Effects of Irregular Topology in Spherical Self-Organizing Maps

Charles R. Schmidt
MS Candidate
Department of Geography
San Diego State University

September 19, 2007

Committee Members
Dr. Sergio Rey, Geography
Dr. André Skupin, Geography
Dr. Robert Malouf, Linguistics

A Thesis Proposal Presented to the Faculty of San Diego State University.

Contents

1 Introduction **3**

2 Research Objectives **3**

3 Background **4**
3.1 Training and the Boundary Effect 4
3.2 Spherical SOM 4
3.3 Network Size 6
3.4 Uniformity 6

4 Data and Methodology **8**
4.1 Diagnostics 8
4.1.1 Internal variance vs. first-order neighborhood size 8
4.1.2 Internal variance vs. topological regularity 9
4.1.3 Visualize internal variance mapping 9
4.2 Empirical Analysis 9
4.2.1 Synthetic Data 10
4.2.2 Training 10
4.2.3 Diagnostics 10

5 Significance and Limitations **10**
5.1 Significance 10
5.2 Limitations 11

PROJECT SUMMARY

The Self-Organizing Map (SOM) is a widely used technique in information visualization and exploration. The development of the spherical SOM has been driven by the border effects observed in traditional SOM. This project will (1) investigate how irregular network topologies affect the SOM, (2) examine any benefits spherical topologies offer over traditional planar topologies, and (3) examine the trade off between increasingly regular topology versus greater control over network size.

Intellectual Merit

I introduce a new method for investigating topologically induced errors. The method first analyzes the neural network to find topological mismatches, then trains the network with synthetic data to produce a comparable measure for each neuron's internal variance. I hypothesize that edge neurons will display greater internal variances.

Broader Impact

This project may offer a greater understanding of how irregularities in network topology affect the SOM. This will help serve as a guide for researchers in choosing a suitable network topology for their SOM.

1 Introduction

The Self-Organizing Map (SOM) is an unsupervised competitive learning process developed by Teuvo Kohonen as a technique to analyze and visualize high dimensional data sets. The applications of SOM are far reaching; Kohonen (2000) provides a thorough review of the SOM literature including applications of SOM. SOM has been used in applications ranging from speech recognition and image classification to breast cancer detection and gene expression clustering. Skupin and Agarwal (forthcoming) outline the growing interest of SOM in the GISciences, and propose that the relationship between SOM and GIScience should be bidirectional. The SOM offers a powerful method for exploring and visualizing geographic data and GIScience offers a wide array of tools and methods to enable the exploration of the SOM itself. This thesis will take advantage of GeoVisualization and GeoCompution in order to explore some of the basic properties of the SOM.

The SOM is a type of artificial neural network in which neurons are "organized" in such a way as to project the high-dimensional relationships of a set of training data onto a low-dimensional network structure. The traditional SOM uses a rectangular or hexagonal network topology (Kohonen, 2000). These topologies create a well-known problem in SOM called the boundary or edge effect. Neurons on the boundary of the hexagonal and rectangular lattices have fewer neighbors, which reduces their ability to interact with other neurons during the self-organizing process. Using a spherical lattice has been widely suggested as a solution to the problem (Ritter, 1999; Boudjemai et al., 2003; Sangole and Knopf, 2003; Nishio et al., 2006; Wu and Takatsuka, 2006). The use of the spherical lattice, however, does not completely overcome the boundary problem, and the choice of which spherical topology to use for the network can be difficult to make.

A regular network topology is one in which every node on the network has exactly the same number of adjacent nodes. Any topology involving an edge is irregular. Arranging our lattice on the surface of a sphere seems to be an obvious way to overcome the edge. However, there exist only five arrangements on the sphere which are completely regular; these are the five platonic solids (Ritter, 1999; Harris et al., 2000). Any other arrangement of neurons on the surface of the sphere will result in an irregular topology, as not all neurons will have the same number of neighbors. The classic method for minimizing this irregularity is to generate the spherical lattice by tessellating the sides of the icosahedron (Nishio et al., 2006). While this method will always result in a highly regular spherical topology, the main drawback is that the number of neurons in the network (the network size), N, grows exponentially as tessellations are applied. That results in only very coarse control over network size. Other methods for arranging neurons on the sphere allow for unlimited control over network size, but yield topologies with increased irregularity (Harris et al., 2000; Wu and Takatsuka, 2005; Nishio et al., 2006). To date the literature has largely ignored the more irregular methods in favor of the aforementioned tessellation-based methods. A topology which yields a more flexible network size may be desirable. However, in order to address this issue of network size, we must first determine the degree to which irregularity effects the SOM.

2 Research Objectives

The objective of this research is to determine the utility of certain irregular spherical topologies beyond offering greater control over network size. Toward that end, I will develop and test new diagnostics to measure and visualize topology-induced errors in SOM. The following diagnostics will be developed and implemented:

1. Compare the internal variance of observations captured by a given neuron to that neuron's first-order neighborhood size.

2. For different topologies, compare the internal variance of each neuron against a composite measure of topological regularity.

3. Develop a SOM-based visualization of the internal variance.

These diagnostics will help facilitate the evaluation of both traditional and spherical SOMs. To satisfy the objective of this research, I will apply these diagnostics to a series of comparable SOMs. Each SOM will be trained using the same synthetic data and training parameters, but will utilize different network topologies. By formally testing for difference of means and variance in the results of the diagnostics, the following questions will be addressed:

1. Does the internal variance of a neuron decrease as its first-order neighborhood size, or degree, increases?

2. Is the average internal variance of a SOM higher when a more irregular topology is used?

3. Which insights, if any, can be gained from a SOM-based visualization of internal variance?

3 Background

This section is divided into four parts. The first provides a general introduction to the SOM algorithm and the problems created by using irregular topology. The second reviews the current spherical topologies used with SOMs. The third examines the flexibility of the various topologies with regard to network size. The fourth takes a look at the limitations of using "uniformity" to evaluate potential topologies.

3.1 Training and the Boundary Effect

The SOM algorithm uses an artificial neural network to organize high-dimensional data onto a low-dimensional lattice, or map, of neurons. Each neuron contains a reference vector that models the input data. Before training, these neuron-vectors are initialized, most commonly to random values. During the training process a randomly selected observation (input vector x) searches all neurons (reference vectors m_i) to find the one to which it is most similar, referred to as its Best Matching Unit (BMU c). The BMU and its neighborhood (N_c) are then adjusted to better match that observation (Kohonen, 2000). The training process is repeated a predefined number of times, or ideally until the map converges. The traditional SOM is laid out on a two-dimensional plane using either a rectangular or hexagonal topology. According to Wu and Takatsuka (2006), the hexagonal structure is more uniform and generally preferred.

One drawback of building the neural lattice in a discrete Euclidean plane is the boundary of the resulting lattice. A neuron located on the boundary has fewer neighbors and thus fewer chances of being updated (Wu and Takatsuka, 2006). As observed in Figure 1, neurons in the center of the map tend to better represent the mean of the input-space. This is arguably caused by outliers being pushed to the edges of the map, where they encounter fewer competing signals.

3.2 Spherical SOM

One way to eliminate the boundary effect is to wrap the lattice around a three-dimensional object such as a sphere or torus, thereby removing the edge entirely. The toroidal SOM was introduced by Li et al. (1993), however the torus is not effective for visualization, as maps generated from a torus are not very intuitive (Ito et al., 2000; Wu and Takatsuka, 2006). Ritter (1999) describes the torus as being topologically flat

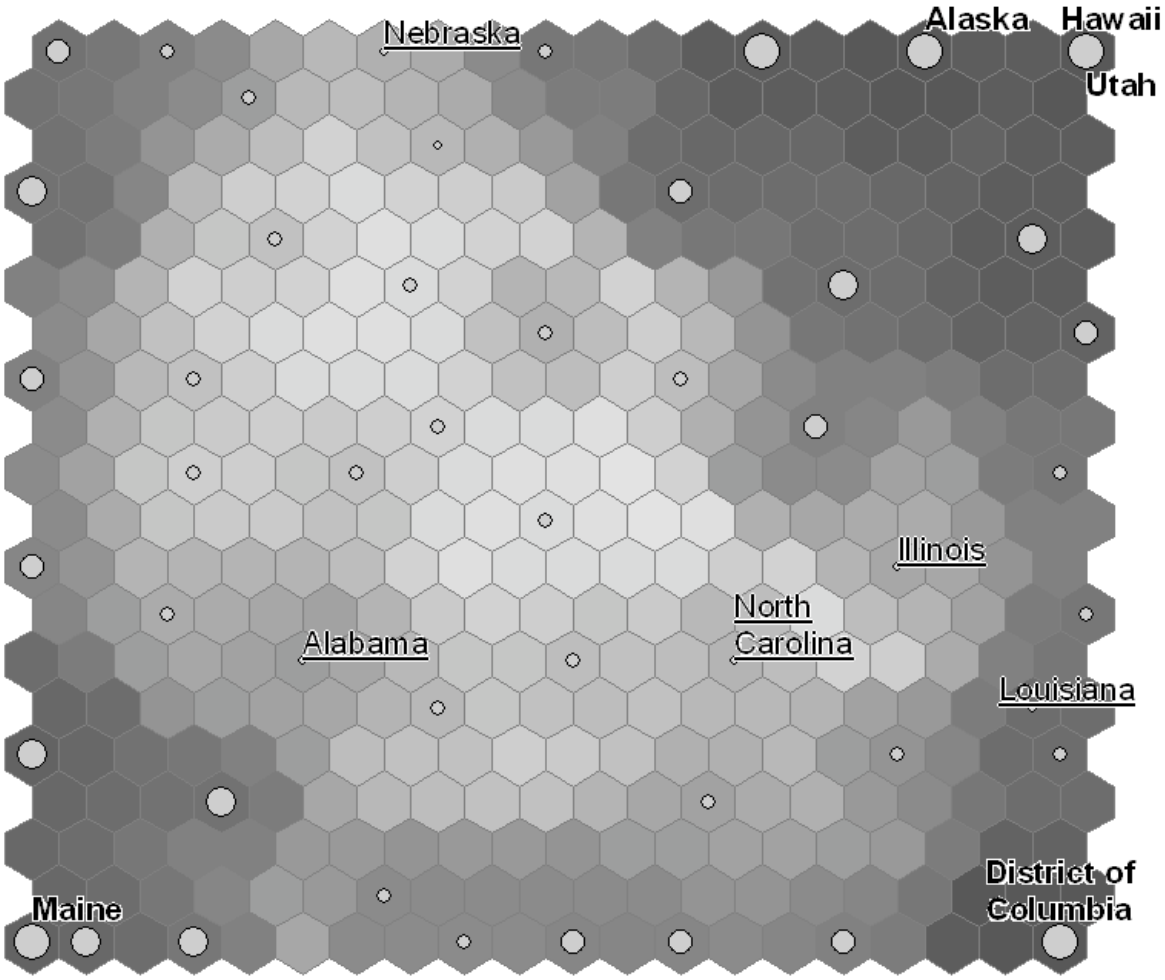

Figure 1: Fifty States and the District of Columbia mapped onto a SOM trained with thirty-two population census variables. Darker neurons have a relatively larger difference from the mean of the states, while lighter neurons are relatively closer. Smaller point symbols show states that are closer to the mean, while large symbols show outliers. The five states closest to the average are shown with underlined labels and the five states furthest from the mean are shown with bold labels.

and suggests that a curved topology, such as that of a sphere, may better reflect directional data. A sphere also results in a more intuitive map, since we are accustomed to looking at geographic maps based on a sphere.

Ritter (1999) first introduced the spherical SOM, and several enhancements have since been suggested (Boudjemai et al., 2003; Sangole and Knopf, 2003; Nishio et al., 2006; Wu and Takatsuka, 2006). A good comparison of these enhancements can be found in Wu and Takatsuka (2006). All of these methods derive their spherical structure through the tessellation of a polyhedron as originally proposed by Ritter (1999). Wu and Takatsuka (2006) point out the importance of a uniform distribution on the sphere, and that it is preferable for all neurons to have an equal number of neighbors and to be equally spaced. They find generally that the tessellation method best satisfies these conditions, and specifically that the icosahedron is the best starting point (Wu and Takatsuka, 2005). Tessellation of the icosahedron results in a network of neurons, each of which having exactly six neighbors, save the original twelve which each have five neighbors. This is very close to the ideal structure in which every neuron would have exactly six neighbors. This structure has very low variances in both neuron spacing and neighborhood size.

3.3 Network Size

The literature offers little theoretical guidance on choosing an appropriate network size for a given dataset (Cho et al., 1996). Vesanto (2005) suggests simply using a network size of $5\sqrt{n}$, where n is the number of observations. Given this lack of theoretical development, researchers should be cautious when using methods that limit the control of network size. Having a high level of control over network size allows support for such very different SOM applications as clustering versus low-dimensional spatial layout. Skupin and Agarwal (2007) demonstrate this when they use the same data to train two SOMs of different sizes. In the three-by-three (9) case the neurons act as containers clustering similar states, while in the twenty-by-twenty (400) case relationships are expressed with much finer granularity. As shown in Figure 2, methods for arranging an arbitrary number of points on a sphere provide a much higher degree of flexibility when choosing a network size.

The cost of relying on Ritter's tessellation method is decreased control over network size. Ritter's tessellation method results in a network size that grows at a rate of $N = 2 + 10 * 4^f$, where f is the frequency of tessellation. Wu and Takatsuka (2006) offer a slight improvement. Rather than recursively subdividing the faces, they redivide the original icosahedron with each step, resulting in $N = 2+10*f^2$. In practice, 2D Euclidean SOMs also offer limited control over network size because it is undesirable to have one dimension dramatically larger than the other. Nishio et al. (2006) try to address the issue of network size granularity by departing from the tessellation method and suggesting the use of a partitioned helix to uniformly distribute any number of neurons on a sphere. A similar method proposed by Rakhmanov et al. (1994) was dismissed by Wu and Takatsuka (2005) for failing to satisfy the uniformity conditions described above.

3.4 Uniformity

Wu and Takatsuka state that "[f]or SOM, it is desirable to have all neurons receive equal geometrical treatment" (Wu and Takatsuka, 2006, p. 900). To satisfy this constraint, two conditions must be met. First, each neuron should occupy the same amount of space on the given surface. Second, each neuron should be bordered by the same number of surrounding neurons, and we should maximize that number. The first condition may be important for visualization, but is irrelevant for training. During the training of the SOM only the topology of the neurons is considered.

Based solely on measures of neuron spacing, Wu and Takatsuka (2005) dismissed a method proposed

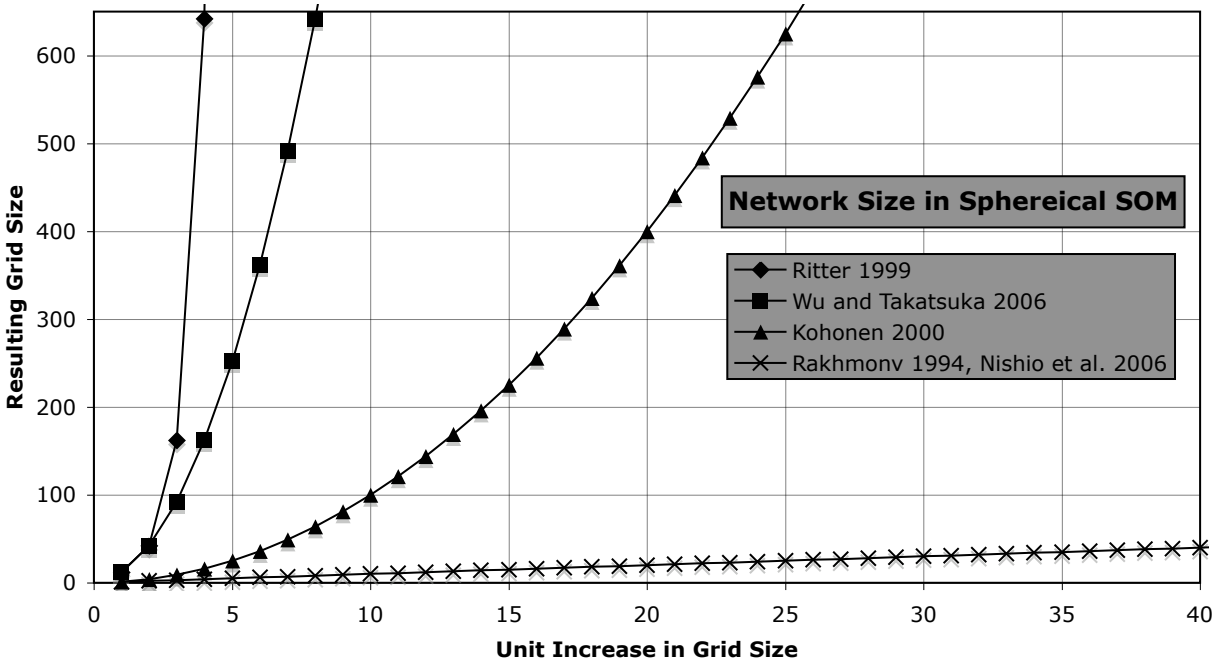

Figure 2: This figure demonstrates the achievable network size using various spherical topologies, in comparison with the traditional SOM. The Y-axis represents the achievable network size, the meaning of the X-axis is dependent on the topology. For the tessellation methods the X-axis represents the frequency of the tessellation. For the traditional Kohonen method the X-axis represents the size of both dimensions of the grid; for comparability the ratio between the dimensions was fixed at one ($X_{dim} = Y_{dim}$). For the Rakhmanov et al. (1994) and Nishio et al. (2006) methods the X-axis represents the exact network size.

by Rakhmanov et al. (1994) for distributing points on a sphere. Similarly Nishio et al. (2006) use these variance measures to support their helix algorithm for distributing points on a sphere. Table 1 shows that these metrics can be misleading and comparison across topologies may not be consistent. The traditional rectangular and hexagonal topologies have no variance in neuron spacing, and the generally preferred hexagonal structure displays greater variance in neighborhood size than the rectangular structure. The torus, by comparison, would have variance in neuron spacing, yet no variance in neighborhood size. The distance between two neurons is only considered during the formation of the neural network. At this stage the spacing is significant as it plays a part in determining neuron adjacency. However, using this measure to evaluate potential topologies for use in SOM may be misleading.

Table 1: Variances in Topologies

Topology	Grid Size	Neuron Spacing	Variance in Neighborhood Size
Rectangular	9x18	1	0.2716
Hexagonal	9x18	1	1.2138
Tessellation	162	0.25319 - 0.31287	0.0686
Rakhmanov	162	0.15779 - 0.30069	0.2908

Methods for distributing points on the sphere, which allow for fine-grained control over network size, produce slightly more irregular topologies. However, no substantive discussion of these irregularities or their effects on SOM training exists in the literature. Given that limited theoretical guidance is available for choosing network size, the desire for finer control over the network size should not be overlooked. Particularly for larger SOMs, the desired network size may not be achievable via tessellation of the icosahedron.

4 Data and Methodology

This section is composed of two parts. The first describes the diagnostics developed for evaluating network topologies. The second describes an empirical study that will implement these methods in order to evaluate the utility of topologies that allow for greater control over network size.

4.1 Diagnostics

Three diagnostics are developed to explore the effect of irregular topology on spherical SOM. The first diagnostic will be used to address the research question regarding the internal variance and neighborhood size. The second diagnostic will address the question concerning internal variance and topological irregularity. The third tool will help visualize the patterns between internal variance and topology.

4.1.1 Internal variance vs. first-order neighborhood size

This diagnostic will compare the internal variance of each neuron against its first-order neighborhood size. In traditional SOMs, outlying observations are pushed to the edge of the map where they encounter fewer competing signals. A prime example of this is the "Utah-Hawaii" case shown in Figure 1. Relying only on the SOM, one would be left to believe that the two states are similar. Upon closer inspection we see that the QError from Utah to the neuron is 1.509, the QError from Hawaii to the neuron in 1.505, but

the QError from Utah to Hawaii is 3.014. In this case only Utah and Hawaii were mapped to that neuron. In a case where multiple observations land on the same neuron, it is possible to measure average pairwise QErrors between those observations. This gives us a notion of internal variance for each neuron. It would be expected that in traditional SOMs neurons closer to the edge will have higher internal variances. This can be extended to spherical SOMs by comparing the degree of a neuron ($deg(m_i)$ or the number of adjacent neurons) to its internal variance. The degree of each neuron can easily be calculated by taking the column sums of the first-order adjacency matrix (A).

Once the internal variance ($var(m_i)$) and degree ($deg(m_i)$) of each neuron have been calculated, the neurons can be separated into a small number of groups based on the degree [1]. The variance and mean will be calculated for each of these groups. The expected result is that variances and means of the groups will decrease as the degree increases. This hypothesis will be tested using a ratio of variances test and a difference of means test. The result will also be visualized using a box plot.

4.1.2 Internal variance vs. topological regularity

This diagnostic will compare the internal variance of each neuron against a measure of regularity for its associated topology. As mentioned above the degree of each neuron can be calculated by taking the column sums of A. A completely regular network topology (i.e. the torus) will have no variance between these column sums. For irregular networks the variance between these column sums gives us a measure of irregularity. There are many alternative ways to classify the connectivity of a network; Florax and Rey (1995) outline four such summary measures, which will be evaluated for use in this diagnostic. For each topology we can compare the internal variances as described above against a measure that summarizes the given topology's regularity.

This diagnostic is evaluated in much the same way as the last diagnostic. The internal variances are this time grouped by their topology. We can then compare the variances of internal variances and the means of the internal variances across topologies. It is expected that the distribution of internal variances will be narrower for groups trained on more regular topologies. It is further hypothesized that the means of these internal variances will decrease when the network is more regular, or when there is less variance in the column sums of A. These assumptions will be tested using a *t-test* on the means and an *F-test* on the variances.

4.1.3 Visualize internal variance mapping

Visualizing the internal variance may yield insight into how irregular topology effects the SOM. Once the internal variance of each neuron has been calculated we can use the values to color or shade a map of the given topology. The degree to the neurons can be visualized using proportional symbols to help show patterns between internal variance and irregularity.

4.2 Empirical Analysis

The empirical analysis consists of three main tasks. The first is to create synthetic data suitable for the diagnostics described above. The second task is to train multiple SOMs, each of with a different topology type. The third task will be to apply the diagnostics and interpret the results.

[1] For most topologies the number of different degrees will be limited to three or four.

4.2.1 Synthetic Data

In order to test the methods described above, I will generate high-dimensional synthetic data with known properties. Knowing the properties of the training data allow us to systematically compare the diagnostics under several different topologies. To ensure that we can calculate an internal variance for each neuron, I will generate the data so as to increase the probability that each neuron will be occupied by more than one observation.

4.2.2 Training

The diagnostics must be given a trained SOM for each topology to be compared. To yield any meaningful results those SOMs must be trained with comparable parameters. Most parameters can simply be set to the same value for each SOM. However, special consideration must be given to network size. As shown in Figure 2, topologies differ in terms of achievable network size. This analysis will include the following topologies:

- Rectangular
- Hexagonal
- A topology based on Rakhmanov et al. (1994)
- The Helix topology proposed by Nishio et al. (2006)[2]
- The Geodesic topology proposed by Wu and Takatsuka (2006)[2]

4.2.3 Diagnostics

The first diagnostic will yield a set of results for each topology test. These results will be analyzed in order to address the first research question of this paper. The second diagnostic provides one set of results. These results will be analyzed to address the second research question. The final diagnostic will return a visualization for each topology tested. The usefulness of these visualizations is a research question in itself. The expectation is that the visualizations will show patterns of internal variance related to irregularities in the network topology.

5 Significance and Limitations

5.1 Significance

The commonly used tessellated icosahedron based topology offers the most regular topology. However, the main disadvantage of this topology type is that it offers a limited control over network size. Alternative methods for generating the spherical topology, which can create a network of any size, have been reviewed or suggested by Wu and Takatsuka (2005) and Nishio et al. (2006). These alternative methods produce network structures that are more irregular. This research will take a closer look at the impact that the irregularity has on the training process in an attempt to address the suitability of these topologies for use in SOM.

[2]Currently there is some uncertainty about the ability to include the Geodesic and Helix type topologies given the complexities involved with their implementation. However, an additional goal of this project to provide a framework on which new topologies can be easily implemented and tested by future researchers.

5.2 Limitations

This research will look at the relationship between regularity in neuron connectedness and the training of a SOM. The relationship between topology and SOM visualization is not addressed. The topology chosen for a SOM has a direct link with how that SOM is visualized. When representing the topology on the surface of a sphere issues arise with the uniformity in neuron spacing and sizing. Future work may be needed to address these issues in visualization.

References

Boudjemai, F., Enberg, P. B., and Postaire, J. G. (2003). Surface modeling by using self organizing maps of Kohonen. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 3, pages 2418–2423 vol.3.

Cho, S., Jang, M., and Reggia, J. A. (1996). Effects of varying parameters on properties of self-organizing feature maps. *Neural Processing Letters*, V4(1):53–59.

Florax, R. J. and Rey, S. (1995). The impacts of misspecified spatial interaction in linear regresion models. In *New directions in spatial econometrics*, Advances in Spatial Science. Springer.

Harris, J. M., Hirst, J. L., and Mossinghoff, M. J. (2000). *Combinatorics and graph theory*. Springer, New York.

Ito, M., Miyoshi, T., and Masuyama, H. (2000). The characteristics of the torus self organizing map. In *Proceedings of 6th International Conference ON Soft Computing (IIZUKA2000)*, volume A-7-2, pages pp.239–244, Iizuka, Fukuoka, Japan.

Kohonen, T. (2000). *Self-Organizing Maps*. Springer, 3rd edition.

Li, X., Gasteiger, J., and Zupan, J. (1993). On the topology distortion in self-organizing feature maps. *Biological Cybernetics*, V70(2):189–198.

Nishio, H., Altaf-Ul-Amin, M., Kurokawa, K., and Kanaya, S. (2006). Spherical SOM and arrangement of neurons using helix on sphere. *IPSJ Digital Courier*, 2:133–137.

Rakhmanov, E. A., Saff, E. B., and Zhou, Y. M. (1994). Minimal discrete energy on the sphere. *Mathematical Research Letters*, 1:647–662.

Ritter, H. (1999). Self-organizing maps on non-euclidean spaces. In Oja, E. & Kaski, S., editor, *Kohonen Maps*, pages 97–110. Elsevier, Amsterdam.

Sangole, A. and Knopf, G. K. (2003). Visualization of randomly ordered numeric data sets using spherical self-organizing feature maps. *Computers & Graphics*, 27(6):963–976.

Skupin, A. and Agarwal, P. (2007). (In preparation) Introduction: What is a self-organizing map? In Agarwal, P. and Skupin, A., editors, *Self-Organizing Maps: Applications in Geographic Information Science*. In preparation for publication by Wiley.

Vesanto, J. (2005). Som toolbox: implementation of the algorithm. `http://www.cis.hut.fi/projects/somtoolbox/documentation/somalg.shtml`.

Wu, Y. and Takatsuka, M. (2005). Geodesic self-organizing map. In Erbacher, R. F., Roberts, J. C., Grohn, M. T., and Borner, K., editors, *Proc. SPIE Vol. 5669*, volume 5669 of *Visualization and Data Analysis 2005*, pages 21–30. SPIE.

Wu, Y. and Takatsuka, M. (2006). Spherical self-organizing map using efficient indexed geodesic data structure. *Neural Networks*, 19(6-7):900–910.