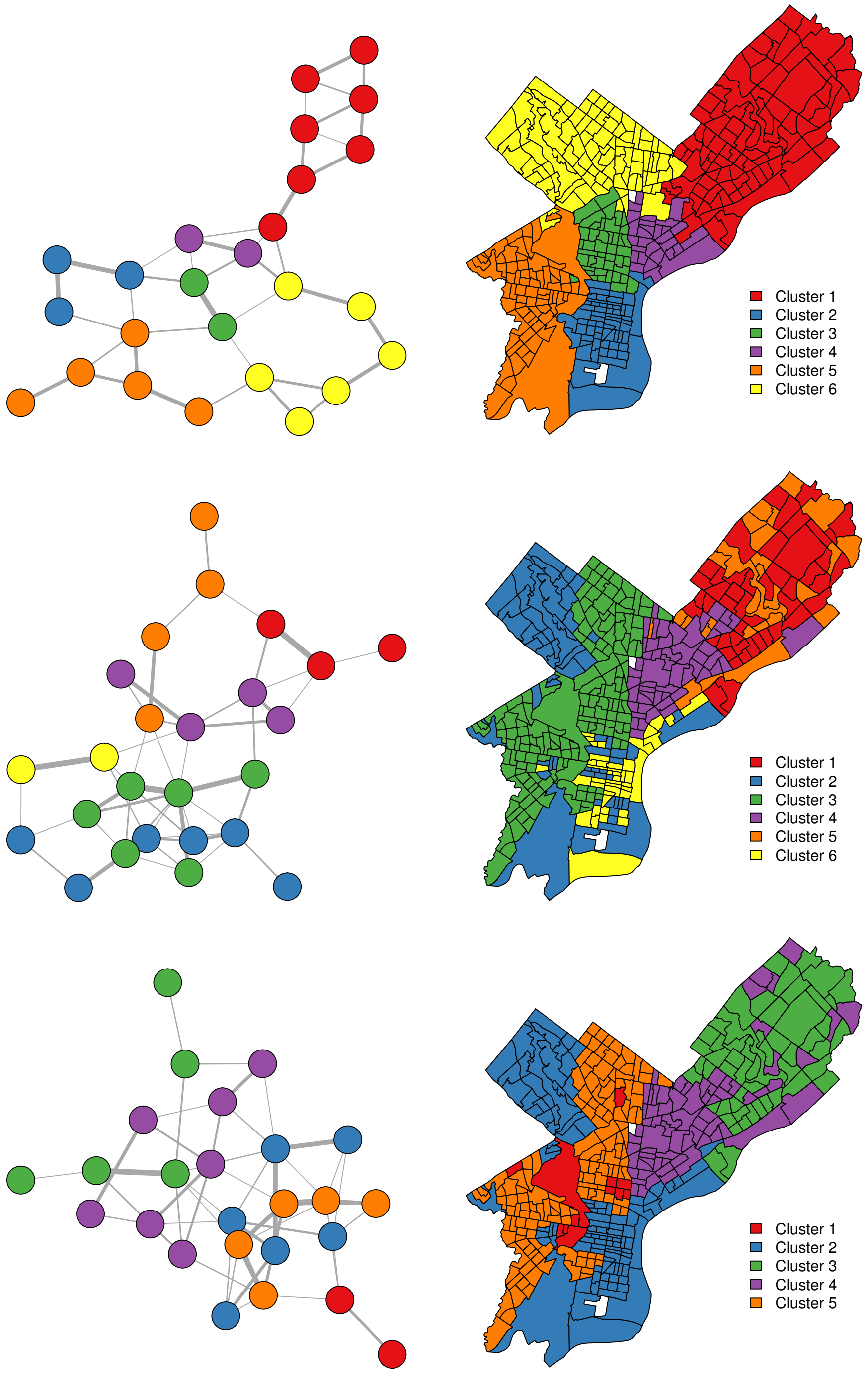


Figure 5.7.: Clustering results for $l = 2$ (top), $l = 10$ (middle), and $l = 13$ (bottom). The census tracts (right) are colored according to the colors of the detected clusters of the graph (left). The thickness of the graph's edges corresponds to the weights of the connections.

Table 5.1.: Mean homogeneity of the attributes for different clusterings.

	$l = 2$	$l = 10$	$l = 13$	PAS
pop	1157.834	1044.499	1006.730	1257.946
0to24	0.055	0.037	0.035	0.071
25to64	0.046	0.033	0.027	0.058
65older	0.041	0.032	0.026	0.026
white	0.122	0.103	0.092	0.184
black	0.119	0.101	0.090	0.090
asian	0.039	0.038	0.038	0.043
hispanic	0.040	0.037	0.033	0.033
avgHHSize	0.241	0.242	0.228	0.295
occup	0.043	0.043	0.040	0.400
rentOccup	0.112	0.113	0.106	0.106

The proposed method does not require prior knowledge about the actual number of clusters in the data, because it utilizes the modularity score when clustering the learned topology. Two experiments, one using an synthetic data set and another one using a demographic data set of Philadelphia/Pennsylvania, confirmed the usefulness of the method for delineating homogeneous clusters. Because of this property, the proposed method is in particular well-suited for spatial analysis and planning tasks.

There are some considerations that must be taken into account when applying the proposed method. The CNG algorithm uses a non-local update scheme, which prevents it from being easily stuck in local optima. However, repeated runs of the experiments have shown that the final positions of the neurons and consequently the learned topology can differ slightly with each run. This difference can possibly affect the clustering of the topology.

The results of the proposed method depend also on its parametrization. The method combines multiple algorithms, and each one's parameter setting can critically affect the final results. It is unclear how to choose the parameters so that the final results meet the analyst's requirements. In particular, the choice of the parameter l, which controls the degree of spatial dependence incorporated into the clustering, has a significant impact on the homogeneity and spatial contiguity of the clustering. However, although the chance that the resulting clusters are contiguous increases with high values of l, there is no guarantee that the clusters will ever be spatially contiguous.

In this study the proposed method utilizes a modified version of the CHL algorithm to learn a topology from CNG, but other approaches might be also reasonable. For example, instead of connecting the first and second neuron of the rank ordering to form a topology, it is also possible to remove the first neuron temporarily from the set of neurons, determine a new rank order using the CNG's ordering scheme, and then connect the first neuron of the resulting rank order with the previously found first neuron. How this strategy performs in comparison to the one used in this study is unclear and deserves further research. Additionally, CHL is sensitive to noisy data and outliers (Aupetit, 2005). Using alternative algorithms for topology learning instead bears potential to improve the results.

This study uses the MLMO algorithm to cluster the CNG's topology. The MLMO algorithm uses a greedy heuristic to optimize the modularity score of the graph. Although the algorithm has been shown to generally perform very well, it lacks accuracy, like any greedy optimization compared with other clustering methods (Fortunato, 2010). In principle any other graph clustering algorithm can be applied within the graph clustering step of the method.

The proposed method combines different methods from different but related disciplines for clustering spatial data. As scientific research for each of these disciplines is going to continue, it can be expected that more powerful methods will be developed. Utilizing these methods has the potential to further increase the value of the proposed method. In particular, improving the CNG algorithm with regard to convergence and parametrization seems worth pursuing.

Finally, the presented method is rather technical and difficult to understand and apply by non-experts. In order to be of real practical value for spatial planners and policy makers, it is necessary to integrate the method into a combined software toolkit which provides powerful analytical and visual means in order to validate the results and which is also easy to use.

Bibliography

Amedeo, D. (1969). An optimization approach to the identification of a system of regions. *Papers in Regional Science*, 23(1), 25–44.

Arribas-Bel, D. and Schmidt, C. R. (2013). Self-organizing maps and the US urban spatial structure. *Environment and Planning B: Planning and Design*, 40(2), 362–371.

Aupetit, M. (2005). Learning topology with the generative gaussian graph and the EM algorithm. In: *Advances in neural information processing systems*, 83–90.

Bação, F., Lobo, V., and Painho, M. (2005). The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computational Geosciences*, 31 (2), 155–163.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis.* Longman Scientific & Technical.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2), 172–188.

Costa, J. and De Andrade Netto, M. L. (1999). Cluster analysis using self-organizing maps and image processing techniques. In: *Proceedings of 1999 IEEE International Conference on Systems, Man, and Cybernetics.* Vol. 5, 367–372 vol.5.

Costa, J. A. F. and Oliveira, R. S. (2007). Cluster analysis using growing neural gas and graph partitioning. In: *Proceedings of International Joint Conference on Neural Networks*, 3051–3056.

De Silva, V. and Carlsson, G. (2004). Topological estimation using witness complexes. In: *Proceedings of the First Eurographics conference on Point-Based Graphics.* Eurographics Association, 157–166.

Edelsbrunner, H. and Shah, N. R. (1997). Triangulating topological spaces. *International Journal of Computational Geometry & Applications*, 7(04), 365–378.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. *Intelligent Data Analysis 5*, 5, 373–384.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5), 75–174.

Fortunato, S. and Barthelemy, M. (2006). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), 36–41.

Fritzke, B. (1995). A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems*, 7, 625–632.

Goodchild, M. F. (1986). *Spatial autocorrelation.* CATMOG. Norwich, United Kingdom: Geo Books.

Guo, D., Peuquet, D., and Gahegan, M. (2003). ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. *GeoInformatica*, 7(3), 229–253.

Hagenauer, J. and Helbich, M. (2013). Contextual neural gas for spatial clustering and analysis. *International Journal of Geographical Information Science*, 27(2), 251–266.

Hagenauer, J., Helbich, M., and Leitner, M. (2011). Visualization of crime trajectories with self-organizing maps: A case study on evaluating the impact of hurricanes on spatio-temporal crime hotspots. In: *Proceedings of the 25th International Cartographic Conference.* Paris, France: International Cartographic Association.

Han, J., Kamber, M., and Tung, A. K. H. (2001). Spatial clustering methods in data mining: A survey. In: H. J. Miller and J. Han (eds.). *Geographic Data Mining and Knowledge Discovery.* London: Taylor and Francis, 188–217.

Helbich, M., Brunauer, W., Hagenauer, J., and Leitner, M. (2013). Data-driven regionalization of housing markets. *Annals of the Association of American Geographers*, 103(4), 871–889.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3) (3), 264–323.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43 (1). 10.1007/BF00337288, 59–69.

Kohonen, T. (2001). *Self-organizing maps.* 3rd. Secaucus, NJ: Springer, New York.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, 80 (5), 056117.

Logan, J. R. (2011). *Separate and Unequal: The Neighborhood Gap for Blacks, Hispanics and Asians in Metropolitan America.* `http://www.s4.brown.edu/us2010/Data/Report/report0727.pdf` (Accessed on 28 march 2013).

Martinetz, T. and Schulten, K. (1991). A "Neural-Gas" network learns topologies. *Artificial Neural Networks*, 1, 397–402.

Martinetz, T. M., Berkovich, S. G., and Schulten, K. J. (1993). "Neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4), 558–569.

Martinetz, T. (1993). Competitive Hebbian learning rule forms perfectly topology preserving maps. In: S. Gielen and B. Kappen (eds.). *ICANN '93*. Springer London, 427–434.

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2), 284–289.

Murtagh, F. (1995). Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Letters*, 16(4), 399–408.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69 (2), 026113.

Openshaw, S. (1984). *Thoe modifiable areal problem.* CATMOG. Norwich, United Kingdom: Geo Books.

Pearsall, H. and Christman, Z. (2012). Tree-lined lanes or vacant lots? Evaluating non-stationarity between urban greenness and socio-economic conditions in Philadelphia, Pennsylvania, USA at multiple scales. *Applied Geography*, 35(1), 257–264.

Philadelphia City Planning Commission (2004). *The political and community service boundaries of Philadelphia.* `http://www.phila.gov/CityPlanning/resources/Publications/Political_boundaries.pdf` (Accessed on 26 march 2013).

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1), 27–64.

Skupin, A. and Esperbé, A. (2011). An alternative map of the United States based on an n-dimensional model of geographic space. *Journal of Visual Languages & Computing*, 22(4), 290–304.

The Philadelphia Research Initiative (2011). *A City Transformed - The Racial and Ethnic Changes in Philadelphia Over the Last 20 Years.* `http://www.pewtrusts.org/uploadedFiles/wwwpewtrustsorg/Reports/Philadelphia_Research_Initiative/Philadelphia-Population-Ethnic-Changes.pdf` (Accessed on 26 march 2013).

Ultsch, A. and Siemon, H. P. (1990). Kohonen's self organizing feature maps for exploratory data analysis. In: *Proceedings of International Neural Networks Conference.* Paris, France: Kluwer Academic Press, 305–308.

Van Der Laan, L. and Schalke, R. (2001). Reality versus policy: The delineation and testing of local labour market and spatial policy areas. *European Planning Studies*, 9(2), 201–221.

Wolfgang, M. E., Figlio, R. M., and Sellin, T. (1987). *Delinquency in a birth cohort.* Studies in Crime and Justice. University of Chicago Press.

6. Weighted merge context for clustering and quantizing spatial data with self-organizing neural networks

Authors

Julian Hagenauer

Journal

Journal of geographical systems

Status

Submitted 7 April 2014; accepted 16 October 2015; published 2 November 2015

Contribution statement

Julian Hagenauer has developed the methods, designed the experiments, and has written the manuscript for the study.

Abstract

This publication presents a generalization of merge context, named weighted merge context (WMC), which is particularly useful for clustering and quantizing spatial data with self-organizing neural networks. In contrast to merge context, WMC does not depend on a predefined (sequential) ordering of the data; distance is evaluated by recursively taking neighboring observations into account. For this purpose, WMC utilizes a weight matrix that describes the neighborhood relationships between observations. This property distinguishes WMC from existing approaches like contextual neural gas or the GeoSOM, which force spatially close observations to be represented by similar prototypes, but neglected the similarity of the observations' neighborhoods.

For practical studies, WMC is combined with the neural gas algorithm (NG) to obtain weighted merging neural gas (WMNG). The properties of WMNG and its usefulness for clustering and quantizing spatial data is investigated on two different case studies which utilize an artificial binary grid and a real-world continuous data set.

Keywords: Cluster analysis; self-organizing neural networks; spatial dependence

6.1. Introduction

Spatial as well as temporal data have become increasingly important in our everyday life. This trend is closely related to the advent of advanced technologies, which facilitate the acquisition, distribution, and storing of huge amounts of data. Well-known examples of such technologies include global positioning systems, remote sensing, geosensor networks, spatial data infrastructures, and location-based services. However, due to the effort needed to categorize or label observations, these large data sets are rarely structurally organized, even though such organization enhances the datas' value for many applications.

One computational approach for structurally organizing observations is clustering. Clustering assigns observations to clusters such that the similarity of the clusters' observations is maximized and the similarity between different clusters is minimized (Jain, 2010). Many different clustering algorithms have been developed in the past, mainly in the fields of statistics and machine learning (see, e.g., Han and Kamber, 2006; Jain, 2010). These algorithms can be broadly classified by the used paradigm. One of the most prominent and widely used clustering paradigms is partitional clustering. Partitional clustering algorithms divide a set of observations into a (smaller) set of non-overlapping and non-empty clusters. These clusters are iteratively refined by repeatedly reassigning the observations to the cluster to which they are closest. For large data sets, partitional clustering algorithms are typically more computationally effective than, e.g., hierarchical clustering algorithms (Jain et al., 1999).

Partitional clustering is closely related to vector quantization. Vector quantization forms a finite number of prototypes that approximate the probability density function of the input space. It can be practically illustrated by a Voronoi tessellation: The prototypes partition the input space into a set of Voronoi regions. Each of these regions consists of the observations that are closest to the corresponding prototype.

Since common clustering algorithms like k-means, the self-organizing map (SOM) (Kohonen, 2001; Kohonen, 1982), or neural gas (NG) (Martinetz and Schulten, 1991) find such prototypes in the process of forming clusters, they can be considered both as clustering as well as vector quantization algorithms. Methods based on artificial neural networks are particularly well suited for clustering and vector quantization tasks because they can efficiently model nonlinear relationships with few or no assumptions, which is useful when dealing with very large and complex data sets. Consequently, a large variety of powerful clustering algorithms based on artificial neural networks have been developed in the past (see Du, 2010).

When clustering and quantizing data that has a structure imposed on it, it is important to account for this structure in order to exploit the full wealth of the data. A common example of such data is temporal data, in which the state of an observation depends on the states of previous observations. This property is commonly referred to as temporal dependence. Many different self-organizing neural network models have been proposed for considering temporal dependence (e.g., Chappell and Taylor, 1993; Euliano and Principe, 1996; Kangas, 1992; Voegtlin, 2002). These models mainly differ in how they internally represent time, affecting the capacity of the model, their flexibility with respect to the network topology, and their computational performance. One approach that has been shown to be superior to those is merging neural gas (MNG) (Strickert and Hammer, 2003). MNG combines merge context, a recursive distance measure for temporal data that takes into account the state of previous observations, with the basic NG algorithm.

Another example of structural data is spatial data, in which observations are spatially located. Analogous to temporal data, these observations are seldom independent of each other; spatial observations that are spatially close to each other tend to have similar characteristics (Getis, 2010). This property is termed spatial dependence. In general, it is more complex than temporal dependence, because spatial closeness is measured in (geographic) space, which is multi-dimensional and not unidirectional. Spatial dependence is essentially important to spatial sciences because without it the variation of phenomena would be independent of location, and thus the notion of region would be less meaningful (Goodchild, 1986).

There is broad agreement that spatial sciences urgently need new and powerful methods that can handle the special properties of spatial data (see, e.g., Miller, 2010; Yuan et al., 2004) and the potential of neural networks for solving complex spatial tasks is well known (see, e.g., Fischer, 2006; S. Openshaw and C. Openshaw, 1997). However, there exist only very few self-organizing neural network models for spatial data that actually take spatial dependence into account (e.g., Bação et al., 2005; Hagenauer and Helbich, 2013). The representational capabilities of these approaches are limited, because they impose strict assumptions on the data and do barely utilize information about spatial relationships in the data.

Therefore, this study proposes a generalization of merge context, termed weighted merge context (WMC), which takes into account the distance relationships between observations in a recursive manner utilizing a weight matrix. Furthermore, it combines WMC with NG to obtain weighted merge neural gas (WMNG) and demonstrates its usefulness for clustering and quantizing spatial data.

The rest of the paper is structured as follows. Section 6.2 outlines the NG algorithm and merge context. Then, Section 6.3 introduces WMC. Section 6.4 presents WMNG on two case studies to illustrate its working and usefulness. Finally, Section 6.5 concludes with some important remarks and identifies future research.

6.2. Background

This section gives background information on NG and merge context. Merge context is of importance, because it is a specific case of WMC, which is combined with NG to obtain WMNG for practical purposes.

6.2.1. Neural gas

The NG algorithm (Martinetz and Schulten, 1991) is closely related to the SOM algorithm. However, while the latter maps a high-dimensional input space to a fixed two-dimensional output space, the NG's output space is of the same dimensions as the input space and not subject to any topological restrictions. As a consequence, the NG's quantitative performance is usually superior to the one of the SOM (e.g., Cottrell et al., 2006; Martinetz et al., 1993).

NG consists of a set of neurons M. Each neuron $k \in M$ is associated with a prototype vector m_k. It is the aim of the training of NG to distribute the prototype vectors in the input space so that they approximate its probability density function. However, the distribution of the input space is not *a priori* known, but only a sample of input vectors is given. At each training step t, a random input vector x is selected. The prototype vectors of the neurons are then moved in the direction of x, whereupon closer prototype vectors are moved more than distant ones. This update scheme enables a more robust convergence compared to, e.g., the k-means algorithm, which merely updates the closest prototype vectors at each training step (Labusch et al., 2009).

Formally, the NG algorithm can be described as follows: Let i_k denote the number of neurons in M whose prototype vectors are closer to x than m_k:

$$i_k = |\{j \in M : \|x - m_j\| < \|x - m_k\|\}| \tag{6.1}$$

Hence, i_k describes a rank order of the neurons in M. Then, each prototype vector m_k is moved towards the presented input vector x with respect to the neuron's rank i_k as follows:

$$m_k(t+1) = m_k(t) + \epsilon(t)e^{-i_k(M)/\lambda(t)}(x - m_k(t)) \tag{6.2}$$

The function $\epsilon(t)$ is the adaptation rate and $\lambda(t)$ the range of neighboring neurons to be adapted. Both functions are typically decreasing monotonically during training. A common choice of the functions which have been proven to be effective (see Martinetz et al., 1993) is the following:

$$\lambda(t) = \lambda_{init}(\lambda_{final}/\lambda_{init})^{t/t_{max}} \tag{6.3}$$

$$\epsilon(t) = \epsilon_{init}(\epsilon_{final}/\epsilon_{init})^{t/t_{max}} \tag{6.4}$$

where t_{max} denotes the maximum number of training steps. Suitable initial values (λ_{init}, ϵ_{init}) and final values (λ_{final},ϵ_{final}) must be determined for both functions. However, NG is typically not very sensitive to the particular choice of these parameters (Martinetz et al., 1993). After a sufficient number of training steps, the prototype vectors approximate the probability density function of the input space with near-minimum quantization error.

6.2.2. Merge context

Merge context (Strickert and Hammer, 2003) is a recursive distance measure for sequential data that can be used in conjunction with a wide variety of different vector quantization algorithms. It evaluates distance by not only taking into account the similarity between prototype vectors and input vectors, but also the prototypes of previously presented input vectors. For this purpose, merge context requires that each prototype k of a vector quantizer has, in addition to the mandatory prototype vector m_k, a context vector c_k of the same dimensions as the prototype vector assigned. Both vectors are updated in the course of the training. The prototype vector is moved in the direction of the presented input vector, whereas the context vector is moved in the direction of the context descriptor, which is a recursively expressed reference to the input vector's recent past. Consequently, if two prototypes have similar context vectors, then the sequences of recently presented input vectors are also similar for all input vectors that are closest to these prototypes.

In more detail, let $x_1, ..., x_t$ be a sequence of input vectors. Then, the distance between prototype k and input vector x_t is determined by:

$$d_k(x_t) = (1 - \alpha)\|x_t - m_k\| + \alpha\|C_t - c_k\| \tag{6.5}$$

The parameter $\alpha \in [0, 1]$ denotes the importance of the current input vector over the past. The past is represented by the context descriptor C_t, which is a linear combination of the context and prototype vectors of the prototype that has been closest to x_{t-1}.

More formally, let r_{t-1} be the prototype that has been closest to x_{t-1} in the previous time step. Then, the context descriptor C_t is the linear combination of the prototype vector $m_{r_{t-1}}$ and context vector $c_{r_{t-1}}$ of prototype r_{t-1}:

$$C_t = (1 - \beta)m_{r_{t-1}} + \beta c_{r_{t-1}} \tag{6.6}$$

The parameter $\beta \in [0, 1]$ determines the distant past's influence over the recent past. Basically, the context descriptor C_t constitutes an exponentially decayed sum of all previously closest prototypes. It has been shown by Strickert and Hammer (2005) that C_t converges to a global optimum.

6.3. Weighted merge context

WMC is a generalization of merge context. In contrast to merge context, it does not depend on a predefined (sequential) ordering of the data. WMC can be used for virtually any structured data set where the observations' distances can be measured. In particular, it is useful for clustering and quantizing spatial data. WMC evaluates distance by not only taking into account the similarity between prototype vectors and input vectors, but also the prototypes of input vectors that are, e.g., spatially close.

For this purpose, WMC requires that, analogous to basic merge context, each prototype k must have, in addition to the mandatory prototype vector m_k for representing patterns, a context vector c_k with the same dimensions as the prototype vector assigned. Analogous to basic merge context, both vectors are updated in the course of the training. The prototype vector is moved in the direction of the presented input vector, whereas the context vector is moved in the direction of the context descriptor, which is a recursively expressed reference to the input vector's neighborhood. Consequently, if two prototypes have similar context vectors, then the neighborhoods of the input vectors that are closest to these prototypes are also similar.

In more detail, let x_i be an arbitrary input vector. WMC does not require the input vectors to be presented in a fixed order. Then, the distance between prototype k and input vector x_i is determined as follows:

$$d_k(x_i) = (1-\alpha)\|x_i - m_k\| + \alpha\|C_i - c_k\| \tag{6.7}$$

The parameter $\alpha \in [0,1]$ weights the importance of the current input vector over the context. The context is represented by the context descriptor C_i, which is a weighted linear combination of the context and prototype vectors of the prototypes that have been recently closest to the input vectors in x_i's vicinity.

In more detail, let W be a weight matrix and $w_{i,j}$ the element in the i^{th} row and j^{th} column of the matrix, referring to the weight (e.g., spatial distance) describing the strength of the relationship of input vectors x_i and x_j, and let $r(i)$ be a function that maps an input vector x_i to the prototype that has been closest the last time x_i was presented to the quantifier. Then, the context descriptor C_i is formally defined by:

$$C_i = \frac{\sum_j w_{i,j}((1-\beta)m_{r(j)} + \beta c_{r(j)})}{\sum_j w_{i,j}} \tag{6.8}$$

The weight $w_{i,j}$ determines the importance of the prototype vector $m_{r(j)}$ and context vector $c_{r(j)}$ of prototype $r(j)$ for the description of the whole context of x_i, whereas the parameter $\beta \in [0,1]$ controls the importance of close context over more distance context.

In principle, the distance calculation of basic merge context and WMC is similar; the only difference is that the former assumes that the input vectors are part of a fixed input sequence, while the latter does not. The main difference between merge context and WMC lies in the definition of the context descriptor. The context descriptor of basic merge context only considers the prototype that has been closest to the most recent input vector, whereas WMC considers all prototypes that have recently been closest to nearby input vectors.

In addition, WMC also weights the importance of the prototypes depending on the importance of the input vectors, which are given by weight matrix. Thus, the choice of the weight matrix critically affects the results of WMC. When applying WMC to temporal sequential data, it is purposeful to set $w_{i,j}$ to one, if and only if observation i directly precedes observation j. In this way WMC resembles basic merge context. When applying WMC to spatial data, the choice of the weight matrix is usually more complex. Among the numerous different types of weight matrices, common choices in applied spatial analysis are adjacency matrices or inverse distance matrices (Getis, 2009; Getis and Aldstadt, 2004).

6.4. Case studies

This section describes two different case studies to demonstrate and discuss the usefulness of WMC. The first case study (Section 6.4.1) utilizes an artificially created binary grid data set. The advantage of using such a data set is that its settings can be precisely controlled.The second case study (Section 6.4.2) uses a real-world data set. This data set allows the evaluation of the results in a practical setting, where the data is subject to complex nuisances which typically hamper the analysis process (see, e.g., Haining, 2003).

In both case studies, WMNG, which is obtained by combining NG with WMC, is applied. The basic network settings of both networks are chosen identically for both case studies: Training time for both networks is set to $150,000$ presentations. The neighborhood range parameters are set to $\lambda_{init} = n/2$ and $\lambda_{final} = 0.01$, where n describes the number of neurons. Following Martinetz et al. (1993), the adaptation rate parameters are set to $\epsilon_{init} = 0.5$ and $\epsilon_{final} = 0.005$. The parameters $\lambda(t)$ and $\epsilon(t)$ are chosen as described in Section 6.2.1.

6.4.1. Quantization of binary grid data

For this case study, a binary grid of size 50×50 is created. In order to avoid edge effects, the opposite sides of the grid are connected to form a toroid. Initially, all cells of the grid are colored white. Then, white cells are randomly selected and colored black until half of all cells are black. The probability for getting selected thereby increases exponentially with the number of neighboring black cells. As a consequence, black cells tend to be located close to other black cells. Join count statistics confirm that the resulting black and white patterns are strongly spatially dependent ($p < 0.05$). Figure6.1 depicts the resulting binary grid.

Generally, the quality of a context quantifier is defined as the expected number of

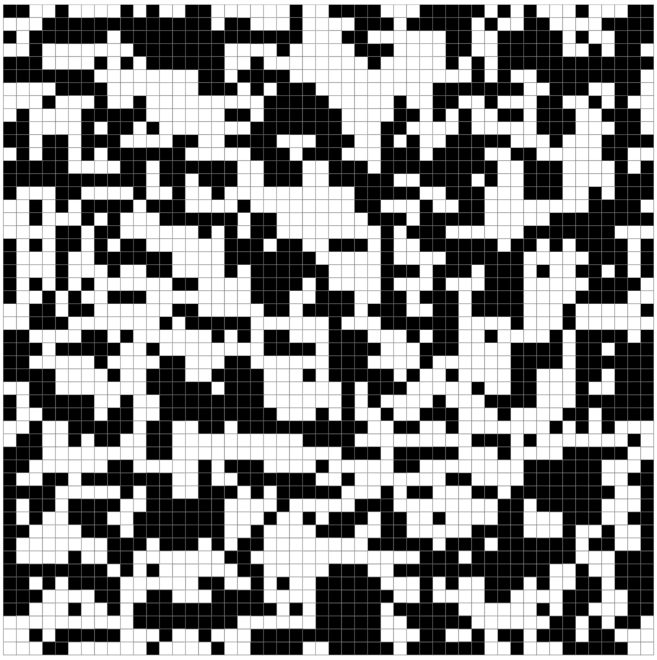


Figure 6.1.: Artificially created binary grid data set.

observations that can be correctly reconstructed from its prototypes (Voegtlin, 2002). Based on this argument, Voegtlin (2002) proposed a performance measure for binary temporal data. However, this measure cannot directly be applied to a spatial context quantifier, because it requires sequential ordering of the data. Furthermore, it does not take into account probabilities for correctly reconstructing observations. Therefore, a new measure which permits to evaluate the performance of a spatial context quantifier on binary grid data is proposed.

Let c_i^l be the set of grid cells whose distance to the grid cell i is exactly l. c_i^l is termed the spatial neighborhood of i for distance l. For example, c_i^0 consists only of grid cell i, whereas the set c_i^1 consists of the direct neighbors of cell i. Furthermore, associate with each grid cell in c_i^l its position in the grid, relative to the grid cell i, and let R_k be the set of all grid cells to which prototype k is closest to. R_k is termed the receptive field of prototype k. Then, the prototype k's probability of reconstructing the value of one at a certain relative position can be determined by calculating the arithmetic mean of the grid cells of c_i^l with $i \in R_k$ for this specific position. The representation of these probabilities for a certain distance l and prototype k as grid cells is termed *probability field* (PF_k^l).

Exemplary probability fields of a prototype k for distance zero, one, and two are depicted in Figure6.2. In this figure, the prototype's probability of reconstructing the

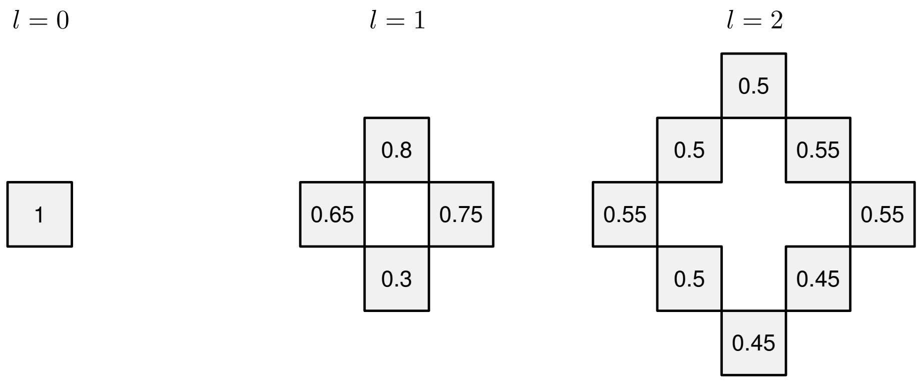


Figure 6.2.: Exemplary probability fields of a prototype k for different distances l.

value one at the center cell is one, whereas it is significantly lower for grid cells at distance one. For grid cells at distance two, the probability of reconstructing the value one is close to 0.5, indicating that at this distance the prototype reconstructs the value one and two with basically the same probability.

The *entropy rate* h_k^l of a prototype k for distance l can then be calculated as follows:

$$h_k^l = -\frac{1}{|PF_k^l|} \sum_{p \in PF_k^l} p \log(p) + (1-p)\log(1-p) \tag{6.9}$$

where PF_k^l is the probability field of prototype k for distance l and p refers to the probabilities of its grid cells.

h_k^l is a quality measure which quantifies the prototype k's uncertainty of reconstructing observations at distance l. Its values range from zero (no uncertainty) to one (total uncertainty). For example, the prototype of the probability fields that are depicted in Figure 6.2 is able to perfectly reconstruct the center grid cell without any uncertainty ($h_k^0 = 0$), whereas its uncertainty of reconstructing grid cells with distance one is significantly higher ($h_k^1 = 0.837$). At distance two, the prototype's uncertainty is close to maximum ($h_k^2 = 0.996$).

In practice, not all prototypes of a vector quantifier are equally probable. Therefore, it is necessary to take into account each prototype's probability p_k in order to evaluate a spatial context quantifier:

$$\overline{n}_l = \sum_k p_k h_k^l \tag{6.10}$$

$\overline{n}_l$ is termed the *uncertainty* of a spatial context quantifier. It is a quality measure which reflects the overall quantifiers' ability to represent observations at a certain distance l.

WMNG consists of ten neurons for this case study. Furthermore, following the grid topology of the data set, it utilizes a simple rook's case adjacency matrix for weighting observations: the distance between two grid cells is one if they share one edge, otherwise it is zero. Rook's case adjacency is also used to evaluate $\overline{n}_l$ throughout this experiment.

In order to investigate the influence of WMNG's parameters α and β on modeling of spatial context, WMNG is trained 16 times for each setting of α and β between zero and one in steps of 0.01 and the mean uncertainty for distances zero to five is calculated. The results are shown in Figure6.3.

The figure reveals several important insights: Firstly, because grid cells at distance zero refer to the grid cells themselves and not to their context, context quantization is not relevant at all and hence the results for this distance are considerably different from the results for larger distances. In particular, the uncertainty for all settings of α and β, except for very large values of α, is close to zero, indicating that WMNG is perfectly able to reconstruct grid cells at distance zero as long as WMNG does not focus too much on context modeling.

Secondly, the minimum uncertainty increases rapidly from distance zero to five. This shows that the reconstruction of close grid cells is more easier than of distance grid cells. The reason for this is that the probability fields for large distances consist of more cells than the ones for small distances and are consequently more difficult to reconstruct. Moreover, at distance four and five uncertainty is generally close to maximum for all parameter settings, indicating that WMNG does not significantly represent context information for these distances.

Finally, the parameter β has more influence on the results, than parameter α which is constant for a wide range of different values. In order to inspect the role of the β-parameter in more detail, two settings have been marked in the figure: $\alpha = 0.5$, $\beta = 0.9$ with a triangle and $\alpha = 0.5$, $\beta = 0.4$ with a circle. Particular interesting are these settings for distance one and two: At distance one, the circle-setting has a lower uncertainty, while at distance two the triangle-setting has a lower uncertainty. This is because for low values of β, WMNG focuses more on the quantization of the close context, whereas it focuses more on the distant context for large values of β.

It can be concluded from this case study that WMNG is able to effectively quantize the spatial context of binary grid data for a very wide range of different parameter settings. However, while its effectiveness is high for low distances, it rapidly decreases with distance.

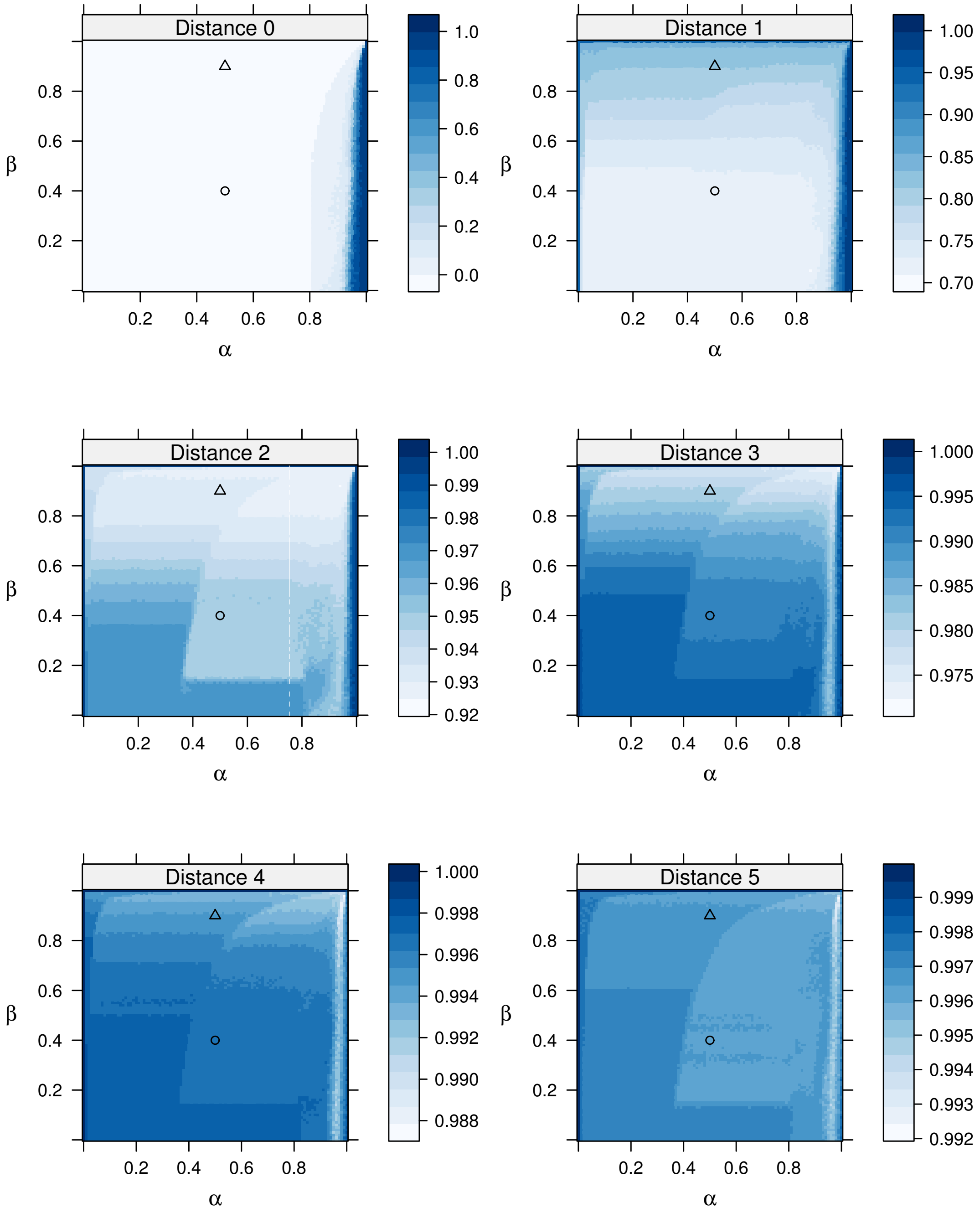


Figure 6.3.: Uncertainty of WMNG for the binary grid data set. Two settings are highlighted: $\alpha = 0.5$, $\beta = 0.9$ (△) and $\alpha = 0.5$, $\beta = 0.4$ (○).

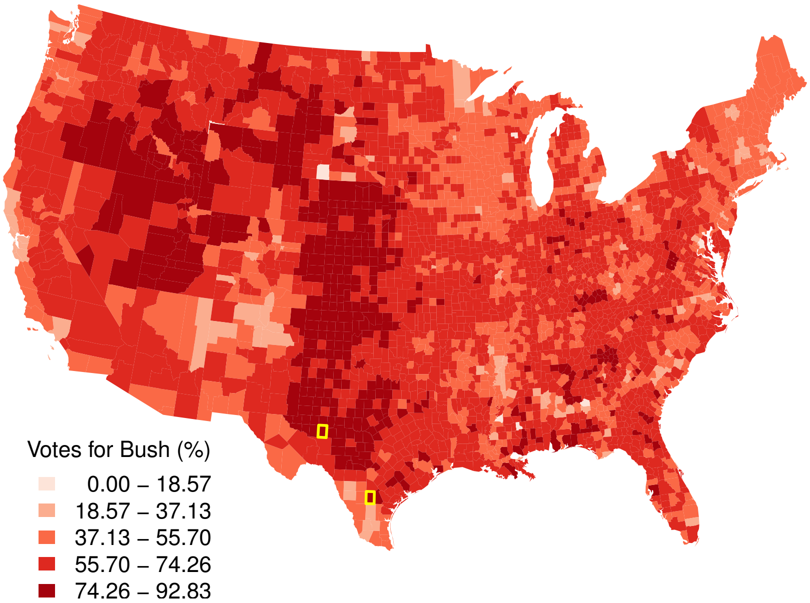


Figure 6.4.: Votes for Bush (%) in the 2004 US presidential election. The counties Reagan, TX, and McMullen, TX, are outlined in yellow.

6.4.2. Quantization and clustering of (continuous) real-world vector data

This case study uses the 2004 US presidential election data for $3,111$ US counties. Even though the data set consists of multiple variables, only the percentage of total votes for Bush in each county is used for ease of visualization and traceability. The distribution of this variable is depicted in Figure6.4. In addition, the counties Reagan, TX, and McMullen, TX, are highlighted. These counties are spatially close and have very similar votes for Bush (83.845% and 83.096%), but have very different neighboring counties. Reagan is surrounded by counties with very high percentages of votes for Bush, while the percentages of the counties surrounding McMullen are generally low.

WMNG consists of eight neurons for this case study. It utilizes a queen's case adjacency matrix for weighting observations: the distance between two observations is one if their assigned boundary touches at at least one point, otherwise it is zero. Prior to the training, the adjacency matrix is row-normalized to make the calculation of the weighted sums for the context descriptor independent from the number of neighboring observations.

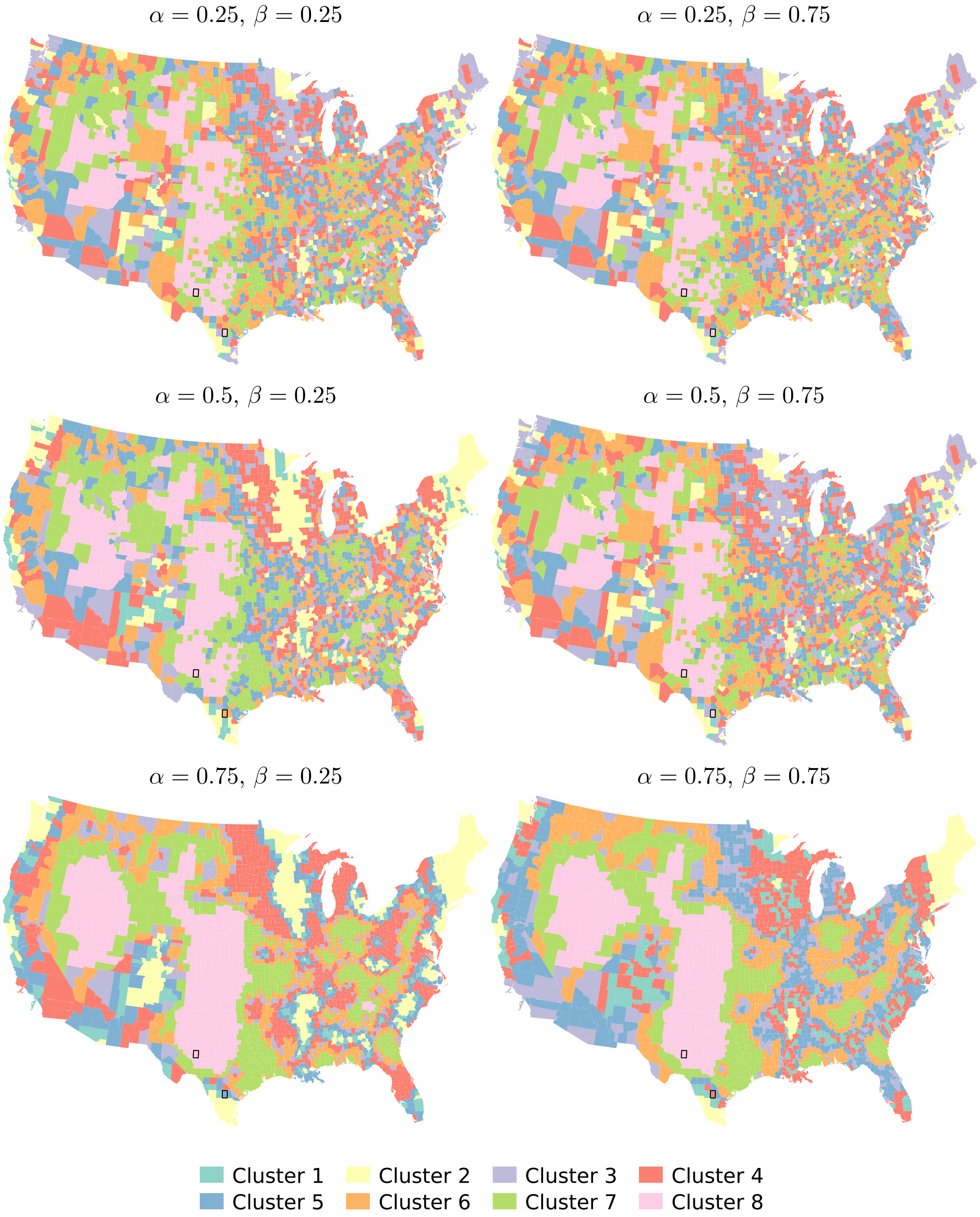
Figure6.5 depicts the resulting clusters of WMNG for different values of α and β. Table 6.1 lists several relevant statistics of the WMNG's results: The quantization error (QE) is a measure of quantization performance. It is calculated by determining the average distance between the prototype vectors and their closest observations. The within-sum of squares (WSS) evaluates the quality of a clustering. It is calculated by summing the squared distances of the clusters' means and their assigned observations. Each cluster consists of one or more contiguous groups of observations which are spread over the map. The total number of such contiguous groups expresses the spatial contiguity of a clustering. The last column of the table indicates if the counties Reagan, TX, and McMullen, TX, are mapped to different clusters.

Table 6.1.: Selected relevant statistics for the trained WMNG.

α	β	QE	WSS	Contiguous groups	Different cluster?
0.25	0.25	1.955	19,033.593	1,018	No
	0.75	1.930	18,369.907	1,041	No
0.50	0.25	3.040	47,991.965	777	Yes
	0.75	2.060	21,596.097	965	No
0.75	0.25	4.975	137,636.670	515	Yes
	0.75	4.597	121,462.489	414	Yes

The table reveals that QE and WSS increase with increasing values of α. This is because the lower the values of α of WMNG, the less it considers the spatial neighborhoods of observations and the more it focuses on the representation of the actual observations, but only the latter is evaluated by the measures. In addition, it can be seen that QE and WSS decrease with increasing values of β. The reason for this is that for large values of β, WMNG considers very distant spatial neighborhoods for the representation of counties. However, these distant neighborhoods are mostly unrelated to the actual counties. As a result, the variance of the trained context vectors is lowered, which increases the importance of the prototype vectors for the distance evaluation and thus improves the representation of the actual counties.

The figure shows that many clusters of WMNG appear spatially nested. This nestedness increases with the values of α and is a result of considering the neighborhood relationships of observations in the clustering process. For example, for WMNG with $\alpha = 0.75$ and $\beta = 0.25$ the clusters 4 and 5 have very similar characteristics ($57.404 \pm 6.426\%$ and $57.989 \pm 6.694\%$) and are generally close to each other. Nevertheless, they are considered different by WMNG, because the counties of cluster 4 are spatially closer to the ones of cluster 6, which have higher percentages of votes for Bush ($66.820 \pm 5.025\%$), whereas the counties of cluster 5 are spatially closer to the ones of cluster 2, which have lower percentages ($44.946 \pm 8.317\%$). As a consequence, these



$\alpha = 0.25$, $\beta = 0.25$

$\alpha = 0.25$, $\beta = 0.75$

$\alpha = 0.5$, $\beta = 0.25$

$\alpha = 0.5$, $\beta = 0.75$

$\alpha = 0.75$, $\beta = 0.25$

$\alpha = 0.75$, $\beta = 0.75$

Figure 6.5.: Cartographic maps of the clusters outlined by WMNG. The counties Reagan, TX, and McMullen, TX, are outlined in black.

clusters appear nested. The parameter β has little to no influence on this characteristic of WMNG.

Table 6.1 and Figure6.5 show that the number of contiguous regions decreases with increasing values of α. This is because the importance of the spatial neighborhoods in the clustering process increases with the value α. No clear trends of the number of contiguous regions are notable for parameter β.

Finally, the counties Reagan, TX, and McMullen, TX, are of particular interest, because they have very similar characteristics and are spatially close, but have very different spatial neighborhoods. Consequently, distance-based spatial clustering algorithms that do not consider the similarity of the counties' neighborhoods like contextual neural gas (Hagenauer and Helbich, 2013) or the GeoSOM (Bação et al., 2005) tend to assign both counties to the same cluster. By contrast, Table 6.1 and Figure6.5 show that for large values of α, WMNG is able to distinguish between both counties.

To conclude, this case study demonstrates the properties and advantages of WMNG for clustering continuous data sets. Varying with the degree to which neighborhood relationships are considered, WMNG can produce a wide range of different clusterings in a practical setting that make it a viable alternative to existing approaches based on self-organizing neural networks.

6.5. Closing remarks

This study proposed WMC, a generalization of merge context, as a powerful procedure for quantizing and clustering spatial data. WMC takes into account neighboring observations in a recursive manner to evaluate the similarity between observations. In this way WMC provides a richer representation of spatial context compared to spatial clustering algorithms that only implicitly consider the neighborhood of observations by mapping spatially nearby observations to close prototypes, assuming that their neighborhood characteristics are also similar.

Furthermore, WMC also does not require that the location of observations is explicitly represented by a single coordinate vector for calculating the spatial distances between them. A weight matrix denoting the spatial distances is sufficient. This makes WMC in particular useful for clustering complex spatial data, where the distance between observations cannot be simply evaluated as the Euclidean distance between single coordinate vectors, e.g., data where observations refer to areas.

Since WMC is merely a distance measure, it can be combined with practically any quantization algorithm. The combination with the k-means algorithm, which is arguably the most often applied quantization and clustering algorithm in science, is straight-forward. In addition, the combination of WMC with SOMs seems particularly beneficial, since SOMs are well suited for visualizing and mapping of high-dimensional data (Flexer, 2001; Vesanto, 1999), which makes them particularly useful for spatial sciences (see Agarwal and Skupin, 2008).

WMC does not distinguish between distant observations as long as the observations themselves and their neighborhoods are sufficiently similar. Depending on the analysis task at hand, this property of WMC might not always be desired. However, WMC can easily be modified to differentiate between distant observations by, e.g., introducing a similarity threshold for distant observations. How such a modification affects the results of WMC deserves further research.

A main concern when using WMC is the number of parameters that must be specified. WMC depends on two parameters, α and β, which determine the extent to which spatial context is taken into account. In practice, WMC produces comparable results for a wide range of different parameter settings. Nevertheless, the parameter α could be eliminated by utilizing an entropy-based adaption scheme (Strickert and Hammer, 2005) to optimize quantization performance. Additionally, it is necessary to determine an appropriate weight matrix that works well for the given task. General guidelines for specifying the weight matrix of a spatial model have been proposed by Griffith (1996).

Similar to basic merge context, WMC is subject to the sequence of presentations, since it recursively evaluates the neighborhood of observations based on observations that have previously been presented. However, when applying WMC to spatial data, neighboring observations are not temporally but spatially dependent. In this case it is assumed, given that WMC is combined with a self-organizing neural network, that the sequence of presentations becomes more and more irrelevant for the quantization process as the prototypes converge to their final states. In fact, the experiments in this study have confirmed that WMC combined with neural gas produces reasonable results for spatial data. However it is necessary for future research to investigate the temporal dynamics of WMC in more detail.

Bibliography

Agarwal, P. and Skupin, A. (eds.) (2008). *Self-organising maps: Applications in geographical information science*. Chichester, United Kingdom: John Wiley & Sons. Chap. Introduction: What is a self-organizing map?

Bação, F., Lobo, V., and Painho, M. (2005). The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computational Geosciences*, 31(2) (2), 155–163.

Chappell, G. J. and Taylor, J. G. (1993). The temporal Kohønen map. *Neural Networks*, 6(3), 441–445.

Cottrell, M., Hammer, B., Hasenfuß, A., and Villmann, T. (2006). Batch and median neural gas. *Neural Networks*, 19(6), 762–771.

Du, K.-L. (2010). Clustering: A neural network approach. *Neural Networks*, 23(1), 89–107.

Euliano, N. R. and Principe, J. C. (1996). Spatio-temporal self-organizing feature maps. In: *Proceedings of the IEEE International Conference on Neural Networks*. Vol. 4. Washington, DC: IEEE, 1900–1905.

Fischer, M. M. (2006). Neural networks. A general framework for non-linear function approximation. *Transactions in GIS*, 10(4), 521–533.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. *Intelligent Data Analysis*, 5(5), 373–384.

Getis, A. (2009). Spatial weights matrices. *Geographical Analysis*, 41(4), 404–410.

Getis, A. (2010). Spatial autocorrelation. In: M. M. Fischer and A. Getis (eds.). *Handbook of Applied Spatial Analysis*. Springer Berlin Heidelberg, 255–278.

Getis, A. and Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36(2), 90–104.

Goodchild, M. F. (1986). *Spatial autocorrelation*. CATMOG. Norwich, United Kingdom: Geo Books.

Griffith, D. A. (1996). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In: S. L. Arlinghaus (ed.). *Practical Handbook of Spatial Statistics*. Boca Raton, FL: CRC Press, 65–82.

Hagenauer, J. and Helbich, M. (2013). Contextual neural gas for spatial clustering and analysis. *International Journal of Geographical Information Science*, 27(2), 251–266.

Haining, R. P. (2003). *Spatial data analysis: Theory and practice.* Cambridge University Press.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and techniques.* San Francisco, CA: Morgan Kaufmann Publishers.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3) (3), 264–323.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.

Kangas, J. (1992). Temporal knowledge in locations of activations in a self-organizing map. In: *Artificial Neural Networks 2.* Ed. by I. Aleksander and J. Taylor. Vol. 1. Amsterdam, Netherlands: North-Holland, 117–120.

Kohonen, T. (2001). *Self-organizing maps.* New York: Springer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1) (1), 59–69.

Labusch, K., Barth, E., and Martinetz, T. (2009). Sparse coding neural gas: Learning of overcomplete data representations. *Neurocomputing*, 72(7) (7-9), 1547–1555.

Martinetz, T. and Schulten, K. (1991). A "Neural-Gas" network learns topologies. In: *Artificial Neural Networks.* Ed. by T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas. Amsterdam, The Netherlands: Elsevier, 397–402.

Martinetz, T., Berkovich, S., and Schulten, K. (1993). "Neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4), 558–569.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, 50(1), 181–201.

Openshaw, S. and Openshaw, C. (1997). *Artificial intelligence in geography.* New York, NY: Wiley.

Strickert, M. and Hammer, B. (2003). Neural gas for sequences. In: *Proceedings of the Workshop on Self-Organizing Networks (WSOM).* Ed. by T. Yamakawa. Kyushu, Japan: Kyushu Institute of Technology, 53–57.

Strickert, M. and Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing*, 64, 39–71.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2), 111–126.

Voegtlin, T. (2002). Recursive self-organizing maps. *Neural Networks*, 15(8-9), 979–991.

Yuan, M., Buttenfield, B., Gahegan, M., and Miller, H. (2004). Geospatial data mining and knowledge discovery. In: R. B. McMaster and E. L. Usery (eds.). *A Research Agenda for Geographic Information Science.* Boca Raton, FL: CRC Press, 365–388.

7. SPAWNN: A toolkit for ***SP***atial ***A***nalysis ***W***ith self-organizing ***N***eural ***N***etworks

Authors

Julian Hagenauer and Marco Helbich

Journal

Transactions in GIS

Status

Published online 2 February 2016; published on 16 October 2016

Contribution statement

Julian Hagenauer has developed the methods, designed the experiments, and has written the manuscript for the study. Marco Helbich supported this publication by continuously discussing the design and results of the study and by proofreading the manuscript.

Abstract

This article introduces the SPAWNN toolkit, an innovative toolkit for spatial analysis with self-organizing neural networks, which is published as free and open-source software (`http://www.spawnn.org`). It extends existing toolkits in three important ways. First, the SPAWNN toolkit distinguishes between self-organizing neural networks and spatial context models with which the networks can be combined to incorporate spatial dependence and provides implementations for both. This distinction maintains modularity and enables a multitude of useful combinations for analyzing spatial data with self-organizing neural networks. Second, SPAWNN interactively links different self-organizing networks and data visualizations in an intuitive manner to facilitate explorative data analysis. Third, it implements cutting-edge clustering algorithms for identifying clusters in the trained networks. Toolkits such as SPAWNN are particularly needed when researchers and practitioners are confronted with large amounts of complex and high-dimensional data. The computational performance of the implemented algorithms is empirically demonstrated using high-dimensional synthetic data sets, while the practical functionality highlighting the distinctive features of the toolkit is illustrated with a case study using socioeconomic data of the city of Philadelphia, Pennsylvania.

Keywords: Self-organizing neural networks; spatial clustering; spatial analysis

7.1. Introduction

The amount of available spatial data has increased rapidly in recent years due to advances in spatially explicit technologies for acquiring, sharing, and storing spatial information (Miller and Goodchild, 2014). This data often contains hidden and a priori unexpected information, which can hardly be explored using traditional statistical methods that require hypothesis testing and are not developed to handle such large amounts of data (Miller and Han, 2009). Spatial data mining explicitly addresses these issues by adopting state-of-the-art methods from the fields of artificial intelligence, machine learning and spatial database systems, among others, in order to extract information and to ultimately transform it into new and potentially useful knowledge (Yuan et al., 2004).

Clustering is a particularly useful method in spatial data mining, because it organizes observations into clusters such that the similarity within a cluster is maximized while the similarity between different clusters is minimized (Jain, 2010). In this way, it imposes a structural organization on the data, which facilitates further analysis and alleviates data exploration. This analysis and data exploration are often performed by a human analyst, whose ability to perceive and understand patterns — through visual representations — exceeds the capabilities of computational algorithms (Keim, 2002; Ware, 2012). Therefore, it is convenient and efficient to combine clustering methods with appropriate visualizations and interactive means in a combined toolkit.

Spatial clustering is the task of clustering spatial data, which is fundamentally different from non-spatial data (see Grubesic et al., 2014). An essential property of spatial data is that observations are usually spatially dependent (Sui, 2004), meaning that observations that are spatially close to each other tend to have similar characteristics. Without this property, the variation of phenomena would be independent of location, and thus, the notion of region would be less meaningful (Goodchild, 1986). However, the available data might not be sufficient to accurately model the spatial varying phenomena and thus, if spatial data is clustered while spatial dependence is neglected, the results may lead to an incomplete understanding of the spatial patterns (Openshaw, 1999).

While many different clustering algorithms for spatial and non-spatial data have been proposed in the literature (e.g., Guo, 2008; Jain, 2010; Parimala et al., 2011), few neural network-based clustering approaches that explicitly account for spatial dependence have been developed. Two notable exceptions are the GeoSOM (Bação et al., 2005) and contextual neural gas (CNG) (Hagenauer and Helbich, 2013). Both are adaptations of basic self-organizing network algorithms that utilize the spatial arrangement of the neurons to account for spatial dependence. However, these approaches are purely computational; they still require a human analyst to interpret the clustering results in the light of domain-specific knowledge and, if necessary, to adjust the parameter settings, which involves repeating the analysis. To facilitate this task, it is necessary to integrate different self-organizing neural network-based clustering methods, where each comes with its unique advantages, in an interactive toolkit with other computational, visual, and geographic methods. Such a toolkit should be intuitive and easy to use so that its usage is promoted across different spatial disciplines.

To address the lack of such toolkit, this article introduces SPAWNN, an innovative toolkit for SPatial Analysis With self-organizing Neural Networks, which implements the self-organizing map (SOM) (Kohonen, 2001; Kohonen, 1982) and neural gas (NG) (T. Martinetz and Schulten, 1991; T. M. Martinetz et al., 1993) algorithms. The toolkit extends existing toolkits in three important ways: First, it is the first toolkit that allows these self-organizing networks to be combined with either the CNG or the GeoSOM approach, or with alternative spatial context models, in order to account for spatial dependence. Second, the toolkit provides different visualizations and links between the neurons and a geographic map. This permits the analyst to interactively select neurons or observations and to visually inspect the mapping between them in order to explore the results of the trained networks in detail. Third, the toolkit provides a set of powerful clustering algorithms for post-processing the network models. The article demonstrates the usefulness of the presented toolkit with a case study exploring census data in Philadelphia, Pennsylvania.

The article is structured as follows. Section 7.2 reviews existing toolkits for spatial cluster analysis. Section 7.3 discusses self-organizing neural networks, while Section 7.4 introduces different models for incorporating spatial dependence into the networks. The SPAWNN toolkit is presented in Section 7.5. In Section 7.6 the computational demand of the implemented spatial context models is analyzed, while Section 7.7 illustrates the application of the SPAWNN toolkit to practical analytical problems. Section 7.8 concludes the article and discusses future work.

7.2. Related work

The process of exploring and analyzing spatial patterns usually involves the application of diverse methods from the fields of spatial data mining and geographic information systems (GIS) (Mennis and Guo, 2009). In order to facilitate this process, numerous software toolkits have been developed that combine different methods from both fields in an integrated and user-friendly environment.

One of the first of such toolkits is GeoMiner, introduced by Han et al. (1997), which enhances the relational data mining system DBMiner (Han et al., 1996). GeoMiner's main feature is its ability to mine three kinds of knowledge rules in spatial databases. For this purpose, the authors proposed a geographic query language. Other features of GeoMiner include the integration of data warehousing and GIS technologies, a user interface, and multiple forms of outputs, including generalized maps, generalized relations, cross-tabulation, and charts. Another prototypical approach that integrates data mining methods and geographic visualization is KGConstruct (MacEachren et al., 1999). It provides three dynamically linked forms of representation: geographic maps, 3D scatter plots and parallel coordinate plots, which can be independently or simultaneously manipulated through applications of different interaction forms in order to explore the spatial data. More recently, Anselin et al. (2006) developed the popular GeoDA tool. It comprises a variety of different approaches for analyzing spatial data, including histograms, box plots, scatter plots and chloropleth maps. Dynamically linked windows, which combine geographic maps and statistical plots, are used for exploratory analysis. GeoDA also provides univariate cluster detection methods, such as local indicators of spatial association (Anselin, 1995). However, similar to KGConstruct, it does not offer multivariate clustering algorithms. Körting et al. (2013) proposed GeoDMA, a toolkit that combines remote sensing image analysis capabilities with spatial data mining techniques. More specifically, the toolkit includes methods for image segmentation, feature extraction, feature selection, classification, landscape metrics and multitemporal methods for change detection and analysis. In order to provide access to common GIS functions, the toolkit is tightly integrated into a freely available GIS software.

The aforementioned combined toolkits have demonstrated that the linkage of different representations of spatial data and data mining methods is useful for many complex spatial analysis tasks. However, none of them supports SOMs, a data mining method that has shown to be very useful for visualization, clustering, and data analysis tasks (e.g., Estévez and Figueroa, 2006; Flexer, 2001; Tasdemir and Merényi, 2009). The two-dimensional topology of SOMs particularly promotes their integration with other GIS methods in an interactive environment for spatial data analysis (Skupin and Agarwal, 2008).

One of the first toolkits that effectively made use of SOMs is GeoVISTA Studio, introduced by Takatsuka and Gahegan (2002). While its arguably most distinctive feature is its component-oriented design, which embraces visual programming to facilitate the development of data analysis and visualization programs, the toolkit also provides means for training SOMs and for linking the SOMs' results with different visualizations and statistical analysis methods. Following a similar approach, Guo et al. (2005) developed SOMVIS, an integrated environment which consists of four major components: SOMs, parallel coordinate plots, geographic maps, and a two-dimensional color design tool. The combination of the computational algorithms and visual methods ought to mitigate each other's weaknesses and thus to facilitate the exploration and discovery of spatial patterns. Because SOMVIS focuses on spatial data, Guo et al. (2006) have extended the toolkit to accommodate for the temporal dimension as well, in order to explore spatiotemporal mappings. SOM Analyst (Lacayo-Emery, 2011) comprises a basic set of tools for using SOMs within the proprietary but widely used ArcGIS platform. It includes tools for data preprocessing, SOM computation and SOM visualization that extend common GIS functions. Furthermore, while the toolkit does not implement direct linkage between SOMs and the data, it supports the mapping of data to an existing SOM. Andrienko et al. (2010) proposed a framework based on SOMs combined with a set of interactive visual tools that support different analytic perspectives for the analysis of spatiotemporal data. The SOM visualization is linked to a geographic map, a time series graph, and a periodic pattern view. In this way, the analysis of SOM results in both the spatial and temporal dimensions is supported. Finally, the GeoSOM suite, introduced by Henriques et al. (2012), mainly differs from the above mentioned approaches in that it does link the GeoSOM, which is particularly tailored to spatial data, instead of a basic SOM to different visualizations and analysis methods. Therefore, the effective coupling of the GeoSOM with a geographic map is even more important.

The above literature review showed that currently only a few toolkits exist that combine SOMs, spatial analysis and GIS within an interactive and user-friendly environment. However, these toolkits have two important deficiencies: First, while basic SOMs are well supported, other self-organizing neural networks are not implemented in the toolkits, even though it has been shown that SOMs are less appropriate for certain analysis tasks such as vector quantization (e.g., Hagenauer and Helbich, 2013; Strickert and Hammer, 2005). Furthermore, as the application of different neural network algorithms can potentially yield different results, a direct comparison between several network models is useful to enhance the understanding of the data. Second, besides the GeoSOM suite, all toolkits use the basic SOM algorithm, which does not account for spatial dependence at all. The GeoSOM suite, on the other hand, is exclusively restricted to the GeoSOM algorithm; it does not support alternative approaches for considering spatial dependence, even though those can produce valuable results for certain analysis tasks (see Hagenauer and Helbich, 2013).

7.3. Self-organizing neural networks

Self-organizing neural networks represent a class of artificial neural networks (ANNs) that are trained in an unsupervised manner. This means that they optimize some task-independent performance criterion, which is defined in terms of the neuronal activity, to detect similarities in the input data (Fischer, 1998). After the training, the network represents the learned data in a more explicit or simple form (Becker, 1991), which is useful for clustering and analysis tasks. Among the large variety of self-organizing neural networks (e.g., Carpenter and Grossberg, 1991), it has been demonstrated that the SOM is particularly useful for a wide range of analytical problems (e.g., Chon, 2011; Kalteh et al., 2008; Kaski et al., 1998; Liu and Weisberg, 2011; Oja et al., 2003). This subsection briefly introduces the SOM and the NG algorithm, a closely related algorithm that is particularly useful for clustering tasks.

7.3.1. Self-organizing map

The SOM (Kohonen, 2001; Kohonen, 1982) consists of an arbitrary number of neurons that are connected to adjacent neurons by a neighborhood relation, defining the topology of the map. In principle, the dimension of a SOM is arbitrary, but in practice, two-dimensional SOMs are commonly used for visualization purposes. Associated with each of these neurons is a prototype vector of the same dimension as the input space. During the training, input vectors are presented to the SOM, and the neuron with the smallest distance to the input vector, referred to as the best matching unit (BMU), is identified. Then, the prototype vector of the BMU and the prototype vectors within a certain neighborhood on the map are moved in the direction of the input vector. The magnitude of the displacement depends on the distance of the neurons to the BMU on the map and on the actual learning rate. Both the size of the neighborhood and the learning rate decrease monotonically during the learning process. Thus, in the beginning of the learning phase, the arrangement of neurons on the map can be altered significantly, while at the end of the training phase, only small changes are made to fine-tune the map. The trained SOM represents a low-dimensional map of the input space, where each neuron represents some portion of the input space and where the distance relationships of the input space are mostly preserved.

7.3.2. Neural gas

Similar to the SOM, the NG algorithm (T. Martinetz and Schulten, 1991) consists of an arbitrary number of neurons. However, in contrast to the SOM, the NG's neurons are not subjected to any topological restrictions, which typically results in a quantitative performance superior to that of the SOM (Cottrell et al., 2006; T. M. Martinetz et al., 1993). Associated with each of the NG's neurons is a prototype vector of the same dimension as the input space. During the training, input vectors are presented to the NG and each neuron is moved in the input vector's direction. The magnitude of the displacement depends on the neurons' ranking order with respect to the distance to the input vector, the learning rate and the neighborhood range. Both the neighborhood range and learning rate are typically set to decrease with training time. After a sufficient number of training steps, the prototype vectors typically approximate the probability density function of the input space with near-minimum quantization error.

In contrast to SOM, NG does not have a predefined topology, which represents the similarity relationships between the neurons (T. Martinetz and Schulten, 1991). However, a topology is particularly useful, because it can reveal valuable information about the underlying data. In order to learn a topology, competitive Hebbian learning

(T. Martinetz and Schulten, 1991; T. Martinetz, 1993) can be applied to NG in a post-processing step as follows: For each input vector, the two closest neurons are identified and a connection between these two neurons is added to the total set of connections, whereas closeness is usually measured with the Euclidean distance. When all input vectors have been processed, the resulting set of connections represents the learned topology. The number of connections that have been added between two neurons indicates the strength of their relationship (Hagenauer, 2014).

7.4. Spatial context models

This article introduces the concept of a spatial context model. A spatial context model describes the relationships between spatial observations and the neurons of a self-organizing neural network during the training or when applying the trained network to data. Because the neurons are constantly moved during the training, their spatial locations are not fixed and, hence, it is difficult to describe their spatial relationships using a spatial weights matrix, a formalization of spatial relationships which is frequently used in spatial statistics (see, e.g., Bavaud, 1998; Getis, 2009). Instead, with the exception of Weighted Merge Context (WMC) (Public, 2015), spatial context models evaluate the spatial relationships between neurons and spatial observations by utilizing different distance measures.

Spatial context models have previously been considered as a integral part of a self-organizing network (see, e.g., Bação et al., 2005; Hagenauer and Helbich, 2013). This article distinguishes between self-organizing neural networks and spatial context models. Such a distinction has several advantages. First, it maintains the modularity of the toolkit. This is desired because it facilitates reuse of existing code, the implementation of new features and its further extension. Second, and more important, it allows the combination of different self-organizing networks with different spatial context models and thus increases the analytical capabilities of the toolkit.

In the following, this section briefly describes the most common spatial context models that can be derived from existing literature.

7.4.1. Augmented input vectors

The simplest context model for considering spatial dependence during the training of a neural network consists in concatenating each input vector with the coordinate vector that represents the location of the corresponding observation. Hence, since the coordinates are treated as regular attributes, this approach can be used with virtually every neural network algorithm. By scaling the coordinates, the relative importance of the coordinates in comparison to the other attributes can be adjusted.

7.4.2. Weighted distance

Another approach that also considers the mutual dependence of the spatial coordinates is to measure the distance between the prototype and the input vectors by calculating the weighted sum of attribute similarity and spatial closeness, both commonly expressed by Euclidean distances but measured according to different scales (e.g., Murray and Shyy, 2000). The weighting of the two addends determines the relative importance of spatial closeness when evaluating the similarity.

7.4.3. GeoSOM

The GeoSOM (Bação et al., 2005) is a variant of the SOM algorithm that adapts the idea of Kangas (1992) for quantizing, clustering and visualizing spatial data. The main difference with the basic SOM is that the GeoSOM uses a two-step procedure to determine the BMU. In the first step, the neuron that is spatially closest to the input vector is identified. In the second step, the closest neuron to the input vector, but within a fixed radius of this neuron in terms of map distance, is identified.[1] This neuron is then designated as the final BMU. The size of the radius affects the strength of spatial dependence that is incorporated into the learning process. The smaller the radius, the more the final ordering of the map is determined by spatial closeness.

7.4.4. Contextual neural gas

Contextual Neural Gas (CNG) (Hagenauer and Helbich, 2013) is a vector quantization and clustering algorithm that combines the concepts of the GeoSOM with the NG algorithm. Analogous to the GeoSOM, CNG enforces spatial proximity between the observations and neurons by exploiting the spatial arrangement of the neurons. However, since its neurons are not topologically ordered in a grid, CNG applies a two-step procedure for determining a rank ordering. In the first step, the neurons are ordered according to spatial closeness. In the second step, the first k neurons of the resulting spatial ordering are reordered within their ranks with respect to input vector similarity. The parameter k controls the degree of spatial dependence that is incorporated in the adaptation process: The smaller the parameter k, the more is the adaptation of neurons determined by spatial closeness.

---

[1] Since the NG's neurons are not arranged in a fixed map and, consequently, the map distance between neurons is not defined for NG, this approach cannot be used with this algorithm.

7.4.5. Weighted merge context

Weighted Merge Context (WMC) (Public, 2015) is a generalization of merge context (Strickert and Hammer, 2003), a method for clustering temporal data. WMC evaluates distance by considering not only the similarity between prototype and input vectors, but also the prototypes of spatially close input vectors, which are represented by context vectors. Both vectors are updated in the course of the training. The prototype vector is moved in the direction of the input vector, whereas the context vector is moved in the direction of the context descriptor, which is a recursively expressed reference to the input vector's neighborhood. The weighting of the context vectors' similarity in the process of learning then basically determines the importance of spatial context information over the current input vectors' similarity.

7.5. The SPAWNN toolkit

The general architecture of the SPAWNN toolkit, depicted in Figure 7.1, is organized in four layers, which roughly correspond to the well-known steps in the process of knowledge discovery in databases (see Fayyad et al., 1996). While Figure 7.1 suggests a sequential execution order, it is not mandatory for application purposes: Steps can be skipped or repeated, depending on the decision of the analyst guiding the process (see Miller and Han, 2009).

- In the input layer, spatial data is first loaded into the application. The general data format is not specified and the data can stem from different sources, e.g., other GIS software or spatial databases. Next, the input data is preprocessed. The preprocessing is a crucial step, because it can significantly affect the results of the analysis.

- In the modeling layer, a (self-organizing) ANN is combined with a spatial context model and trained using the preprocessed input data. There are no general constraints on the kind of network or spatial context model that can be applied. In fact, depending on the problem at hand, multiple combinations can be useful. After training, it is usually desirable to post-process the resulting network data. A typical procedure is to form clusters of neurons in order to summarize their properties.

- The visualization and interaction layer is one of the central components of the toolkit. It provides two different visualizations: a view of the resulting self-organizing neural network and a map view of the spatial data applied to the

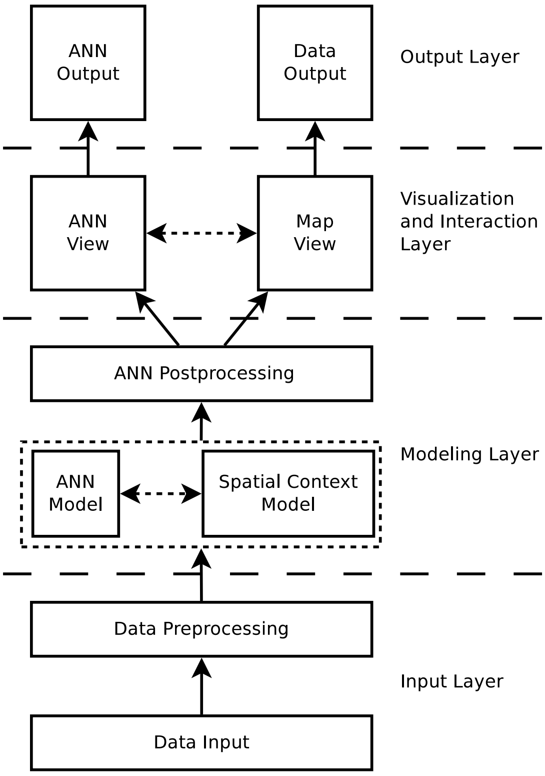


Figure 7.1.: General architecture of the SPAWNN toolkit.

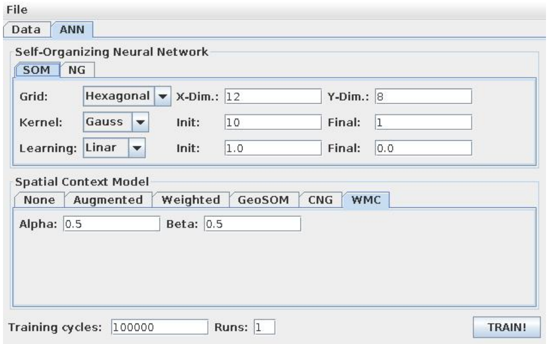
network. Both views are tightly linked to each other. The analyst can interactively select color or group neurons on the network view and immediately see the effects on the observations on the map view and vice versa. This gives the analyst a deeper understanding of the relationship between the network and the spatial data, which in turn facilitates the formulation of hypotheses and helps to gain new insights.

- The output layer is responsible for exporting the results in common data formats. Using these formats, the data can then straightforwardly be imported to other GIS or statistical software for further analysis. This is important because the presented SPAWNN toolkit focuses on self-organizing neural network analysis; it is not intended to replace but to complement existing analytical frameworks.

The actual SPAWNN toolkit is an independent standalone application that is written in Java and distributed as open-source software under the GNU General Public License (GPL). This has several advantages. First, because the toolkit is independent of other software rather than integrated into a GIS or statistical software, the user does not have to deal with different software products and licenses. Second, the implementation in a platform-independent language permits the toolkit to be run on a multitude of different platforms. Third, because the toolkit is developed as free and open-source software, the scientific community has access to the source code, allowing other researchers to modify the toolkit for subsequent integration in their own toolkits or to participate in its further development (see Rey, 2009). The toolkit is available free-of-charge and can be downloaded from the following website: `http://www.spawnn.org`.

Currently, the toolkit supports comma-separated files and ESRI shapefiles for importing spatial data. After importing the data, the analyst can select attributes and normalize them either by scaling them to the zero-to-one range or to have zero mean and unit standard deviation. Furthermore, attributes that represent geographic coordinates can be flagged. The presence of coordinate vectors is necessary for all spatial context models, except WMC, which uses a weight matrix to determine the distance between observations.

The SPAWNN toolkit implements the SOM algorithm as well as the NG algorithm. To account for spatial dependence, these neural networks can be combined with different spatial context models (see Section 7.4). 7.2 shows the graphical user interface for the selection and configuration of the self-organizing network and the spatial context model that should be applied. Once the network is learned, a split view that displays a representation of the trained network (ANN view) as well as of the data (Data view) is



Grid: Hexagonal X-Dim.: 12 Y-Dim.: 8
Kernel: Gauss Init: 10 Final: 1
Learning: Linar Init: 1.0 Final: 0.0
Alpha: 0.5 Beta: 0.5
Training cycles: 100000 Runs: 1

Figure 7.2.: Self-organizing neural network and spatial context model selection window.

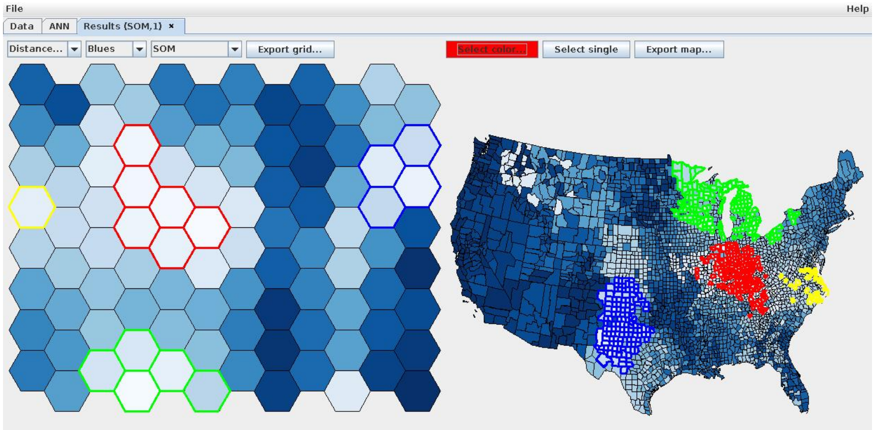


Figure 7.3.: Linked view of a self-organizing neural network (left) and a geographic map of the data (right). The colored neurons of the network correspond to the colored observations on the geographic map.

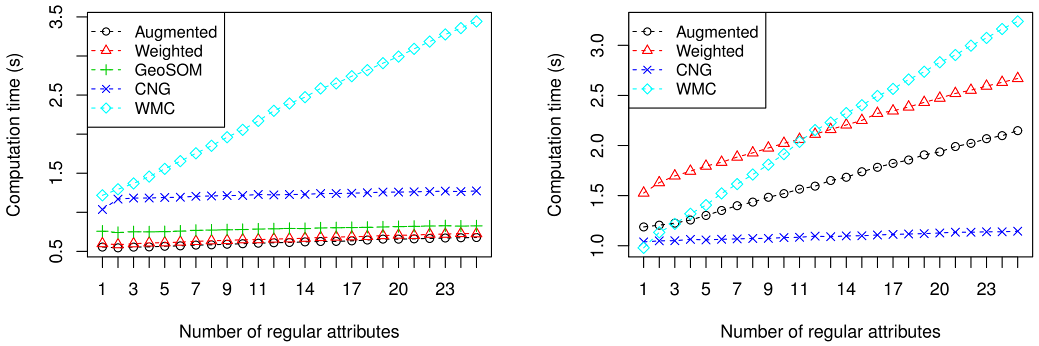
presented to the user 7.3). Utilizing these views, the analyst can explore and highlight either the neurons or observations in order to investigate the mapping in detail.

If the analyst has trained a SOM, a grid view of the neurons or a graph view in which the neurons are arranged according to their geographic coordinates can be displayed. The coloring of the neurons can be chosen to either represent the value of a certain attribute of the neurons' prototype vectors (component plane) or represent a distance-related statistic of each neurons to its immediate neighbors (distance matrix or distance-based representation, see, e.g., Vesanto, 1999). The distance-related statistic can either be the mean, median, mode or some other central or typical values. Moreover, the neurons can also be clustered and colored according to cluster membership. This is useful because the trained networks often consist of far more neurons than there are actual clusters in the data. The toolkit provides several algorithms to detect clusters of neurons, e.g., the basic k-means algorithm, watershed clustering (Vincent and Soille, 1991), a subset of the Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP) algorithms (Guo, 2008) and the Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) algorithm (Assunção et al., 2006). The later three are particularly useful because they can identify spatially contiguous clusters of neurons.

If the trained network is NG, the toolkit displays a graph view of the neurons. The neurons can thereby be arranged either according to their geographic coordinates or by using common graph layout algorithms, including the Fruchterman-Reingold (Fruchterman and Reingold, 1991) or Kamada-Kawai algorithm (Kamada and Kawai, 1989), which both try to automatically arrange the neurons in an aesthetically pleasing and meaningful way. In addition, the neurons in the graph view can be colored according to the same criteria as when displaying a SOM, such as some distance-related statistic of the neurons or the values of the prototype vectors for chosen attributes. The user can choose from multiple color schemes, which include sequential, diverging, and qualitative schemes that have been particularly designed for thematic mapping (Harrower and Brewer, 2003). Moreover, the strength of the connections between neurons can be displayed by varying their line width, depending on either the distance between the connected neurons or the number of times the connected neurons have been closest to each other in the mapping process. Thus, the analyst can explore the similarity relationships between the neurons in detail. In addition to the aforementioned algorithms for clustering neurons, the SPAWNN toolkit also provides powerful community detection algorithms that exploit the NG's network topology for this purpose. Among these algorithms are the Girvan-Newman algorithm (Girvan and Newman, 2002), which evaluates the number of shortest paths between neurons and then progressively removes connections to detect communities, as well as the multi-level modularity optimization (MLMO) algorithm (Blondel et al., 2008), a greedy hierarchical optimization heuristic that is particularly well suited for large networks.

Depending on the results, the analyst can re-run the network training and visualization using different parameters *ad libitum*. The results of previous runs are retained, so that they can be related to the current ones to facilitate the analysis process.

Finally, the toolkit supports the export of the results in different data formats. The trained networks can be saved as, among others, a GraphML (Brandes et al., 2002) file or in the data format of the Java SOMToolkit (Mayer et al., 2011). The spatial data, enriched with the results of the analysis, can be exported as a common ESRI shapefile. Since these data formats are also supported by a wide range of other software products, interoperability is promoted.



(a) SOM

(b) NG

Figure 7.4.: Mean computation times of the spatial context models for different numbers of regular attributes.

7.6. Performance study

Computation time is a concern when clustering large and complex spatial data sets. Therefore, this section investigates the computational demand of the different spatial context models with respect to the number of regular non-spatial and location attributes. For this purpose, synthetic data sets are created, whose observations consist of n normally distributed random regular attributes and m normally distributed location attributes. Each spatial context model is combined with a 6×6 SOM as well as a NG consisting of 36 neurons, except the GeoSOM approach, which is only combined with the SOM, because it requires a fixed map topology.

Then, the resulting networks are trained for $100,000$ steps using the created data sets on a standard laptop PC that is equipped with an Intel® Core ™ i5-3320M CPU@2.4GHz and running on Debian Linux 8.0.

Figure 7.4 shows the mean computation times over 32 runs for $1 \leq n \leq 25$ and $m = 2$, while Figure 7.5 shows these for $n = 5$ and $1 \leq m \leq 25$. In more detail, Figure 7.4a reveals that, except for WMC, the computation time increases only very slightly with the number of regular attributes when training a SOM that is combined with a spatial context model. WMC is much more computationally demanding than the other models, because it considers the regular attributes of all neighbors for each observation in order to find the BMU. When training a NG that is combined with a spatial context model, the computation time increases significantly for all models, except CNG. This is because they consider the regular attributes of all neurons in the sorting procedure, whereas CNG considers only a small subset of neurons.

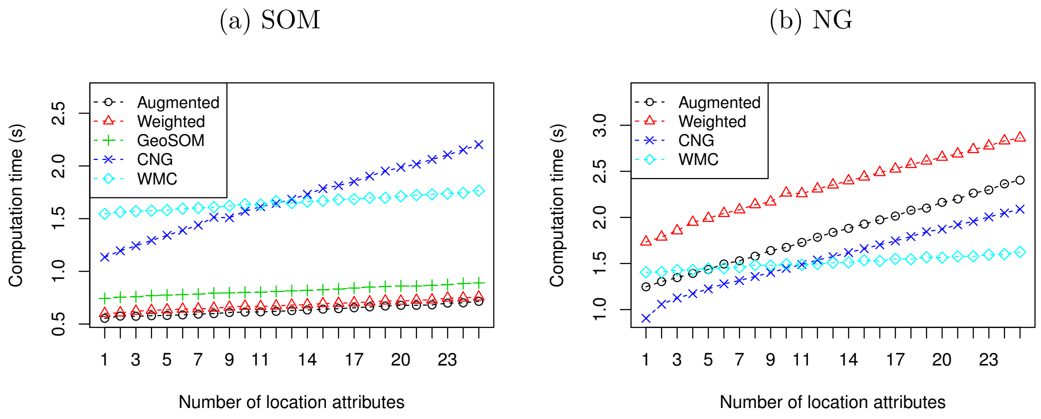

(a) SOM
(b) NG

Figure 7.5.: Mean computation times of the spatial context models for different numbers of location attributes.

Concerning the number of location attributes, Figure 7.5a shows that when training a SOM that is combined with a spatial context model, the CNG is the only model for which the computation time significantly increases with the number of location attributes. This is because, to find the BMU, the CNG approach uses a sorting procedure that frequently evaluates the location attributes of the observations, while the BMU search of the other models does not include such a sorting procedure and evaluates these attributes far less often. By contrast, the only spatial context model for which the computation time does not significantly increase when combined with a NG is WMC. The reason for this is that WMC does not utilize the location attributes of the observations to evaluate spatial relationships, but uses a weight matrix instead. Nevertheless, in real-world applications the performance with regard to the number of location attributes is rarely a concern, because most observations are typically measured in two- or three-dimensional space.

To conclude, the computation times of the spatial context models are kept at perfectly reasonable levels, which gives evidence for the appropriateness of the SPAWNN toolkit for analyzing and clustering high-dimensional spatial data sets.

7.7. Case study

To illustrate the practical applicability of the SPAWNN toolkit and to highlight the advantages of different self-organizing neural networks for different tasks, this section presents a case study which consists of three common analysis steps: an outlier analysis, a correlation analysis and a cluster analysis. The case study uses socioeconomic data of the city of Philadelphia, Pennsylvania. The city is situated in the northeastern US along the Delaware and Schuylkill rivers and consists of an area of approximately 369 square kilometers. Philadelphia is currently the fifth largest city in the US, with an estimated population of 1.5 million people in 2012, and is the economic and cultural center of the Delaware Valley. The city is of particular interest because it is one of the most segregated cities in the US while having spatially varying socio-demographics; even the most affluent African Americans live in neighborhoods that are close to majority African American (Logan, 2011). Hence, it can be expected that these neighborhoods emerge as distinct clusters in the analysis results.

The case study uses freely available tract-level data extracted from the 2010 US Census about ethnicity, age, housing, and households in Philadelphia. While census tracts are mostly homogeneous with respect to population characteristics, economic status, and living conditions, there exist census tracts, in particular in South Philadelphia, which exhibit pronounced ethnic heterogeneity below the census tract level. These tracts do not affect the applicability of the toolkit, but they must be considered when interpreting the analysis results. The following nine variables are used: (1) percentage of white population; (2) percentage of African Americans; (3) percentage of Asians; (4) percentage of Hispanics; (5) percentage of renter-occupied houses; (6) percentage of population younger than 25 years old; (7) percentage of population between 25 and 64 years old; (8) percentage of population older than 64 years; and (9) the average size of households. Tracts without population are removed from the data set beforehand, and all attributes are standardized to zero mean and unit variance to make them comparable. The study site consists of 380 census tracts in total. While the SPAWNN toolkit can generally applied to data sets of arbitrary size, the small number of census tracts in this case study facilitates the visualization and discussion of the results.

7.7.1. Outlier analysis

First, a GeoSOM and CNG are applied for outlier detection. The identification of outliers is a crucial task, because outliers distort the distribution of the data and thus can significantly affect the results of subsequent analysis. The GeoSOM consists of 12×8 neurons and the CNG accordingly of 96 neurons. Preliminary tests have shown that these numbers represent a fair compromise between computational effort for training the networks and quantization performance. Both networks are trained for $100,000$ iterations.

In the distance matrix representation of the resulting GeoSOM (Figure 7.6), outliers can be identified by neurons that have high median distance to neighboring values

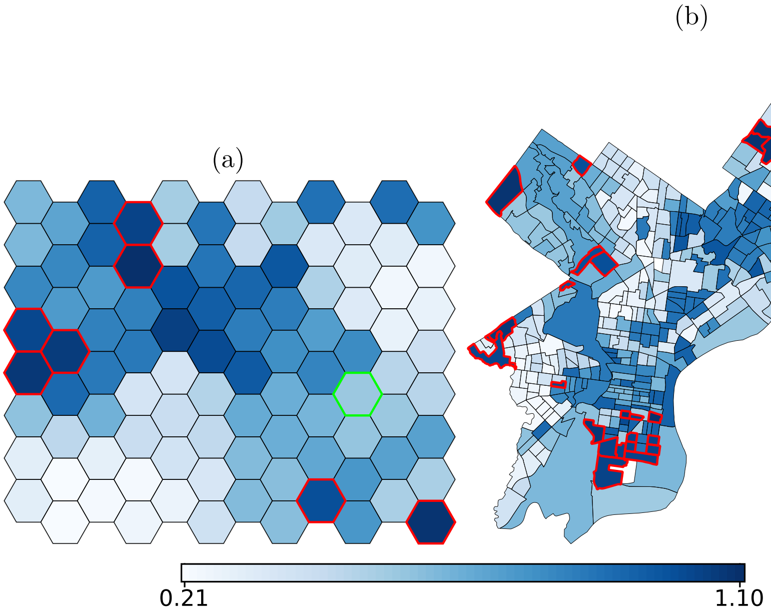


Figure 7.6.: Distance matrix (a) and cartographic map (b) of the GeoSOM. Identified outliers are outlined in red. The neuron that maps the census tract where the Northeast Philadelphia Airport is located and the tract itself are outlined in green. For clarity, other tracts that this particular neuron maps are not outlined.

and, given that the size of the map is sufficiently large, are located at the border of the matrix (Muñoz and Muruzábal, 1998). In the distance-based representation of the resulting CNG (Figure 7.7), outliers can be identified by also having a high median distance to neighboring neurons while being sparsely connected to other neurons.

Comparing the identified outliers (outlined in red) in both representations shows that while they partly correspond, there also exist some notable differences. A reason for these differences might be that selecting a median distance threshold is a rather subjective task. This matter is typically less crucial for the CNG, because the learned topology provides additional and more useful guidance for the identification of outliers than the a priori fixed topology of the GeoSOM. Indeed, in-depth inspection of the identified outliers revealed that the results for the CNG are more consistent than those for the GeoSOM. For instance, the census tract where the Northeast Philadelphia Airport is located (outlined in green in Figure 7.6) has a extremely low population density and its attributes are consequently very skewed (e.g., a 100% rate of white population). This tract is barely recognizable as an outlier on the distance matrix of the GeoSOM because the mapping neuron's median distance to neighboring neurons is

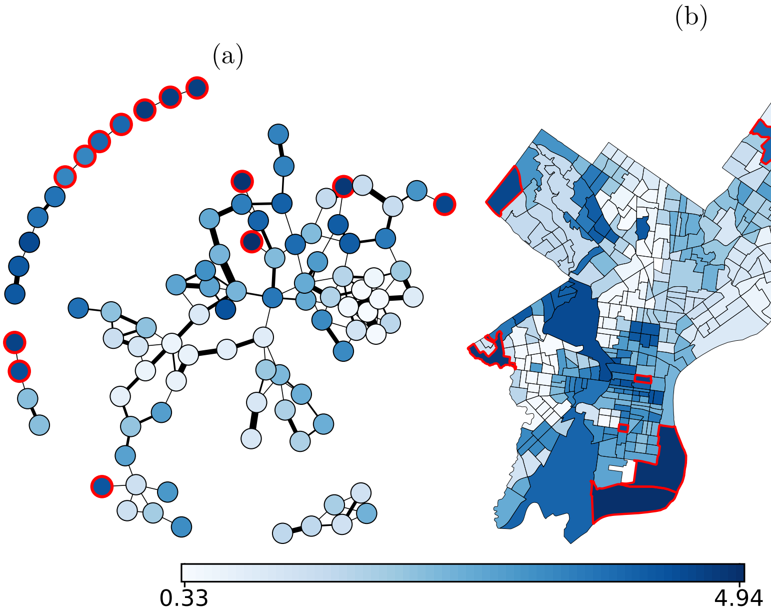


Figure 7.7.: Distance-based representation of the CNG (a) and the geographic map (b). Identified outliers are outlined in red.

rather small (the neuron is also colored in green in Figure 7.6). In contrast, the census tract is clearly identifiable as an outlier on the CNG representation (see Figure 7.7). Hence, it can be concluded from this section that the CNG is more appropriate for identifying outliers than the GeoSOM.

7.7.2. Correlation analysis

As a second analysis step, correlation analysis is performed which identifies and evaluates the associations between different attributes of the data. For this purpose, the identified outliers are removed from the data set first. Then, a GeoSOM and a CNG of the same sizes as in the preceding section are trained.

A common approach for identifying correlations in the data is to compare component planes (e.g., Barreto S. and Pérez-Uribe, 2007; Vesanto and Ahola, 1999). Correlations become apparent by similar (positive correlation) or complementary (negative correlation) patterns in identical areas of the network.

This approach for identifying correlations has several advantages over standard correlation analyses. First, the SOM as well as the NG provide a nonlinear map of the data which allows the identification of nonlinear correlations. Second, by comparing multiple component planes multivariate correlations become apparent. Third, local correlations can be identified by partially matching patterns.

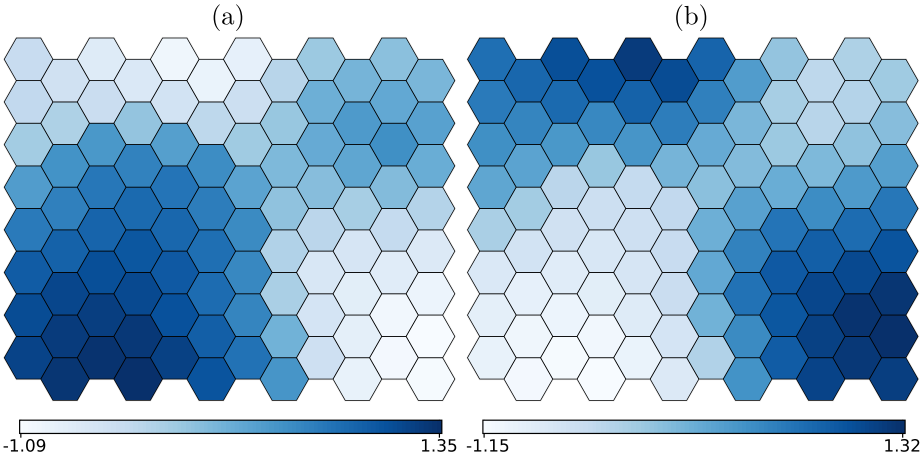


Figure 7.8.: GeoSOM component planes for the rates of African Americans (a) and white population (b).

Figure 7.8 exemplarily depicts the GeoSOM component planes for the rates of African Americans and white population. The component planes reveal rather complementary patterns, indicating a strong negative correlation and therefore high segregation between the African American and white populations of the city. Furthermore, it can be seen that two distant areas of the network both have very high rates of white population, indicating that other discriminating factors determined the separation of these areas.

Figure 7.9 shows the neurons of the CNG, which are also colored according to the percentage of African Americans (a) and the percentage of white population b). Even though the complementary patterns are also present, these patterns are more difficult to perceive due to the seemingly unordered arrangement of the neurons. In fact, it is hardly feasible to arrange the CNG's neurons on a two-dimensional plane while preserving the neurons' topological relationships. This problem typically becomes even more severe as the dimension of the input space increases. To conclude, the GeoSOM is more appropriate for correlation analysis than the CNG.

7.7.3. Cluster analysis

As a last step showing the use of the SPAWNN toolkit, the trained GeoSOM and CNG from the preceding section are used to detect spatially contiguous clusters within the study area. For this purpose, the SPAWNN toolkit provides several powerful clustering algorithms as well as means for manually outlining and visualizing clusters. Here, clustering algorithms are used because they depend less on the subjective decisions of an analyst and are more convenient for complex networks.

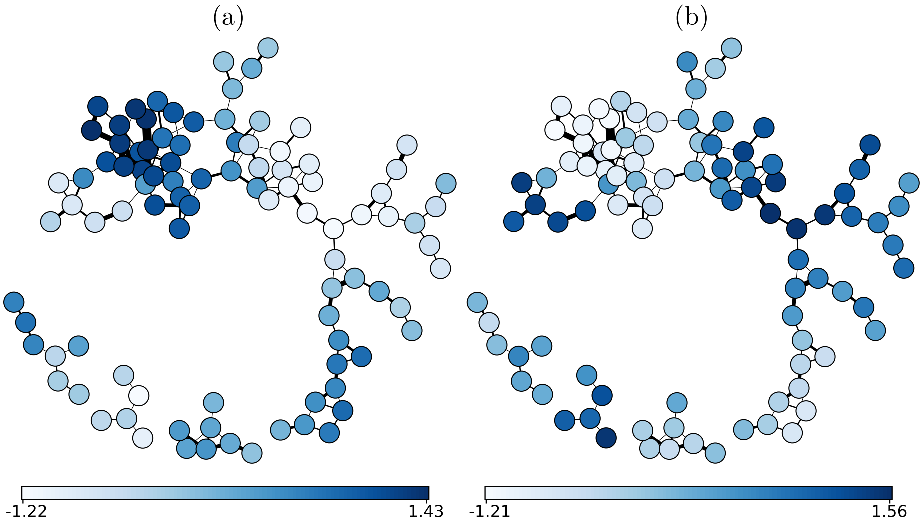


Figure 7.9.: Neurons of the CNG, colored according to the rates of African Americans (a) and white population (b).

In order to cluster the CNG and GeoSOM, contiguity-constrained hierarchical clustering using Ward's criterion is applied (Murtagh, 1995). Figure 7.10 depicts the results for the clustering of the CNG, while the results for the GeoSOM are shown in Figure 7.11.

Both algorithms detected similar clusters, even though there are also some notable differences. For example, both algorithms outlined a separate cluster (cluster 3) that captures the neighborhoods of the city's universities, e.g., Drexel, Temple, and Saint Joseph's University. However, cluster 3 of the GeoSOM does not include La Salle University in the North.

Cluster 1, which is predominantly characterized by high rates of African American population, is also very similar for the clustering of the CNG and GeoSOM. In fact, comparing the outline of this particular cluster in the network representation of the GeoSOM (Figure 7.11, left panel) with the component plane of African American Population in Figure 7.8a, reveals that the clusterings essentially follow the distribution of African American population. This exemplifies that high segregation tendencies of African Americans are still present in the city.

By contrast, the few existing predominantly Asian neighborhoods do not appear in the clustering of the GeoSOM, even though they are clearly outlined by the CNG (cluster 4). Analogously, while the CNG outlines neighborhoods with very high rates of senior citizens (cluster 8), these neighborhoods are not demarcated by the GeoSOM. Also neighborhoods with high rates of Hispanic population, in particular around Fairhill,

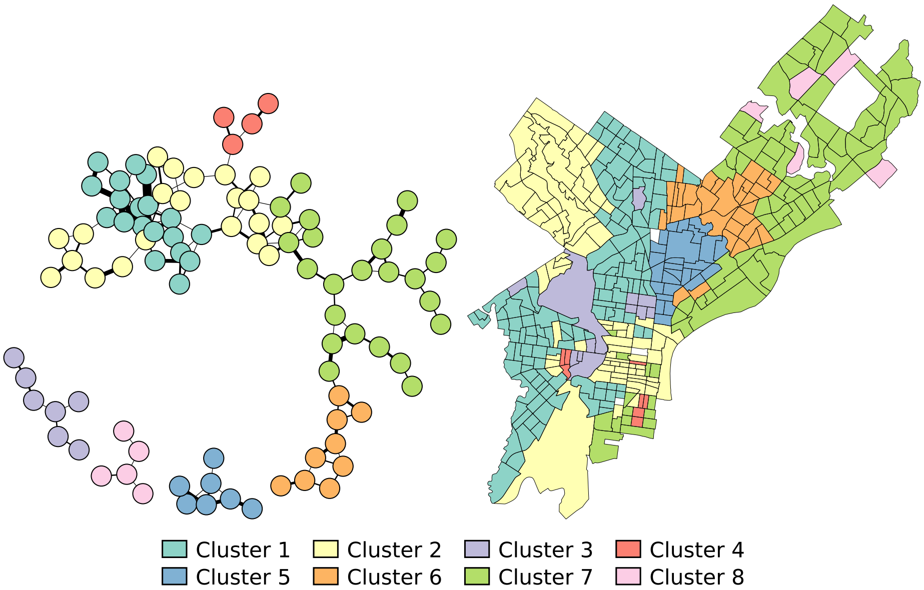


Figure 7.10.: Clustering results for CNG.

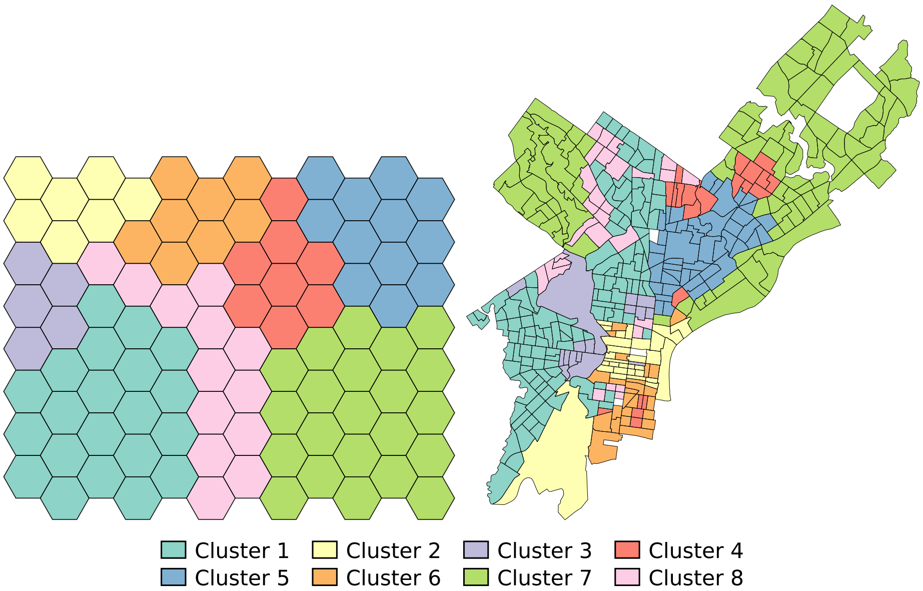


Figure 7.11.: Clustering results for the GeoSOM.

which serves as the center of the Hispanic community in Philadelphia, are more clearly identified by the CNG than by the GeoSOM (cluster 5).

In addition, the figures show that cluster 7 of the GeoSOM covers two distinct parts of the city in the northeast and northwest, while for the CNG these parts are covered by two spatially disjoint clusters (clusters 2 and 7). From a geographical perspective, a distinction between the northeast and northwest parts of Philadelphia can indeed be expected, because in the past the northeast has undergone very different economic and sociological developments from the remainder of the city (see, e.g., Adams, 1993).

In conclusion, while the GeoSOM is particularly useful for relating clusters to component planes in order to inspect data relationships, the clustering of the CNG is geographically more accurate.

7.8. Conclusion

This article presented the SPAWNN toolkit, a new and powerful exploratory toolkit for spatial analysis and clustering, which is not embedded in a standard GIS. The toolkit is innovative in several ways. It distinguishes between different self-organizing neural networks and spatial context models and provides implementations of different kinds of both. In this way, the analyst can combine different self-organizing networks with different spatial context models and modularity is maintained. In addition, the toolkit provides powerful clustering methods for post-processing the networks. Moreover, it provides linkage between the different network and data visualizations, which allows strong interaction between the analyst, the data and the trained networks, and thus helps to improve understanding of the data. Apart from these contributions, the toolkit has been developed with the objective of enabling non-expert users without programming skills to use cutting-edge clustering methods. In these respects, the SPAWNN toolkit complements existing toolkits and makes a significant contribution.

How an analyst can take advantage of the distinguishing features of the toolkit's different self-organizing networks, spatial context models, and visualizations was demonstrated through a case study analyzing socioeconomic census data of the city of Philadelphia. In particular, it has been shown how the complementary advantages of CNG and GeoSOM can be used to get a better understanding of the data. The results underscore the fact that Philadelphia is faced with segregation across the cityscape. The spatial analysis capabilities of the toolkit are not restricted to geography, but are also relevant to a variety of other domains, including crime (e.g., Hagenauer et al., 2011; Helbich et al., 2013b), health (e.g., Augustijn and Zurita-Milla, 2013), real estate (e.g., Helbich et al., 2013a), and ecology (e.g., Stojkovic et al., 2013), among others. Moreover, it has been shown that the computation times of the SPAWNN toolkit are kept at perfectly reasonable levels, even for high-dimensional spatial data sets.

Future research will focus on how the toolkit can be extended to further increase its usefulness for spatial analysis. Currently, the toolkit provides linkage between a single network model and a single geographic map (one-to-one linkage). One way to extend the toolkit is to provide linkage between multiple network models and multiple geographic maps (n-to-n linkage). In this way, the analyst could train several networks (i.e., non-spatial vs. contextual networks, or investigate the impact of different parameter settings on the output) for different parts of a study area and then interactively inspect via n-to-n linkage how the networks relate to the different parts.

Bibliography

Adams, C. (1993). *Philadelphia: Neighborhoods, division, and conflict in a postindustrial city.* Temple University Press.

Andrienko, G., Andrienko, N., Bremm, S., Schreck, T., Landesberger, T. v., Bak, P., and Keim, D. (2010). Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum*, 29, 913–922.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, 27(2), 93–115.

Anselin, L., Syabri, I., and Kho, Y. (2006). GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, 38(1), 5–22.

Assunção, R. M., Neves, M. C., Câmara, G., and Costa Freitas, C. da (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7), 797–811.

Augustijn, E.-W. and Zurita-Milla, R. (2013). Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. *International Journal of Health Geographics*, 12(1), 60.

Bação, F., Lobo, V., and Painho, M. (2005). The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computational Geosciences*, 31 (2), 155–163.

Barreto S., M. A. and Pérez-Uribe, A. (2007). Improving the Correlation Hunting in a large quantity of SOM component planes. In: J. M. de Sá, L. Alexandre, W. Duch, and D. Mandic (eds.). *Artificial Neural Networks – ICANN 2007.* Vol. 4669. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 379–388.

Bavaud, F. (1998). Models for spatial weights: A systematic look. *Geographical Analysis*, 30(2), 153–171.

Becker, S. (1991). Unsupervised learning procedures for neural networks. *International Journal of Neural Systems*, 2(1 & 2), 17–33.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.

Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., and Marshall, M. S. (2002). GraphML progress report structural layer proposal. In: *Graph Drawing*. Springer, 501–512.

Carpenter, G. A. and Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks*. MIT Press.

Chon, T.-S. (2011). Self-organizing maps applied to ecological sciences. *Ecological Informatics*, 6(1), 50–61.

Cottrell, M., Hammer, B., Hasenfuß, A., and Villmann, T. (2006). Batch and median neural gas. *Neural Networks*, 19(6), 762–771.

Estévez, P. A. and Figueroa, C. J. (2006). Online data visualization using the neural gas network. *Neural Networks*, 19(6), 923–934.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In: U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press, 1–34.

Fischer, M. M. (1998). Computational neural networks: A new paradigm for spatial analysis. *Environment and Planning A*, 30(10), 1873–1891.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. *Intelligent Data Analysis 5*, 5, 373–384.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11), 1129–1164.

Getis, A. (2009). Spatial weights matrices. *Geographical Analysis*, 41(4), 404–410.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821–7826.

Goodchild, M. F. (1986). *Spatial autocorrelation*. CATMOG. Norwich, United Kingdom: Geo Books.

Grubesic, T. H., Wei, R., and Murray, A. T. (2014). Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers*, 104(6), 1134–1156.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22(7), 801–823.

Guo, D., Chen, J., MacEachren, A. M., and Liao, K. (2006). A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, 12(6), 1461–1474.

Guo, D., Gahegan, M., MacEachren, A. M., and Zhou, B. (2005). Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science*, 32(2), 113–132.

Hagenauer, J. (2014). Clustering contextual neural gas: A new approach for spatial planning and analysis tasks. In: M. Helbich, J. Jokar Arsanjani, and M. Leitner (eds.). *Computational Approaches for Urban Environments*. Springer.

Hagenauer, J. and Helbich, M. (2013). Contextual neural gas for spatial clustering and analysis. *International Journal of Geographical Information Science*, 27(2), 251–266.

Hagenauer, J., Helbich, M., and Leitner, M. (2011). Visualization of crime trajectories with self-organizing maps: A case study on evaluating the impact of hurricanes on spatio-temporal crime hotspots. In: *Proceedings of the 25th International Cartographic Conference*. Paris, France: International Cartographic Association.

Han, J., Koperski, K., and Stefanovic, N. (1997). GeoMiner: a system prototype for spatial data mining. In: *ACM SIGMOD Record*. Vol. 26. 2. ACM, 553–556.

Han, J., Fu, Y., Chiang, W. W. J., Gong, W., Koperski, K., Li, D., Lu, Y., Rajan, A., Xia, N. S. B., and Zaiane, O. R. (1996). DBMiner: A system for mining knowledge in large relational databases. In: *Proc. Intl. Conf. on Data Mining and Knowledge Discovery (KDD '96)*, 250–255.

Harrower, M. and Brewer, C. A. (2003). Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1), 27–37.

Helbich, M., Brunauer, W., Hagenauer, J., and Leitner, M. (2013a). Data-driven regionalization of housing markets. *Annals of the Association of American Geographers*, 103(4), 871–889.

Helbich, M., Hagenauer, J., Leitner, M., and Edwards, R. (2013b). Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach. *Cartography and Geographic Information Science*, 40(4), 326–336.

Henriques, R., Bacao, F., and Lobo, V. (2012). Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems*, 36(3), 218–232.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.

Kalteh, A. M., Hjorth, P., and Berndtsson, R. (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. *Environmental Modelling & Software*, 23(7), 835–845.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1), 7–15.

Kangas, J. (1992). Temporal knowledge in locations of activations in a self-organizing map. In: *Artificial Neural Networks, 2*. Ed. by I. Aleksander and J. Taylor. Vol. 1. Amsterdam, Netherlands: North-Holland, 117–120.

Kaski, S., Kangas, J., and Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, 1, 102–350.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 1–8.

Kohonen, T. (2001). *Self-organizing maps*. New York, NY: Springer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43 (1), 59–69.

Körting, T. S., Garcia Fonseca, L. M., and Câmara, G. (2013). GeoDMA — Geographic data mining analyst. *Computers & Geosciences*, 57, 133–145.

Lacayo-Emery, M. A. (2011). *SOM Analyst*. `https://code.google.com/p/somanalyst/` (Accessed on 5 november 2014).

Liu, Y. and Weisberg, R. H. (2011). *A review of self-organizing map applications in meteorology and oceanography*. INTECH Open Access Publisher.

Logan, J. R. (2011). *Separate and unequal: The neighborhood gap for Blacks, Hispanics and Asians in metropolitan America*. `http://www.s4.brown.edu/us2010/Data/Report/report0727.pdf` (Accessed on 28 march 2013).

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D., and Masters, R. (1999). Constructing knowledge from multivariate spatiotemporal data: Integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, 13(4), 311–334.

Martinetz, T. and Schulten, K. (1991). A "Neural-Gas" network learns topologies. *Artificial Neural Networks*, 1, 397–402.

Martinetz, T. M., Berkovich, S. G., and Schulten, K. J. (1993). "Neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4), 558–569.

Martinetz, T. (1993). Competitive Hebbian learning rule forms perfectly topology preserving maps. In: S. Gielen and B. Kappen (eds.). *ICANN '93*. Springer London, 427–434.

Mayer, R., Dittenbach, M., Frank, J., Neumayer, R., and Lidy, T. (2011). *Data mining with the Java SOMToolbox*. `http://www.ifs.tuwien.ac.at/dm/somtoolbox/` (Accessed on 5 november 2014).

Mennis, J. and Guo, D. (2009). Spatial data mining and geographic knowledge discovery — An introduction. *Computers, Environment and Urban Systems*, 33(6), 403–408.

Miller, H. J. and Goodchild, M. F. (2014). Data-driven geography. *GeoJournal*, 1–13.

Miller, H. J. and Han, J. (2009). *Geographic data mining and knowledge discovery*. Boca Raton, FL: CRC Press, 486pp.

Muñoz, A. and Muruzábal, J. (1998). Self-organizing maps for outlier detection. *Neurocomputing*, 18(1), 33–60.

Murray, A. T. and Shyy, T.-K. (2000). Integrating attribute and space characteristics in choropleth display and spatial data mining. *International Journal of Geographical Information Science*, 14(7), 649–667.

Murtagh, F. (1995). Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Letters*, 16(4), 399–408.

Oja, M., Kaski, S., and Kohonen, T. (2003). Bibliography of self-organizing map (SOM) papers: 1998–2001 addendum. *Neural Computing Surveys*, 3, 1–156.

Openshaw, S. (1999). Geographical data mining: Key design issues. In: *4th International Conference on Geocomputation*. Mary Washington College. Fredericksburg, VA: GeoComputation CD-ROM.

Parimala, M., Lopez, D., and Senthilkumar, N. C. (2011). A survey on density based clustering algorithms for mining large spatial databases. *International Journal of Advanced Science and Technology*, 31(1).

Public, J. Q. (2015). Weighted merge context for clustering spatial data with self-organizing neural networks. (under review).

Rey, S. J. (2009). Show me the code: spatial analysis and open source. *Journal of Geographical Systems*, 11(2), 191–207.

Skupin, A. and Agarwal, P. (2008). Introduction: What is a self-organizing map? In: P. Agarwal and A. Skupin (eds.). *Self-organising maps: Applications in geographic information science*. Wiley Online Library, 1–20.

Stojkovic, M., Simic, V., Milosevic, D., Mancev, D., and Penczak, T. (2013). Visualization of fish community distribution patterns using the self-organizing map: A case study of the Great Morava river system (Serbia). *Ecological Modelling*, 248, 20–29.

Strickert, M. and Hammer, B. (2003). Neural gas for sequences. In: *Proceedings of the Workshop on Self-Organizing Networks (WSOM)*. Kyushu Institute of Technology, 53–57.

Strickert, M. and Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing*, 64(0), 39–71.

Sui, D. Z. (2004). Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers*, 94(2), 269–277.

Takatsuka, M. and Gahegan, M. (2002). GeoVISTA Studio: A codeless visual programming environment for geoscientific data analysis and visualization. *Computers & Geosciences*, 28(10), 1131–1144.

Tasdemir, K. and Merényi, E. (2009). Exploiting data topology in visualization and clustering of self-organizing maps. *IEEE Transactions on Neural Networks*, 20(4), 549–562.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 3, 111–126.

Vesanto, J. and Ahola, J. (1999). Hunting for correlations in data using the self-organizing map. In: *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA '99)*, 279–285.

Vincent, L. and Soille, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, 583–598.

Ware, C. (2012). *Information visualization: perception for design*. Elsevier.

Yuan, M., Buttenfield, B., Gahegan, M., and Miller, H. (2004). Geospatial data mining and knowledge discovery. In: R. B. McMaster and E. L. Usery (eds.). *A Research Agenda for Geographic Information Science*. Boca Raton, FL: CRC Press.

8. A comparative study of machine learning classifiers for modeling travel mode choice

Authors

Julian Hagenauer and Marco Helbich

Journal

Expert systems With applications

Status

Received 12 October 2016; revised 14 January 2017; accepted 30 January 2017; published online 13 February 2017

Contribution statement

Julian Hagenauer has developed the methods, designed the experiments, and has written the manuscript for the study. Marco Helbich supported this publication by continuously discussing the design and results of the study and by proofreading the manuscript.

Abstract

The analysis of travel mode choice is an important task in transportation planning and policy making in order to understand and predict travel demands. While advances in machine learning have led to numerous powerful classifiers, their usefulness for modeling travel mode choice remains largely unexplored. Using extensive Dutch travel diary data from the years 2010 to 2012, enriched with variables on the built and natural environment as well as on weather conditions, this study compares the predictive performance of seven selected machine learning classifiers for travel mode choice analysis and makes recommendations for model selection. In addition, it addresses the importance of different variables and how they relate to different travel modes. The results show that random forest performs significantly better than any other of the investigated classifiers, including the commonly used multinomial logit model. While trip distance is found to be the most important variable, the importance of the other variables varies with classifiers and travel modes. The importance of the meteorological variables is highest for support vector machine, while temperature is particularly important for predicting bicycle and public transport trips. The results suggest that the analysis of variable importance with respect to the different classifiers and travel modes is essential for a better understanding and effective modeling of people's travel behavior.

Travel mode choice; classification; machine learning; the Netherlands

8.1. Introduction

The accurate modeling of travel mode choice is important for transportation planning and policy makers to predict travel demand and understand the underlying factors (see Dios Ortúzar and Willumsen, 2011). In fact, a large body of literature shows that travel mode choice is affected by a variety of factors including individual and household characteristics (e.g., Böcker et al., 2016; Dieleman et al., 2002; Schwanen and Mokhtarian, 2005) as well as the built environment (e.g., Ewing and Cervero, 2010; Helbich, 2016) and weather conditions (e.g., Böcker et al., 2013).

Models of travel mode choice have traditionally been estimated using the discrete choice framework, where travel modes represent mutually exclusive and collectively exhaustive alternatives (Ben-Akiva and Lerman, 1985). The most widely used discrete choice model is the multinomial logit (MNL) model (McFadden, 1973). It is based on the principles of utility maximization and has a mathematic structure which eases parameter estimation (Koppelman and Wen, 1998). For this reasons it has been widely adopted in transportation research (see e.g., Böcker et al., 2016; Ewing et al., 2004).

A limitation of MNL models is that they assume that the probabilities of each pair of alternatives are independent of the presence or characteristics of all other alternatives (McFadden, 1973). Consequently, the introduction of any alternative has the same proportional impact on the probability of each other alternative. Violation of this assumption yields inconsistent parameter estimates and biased predictions (McFadden, 1973). Other discrete choice models, such as the multinomial probit model (MNP), do not make this independence assumption, but parameter estimation is more difficult than for the MNL model, which hampers their usefulness (Dow and Endersby, 2004).

Methods from the field of machine learning are a promising alternative to statistical approaches for modeling travel mode choice. Instead of making strict assumptions about the data, machine learning models learn to represent complex relationships in a data-driven manner (e.g., Bishop, 2006). The usefulness of machine learning models has already been demonstrated for different areas in transportation research. For example, machine learning models are particularly useful for classifying travel modes and inferring trip purposes from global position system and acceleration data (e.g., Gong et al., 2014; Shafique and Hato, 2015; Shen and Stopher, 2014). Other examples include the prediction of railway passenger demand (e.g., Tsai et al., 2009) and bimodal modeling of freight transportation (e.g., Tortum et al., 2009). However, machine learning is still under represented in research of travel mode choice modeling. Existing studies are limited to a small number of machine learning methods and do not provide comprehensive model comparisons.

Cantarella and De Luca (2003), for example, trained two artificial neural networks (ANNs) with different architectures to model people's travel mode choice behavior. They found that both ANNs clearly outperform a MNL model. Celikoglu (2006) showed that ANNs are effective for calibrating the utility function in travel choice modeling. Zhao et al. (2010) demonstrated that the accuracy of probabilistic ANNs is similar to basic ANNs for travel mode choice prediction, whereas Omrani et al. (2013) showed that ANNs are more accurate than the other investigated alternatives. A few studies report less promising results for ANNs in comparison to traditional models. Hensher and Ton (2000), for instance, compared the predictive capabilities of ANNs and nested logit models in the context of commuter mode choice and found no performance advantage for ANNs. Similarly, Sayed and Razavi (2000) reported that the classification performance of fuzzy ANNs, MNL, and MNP models is similar.

Classification trees (CTs) have also been applied for travel mode choice analysis. C. Xie et al. (2003), for instance, compared CTs and ANNs with MNL models. They conclude that CTs and ANNs perform better than MNL. Moreover, they state that

CTs are more efficient and provide better interpretability than ANNs. Rasouli and Timmermans (2014) investigated the relationship between predictive performance and the number of CTs when using ensemble learning. They showed that the accuracy increases nonmonotonically with the size of the ensemble. Hierarchical treebased regression is used by Zhan et al. (2016) to investigate the travel characteristics of Chinese students and to determine variables that affect students' travel behavior. Tang et al. (2015) used CTs to explore travel mode choice for the case where the choice is restricted to two modes in order to investigate people's mode-switching behavior. They confirmed the superior predictive capability of a CT to an MNL model.

Support-vector machines (SVMs) have also been applied in numerous studies. For example, when Zhang and Y. Xie (2008) compared SVM, ANNs, and MNL for modeling travel mode choice, they found that SVM provided the highest accuracy. By contrast, Omrani (2015) showed that ANNs are more accurate than SVMs and MNL models for modeling the travel mode choice behavior of commuters. Xian-Yu (2011) reported that the performance of SVM is superior to ANN and nested logit models.

While the aforementioned studies represent important contributions to the application of machine learning in transportation research, they also have some major limitations. First, these studies deal only with a limited set of machine learning classifiers, even though the number of available classifiers is large (see Fernández-Delgado et al., 2014). Advanced classifiers such as random forests or ensemble learners have not been considered in a comparative study, even though it has been shown that these classifiers can produce highly accurate results for many applications (see e.g., Fernández-Delgado et al., 2014). Second, model comparisons are not done in a systematically quantitative way using statistical test procedures which take the sampling variability into account (see Hothorn et al., 2005). Third, previous studies do not consider characteristics of the built and natural environment and meteorological conditions, even though these factors substantially influence travel behavior (see e.g., Helbich et al., 2014; Liu et al., 2015). Finally, these studies do not thoroughly investigate the importance of variables, particularly with regard to the different models and travel modes, even though such an analysis supports the interpretation and understanding of the results (e.g., Murray and Conner, 2009).

This study addresses the identified shortcomings and adds to the literature as follows: First, it presents a comprehensive comparison of seven machine learning classifiers. Second, the article systematically evaluates the classifiers using strict model validation techniques and test statistics. Third, in addition to individual and household characteristics, it considers characteristics of the built and natural environmental as well as meteorological conditions for model building. Finally, the article investigates the importance of each variable for each classifier and travel mode in detail.

The rest of this article is structured as follows: Section 8.2 outlines the data and methods used. Section 8.3 describes the results, followed by a discussion in Section 8.4. Finally, Section 8.5 closes the paper with concluding remarks.

8.2. Materials and methods

8.2.1. Data

The primary data source for this study is the Dutch national travel survey (NTS) conducted from 2010 to 2012. It is supplied by Onderzoek Verplaatsingen in Nederland (2014) and is based on individual travel diaries. The survey participants were asked to record every trip over the course of six days, which have been randomly selected to cover a whole year in order to account for seasonal effects. To compensate for the lower response rates of nonnatives and older participants, both groups were oversampled. In addition to trip-specific data (e.g., travel mode and trip distance), the NTS also provides socio-economic data about the participants (e.g., gender, age, and ethnicity) as well as information on households (e.g., income, number of cars and bicycles). The present study considers only records of participants aged 18 and over to exclude the distinct travel behavior of younger people. Furthermore, records that contain incomplete or erroneous information are also excluded. The resulting sampled data set consists of 69, 918 individuals and a total of 230, 608 trips. These trips are spatially distributed across all regions of the Netherlands and represent the travel behavior of the Dutch population as a whole. The NTS data can be accessed free of charge from DANS (Data Archiving and Networked Services) through the following link: `https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:54132/tab/1`.

The data is additionally enriched with environmental data (see Fishman et al., 2015). For this purpose, the residential locations of the participants are geocoded by postal codes using the nationwide cadastre database (Basisregistraties Adressen en Gebouwen). The locations are then utilized to derive variables that characterize the built and natural environments and weather conditions. The resulting data set consists of 17 variables, described in Table 8.1. The proportion of green space and the land use diversity are calculated using the Dutch land use model (Landelikj Grandgebruiksbestand Nederland 20 08/20 09). The meteorological variables are derived from the daily reports of the nearest weather station of the Royal Dutch Meteorological Institute. There are 36 such stations. Descriptive statistics of each variable are given in Table 8.2. The data set used for this model competition is provided through the journal's website.

Table 8.1.: Description of the variables.

Variable	Description
Trip	
distance	Total trip distance in km
weekend	Trip is done at the weekend
mode	Main travel mode (walk, bike, pt, car). pt refers to public transport.
Individual	
age	Age of participant in years
education	Education of participant (lower, middle, higher)
ethnicity	Ethnicity of participant (native, western, other)
license	Participant owns a driver's license (yes, no)
male	Male participant (yes, no)
Household	
bicycles	Number of bicycles per household
cars	Number of cars per household
income	Net annual household income in $1,000$€ (<20,≥ 20–40,≥ 40)
Built environment	
density	Address density, aggregated over post codes, in $1,000$ addresses per km^2
diversity	Shannon diversity index of land use classes
green	Proportion of green space per post code area in %
Weather	
precip	Daily precipitation sum in mm
temp	Daily maximum temperature in °C
wind	Daily average wind speed in m/s

Table 8.2.: Descriptive statistics of the variables of the trip data set.

Variable	Category	%	Min.	Max.	Mean	Std. Dev.
Trip						
distance			0.100	400.000	12.218	23.546
weekend	no	82.066				
mode	walk	20.935				
	bike	24.473				
	pt	2.316				
	car	52.276				
Individual						
age			18.000	98.000	47.661	15.935
education	low	27.370				
	middle	38.293				
	high	34.337				
ethnicity	native Dutch	87.404				
	western	7.707				
	other	4.889				
license	no	10.243				
male	no	54.498				
Household						
bicycles			0.000	10.000	3.357	1.937
cars			0.000	10.000	1.383	0.822
income	<20	11.832				
	≥20–40	42.123				
	≥40	46.044				
Build and natural environment						
density			0.002	11.443	1.569	1.593
diversity			0.000	2.828	1.775	0.493
green			0.000	97.813	54.939	22.172
Weather						
precip			0.000	142.300	2.185	4.675
temp			−9.000	35.900	13.317	7.566
wind			0.400	16.300	4.098	1.915

8.2.2. Classifiers

This article compares seven machine learning methods to classify travel mode choice. These methods have either already been successfully used in transportation research or have shown promising results in other fields (see e.g., Xu et al., 2014). The parameters of each classifier are determined by systematically testing values from a manually specified subspace (see e.g., Hsu et al., 2003). For computational reasons, a random sample (without replacement) of $100,000$ trips is used for this purpose. The results can be downloaded from the journal's website.

Because MNL models are frequently used in discrete choice modeling and classification of travel mode choice (Ben-Akiva and Lerman, 1985), they serve in this study as a baseline classifier. The MNL model is estimated using an ANN-based approach (see Ripley, 2007). The ANN used has no hidden layers and is trained by back propagation with a weight decay constant of 0.01.

Naive Bayes (NB) is a simple machine learning method that calculates class probabilities using Bayes theorem while assuming that the features are independent. Predictions are then made for the class with the highest probability. In order to calculate probabilities from continuous features, their probability distributions must be estimated. This is typically done using kernel density estimation (John and Langley, 1995). Even though the independence assumption of NB rarely holds in practice, the classifier has shown to be competitive with more advanced classifiers (e.g., Huang and Ling, 2005). In this study, kernel density estimation with a Laplace correction factor of 0.001 is used.

SVM is a machine learning method for binary classification. It classifies observations by projecting the independent variables into a high-dimensional feature space, where the classes are linearly separable (Cortes and Vapnik, 1995). Since the basic SVM is a binary classifier, a one-against-one-approach is used for multiclass classification. In this approach, $k(k-1)/2$ binary classifiers are trained, with each classifier learning to distinguish a different pair of k classes. For prediction, the class that receives the most votes from all classifiers is chosen. Here, a SVM with a Gaussian kernel is used. The cost of constraint violation is set to 1.25 and the kernel bandwidth is set to 0.4.

Inspired by the biological brain, ANNs consist of a set of artificial neurons and directed connections between them (e.g., Rojas, 2013). Input data is passed through the network where it is summarized and processed by the neurons and weighted by the connections to give a network output. During the training of an ANN, the weights of the connections are adapted to produce a desired network output. The prediction of class membership is determined by the neuron with the largest output value. Hornik et al. (1989) showed that ANNs can approximate arbitrary continuous functions in

Euclidean space to any degree of accuracy. In this study, an ANN with a single hidden layer of 48 neurons is used. The connection weights are trained by back propagation with a weight decay constant of 0.1.

CTs utilize a tree-like data structure for classification. The nodes of the tree represent binary decision rules which recursively split the feature space, while the leaves of the tree represent the classes (Breiman et al., 1984). Classification trees are easy to interpret and can effectively deal with nonlinear relationships and interactions between variables. However, they are sensitive to noisy data and also have a tendency to overfit (Quinlan, 2014). Tree-based ensemble techniques combine many classification trees in order to form more stable and accurate classifiers than single CTs (Breiman, 1996).

The first tree-based ensemble method selected for comparison is boosting (BOOST). Here, the general idea is to build a sequence of CTs, where each successive tree aims to improve the previously wrong classifications of the preceding trees. Prediction is accomplished by a weighted voting among all CTs. Here, the gradient boosting machine variant (Friedman, 2001) is used. 300 trees are fitted in total. The shrinkage parameter is set to 0.2 and the interaction depth to 48. Additionally, each leaf node must have at least 10 observations.

Bagging (BAG) is a straight-forward application of an ensemble of trees, whereby many CTs are trained in parallel using bootstrap samples of the data. For prediction, class assignment is determined by majority voting among all trees. In this study, 350 classification trees are bagged. Each tree is grown without pruning until the class assignment at each node is unambiguous.

RF is another tree-based ensemble method which is closely related to bagging. While RF also trains many CTs in parallel using bootstrap samples, each split at the nodes of the trees is determined by a random subset of variables (Breiman, 2001). Again, for prediction, a majority vote among all trees determines class membership. In this study, an RF consisting of 450 trees is used and three randomly selected variables are considered for each split at the trees nodes.

All modeling and analyses is done in the R programming environment (R Core Team, 2015) using the 'caret' package (Kuhn, 2008). The 'caret' package provides a common interface for several modeling packages. The relevant modeling packages for this study are 'nnet' (Venables and Ripley, 2002), 'klaR' (Weihs et al., 2005), 'ipred' (Peters and Hothorn, 2015), 'e1071' (Meyer et al., 2015), 'randomForest' (Liaw and Wiener, 2002), and 'gbm' (Ridgeway, 2015).

8.2.3. Model comparison

The performance of each classifier is estimated in this study using 10-fold cross-validation (Kohavi, 1995). This procedure randomly partitions the data into 10 disjoint subsets. One subset at a time is then used for testing the model, while the remaining sets are used to build the model. Consequently, since the testing and training data sets are independent of each other, bias in performance estimation is reduced (e.g., Kohavi, 1995).

The distribution of the dependent variable is imbalanced. For instance, trips by car are done very frequently, while trips by public transport are rare. To account for the class-imbalance, the following procedure is suggested for each training subset of the validation procedure. First, the mean number of trips per travel mode, denoted by n, is calculated. Then, for classes which have less than n cases, observations are sampled with replacement from this class and added to the data set until the class consists of n cases. For classes which have more than n cases, observations are randomly removed from the data set until the class consists of n cases. After this procedure, every class is represented by exactly n observations. Thus, the total size of the data set is not changed by this procedure.

The classification performance is measured using the accuracy and sensitivity statistics. Accuracy measures the overall proportion of correctly classified observations, while sensitivity evaluates the proportion of correctly assigned observations for each class (Japkowicz and Shah, 2011). Hence, sensitivity is particular useful for evaluating classification performance on imbalanced data sets. These statistics are calculated for each model built during the validation procedure and for each repetition.

To evaluate and compare the different classifiers, it is useful to take into account the distribution of the performance statistics (e.g., Hothorn et al., 2005). Following Hothorn et al. (2005), this study evaluates the statistical significance of the classifiers' differences in accuracy as follows. First, the Kruskal–Wallis test with a 5% significance level is used to test the null hypothesis that the performance estimates of all classifiers are not systematically different from each other. Then, the two-sided Wilcoxon rank-sum test is applied to determine the statistical significance of systematic pairwise differences between classifiers. To control the false discovery rate at the 5% level, the p-values are adjusted using the Benjamini–Hochberg procedure (see Benjamini and Hochberg, 1995).

8.2.4. Variable importance

The assessment of variable importance is generally an important analysis task, because it allows variable selection and supports meaningful interpretation. However, this remains a complex task due to interactions and correlations among the variables. Seemingly irrelevant variables may become important only in the context of others, while redundancies between variables may lead to an overestimation of importance (e.g., Strobl et al., 2008). In addition, the assessment of variable importance depends strictly on the model under consideration. If a classifier is incapable of modeling a variable's relationships with a response variable, its importance for the classifier is generally low, while its importance might be high for more powerful classifiers.

Numerous approaches for quantifying variable importance for different models have been proposed (e.g., Hagenauer and Helbich, 2012; Nathans et al., 2012; Olden et al., 2004). In the RF framework, the importance of a variable is commonly evaluated by measuring the change in model performance when randomly permuting the variable in the test data (e.g., Breiman, 2001; Strobl et al., 2007). The more the performance decreases under permutation of a variable, the higher is its importance. This approach can be applied to arbitrary prediction models, given that independent test data for evaluation purposes is available (e.g., Goetz et al., 2015; Knudby et al., 2010; Xu et al., 2014).

In this study, a permutation-based approach to measure the overall importance of each variable is used. This is done by permuting each variable within the test data 10 times for each fold and repetition of the validation procedure and reporting the resulting differences in accuracy. However, in a multiclass classification problem such as travel mode choice, the importance of the variables for the prediction of different classes is also of interest. For example, the ownership of a driver's license might be relevant for predicting car trips, but might be less relevant for the prediction of other travel modes. This study is the first that uses the permutation-based approach for analyzing such importances by reporting the differences in sensitivity for each travel mode under permutation.

8.3. Results

8.3.1. Classification performance

The accuracy of each classifier is shown in Figure 8.1. With respect to median accuracy, RF achieved the best results (0.914), closely followed by BAG (0.906). The third and fourth best classifiers are SVM (0.825) and BOOST (0.801). The accuracy of the other

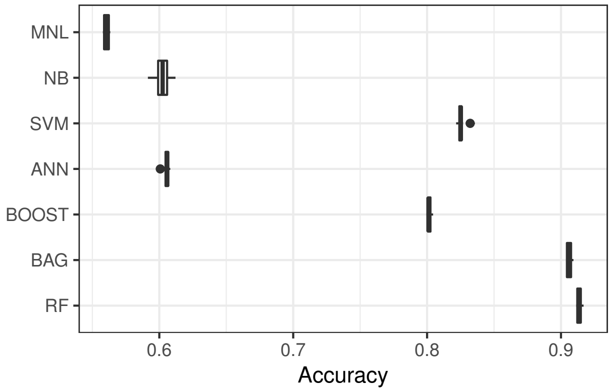


Figure 8.1.: Accuracy for each classifier.

classifiers is substantially lower. The accuracy of ANN (0.606) is only slightly higher than NB (0.602). MNL has the lowest accuracy of all classifiers with 0.561.

The null hypothesis of no performance differences between the classifiers was rejected by the Kruskal–Wallis test at 5% significance level. Table 8.3 shows the results of the two-sided Wilcoxon rank-sum test with adjusted p-values. For ANN and NB the null hypothesis that the results are drawn from the same continuous distributions is not rejected. Hence, these classifiers are the only ones whose accuracy is not significantly different.

Table 8.3.: Results of Wilcoxon rank-sum tests for differences in accuracy. Numbers below the diagonal are p-values, the numbers above the estimated differences. The false discovery rate is controlled at 5%.

	MNL	NB	SVM	ANN	BOOST	BAG	RF
MNL		−0.042	−0.265	−0.045	−0.241	−0.346	−0.353
NB	0.000		−0.223	−0.003	−0.199	−0.303	−0.311
SVM	0.000	0.000		0.022	0.024	−0.081	−0.088
ANN	0.000	0.257	0.000		−0.196	−0.301	−0.308
BOOST	0.000	0.000	0.000	0.000		−0.105	−0.122
BAG	0.000	0.000	0.000	0.000	0.000		−0.008
RF	0.000	0.000	0.000	0.000	0.000	0.000	

The sensitivity of the classifiers for each travel mode is shown in Figure 8.2. Notably, all classifiers, except NB, predict public transport trips more accurately than other travel modes. NB, however, predicts car trips slightly more accurately than public transport trips. Bike trips are generally less accurately predicted than the other travel modes, though the absolute sensitivity values of the classifiers differ. For instance,

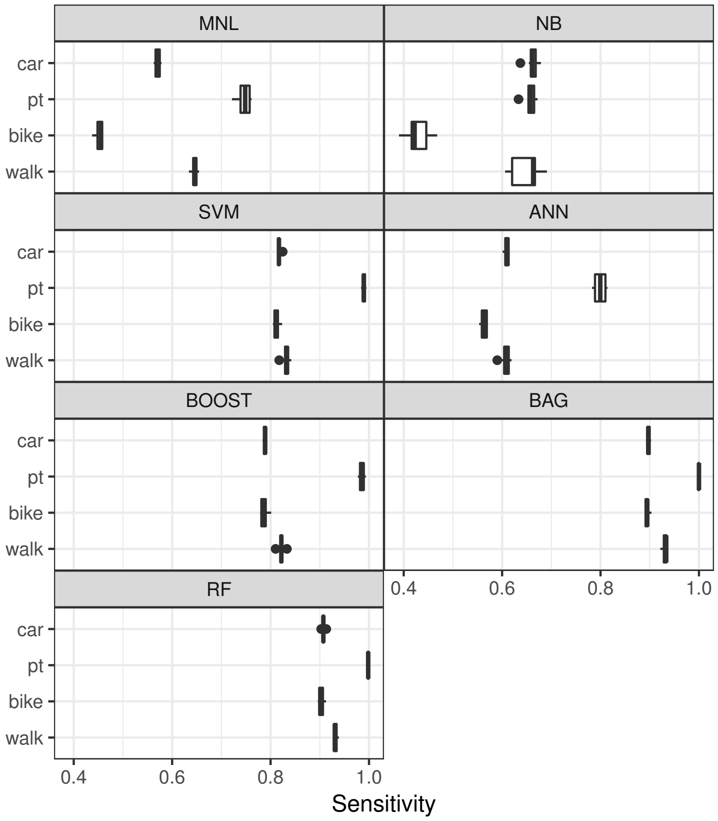


Figure 8.2.: Sensitivity for each classifier.

NB predicts bike trips substantially less accurately than the other travel modes, while for RF and SVM the difference in sensitivity between bike and walking trips is only marginal. Moreover, it can be seen that the sensitivity values of SVM, BOOST, BAG, ANN, and RF follow the same patterns. That is, public transport is predicted most accurately, followed by walking trips. The difference in sensitivity between bike and car trips is only marginal. By contrast, the sensitivity values of MNL and NB follow very different patterns.

8.3.2. Variable importance

Figure 8.3 shows boxplots of the importance of each variable for each classifier with respect to accuracy. By far the most important variable for all classifiers is trip distance. For the other variables the ranking of importance is more complex, though address density, age, number of cars and bicycles per household, and the possession of a driving license are of importance for most classifiers. Exceptions are MNL and NB, for which age, address density, and number of bicycles (only NB) are not important. Generally, the number of important variables is smaller for NB, MNL, and ANN than for the other classifiers. For SVM, by contrast, all variables bear substantial importance. In particular, while education and household income are only marginally important for the other classifiers, these variables are the second and third most important variable for SVM. In addition, while in general the meteorological variables are more important for SVM than for the other classifiers, temperature is generally the most important meteorological variable.

Exemplarily, Figure 8.4 depicts the importance of the variables with respect to sensitivity for MNL (lowest accuracy), Figure 8.5 for BOOST (moderate accuracy), and Figure 8.6 for RF (highest accuracy). While trip distance is the most important variable for all travel modes and classifiers, there exist numerous notable differences between classifiers and travel modes. First, the number of important variables varies substantially with travel mode. For example, three variables are substantially important (ΔSensitivity < -0.2) for predicting public transport trips by RF (distance, number of cars, age), but only a single variable (distance) for predicting other trips by RF. Consequently, and second, some variables are more important for certain travel modes than others. For instance, while the address density is important for predicting public transport trips by any classifier, this variable is basically of no importance for predicting car trips. Third, the importance of variables also varies between classifiers. For example, in contrast to MNL, BOOST and RF identify along temperature the proportion of green space as an important variable for predicting public transport and bicycle trips.

8.4. Discussion

8.4.1. Classification performance

The tree-based ensemble classifiers performed exceptionally well. This indicates that the flexibility which is obtained by combining multiple CTs is particularly useful for modeling travel mode choice. The performance of RF is significantly better than BAG. This difference can be attributed to the larger diversity among the learned trees of RF, which is a result of the RF's procedure for randomized splitting at nodes. Generally, ensemble classifiers perform better if there is significant diversity among the models (Kuncheva and Whitaker, 2003). However, the performance of BOOST is inferior to both RF and BAG. One explanation can be that boosting methods are primarily designed to minimize model bias and are therefore more prone to overfitting, while RF and BAG conceptually aim to reduce model variance (Ganjisaffar et al., 2011).

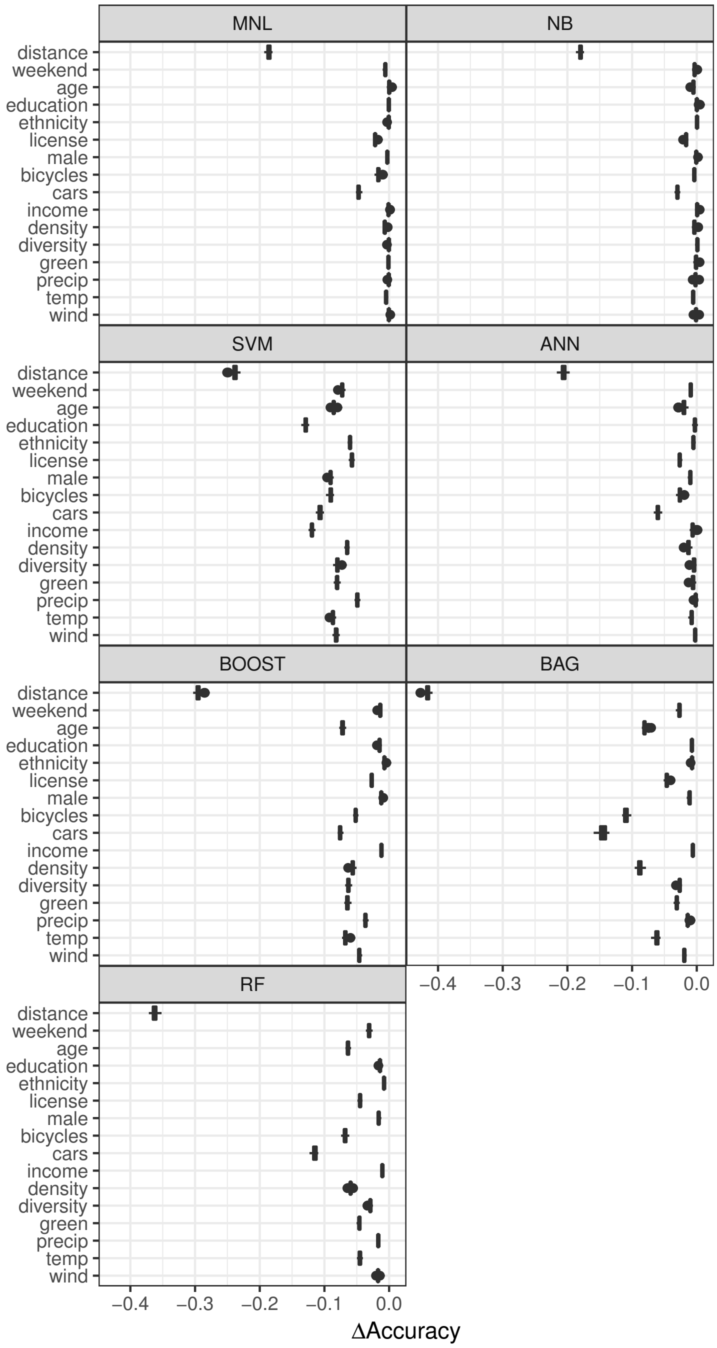


Figure 8.3.: Overall variable importance.

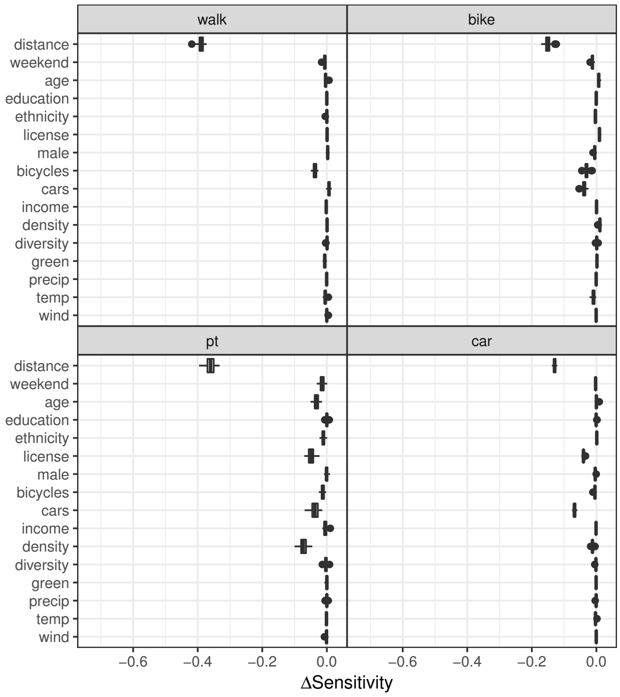


Figure 8.4.: Variable importance for MNL.

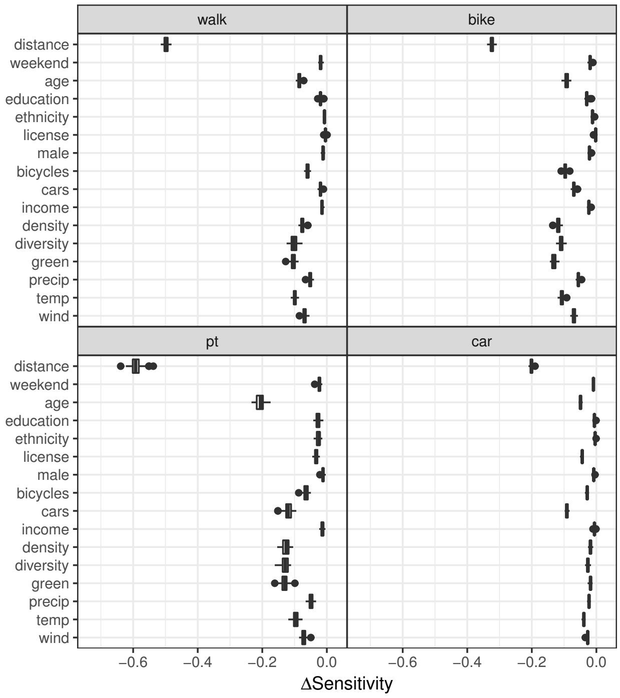


Figure 8.5.: Variable importance for BOOST.

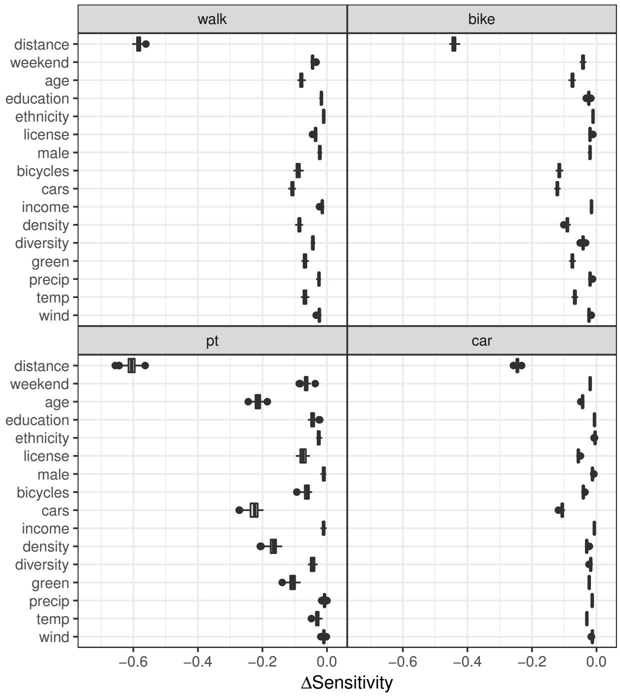


Figure 8.6.: Variable importance for RF.

Furthermore, because boosting methods try to improve previously misclassified data iteratively, outliers can have a critical effect on their performance (Rätsch et al., 2001).

SVM uses the one-against-one approach for multiclass classification. This approach generally tends to increase the classifier's variance, because only small subsets of the data are used to learn to distinguish between each pair of classes (Lee et al., 2004). In addition, it can lead to inconsistent results in which observations are assigned to multiple classes simultaneously (Bishop, 2006). The similar performance of SVM and BOOST, which is a true multiclass classifier, indicates that these issues do not substantially effect the performance of SVM.

The lowest accuracy was provided by MNL, indicating that its modeling capabilities are generally less effective for modeling travel mode choice and/or that the models' assumptions are substantially violated. These findings are in line with previous studies (e.g., Omrani, 2015; Sayed and Razavi, 2000; C. Xie et al., 2003). The accuracy of NB, on the other hand, is significantly higher than MNL and close to ANN, indicating that despite the strict independence assumption of NB (see e.g., Hand and Yu, 2001) it can be useful for the modeling of travel mode choice.

The results of the sensitivity analysis allow a more detailed investigation of accuracy results. Overall, RF predicts all travel modes with high sensitivity. No classifiers predicts any travel mode more accurately than RF. Thus, RF can be considered the most appropriate classifier for modeling travel mode choice. In addition, since the sensitivity of SVM, BOOST, BAG, ANN, and RF basically follow the same patterns, it can be concluded that neither of these classifiers has distinct properties that make it substantially more useful for predicting certain travel modes.

8.4.2. Variable importance

The results of the analysis of variable importance with respect to accuracy show that the considered classifiers, except SVM, generally correspond well with regard to the most important variables, though the magnitudes of variable importance between classifiers differ. In particular, even simple classifiers such as MNL and NB are able determine the most important variables. However, because MNL and NB do only consider a rather small set of variables as important for classification and their generally low classification performance, it can be concluded that more advanced and flexible classifiers are required to model the complex interactions and relationships of most variables.

The similar patterns of variable importance of RF and BAG indicates that these classifiers model relationships between variables in a similar manner. However, the magnitude of importance of some variables is different for RF and BAG, even though both classifiers are based on an ensemble of CTs. For instance, address density is more important for BAG than RF, while the proportion of green space and education is more important for RF than BAG. One explanation for these differences can be that RF, in contrast to BAG, creates CTs on random subsets in order to avoid overfitting (Ho, 1998).

The results also confirm the findings of Rasouli and Timmermans (2014), who found that trip distance is the most important variable and, furthermore, that age is more important than income or education when classifying travel mode choice using tree-based ensemble classifiers. While in this article the number of cars is ranked as the second most important variable by all classifiers, except SVM, Rasouli and Timmermans (2014) reported that car availability is not of substantial importance. A reason for this difference could be that this article considered the total number of available cars, while Rasouli and Timmermans (2014) merely considered the general availability of a car, disregarding the total number of cars available.

Previous studies have also identified weather conditions as important variables for making decision about travel modes (e.g., Garvill et al., 2003; Helbich et al., 2014; Liu et al., 2015; Verplanken et al., 1997), which is confirmed by the results of the present study. However, the results also show that weather variables do not play a dominating role for travel mode choice classification and that temperature is generally more important than precipitation or wind speed. An explanation for the latter can be that temperature also reflects seasonal effects, which are not directly related to weather conditions but nevertheless influence travel mode choice (see Clifton et al., 2011).

In addition, the results emphasize the overall importance of address density for most classifiers. An explanation for this can be that address density is closely related to variables such as parking costs, distance to public transport stations, and travel time, which have been shown to significantly influence the choice of travel modes (see e.g., Frank et al., 2008; Susilo et al., 2012).

The analysis of sensitivity allows a more detailed view of the importance of the variables for the different classifiers. For instance, the results show that address density is generally more important for the prediction of public transport trips than for the other travel modes. Considering that address density is a proxy for population density, these results correspond to Limtanakool et al. (2006), who determined that high population density is associated with an increased use of public transport.

Finally, the results also indicate that temperature and proportion of green space are particularly important for predicting bicycle trips by RF. This supports the results of Winters et al. (2010) and Helbich et al. (2014), who showed that the natural and built environment, as well as temperature, substantially affect bicycle behavior. In addition, the results show that these variables are also important for predicting public transport trips. This is in line with the findings of Nankervis (1999), who showed that public transport is a common alternative to cycling, particularly during bad weather conditions.

8.5. Conclusion

This article presented a systematic comparison of seven different machine learning classifiers for travel mode choice prediction using Dutch travel diary data from the years 2010 to 2012. For this purpose, accuracy and sensitivity analyses have been performed utilizing repeated k-fold cross validation.

The results showed that among the investigated classifiers, RF produced the most accurate predictions. The performance of MNL, arguably the most common model for analyzing travel mode choice, is low. In-depth sensitivity analysis revealed that public transport and car trips are predicted with the highest sensitivity by all classifiers, while walking and bicycles trips are predicted with the lowest sensitivity.

Using a permutation-based approach to measure variable importance, the article showed that with regard to accuracy the most important variable is trip distance, followed by the number of cars per household. The importance of the other variables varies with the applied classifiers. Though generally of little importance, the meteorological variables are more important for SVM than for the other classifiers. Furthermore, a detailed analysis of variable importance with regard to sensitivity has shown that variable importance also varies strictly with the travel mode being predicted. Temperature is more important for predicting public transport and bicycle trips than for other travel modes.

The results suggest that it is necessary to analyze alongside the overall classification performance the importance of variables for the different classifiers and travel modes in order to get a better understanding of the relationships within the data and to allow effective modeling of travel mode choice.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at `https://doi.org/10.1016/j.eswa.2017.01.057`.

Bibliography

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*. Vol. 9. MIT press.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Böcker, L., Amen, P. van, and Helbich, M. (2016). Elderly travel frequencies and transport mode choices in Greater Rotterdam, the Netherlands. *Transportation*, 1–22.

Böcker, L., Dijst, M., and Prillwitz, J. (2013). Impact of everyday weather on individual daily travel behaviours in perspective: A literature review. *Transport reviews*, 33(1), 71–91.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression zrees*. Monterey, CA: Wadsworth and Brooks.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Cantarella, G. E. and De Luca, S. (2003). Modeling transportation mode choice through artificial neural networks. In: *Fourth International Symposium on Uncertainty Modeling and Analysis*, 84–90.

Celikoglu, H. B. (2006). Application of radial basis function and generalized regression neural networks in non-linear utility function specification for travel mode choice modelling. *Mathematical and Computer Modelling*, 44(7), 640–658.

Clifton, K. J., Chen, R. B., and Cutter, A. (2011). Representing weather in travel behaviour models: A case study from Sydney, AUS. In: *Australasian Transport Research Forum 2011 Proceedings*, 28–30.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297.

Dieleman, F. M., Dijst, M., and Burghouwt, G. (2002). Urban form and travel behaviour: Micro-level household attributes and residential context. *Urban Studies*, 39(3), 507–527.

Dios Ortúzar, J. de and Willumsen, L. G. (2011). *Modelling transport.* John Wiley & Sons.

Dow, J. K. and Endersby, J. W. (2004). Multinomial probit and multinomial logit: A comparison of choice models for voting research. *Electoral studies*, 23(1), 107–122.

Ewing, R. and Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American planning association*, 76(3), 265–294.

Ewing, R., Schroeer, W., and Greene, W. (2004). School location and student travel analysis of factors affecting mode choice. *Transportation Research Record: Journal of the Transportation Research Board*, (1895), 55–63.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1), 3133–3181.

Fishman, E., Böcker, L., and Helbich, M. (2015). Adult active transport in the Netherlands: An analysis of its contribution to physical activity requirements. *PloS one*, 10(4), e0121871.

Frank, L., Bradley, M., Kavage, S., Chapman, J., and Lawton, T. K. (2008). Urban form, travel time, and cost relationships with tour complexity and mode choice. *Transportation*, 35(1), 37–54.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Ganjisaffar, Y., Caruana, R., and Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 85–94.

Garvill, J., Marell, A., and Nordlund, A. (2003). Effects of increased awareness on choice of travel mode. *Transportation*, 30(1), 63–79.

Goetz, J. N., Brenning, A., Petschko, H., and Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences*, 81, 1–11.

Gong, L., Morikawa, T., Yamamoto, T., and Sato, H. (2014). Deriving personal trip data from GPS data: A literature review on the existing methodologies. *Procedia-Social and Behavioral Sciences*, 138, 557–565.

Hagenauer, J. and Helbich, M. (2012). Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *International Journal of Geographical Information Science*, 26(6), 963–982.

Hand, D. J. and Yu, K. (2001). Idiot's Bayes — not so stupid after all? *International statistical review*, 69(3), 385–398.

Helbich, M. (2016). Children's school commuting in the Netherlands: Does it matter how urban form is incorporated in mode choice models? *International Journal of Sustainable Transportation.*

Helbich, M., Böcker, L., and Dijst, M. (2014). Geographic heterogeneity in cycling under various weather conditions: Evidence from Greater Rotterdam. *Journal of Transport Geography*, 38, 38–47.

Hensher, D. A. and Ton, T. T. (2000). A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. *Transportation Research Part E: Logistics and Transportation Review*, 36(3), 155–172.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832–844.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359–366.

Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3), 675–699.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). *A practical guide to support vector classification.* Tech. rep. Department of Computer Science, National Taiwan University.

Huang, J. and Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3), 299–310.

Japkowicz, N. and Shah, M. (2011). *Evaluating learning algorithms: A classification perspective.* Cambridge University Press.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc., 338–345.

Knudby, A., Brenning, A., and LeDrew, E. (2010). New approaches to modelling fish–habitat relationships. *Ecological Modelling*, 221(3), 503–511.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence.* Vol. 14. 2, 1137–1145.

Koppelman, F. S. and Wen, C.-H. (1998). Alternative nested logit models: structure, properties and estimation. *Transportation Research Part B: Methodological*, 32(5), 289–298.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, 28(5).

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2), 181–207.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465), 67–81.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.

Limtanakool, N., Dijst, M., and Schwanen, T. (2006). The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium- and longer-distance trips. *Journal of Transport Geography*, 14(5), 327–341.

Liu, C., Susilo, Y. O., and Karlström, A. (2015). The influence of weather characteristics variability on individual's travel mode choice in different seasons and regions in Sweden. *Transport Policy*, 41, 147–158.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In: P. Zarembka (ed.). *Frontiers in Econometrics*. New York, NY: Academic Press.

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*. R package version 1.6-7.

Murray, K. and Conner, M. M. (2009). Methods to quantify variable importance: implications for the analysis of noisy ecological data. *Ecology*, 90(2), 348–355.

Nankervis, M. (1999). The effect of weather and climate on bicycle commuting. *Transportation Research Part A: Policy and Practice*, 33(6), 417–431.

Nathans, L. L., Oswald, F. L., and Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, 17(9), 1–19.

Olden, J. D., Joy, M. K., and Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3), 389–397.

Omrani, H. (2015). Predicting travel mode of individuals by machine learning. *Transportation Research Procedia*, 10, 840–849.

Omrani, H., Charif, O., Gerber, P., Awasthi, A., and Trigano, P. (2013). Prediction of individual travel mode with evidential neural network model. *Transportation Research Record: Journal of the Transportation Research Board*, (2399), 1–8.

Onderzoek Verplaatsingen in Nederland (2014). *Onderzoeksbeschrijving OViN 2010–2014. Data archiving and networked services (DANS).* `http://www.cbs.nl/nl-NL/menu/themas/verkeer-vervoer/methoden/dataverzameling/korte-`

`onderzoeksbeschrijvingen/ovin-beschrijving-art.htm`. Accessed on 27th January 2016.

Peters, A. and Hothorn, T. (2015). *ipred: Improved predictors*. R package version 0.9-5.

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rasouli, S. and Timmermans, H. J. P. (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *European Journal of Transportation and Infrastructure Research*, 14(4), 412–424.

Rätsch, G., Onoda, T., and Müller, K.-R. (2001). Soft margins for AdaBoost. *Machine learning*, 42(3), 287–320.

Ridgeway, G. (2015). *gbm: Generalized boosted regression models*. R package version 2.1.1.

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

Rojas, R. (2013). *Neural networks: A systematic introduction*. Springer Science & Business Media.

Sayed, T. and Razavi, A. (2000). Comparison of neural and conventional approaches to mode choice analysis. *Journal of Computing in Civil Engineering*, 14(1), 23–30.

Schwanen, T. and Mokhtarian, P. L. (2005). What affects commute mode choice: neighborhood physical structure or preferences toward neighborhoods? *Journal of Transport Geography*, 13(1), 83–99.

Shafique, M. A. and Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, 42(1), 163–188.

Shen, L. and Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, 34(3), 316–334.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1), 1.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1), 1.

Susilo, Y. O., Williams, K., Lindsay, M., and Dair, C. (2012). The influence of individuals' environmental attitudes and urban design features on their travel patterns in sustainable neighborhoods in the UK. *Transportation Research Part D: Transport and Environment*, 17(3), 190–200.

Tang, L., Xiong, C., and Zhang, L. (2015). Decision tree method for modeling travel mode switching in a dynamic behavioral process. *Transportation Planning and Technology*, 38(8), 833–850.

Tortum, A., Yayla, N., and Gökdağ, M. (2009). The modeling of mode choices of intercity freight transportation with the artificial neural networks and adaptive neuro-fuzzy inference system. *Expert Systems with Applications*, 36(3), 6199–6217.

Tsai, T.-H., Lee, C.-K., and Wei, C.-H. (2009). Neural network based temporal feature models for short-term railway passenger demand forecasting. *Expert Systems with Applications*, 36(2), 3728–3736.

Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S.* Fourth. ISBN 0-387-95457-0. New York: Springer.

Verplanken, B., Aarts, H., and Van Knippenberg, A. (1997). Habit, information acquisition, and the process of making travel mode choices. *European Journal of Social Psychology*, 27(5), 539–560.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klaR Analyzing German business cycles. In: *Data Analysis and Decision Support.* Ed. by D. Baier, R. Decker, and L. Schmidt-Thieme. Berlin: Springer-Verlag, 335–343.

Winters, M., Brauer, M., Setton, E. M., and Teschke, K. (2010). Built environment influences on healthy transportation choices: bicycling versus driving. *Journal of Urban Health*, 87(6), 969–993.

Xian-Yu, J.-C. (2011). Travel mode choice analysis using support vector machines. In: *ICCTP 2011: Towards Sustainable Transportation Systems.* 360–371.

Xie, C., Lu, J., and Parkany, E. (2003). Work travel mode choice modeling with data mining: Decision trees and neural networks. *Transportation Research Record: Journal of the Transportation Research Board*, (1854), 50–61.

Xu, L., Li, J., and Brenning, A. (2014). A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery. *Remote Sensing of Environment*, 141, 14–23.

Zhan, G., Yan, X., Zhu, S., Wang, Y., et al. (2016). Using hierarchical tree-based regression model to examine university student travel frequency and mode choice patterns in China. *Transport Policy*, 45(C), 55–65.

Zhang, Y. and Xie, Y. (2008). Travel mode choice modeling with support vector machines. *Transportation Research Record: Journal of the Transportation Research Board*, (2076), 141–150.

Zhao, D., Shao, C., Li, J., Dong, C., and Liu, Y. (2010). Travel Mode Choice Modeling Based on Improved Probabilistic Neural Network. In: *Traffic and Transportation Studies 2010.* ASCE, 685–695.

9. Local modelling of land consumption in Germany with RegioClust

Authors

Julian Hagenauer and Marco Helbich

Journal

International journal of applied earth observation and geoinformation

Status

Received 17 August 2017; received in revised form 13 October 2017; accepted 13 October 2017

Contribution statement

Julian Hagenauer has developed the methods, designed the experiments, and has written the manuscript for the study. Marco Helbich supported this publication by continuously discussing the design and results of the study and by proofreading the manuscript.

Abstract

Germany is experiencing extensive land consumption. This necessitates local models to understand actual and future land consumption patterns. This research examined land consumption rates on a municipality level in Germany for the period 2000–10 and predicted rates for 2010–20. For this purpose, RegioClust, an algorithm that combines hierarchical clustering and regression analysis to identify regions with similar relationships between land consumption and its drivers, was developed. The performance of RegioClust was compared against geographically weighted regression (GWR). Distinct spatially varying relationships across regions emerged, whereas population density is suggested as the central driver. Although both RegioClust and GWR predicted an increase in land consumption rates for east Germany for 2010–20, only RegioClust forecasts a decline for west Germany. In conclusion, both models predict for 2010–20 a rate of land consumption that suggests that the policy objective of reducing land consumption to 30 ha per day in 2020 will not be achieved. Policymakers are advised to take action and revise existing planning strategies to counteract this development.

Keywords: Land use; drivers; spatial clustering; spatial heterogeneity; regression; Germany

9.1. Introduction

Urbanization is occurring at an unprecedented rate (United Nations, 2015), whereby natural, agricultural, and forestry landscapes are converted into built-up areas. This irreversible anthropogenic process, commonly termed land consumption (Nuissl and Schroeter-Schlaack, 2009), has severe and long-lasting consequences for natural habitats, causing a loss of biodiversity, atmospheric pollution, etc. (Foley et al., 2005; McKinney, 2002; Seto et al., 2012).

Land consumption is of particular relevance in Germany, which has one of the highest rates within the European Union (Kroll and Haase, 2010; Siedentop and Kausch, 2004). Even in areas with a declining population, the expansion of built-up areas continues across Germany (Haase et al., 2013). To prevent a further increase, the federal government implemented policies to limit the land consumption rate (LCR) to 30 ha per day up to the year 2020 (Die Bundesregierung, 2016). Although these policies are implemented at the national level, local authorities at the municipal level are still granted spatial planning autonomy by the government. As a consequence, restricting land consumption is a continuous process of reconciling interests across different administrative hierarchies (Jakubowski and Zarth, 2003; Malburg-Graf et al., 2007). Both economic incentives for stakeholders to promote the reuse of formerly built-up land (Borchard, 2011; Schultz and Dosch, 2005) and evidence-based local policymaking are prerequisites to counteract uncoordinated and excessive land consumption (Shafizadeh-Moghadam and Helbich, 2015). Long-term policies need to be founded on precise and data-driven land consumption models (Veldkamp and Lambin, 2001; Verburg et al., 2004; Vliet et al., 2016).

While numerous studies deal with urban growth of specific metropolitan areas (Basse et al., 2016; Zeng et al., 2015), to the best of our knowledge, this is the first study addressing LCRs on a nationwide level using local regression-based modeling. A wide spectrum of approaches to modeling land use change has been proposed (Brown et al., 2004; Shafizadeh-Moghadam et al., 2017b; Triantakonstantis and Mountrakis, 2012; Verburg et al., 2004). Markov-cellular automata is a frequently applied model (Aburas et al., 2017; Arsanjani et al., 2013; de Noronha Vaz et al., 2012; Guan and Rowe, 2016; Li et al., 2017). However, the calibration and validation of cellular automata together with the development of transition rules (e.g., neighborhood definitions) is challenging and relies mostly on ad hoc definitions and heuristics (Li et al., 2017; Shafizadeh-Moghadam et al., 2017a). Further, Markov-cellular automata does not consider the underlying drivers. To circumvent these limitations, artificial neural networks (Shafizadeh-Moghadam et al., 2017b), random forests (Haas and Ban, 2014), and support vector machines (Samardžić-Petrović et al., 2016), or a combination of these approaches (Arsanjani et al., 2013; Omrani et al., 2017), have been suggested to determine the effects of environmental and socioeconomic drivers on land consumption.

Although these machine learning approaches have methodical advantages over statistical methods like regression (e.g., being free of assumptions concerning the input data) (Haykin, 2009), there is no consensus on how to integrate spatial autocorrelation and spatial heterogeneity in these models. Spatial autocorrelation means that areas that are close to each other are subject to similar land consumption processes (Anselin, 2010), while spatial heterogeneity (Fotheringham et al., 2003) means that the associations between land consumption and its drivers vary spatially (Shafizadeh-Moghadam and Helbich, 2015). As stressed by several studies (Anselin, 2010; Brunsdon et al., 1996), ignoring either issue can seriously bias results and might lead to false conclusions and inappropriate policies. Nevertheless, non-spatial regression models are still often applied to assess drivers of land use change (Achmad et al., 2015; Arsanjani et al., 2013; Hu and Lo, 2007; Van Dessel et al., 2011). Whereas spatial autocorrelation has received some attention in the literature (Ay et al., 2017; Dendoncker et al., 2007; Ku, 2016), far less has been devoted to spatial heterogeneity (exceptions are Bagan and Yamagata, 2015; Luo and Wei, 2009; Maimaitijiang et al., 2015; Shafizadeh-Moghadam and Helbich, 2015). Explicitly modeling spatial heterogeneity is particularly important when conducting nationwide studies, where relationships can be a priori expected to vary across space due to different levels of regional economic wealth, environmental differences, and local planning policies.

Only a few models exist to explore spatially heterogeneity. A widespread approach is geographically weighted regression (GWR) (Brunsdon et al., 1996; Fotheringham et al., 2003). GWR uses the spatial distance of neighboring observations in order to estimate local coefficients. Both Luo and Wei (2009) as well as Shafizadeh-Moghadam and Helbich (2015) estimated GWR-based urban growth models and confirmed that GWR not only produces more accurate results compared to a global (i.e., study area-wide) regression model, but also reduces residual spatial autocorrelation. No less important, there is statistical evidence that the underlying drivers vary significantly across space (Hennig et al., 2015), which is paramount for place-based planning — a fact that global models had not uncovered (Achmad et al., 2015).

Despite these appealing advantages, GWR is subject to methodological debate. In addition to the high volatility of the resulting coefficient surfaces, multicollinearity amongst the estimated GWR coefficients is reported (Griffith, 2008; Wheeler and Tiefelsdorf, 2005). While mild correlations obfuscate coefficient interpretation, strong correlations make it hardly possible to make a reliable separation of individual variable effects (Helbich and Griffith, 2016; Wheeler and Tiefelsdorf, 2005). Others note that GWR itself artificially introduces correlations among coefficients, though the input variables are uncorrelated, potentially artificially causing sign reversals (Páez et al., 2011). Fotheringham and Oshan (2016) refuted this critique through simulations demonstrating that GWR is rather robust against coefficient multicollinearity. However, GWR is not recommended as an inferential tool (Páez et al., 2011) because model calibration (e.g., bandwidths selection) as well as interpretation of the model output (e.g., continuous parameter surfaces) remain challenging.

To circumvent some of these limitations, an alternative approach, termed RegioClust, was developed. RegioClust identifies regions with similar associations between the dependent and the independent variables and calculates local parameter estimates for each region. Such a region-based approach is useful because it facilitates the definition of place-based policies, ensures that local policies have a homogeneous impact, and supports scenario development (de Noronha Vaz et al., 2012; Fischer, 1980). In addition, this study addressed the local drivers of LCRs in Germany at the level of municipalities for the period 2000–10, something that had not been done before, and the predicted rates for 2010–20. The research questions were as follows:

- To what extent did the relationships between LCRs and the drivers vary across Germany in 2000–10?
- Does RegioClust predict actual and future LCRs more accurately than GWR?
- Will the predicted LCRs in 2010–20 be below the targeted 30 ha per day?

The rest of the article is structured as follows. Section 9.2 outlines the materials and methods; Section 9.3 summarizes the results; Section 9.4 discusses the results in the context of the existing literature, and Section 9.5 presents the conclusions.

9.2. Materials and methods

9.2.1. Study area

Germany is the most populous country in Europe: In 2010, the country's $357,375$ km^2 of land was home to about 81 million people. The present study was longitudinal and based on the administrative units of municipalities. Municipalities are an appropriate analyses level, as they are small in size and represent the lowest planning level in Germany. Non-contiguous regions, such as islands or exclaves (e.g., Sylt), and unincorporated regions without populations (e.g., Sachsenwald) were removed. This resulted in a total of $11,357$ municipalities.

9.2.2. Data

The central variable was LCR per territorial unit. The built-up areas comprised settlement and transportation infrastructure for the years 2000 and 2010. Data were extracted from the IÖR Monitor (Meinel and Schumacher, 2010), which is based on the official German digital landscape model ATKIS®–Basis–DLM (Bundesamt für Kartographie und Geodäsie, 2016). The LCR, subsequently serving as continuous response variable, was computed by dividing the difference between the consumed land (e.g., built-up areas, transportation infrastructure) in 2010 and 2000 by the total area of the municipality.

Selecting the explanatory variables was guided by data availability and a literature review (Dubovyk et al., 2011; Kretschmer et al., 2015). Six area-level covariates were collected for 2000 and 2010. As strong evidence exists that accessibility is one of the major drivers of urban growth (Duranton and Turner, 2012; Iacono et al., 2008), the spatial distance (in km) from the center of each municipality to the nearest major highway was calculated. To approximate a Gaussian distribution, the square root (srDistHwy) was taken. The highway data were retrieved from the ATKIS®–Basis–DLM. Employment rate (EmplRate) served as proxy variable for wealth, which is highly correlated with urbanization (Bloom et al., 2008). In order to differentiate between municipalities with different housing characteristics, the proportion of family houses (FamHouse) was included. To model urbanization pressure through population inflow, the net migration rate (NetMig) was incorporated. To control for the degree of urbanity, population density in $1,000$ people per km^2 was included. To approximate a Gaussian-like distribution, the variable was log transformed (logPopDens). Data on the average tax revenue in $1,000$€ per capita (TaxRev) was collected to represent social deprivation. All these data were obtained from the Federal Institute for Research on Building, Urban Affairs and Spatial Development. Figure 9.1 depicts the spatial distribution of each covariate for the year 2000.

9.2.3. Methods

RegioClust

RegioClust consists of two steps (Figure 9.2): While the first step of RegioClust determines spatial clusters, the second step determines regions with similar relationships between the dependent variable and independent variables whereas a separate local regression model is estimated for each region. Thus, RegioClust combines spatial clustering with local modeling.

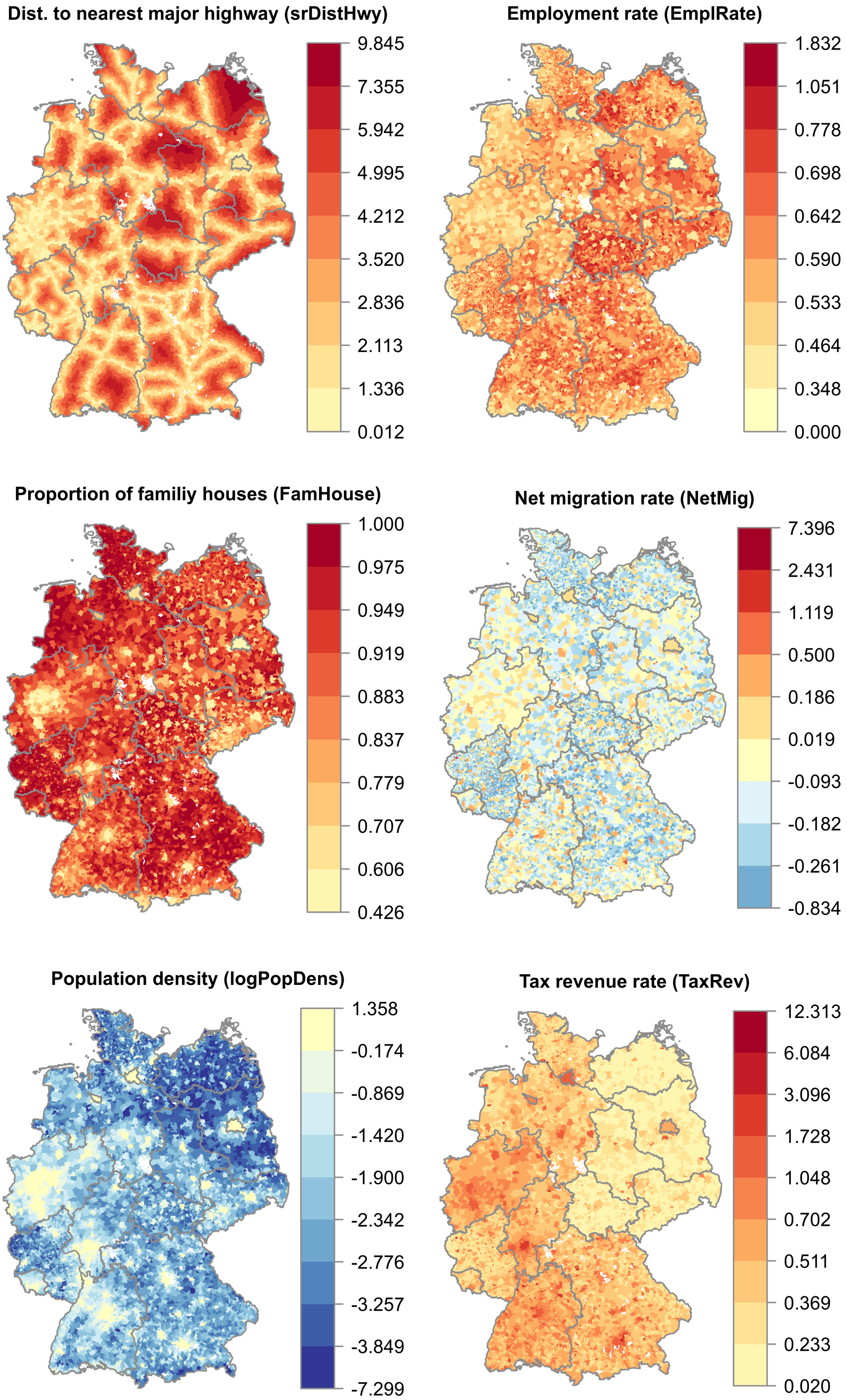


Figure 9.1.: Drivers of land consumption for the year 2000 on the municipality level.

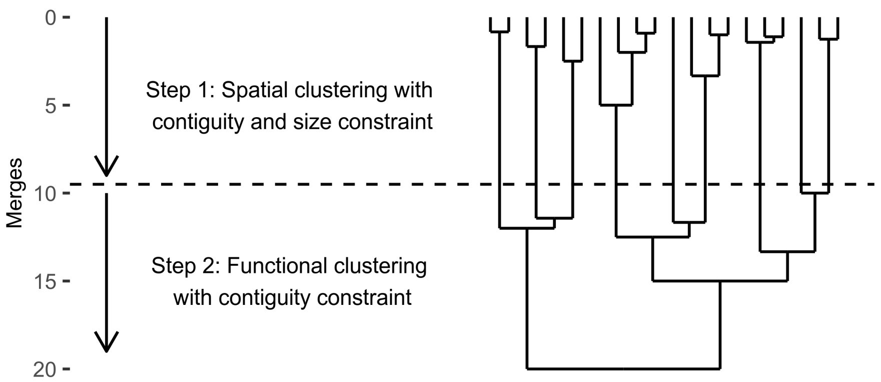

0
5
10
15
20
Merges
Step 1: Spatial clustering with
contiguity and size constraint
Step 2: Functional clustering
with contiguity constraint

Figure 9.2.: The hierarchical clustering steps of RegioClust.

In detail, in the first step, each observation is considered as a single cluster. Then, the clusters are repeatedly merged. Which clusters are merged is determined by Ward's cluster criterion (Ward, 1963). This criterion determines the clusters that minimize the increase in the total within-cluster variance after merging. The variance is measured by evaluating the spatial distances between the observations (i.e., Euclidean distances between the municipality-based centroids). As a consequence, observations within the same cluster tend to be spatially nearby and thus tend to have similar relationships between the dependent variable and its drivers. As spatially contiguous clusters are essential for many applications (Helbich et al., 2013; Spielman and Folch, 2015), the merging criterion is modified to only consider spatially adjacent clusters (Guo, 2009; Murtagh, 1985; Ruß and Kruse, 2011) whose size is smaller than i observations. The clustering ends when no more clusters can be merged, which is typically when all clusters consist of at least i observations.

In the second step, to obtain spatial clusters with a similar response–covariate relationship, the resulting clusters from the first step are merged employing hierarchical clustering with a contiguity (but not size) constraint. The merging criterion is different from the one in the first step. For each cluster a linear model is estimated, and the clusters with the lowest increase in the residual sum of squares (RSS) are merged. The clustering stops when a given number of clusters j is obtained or no more clusters can be merged. Since each resulting cluster refers to a contiguous geographic region for which a uniquely associated linear model exists, RegioClust refers to a pooled piece-wise linear model.

An open-source software implementation of RegioClust can be downloaded from `https://github.com/jhagenauer/regioclust`.

Geographically weighted regression

GWR is a locally weighted regression (Brunsdon et al., 1996) that was used as a benchmark for RegioClust to model spatial heterogeneity. Briefly, GWR estimates coefficients for a sub-set of locations by taking the distance of observations into account. The local coefficients are estimated by solving a location-specific weighted least squares model (Fotheringham et al., 2003). The weights are given by a weight matrix that is specified using a local kernel function that models a distance decay between locations. Nearby observations receive higher weights than distant ones. The choice of the kernel function generally has little impact on the results, given that it is smooth. Either a Gaussian or a bisquare kernel function is commonly used. More crucial than the kernel type is the choice of the bandwidth, which is frequently determined using cross validation (Brunsdon et al., 1996; Fotheringham et al., 2003). Moreover, the bandwidth can vary across space depending on the distribution of the data. If the regression points are sparsely distributed across space, a larger bandwidth is selected, and vice versa. Using an adaptive rather than a predefined bandwidth has the advantage that it reduces the number of extreme coefficients (Fotheringham et al., 2003). The GWR models were estimated by means of the ’GWmodel’ package (Lu et al., 2014) using the R programming environment (R Core Team, 2015).

9.3. Results

9.3.1. Descriptive statistics

The descriptive statistics show that in the year 2000, 9.800% of Germany’s total area was covered by built-up areas and transportation infrastructure. In 2010, the proportion had increased to 10.603% (+0.803%). This corresponds to an LCR of 77.455 ha per day. The amount of built-up area differs between east and west Germany. In west Germany, the proportion of covered area increased from 10.555% in 2000 to 11.651% in 2010 (+1.097%), whereas in east Germany the proportion increased from 8.094% in 2000 to 8.233% in 2010 (+0.014%). This corresponds to an LCR of 73.330 ha per day in west Germany and 4.126 ha per day in east Germany.

Figure 9.3 depicts the LCRs for each municipality. A value of zero indicates that the amount of consumed land did not change between 2000 and 2010, while a value of 50%, for example, indicates that the amount of consumed land increased by 50% of the municipality’s total area. Using first-order queen contiguity, the Moran’s I statistic confirms that the LCRs are not randomly distributed across Germany and that significant regional differences exist ($I = 0.240$, $p < 0.05$).

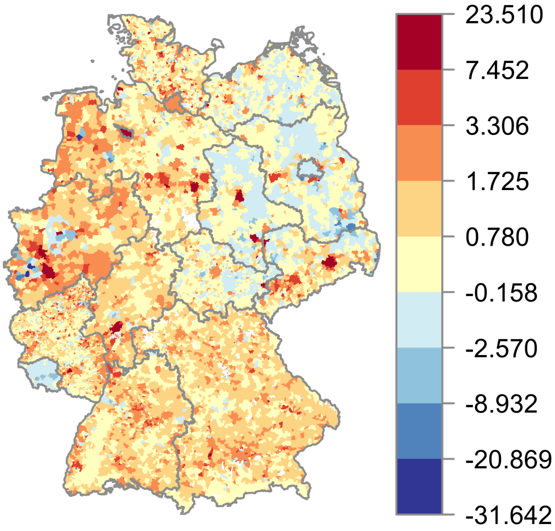


Figure 9.3.: LCRs in 2000–10 (in % of the municipalities' areas).

9.3.2. Model fit

Figure 9.4 shows the Akaike information criterion (corrected for finite sample sizes) (AICc) (Burnham and Anderson, 2004) of RegioClust for $i = 8$ to 12 (i.e., the minimum number of observations per cluster) and for $j = 1$ to 250 (i.e., the total number of clusters). A low AICc value refers to a good compromise between model fit and model complexity (i.e., number of parameters). The Figure illustrates that for fixed values of j, the AICc mostly increases with i. Also, for each value of i and beginning with $j = 1$, the AICc score decreases with increasing j until its minimum is reached. Beyond this, the AICc increases with j. The value of j for which the minimum is obtained is generally lower the smaller the value of i. Figure 9.5 shows the number of outlying coefficients of RegioClust for $i = 8$ to 12 and for $j = 1$ to 250. A coefficient is considered an outlier if its value is three times the interquartile range below the first quartile or above the third quartile of the ranked values. The presence of outlying coefficients is an indicator of local overfitting which is not directly taken into account by the AICc. Generally, the number of outlying coefficients increases with i and j. Also, RegioClust tends to estimate extreme coefficients for population density, in particular for large values of j. Overall, RegioClust represents for a wide range of i and j a reasonable trade-off between model fit, model complexity, and local overfitting. In the following,

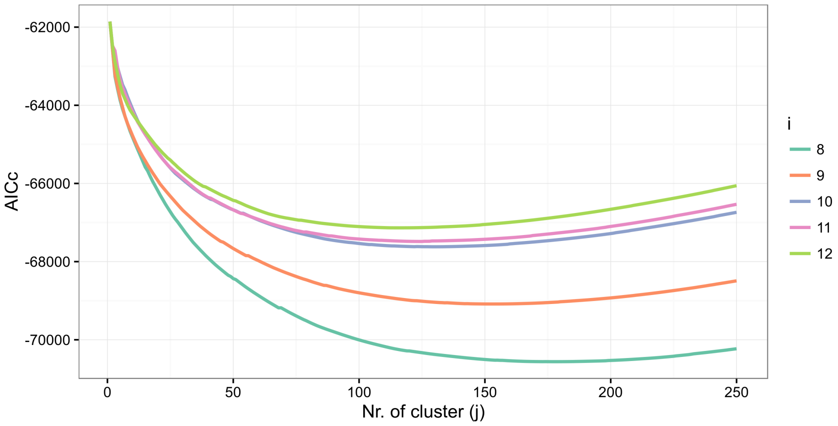


Figure 9.4.: AICc statistics for RegioClust.

the RegioClust model with $i = 10$ and $j = 130$ will be considered. The Moran's I test statistic shows marginal significant spatial dependence of the residuals ($I = 0.030$, $p < 0.05$).

For comparison purposes, a GWR model with Gaussian kernel and adaptive bandwidth using the 20 nearest neighbors was estimated. The number of neighbors was selected by optimizing the AICc. Monte Carlo tests confirmed the significant spatial variability of all coefficients ($p < 0.05$). The AICc is $-63,857.010$. The Moran's I test statistic refers to marginal spatial dependence of the residuals ($I = 0.033$,$p < 0.05$).

On a Lenovo Thinkpad X230, Intel® Core ™ i5–3320M CPU@2.60 GHz with 16 GB RAM, it took 32.510 min to estimate the GWR model and 72.026 min to compute the RegioClust model.

Figure 9.6 depicts the spatial regions outlined by the RegioClust model and the local model fits for both RegioClust and GWR. It shows that in the southwest of Germany (i.e., Rhineland–Palatinate and Saarland), the regions are smaller than in the rest. Many of the larger regions consist of rural municipalities as well as medium-sized cities (e.g., Nürnberg). A few small regions consist of only a single large city and its close surroundings (e.g., Hamburg and Dresden). The largest region, with an area of about $45,133$ km^2, is located in the west and comprises the cities of Kassel and Münster, and some outskirts of the Rhine–Ruhr metropolitan region.

Both models show a pronounced volatility of the model fit in the southwest (i.e., Rhineland–Palatinate and Saarland) while fitting the data particularly well in eastern Bavaria (i.e., east of Augsburg and Nürnberg). However, some differences are notable. For example, whereas RegioClust provides a better fit for Berlin, Dresden, and Hamburg, GWR shows an improved fit for the south of Kassel and the northwestern surroundings

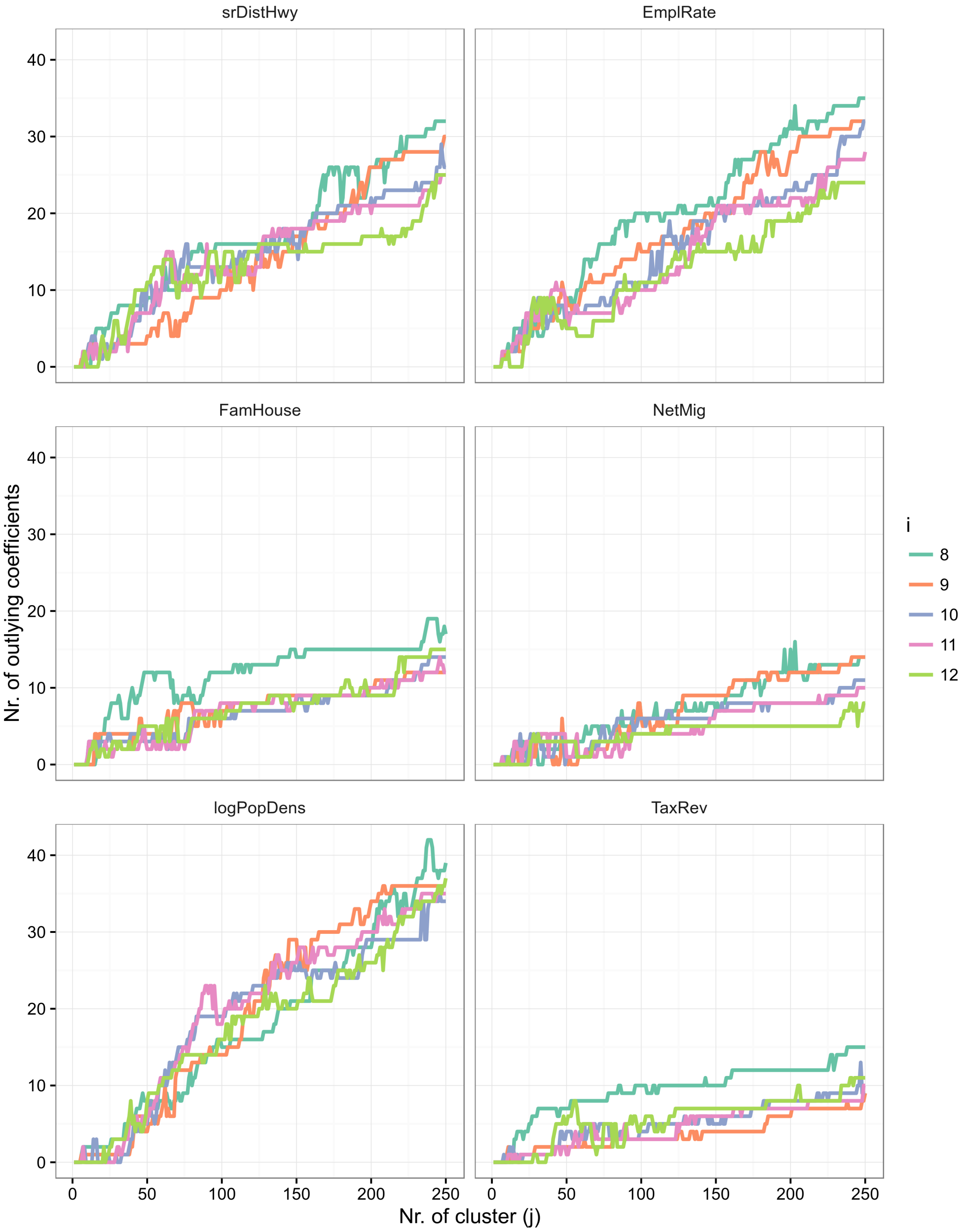


Figure 9.5.: Number of outlying coefficients for RegioClust.

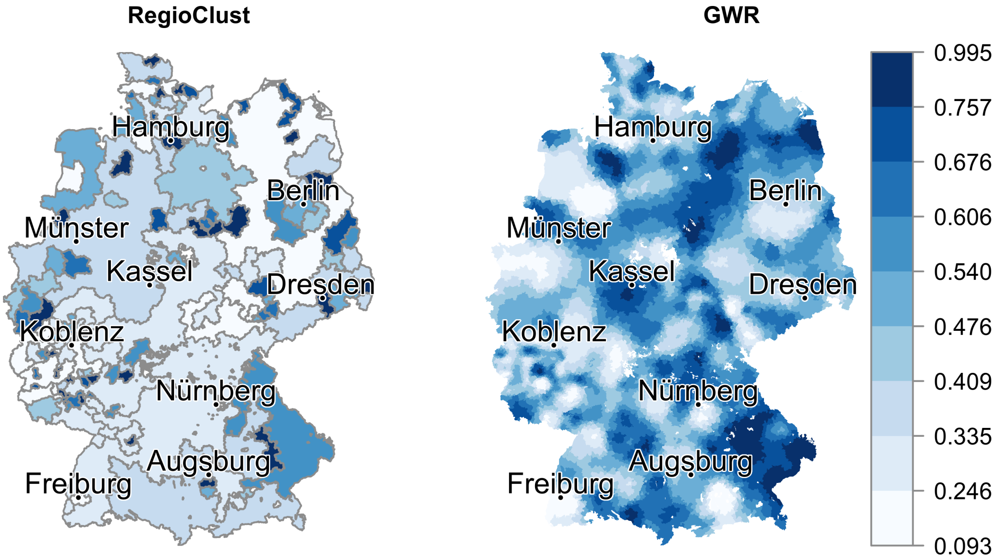


Figure 9.6.: LCRs in 2000–10 (in % of the municipalities' areas).

of Berlin. Although the R^2 is spatially randomly distributed, smaller RegioClust regions tend to have higher R^2 than large regions. This was confirmed by analysis of the Pearson's correlation coefficient ($\rho = -0.272$, $p < 0.05$).

9.3.3. Coefficients of RegioClust and GWR

The significant coefficients (excluding the intercept, $p < 0.05$) are shown for RegioClust in Figure 9.7 and for GWR in Figure 9.8. Spatial heterogeneity in the associations is evident. For both RegioClust and GWR, population density is a key driver of LCR at the national level. It reaches significance more often than any other coefficient. In contrast, employment rate and distance to nearest major highway are relevant drivers for only a few municipalities. They less often reach significance compared to other coefficients. Overall, with the exception of population density, the number of municipalities showing a significant relationship is higher for the coefficients of RegioClust than for those of GWR.

A detailed inspection of Figure 9.7 and 9.8 reveals some notable similarities across the coefficient surfaces. For instance, whereas for most municipalities population density is positively related to LCR, RegioClust and GWR estimate a significant negative relationship for the federal state of Saarland (g) and for the north of the district of Dennim (a). Analogously, for the federal state of Bremen including its surroundings (b), both models estimate a strong and significant negative association between proportion

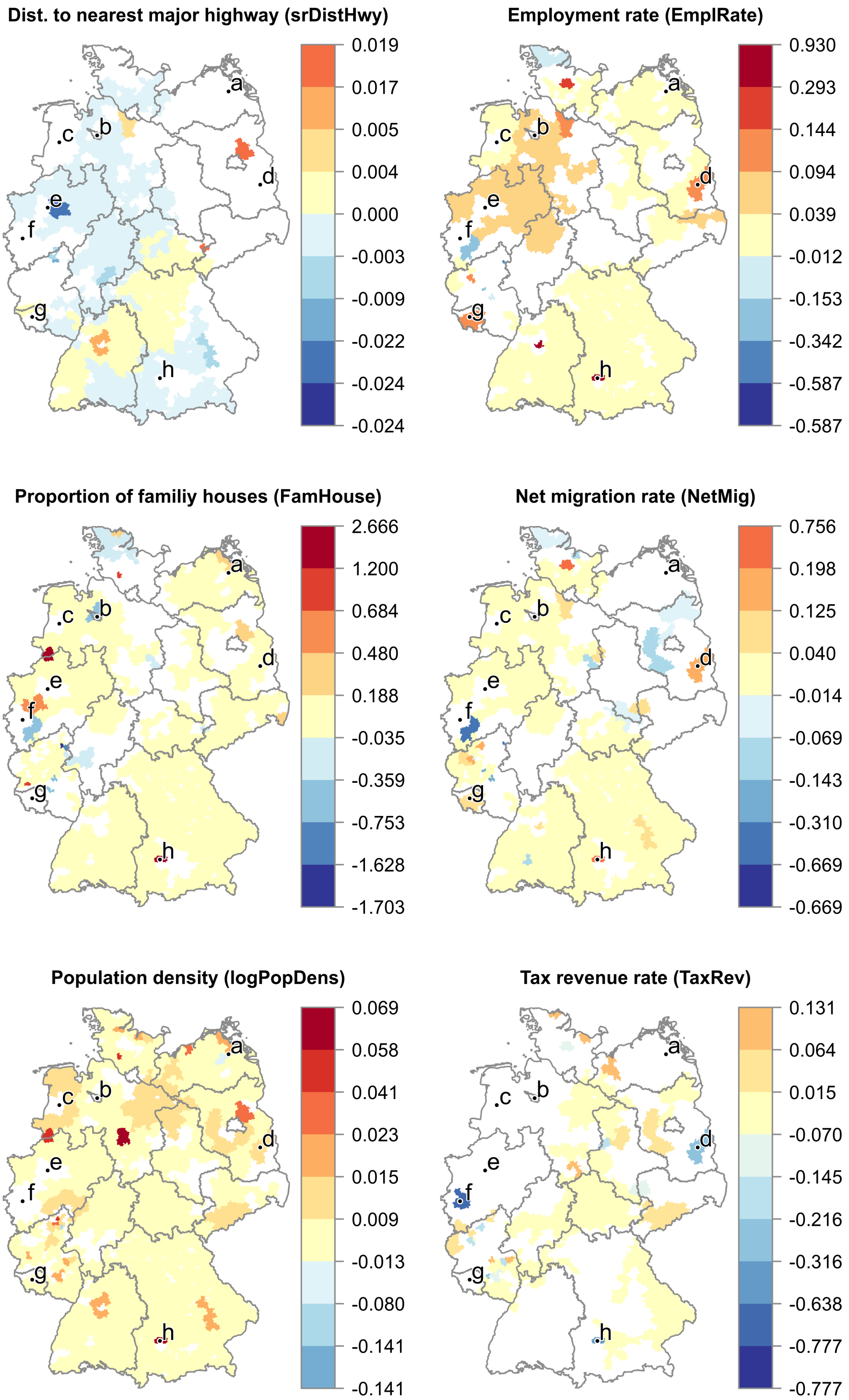


Figure 9.7.: Estimated local coefficients of RegioClust (areas colored in white refer to insignificant associations at the 0.05 level).

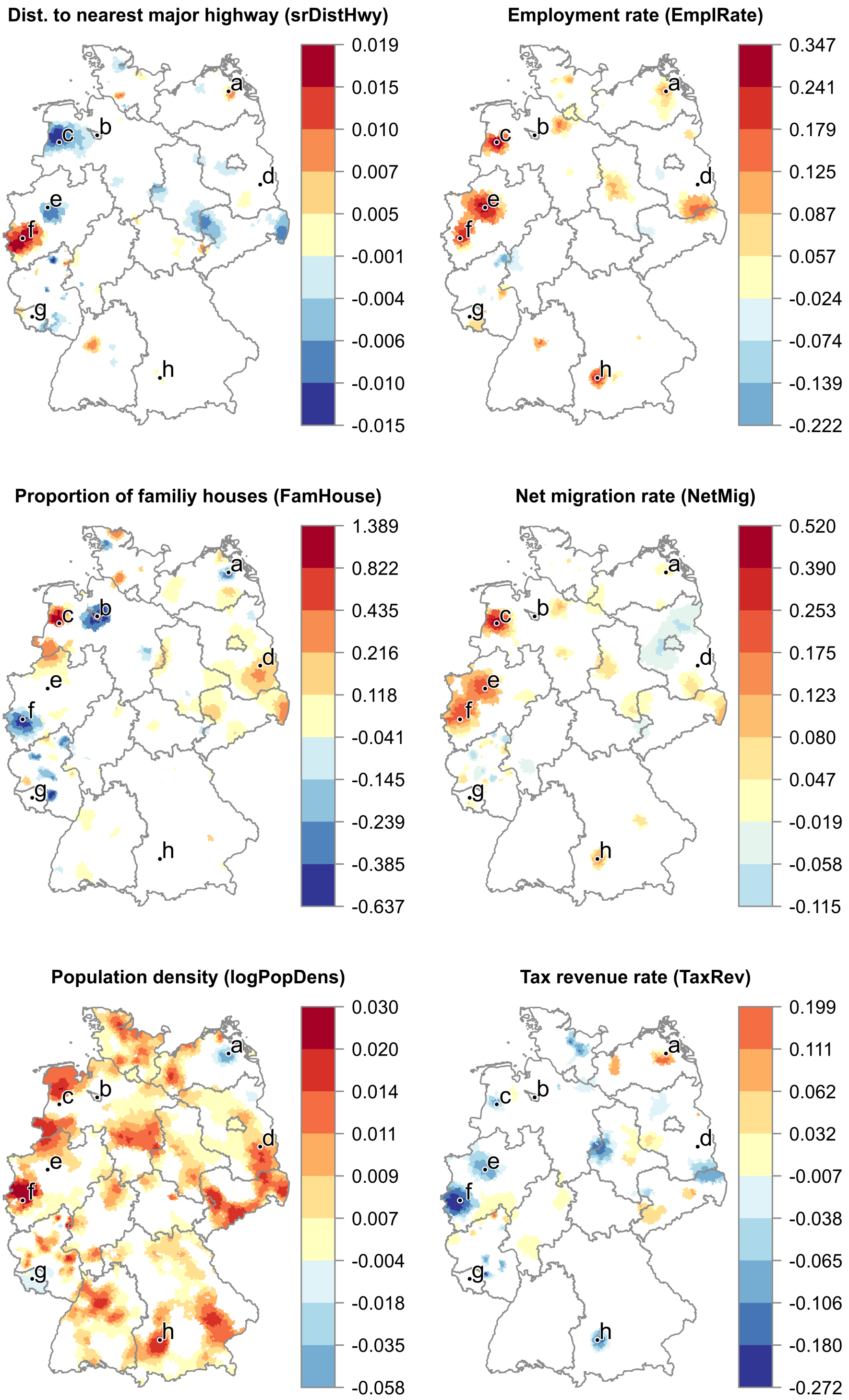


Figure 9.8.: Estimated local coefficients of GWR (areas colored in white refer to insignificant associations at the 0.05 level).

of family houses and LCR. Other coefficients are not significant for this region. In the south of Augsburg (h), RegioClust and GWR estimate a significant positive effect of employment rate, net migration rate, and population density on LCR, whereas the tax revenue rate has a negative effect. Distance to nearest major highway, however, is only weakly positive and statistically significant for GWR, while proportion of family houses shows a strong positive and significant association only for RegioClust.

The two algorithms often estimate different coefficients for some regions. For example, GWR estimates substantially strong and significant coefficients for most variables in the north of the district of Emsland (c), the district of Recklinghausen (e), and the north of the districts of Düren and Rhein–Erft Kreis (f). RegioClust, by contrast, does not estimate any significant coefficient for regions (c) and (e), and for region (f) only the variable tax revenue rate is significant. In the east of Germany, RegioClust identifies region (d), which comprises the municipalities between Frankfurt/Oder and Kolkwitz, close to Cottbus. For this region, RegioClust estimates a significant positive effect for employment rate, proportion of family houses, population density, and net migration rate, while it estimates a significant negative effect for tax revenue rate. For most municipalities in this region, GWR estimates positive coefficients (i.e., proportion of family houses and population density), whereas most drivers are not significant. Moreover, the distinct outline of the region identified by RegioClust is not identified by GWR.

9.3.4. Predictions of RegioClust and GWR

For the period 2010–20, RegioClust predicts that in Germany 71.011 ha of land will be consumed per day. This corresponds to a decrease of 6.444 ha per day compared to 2000–10. The predicted LCRs differ substantially between east and west Germany. For west Germany, RegioClust forecasts an LCR of 63.385 ha per day, whereas for east Germany the prediction is 7.626 ha per day. Thus, RegioClust predicts a decrease of 9.945 ha per day for west Germany and an increase of 2.318 ha per day for east Germany. The predictions differ when GWR is applied. GWR forecasts that 78.529 ha of land will be consumed per day for the entire country for 2010–20. This corresponds to an increase of 1.074 ha per day in comparison to 2000–10. Again, the LCRs differ substantially between west (73.516 ha per day) and east Germany (5.013 ha per day). This corresponds to an increase of 0.186 ha per day in west Germany and 0.887 ha per day in east Germany.

Figure 9.9 compares the predicted LCRs of RegioClust and GWR. It becomes apparent that both models predict a smaller increase in LCR in east than in west

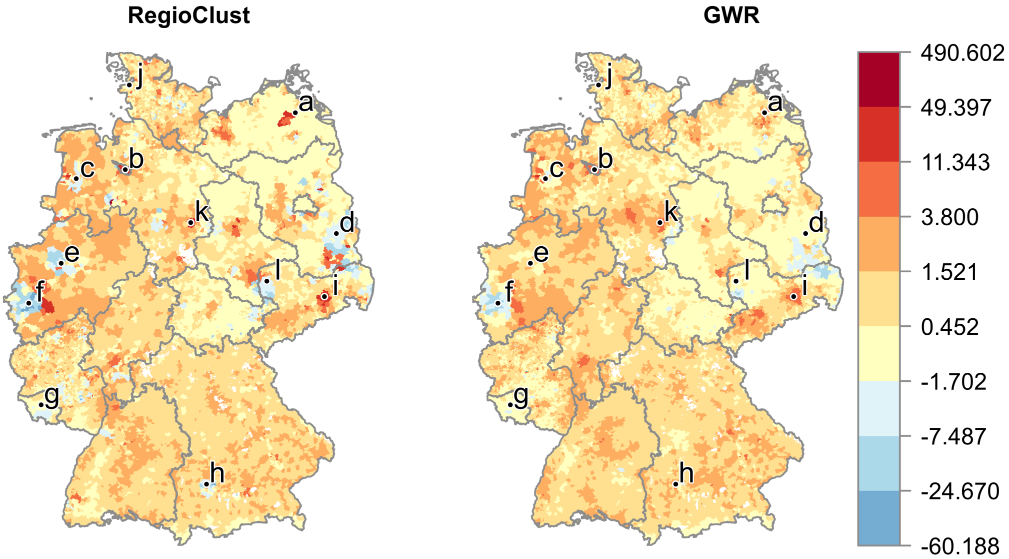


Figure 9.9.: Predicted LCRs in 2010–20 (in % of the municipalities' areas).

Germany. Saxony, consisting of the major cities of Dresden (i) and Leipzig, has the highest predicted increase in east Germany. In west Germany, the increase in land consumption is particularly low for the federal state of Saarland (g). In addition, the predicted increase tends to be higher for urban districts, such as Dresden (i), Braunschweig (k), and Bremen (b), than for rural ones. Some differences are also noticeable between RegioClust and GWR. RegioClust tends to predict more pronounced LCRs compared to GWR. This is observable for the Dresden region (i), the district of Düren (f), and the district of Dennim (a). Because unlike GWR RegioClust does not embody a distance-based smoothing, the predictions are often substantially different for adjacent regions. This is in particular noticeable for the south of Lübbenau (d), the district of Dennim (a), and the east of Rhein–Erft–Kreis, next to the district of Düren (f). For the municipalities between Recklinghausen and Dortmund (e) and south of Augsburg (h), RegioClust predicts negative LCRs whereas GWR predicts positive rates. A reverse effect appears, for example, for municipalities south of Lübbenau (d). In contrast to RegioClust, GWR is prone to making extreme predictions. For example, GWR predicts for Norderfriedrichskoog (j) an excessive increase in the LCR (+490.6%), while RegioClust predicts a moderate decrease (−12.7%).

9.4. Discussion

Numerous studies emphasize the socioeconomic and demographic differences between east and west Germany that translate into diverse patterns of land consumption (Kroll and Haase, 2010; Nuissl et al., 2009; Schmidt, 2011). The present findings are congruent with these studies and provide statistical evidence that the LCRs between east and west Germany differed for the period 2000–10. The predictions of RegioClust and GWR indicate that different LCRs can also be expected for the period 2010–20. In particular, both models refer to a further increase in the LCRs for east Germany. For west Germany, however, RegioClust predicts a substantial decrease, whereas GWR predicts a marginal increase in the LCRs. Both models agree that the predicted LCRs for the period 2010–20 exceeds the goal of the German federal government to reduce the LCR to 30 ha per day until 2020 (Die Bundesregierung, 2016), thus bringing into question the effectiveness of currently implemented planning policies to reduce the LCR (Kretschmer et al., 2015).

Due to the diverse pattern of LCRs across Germany, decision-makers are advised to formulate spatially tailored spatial planning strategies to counteract the predicted future increase in LCRs at both the municipal and the federal state level. Complementing a quantitative restriction of land consumption, Siedentop and Kausch (2004) propose qualitative regulations concerning the determination of locations for urban expansion. Such a strategy seems to support the prevention of the further expansion of built-up areas, which increases the transportation infrastructure demand in rural areas. Besides, brown field recycling and urban renewal seem to be feasible approaches to mitigating land consumption (Borchard, 2011; Schultz and Dosch, 2005).

The comparison of the performance of RegioClust and GWR showed that the former provides a better model fit with respect to the AICc. Even more important, the residuals of RegioClust exhibit less spatial dependence, which is essential for model estimation and inference (Ay et al., 2017; Brady and Irwin, 2011), compared to GWR. Heuristic optimization of the RegioClust's clusters has the potential to further improve model performance (Guo, 2009). However, it must be noted that the performance of both models depends on the selected parameters. RegioClust is affected by the parameters i and j. Generally speaking, low values of i and high values of j result in small but spatially homogeneous clusters with respect to the within-cluster sum of spatial distances between locations. This permits the modeling of small-scale spatial variation. Smaller clusters, however, also increase the risk of estimating linear models that only represent small-scale noise or are based on a low number of unrepresentative data points (local overfitting). High values of i and low values of j typically result in larger but spatially inhomogeneous clusters. Linear model estimated for larger clusters, however, may fail to represent regularities that exist only on a small scale (underfitting). To circumvent a subjective parameter selection, the parameters i and j of RegioClust are chosen by considering the AICc score, which is a well-established measure for this purpose (Symonds and Moussalli, 2011) and the number of outlying coefficients.

As a result of local overfitting, outlying coefficients indicate unstable local estimates, which hamper a meaningful interpretation and make predictions less reliable. However, outlying coefficients can also be an issue for GWR. This becomes apparent for the municipality of Norderfriedrichskoog (j). Because this municipality was a tax haven until 2004 for foreign and domestic enterprises (von Schwerin and Buettner, 2016), the tax revenue for the year 2000 was high in absolute terms compared to the adjacent rural municipalities. The high tax revenue did not substantially affect the coefficient estimation of RegioClust, because the municipality is part of a large cluster/region in which each municipality is considered similarly. In contrast, the optimized kernel width of GWR is too small to compensate for the high tax rate, which results in a biased estimate of the coefficient. As a consequence, when using data of 2010, where the tax revenue rate of the Norderfriedrichskoog was already increased due to the introduction of higher taxes in 2004, GWR substantially overestimates the LCRs, whereas the LCR estimates of RegioClust are still within a reasonable range.

This study was innovative in developing a new approach to modeling spatially varying coefficients. RegioClust has the strength to identify regions with a clearly defined boundary and non-volatile coefficients, which supports visual analysis and model interpretation. For instance, the outlined regions of RegioClust show that several urban municipalities are clearly separated from surrounding rural areas while others are not. This means that for some urban municipalities, the relationship between LCR and its drivers is similar to that of its neighboring rural municipalities. This finding is remarkable, because land consumption itself is generally considered to be different for urban and rural areas (Siedentop and Kausch, 2004). As all previous studies focused on selected regions (Bieling et al., 2013; Rienow and Goetzke, 2015), this was the first regional analysis dealing with land consumption and its drivers at the national level, not only retrospectively but also prospectively.

However, this study also had some limitations. Because aggregated data at the municipal level were used, bias due to the size and shape of the municipalities might have affected the results (Openshaw and Taylor, 1979). Although this problem is hardly avoidable, it is important to be aware of it when interpreting the results. Also, even though this study went beyond the usage of distance-based drivers only (Achmad et al., 2015; Arsanjani et al., 2013; Hu and Lo, 2007; Shafizadeh-Moghadam et al., 2017b) and considered socio-demographic, economic, and environmental variables as well, the whole spectrum of possible drivers is still to be explored (Kretschmer et al., 2015). In particular, the effect of other road types besides major highways or means of transportation on LCRs calls for further research. Finally, LCRs were considered only on a single spatial and temporal scale. Land consumption, however, is closely related to urbanization, which is known to occur on many spatial and temporal scales (Wu, 2007). A multi-scale analysis has the potential for more accurate modeling of LCRs and thus a better understanding of land consumption processes (Grant et al., 2015).

9.5. Conclusion

Germany is faced with exceptionally high LCRs, which are a challenge to sustainability. This study addressed this issue by examining the drivers of LCRs at the municipal level for the period 2000–10 and predicting rates for 2010–20. For this purpose, a new method for modeling spatial varying relationships, termed RegioClust, was developed. Empirical comparison indicated that RegioClust provides better model fits (i.e., AICc scores) than GWR, but tends toward minor local overfitting if parameters are not chosen appropriately. Both models provided clear evidence that LCR drivers vary substantially across Germany and that population density is of the utmost importance. For 2010–20, RegioClust and GWR predicted substantially different LCRs for east and west Germany. Most important, the forecasts provide evidence that the policy target of reducing the LCR to 30 ha per day in 2020 will not be achieved. In order to counteract this development, it is advised to revise local planning policies while intensifying brown field recycling and urban renewal, particularly in those municipalities that show an excess of LCRs for 2010–20.

Acknowledgements

Julian Hagenauer thanks Martin Behnisch from the Leibniz Institute of Ecological Urban and Regional Development for providing the data for this study.

Bibliography

Aburas, M. M., Ho, Y. M., Ramli, M. F., and Ash'aari, Z. H. (2017). Improving the capability of an integrated CA-Markov model to simulate spatio-temporal urban growth trends using an analytical hierarchy process and frequency ratio. *International Journal of Applied Earth Observation and Geoinformation*, 59, 65–78.

Achmad, A., Hasyim, S., Dahlan, B., and Aulia, D. N. (2015). Modeling of urban growth in tsunami-prone city using logistic regression: Analysis of Banda Aceh, Indonesia. *Applied Geography*, 62, 237–246.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1), 3–25.

Arsanjani, J. J., Helbich, M., Kainz, W., and Boloorani, A. D. (2013). Integration of logistic regression, Markov chain and cellular automata models to simulate urban expansion. *International Journal of Applied Earth Observation and Geoinformation*, 21, 265–275.

Ay, J.-S., Chakir, R., and Gallo, J. L. (2017). Aggregated versus individual land-use models: Modeling spatial autocorrelation to increase predictive accuracy. *Environmental Modeling & Assessment*, 22(2), 129–145.

Bagan, H. and Yamagata, Y. (2015). Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data. *GIScience & Remote Sensing*, 52(6), 765–780.

Basse, R. M., Charif, O., and Bodis, K. (2016). Spatial and temporal dimensions of land use change in cross border region of Luxembourg. Development of a hybrid approach integrating GIS, cellular automata and decision learning tree models. *Applied Geography*, 67, 94–108.

Bieling, C., Plieninger, T., and Schaich, H. (2013). Patterns and causes of land change: Empirical results and conceptual considerations derived from a case study in the Swabian Alb, Germany. *Land Use Policy*, 35, 192–203.

Bloom, D. E., Canning, D., and Fink, G. (2008). Urbanization and the wealth of nations. *Science*, 319(5864), 772–775.

Borchard, K. (2011). *Grundriss der Raumordnung und Raumentwicklung.* Verlag der Akademie für Raumforschung und Landesplanung.

Brady, M. and Irwin, E. (2011). Accounting for spatial effects in economic models of land use: recent developments and challenges ahead. *Environmental and Resource Economics*, 48(3), 487–509.

Brown, D. G., Walker, R., Manson, S., and Seto, K. (2004). Land-use and land-cover change. In: *Land Change Science: Observing, Monitoring and Understanding Trajectories of Change on the Earth's Surface*. Ed. by G. Gutman, A. C. Janetos, C. O. Justice, E. F. Moran, J. F. Mustard, R. R. Rindfuss, D. Skole, B. L. Turner, and M. A. Cochrane. Dordrecht: Springer Netherlands. Chap. Modeling land use and land cover change, 395–409.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4), 281–298.

Bundesamt für Kartographie und Geodäsie (2016). *Digitales Basis-Landschaftsmodell*. `http://www.geodatenzentrum.de/docpdf/basis-dlm-aaa.pdf`.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference. *Sociological Methods & Research*, 33(2), 261–304.

de Noronha Vaz, E., Nijkamp, P., Painho, M., and Caetano, M. (2012). A multi-scenario forecast of urban change: A study on urban growth in the Algarve. *Landscape and Urban Planning*, 104(2), 201–211.

Dendoncker, N., Rounsevell, M., and Bogaert, P. (2007). Spatial analysis and modelling of land use distributions in Belgium. *Computers, Environment and Urban Systems*, 31(2), 188–205.

Die Bundesregierung (2016). *Deutsche Nachhaltigkeitsstrategie*. `http://www.bundesregierung.de/Content/DE/_Anlagen/Nachhaltigkeit-wiederhergestellt/2017-01-11-nachhaltigkeitsstrategie.pdf?__blob=publicationFile&v=12`.

Dubovyk, O., Sliuzas, R., and Flacke, J. (2011). Spatio-temporal modelling of informal settlement development in Sancaktepe district, Istanbul, Turkey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(2), 235–246.

Duranton, G. and Turner, M. A. (2012). Urban growth and transportation. *The Review of Economic Studies*, 79(4), 1407–1440.

Fischer, M. M. (1980). Regional taxonomy: A comparison of some hierarchic and non-hierarchic strategies. *Regional Science and Urban Economics*, 10(4), 503–537.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowski, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., Prentice, I. C., Ramankutty, N., and Snyder, P. K. (2005). Global consequences of land use. *Science*, 309(5734), 570–574.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2003). *Geographically weighted regression: The analysis of spatially varying relationships.* John Wiley & Sons.

Fotheringham, A. S. and Oshan, T. M. (2016). Geographically weighted regression and multicollinearity: Dispelling the myth. *Journal of Geographical Systems*, 18(4), 303–329.

Grant, L. P., Gennings, C., and Wheeler, D. C. (2015). Selecting spatial scale of covariates in regression models of environmental exposures. *Cancer informatics*, 14, CIN–S17302.

Griffith, D. A. (2008). Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A*, 40(11), 2751–2769.

Guan, C. and Rowe, P. G. (2016). Should big cities grow? Scenario-based cellular automata urban growth modeling and policy applications. *Journal of Urban Management*, 5(2), 65–78.

Guo, D. (2009). Greedy optimization for contiguity-constrained hierarchical clustering. In: *IEEE International Conference on Data Mining Workshops.* IEEE, 591–596.

Haas, J. and Ban, Y. (2014). Urban growth and environmental impacts in Jing-Jin-Ji, the Yangtze, river delta and the Pearl river delta. *International Journal of Applied Earth Observation and Geoinformation*, 30, 42–55.

Haase, D., Kabisch, N., and Haase, A. (2013). Endless urban growth? On the mismatch of population, household and urban land area growth and its effects on the urban debate. *PloS one*, 8(6), e66531.

Haykin, S. S. (2009). *Neural networks and learning machines.* Neural networks and learning machines Bd. 10. Prentice Hall.

Helbich, M., Brunauer, W., Hagenauer, J., and Leitner, M. (2013). Data-driven regionalization of housing markets. *Annals of the Association of American Geographers*, 103(4), 871–889.

Helbich, M. and Griffith, D. A. (2016). Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches. *Computers, Environment and Urban Systems*, 57, 1–11.

Hennig, E. I., Schwick, C., Soukup, T., Orlitová, E., Kienast, F., and Jaeger, J. A. G. (2015). Multi-scale analysis of urban sprawl in Europe: Towards a European desprawling strategy. *Land Use Policy*, 49, 483–498.

Hu, Z. and Lo, C. P. (2007). Modeling urban growth in Atlanta using logistic regression. *Computers, Environment and Urban Systems*, 31(6), 667–688.

Iacono, M., Levinson, D., and El-Geneidy, A. (2008). Models of transportation and land use change: A guide to the territory. *CPL bibliography*, 22(4), 323–340.

Jakubowski, P. and Zarth, M. (2003). Nur noch 30 Hektar Flächenverbrauch pro Tag. *Raumforschung und Raumordnung*, 61(3), 185–197.

Kretschmer, O., Ultsch, A., and Behnisch, M. (2015). Towards an understanding of land consumption in Germany — outline of influential factors as a basis for multidimensional analyses. *Erdkunde*, 69(3), 267–279.

Kroll, F. and Haase, D. (2010). Does demographic change affect land use patterns?: A case study from Germany. *Land Use Policy*, 27(3), 726–737.

Ku, C.-A. (2016). Incorporating spatial regression model into cellular automata for simulating land use change. *Applied Geography*, 69, 1–9.

Li, X., Chen, Y., Liu, X., Xu, X., and Chen, G. (2017). Experiences and issues of using cellular automata for assisting urban and regional planning in China. *International Journal of Geographical Information Science*, 31(8), 1606–1629.

Lu, B., Harris, P., Charlton, M., and Brunsdon, C. (2014). The GWmodel R package: Further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-spatial Information Science*, 17(2), 85–101.

Luo, J. and Wei, Y. H. D. (2009). Modeling spatial variations of urban growth patterns in Chinese cities: the case of Nanjing. *Landscape and Urban Planning*, 91(2), 51–64.

Maimaitijiang, M., Ghulam, A., Sandoval, J. S. O., and Maimaitiyiming, M. (2015). Drivers of land cover and land use changes in St. Louis metropolitan area over the past 40 years characterized by remote sensing and census population data. *International Journal of Applied Earth Observation and Geoinformation*, 35, 161–174.

Malburg-Graf, B., Jany, A., Lilienthal, M., and Ulmer, F. (2007). Strategies and instruments to limit excessive land use in Germany — A proposal to the German Council for Sustainable Development. In: *Proceedings of the 2nd international conference on managing urban land*. Dresden, Germany.

McKinney, M. L. (2002). Urbanization, biodiversity, and conservation the impacts of urbanization on native species are poorly studied, but educating a highly urbanized human population about these impacts can greatly improve species conservation in all ecosystems. *BioScience*, 52(10), 883–890.

Meinel, G. and Schumacher, U. (2010). Konzept, Funktionalität und erste exemplarische Ergebnisse des Monitors der Siedlungs-und Freiraumentwicklung (IÖR-Monitor). In: *Flächennutzungsmonitoring II*. Rhombos-Verlag.

Murtagh, F. (1985). A survey of algorithms for contiguity-constrained clustering and related problems. *The Computer Journal*, 28(1), 82–88.

Nuissl, H., Haase, D., Lanzendorf, M., and Wittmer, H. (2009). Environmental impact assessment of urban land use transitions — A context-sensitive approach. *Land Use Policy*, 26(2), 414–424.

Nuissl, H. and Schroeter-Schlaack, C. (2009). On the economic approach to the containment of land consumption. *Environmental Science & Policy*, 12(3), 270–280.

Omrani, H., Tayyebi, A., and Pijanowski, B. (2017). Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based land transformation model: An integrated ML-CA-LTM modeling framework. *GIScience & Remote Sensing*, 54(3), 283–304.

Openshaw, S. and Taylor, P. J. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Sciences*, 21, 127–144.

Páez, A., Farber, S., and Wheeler, D. (2011). A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A*, 43(12), 2992–3010.

R Core Team (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria.

Rienow, A. and Goetzke, R. (2015). Supporting SLEUTH — Enhancing a cellular automaton with support vector machines for urban growth modeling. *Computers, Environment and Urban Systems*, 49, 66–81.

Ruß, G. and Kruse, R. (2011). Exploratory hierarchical clustering for management zone delineation in precision agriculture. In: *Industrial Conference on Data Mining.* Springer, 161–173.

Samardžić-Petrović, M., Dragićević, S., Kovačević, M., and Bajat, B. (2016). Modeling urban land use changes using support vector machines. *Transactions in GIS*, 20(5), 718–734.

Schmidt, S. (2011). Sprawl without growth in eastern Germany. *Urban Geography*, 32(1), 105–128.

Schultz, B. and Dosch, F. (2005). Trends der Siedlungsflächenentwicklung und ihre Steuerung in der Schweiz und Deutschland. *disP - The Planning Review*, 41(160), 5–15.

Seto, K. C., Güneralp, B., and Hutyra, L. R. (2012). Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences*, 109(40), 16083–16088.

Shafizadeh-Moghadam, H., Asghari, A., Taleai, M., Helbich, M., and Tayyebi, A. (2017a). Sensitivity analysis and accuracy assessment of the land transformation model using cellular automata. *GIScience & Remote Sensing*, in press(0), 1–18.

Shafizadeh-Moghadam, H., Asghari, A., Tayyebi, A., and Taleai, M. (2017b). Coupling machine learning, tree-based and statistical models with cellular automata to simulate urban growth. *Computers, Environment and Urban Systems*, 64, 297–308.

Shafizadeh-Moghadam, H. and Helbich, M. (2015). Spatiotemporal variability of urban growth factors: A global and local perspective on the megacity of Mumbai. *International Journal of Applied Earth Observation and Geoinformation*, 35, 187–198.

Siedentop, S. and Kausch, S. (2004). Die räumliche Struktur des Flächenverbrauchs in Deutschland. *Raumforschung und Raumordnung*, 62(1), 36–49.

Spielman, S. E. and Folch, D. C. (2015). Reducing uncertainty in the American Community Survey through data-driven regionalization. *PloS one*, 10(2), e0115626.

Symonds, M. R. E. and Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, 65(1), 13–21.

Triantakonstantis, D. and Mountrakis, G. (2012). Urban growth prediction: A review of computational models and human perceptions. *Journal of Geographjicl Information System*, 4, 555–587.

United Nations (2015). *World urbanization prospects: The 2014 revision.* `https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Report.pdf`. Department of Economic and Social Affairs, Population Division. (ST/ESA/SER.A/366).

Van Dessel, W., Van Rompaey, A., and Szilassi, P. (2011). Sensitivity analysis of logistic regression parameterization for land use and land cover probability estimation. *International Journal of Geographical Information Science*, 25(3), 489–508.

Veldkamp, A. and Lambin, E. F. (2001). Predicting land-use change. *Agriculture, Ecosystems & Environment*, 85(1). Predicting Land-Use Change, 1–6.

Verburg, P. H., Schot, P. P., Dijst, M. J., and Veldkamp, A. (2004). Land use change modelling: Current practice and research priorities. *GeoJournal*, 61(4), 309–324.

Vliet, J. van, Bregt, A. K., Brown, D. G., Delden, H. van, Heckbert, S., and Verburg, P. H. (2016). A review of current calibration and validation practices in land-change modeling. *Environmental Modelling & Software*, 82, 174–182.

von Schwerin, A. and Buettner, T. (2016). Constrained tax competition — Empirical effects of the minimum tax rate on the tax rate distribution. In: *Beiträge zur Jahrestagung des Vereins für Sozialpolitik 2016: Demographischer Wandel.* Session: Fiscal Competition G10-V1. Kiel und Hamburg: ZBW – Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

Wheeler, D. and Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7(2), 161–187.

Wu, J. (2007). Scale and scaling: A cross-disciplinary perspective. In: J. Wu and R. Hobbs (eds.). *Key Topics in Landscape Ecology.* Cambridge University Press, Cambridge, 115–142.

Zeng, C., Liu, Y., Stein, A., and Jiao, L. (2015). Characterization and spatial modeling of urban sprawl in the Wuhan metropolitan area, China. *International Journal of Applied Earth Observation and Geoinformation*, 34, 10–24.

10. Assessing the performance of 38 machine learning models: the case of land consumption rates in Bavaria, Germany

Authors

Julian Hagenauer, Hichem Omrani, and Marco Helbich

Journal

International journal of geographical information science

Status

Received 17 June 2018; accepted 3 February 2019

Contribution statement

Julian Hagenauer has developed the methods, designed the experiments, and has written the manuscript for the study. Marco Helbich and Hichem Omrani supported this publication by continuously discussing the design and results of the study and by proofreading the manuscript.

Abstract

Machine learning (ML) is at the forefront of land-use change modeling. Due to numerous available ML approaches, the model choice is complex and usually based on ad hoc decisions, though informed through a few comparative studies that considered a limited number of models. This study contributes a comprehensive comparison of 38 ML models to examine land consumption rates (LCR) (i.e., the transition of landscapes to built-up areas). We modeled LCR for 2009–2015 in Bavaria, Germany, and predicted rates for 2015–2021 at a municipality level. To assess the performance of each approach, we measured the mean absolute error (MAE), the root-mean-square error (RMSE), and the coefficient of determination (R^2) using crossvalidation. All algorithms consistently predicted that the land consumption rate for Bavaria will increase. eXtreme gradient boosting decision trees performed best with respect to the RMSE (0.500) and R^2 (0.183), while the support vector machine with polynomial kernel has the lowest MAE (0.288). The generalized additive model and the random forest models also performed well. We recommend these ML approaches for future land consumption and land-use change studies. A poor performance was found for recursive partitioning by decision trees, self-organizing maps, and partitioning using deletion, substitution, and addition moves.

Keywords: Land consumption; land-use; machine learning; model comparison; Germany

10.1. Introduction

Land-use mapping and modeling of environmental data using machine learning (ML) (Witten et al., 2016) have gained increasing interest within the geospatial community (Lary et al., 2016). It has become a vital methodology to(Heung et al., 2016; Omrani et al., 2019; Qian et al., 2014; Rogan et al., 2008), and forecast land-use change (Du et al., 2018; Hagenauer and Helbich, 2018; Samardžić-Petrović et al., 2016; Shafizadeh-Moghadam et al., 2017). ML comprises a set of inductive models that recognize patterns and/or minimize the prediction error of complex regression functions, by means of a repeated learning strategy from training data, linking an output such as land-use change to several underlying drivers. Once learned, the model can then be used to estimate previously unseen cases and predict future land-use change (Brown et al., 2013; Lary et al., 2016; Maxwell et al., 2018; Tayyebi and Pijanowski, 2014).

Statistical land-use models such as ordinary least squares regression or logistic binary regression are vital to understand land-use change processes and have thus gained popularity (Arsanjani et al., 2013; Lambin and Geist, 2008; Mustafa et al., 2018; Shafizadeh-Moghadam and Helbich, 2015; Veldkamp and Lambin, 2001). While these basic statistical models have a long tradition and have contributed to our understanding of how land-use change takes place across space and over time due to anthropogenic influences (Brown et al., 2004; Seto et al., 2012), the results are only reliable when model assumptions are fulfilled. These assumptions include that the drivers underlying landuse change are not correlated and that residuals (i.e., the difference between observed and predicted values) follow a specific distribution while being homoscedastic (i.e., having similar variance) (Fahrmeir et al., 2013). Of equal importance, these models are frequently restricted to linear associations (Arsanjani et al., 2013; Dubovyk et al., 2011), which seems a rather implausible assumption, as shown by a few studies (Tayyebi and Pijanowski, 2014). Although non-linearity can be incorporated through polynomial terms or splines (Fahrmeir et al., 2013), some prior knowledge about the underlying correlations is still required, though it is rarely available. Similarly, it seems of ultimate importance to consider not only the drivers' main effects but also the interactions between them (Pijanowski et al., 2002). Both issues can be addressed through ML.

Data-driven models from the field of ML have recently emerged as powerful alternatives to parametric land-use models, and they hold great promise (Du et al., 2018; Hagenauer and Helbich, 2012; Omrani et al., 2019, 2017; Rienow and Goetzke, 2015; Samardžić-Petrović et al., 2016; Shafizadeh-Moghadam et al., 2017; Tayyebi and Pijanowski, 2014). ML models have manifold benefits. For example, they are capable of dealing with massive amounts of data and a large number of variables, and they have the ability to model complex non-linear relationships as well as interactions between drivers, while not being grounded in restrictive distributional assumptions of the input data that are hard to achieve in practice (Witten et al., 2016).

The repertoire of ML is large (Fernández-Delgado et al., 2014; Lim et al., 2000), which makes the selection of a well-performing model a challenging task. However, frequently only a single approach is applied (Azari et al., 2016; Hagenauer and Helbich, 2012; B. Huang et al., 2010; Linard et al., 2013; Omrani et al., 2019; Samardžić-Petrović et al., 2016), often either artificial neural networks (S. S. Haykin, 2009), random forests (Breiman, 2001), and support vector machines (Scholkopf and Smola, 2001). To support an evidence-based algorithmic selection, a small number of land-use studies have compared multiple ML algorithms (Rogan et al., 2008; Samardžić-Petrović et al., 2017; Shafizadeh-Moghadam et al., 2017; Tayyebi and Pijanowski, 2014). While supportive, these studies assessed the performance of only a small number of algorithms. For example, Shafizadeh-Moghadam et al. (2017) compared six models to generate land-use transition probability maps. Multivariate adaptive regression splines and artificial neural networks were found to predict most accurately. Another study (Samardžić-Petrović et al., 2017) concluded that support vector machines outperform their competitors, including neural networks and logistic regression (B. Huang et al., 2010). While consensus is building that ML outperforms parametric models (e.g., regression), findings are less uniform and partly contradictory when ML models are contrasted. However, since many other ML models exist (Fernández-Delgado et al., 2014), more powerful and accurate approaches than the established ones might remain unacknowledged.

Previous studies formulate land-use change as a classification problem based on binary raster cells (Samardžić-Petrović et al., 2017; Shafizadeh-Moghadam et al., 2017; Tayyebi and Pijanowski, 2014), however, research dealing with land-use change from a regression-based perspective is virtually not existent. A special case of land-use change, which can analytically be explored on an area level, is the proportion of land consumption per municipality. Land consumption refers to the conversion of agricultural and forestry landscapes into built-up areas (Nuissl et al., 2009). As land consumption proceeds and causes an irreversible anthropogenic process (e.g., a loss of biodiversity, atmospheric pollution) (Bren d'Amour et al., 2016; Seto et al., 2012), it is of importance for policymakers to have the most accurate forecasts available in order to formulate sustainable planning policies.

To address these research gaps, a comprehensive comparison of numerous models is needed to evaluate the suitability of individual regression-based models and to identify their individual strengths and weaknesses. The present study rigorously cross-compared 38 ML models based on data on land consumption data in Bavaria, Germany. Our selected algorithms included tree-based models, artificial neural networks, support vector machines, linear regression models, nearest neighbor algorithms, and rule-based learners. The selection was guided by their proliferation in other domains. Yet, to the best of our knowledge, some of them had never been tested, and such an extensive model comparison had neither been conducted in the context of modeling land-use change nor modeling land consumption. The research question was as follows: Which algorithm predicts land consumption rates most accurately, while still being computationally moderately intense?

This article is structured as follows: Section 10.2 outlines the study area and the data; Section 10.3 introduces the methods; Section 10.4 summarizes the results; Section 10.5 discusses the findings in the context of the existing literature; and Section 10.6 draws conclusions

10.2. Materials

10.2.1. Study area

Germany faces one of the highest rates of land consumption of all EU member states (Kroll and Haase, 2010; Siedentop and Kausch, 2004). The federal government of Germany aims to limit the amount of land consumption to 30 ha per day up to the year 2020 (Bundesamt, 2012). This corresponds to a land consumption rate (LCR) of 0.43% of Germany's total area in six years. For this model comparison, we selected the federal state of Bavaria, Germany, as the study area. Its approximately 12.977 million inhabitants make it the second most populated federal state in Germany, and its $70,550$ km^2 make it the largest one. The area of Bavaria accounts for 19.47% of the total area of Germany.

10.2.2. Data

We analyzed LCR at a municipality level for 2009–2015. After removing unincorporated municipalities (e.g., Chiemsee and Veldensteiner Forst), the dataset included $2,056$ municipalities.[1] For each municipality, the LCR for both time periods were determined by dividing the difference between the consumed land (e.g., built-up areas, transportation infrastructure) in the beginning and end of the period by the total area of the municipality and multiplying it by 100 to obtain percentages. Data were

---

[1] `https://opendata.bayern.de/detailansicht/datensatz/verwaltungsgebiete`

extracted from the IÖR Monitor,[2] which is based on the digital landscape model ATKIS®-Basis-DLM (Meinel and Schumacher, 2010). A set of 10 demographic, socioeconomic, and environmental covariates were used as explanatory variables that were assumed to influence land-use change (Dubovyk et al., 2011; Hagenauer and Helbich, 2018; Kretschmer et al., 2015). Unless stated otherwise, the data were obtained from Germany's regional database (Regionaldatenbank Deutschland),[3] whose tables are based on the Regional Statistical Data Catalog and the Regional Statistical special program of the Federal Statistical Office and the statistical offices of the federal states. The demographic and socioeconomic variables were collected for 2009 and 2015. In line with urban theory (Duranton and Turner, 2012), the spatial distance (in km) from each municipality to the nearest regional center (Oberzentrum) was considered to reflect that accessibility stimulates urban growth. To incorporate the wealth per municipality which was found to be associated with urban growth (Bloom et al., 2008), employment rate (in %) served as a proxy variable. We used the in- and out-commuter rates (in %) as indicators to adjust for urbanization pressure through the population in- and outflows. As urban and rural municipalities may be affected differently, we controlled for the degree of urbanity by means of the logged population density (in $1,000$ people per km^2) (Hagenauer and Helbich, 2018). Data on the trade tax (in $1,000$€ per capita) were included to represent the municipalities' economic prosperity. Since the potential for further urban change decreases with the amount of existing built-up land, we considered the proportion of built-up area per municipality (in %). The variable was log transformed to make it Gaussian-like distributed. We included the mean of the terrain ruggedness index (Riley et al., 1999) to account for different building potentials due to topologic variation. Elevation data with a resolution of 90 m obtained from the Shuttle Radar Topology Mission served as the basis. Finally, the longitude and the latitude of the center of each municipality were included in order to account for the (unmeasured) locational characteristics (Arsanjani et al., 2013). Both longitude and latitude were scaled to zero and one, maintaining the aspect ratio. Table 10.1 provides descriptive statistics for all variables.

---

[2] https://monitor.ioer.de

[3] https://www.regionalstatistik.de

Table 10.1.: Descriptive statistics for the response and the covariates for 2009 and 2015.

Response variable	Min.	1st quartile	Median	Mean	3rd quartile	Max.	NAs
Land consumption rate for 2009–2015	−5.200	0.300	0.500	0.611	0.800	5.500	0
Covariates 2009							
Distance to regional center	0.000	8.780	15.600	16.800	23.800	54.400	0
In-commuter rate	0.238	0.611	0.685	0.682	0.759	0.952	27
Latitude	0.000	0.272	0.480	0.499	0.741	1.000	0
Longitude	0.000	0.298	0.447	0.464	0.627	0.980	0
Out-commuter rate	0.255	0.781	0.873	0.828	0.917	0.988	0
Population density (log)	0.766	1.830	2.000	2.080	2.260	3.630	0
Proportion of built-up area (log)	0.000	0.778	0.875	0.920	1.030	1.920	0
Terrain ruggedness (log)	2.938	11.931	15.896	18.525	21.567	112.918	0
Trade tax	−1.320	0.082	0.149	0.254	0.280	19.500	0
Unemployment rate	0.003	0.014	0.018	0.019	0.023	0.048	0
Covariates 2015							
Distance to regional center	0.000	8.780	15.600	16.800	23.800	54.400	0
In-commuter rate	0.228	0.634	0.702	0.701	0.774	0.964	22
Latitude	0.000	0.272	0.480	0.499	0.741	1.000	0
Longitude	0.000	0.298	0.447	0.464	0.627	0.980	0
Out-commuter rate	0.230	0.790	0.875	0.834	0.918	0.993	0
Population density (log)	0.817	1.830	2.010	2.080	2.270	3.670	0
Proportion of built-up area (log)	0.079	0.806	0.908	0.950	1.060	1.910	0
Terrain ruggedness (log)	2.938	11.931	15.896	18.525	21.567	112.918	0
Trade tax	−0.026	0.147	0.246	0.380	0.413	16.500	0
Unemployment rate	0.002	0.011	0.014	0.015	0.017	0.040	0

10.3. Methods

This section briefly describes each of the 38 ML models and their tunable parameters. All models have either already been used in land-use research or have shown promising results in other model comparisons (Fernández-Delgado et al., 2014; Xu et al., 2014).

All analysis steps were performed in the R programming environment (R Core Team, 2018) using the caret package (Kuhn, 2008). The latter is particularly useful for model comparisons as it harmonizes and streamlines the workflow for predictive models, parameter tuning, and validation, while providing a unified interface for each algorithm. A list of the tested models together with a short description is presented in Table 10.2. Note that the tested parameter values are given in the notation A:B:C, referring to the integer values A to B with a step size C. We also extended the default values of the tunable parameter values of caret to provide more comprehensive parameter tuning. The R code is provided as supplementary materials.

Table 10.2.: Summary of the tested models.

Label	Description	R package	Tested parameters	References	Model class
brnn	Bayesian regularized neural network	brnn	$\text{neurons} = 2^{(0:5:1)}$	Dan Foresee and Hagan (1997) and MacKay (1992)	NN
cforest	Random forest using conditional inference trees	party	$\text{mtry} = 1:10:1$	Strobl et al. (2008)	Bag,Tree
ctree	Conditional decision tree	party	$\text{mincriterion} = 0.8:0.99:0.01$	Hothorn et al. (2006)	Tree
cubist	Rule-based model that extends Quinlan's M5 model tree	cubist	$\text{committees} = 1:10:1$; $\text{neighbors} = 0:9:1$	Quinlan (1992, 1993a,b)	Boost, Tree, LM
earth	Multivariate adaptive regression splines	earth	$\text{Degree} = 1:4:1$; $\text{nprune} = 2^{(0:5:1)}$	J. H. Friedman (1991)	Spline
elm	Extreme learning machine	elmNN	$\text{actfun} = \text{radbas}, \text{sind}, \text{purelin}, \text{tansig}$; $\text{nhid} = 2^{(0:5:1)}$	G.-B. Huang et al. (2012)	NN
enet	Elastic Net	elasticnet	$\text{fraction} = 0:1:0.05$; $\text{lambda} = 10^{(-6:-1:1)}$	Zou and Hastie (2005)	LM
gam	Generalized additive model	mgcv	$\text{select} = \text{True}, \text{False}$; $\text{method} = \text{GCV.Cp}$	Hastie and Tibshirani (1986)	Add

Summary of the tested models.

Label	Description	R package	Tested parameters	References	Model class
gbm	Generalized boosting regression machine	gbm	n.trees = 500; interaction.depth = $4:16:4$, shrinkage = $0.01, 0.02, 0.05$; n.minobsnode = $4:32:4$	J. H. Friedman (2001)	Boost,Tree
gcvEarth	Earth using generalized cross-validation	earth	degree = $1:3:1$	J. H. Friedman (1991)	Spline
glm	Generalized linear model	stats		McCullagh and Nelder (1989)	LM
glmnet	Generalized linear model using penalized maximum likelihood with L1 and/or L2 regularization	glmnet	alpha = $0:1:0.05$; lambda = $10^{(-6:-1:1)}$	J. Friedman et al. (2010) and Simon et al. (2011)	LM
kknn	Weighted sum of k-nearest neighbors	kknn	kernel = gaussian, optimal, rank, inv; $k = 10:90:2$; distance = c	Hechenbichler and Schliep (2004) and Samworth (2012)	Avg
knn	Average of k-nearest neighbors	class	$k = 1:100:1$		Avg
lars	Least angle regression	lars	fraction = $0:1:0.05$	Efron et al. (2004)	LM

Summary of the tested models.

Label	Description	R package	Tested parameters	References	Model class
lasso	Linear regression with L1 regularization	elasticnet	fraction $= 0 : 1 : 0.05$	Zou and Hastie (2005)	LM
monmlp	Ensemble of neural networks with monotonic constraints	monmlp	hidden1 $= 2^{(0:5:1)}$; n.ensemble $= 1 : 4 : 1$	Lang (2005) and H. Zhang and Z. Zhang (1999)	Bag, NN
nnet	Neural network	nnet	size $= 2^{(0:5:1)}$; decay $= 10^{(-6:-1:1)}$	S. Haykin (2004)	NN
partDSA	Piecewise constant estimation sieve of candidate estimators based on a comprehensive search over the covariate space	partDSA	cut.off.growth $= 1 : 10 : 1$;MPD $= 0.1 : 1 : 0.1$	Molinaro et al. (2010)	Tree
pcaNNet	Neural network with principal component preprocessing	nnet	size $= 2^{(0:5:1)}$; decay $= 10^{(-6:-1:1)}$	Eleyan and Demirel (2005)	NN
pcr	Principal component regression	pls	n.comp $= 1 : 9 : 1$	Jolliffe (1982)	LM
penalized	Generalized linear model with L1 and/or L2 regularization	penalized		Goeman (2010)	LM
pls	Partial least squares regression	pls	ncomp $= 1 : 4 : 1, 8$	Wold et al. (1984)	LM

Summary of the tested models.

Label	Description	R package	Tested parameters	References	Model class
ppr	Projection pursuit regression	ppr	nterms = 1 : 9 : 1 J. H. Friedman and Stuetzle (1981)	Add	
ranger	Random forest	ranger	mtry = 1 : 10 : 1; min.obs.size = 5; splitrule = variance, extratrees	Wright and Ziegler (2017)	Bag, Tree
rf	Random forest	randomForest	mtry = 1 : 10 : 1	Breiman (2001)	Bag, Tree
ridge	Linear regression with L2 regularization	elasticnet	lambda = 0 : 1 : 0.05	Zou and Hastie (2005)	LM
rlm	Robust linear regression that uses Huber's M-estimators	MASS	intercept = True, False; psi = humber, hampel, bisquare	Huber (1981)	LM
rpart	Decision tree	rpart	complexity = 0.001 : 0.1 : 0.001 Breiman et al. (1984)	Tree	
rqnc	Non-convex penalized quantile regression	rqPen	penalty = SCAD, MCP; lambda = $10^{(-6:-1:1)}$	Fan and Li (2001) and Tibshirani (1996)	LM
spikeslab	Spike and slab model using a continuous bimodal prior	spikeslab	var = 1 : 9 : 1	Ishwaran et al. (2010)	LM
spls	Sparse partial least squares regression	spsl	k = 1 : 9 : 1; kappa = 0 : 05 : 01; eta = 0 : 1 : 0.005	Chun and Keleş (2010)	LM

Summary of the tested models.

Label	Description	R package	Tested parameters	References	Model class
svmLinear	Support vector machine with linear kernel	kernlab	$C = 0.05 : 1 : 0.05$	Cortes and Vapnik (1995) and Drucker et al. (1997)	SVM
svmPoly	Support vector machine with polynomial kernel	kernlab	degree $= 1 : 3 : 1$; scale $= 0.01, 0.05, 0.1, 0.2$; $C = 0.1 : 1 : 0.1$	Cortes and Vapnik (1995) and Drucker et al. (1997)	SVM
svmRadial	Support vector machine with radial basis kernel	kernlab	$C = 0.1 : 1 : 0.1$; sigma $= 0.01, 0.02, 0.05, 0.1, 0.15, 0.2$	Cortes and Vapnik (1995) and Drucker et al. (1997)	SVM
treebag	Bagging ensemble of decision trees	ipred		Breiman (1996) and Hothorn et al. (2005)	Bag, Tree

Summary of the tested models.

Label	Description	R package	Tested parameters	References	Model class
xgbTree	eXtreme gradient boosting decision trees	xgboost	eta = 0.02; gamma = 0; nrounds = 500; max$_{\mathrm{d}}$epth = 8 : 24 : 4; colsample$_{\mathrm{b}}$ytree = 0.3 : 0.7 : 0.2; subsample = 08 : 1.0 : 0.1, min$_{\mathrm{c}}$hild$_{\mathrm{w}}$eight = 0.8 : 1.0 : 0.1	Chen and Guestrin (2016)	Boost, Tree
xyf	Self-organizing map	kohonen	topo = hexagonal; xdim = $1, 2, 4, 6, 8, 12$; ydim = $1, 2, 4, 6, 8, 12$; user.weights = 0.2 : 0.8 : 0.2	Kohonen (2001)	Avg

LM = linear model, NN = neural network, Avg = local averaging, SVM = support vector machine, Tree = decision tree, Bag = bagging, Boost = bootstrap aggregating, Add = additive model, Spline = model using splines.

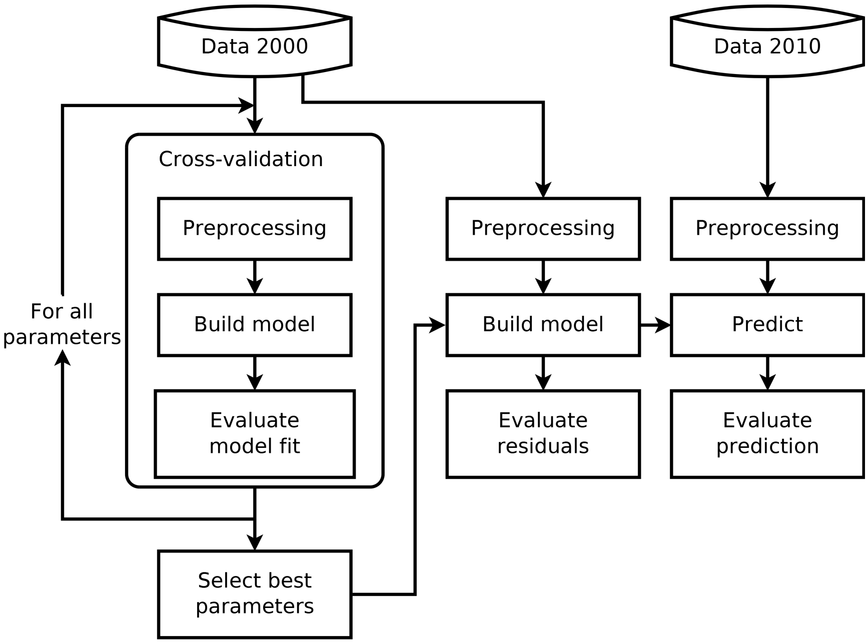


Figure 10.1.: Workflow of parameter selection, model building, and prediction.

Before the ML models were built, not available values (NAs) were imputed using a bagging ensemble of trees (Feelders, 1999). The input variables were then scaled to have zero mean and unit variance in order to make them comparable. The out-of-sample performance of the models (i.e., MAE, RMSE, and R2) with selected parameters was evaluated through 10-fold cross-validation (Kuhn, 2008). This procedure randomly partitions the data into 10 disjoint subsets. One subset at a time is then used for testing the model, while the remaining sets are used to build the model. This reduces the bias in performance estimation since the testing and training data sets are independent of each other (Kohavi, 1995). Cross-validation was repeated four times to lower the variance of the estimated performance and the results were averaged. We used the covariates of 2009 to perform cross-validation and determine the best parameters. Then, using the best parameters, the final model was built using the same data. Finally, the covariates of 2015 were applied to the model to predict LCRs for 2015–2021. Figure 10.1 summarizes the workflow.

To evaluate the different classifiers, the distribution of the performance statistics needed to be taken into account (Hothorn et al., 2005). We evaluated the statistical significance of the models' differences in performance as follows: Kruskal–Wallis tests with a 5% significance level were employed to test the null hypothesis that the performance estimates of the models are not systematically different from each other.

Two-sided Wilcoxon rank-sum tests were applied to determine the statistical significance of systematic pairwise differences between models. To control for a false discovery rate at the 5% level, the p-values were adjusted by means of the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

10.4. Results

Figure 10.2 summarizes the performance of the classifiers. The lowest mean MAE was achieved by *svmPoly* (0.288), while *xgbTree* had the highest mean R^2 (0.183) and the lowest mean RMSE (0.500). However, the differences between *xgbTree* and *svmPoly* in mean MAE were minor. The *partDSA* model generally achieved the lowest performance; its mean RMSE (0.535) and mean MAE (0.321) were higher and its mean R^2 (0.075) was lower than any other model.

The null hypothesis of no performance differences between the classifiers was rejected by the Kruskal–Wallis test at 5% significance level for MAE and R^2 but not for RMSE ($p < 0.05$). The null hypothesis of the Wilcoxon rank-sum of equal medians was rejected for many pairs of models for MAE and RMSE ($p < 0.05$); for RMSE, the null hypothesis was not rejected for any pair of models. With respect to RMSE and R^2, the null hypothesis of equal medians was not rejected for any pair of *gam*, *svmRadial*, and the tree-based ensemble models *rf*, *xgbTree*, *ranger*, *gbm*, and *cforest*. However, the null hypothesis was mostly rejected when comparing these models with linear models (e.g., *glm*, *lasso*, and *ridge*) or single tree models (e.g.,*rpart* and *ctree*), indicating significant performance differences. Detailed results are given in the supplementary materials (Table S1-S3).

As an example, Figure 10.3 depicts the residuals of the *xgbTree* model (left panel), which is among the best performing models and, for comparison purposes, the residuals of *nnet* (right panel). While in general the residual means were similar across the study area, some minor patterns were observable. For instance, the figure shows *xgbTree* and *nnet* both underestimate the LCR for the city of Straubing and its surroundings (a), while they overestimate the LCR for the city of Augsburg and its surroundings (b). However, despite the overall better performance of *xgbTree* (Figure 10.2), it underestimates the LCR for region (a) more than *nnet*. Furthermore, *xgbTree* underestimates and *nnet* overestimates the LCR for the municipality of Hof (c).

Figure 10.4 shows the predicted mean LCRs for 2015–2021 in percentage (left panel) and the absolute LCR for the same period in ha per day (right panel). The latter refers to an absolute measure that takes the municipality's area into account. The lowest mean LCR and absolute LCR were predicted by *cubist* (0.580% and 16.924 ha

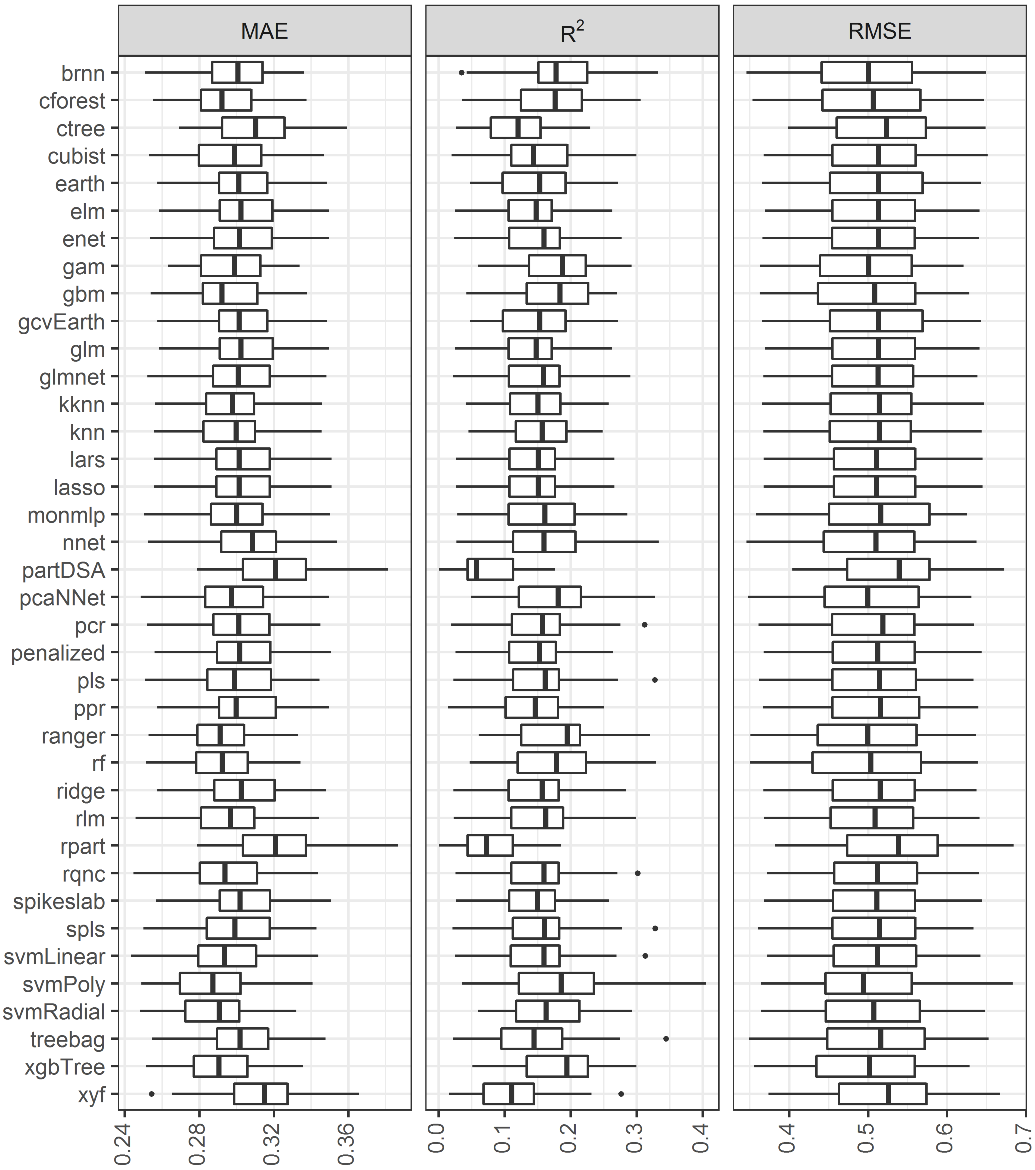


Figure 10.2.: Performance boxplots calculated from the cross-validation results.

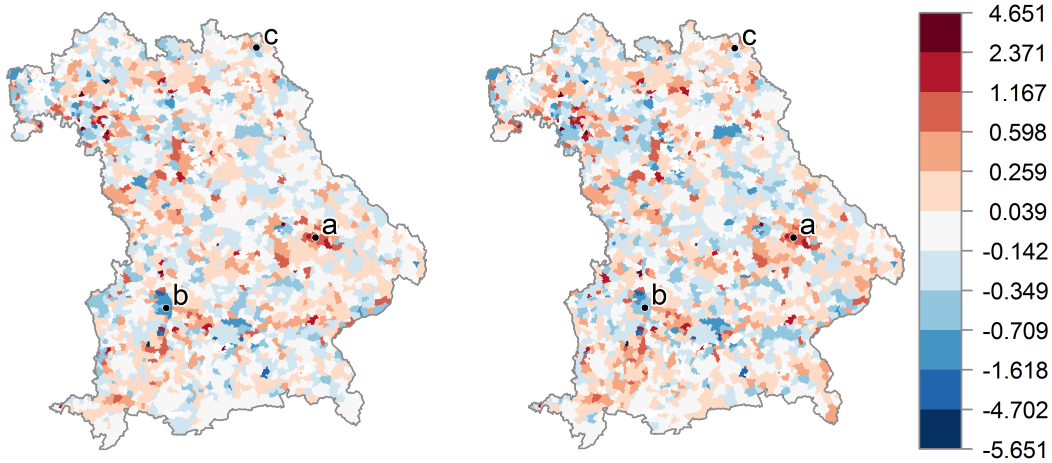


Figure 10.3.: Residuals of the *xgbTree* model (left) and *nnet* model (right).

per day). The highest mean LCR was predicted by *ridge* (0.646%), while the highest absolute LCR was predicted by *rpart* (19.056 ha per day).

For 2009–2015, the observed mean LCR for Bavaria was 0.611%, which corresponds to an absolute LCR of 17.444 ha per day. For 2015–2021, almost all models predicted an increase in mean LCR and absolute LCR compared to 2009–2015. Since Bavaria is the federal state of Germany with the largest area, this result seriously challenges the German policy goal to limit LCR to 30 ha per day until 2020 (Bundesamt, 2012). The few models that predicted a decrease in both measures were *cubist*, *rqnc*, and *svmLinear*, *svmRadial*, and *svmPoly*. A decrease in mean LCR and an increase in absolute LCR was predicted only by *rlm*.

Figure 10.5 compares the observed LCR for 2009–2015 (i.e., the response variable, left panel) and the predicted LCRs for 2015–2021 based on the *xgbTree* model (right panel). The model predicted for most municipalities an increase in LCR. A decrease in LCR was predicted for Straubing (a) and Nuremberg (c) and their vicinities, while a substantial increase was predicted for Augsburg (b) and its vicinity. Moreover, it can be seen that Augsburg's increase in LCR is much higher than that of Munich (d).

To show the influence of the different covariates on LCR, the importance of the variables for the *xgbTree* model is shown in Figure 10.6. Gain refers to the improvement in performance that is brought by a variable to the branches it is on. Cover measures the relative number of observations related to a variable. Frequency is the proportion of how many times a variable decides on a split in the trees. The figure shows that the present amount of built-up area, population density, and terrain ruggedness are the

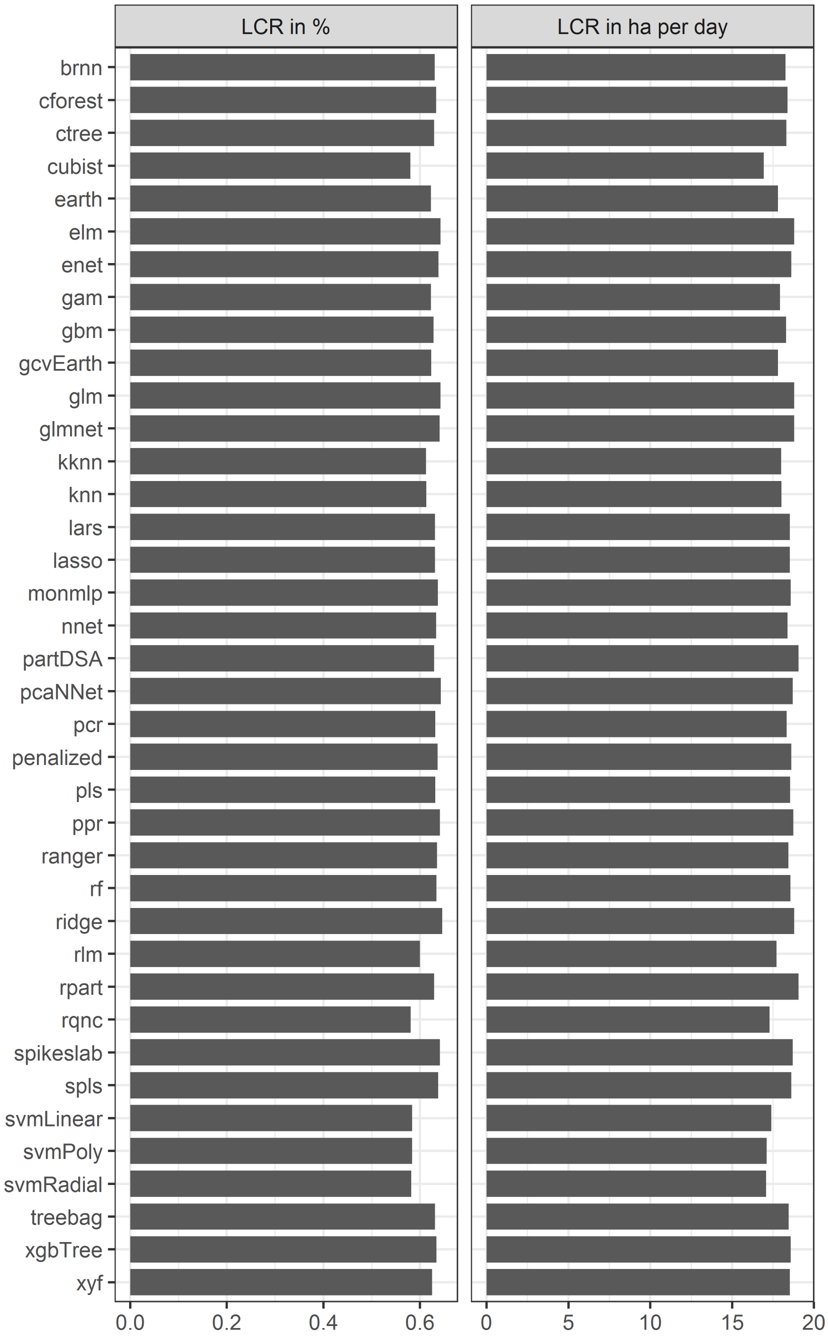


Figure 10.4.: Predicted mean LCRs for 2015–2021 in percentage (left) and absolute LCR for 2015–2021 in ha per day (right).

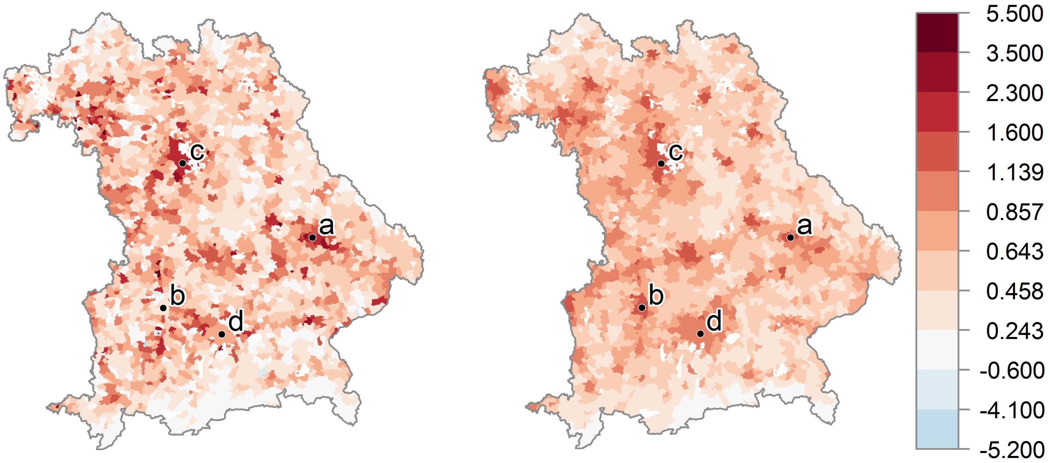


Figure 10.5.: Observed LCRs for 2009–2015 (left) and predicted LCRs for 2015–2021 using *xgbTree* (right).

most important variables, while the least important are employment rate and trade tax. Latitude seems to be more important than longitude.

Figure 10.7 shows one-way partial dependence plots (J. H. Friedman, 2001) for the four most relevant variables and a joint partial dependence plot for longitude and latitude (lower panel). Partial dependence plots visualize the change in the average predicted value as one or more variables vary over their marginal distribution (Goldstein et al., 2015). While all variables show a non-linear association with LCR, the strength of the association varies. The LCR increases with commuter rates substantially only from values higher than 0.85. Similarly, population density shows a positive association with LCR for values lower 2.0. While the LCR increases substantially with the proportion of built-up area, in particular from values 1.0 onwards, the LCR decreases marginally with terrain ruggedness. The LCR of municipalities close to the east of Upper Palatine (a) is associated with a lower LCR than municipalities that are far apart, in particular in the far south and north-west.

Finally, Figure 10.8 compares the required time to build the final models. All computations were performed on a standard laptop PC equipped with 32 GB of RAM and an Intel Core i7-6820HQ CPU @ 2.7 GHz. The most computational intensive was *cforest*, while the *rqnc* model was the fastest. As expected, ensemble models (e.g., *rf*, *treeBag*, *xgbTree*, monmlp) and support vector machine models generally took longer to build, while models based on linear regression were far less computationally intensive.

Among the boosting models, running *cubist* took less time than *gbm* or *xgbTree*.

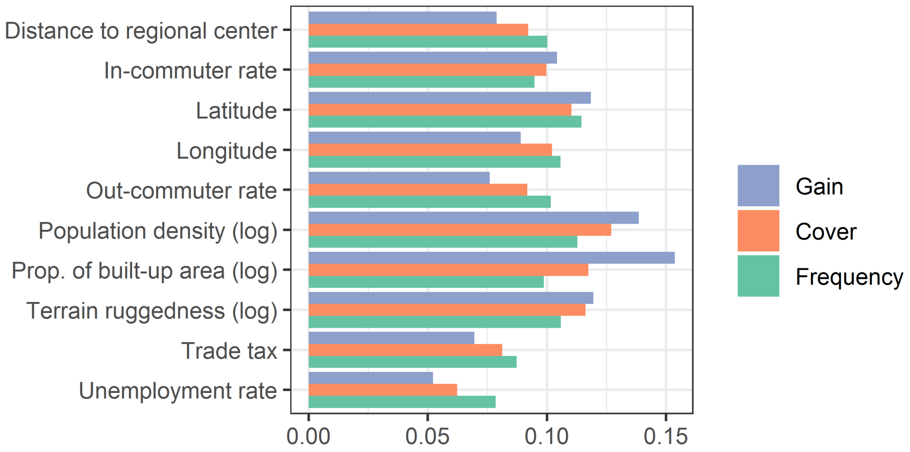


Figure 10.6.: Variable importance for *xgbTree*.

In addition, *ranger* was much faster than other random forest models, such as*rf*and *cforest*.

10.5. Discussion

Rapid progress in ML has resulted in numerous models, necessitating comparative studies to guide application-oriented research (Fernández-Delgado et al., 2014). This is particularly true for land-use change science. The present study has contributed, to the best of our knowledge, the most comprehensive comparison of regression-based ML algorithms for a continuous outcome variable, namely LCR.

Generally speaking, ML has performed well in several situations, including land-use modeling where the input data is usually multidimensional (Shafizadeh-Moghadam et al., 2017). Our findings confirm this conclusion by indicating that *xgbTree*, *gbm*, and*gam* achieved the highest predictive accuracy compared to the other 35 models. Although the *treeBag* algorithm is also a tree-based ensemble learner, it did substantially worse. We assume that this can be attributed to the tendency of *treeBag* to create correlated trees, which increases the upper bound for the generalization error (Breiman, 2001). An ensemble model performs better if there is significant diversity among the sub-models forming the ensemble (Kuncheva and Whitaker, 2003). Further, it is remarkable that *knn*, arguably one of the simplest models considered, performed well: It out-performed both single-tree models (e.g., *ctree* and*rpart*) and linear models (e.g., *lasso* and *glm*). This could imply that covariate interactions were minor, because, in contrast to*rpart*,

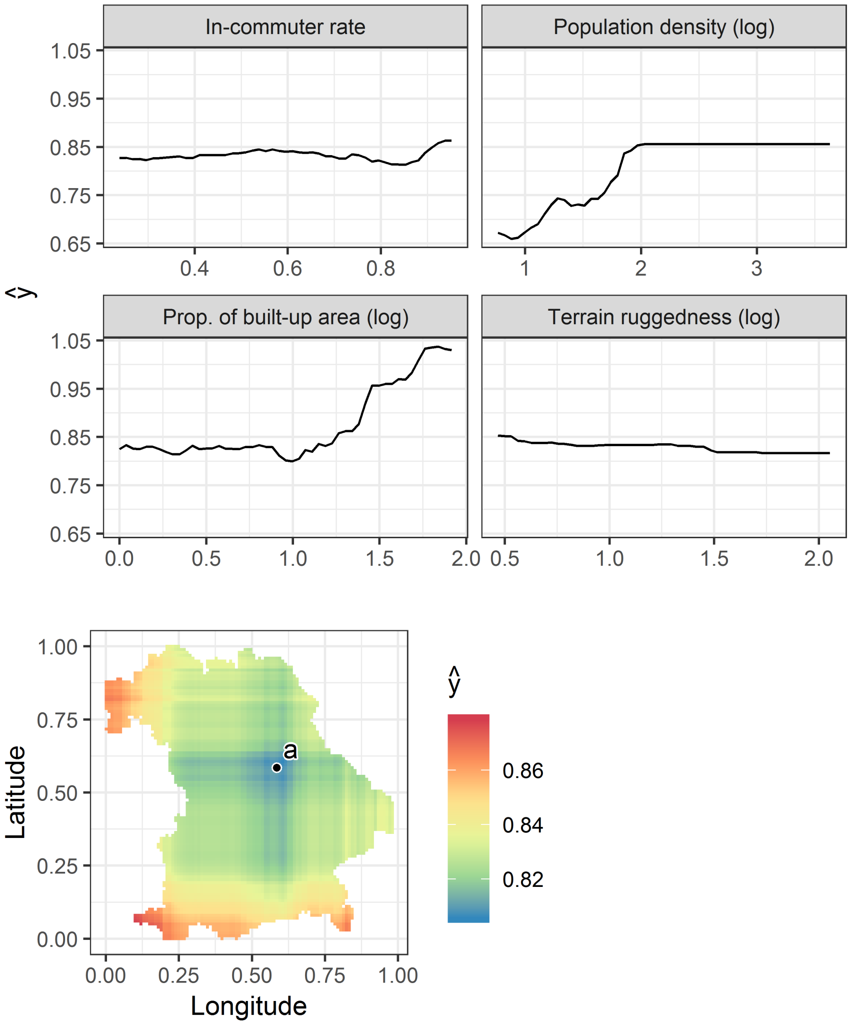


Figure 10.7.: Partial dependence plots for the most relevant variables.

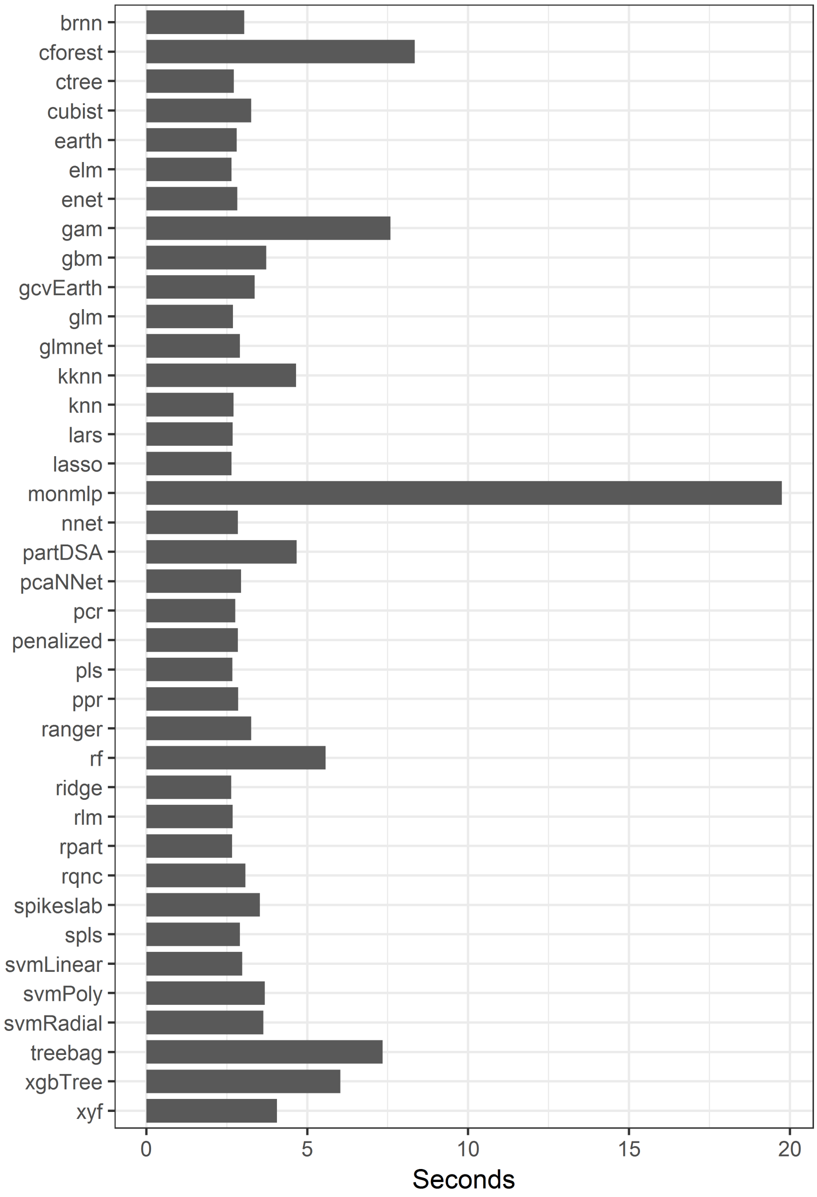


Figure 10.8.: Computational times to build the final models.

knn is not able to model interactions among variables. This hypothesis is also supported by the very good performance of *gam*, which also does not model interactions.

As we are not aware of a study similar to ours using LCR as an outcome, we discuss our findings in the general context of land-use modeling (Du et al., 2018; Samardžić-Petrović et al., 2017; Shafizadeh-Moghadam et al., 2017; Tayyebi and Pijanowski, 2014). While existing model comparisons consider rather similar models (i.e., artificial neural networks, multivariate adaptive regression splines, classification and regression trees), the literature is inconclusive about the performance of ML algorithms, although it does acknowledge that logit regressions (Arsanjani et al., 2013) are out-performed by ML (Shafizadeh-Moghadam et al., 2017). This is in line with our findings that generalized linear models had a weak predictive performance. The results were less clear when comparing different ML algorithms. For example, others (Tayyebi and Pijanowski, 2014) report that an artificial neural network provided higher accuracy, given limited model insights, than tree-based models and multivariate adaptive regression splines. Focusing on tree-based learners, Du et al. (2018) found that artificial neural networks are less suited and that extremely randomized trees perform best. In our case, artificial neural networks also had only an average accuracy. Confirming previous studies (Samardžić-Petrović et al., 2017), we also found that support vector machines (i.e., *svmRadial*) perform better than decision trees and artificial neural networks, though our study indicates that ensemble and boosting models can achieve better results. In another multi-label classification study (Omrani et al., 2015), *knn* was also found to be promising.

In total, no particular model in the reviewed ML studies stands out. Moreover, none of the studies made an effort to consider spatial autocorrelation (i.e., the fact that adjacent units are more similar than distant ones (Anselin, 2010)), as is done in econometric-based land-use models (Ay et al., 2017; Hagenauer and Helbich, 2018; Shafizadeh-Moghadam and Helbich, 2015). By considering the longitude and latitude per area, we attempted to adjust for unexplained locational effects. Research should prioritize exploring the impact of autocorrelation on ML models.

10.5.1. Strengths and limitations

Only a limited number of land-use studies have compared ML models, and a large majority of previous studies considered land-use change as a classification problem and used raster data as input (Shafizadeh-Moghadam et al., 2017). We have extended this body of knowledge towards regression-based approaches by means of modeling changes in land consumption. Another strength is that the set of tested algorithms (38 models) is significantly larger compared to available land-use change studies. This supports future ML-based studies in selecting a model that will potentially have a high predictive accuracy. Further, the models were challenged through a unified framework using optimized tuning parameters and repeating cross-validation to adjust for sampling fluctuations coupled with several goodness-of-fit metrics. In contrast to others (Du et al., 2018), we also compared the algorithms in terms of their computation time. We used not only distance-based variables, which are without doubt essential drivers, but also demographic, socioeconomic, and environmental variables. Finally, we used data available to the public and share the used code to ensure the reproducibility of our findings.

Despite these strengths, the following limitations need to be acknowledged. Most of these ML models use several parameters, which influences their performance. To circumvent subjective parameter selections beyond the default values, we systematically investigated values from a manually specified subspace. A more exhaustive parameter search is computationally expensive. Though done with care, we cannot entirely rule out that the optimal parameters are covered within the considered search space. As our focus relied on land consumption, an infrequently used but highly relevant outcome variable for spatial planning, it remains unclear how generalizable our results are for other land-use change outcomes. Further, data for all German municipalities were not available to us and we cannot rule out that other relevant explanatory variables were missing. Different study areas as well as diverse input data may have an impact on the model performance. We caution against blindly using a single algorithm; instead, we recommend testing at least a few of the ones that perform well. Despite this restriction, our results provide applied land scientists with essential insights into ML model performance. To address this restriction, we advise developing a benchmark data to support future model assessments. Because we used municipality-level data, we cannot rule out that the aggregation to municipalities affected the results (Openshaw and Taylor, 1979). Moreover, although our study grounds on the smallest territorial level for which data were available, the size of the tested data is moderate, as is the complexity through the number of considered covariates. Future studies should consider higher dimensional input data.

10.6. Conclusion

The selection of ML algorithms is not a straightforward task due to the large number of available alternatives. The present study performed a systematic and comprehensive comparison of 38 regression-based ML algorithms for modeling land consumption, using a variety of performance measures grounded on repeated 10-fold cross-validation.

While the ranking of the models depended to a minor extent on the consulted goodness-of-fit metrics when assessing land consumption data, our results showed that eXtreme gradient boosting decision trees (*xgbTree*) performed substantially better with respect to the RMSE (0.500) and R^2 (0.183), while the support vector machine with radial basis kernel (*svmPoly*) had the lowest MAE (0.288). Of similar importance, other frequently applied ML models in land-use science only performed either moderately (e.g., *earth*, *nnet*, and *glm*) or poorly (e.g.,*rpart* and *xyf*). Rarely used models such as partitioning using deletion, substitution, and addition moves (*partDSA*) and*rpart* did not perform well, independent of the performance measure. Due to their outstanding predictive performance, *xgbTree*, *gam*, and random forest models like *ranger* and *rf* seem to be a good initial choice when conducting a case study. However, we also suspect that they will not outperform the other models in all situations, and advise evaluating at least a few alternative models. Though not yet used, these well-performing ML approaches should play a major role in future land consumption and land-use studies to explore under which conditions they perform well and what results in a limited performance.

Acknowledgments

We thank the anonymous reviewers for their constructive comments that greatly improved the article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Bibliography

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1), 3–25.

Arsanjani, J. J., Helbich, M., Kainz, W., and Boloorani, A. D. (2013). Integration of logistic regression, Markov chain and cellular automata models to simulate urban expansion. *International Journal of Applied Earth Observation and Geoinformation*, 21, 265–275.

Ay, J.-S., Chakir, R., and Gallo, J. L. (2017). Aggregated versus individual land-use models: Modeling spatial autocorrelation to increase predictive accuracy. *Environmental Modeling& Assessment*, 22(2), 129–145.

Azari, M., Tayyebi, A., Helbich, M., and Reveshty, M. A. (2016). Integrating cellular automata, artificial neural network, and fuzzy set theory to simulate threatened orchards: Application to Maragheh, Iran. *GIScience& Remote Sensing*, 53(2), 183–205.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

Bloom, D. E., Canning, D., and Fink, G. (2008). Urbanization and the wealth of nations. *Science*, 319(5864), 772–775.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.* CRC press, 368.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Bren d'Amour, C., Reitsma, F., Baiocchi, G., Barthel, S., Güneralp, B., Erb, K.-H., Haberl, H., Creutzig, F., and Seto, K. C. (2016). Future urban land expansion and implications for global croplands. *Proceedings of the National Academy of Sciences*, 201606036.

Brown, D. G., Verburg, P. H., Pontius Jr, R. G., and Lange, M. D. (2013). Opportunities to improve impact, integration, and evaluation of land change models. *Current Opinion in Environmental Sustainability*, 5(5), 452–457.

Brown, D. G., Walker, R., Manson, S., and Seto, K. (2004). Modeling land use and land cover change. In: G. Gutman, A. C. Janetos, C. O. Justice, E. F. Moran, J. F. Mustard, R. R. Rindfuss, D. Skole, B. L. Turner, and M. A. Cochrane (eds.). *Land change science: Observing, monitoring and understanding trajectories of change on the earth's surface*. Dordrecht: Springer Netherlands, 395–409.

Bundesamt, S. (2012). *Nachhaltige Entwicklung in Deutschland — Indikatorenbericht 2012*. Tech. rep. Statistisches Bundesamt.

Chen, T. and Guestrin, C. (2016). XGBoost : Reliable large-scale tree boosting system. *arXiv*, 1–6.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 72(1), 3–25.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Dan Foresee, F. and Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian learning. In: *IEEE International Conference on Neural Networks - Conference Proceedings*. Vol. 3, 1930–1935.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9(x), 155–161.

Du, G., Shin, K. J., Yuan, L., and Managi, S. (2018). A comparative approach to modelling multiple urban land use changes using tree-based methods and cellular automata: the case of greater Tokyo area. *International Journal of Geographical Information Science*, 32(4), 757–782.

Dubovyk, O., Sliuzas, R., and Flacke, J. (2011). Spatio-temporal modelling of informal settlement development in Sancaktepe district, Istanbul, Turkey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(2), 235–246.

Duranton, G. and Turner, M. A. (2012). Urban growth and transportation. *Review of Economic Studies*, 79(4), 1407–1440.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., Loubes, J. M., Massart, P., Madigan, D., Ridgeway, G., Rosset, S., Zhu, J. I., Stine, R. A., Turlach, B. A., Weisberg, S., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2), 407–499.

Eleyan, A. and Demirel, H. (2005). Face recognition system based on PCA and feedforward neural networks. In: *Lecture Notes in Computer Science*. Vol. 3512.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, methods and applications*. Springer.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

Feelders, A. (1999). Handling missing data in trees: Surrogate splits or statistical imputation? *European Conference on Principles of Data Mining and Knowledge Discovery*, 329–34.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., and Amorim Fernández-Delgado, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15, 3133–3181.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1–67.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376), 817–823.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1), 70–84.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65.

Hagenauer, J. and Helbich, M. (2012). Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *International Journal of Geographical Information Science*, 26, 963–982.

Hagenauer, J. and Helbich, M. (2018). Local modelling of land consumption in Germany with RegioClust. *International Journal of Applied Earth Observation and Geoinformation*, 65, 46–56.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3), 297–310.

Haykin, S. S. (2009). *Neural networks and learning machines.* Bd. 10. Prentice Hall.

Haykin, S. (2004). A comprehensive foundation. *Neural Networks*, 2(2004), 41.

Hechenbichler, K. and Schliep, K. (2004). *Weighted k-nearest-neighbor techniques and ordinal classification.* Tech. rep. LMU Munich.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., and Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, 265, 62–77.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674.

Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3), 675–699.

Huang, B., Xie, C., and Tay, R. (2010). Support vector machines for urban growth modeling. *Geoinformatica*, 14(1), 83.

Huang, G.-B., Zhou, H., Ding, X., and Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2), 513–529.

Huber, P. J. (1981). *Robust statistics*. Vol. 82. 3. John Wiley & Sons, 308.

Ishwaran, H., Kogalur, U. B., and Rao, J. S. (2010). spikeslab: Prediction and variable selection using spike and slab regression. *The R Journal*, 2(2), 68–73.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, 31(3), 300.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence*. Vol. 2, 1137–1143.

Kohonen, T. (2001). *The self-organizing map*. Springer.

Kretschmer, O., Ultsch, A., and Behnisch, M. (2015). Towards an understanding of land consumption in Germany — Outline of influential factors as a basis for multidimensional analyses. *Erdkunde*, 69(3), 267–279.

Kroll, F. and Haase, D. (2010). Does demographic change affect land use patterns? A case study from Germany. *Land Use Policy*, 27(3), 726–737.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal Of Statistical Software*, 28(5), 1–26.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2), 181–207.

Lambin, E. F. and Geist, H. J. (2008). *Land-use and land-cover change: local processes and global impacts*. Springer.

Lang, B. (2005). Monotonic multi-layer perceptron networks as universal approximators. In: *15th International Conference on Artificial Neural Networks: Biological Inspirations*. Vol. 3697 LNCS, 31–37.

Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1), 3–10.

Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning*, 40(3), 203–228.

Linard, C., Tatem, A. J., and Gilbert, M. (2013). Modelling spatial patterns of urban growth in Africa. *Applied Geography*, 44, 23–32.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, 4(3), 415–447.

Maxwell, A. E., Warner, T. A., and Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9), 2784–2817.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall/CRC, 500.

Meinel, G. and Schumacher, U. (2010). Konzept, Funktionalitaet und erste exemplarische Ergebnisse des Monitors der Siedlungs-und Freiraumentwicklung (IOER-Monitor). In: *Flaechennutzungsmonitoring II*. Rhombos-Verlag.

Molinaro, A. M., Lostritto, K., and Laan, M. van der (2010). partDSA: Deletion/substitution/addition algorithm for partitioning the covariate space in prediction. *Bioinformatics*, 26(10), 1357–1363.

Mustafa, A., Heppenstall, A., Omrani, H., Saadi, I., Cools, M., and Teller, J. (2018). Modelling built-up expansion and densification with multinomial logistic regression, cellular automata and genetic algorithm. *Computers, Environment and Urban Systems*, 67, 147–156.

Nuissl, H., Haase, D., Lanzendorf, M., and Wittmer, H. (2009). Environmental impact assessment of urban land use transitions — A context-sensitive approach. *Land Use Policy*, 26(2), 414–424.

Omrani, H., Abdallah, F., Charif, O., and Longford, N. T. (2015). Multi-label class assignment in land-use modelling. *International Journal of Geographical Information Science*, 29(6), 1023–1041.

Omrani, H., Parmentier, B., Helbich, M., and Pijanowski, B. (2019). The land transformation model-cluster framework: Applying k-means and the Spark computing environment for large scale land change analytics. *Environmental Modelling& Software*, 111, 182–191.

Omrani, H., Tayyebi, A., and Pijanowski, B. (2017). Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based Land Transformation Model: an integrated ML-CA-LTM modeling framework. *GIScience & Remote Sensing*, 54(3), 283–304.

Openshaw, S. and Taylor, P. J. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Sciences*, 21, 127–144.

Pijanowski, B. C., Brown, D. G., Shellito, B. A., and Manik, G. A. (2002). Using neural networks and GIS to forecast land use changes: A land transformation model. *Computers, Environment and Urban Systems*, 26(6), 553–575.

Qian, Y., Zhou, W., Yan, J., Li, W., and Han, L. (2014). Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sensing*, 7(1), 153–168.

Quinlan, J. R. (1992). Learning with continuous classes. *Machine Learning*, 92, 343–348.

Quinlan, J. R. (1993a). *C4.5: Programs for machine learning.* Vol. 240. Morgan Kaufmann, 302.

Quinlan, J. R. (1993b). Combining instance-based and model-based learning. *Machine Learning*, 76, 236–243.

R Core Team (2018). *R: A language and environment for statistical computing.*

Rienow, A. and Goetzke, R. (2015). Supporting SLEUTH–Enhancing a cellular automaton with support vector machines for urban growth modeling. *Computers, Environment and Urban Systems*, 49, 66–81.

Riley, S. J., DeGloria, S. D., and Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences.*

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., and Roberts, D. (2008). Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sensing of Environment*, 112(5), 2272–2283.

Samardžić-Petrović, M., Dragićević, S., Kovačević, M., and Bajat, B. (2016). Modeling urban land use changes using support vector machines. *Transactions in GIS*, 20(5), 718–734.

Samardžić-Petrović, M., Kovačević, M., Bajat, B., and Dragićević, S. (2017). Machine learning techniques for modelling short term land-use change. *ISPRS International Journal of Geo-Information*, 6(12), 387.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Annals of Statistics*, 40(5), 2733–2763.

Scholkopf, B. and Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* MIT press.

Seto, K. C., Guneralp, B., and Hutyra, L. R. (2012). Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences*, 109(40), 16083–16088.

Shafizadeh-Moghadam, H., Asghari, A., Tayyebi, A., and Taleai, M. (2017). Coupling machine learning, tree-based and statistical models with cellular automata to simulate urban growth. *Computers, Environment and Urban Systems*, 64, 297–308.

Shafizadeh-Moghadam, H. and Helbich, M. (2015). Spatiotemporal variability of urban growth factors: A global and local perspective on the megacity of Mumbai. *International Journal of Applied Earth Observation and Geoinformation*, 35, 187–198.

Siedentop, S. and Kausch, S. (2004). Die räumliche Struktur des Flächenverbrauchs in Deutschland. *Raumforschung und Raumordnung*, 62(1), 36–49.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5).

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9.

Tayyebi, A. and Pijanowski, B. C. (2014). Modeling multiple land use changes using ANN, CART and MARS: Comparing tradeoffs in goodness of fit and explanatory power of data mining tools. *International Journal of Applied Earth Observation and Geoinformation*, 28, 102–116.

Tibshirani, R. (1996). Regression selection and shrinkage via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

Veldkamp, A. and Lambin, E. F. (2001). Predicting land-use change. *Agriculture, Ecosystems and Environment*, 85(1-3), 1–6.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. J. (1984). The collinearity problem in linear regression. The partial peast squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3), 735–743.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.

Xu, L., Li, J., and Brenning, A. (2014). A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery. *Remote Sensing of Environment*, 141, 14–23.

Zhang, H. and Zhang, Z. (1999). Feedforward networks with monotone constraints. *International Joint Conference on Neural Networks*, 3, 1820–1823.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(2), 301–320.

11. Relative importance of perceived physical and social neighborhood characteristics for depression: a machine learning approach

Authors

Marco Helbich, Julian Hagenauer, and Hannah Roberts

Journal

Social psychiatry and psychiatric epidemiology

Status

Received 30 May 2019; accepted 9 November 2019; published online 14 November 2019

Contribution statement

Marco Helbich developed the idea for this study and has written the manuscript in large parts. Julian Hagenauer developed the methods, designed the experiments, and wrote some parts of the manuscript. Hannah Roberts supported this publication by continuously discussing the results of the study and by proofreading the manuscript.

Abstract

Purpose The physical and social neighborhood environments are increasingly recognized as determinants for depression. There is little evidence on combined effects of multiple neighborhood characteristics and their importance. Our aim was (1) to examine associations between depression severity and multiple perceived neighborhood environments; and (2) to assess their relative importance.

Methods Cross-sectional data were drawn from a population-representative sample (N=9435) from the Netherlands. Depression severity was screened with the Patient Health Questionnaire (PHQ-9) and neighborhood perceptions were surveyed. Supervised machine learning models were employed to assess depression severity-perceived neighborhood environment associations.

Results We found indications that neighborhood social cohesion, pleasantness, and safety inversely correlate with PHQ-9 scores, while increasing perceived distance to green space and traffic were correlated positively. Perceived distance to blue space and urbanicity seemed uncorrelated. Young adults, low-income earners, low-educated, unemployed, and divorced persons were more likely to have higher PHQ-9 scores. Neighborhood characteristics appeared to be less important than personal attributes (e.g., age, marital and employment status). Results were robust across different ML models.

Conclusions This study suggested that the perceived social environment plays, independent of socio-demographics, a role in depression severity. Contrasted with person-level and social neighborhood characteristics, the prominence of the physical neighborhood environment should not be overstated.

Keywords: Depression; social neighborhood environment; physical neighborhood environment; the Netherlands

11.1. Introduction

It is gradually established that people's mental health is shaped, in addition to person-level attributes (Malhi and Mann, 2018), by the neighborhood environment, which can be broadly categorized into physical and social characteristics (Diez Roux and Mair, 2010). Since mental illness contributes 13% of disability adjusted life-years lost to the global burden of mental disorders (Vigo et al., 2016), it is necessary to understand how and to what extent the physical and social neighborhoods affect depression.

Recent reviews, mainly including cross-sectional and limited longitudinal evidence, suggest that socio-spatial aspects of people's living environment can contribute to or be protective against depression (Ehsan and De Silva, 2015; Rautio et al., 2018; Richardson et al., 2015). It was found that, for example, traffic-related air pollution (Gu et al., 2019), noise (Orban et al., 2016), safety concerns (Wilson-Genderson and Pruchno, 2013), and urbanicity (Purtle et al., 2019) were harmful for mental health because they are usually experienced as undesirable and stressful for residents which may, in turn, promote depressive mood (Gong et al., 2016; Rautio et al., 2018).

In contrast, it is theorized that green and blue space (Helbich et al., 2018), and social capital (Kawachi and L. F. Berkman, 2001) were beneficial because such factors may be stress-reducing and buffer against negative thoughts (Hartig et al., 2014), while neighborhood safety and social cohesion could act as coping mechanisms to safeguard from psychological distress (Ehsan and De Silva, 2015). Associations such as these are, however, not universally confirmed, and the mechanisms are yet to be fully understood.

Our present knowledge mainly originates from studies incorporating a single neighborhood characteristic (De Vries et al., 2016; Klijs et al., 2016; Zijlema et al., 2016), which may have resulted in misestimated neighborhood effects. In fact, as put forward by the socio-ecological model of health (Stokols, 1992), multiple physical and social neighborhood characteristics may be involved at the same time, implying a complex interplay. Therefore, when assessing correlations, either directly or in interaction with person-level attributes, it is rational to assume that multiple neighborhood characteristics may re-inforce or level-out each other. Supportive empirical evidence is, however, scarce (Generaal et al., 2019b; Groenewegen et al., 2018; Zock et al., 2018), as are studies that assess the relative importance of such characteristics in such a constellation (Gidlow et al., 2010).

Furthermore, it is suggested that neighborhood characteristics interact and are potentially non-linearly associated with depression (Helbich et al., 2018). Current insights from conventional (multilevel) regressions may be limited in that they assume correlations are linear (Generaal et al., 2019b; Groenewegen et al., 2018; Zock et al., 2018). This cannot be substantiated by theoretical considerations (Stokols, 1992), and potentially results in overly simplistic models which may contribute to contradictory findings.

These issues might be overcome through machine learning (ML) (Fernández-Delgado et al., 2014). ML includes a broad set of inductive models that learn to approximate unknown target functions from training data without being explicitly designed for a specific task. Echoing recent calls for methodological advances (Helbich et al., 2018), many of the models allow for non-linear correlations, routinely assess variable importance, and explore interactions between person-level and neighborhood characteristics.

Given the inconsistencies between and the methodological limitations of studies conducted to date, this large-scale explorative study in the Netherlands aimed (1) to examine the associations of physical and social neighborhood characteristics on people's depression severity; and (2) to assess the relative importance of people's perception of physical and social neighborhood characteristics on depression severity through ML approaches.

11.2. Materials and methods

11.2.1. Study setting and participants

This cross-sectional study reports on a nationally representative population sample in the Netherlands. In the course of the NEEDS project (Helbich, 2019a), an online survey was carried out with Statistics Netherlands between September and December 2018.

Participants needed to fulfill the following eligibility criteria: to be aged between 18 and 65 years, and living in a private household. Through systematic sampling with probabilities proportional to the target population size, sub-municipalities were first selected from each COROP region (i.e., a regional Dutch division). Next, from those regions individuals registered in the Dutch National Personal Records Database were randomly sampled. Incentives were offered to increase the response. Of those $45,000$ invited people, 11,524 completed the questionnaire resulting in an overall response rate of 25.6%. We conducted a complete case analysis excluding those with any missing information. After exclusions due to incomplete variables ($N = 2089$), the final sample size was $9,435$. A full description of the study protocol (Helbich, 2019a) and the survey is available (de Groot et al., 2018).

Ethical approval of the study design was obtained from the Ethics Review Board of the Faculty of Social and Behavioral Sciences of Utrecht University (FETC17–060).

Informed consent was implied by conducting the questionnaire.

11.2.2. Data

Our survey in the Dutch language comprised various modules including sociodemographics, mental health, and perception of the residential neighborhood. Other questions were asked but were not included here. The survey was further enriched with selected register data, namely urbanicity and income, available through Statistics Netherlands. If not mentioned otherwise, register data refer to 1st July 2018.

Severity of depression

Depression severity was operationalized through the depression module of the Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001). This instrument was recognized as having good diagnostic performance, good sensitivity, and good specificity in a meta-analyses (Gilbody et al., 2007). The 9-item long screener assesses people's experience within the last two weeks. The statements address whether a respondent felt down or depressed, had pleasure in doing things, had thoughts of suicide, etc. Each item is on a 4-point Likert scale ranging from "not at all" to "nearly every day". We summed the individual item scores per question. The total score was our outcome measure, assumed to be continuous. A PHQ-9 total score of one refers to no evidence of depressive symptoms, while 27 represents highest depression severity. The internal consistency of the PHQ-9 in our sample had a Cronbach's alpha of 0.887.

Physical neighborhood environment

Residential exposure to natural environments was measured twofold: first, we asked for the perceived distance to the nearest green space (defined as parks, play areas, sports fields or forests), and second, perceived distance to the nearest blue space (rivers, lakes, beaches). Distances were categorized into < 300 m, $300 - -1$ km, $> 1 - -5$ km, and > 5 km. These were in line with others (Reklaitiene et al., 2014; Stigsdotter et al., 2010).

To capture the perceived density of traffic, respondents were asked to evaluate traffic in their neighborhood based on their experiences in the past six months. The variable was on a 4-point Likert scale from "very busy/congestion" to "very quiet", with a greater higher score indicating less perceived density of traffic.

Pleasantness was operationalized using four questions from the ALPHA questionnaire (Spittaels et al., 2010). Respondents were asked to what extent they agree that the environment is pleasant for walking and cycling, the amount of incivilities (e.g., litter, graffiti) present, the number of trees in the street and the maintenance of buildings. Each question was answered on a 4-point Likert scale, with inverse scoring on negatively stated items. A greater score overall indicated a more pleasant neighborhood. The Cronbach's alpha was with 0.620 low. It is argued elsewhere (Spittaels et al., 2010) that for such environmental constructs a reduced Cronbach's alpha is acceptable because the involved indicators are often not intercorrelated.

Data on urbanicity refer to the urbanicity of the neighborhood ('buurt') of the person's home according to Statistics Netherlands. The variable was grouped into quintiles ranging from "not urban" (< 500 addresses/km^2) to "very strongly urban" (> 2500 addresses/km^2). Within this range, class breaks were set every additional 500 addresses/km^2.

Social neighborhood environment

To operationalize social cohesion, participants were asked to rate their agreement on a 5-point scale ranging from one (totally agree) to five (totally disagree), with the following statements (Sampson et al., 1997): 'People around here are willing to help their neighbors', 'I live in a cozy neighborhood, people in this neighborhood can be trusted', 'people in this neighborhood generally cannot get along so well', and 'people in this neighborhood do not share the same values'. Negatively stated items were inversely scored, with a higher score overall indicating greater social cohesion. The internal consistency was with a Cronbach's alpha of 0.829.

Questions on perceived safety were drawn from the "neighborhood safety" module within the ALPHA questionnaire (Spittaels et al., 2010). Participants rated their level of agreement with a set of five statements on a 4-point Likert scale from one (strongly disagree) to four (strongly agree). Statements included: 'It is dangerous to leave a bicycle locked in my neighborhood' and 'it is dangerous in my neighborhood during the day because of the level of crime'. Responses were reverse coded and then summed. The Cronbach's alpha was 0.822.

Covariates

The following routinely considered covariates were included (Generaal et al., 2019b; Lorant et al., 2003; Malhi and Mann, 2018): age (grouped into 5-year categories), gender (men, woman), ethnicity (Dutch, Western background, non-Western background), marital status (married, divorced, widowed, unmarried), employment status (employed, unemployed), and education (re-coded into low (up to lower secondary education), medium (up to upper secondary education), and high (university education and further)). Household income was obtained via Statistics Netherlands; the most recent data available are from 1st January 2016. The data were classified into quintiles (1 = lowest, 5 = highest).

11.2.3. Statistical analyses

Machine learning models

We undertook a supervised machine learning (ML) (Hastie et al., 2017) approach to assess associations between depression severity and neighborhood characteristics while adjusting for person-level attributes. Generally spoken, supervised ML models seek for patterns (i.e., complex relationships and interactions) in training data and use this information to conduct inference or predictions for unseen data without relying on strict model assumptions.

Since the repertoire of available regression-based ML algorithms is large and the performance may depend on the data set at hand (Fernández-Delgado et al., 2014), we selected well-established regression-based models. We fitted a generalized linear model (GLM) (McCullagh and Nelder, 1989) as base model, and the following three ML models: an artificial neural network (NNET) (Haykin, 1994), a random forest (RF) (Breiman, 2001), and a gradient boosting machine (GBM) (Friedman, 2001). For brief model descriptions, see the supplementary materials. Our model pre-selection was also guided by benchmark studies (Fernández-Delgado et al., 2014). Each model was fitted with depression severity as outcome variable and full covariate adjustment. All input variables were scaled to have zero mean and unit variance to make them comparable.

The goodness-of-fit of each model depends on the chosen hyper-parameters. We tested different settings from and evaluated the models' root-mean-square error (RMSE), the mean absolute error (MAE), and R^2 using 10 times repeated 10-fold cross-validation (CV). CV randomly partitions the data into 10 disjoint subsets. Subsets are used one at a time for model testing while the remaining ones are used for model building. Table A1 in the supplementary materials lists the final parameter settings. All analyses were carried out in the R programing environment (R Core Team, 2018) and the caret package (Kuhn, 2008) facilitated parameter tuning and model validation, while providing a unified interface for each algorithm.

Model interpretability

Different strategies were conducted for an in-depth model understanding. First, we assessed the variable importance relative to each other by permuting one variable at a time and measuring the change in performance (Breiman, 2001). To explore commonalities in variable importance, a heat map was generated. Second, to investigate the directions of the relationships and possible non-linearities, partial dependence plots were used. These plots show the change in the average predicted value as one or more variables vary over their marginal distribution (Goldstein et al., 2015). Third, through the H statistic we quantified either the total interaction of one variable with all others or the interaction of two variables (Friedman and Popescu, 2008). A value of zero means no interaction; one means that the entire variance is explained by the partial dependence functions.

11.3. Results

11.3.1. Sample description

Of $11,524$ participants, $9,435$ (81.8%) had complete data. The Mann–Whitney–Wilcoxon test confirmed that omitting survey respondents with incomplete information resulted in no significant differences ($p = 0.333$) in depression severity between the full and retained sample. Our sample had a mean PHQ-9 score of 4.857 with a standard deviation (SD) of 4.913. Table 11.1 summarizes the socio-demographic characteristics of our sample.

11.3.2. Model fits

Figure 11.1 shows the cross-validated model fits. The median magnitudes of model performances were, independent of the fit measure, rather similar. More specifically, the lowest median MAE and RMSE were achieved by GBM, while GLM had the highest errors. GBM also achieved the highest R^2, while GLM had the lowest. The R^2s were modestly high. Wilcoxon tests showed that the median performance of GBM was always significantly better than the one of GLM ($p < 0.050$). Generally, no significant difference in median performance were found for RF and NNET. The Table A2 in the supplementary materials lists the detailed test results.

11.3.3. Variable importance

The clustered heat map in Figure 11.2 shows two groups of variables with different levels of importance. The three most important variables for predicting depression severity were social cohesion, age, and employment status. Of minor importance were urbanicity, ethnicity, and perceived distance to green and blue spaces. Only minor

Table 11.1.: Sample characteristics.

Variable	Category	%	Min.	Max.	Mean	SD
PHQ-9			0.000	27.000	4.797	4.861
Physical neighborhood						
Green space	<300m	74.118				
	300m to 1km	21.293				
	>1km to 5km	4.240				
	>5km	0.350				
Blue space	<300m	21.791				
	300m to 1km	29.889				
	>1km to 5km	33.333				
	>5km	14.987				
Traffic	Very low	22.946				
	Low	43.932				
	High	29.995				
	Very high	3.127				
Pleasantness			4.000	16.000	13.513	2.009
Urbanicity	Very low	9.274				
	Low	24.123				
	Middle	16.789				
	High	30.069				
	Very high	19.746				
Social neighborhood						
Safety			6.000	24.000	19.912	2.947
Social cohesion			5.000	25.000	19.388	3.250
Other						
Age			18.000	65.000	44.904	13.815
Education	Low	19.194				
	Middle	35.601				
	High	45.204				
Employment	Unemployed	24.759				
Ethnicity	Dutch	87.790				
	Western	7.451				
	Non-Western	4.759				
Gender	Male	47.843				
Income	Very low	9.126				
	Low	10.429				
	Middle	18.654				
	High	26.762				
	Very high	35.029				
Marital status	Married	53.672				
	Divorced	8.511				
	Widowed	1.526				
	Unmarried	36.290				

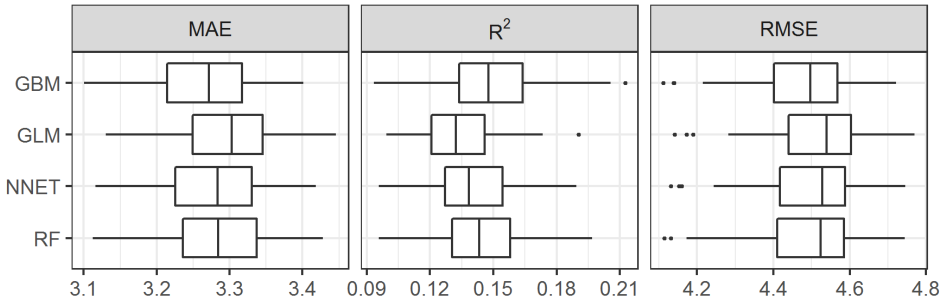


Figure 11.1.: Summary of the crossvalidated model fits. RMSE=root-mean-square error, MAE=mean absolute error.

changes appeared across the models.

11.3.4. Correlation assessments

Correlations between PHQ-9 and the person-level and neighborhood characteristics are displayed in Figure 11.3. As before, model differences were mostly small. Perceived safety and pleasantness were both negatively, and roughly linearly, associated with PHQ-9 scores. Social cohesion was inversely correlated with PHQ-9 scores. Perceived traffic was positively correlated with PHQ-9 scores. Both RF and GBM suggested that perceived distance to green space was positively correlated with PHQ-9 scores. Yet, the results for the distances > 5 km are inconclusive and diverge across the models. However, the category perceived distance to green space > 5 km was sparsely populated (11.1). Blue space seemed to be uncorrelated with PHQ-9; as was urbanicity. Unemployed, female, and divorced people showed higher PHQ-9 scores. PHQ-9 scores were substantially higher also for low earners and lower educated people. No differences were observed across ethnicities.

11.3.5. Variable interactions

Figure 11.4 shows pronounced overall variable interactions for social cohesion, age, employment, and education; neighborhood characteristics showed only little interaction.

Further, Figures A1 and A2 detail with which variables social cohesion and age interact most (e.g., education, employment), which is the basis for Figure 11.5 showing the bivariate interactions of these variables. The effect of employment status on the PHQ-9 scores varied over age, with unemployed persons always having a higher risk.

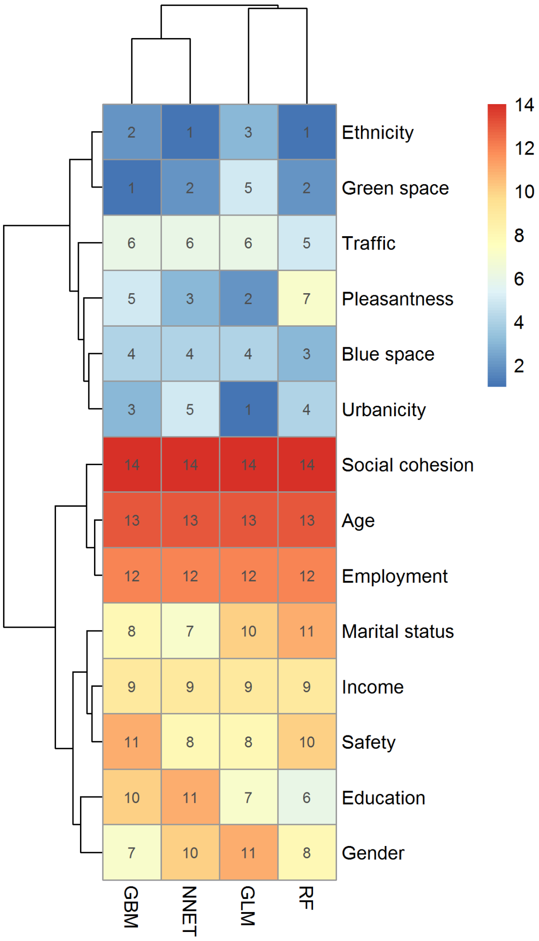


Figure 11.2.: Clustered heat map of the variable importance. The number per cell refers to the variable rank. Higher ranks indicate more important variables.

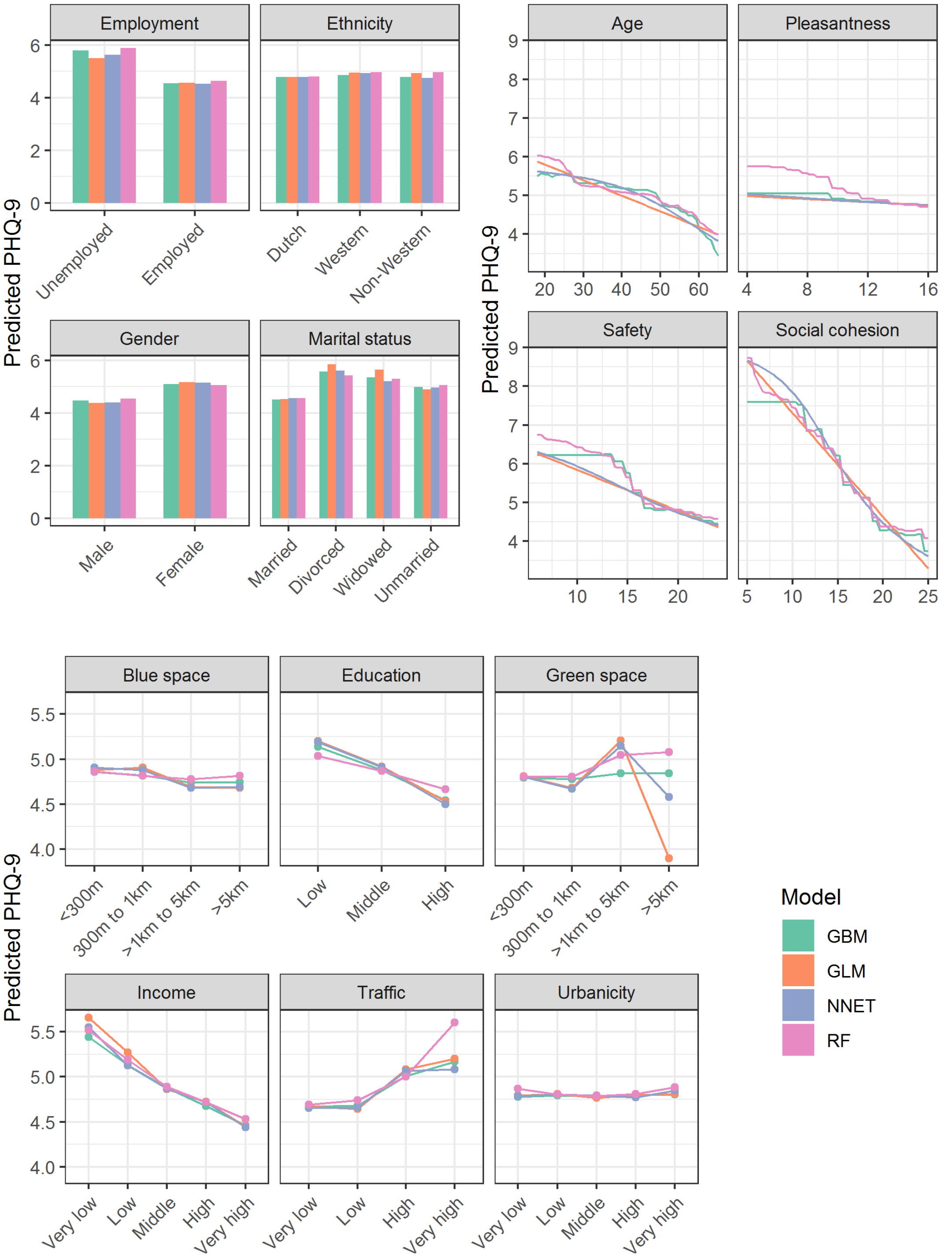


Figure 11.3.: Partial dependence plots relating each predictor to the PHQ-9 scores. Models are based on full covariate adjustment.

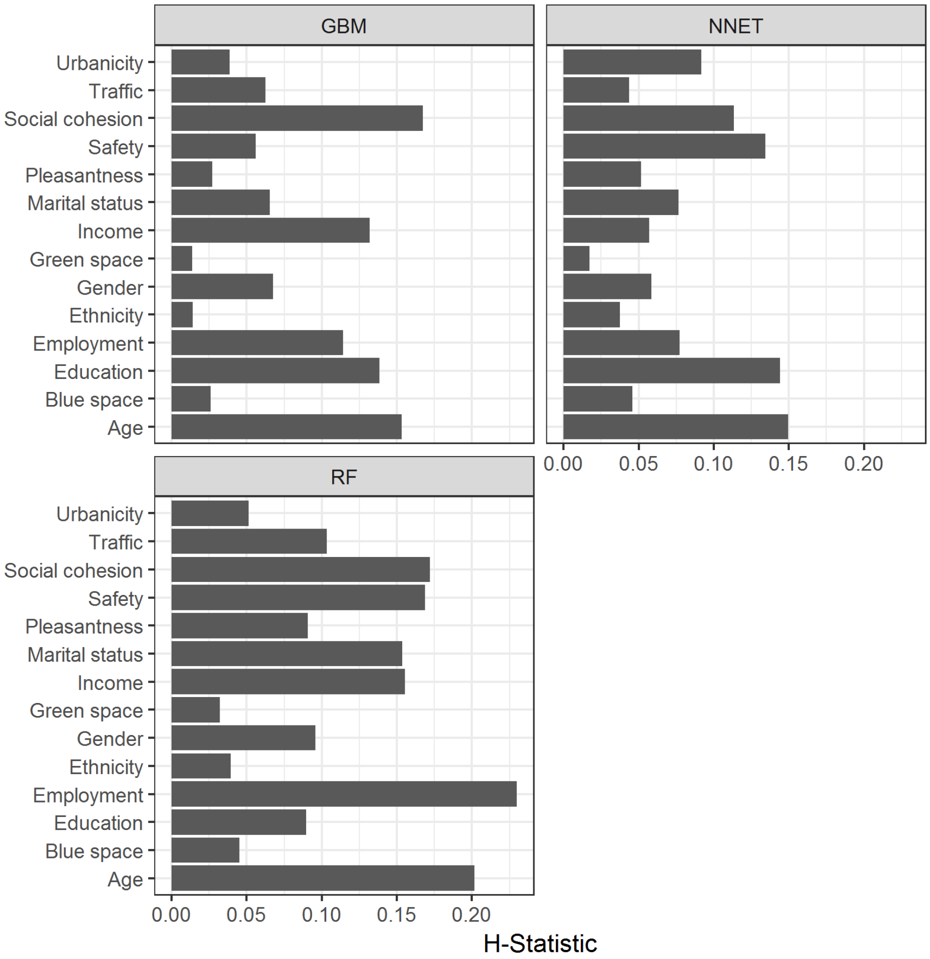


Figure 11.4.: Overall variable interaction (note that the H statistic is not available for GLMs).

Between 30 and 50 years of age, the differences were most pronounced; this gap in risk shrinks from 50 years onwards. Independent of the model, respondents that were less well educated and younger than 30 years old had pronounced PHQ-9 scores. With increasing age, the differences between lower and higher educated groups aligned each other.

Similar patterns were observed for social cohesion. High education and income were associated with lower PHQ-9 scores. However, while the differences in PHQ-9 scores between education groups decreased with social cohesion score, a decrease in the differences of the PHQ-9 scores between income groups was not always observable. For instance, there was always a notable gap in PHQ-9 scores between low and middle income groups, regardless of the social cohesion score. Moreover, when comparing the different models, it can be seen that for GBM, no decrease in PHQ-9 scores can be observed for a social cohesion score from 5 to 11.

11.4. Discussion

11.4.1. Main findings in the context of available evidence

This study assessed how multiple physical and social neighborhood characteristics together are correlated with depression severity after adjusting for individual sociodemographic factors, using a ML approach. The four models fitted on our large-scale data resulted in robust evidence that demonstrates which perceived neighborhood characteristics are cross-sectionally correlated with depression severity. All ML models showed a better fit than basic regression, however, the differences were more of a statistical nature (Figure 11.1).

We went a step further than previous studies (Generaal et al., 2019b; Groenewegen et al., 2018; Zock et al., 2018) in also assessing the relative importance of individuals' perceptions of the physical and social neighborhood with respect to depression severity. Our models consistently showed that perceived physical neighborhood environment only played a minor role in explaining depression severity (Figure 11.2). In contrast, social cohesion and safety were found to be important overall. Our result that the neighborhood social environment is of greater importance than the physical one replicates a study from the UK (Gidlow et al., 2010).

In line with a systematic review (Jorm, 2000), we observed a negative relation between depression and age. It seems that older aged people's susceptibility to depression declines which could result from diminishing emotional responsiveness or psychological immunization against stressful situations (Gross, 1998). Moreover, age was found to

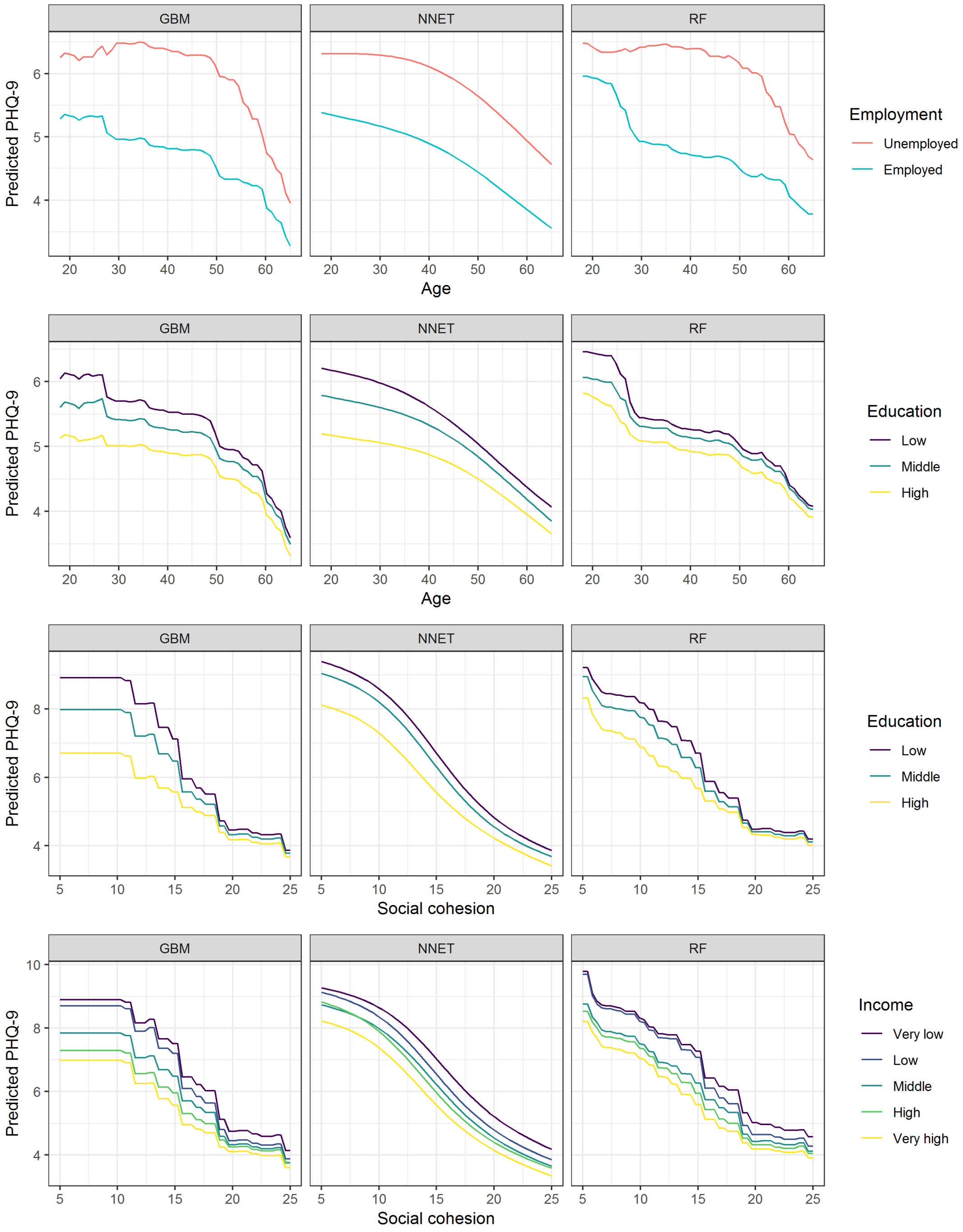


Figure 11.5.: Two-dimensional partial dependence plots relating predictors to the outcome PHQ-9 scores across the ML models. Only variables with a pronounced interaction are shown. Models are based on full covariate adjustment.

interact strongly with other variables, primarily personal-level (e.g., employment status (Jefferis et al., 2011)) and to a minor extent with environmental ones (e.g., perceived green space) (Figure A1), over the life span. Such a co-variation (Jorm, 2000) is not surprising because, for instance, unemployment may pose a higher risk for a young adult than someone close to retirement.

Perceived physical neighborhood characteristics including green and blue space, pleasantness, and urbanicity were found to be less important. This may partly be due to the way we assessed neighborhood features; some variables (e.g., green space) also showed limited variance. To circumvent methodological issues we employed, as frequently done (Gidlow et al., 2010; Ruijsbroek et al., 2017), people's neighborhood perceptions instead of geographic information system (GIS)-based measures per administrative area or buffer. Both ways cause spatial (Kwan, 2012) and temporal context uncertainties (i.e., temporally ill-aligned GIS and survey data) (Helbich, 2019b) potentially translating into biased outcomes. Work undertaken in metropolitan Chicago found that perceived but not objectively measured neighborhood deterioration was correlated with higher depressive symptoms, which further supports our reasoning (Wilbur et al., 2009).

Some neighborhood characteristics were identified as relevant, but not all turned out to be related to depression severity (Figure 11.3). In what follows the neighborhood characteristics are discussed in accordance to their descending order of importance (Figure 11.2). First, our study supports previous findings suggesting that pronounced neighborhood social cohesion seems to correlate with reduced depression severity (Mair et al., 2009; Ruijsbroek et al., 2017). It is assumed that in socially cohesive neighborhoods it is more likely that people help, support, and trust each other, and that a tightly knit social network may facilitate the spread of information among neighbors (Kawachi and L. Berkman, 2000). Through such pathways living in a cohesive environment may promote mental health.

Second, neighborhood safety was confirmed in our study to be negatively associated with depression severity. Another Dutch cross-sectional study has concluded the same (Generaal et al., 2019b), but overall findings are inconclusive (Lorenc et al., 2012). Among different conceivable mechanisms, we speculate that living in an unsafe neighborhood enhances experienced stress, which in turn is a depression risk factor (Chandola, 2001). Alternatively, it has been theorized that a lack of safety limits social cohesion due to mistrusting others in the neighborhood (Kawachi and L. Berkman, 2000).

Third, perceived traffic appeared to be positively correlated with depression severity. While our data did not allow us to disentangle pollutants emitted from traffic, we believe that air pollution and noise are conceivable underlying pathways. This is underpinned by a meta-analysis on air pollution and risk of depression (Gu et al., 2019), but contradicts a European multi-cohort study (Zijlema et al., 2016). Traffic noise is regarded as a psychosocial stressor causing annoyance and negative emotions (Rylander, 2004), and in a German study was significantly related to depressive symptoms (Generaal et al., 2019b; Orban et al., 2016).

Fourth, we found pleasantness was negatively correlated with depression severity. This is in line with previous research concerning neighborhood quality and depression. For example, walkable neighborhoods have previously been associated with reduced depressive symptoms (Koohsari et al., 2019). It is suggested that this is due to increased opportunity for social interaction, which in turn can improve depressive symptoms. Poor maintenance of buildings and incivilities in the street, or neighborhood social disorder, has been linked to increased risk of depression (Diez Roux and Mair, 2010). This may be the result of reduced neighborhood satisfaction (Leslie and Cerin, 2008), or via enhanced stress (Diez Roux and Mair, 2010).

Fifth, we found no indication that depression severity differed between urban and rural areas. While contradicting an international meta-analyses on mood disorders and urbanization (Peen et al., 2010), our results confirm another Dutch study reporting an insignificant correlation (Generaal et al., 2019b). Further, in a recent analysis of eight Dutch cohort studies, inconsistent results were found for the effect of urbanization on depression severity (Generaal et al., 2019a). It is suggested this is due to the use of different research designs, measures of depression, and confounders.

Lastly, we could not confirm that blue space within people's living environment is correlated with depression severity. This finding aligns with a series of others reporting insignificant associations on the 5% level (Generaal et al., 2019b; Zock et al., 2018). However, our findings were suggestive for beneficial mental health effects of perceived closeness to green space, though no causality can be inferred. Similar results were reported elsewhere (Groenewegen et al., 2018; Helbich et al., 2018; Stigsdotter et al., 2010). The assumed mechanisms may operate through stress recovery, attention restoration, physical activity, and social interaction (Hartig et al., 2014).

11.4.2. Strengths and limitations

A number of key strengths of this study need to be emphasized. Our study is innovative in the way correlations were assessed. While earlier studies were limited to linear associations without examining variable interactions and nonlinearities (De Vries et al., 2016; Generaal et al., 2019b; Gidlow et al., 2010), we put these challenges central and fitted flexible ML models in a data-driven manner. Our study also used a large nationally representative data set for the Netherlands. This produced a large sample size where our results are deemed to be robust. However, whether and how our findings can be generalized for a wider European or other cultural contexts needs further, ideally longitudinal (Murphy et al., 1991), exploration.

Despite these strengths, several limitations are recognized. The cross-sectional nature of the data has limited capability to establish causal links. We were unable to assess whether the social causation hypothesis or the social drift hypothesis applies (Lund et al., 2014). While the former posits that adversity linked with low socio-economic status contribute to depression, the latter argues that depressed people experience a downward drift towards neighborhoods with lower socio-economic status (Lund et al., 2014; Ritsher et al., 2001). Our findings may also be biased because depressed people might be more likely to view their environment negatively (Gong et al., 2016).

Our survey benefited from the inclusion of well-tested questionnaires (e.g., PHQ-9), which facilitates comparability with other studies, but they may be subject to self-reporting bias. We cannot eliminate that the perception of depressed people is impaired (Althubaiti, 2016). As some survey questions relate to people's living environment, ambiguities concerning the neighborhood size and the environmental perception may arise; which potentially have attenuated the relationships. Despite the fact we adjusted for several socio-economic characteristics, another final consideration is that we cannot rule out unmeasured and residual confounding. However, our findings were robust to adjustment for many potential confounding factors but some, for example people's physical activity levels (Schuch et al., 2018), were not available to us on a personal level.

11.5. Conclusions

The results reported here are from a large nationally representative sample from the Netherlands and provide support for a relationship between perceived physical and social neighborhood characteristics and people's severity of depression. The importance of the physical neighborhood environment is, however, limited relative to the social environment and individual attributes. We observed specifically that neighborhood social cohesion, pleasantness and safety were inversely correlated with depression severity, while distance to green space and traffic were positively correlated. No association was found for urbanicity and blue space. While confirmation through longitudinal research is required, our study suggests that modification of physical and social neighborhood characteristics could represent an effective intervention to promote mental health.

Acknowledgements

We thank the anonymous reviewers for their valuable comments on an earlier draft of this paper.

Funding

Funding The project was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 714993). In the course of the NEEDS project, JH's contribution was based on a research stay at Utrecht University.

Compliance with ethical standards

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (`http://creativecommons.org/licenses/by/4.0/`), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Bibliography

Althubaiti, A. (2016). Information bias in health research: Definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, 9, 211.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Chandola, T. (2001). The fear of crime and area differences in health. *Health & Place*, 7(2), 105–116.

de Groot, J., Roels, J., Shah, S., and de Witt, S. (2018). Stemming and leefomgeving. In: *Onderzoeksdocumentatie kwaliteitsanalyse (NEEDS)*. Den Haag.

De Vries, S., Ten Have, M., Dorsselaer, S. van, Wezep, M. van, Hermans, T., and Graaf, R. de (2016). Local availability of green and blue space and prevalence of common mental disorders in the Netherlands. *BJPsych Open*, 2(6), 366–372.

Diez Roux, A. V. and Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences*, 1186, 125–45.

Ehsan, A. M. and De Silva, M. J. (2015). Social capital and common mental disorder: A systematic review. *Journal of Epidemiology and Community Health*, 69(10), 1021–1028.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., and Amorim Fernández-Delgado, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1), 3133–3181.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3), 916–954.

Generaal, E., Hoogendijk, E. O., Stam, M., Henke, C. E., Rutters, F., Oosterman, M., Huisman, M., Kramer, S. E., Elders, P. J. M., Timmermans, E. J., Lakerveld, J., Koomen, E., Ten Have, M., de Graaf, R., Snijder, M. B., Stronks, K., Willemsen, G., Boomsma, D. I., Smit, J. H., and Penninx, B. W. J. H. (2019a). Neighbourhood characteristics and prevalence and severity of depression: pooled analysis of eight Dutch cohort studies. *The British Journal of Psychiatry*, 215(2), 468–475.

Generaal, E., Timmermans, E. J., Dekkers, J. E. C., Smit, J. H., and Penninx, B. W. J. H. (2019b). Not urbanization level but socioeconomic, physical and social neighbourhood characteristics are associated with presence and severity of depressive and anxiety disorders. *Psychological Medicine*, 49(1), 149–161.

Gidlow, C., Cochrane, T., Davey, R. C., Smith, G., and Fairburn, J. (2010). Relative importance of physical and social aspects of perceived neighbourhood environment for self-reported health. *Preventive Medicine*, 51(2), 157–163.

Gilbody, S., Richards, D., Brealey, S., and Hewitt, C. (2007). Screening for depression in medical settings with the patient health questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine*, 22(11), 1596–1602.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65.

Gong, Y., Palmer, S., Gallacher, J., Marsden, T., and Fone, D. (2016). A systematic review of the relationship between objective measurements of the urban environment and psychological distress. *Environment International*, 96, 48–57.

Groenewegen, P. P., Zock, J.-P., Spreeuwenberg, P., Helbich, M., Hoek, G., Ruijsbroek, A., Strak, M., Verheij, R., Volker, B., Waverijn, G., and Dijst, M. (2018). Neighbourhood social and physical environment and general practitioner assessed morbidity. *Health & place*, 49, 68–84.

Gross, J. J. (1998). Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology*, 74(1), 224.

Gu, X., Liu, Q., Deng, F., Wang, X., Lin, H., Guo, X., and Wu, S. (2019). Association between particulate matter air pollution and risk of depression and suicide: systematic review and meta-analysis. *The British Journal of Psychiatry*, 215(2), 456–467.

Hartig, T., Mitchell, R., De Vries, S., and Frumkin, H. (2014). Nature and health. *Annual Review of Public Health*, 35, 207–228.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Haykin, S. S. (1994). *Neural networks and learning machines.* Bd. 10. Prentice Hall.

Helbich, M. (2019a). Dynamic Urban Environmental Exposures on Depression and Suicide (NEEDS) in the Netherlands: A protocol for a cross-sectional smartphone tracking study and a longitudinal population register study. *BMJ open*, 9(8), e030075.

Helbich, M. (2019b). Spatiotemporal contextual uncertainties in green space exposure measures: Exploring a time series of the normalized difference vegetation indices. *International Journal of Environmental Research and Public Health*, 16(5), 852.

Helbich, M., Klein, N., Roberts, H., Hagedoorn, P., and Groenewegen, P. P. (2018). More green space is related to less antidepressant prescription rates in the Netherlands:

A Bayesian geoadditive quantile regression approach. *Environmental Research*, 166, 290–297.

Jefferis, B. J., Nazareth, I., Marston, L., Moreno-Kustner, B., Bellón, J. Á., Svab, I., Rotar, D., Geerlings, M. I., Xavier, M., Goncalves-Pereira, M., Vicente, B., Saldivia, S., Aluoja, A., Kalda, R., and King, M. (2011). Associations between unemployment and major depressive disorder: Evidence from an international, prospective study (the predict cohort). *Social Science & Medicine*, 73(11), 1627–1634.

Jorm, A. F. (2000). Does old age reduce the risk of anxiety and depression? A review of epidemiological studies across the adult life span. *Psychological Medicine*, 30(1), 11–22.

Kawachi, I. and Berkman, L. (2000). Social cohesion, social capital, and health. *Social Epidemiology*, 174(7).

Kawachi, I. and Berkman, L. F. (2001). Social ties and mental health. *Journal of Urban Health*, 78(3), 458–467.

Klijs, B., Kibele, E. U. B., Ellwardt, L., Zuidersma, M., Stolk, R. P., Wittek, R. P. M., Leon, C. M. M. de, and Smidt, N. (2016). Neighborhood income and major depressive disorder in a large Dutch population: Results from the LifeLines cohort study. *BMC Public Health*, 16(1), 1–13.

Koohsari, M. J., McCormack, G. R., Nakaya, T., Shibata, A., Ishii, K., Yasunaga, A., Hanibuchi, T., and Oka, K. (2019). Urban design and Japanese older adults' depressive symptoms. *Cities*, 87, 166–173.

Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal Of Statistical Software*, 28(5), 1–26.

Kwan, M.-P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers*, 102(5), 958–968.

Leslie, E. and Cerin, E. (2008). Are perceptions of the local environment related to neighbourhood satisfaction and mental health in adults? *Preventive Medicine*, 47(3), 273–278.

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., and Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. *American Journal of Epidemiology*, 157(2), 98–112.

Lorenc, T., Clayton, S., Neary, D., Whitehead, M., Petticrew, M., Thomson, H., Cummins, S., Sowden, A., and Renton, A. (2012). Crime, fear of crime, environment, and mental health and wellbeing: Mapping review of theories and causal pathways. *Health & Place*, 18(4), 757–765.

Lund, C., Stansfeld, S., and De Silva, M. (2014). Social determinants of mental health. In: *Global mental health: principles and practice.* Oxford University Press, Oxford, 116–136.

Mair, C., Roux, A. V. D., Shen, M., Shea, S., Seeman, T., Echeverria, S., and O'meara, E. S. (2009). Cross-sectional and longitudinal associations of neighborhood cohesion and stressors with depressive symptoms in the multiethnic study of atherosclerosis. *Annals of Epidemiology*, 19(1), 49–57.

Malhi, G. S. and Mann, J. J. (2018). *Depression.* Lancet.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models.* Chapman & Hall/CRC, 500.

Murphy, J. M., Olivier, D. C., Monson, R. R., Sobol, A. M., Federman, E. B., and Leighton, A. H. (1991). Depression and anxiety in relation to social status: A prospective epidemiologic study. *Archives of General Psychiatry*, 48(3), 223–229.

Orban, E., McDonald, K., Sutcliffe, R., Hoffmann, B., Fuks, K. B., Dragano, N., Viehmann, A., Erbel, R., Jöckel, K.-H., Pundt, N., and Moebus, S. (2016). Residential road traffic noise and high depressive symptoms after five years of follow-up: results from the Heinz Nixdorf recall study. *Environmental Health Perspectives*, 124(5), 578–585.

Peen, J., Schoevers, R. A., Beekman, A. T., and Dekker, J. (2010). The current status of urban-rural differences in psychiatric disorders. *Acta Psychiatrica Scandinavica*, 121(2), 84–93.

Purtle, J., Nelson, K. L., Yang, Y., Langellier, B., Stankov, I., and Roux, A. V. D. (2019). Urban–rural differences in older adult depression: A systematic review and meta-analysis of comparative studies. *American Journal of Preventive Medicine*, 56(4), 603–613.

R Core Team (2018). *R: A language and environment for statistical computing.*

Rautio, N., Filatova, S., Lehtiniemi, H., and Miettunen, J. (2018). Living environment and its relationship to depressive mood: A systematic review. *International Journal of Social Psychiatry*, 64(1), 92–103.

Reklaitiene, R., Grazuleviciene, R., Dedele, A., Virviciute, D., Vensloviene, J., Tamosiunas, A., Baceviciene, M., Luksiene, D., Sapranaviciute-Zabazlajeva, L., Radisauskas, R., Bernotiene, G., Bobak, M., and Nieuwenhuijsen, M. J. (2014). The relationship of green space, depressive symptoms and perceived general health in urban population. *Scandinavian Journal of Public Health*, 42(7), 669–676.

Richardson, R., Westley, T., Gariépy, G., Austin, N., and Nandi, A. (2015). Neighborhood socioeconomic conditions and depression: A systematic review and meta-analysis. *Social Psychiatry and Psychiatric Epidemiology*, 50(11), 1641–1656.

Ritsher, J. E. B., Warner, V., Johnson, J. G., and Dohrenwend, B. P. (2001). Intergenerational longitudinal study of social class and depression: A test of social causation and social selection models. *The British Journal of Psychiatry*, 178(S40), s84–s90.

Ruijsbroek, A., Mohnen, S. M., Droomers, M., Kruize, H., Gidlow, C., Gražulevičiene, R., Andrusaityte, S., Maas, J., Nieuwenhuijsen, M. J., Triguero-Mas, M., Masterson, D., Ellis, N., van Kempen, E., Hardyns, W., Stronks, K., and Groenewegen, P. P. (2017). Neighbourhood green space, social environment and mental health: An examination in four European cities. *International Journal of Public Health*, 62(6), 657–667.

Rylander, R. (2004). Physiological aspects of noise-induced stress and annoyance. *Journal of Sound and Vibration*, 277(3), 471–478.

Sampson, R. J., Raudenbush, S. W., and Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328), 918–924.

Schuch, F. B., Vancampfort, D., Firth, J., Rosenbaum, S., Ward, P. B., Silva, E. S., Hallgren, M., Ponce De Leon, A., Dunn, A. L., Deslandes, A. C., Fleck, M. P., Carvalho, A. F., and Stubbs, B. (2018). Physical activity and incident depression: A meta-analysis of prospective cohort studies. *American Journal of Psychiatry*, 175(7), 631–648.

Spittaels, H., Verloigne, M., Gidlow, C., Gloanec, J., Titze, S., Foster, C., Oppert, J.-M., Rutter, H., Oja, P., Sjöström, M., and De Bourdeaudhuij, I. (2010). Measuring physical activity-related environmental factors: Reliability and predictive validity of the European environmental questionnaire. *International Journal of Behavioral Nutrition and Physical Activity*, 7(1), 1–19.

Stigsdotter, U. K., Ekholm, O., Schipperijn, J., Toftager, M., Kamper-Jørgensen, F., and Randrup, T. B. (2010). Health promoting outdoor environments-associations between green space, and health, health-related quality of life and stress based on a Danish national representative survey. *Scandinavian Journal of Public Health*, 38(4), 411–417.

Stokols, D. (1992). Establishing and maintaining healthy environments: Toward a social ecology of health promotion. *American Psychologist*, 47(1), 6.

Vigo, D., Thornicroft, G., and Atun, R. (2016). Estimating the true global burden of mental illness. *The Lancet Psychiatry*, 3(2), 171–178.

Wilbur, J., Zenk, S., Wang, E., Oh, A., McDevitt, J., Block, D., McNeil, S., and Ju, S. (2009). Neighborhood characteristics, adherence to walking, and depressive symptoms in midlife African American women. *Journal of Women's Health*, 18(8), 1201–1210.

Wilson-Genderson, M. and Pruchno, R. (2013). Effects of neighborhood violence and perceptions of neighborhood safety on depressive symptoms of older adults. *Social Science & Medicine*, 85, 43–49.

Zijlema, W. L., Wolf, K., Emeny, R., Ladwig, K.-H., Peters, A., Kongsgård, H., Hveem, K., Kvaløy, K., Yli-Tuomi, T., Partonen, T., Lanki, T., Eeftens, M., de Hoogh, K., Brunekreef, B., BioSHaRE, Stolk, R. P., and Rosmalen, J. G. (2016). The association of air pollution and depressed mood in 70,928 individuals from four European cohorts. *International Journal of Hygiene and Environmental Health*, 219(2), 212–219.

Zock, J.-P., Verheij, R., Helbich, M., Volker, B., Spreeuwenberg, P., Strak, M., Janssen, N. A. H., Dijst, M., and Groenewegen, P. (2018). The impact of social capital, land use, air pollution and noise on individual morbidity in Dutch neighbourhoods. *Environment International*, 121, 453–460.

12. A geographically weighted artificial neural network

Authors

Julian Hagenauer and Marco Helbich

Journal

International journal of geographical information science

Status

Submitted 10 April 2020; accepted 31 December 2020; published online 8 February 2021

Contribution statement

Julian Hagenauer has developed the methods, designed the experiments, and has written the manuscript for the study. Marco Helbich supported this publication by continuously discussing the design and results of the study and by proofreading the manuscript.

Abstract

While recent developments have extended geographically weighted regression (GWR) in many directions, it is usually assumed that the relationships between the dependent and the independent variables are linear. In practice, however, it is often the case that variables are nonlinearly associated. To address this issue, we propose a geographically weighted artificial neural network (GWANN). GWANN combines geographical weighting with artificial neural networks, which are able to learn complex nonlinear relationships in a data-driven manner without assumptions. Using synthetic data with known spatial characteristics and a real-world case study, we compared GWANN with GWR. While the results for the synthetic data show that GWANN performs better than GWR when the relationships within the data are nonlinear and their spatial variance is high, the results based on the real-world data demonstrate that the performance of GWANN can also be superior in a practical setting.

Keywords: *Geographically weighted regression; artificial neural network; spatial heterogeneity; nonlinear relationships; spatial prediction*

12.1. Introduction

Spatial heterogeneity of relationships (i.e., spatial nonstationarity) is an important issue in spatial data analysis (Anselin, 1989). It refers to the notion that for a spatial process, the relationships between variables depend to some degree on the location where the relationships are observed (Fotheringham et al., 2002). If spatial heterogeneity is not appropriately taken into account when calibrating a model, the estimation of the coefficients is likely to be biased, which can lead to inappropriate conclusions (LeSage and Pace, 2009; Páez et al., 2008).

Several approaches have been proposed to model spatially varying relationships. Notable examples include the expansion method (Casetti, 1972), weighted spatial adaptive filtering (Gorr and Olligschlaeger, 1994), Eigenvector spatial filtering (Griffith, 2003), and geographically weighted regression (GWR) (Brunsdon et al., 1999). Of these approaches, GWR has received the most attention and is employed across many disciplines, for example, real estate economics (Bitter et al., 2007; Helbich and Griffith, 2016), ecology (Nelson et al., 2007), criminology (Troy et al., 2012; Waller et al., 2007), health (Choi and Kim, 2017), and land-use science (Hagenauer and Helbich, 2018; Yu et al., 2011).

GWR is an extension of ordinary least squares (OLS), which estimates for each location a weighted least squares regression, where observations that are closer to the regression location are given a higher weight than those farther away. The weighting is determined by a distance–decay kernel function and a bandwidth parameter.

Several extensions and modifications of GWR have been proposed. While in basic GWR all relationships are assumed to vary spatially, in mixed GWR (Brunsdon et al., 1999) only a subset of the coefficients are subject to geographical weighting; the kernel function and bandwidth for each spatially varying coefficient are identical. The latter restriction was addressed by Fotheringham et al. (2017), who proposed a multiscale GWR that uses individual bandwidths for the coefficients to model different scales of spatial heterogeneity. Furthermore, while basic GWR is based on Euclidean distances between observations, the application of different distance metrics has been proposed. Lu et al. (2011), for example, showed that non-Euclidean distance metrics can improve the fit of GWR, whereas Fotheringham et al. (2015) suggested the use of a spatio-temporal distance metric. Lu et al. (2017) combined multiscale GWR with individual distance metrics per coefficient.

GWR has also been criticized for artificially introducing multicollinearity between coefficient pairs (Wheeler and Tiefelsdorf, 2005), which was recently refuted (Fotheringham and Oshan, 2016). To counteract this criticism, penalized forms of GWR were proposed (e.g., geographically weighted lasso (Wheeler, 2009), ridge (Bárcena et al., 2014; Wheeler, 2007), and elastic net regression (Li and Lam, 2018)). Another extension is geographically neural network weighted regression (Du et al., 2020), which utilizes an artificial neural network (ANN) to find appropriate geographical weights when estimating the coefficients of a GWR model.

Despite these efforts, some restrictions of GWR have not yet been addressed. For instance, because GWR resembles a collection of local models where data from neighboring local models are reused, its inferential properties are inferior to a single nonstationary model (Comber et al., 2020). Also, analogous to OLS, when using GWR in its simplest form it is assumed that the relationships between dependent and independent variables are linear. This assumption, however, typically does not hold for complex spatial prediction tasks (Anselin, 1989; Leuenberger and Kanevski, 2015).

To address this issue, we propose a geographically weighted artificial neural network (GWANN), which combines geographical weighting with an ANN. Similar to GWR, GWANN uses a distance-decay kernel function and a bandwidth parameter to geographically weight observations when building the model. However, in contrast to GWR, GWANN is also able to model nonlinear functions in a data-driven manner without making any assumptions.

The rest of this article is structured as follows. Section 12.2 describes GWR and ANN and introduces GWANN. Next, section 12.3 presents experiments that were carried out to compare GWANN with GWR. Finally, section 12.4 gives concluding remarks and proposes future work.

12.2. Methods

12.2.1. Artificial neural network

An artificial neural network (ANN) consists of a set of neurons and unidirectional connections between them, which enables the imitation of the brain's ability to detect patterns and learn relationships within data (Haykin, 2008). Associated with each neuron i is an activation function ϕ_i and each connection between two neurons i,j has a weight w_{ij} assigned that controls the influence of neuron i on neuron j. While the neurons represent the basic computation units of an ANN, the weighted connections between them allow the modeling of complex relationships.

The neurons are typically organized in layers, and each neuron in a layer has directed connections to the neurons in the subsequent layer (Figure 12.1). The first layer is termed "input layer" and the last layer "output layer," while all layers in between are "hidden layers". The input data are passed from the input layer to the first hidden layer, where it is aggregated and transformed as follows:

$$net_j = \sum_{i \in P_j} w_{ij} o_i \tag{12.1}$$

where w_{ij} is the weight of the connection between neuron i and j, o_i the output of neuron i, and P_j the set of neurons that have an outgoing connection to neuron j. The output of a neuron i is calculated as follows:

$$o_i = \phi(net_i) \tag{12.2}$$

where ϕ is the activation function of neuron i. A common activation function is the hyperbolic tangent function, which is defined as $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. This function is particularly useful because it is continuous and differentiable; both are necessary conditions for calculating the network's error gradient (Rojas, 2013).

The output of each neuron is then passed on to the neurons in the next layer. For each subsequent layer, this procedure is repeated until the output layer of the network is reached. The output of the output layer represents the total output of the network.

In order to model nonlinear relationships, the connection weights of an ANN must be adjusted. This is typically done using a two-step procedure. In the first step, the error signal of each neuron for a given observation is calculated using backpropagation

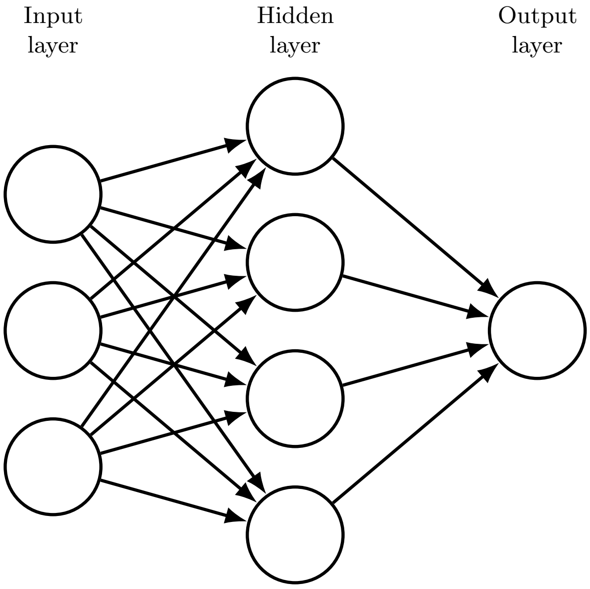


Figure 12.1.: ANN with three layers.

(Rumelhart et al., 1986). The error signal depends on the error function. In the case of regression, the error function is defined as $E = \frac{1}{2}\sum_{i=1}^{n}(t_i - o_i)^2$ where t_i is the target value, o_i the output of the output neuron i, and n the number of the target values. Given this error function, the error signal is calculated as follows:

$$\delta_j = \begin{cases} \phi'(net_j)(o_j - t_j) & \text{if } j \text{ is an output neuron} \\ \phi'(net_j)\sum_k \delta_k w_{jk} & \text{otherwise} \end{cases} \tag{12.3}$$

where o_j is the output of neuron j, t_j the target value of neuron j, w_{jk} the connection weight between neuron j and k, δ_k the error signal for neuron k, net_j the network input to neuron j, and ϕ' the derivative of the activation function.

In the second step, the connection weights are adjusted using gradient descent:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = -\eta \delta_j o_i \tag{12.4}$$

where w_{ij} is the connection weight between neuron i and j, E the error function, o_i the output of neuron i, and δ_j the error signal of neuron j. Both steps are repeated until a terminating condition is reached (e.g., the error rate is below a predetermined threshold value).

Several extensions and variants of gradient descent have been proposed to improve the training of the network. To make the training more robust to noise, the error gradients are in practice summed over a subset of observations, termed a "mini-batch". The connection weights are then updated using the accumulated changes. Also, using Nesterov's accelerated gradient (Nesterov, 1983) when adjusting the connection weights can substantially improve the training performance (Sutskever et al., 2013).

12.2.2. Geographically weighted regression

Geographically weighted regression (GWR) (Brunsdon et al., 1996) estimates for each location a separate local model. Assuming that there are n locations and each location has an observation assigned to it, the GWR model for the location $i \in 1, 2, ..., n$ is:

$$y_i = \sum_{j=0}^{m} \beta_{ij} x_{ij} + \epsilon_i \tag{12.5}$$

where y_i is the dependent variable, x_{ij} the independent variable j, β_{ij} the coefficient for the independent variable j, and ϵ_i the error term, which is assumed to be independent and identically distributed.

GWR weights the observations by their spatial distance when estimating the local coefficients; close observations are given more weight than observations farther away. The estimation is typically done using weighted least squares, the matrix expression of which is:

$$\hat{\beta}_i = (X^T W_i X)^{-1} X^T W_i y \tag{12.6}$$

where X is the design matrix, y the dependent variable, and W_i a column vector of the spatial weights matrix W for location i. To calculate W, a kernel function is applied to the distances between observations and regression locations. Widely used kernels are Gaussian, bisquare, tricube, and boxcar kernels (Brunsdon et al., 1999). The Gaussian kernel, for instance, is defined as $v_{ij} = e^{-0.5(\frac{d_{ij}}{h})^2}$, where d_{ij} is the distance between locations i and j and h is the kernel bandwidth. The bandwidth determines the degree of variation in the local coefficient estimates and is considered to be more important for the performance of GWR than the choice of the kernel function (Fotheringham et al., 2002). The bandwidth can be either fixed or adaptive, where the latter refers to the distance to the k-nearest neighbor of each observation (Brunsdon et al., 2007; Guo et al., 2008).

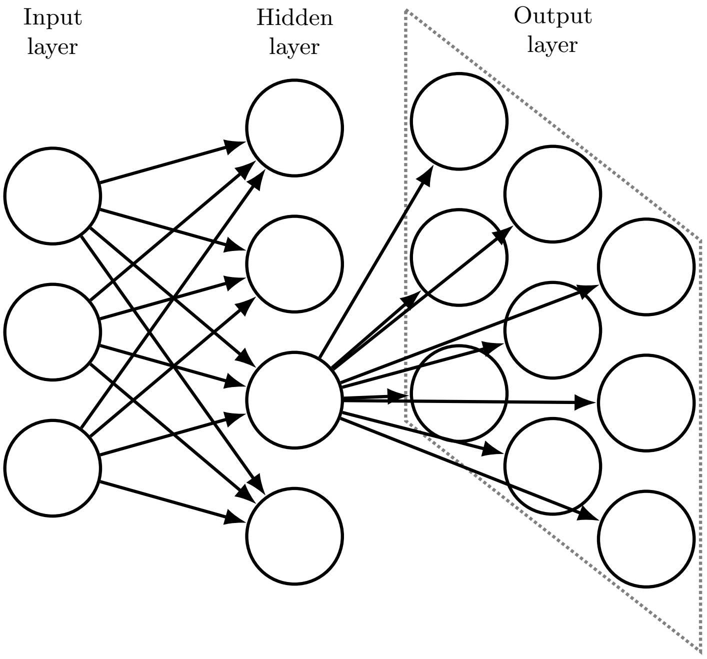


Figure 12.2.: GWANN with three layers. The rectangle indicates that the output neurons are assigned to locations on a plane. Note that although each hidden neuron typically has connections to all output neurons, for the sake of clarity the outgoing connections are shown for a single hidden neuron only.

12.2.3. Geographically weighted artificial neural network

A geographically weighted artificial neural network (GWANN) is a variant of an ANN that incorporates geographical weighting of connection weights. The principle idea is as follows. A basic ANN consists of an input, a hidden, and an output layer. The connection weights of a basic ANN from the hidden to the output layer can be interpreted as the coefficients of a linear model of nonlinearly transformed variables, namely the outputs of the hidden neurons. Thus, when the connection weights between the hidden and the output layer are estimated by utilizing a geographically weighted error function, these weights can be interpreted as a GWR model.

The architecture of GWANN is identical to that of a basic ANN, except that each output neuron of GWANN is assigned to a location in geographic space (Figure 12.2). This allows to calculate the spatial distances between the observations and the locations of the output neurons.

Besides the network architecture, the main difference between GWANN and a basic ANN is that GWANN uses a geographic weighted error function instead of the basic quadratic error function in order to calculate an error signal. In the case of regression, the geographically weighted error function is defined as $E = \frac{1}{2}\sum_{i=1}^{n} v_i(t_i - o_i)^2$, where t_i is the target value, o_i the output of output neuron i, v_i the geographically weighted distance between the observation and the location of output neuron i, and n the number of target values/output neurons. Following this definition, the difference between the output neurons' output and the target values is weighted by the spatial distance between output neurons' location and the observation; when the output neurons' location and observation are close, the difference is given more weight than when they are farther apart. Note that the number of target values must be identical to the number of output neurons. In particular, in a practical example where one wants to calculate the value of the geographic error function for a single target value but multiple output neurons with typically different locations, it is necessary to replicate the target value for each output neuron.

Following the definition of the geographically weighted error function, the calculation of the error signal of backpropagation is modified as follows:

$$\delta_j = \begin{cases} \phi'(net_j)v_j(o_j - t_j) & \text{if } j \text{ is an output neuron} \\ \phi'(net_j)\sum_k \delta_k w_{jk} & \text{otherwise} \end{cases} \tag{12.7}$$

where o_j is the output of neuron j, t_j the target value of neuron j, w_{jk} the connection weight between neuron j and k, δ_k the error signal for neuron k, net_j the network input to neuron j, ϕ' the derivative of the activation function, and v_j the geographically weighted distance between the observation and the location of output neuron j. Geographical weighting is only used for calculating the error signal of the output neurons, whereas all other neurons backpropagate the error signal of the neurons of the next layer. Like ANN, the connection weights of GWANN are adjusted using gradient descent (equation 12.4).

12.3. Experiments

To compare GWR with GWANN, we used four synthetic datasets and one real-world dataset from real-estate economics. The synthetic datasets gave us full control over the characteristics of the data, in particular the nature of the relationships and spatial heterogeneity, which contributed to a better understanding of the different properties of the models. The real-world data allowed us to assess the models in a practical use case.[1]

For all experiments, we scaled the input variables to have zero mean and unit variance to make them comparable. We used Nesterov's accelerated gradient with a momentum coefficient of 0.900 when adjusting the connection weights. We set the learning rate η of GWANN to 0.010 and the mini-batch size to 50. While in principle the number of hidden layers of an ANN is arbitrary, we chose networks with a single hidden layer. Given enough hidden neurons, ANNs with a single hidden layer are able to arbitrary well approximate any continuous function on closed and bounded subsets of n-dimensional Euclidean space (Cybenko, 1989). For each experiment, we tested different numbers of hidden neurons. A bias neuron is always added to the input and the hidden layer, but we did not include them when reporting the number of neurons. The hyperbolic tangent function is used as activation function for the hidden neurons.

We used a Gaussian kernel with GWR and GWANN for geographical weighting. When using an adaptive bandwidth, a grid search is performed to determine an appropriate bandwidth. When using a fixed bandwidth, the following local search approach is used to determine an appropriate bandwidth. The approach initially selects half of the largest distance between two observations as the current bandwidth. Then, a grid search is performed within the neighborhood of the current bandwidth for a bandwidth that results in a better mean performance. When one is found, the process is repeated within a smaller neighborhood of the newly found bandwidth until convergence.

The performance within the bandwidth search is estimated using 10-fold cross-validation (CV). This procedure randomly partitions the data into 10 disjoint subsets. One subset at a time is then used to test the model, while the others are used to build it. Then, the mean performance over all folds is reported. We used the root mean square error (RMSE) as a performance measure.

The number of training iterations of GWANN was also determined using 10-fold CV. Within each fold, the models are trained until the performance for the test data of the current fold does not improve for 1,000 iterations. The purpose of the additional iterations is to give the networks a chance to escape from local minima. This approach, commonly termed "early stopping with patience", substantially reduces the risk of overfitting the training data (Bengio, 2012). Then, the iteration for which the best mean performance over all folds has been obtained as well as the obtained performance value are reported.

---

[1] Two additional experiments are given in the supplemental materials. The first one uses housing benchmark data to predict house prices and the second one traffic and land-use data to predict nitrogen dioxide concentrations.

12.3.1. Experiment 1: Synthetic data

The purpose of this experiment was to investigate the differences between GWR and GWANN when modeling processes with different spatial characteristics. In particular, we were interested in how the model performance depends on the linearity and spatial variation of the relationships. We also examined the visualization of GWANN's connection weights between the hidden and output neurons as surfaces.

Data generating process

We created four artificial datasets. The spatial layout of the datasets was given by a grid of size 25 × 25. The following functions were used to create the datasets:

$$y_i = \beta_0 + \beta_1(u_i, v_i)x_{i1} + \beta_2^1(u_i, v_i)x_{i2} + \epsilon_i \tag{12.8}$$

$$y_i = \beta_0 + \beta_1(u_i, v_i)x_{i1} + \beta_2^2(u_i, v_i)x_{i2} + \epsilon_i \tag{12.9}$$

$$y_i = \beta_0 + 4\tanh(\frac{\beta_1(u_i, v_i)x_{i1}}{3}) + 4\tanh(\frac{\beta_2^1(u_i, v_i)x_{i2}}{3}) + \epsilon_i \tag{12.10}$$

$$y_i = \beta_0 + 4\tanh(\frac{\beta_1(u_i, v_i)x_{i1}}{3}) + 4\tanh(\frac{\beta_2^2(u_i, v_i)x_{i2}}{3}) + \epsilon_i \tag{12.11}$$

For all functions, (u_i, v_i) denotes the position of grid cell i, ϵ_i the error term drawn from $N(0, 0.25)$, x_i a random variable drawn $N(0, 1)$, and β_0, $\beta_1(u_i, v_i)$, and $\beta_2^1(u_i, v_i)$ and $\beta_2^2(u_i, v_i)$, respectively, the coefficients for grid cell i. While the first two functions (equations 12.8 and 12.9) model linear relationships between the dependent and independent variables, the third and fourth functions (equations 12.10 and 12.11) use the hyperbolic tangent function to represent nonlinear relationships.

The coefficients were designed to represent different characteristics of spatial heterogeneity. They were calculated as follows:

$$\beta_0 = 1 \tag{12.12}$$

$$\beta_1(u_i, v_i) = 1 + \frac{u_i + v_i}{12} \tag{12.13}$$

$$\beta_2^1(u_i, v_i) = 1 + 2(\cos(\frac{\pi u_i}{24})\cos(\frac{\pi v_i}{24})) \tag{12.14}$$

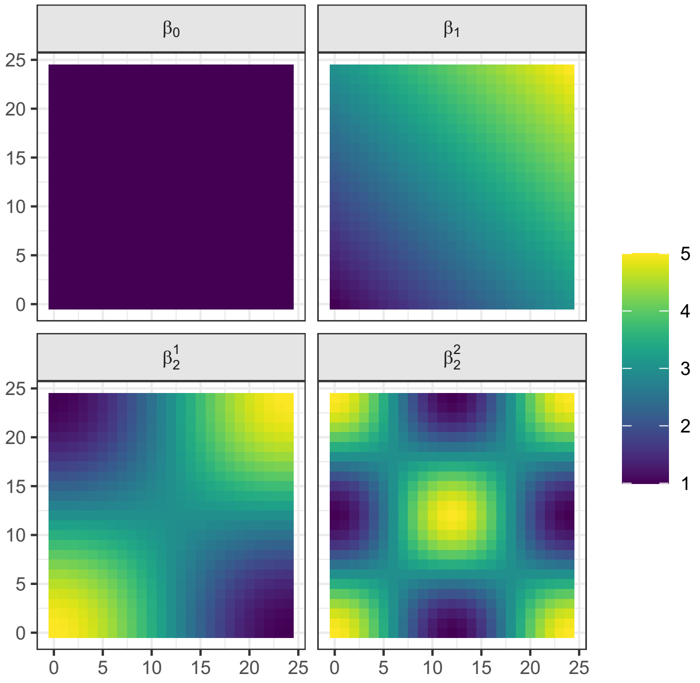


Figure 12.3.: Coefficients' surfaces with different characteristics of spatial heterogeneity.

$$\beta_2^2(u_i, v_i) = 1 + 2(\cos(\frac{\pi u_i}{12})\cos(\frac{\pi v_i}{12})) \tag{12.15}$$

For all coefficients, (u_i, v_i) denotes the position of grid cell i. β_0 represents a constant surface with no spatial heterogeneity. β_1 is a linear trend surface. β_2^1 and β_2^2 vary nonlinearly with location; the spatial variation of β_2^2 is higher than that of β_2^1. In terms of scale, β_2^1 represents small-scale spatial heterogeneity, and β_2^2 large-scale spatial heterogeneity. Figure 12.3 shows the coefficients' surfaces.

Following the definition of the coefficients, the first and third functions (equations 12.8 and 12.10) represent processes with low spatial variance, while the second and fourth functions (equations 12.9 and 12.11) represent processes with high spatial variance of coefficients.

Experimental setup

For all datasets, we used the variable y as the dependent variable and the variables x_1 and x_2 as independent variables. We used fixed bandwidths for GWR and GWANN. This allowed a finer control of the bandwidths when the observations were uniformly arranged in a grid and thus only a few distance classes were present.

To investigate the performance of GWR and GWANN, we used 10-fold CV with 90% of the data to determine an appropriate bandwidth for GWR and GWANN as well as an appropriate number of iterations for GWANN. Then, we used the same data to build a GWR and GWANN model with the hyperparameters determined and used the remaining data to obtain an independent estimate of their performance. We repeated the procedure for 100 random replications of each of the four toy datasets and reported the mean results.

The estimated coefficients of GWR can be visualized as surfaces to explore the spatial variation of the relationships. Analogously to GWR, it is also possible to visualize GWANN's connection weights between the hidden and the output neurons as surfaces. Each surface then refers to a hidden neuron's output, which is a nonlinearly transformed linear combination of the input variables.

To investigate and compare the visualization of the coefficient surfaces of GWR and the connection weight surfaces of GWANN, we built a GWR and a GWANN model using an exemplary replication of the dataset that was created using equation 12.11. This dataset is the most complex one because of the nonlinearity of the relationships and the high spatial variance of the coefficients. The number of hidden neurons of GWANN was set to five because this allowed a comprehensive visualization while providing a good model fit. Since we wanted to visualize the coefficient weights for every observation, the number of output neurons equaled the total number of observations, and each output neuron was assigned the location of an observation. Due to randomness in the data generating process and in the training of GWANN, a different bandwidth and different number of training iterations were determined for most replications. We chose the bandwidth and number of iterations corresponding to the replication for which the median RMSE over all replications had been obtained.

Results & discussion

Figure 12.4 shows the mean number of training iterations GWANN until convergence. The mean number of training iterations of GWANN does not change with the number of hidden neurons when the relationships are nonlinear and the spatial variance of the coefficients is low; otherwise, it decreases with the number of hidden neurons.

Figure 12.5 shows the obtained mean bandwidths. The mean bandwidth of GWANN always decreases with the number of hidden neurons; the decrease, however, is small when the relationships are nonlinear. The mean bandwidth of GWANN is larger than that of GWR when the relationships are linear, whereas it is lower when the relationships are nonlinear. Also, the mean bandwidths of GWANN and GWR are

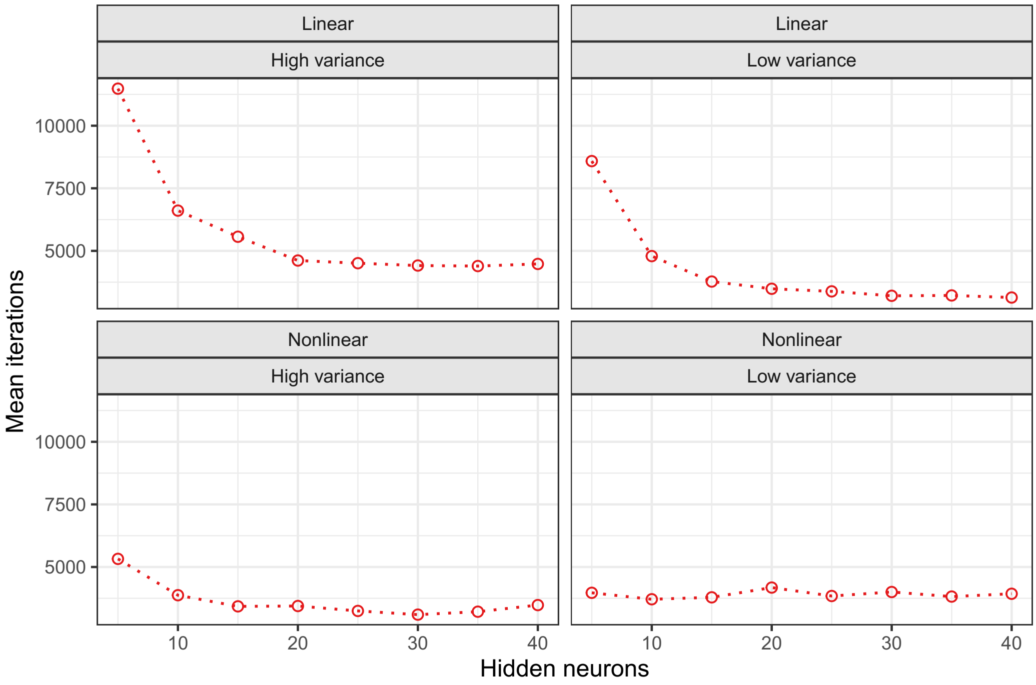


Figure 12.4.: Number of iterations until convergence.

generally higher when the spatial variance of the coefficients is low.

Figure 12.6 shows the mean RMSE of the models for the independent hold-out test datasets (for the proportion of explained variance, see Figure S1 in the supplementary materials). The mean RMSE of GWR is lower than the mean RMSE of GWANN when the relationships are linear. However, when the relationships are nonlinear, the mean RMSE of GWR is substantially higher than the mean RMSE of GWANN. This is not unexpected, because unlike GWANN, GWR is not inherently capable of modeling nonlinear relationships. The mean RMSE of GWANN is substantially lower than the mean RMSE of GWR when the relationships are nonlinear and the spatial variance of the coefficient is high. The mean RMSE of GWANN generally decrease with the number of hidden neurons when the relationships are linear; the decrease is stronger, however, when the spatial variance of the coefficients is high. When the relationships are nonlinear and the spatial variance of the coefficients is low, then the mean RMSE of GWANN only decrease with the number of hidden neurons if that number is low; otherwise the RMSE remains the same. No correlation between the number hidden neurons and mean RMSE is observable when the relationships are nonlinear and the spatial variance of the coefficients is high.

In general, the performance of all tested models depends on the nature of the

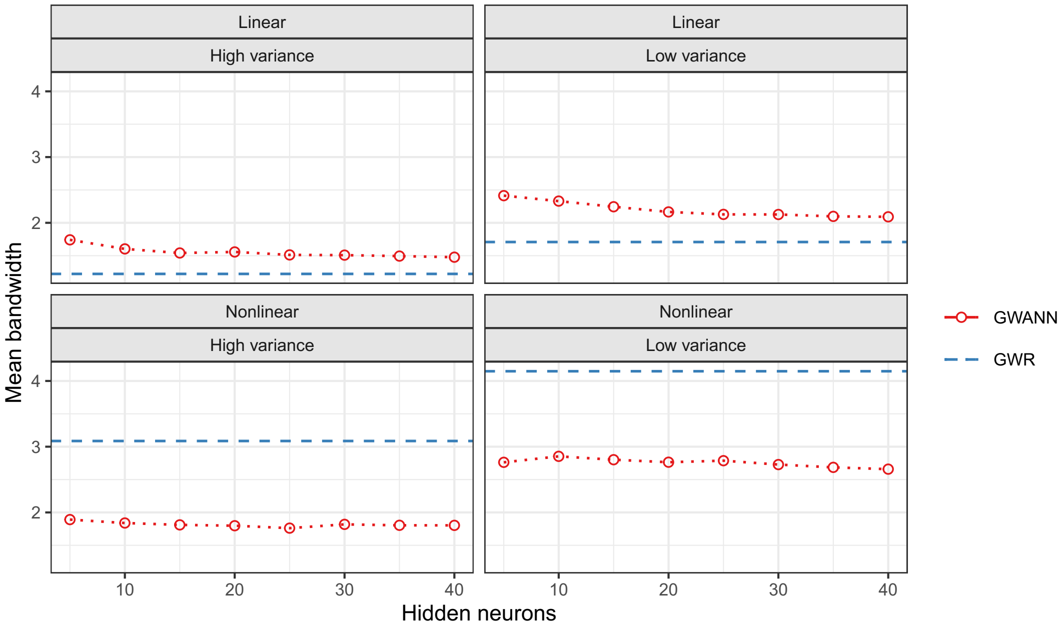


Figure 12.5.: Determined bandwidths (fixed).

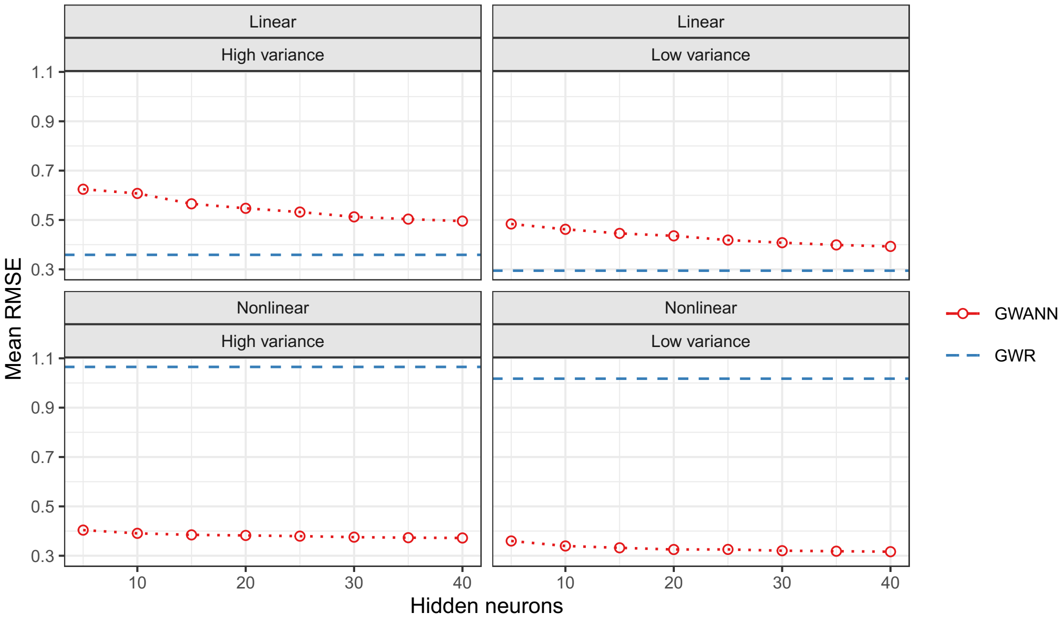


Figure 12.6.: Estimated prediction error (RMSE).

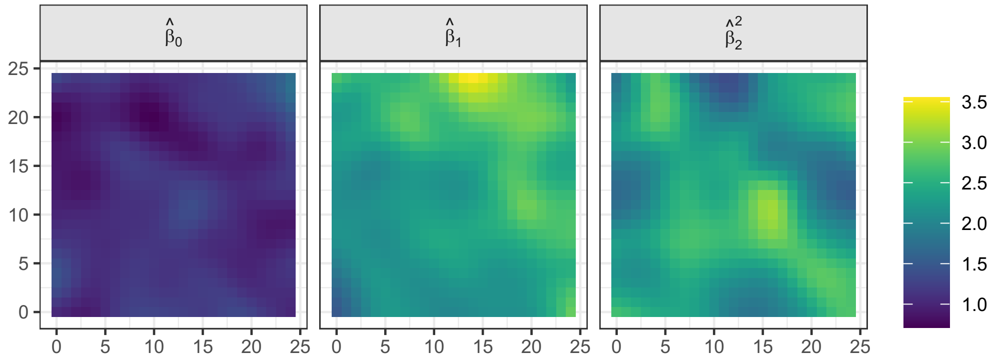


Figure 12.7.: Estimated coefficient surfaces of GWR.

underlying process. When the relationships in the data are nonlinear and the spatial variance of the coefficients is high (i.e., large-scale spatial heterogeneity), GWANN performs substantially better than GWR. In practice, however, the characteristics of the data generating process are usually not known beforehand and therefore it is necessary to empirically assess the performance of the competing models.

Using an exemplary replication of the dataset that was created using equation 12.11, we trained GWANN for $5,610$ iterations with a bandwidth of 1.801 and fitted a GWR model with a bandwidth of 2.000. Figure 12.7 shows the coefficient surfaces of GWR. The coefficient surfaces roughly resemble the coefficients of the original dataset (see Figure 12.3). We calculated Pearson's correlation coefficient between the surfaces of GWR and the coefficients of the dataset to quantify their similarity. The linear trend of β_1 from bottom left to top right is observable for the surface of $\hat{\beta}^1$ ($r = 0.871$, $p \leq 0.05$) as well as the hill and valley patterns of β_2^2 for the surface of $\hat{\beta}_2^2$ ($r = 0.708$, $p \leq 0.05$). However, all surfaces of the estimated coefficients show irregularities and noise.

Figure 12.8 shows the connection weights between the hidden neurons (including the bias neuron) and the output neurons of GWANN as surfaces. Some surfaces of GWANN show patterns that correspond to the coefficients of the original dataset. The linear trend of β_1 from bottom left to top right is visible for the surfaces of neurons 2 (Pearson's correlation coefficient $r = -0.906$, $p \leq 0.05$) and 5 ($r = -0.779$, $p \leq 0.05$), whereas the hill and valley patterns of β_2^2 are noticeable for the surfaces of neuron 4 ($r = 0.960$, $p \leq 0.05$). The surfaces of neuron 1 and 3 and the bias neuron, however, do not resemble any of the coefficient surfaces. Also, none of the neurons' surfaces shows the pattern of β_0 and all surfaces of GWANN exhibit substantial traces of irregularities and noise. With the exception of β_0, we can identify for each coefficient at least one surface of GWANN's connection weights with which it is more correlated than it is

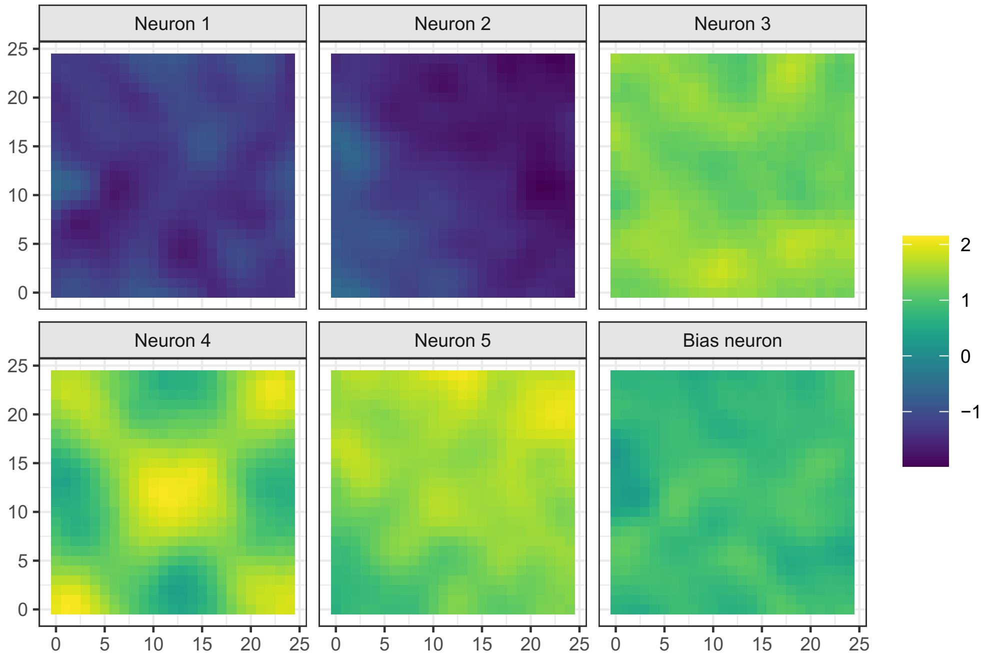


Figure 12.8.: Connection weights between the hidden neurons (including the bias neuron) and the output neurons of GWANN as surfaces.

with any surface of GWR's estimated coefficient.

While for this experiment the visualization of GWANN's surfaces provided evidence that the model learned the spatial relationships of the dataset, a detailed interpretation of the surfaces is difficult due to the complexity of the computations performed within the network. For instance, the surfaces of neurons 1 and 3 do not reveal how the neurons relate to the input data or how they contribute to the overall output of the network. Moreover, the computations performed within the network become less traceable the more input and hidden neurons the network consists of, which further limits the usefulness of GWANN's surfaces for explorative spatial data analysis in a practical setting.

12.3.2. Experiment 2: House prices in Austria

In this experiment, we assessed the differences in the predictive performance of GWR and GWANN using real-world data. We also investigated the effect of different distance matrices on the predictions of GWR and GWANN and evaluated the spatial distribution of the residuals.

We chose housing as a case study because in real-estate economics, regression-based house price assessments are vital (Sopranzetti, 2010). Hedonic theory assumes that a property represents a heterogeneous good that can be decomposed into its utility-bearing characteristics, and that the resulting benefit is reflected in the property price (Rosen, 1974). Both the physical characteristics of a property (e.g., floor area) and the neighborhood characteristics (i.e., a dwelling's surroundings) contribute to the overall price. It is well established in housing research that transaction prices vary spatially and thus hedonic house price models that consider spatial heterogeneity are increasingly applied (e.g., Bitter et al., 2007; Helbich and Griffith, 2016; Lu et al., 2011; Sunding and Swoboda, 2010).

Data

Data on $3,887$ geocoded single-family houses in Austria were provided by UniCredit Bank Austria AG (Helbich et al., 2014). Individual transaction prices of house purchases recorded in euros were collected from 1998 to 2009, along with 11 structural properties of the houses and two temporal variables. Descriptive statistics are listed in Table S1 in the supplementary materials.

Experimental setup

We used the log-transformed transaction prices as the dependent variable and the structural properties and temporal variables as independent variables. We used an adaptive bandwidth for GWANN and GWR, because of the uneven distribution of the housing locations (Figure 12.12). We applied two different distance metrics for geographical weighting, namely Euclidean distance (ED) and travel time distance by car (TTD). TTDs were computed using the Open Source Routing Machine (Huber and Rust, 2016) with OpenStreetMap data.

To investigate the performance of GWR and GWANN, we used 10-fold CV to obtain robust estimates of the performance of the models (outer 10-fold CV). Note that within each fold, 10-fold CV was also used to determine an appropriate bandwidth for GWR and GWANN and number of iterations for GWANN (inner 10-fold CV).

To investigate the predictions and residuals in detail, we built the models using the complete dataset. We chose the number of the networks' hidden neurons according to the number of hidden neurons for which the lowest mean RMSE had been obtained. Since we wanted to predict house prices for the complete dataset, the number of output neurons of GWANN equaled the total number of observations and each output neuron was assigned the location of an observation. Due to randomness in the (outer) 10-fold CV procedure and in the training of GWANN, a different bandwidth and a different

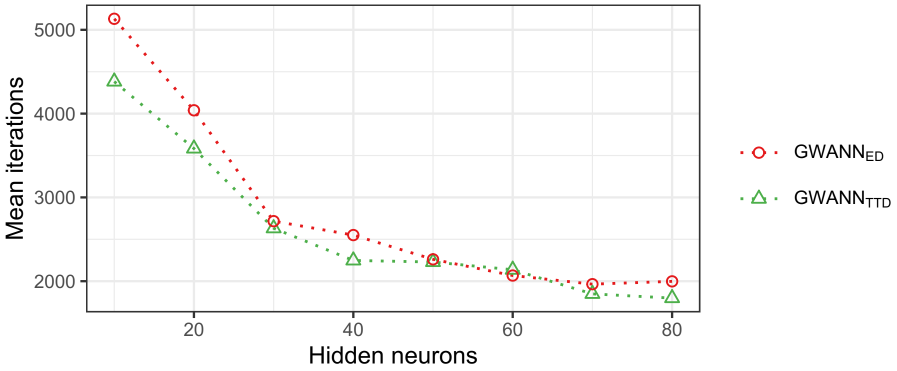


Figure 12.9.: Number of iterations until convergence.

number of training iterations were determined for most folds. We chose the bandwidth and number of iterations corresponding to the fold for which the median RMSE over all folds had been obtained.

If a model is unable to take into account the spatial properties of the data, its residuals tend to be spatially autocorrelated. We tested for residual spatial autocorrelation of the models using Moran's *I*. We calculated the test statistics using inverse EDs and evaluated the significance by means of 999 Monte-Carlo simulation runs.

Results & discussion

Figure 12.9 shows the mean number of training iterations of GWANN until convergence. GWANN generally requires fewer iterations to converge when using TTDs rather than EDs. The mean number of training iterations of GWANN when using EDs and TTDs decreases with the number of hidden neurons; the larger the number of hidden neurons, though, the smaller the decrease.

Figure 12.10 shows the obtained mean bandwidths for GWANN and GWR. While the mean bandwidth of GWANN is independent of the number of hidden neurons, the mean bandwidth is smaller when it uses TTDs rather than EDs. Similarly, the mean bandwidth of GWR is smaller when TTDs are used rather than EDs. Generally, the mean bandwidth of GWANN is considerably smaller than that of GWR, independent of the used distance metric and the number of hidden neurons. This result suggests that GWANN is generally able to model spatial variations in the data on a smaller scale than GWR.

Figure 12.11 shows the models' mean RMSEs obtained by means of (outer) 10-fold CV (for the proportion of explained variance, see Figure S2). While the mean RMSE of GWANN is lower when using EDs rather than TTDs, the difference in mean RMSE

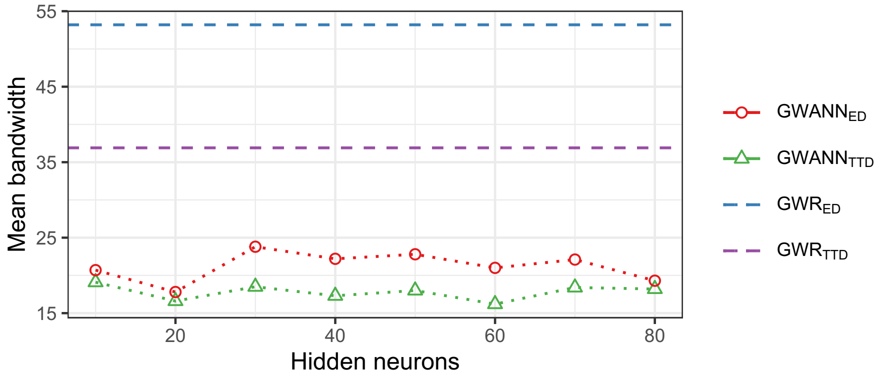


Figure 12.10.: Determined bandwidths (adaptive).

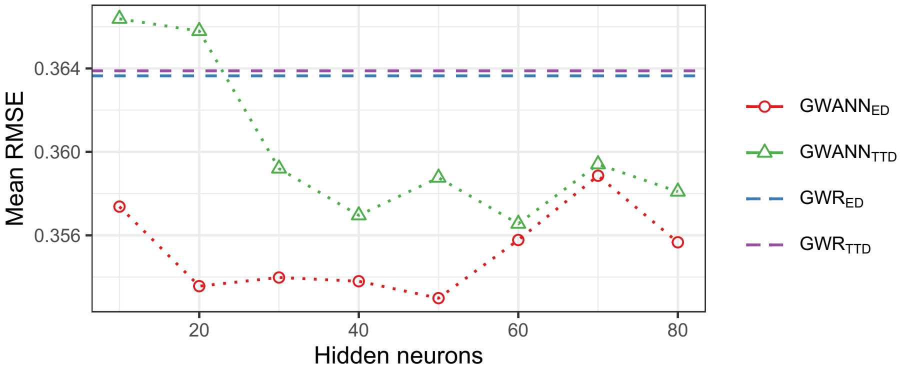


Figure 12.11.: Estimated prediction error (RMSE).

between the distance metrics for GWR is barely observable. This confirms the results of Lu et al. (2017), who also found no substantial difference in the goodness-of-fit of GWR between EDs and TTDs, and also indicates that the predictive performance of GWANN depends more on the choice of the distance metric than is the case with GWR. Moreover, with the exception of GWANN consisting of fewer than 30 hidden neurons and using TTDs, the mean RMSE of GWANN is always lower than that of GWR, independent of the distance metric used for model building. The overall lowest mean RMSE is obtained by GWANN when using EDs and 60 hidden neurons. The results demonstrate that GWANN can make better predictions than GWR when dealing with spatially heterogeneous relationships in a practical setting.

Using the complete dataset, we built GWR and GWANN using the following hyperparameters. When using EDs, GWR was fitted with a bandwidth of 53 and GWANN was trained with 50 hidden neurons and a bandwidth of 27 for 2,304 iterations. When using TTDs, GWR was fitted with a bandwidth of 36 and GWANN was trained with

60 hidden neurons and a bandwidth of 15 for 2,478 iterations.

To compare the influence of the chosen distance metric on the predictions, Figure 12.12 shows the differences in predicted house prices when using TTDs and EDs for GWANN and GWR. When using EDs rather than TTDs, GWANN predicts higher house prices for the city of Linz. For the Graz region, a stark contrast between the city and its surroundings is observable: GWANN predicts higher house prices for the city itself but lower house prices for the surroundings when using EDs rather than TTDs. For the metropolitan areas of Vienna, it can be seen that GWANN predicts higher prices in the eastern surroundings of the city and lower prices in the western surroundings when using EDs rather than TTDs. For the city of Salzburg, no differences in predicted house prices are observable. However, in the northern surroundings of Salzburg, substantially lower house prices are predicted when GWANN uses EDs rather than TTDs. For GWR the differences in predicted house prices resulting from the use of either EDs or TTDs are generally small and no spatial patterns are observable. These results demonstrate that in contrast to GWR, for GWANN the choice of the distance metric has a substantial effect on the spatial distribution of the predictions.

Figure 12.13 shows the Moran's I statistics of the models' residuals. For GWR the Moran's I values are smaller when using EDs rather than TTDs, while for GWANN they are smaller when using TTDs rather than EDs. Independent of the distance metric, the Moran's I values of GWANN are smaller than those of GWR. However, the residuals of both models do not reach statistical significance ($p > 0.05$), suggesting that both models take into account the spatial properties of the data appropriately.

12.4. Conclusion

We introduced GWANN — a method that combines ANNs and geographical weighting for modeling spatially heterogeneous relationships. We used synthetic and real-world data to compare GWANN with GWR. The results of the synthetic data showed that GWANN can have a better predictive performance than GWR when the relationships within the data are nonlinear and their spatial variance is high. The results based on the real-world data demonstrated that the predictive performance of GWANN can also be superior to that of the competing models in a practical setting.

Notwithstanding these promising results, this study had some limitations that should be considered when interpreting the findings or applying GWANN.

First, the results depended on the choice of the models' hyperparameters. While we followed common practices when choosing the hyperparameters and did careful sensitivity analysis, it cannot be guaranteed that we chose the most appropriate

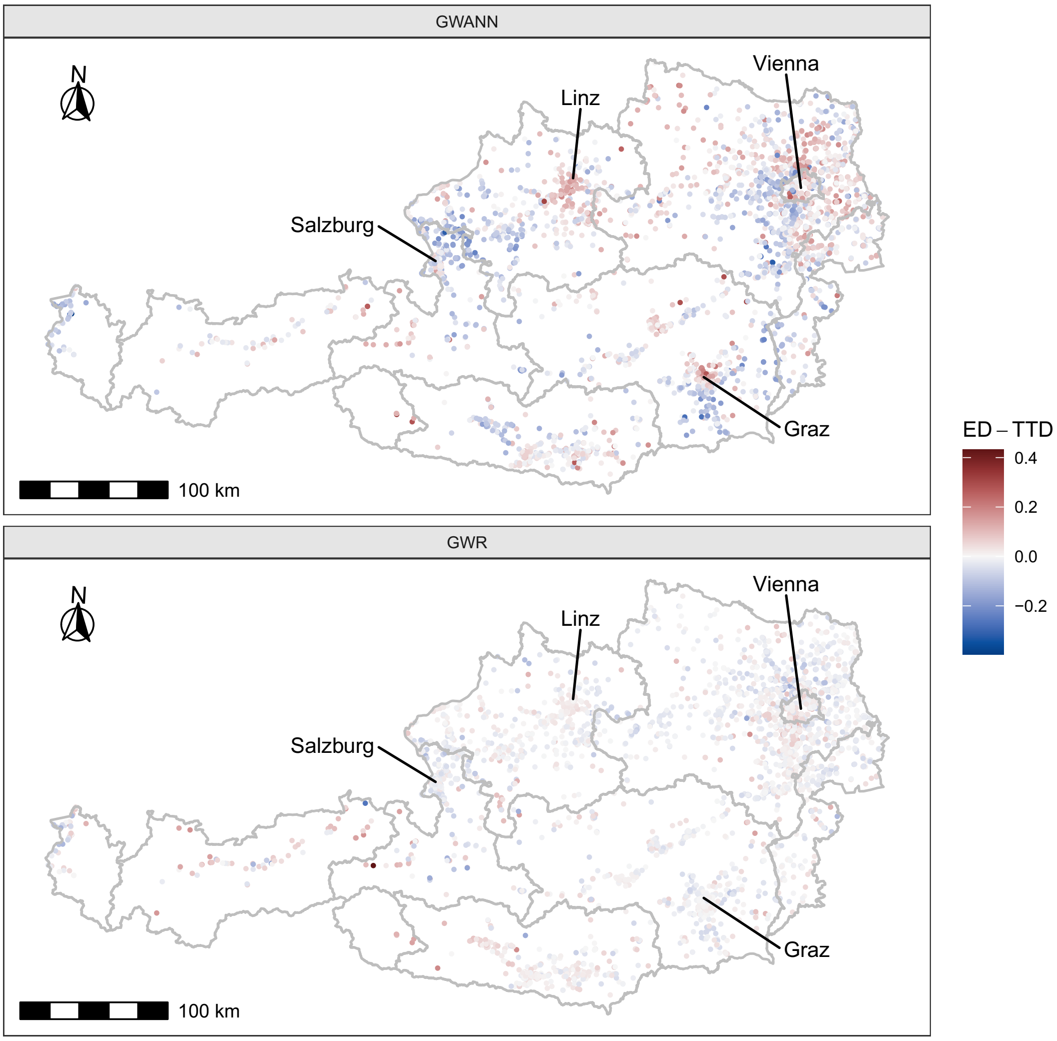


Figure 12.12.: Difference in predicted house prices when using EDs and TTDs for GWR and GWANN. Gray lines demarcate the federal states of Austria.

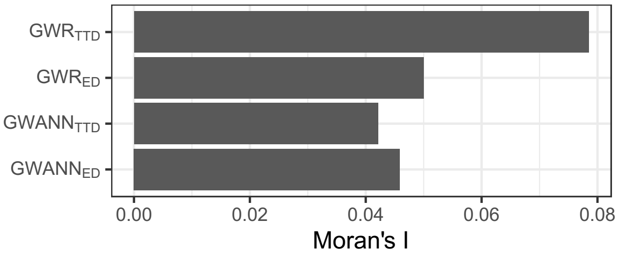


Figure 12.13.: Moran's I statistics of the residuals.

hyperparameters. Comprehensive sensitivity analysis is part of future analysis. Second, while the coefficient surfaces of GWR are useful for analyzing the modeled relationships, the complexity of the computations performed within the network of GWANN makes the interpretation of its surfaces difficult if not impossible. Third, in most practical applications GWANN consists of many output neurons (i.e., one output neuron for each location for which a prediction is to be made). Hence, because each output neuron is connected to each hidden neuron, the number of connection weights can be very large and the adjustment of the connection weights during the training require substantial computational resources. This is particularly a concern when searching for an appropriate bandwidth, which involves training and comparing numerous GWANNs with different bandwidths. More efficient heuristics for finding an appropriate bandwidth have the potential to mitigate this issue. Fourth, in the context of GWR, Fotheringham et al. (2017) showed that it is useful to model spatial heterogeneity at different scales by using individual bandwidths for the coefficients. While such an approach also has the potential to improve the predictive performance of GWANN, it remains open to further research as to whether and, if so, how it can be transferred to GWANN.

Data and codes availability statement

An R package that provides an implementation of GWANN, the source code, and the synthetic datasets can be downloaded from `https://github.com/jhagenauer/gwann`. The real-estate dataset cannot be shared publicly due to data protection restrictions.

Acknowledgement

We thank the anonymous reviewers for their constructive comments, which greatly improved this article. We also acknowledge UniCredit Bank Austria AG, in particular Wolfgang Brunauer, for providing the housing dataset. The opinions expressed by the authors do not reflect the official viewpoint of UniCredit Bank Austria AG.

Funding

This research is an outcome of the first author's research stay at Utrecht University in the course of the NEEDS project. The NEEDS project was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 714993). The funders had no role in the study design, data collection and analysis, interpretation, or dissemination.

Declaration of interest

None.

Conflict of interest

The authors have no conflict of interest to declare.

Bibliography

Anselin, L. (1989). *What is special about spatial data? Alternative perspectives on spatial data analysis*. Tech. rep. 89-4. UC Santa Barbara: National Center for Geographic Information and Analysis.

Bárcena, M. J., Menéndez, P., Palacios, M. B., and Tusell, F. (2014). Alleviating the effect of collinearity in geographically weighted regression. *Journal of Geographical Systems*, 16(4), 441–466.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In: G. Montavon, G. B. Orr, and K.-R. Müller (eds.). *Neural Networks: Tricks of the Trade: Second Edition*. Berlin Heidelberg: Springer, 437–478.

Bitter, C., Mulligan, G. F., and Dall'erba, S. (2007). Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems*, 9(1), 7–27.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1999). Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional science*, 39(3), 497–524.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4), 281–298.

Brunsdon, C., Fotheringham, S., and Charlton, M. (2007). Geographically weighted discriminant analysis. *Geographical Analysis*, 39(4), 376–396.

Casetti, E. (1972). Generating models by the expansion method: applications to geographical research. *Geographical Analysis*, 4(1), 81–91.

Choi, H. and Kim, H. (2017). Analysis of the relationship between community characteristics and depression using geographically weighted regression. *Epidemiology and health*, 39.

Comber, A., Chi, K., Huy, M. Q., Nguyen, Q., Lu, B., Phe, H. H., and Harris, P. (2020). Distance metric choice can both reduce and induce collinearity in geographically weighted regression. *Environment and Planning B: Urban Analytics and City Science*, 47(3), 489–507.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.

Du, Z., Wang, Z., Wu, S., Zhang, F., and Liu, R. (2020). Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. *International Journal of Geographical Information Science*, 34(7), 1353–1377.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships.* Chichester, UK: Wiley.

Fotheringham, A. S., Crespo, R., and Yao, J. (2015). Geographical and temporal weighted regression (GTWR). *Geographical Analysis*, 47(4), 431–452.

Fotheringham, A. S. and Oshan, T. M. (2016). Geographically weighted regression and multicollinearity: Dispelling the myth. *Journal of Geographical Systems*, 18(4), 303–329.

Fotheringham, A. S., Yang, W., and Kang, W. (2017). Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, 107(6), 1247–1265.

Gorr, W. L. and Olligschlaeger, A. M. (1994). Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market modeling. *Geographical Analysis*, 26(1), 67–87.

Griffith, D. A. (2003). *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization.* Springer Science & Business Media.

Guo, L., Ma, Z., and Zhang, L. (2008). Comparison of bandwidth selection in application of geographically weighted regression: A case study. *Canadian Journal of Forest Research*, 38(9), 2526–2534.

Hagenauer, J. and Helbich, M. (2018). Local modelling of land consumption in Germany with RegioClust. *International Journal of Applied Earth Observation and Geoinformation*, 65, 46–56.

Haykin, S. (2008). *Neural networks: A comprehensive foundation.* 3rd. Upper Saddle River, NJ: Prentice Hall.

Helbich, M., Brunauer, W., Vaz, E., and Nijkamp, P. (2014). Spatial heterogeneity in hedonic house price models: The case of Austria. *Urban Studies*, 51(2), 390–411.

Helbich, M. and Griffith, D. A. (2016). Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches. *Computers, Environment and Urban Systems*, 57, 1–11.

Huber, S. and Rust, C. (2016). Calculate travel time and distance with OpenStreetMap data using the Open Source Routing Machine (OSRM). *The Stata Journal*, 16(2), 416–423.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to spatial econometrics.* Statistics: A Series of Textbooks and Monographs. CRC Press.

Leuenberger, M. and Kanevski, M. (2015). Extreme Learning Machines for spatial environmental data. *Computers & Geosciences*, 85, 64–73.

Li, K. and Lam, N. S. N. (2018). Geographically weighted elastic net: A variable-selection and modeling method under the spatially nonstationary condition. *Annals of the American Association of Geographers*, 108(6), 1582–1600.

Lu, B., Brunsdon, C., Charlton, M., and Harris, P. (2017). Geographically weighted regression with parameter-specific distance metrics. *International Journal of Geographical Information Science*, 31(5), 982–998.

Lu, B., Charlton, M., and Fotheringhama, A. S. (2011). Geographically weighted regression using a non-Euclidean distance metric with a study on London house price data. *Procedia Environmental Sciences*, 7, 92–97.

Nelson, A., Oberthür, T., and Cook, S. (2007). Multi-scale correlations between topography and vegetation in a hillside catchment of Honduras. *International Journal of Geographical Information Science*, 21(2), 145–174.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269, 543–547.

Páez, A., Long, F., and Farber, S. (2008). Moving window approaches for hedonic price estimation: An empirical comparison of modelling techniques. *Urban Studies*, 45(8), 1565–1581.

Rojas, R. (2013). *Neural networks: A systematic introduction.* Springer Science & Business Media.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34–55.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.

Sopranzetti, B. J. (2010). Hedonic regression analysis in real estate markets: a primer. In: *Handbook of quantitative finance and risk management.* Springer, 1201–1207.

Sunding, D. L. and Swoboda, A. M. (2010). Hedonic analysis with locally weighted regression: An application to the shadow cost of housing regulation in southern California. *Regional Science and Urban Economics*, 40(6), 550–573.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In: *Proceedings of the 30th International Conference on Machine Learning.* Atlanta, Georgia, 1139–1147.

Troy, A., Grove, J. M., and O'Neil-Dunne, J. (2012). The relationship between tree canopy and crime rates across an urban-rural gradient in the greater Baltimore region. *Landscape and Urban Planning*, 106(3), 262–270.

Waller, L. A., Zhu, L., Gotway, C. A., Gorman, D. M., and Gruenewald, P. J. (2007). Quantifying geographic variations in associations between alcohol distribution and violence: A comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment*, 21(5), 573–588.

Wheeler, D. C. (2007). Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A*, 39(10), 2464–2481.

Wheeler, D. C. (2009). Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted lasso. *Environment and planning A*, 41(3), 722–742.

Wheeler, D. C. and Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7(2), 161–187.

Yu, W., Zang, S., Wu, C., Liu, W., and Na, X. (2011). Analyzing and modeling land use land cover change (LUCC) in the Daqing city, China. *Applied Geography*, 31(2), 600–608.

Appendix

Thesen der Habilitationsschrift

Problemstellung

1. Sowohl die Menge an zur Verfügung stehenden räumlichen Daten, als auch deren Komplexität, Veränderlichkeit und Diversität ist in der bisherigen Geschichte der Geowissenschaften einmalig.

2. Anders als herkömmliche statistische Verfahren sind räumliche maschinelle Lernverfahren besonders zur Anwendung auf solche Datensätze geeignet, da diese Verfahren komplexe Beziehungen ohne grundlegende a-priori zu treffende Annahmen in angemessener Berechnungszeit modellieren können.

3. Räumliche maschinelle Lernverfahren sind jedoch auch mit einigen Problemen behaftet, welche berücksichtigt werden müssen, um ihr Potential vollständig ausnutzen zu können.

4. Besonders hervorhebenswerte Probleme sind die Modellierung von räumlicher Autokorrelation und räumlicher Heterogenität, die Wahl eines geeigneten Lernverfahrens für ein gegebenes räumliches Problem und das Verständnis der inneren Zusammenhänge von komplexen räumlichen maschinell-gelernten Modellen.

Methodik und Forschungsansatz

5. Die Entwicklung von neuen maschinellen Lernverfahren für räumliche Vorhersage- und Clusterungsaufgaben, welche räumliche Autokorrelation und räumliche Heterogenität geeignet berücksichtigen, erlaubt eine genauere Modellierung von komplexen Zusammenhängen in räumlichen Daten.

6. Der umfassende Vergleich unterschiedlicher maschineller Lernverfahren für bestimmte räumliche Probleme offenbart nicht nur, welche Modelle für das zugrundeliegende Problem vielversprechend sind und bietet tiefere Einsicht in dieses, sondern legt auch geeignete Modelle für ähnlich gelagerte räumliche Probleme nahe.

7. Die Untersuchung von unterschiedlichen Ansätzen, welche das Verständnis von maschinellen Lernverfahren unterstützen, ermöglicht die Identifizierung von Ansätzen, welche besonders geeignet sind, vielversprechende Hypothesen zu generieren und dadurch die Entdeckung von neuem Wissen unterstützen.

8. Die Fundierung der Forschungsarbeiten auf Anwendungen aus so unterschiedlichen Feldern wie Gesundheit, Verkehr, Immobilien und Landnutzung bietet eine umfängliche Perspektive auf grundlegend unterschiedliche räumliche Probleme.

Ergebnisse und Schlussfolgerungen

9. Die Berücksichtigung von räumlicher Autokorrelation und/oder räumlicher Heterogenität führt häufig zu einer verbesserten Modellgüte von maschinellen Lernverfahren.
10. Baum-basierte Ensemble-Methoden wie „random forest“ oder „gradient boosting“ sind häufig besonders vielversprechend für die Anwendung in räumlichen Vorhersageaufgaben.
11. Reproduzierbarkeit von Vergleichsstudien ist eine wichtige Eigenschaft um weitere Forschung im Bereich des räumlichen maschinellen Lernens zu begünstigen.
12. Bei der Anwendung räumlicher maschineller Lernverfahren ist die Bestimmung geeigneter Hyperparameter eine entscheidende aber auch komplexe Aufgabe, welche die Modellgüte maßgeblich beeinflussen
13. Die Möglichkeit die Parameter von maschinellen Lernverfahren mittels geographischer Karten zu visualisieren ist besonders nützlich um gelernte räumlichen Beziehungen zu kommunizieren und zu verstehen.
14. Die Flexibilität und Anpassbarkeit von räumlichen maschinellen Lernverfahren, insbesondere von künstlichen neuronalen Netzen, begünstigt ihre Anwendung zur Lösung wichtiger räumliche Probleme.
15. Räumliche maschinelle Lernverfahren sind meist komplexer als nicht-räumliche Verfahren und stellen gesonderte Anforderungen an den Anwender.
16. Räumliche maschinelle Lernverfahren sind nur so nützlich, wie die zugrundeliegenden Daten, welche für das Lernen verwendet werden, repräsentativ sind, das Lernverfahren in der Lage ist, wichtige Datenbeziehungen zu modellieren und der Analyst fähig ist, die geeigneten Entscheidungen für das Lernen zu treffen.
17. Räumliche maschinelle Lernverfahren ersetzen nicht traditionelle statistische Methoden, sondern erweitern den „Werkzeugkasten“ von räumlichen Wissenschaftlern um Verfahren, welche für komplexe und datenreiche räumliche Anwendungen besonders nützlich sind.

Ausblick

18. Die angesprochenen Probleme stellen nur einen kleinen jedoch wichtigen Teil der Probleme dar, mit denen räumliche maschinelle Lernverfahren konfrontiert sind.
19. Die Komplexität der meisten räumlichen Probleme auf der einen Seite und die Flexibilität und Anpassbarkeit von räumlichen maschinellen Lernverfahren auf der anderen Seite bergen umfängliches Potential für weitere Forschungen.
20. Weitere Forschungen im Bereich räumlicher maschineller Lernverfahren bieten die Möglichkeit der Gewinnung neuer Einsichten über räumliche Phänomene und können somit helfen unser Verständnis der Welt in der wir leben zu verbessern.