Transactions in GIS

Research Article Transactions in GIS, 2016, 20(5): 755–774

SPAWNN: A Toolkit for *SP*atial *A*nalysis *W*ith Self-Organizing *N*eural *N*etworks

Julian Hagenauer* and Marco Helbich†

*Leibniz Institute of Ecological Urban and Regional Development
†Department of Human Geography and Planning, Utrecht University

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.

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

Address for correspondence: Julian Hagenauer, Leibniz Institute of Ecological Urban and Regional Development, Weberplatz 1, 01217 Dresden, Germany. E-mail: j.hagenauer@ioer.de
[Correction added on 22 February 2016, after Online publication in 2 February 2016: The affiliation address of author Julian Hagenauer was previously incorrect on first publication due to a production error and has been corrected in this current online version.]

© 2016 John Wiley & Sons Ltd doi: 10.1111/tgis.12180

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 *SP*atial *A*nalysis *W*ith self-organizing *N*eural *N*etworks, which implements the self-organizing map (SOM) (Kohonen 1982, 2001) and neural gas (NG) (Martinetz and Schulten 1991; 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 2 reviews existing toolkits for spatial cluster analysis. Section 3 discusses self-organizing neural networks, while Section 4 introduces different models for incorporating spatial dependence into the networks. The SPAWNN toolkit is presented in Section 5. In Section 6 the computational demand of the implemented spatial context models is analyzed, while Section 7 illustrates the application of the SPAWNN toolkit to practical analytical problems. Section 8 concludes the article and discusses future work.

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

© 2016 John Wiley & Sons Ltd

(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 choropleth 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. Flexer 2001; Estévez and Figueroa 2006; 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

© 2016 John Wiley & Sons Ltd

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 abovementioned 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. Strickert and Hammer 2005; Hagenauer and Helbich 2013). 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).

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. Kaski et al. 1998; Oja et al. 2003; Kalteh et al. 2008; Chon 2011; Liu and Weisberg 2011). This subsection briefly introduces the SOM and the NG algorithm, a closely related algorithm that is particularly useful for clustering tasks.

3.1 Self-Organizing Map

The SOM (Kohonen 1982, 2001) 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

© 2016 John Wiley & Sons Ltd

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.

3.2 Neural Gas

Similar to the SOM, the NG algorithm (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 (Martinetz et al. 1993; Cottrell et al. 2006). 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 (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 (Martinetz and Schulten 1991; 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).

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 (e.g. Bavaud 1998; Getis 2009). Instead, with the exception of Weighted Merge Context (WMC) (Hagenauer 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 (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

© 2016 John Wiley & Sons Ltd

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.

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.

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.

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. (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). 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.

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.

© 2016 John Wiley & Sons Ltd

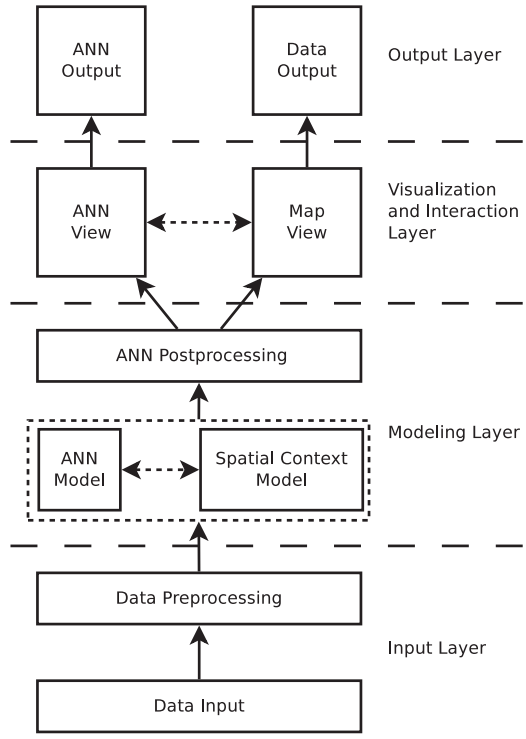

Figure 1 General architecture of the SPAWNN toolkit

4.5 Weighted Merge Context

Weighted Merge Context (WMC) (Hagenauer 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.

5 The SPAWNN Toolkit

The general architecture of the SPAWNN toolkit, depicted in Figure 1, is organized in four layers, which roughly correspond to the well-known steps in the process of knowledge discovery in databases (Fayyad et al. 1996). While Figure 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 (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

© 2016 John Wiley & Sons Ltd

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 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 (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 4). Figure 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 presented to the user (see Figure 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

© 2016 John Wiley & Sons Ltd

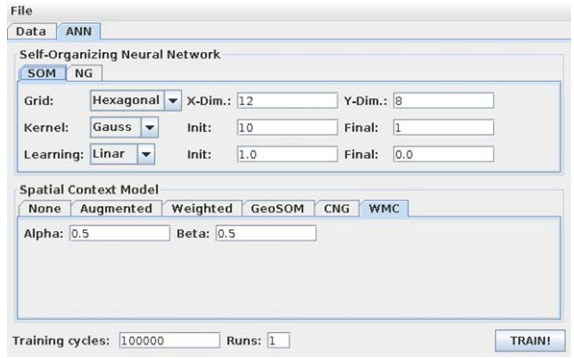

Figure 2 Self-organizing neural network and spatial context model selection window

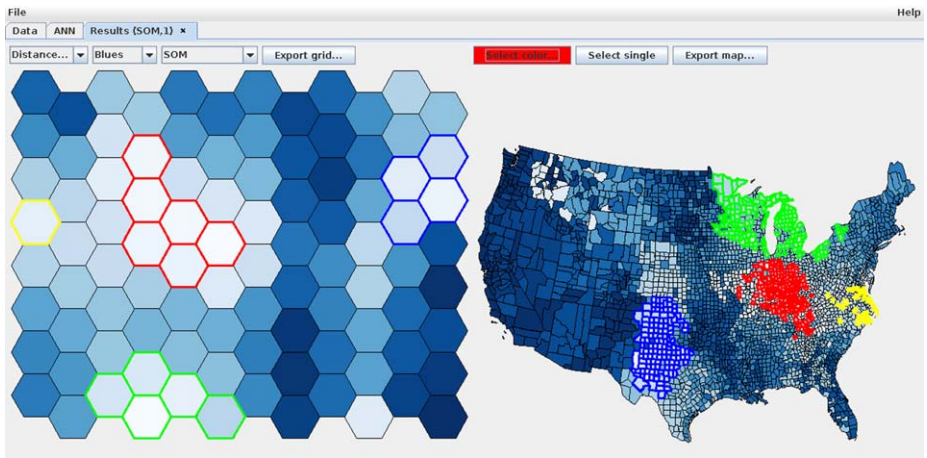

Figure 3 Linked view of a self-organizing neural network (left) and a geographic map of the data (right). The colored neurons of the network correspond to the colored observations on the geographic map

neurons' prototype vectors (component plane) or represent a distance-related statistic of each neurons to its immediate neighbors (distance matrix or distance-based representation; 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

© 2016 John Wiley & Sons Ltd

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.

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 4 shows the mean computation times over 32 runs for $1 \leq n \leq 25$ and $m = 2$, while Figure 5 shows these for $n = 5$ and $1 \leq m \leq 25$. In more detail, Figure 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

© 2016 John Wiley & Sons Ltd

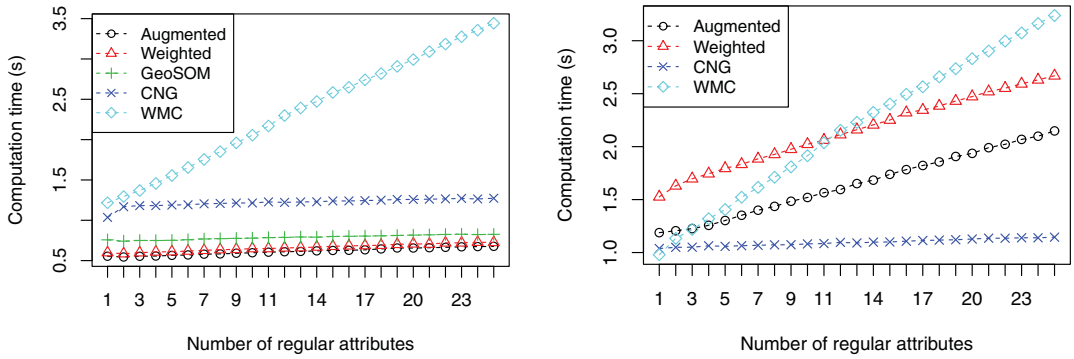

Figure 4 Mean computation times of the spatial context models for different numbers of regular attributes

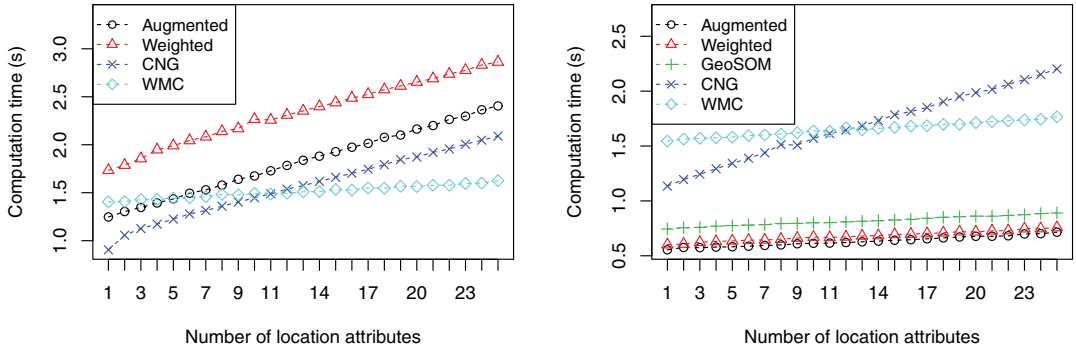

Figure 5 Mean computation times of the spatial context models for different numbers of location attributes

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.

Concerning the number of location attributes, Figure 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.

© 2016 John Wiley & Sons Ltd

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 km^2. 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.

7.1 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 6), outliers can be identified by neurons that have high median distance to neighboring values 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), 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

© 2016 John Wiley & Sons Ltd

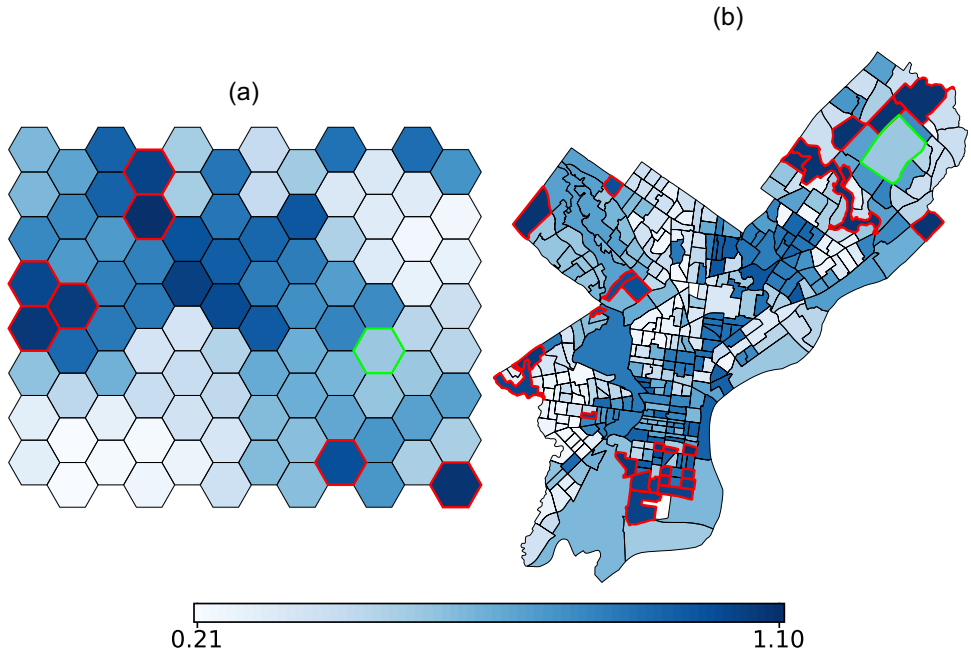

Figure 6 Distance matrix (a) and cartographic map (b) of the GeoSOM. Identified outliers are outlined in red. The neuron that maps the census tract where the Northeast Philadelphia Airport is located and the tract itself are outlined in green. For clarity, other tracts that this particular neuron maps are not outlined

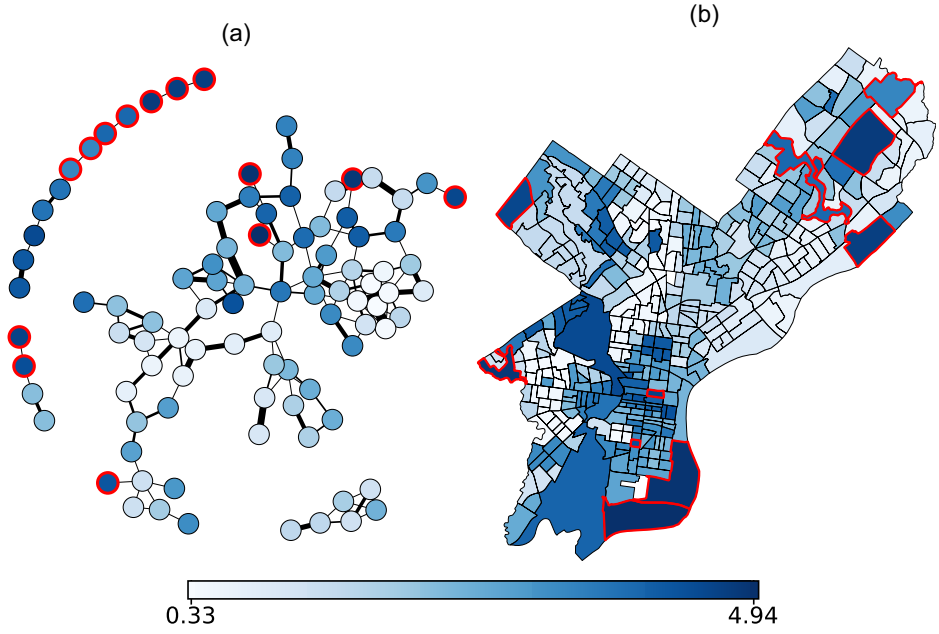

Figure 7 Distance-based representation of the CNG (a) and the geographic map (b). Identified outliers are outlined in red

© 2016 John Wiley & Sons Ltd

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 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 rather small (the neuron is also colored in green in Figure 6). In contrast, the census tract is clearly identifiable as an outlier on the CNG representation (see Figure 7). Hence, it can be concluded from this section that the CNG is more appropriate for identifying outliers than the GeoSOM.

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. Vesanto and Ahola 1999; Barreto-Sanz and Pérez-Uribe 2007). 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 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 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.

7.3 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.

© 2016 John Wiley & Sons Ltd

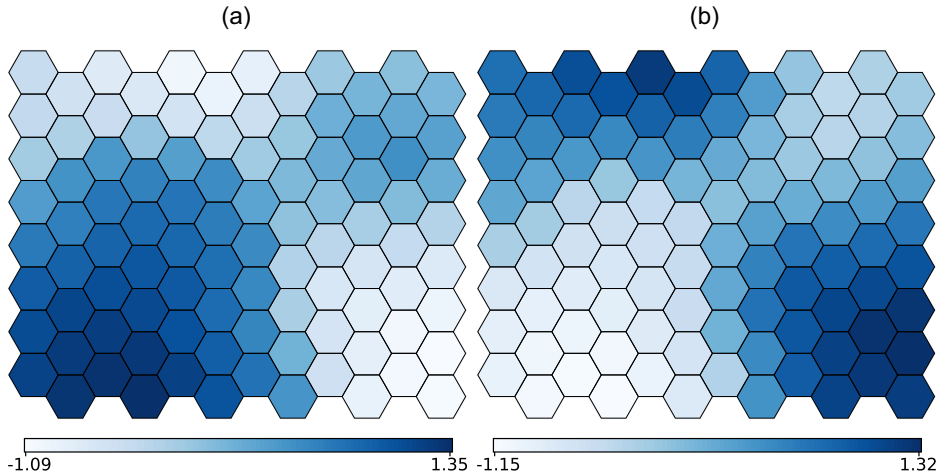

Figure 8 GeoSOM component planes for the rates of African Americans (a) and white population (b)

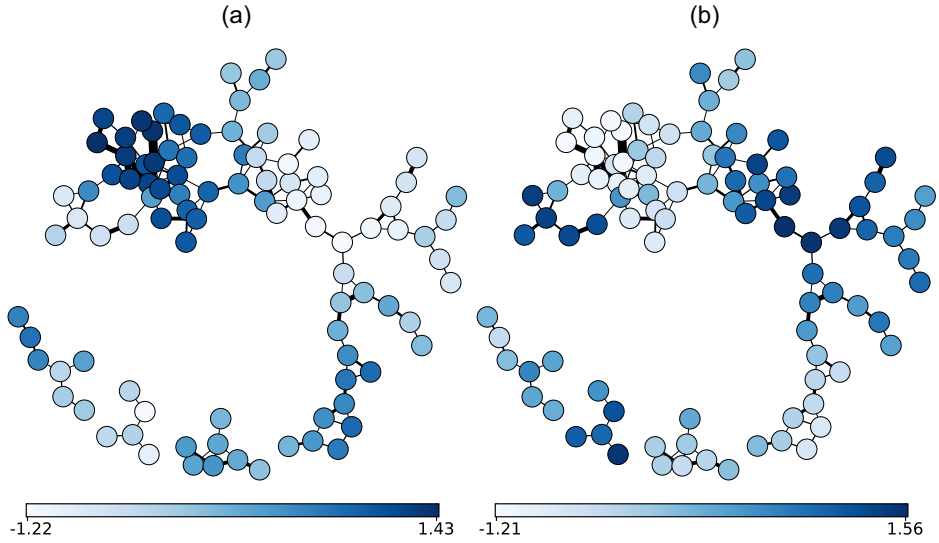

Figure 9 Neurons of the CNG, colored according to the rates of African Americans (a) and white population (b)

In order to cluster the CNG and GeoSOM, contiguity-constrained hierarchical clustering using Ward's criterion is applied (Murtagh 1995). Figure 10 depicts the results for the clustering of the CNG, while the results for the GeoSOM are shown in Figure 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.

© 2016 John Wiley & Sons Ltd

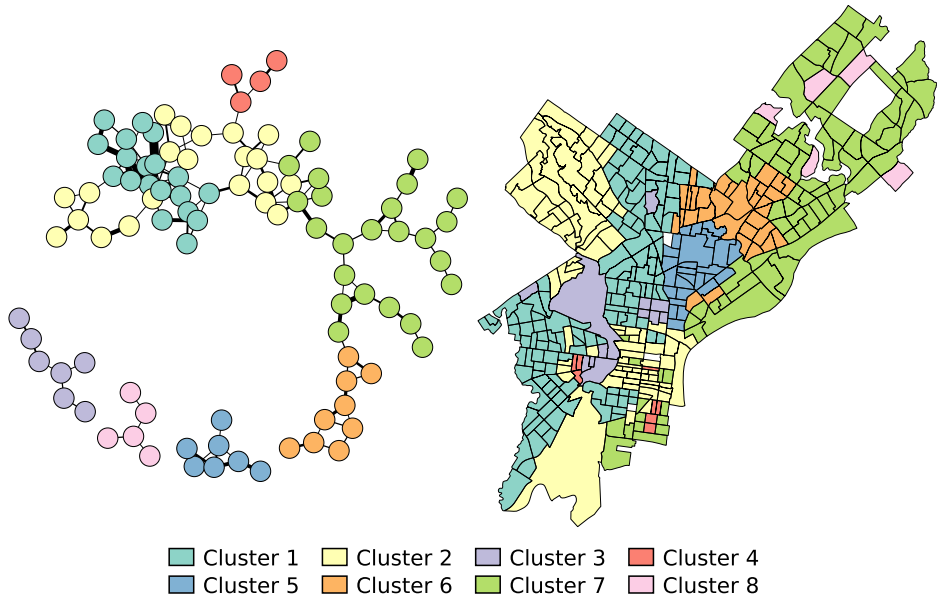

Figure 10 Clustering results for CNG

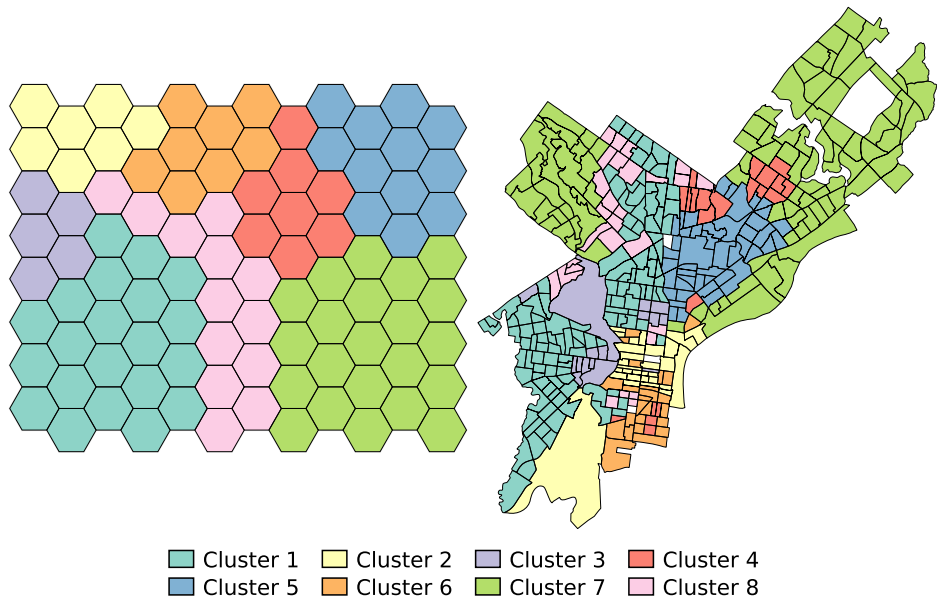

Figure 11 Clustering results for the GeoSOM

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 11, left panel) with the component plane of African American Population in Figure 8a, reveals that the

© 2016 John Wiley & Sons Ltd

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, 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 (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.

8 Conclusions

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 (Hagenauer et al. 2011; Helbich et al. 2013b), health (Augustijn and Zurita-Milla 2013), real estate (Helbich et al. 2013a), and ecology (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

© 2016 John Wiley & Sons Ltd

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.

References

Adams C 1993 *Philadelphia: Neighborhoods, Division, and Conflict in a Postindustrial City*. Philadelphia, PA, Temple University Press

Andrienko G, Andrienko N, Bremm S, Schreck T, van Landesberger T, Bak P, and Keim D 2010 Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum* 29: 913–22

Anselin L 1995 Local indicators of spatial association: LISA. *Geographical Analysis* 27: 93–115

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

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

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

Bacao F, Lobo V, and Painho M 2005 The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computational Geosciences* 31: 155–63

Barreto-Sanz M A and Perez-Uribe A 2007 Improving the correlation hunting in a large quantity of SOM component planes. In de Sa J M, Alexandre L, Duch W, and Mandic D (eds) *Artificial Neural Networks: ICANN 2007*. Berlin, Springer Lecture Notes in Computer Science Vol. 4669: 379–88

Bavaud F 1998 Models for spatial weights: A systematic look. *Geographical Analysis* 30: 153–71

Becker S 1991 Unsupervised learning procedures for neural networks. *International Journal of Neural Systems* 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* 10: P10008

Brandes U, Eiglsperger M, Herman I, Himsolt M, and Marshall M S 2002 GraphML progress report: Structural layer proposal. In Mutzel P, Jünger M, and Leipert S (eds) *Graph Drawing (GD 2001)*. Berlin, Springer Lecture Notes in Computer Science Vol. 2265: 501–12

Carpenter G A and Grossberg S 1991 *Pattern Recognition by Self-organizing Neural Networks*. Cambridge, MA, MIT Press

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

Cottrell M, Hammer B, Hasenfus A, and Villmann T 2006 Batch and median neural gas. *Neural Networks* 19: 762–71

Estevez P A and Figueroa C J 2006 Online data visualization using the neural gas network. *Neural Networks* 19: 923–34

Fayyad U, Piatetsky-Shapiro G, and Smyth P 1996 From data mining to knowledge discovery: An overview. In Fayyad U M, Piatetsky-Shapiro G, Smyth P, and Uthurusamy R (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: 1873–91

Flexer A 2001 On the use of self-organizing maps for clustering and visualization. *Intelligent Data Analysis* 5: 373–84

Fruchterman T M and Reingold E M 1991 Graph drawing by force-directed placement. *Software: Practice and Experience* 21: 1129–64

Getis, A 2009 Spatial weights matrices. *Geographical Analysis* 41(4): 404–10

Girvan M and Newman M E 2002 Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, USA* 99: 7821–26

Goodchild, M. F 1986 *Spatial Autocorrelation*. Norwich, UK, 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: 1134–56

© 2016 John Wiley & Sons Ltd

Guo D 2008 Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science* 22: 801–23

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: 1461–74

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: 113–32

Hagenauer J 2014 Clustering contextual neural gas: A new approach for spatial planning and analysis tasks. In Helbich M, Jokar Arsanjani J, and Leitner M (eds) *Computational Approaches for Urban Environments*. Berlin, Springer: 77–94

Hagenauer J 2015 Weighted merge context for clustering and quantizing spatial data with self-organizing neural networks. *Journal of Geographical Systems* 17: in press

Hagenauer J and Helbich M 2013 Contextual neural gas for spatial clustering and analysis. *International Journal of Geographical Information Science* 27: 251–66

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 Twenty-fifth International Cartographic Conference*, Paris, France

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 1996 DBMiner: A system for mining knowledge in large relational databases. In *Proceedings of the International Conference on Data Mining and Knowledge Discovery*, Portland, Oregon: 250–55

Han J, Koperski K, and Stefanovic N 1997 GeoMiner: A system prototype for spatial data mining. *ACM SIGMOD Record* 26: 553–56

Harrower M and Brewer C A 2003 Colorbrewer.org: An online tool for selecting colour schemes for maps. *Cartographic Journal* 40: 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: 871–89

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: 326–36

Henriques R, Bacao F, and Lobo V 2012 Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems* 36: 218–32

Jain A K 2010 Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31: 651–66

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 and Software* 23: 835–45

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

Kangas J 1992 Temporal knowledge in locations of activations in a self-organizing map. In Aleksander I and Taylor J (eds) *Artificial Neural Networks*, 2. Amsterdam, Netherlands, North-Holland: 117–20

Kaski S 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: 1–8

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

Kohonen T 2001 *Self-Organizing Maps*. New York, Springer

Korting T S, Garcia Fonseca L M, and Camara G 2013 GeoDMA–geographic data mining analyst. *Computers and Geosciences* 57: 133–45

Lacayo-Emery M A 2011 SOM Analyst. WWW document, https://github.com/mlacayoemery/somanalyst

Liu Y and Weisberg R H 2011 *A Review of Self-organizing Map Applications in Meteorology and Oceanography*. Rijeka, Croatia, InTech

Logan J R 2011 Separate and unequal: The neighborhood gap for Blacks, Hispanics and Asians in Metropolitan America. WWW document, http://www.s4.brown.edu/us2010/Data/Report/report0727.pdf

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: 311–34

Martinetz T 1993 Competitive hebbian learning rule forms perfectly topology preserving maps. In Gielen S and Kappen B (eds) *ICANN '93*. London, Springer: 427–34

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: 558–69

Martinetz T and Schulten K 1991 A "neural-gas" network learns topologies. *Artificial Neural Networks* 1: 397–402

© 2016 John Wiley & Sons Ltd

Mayer R, Dittenbach M, Frank J, Neumayer R, and Lidy T 2011 Data mining with the Java SOMToolbox. WWW document, http://www.ifs.tuwien.ac.at/dm/somtoolbox/

Mennis J and Guo D 2009 Spatial data mining and geographic knowledge discovery: An introduction. *Computers, Environment and Urban Systems* 33: 403–08

Miller H J and Goodchild M F 2014 Data-driven geography. *GeoJournal* 80: 449–61

Miller H J and Han J 2009 *Geographic Data Mining and Knowledge Discovery*. Boca Raton, FL, CRC Press

Munoz A and Muruzabal J 1998 Self-organizing maps for outlier detection. *Neurocomputing* 18: 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: 649–67

Murtagh F 1995 Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Letters* 16: 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 *Proceedings of the Fourth International Conference on Geocomputation*, Fredericksburg, Virginia (CD-ROM)

Parimala M, Lopez D, and Senthilkumar N 2011 A survey on density based clustering algorithms for mining large spatial databases. *International Journal of Advanced Science and Technology* 31: 59–66

Rey S J 2009 Show me the code: Spatial analysis and open source. *Journal of Geographical Systems* 11: 191–207

Skupin A and Agarwal P 2008 Introduction: What is a self-organizing map? In Agarwal P and Skupin A (eds) *Self-organising Maps: Applications in Geographic Information Science*. New York, John Wiley and Sons: 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*, Fukuoka, Japan: 53–57

Strickert M and Hammer B 2005 Merge SOM for temporal data. *Neurocomputing* 64: 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: 269–77

Takatsuka M and Gahegan M 2002 GeoVISTA studio: A codeless visual programming environment for geoscientific data analysis and visualization. *Computers and Geosciences* 28: 1131–44

Tasdemir K and Merenyi E 2009 Exploiting data topology in visualization and clustering of self-organizing maps. *IEEE Transactions on Neural Networks* 20: 549–62

Vesanto J 1999 SOM-based data visualization methods. *Intelligent Data Analysis* 3: 111–26

Vesanto J and Ahola J 1999 Hunting for correlations in data using the self-organizing map. In *Proceedings of the International ICSC Congress on Computational Intelligence Methods and Applications*, Rochester, New York: 279–85

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: 583–98

Ware C 2012 *Information Visualization: Perception for Design*. Oxford, UK, Elsevier

Yuan M, Buttenfield B, Gahegan M, and Miller H 2004 Geospatial data mining and knowledge discovery. In McMaster R B and Usery E L (eds) *A Research Agenda for Geographic Information Science*. Boca Raton, FL, CRC Press: 365–88

© 2016 John Wiley & Sons Ltd