Data nonlinearity in exploratory multivariate analysis of language corpora

Hermann Moisl
School of English Literature, Language, and Linguistics
University of Newcastle
Newcastle upon Tyne NE1 7RU
United Kingdom

hermann.moisl@ncl.ac.uk

Abstract

Data nonlinearity has historically not been and currently is not an issue in work on exploratory multivariate analysis of language corpora. However, the presence of nonlinearity in data has a fundamental bearing on the conduct of exploratory analysis. The first part of the discussion explains why this is so in principle, and the second exemplifies the explanation via exploratory analysis of the *Newcastle Electronic Corpus of Tyneside English* (NECTE), an historical speech corpus. The conclusion is that data should be screened for nonlinearity prior to analysis and, if a substantial degree of it is found, a nonlinear analytical method should be used.

1. Introduction

Exploratory multivariate analysis methods are used across a wide range of research disciplines to identify interesting structure in multidimensional data whose characteristics are not well known, and, if structure is found, to generate hypotheses about the domain which the data describes (Andrienko and Andrienko, 2005). Corpus-based linguistics has long been among these disciplines, and, as computational power has increased and ever-larger natural language corpora have become available, the application of exploratory analysis in empirical linguistic research has grown. When one surveys the relevant linguistics literature, it becomes clear that data nonlinearity has historically not been and is not currently an issue. An exhaustive review cannot be undertaken here, but a snapshot of recent literature is symptomatic: neither the relevant papers in the *Literary and Linguistic Computing* journal's special issue on 'Progress in Dialectometry' (2006) nor Manning and Schütze's discussion of clustering in their subject-standard *Foundations of Statistical Natural Language Processing* (2000) refer to it, except perhaps in passing. However, the presence of nonlinearity in data has a fundamental bearing on the conduct of exploratory analysis. The first part of the discussion explains why this is so in principle, and the second exemplifies the explanation via exploratory analysis of the *Newcastle Electronic Corpus of Tyneside English* (NECTE), an historical speech corpus. The conclusion is that data should be screened for nonlinearity prior to analysis and, if a substantial degree of it is found, a nonlinear analytical method should be used.

2. Nonlinearity and exploratory analysis

In physical systems, nonlinearity is the breakdown of proportionality between cause and effect, and it manifests itself in a variety of complex and often unexpected --including chaotic-- behaviours. Since nonlinearity pervades the physical world (see for example Bertuglia, 2005), data that describes it is likely to contain nonlinearity as well. If the data is in vector space representation, such nonlinearity manifests itself as curvature in the data manifold, which can range from simple curves and surfaces to highly convoluted fractals.

Many of the commonly used exploratory multivariate methods, henceforth called 'linear methods', are insensitive to nonlinearity, and as such can generate results that misrepresent the structure of a nonlinear data manifold. This insensitivity stems from the way in which the linear methods measure distance between pairs of vectors in the manifold --as the shortest straight-line

Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 93–100,
Prague, June 2007.

distance between them. This is not, however, the only possible measure. This distance between two cities can be measured linearly as in figure 1a or nonlinearly along the curve of the earth's surface, as in figure 1b:

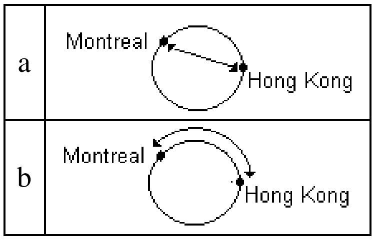


Figure 1: Linear and nonlinear distance measure

Linear distance in this case seriously misrepresents the true distance. The same applies to nonlinear data manifolds. Figure 2 shows an extreme example frequently used in discussions of nonlinear dimensionality reduction (i.e. Tenenbaum et al., 2000), in which linear distance and distance along the surface of the manifold differ markedly.

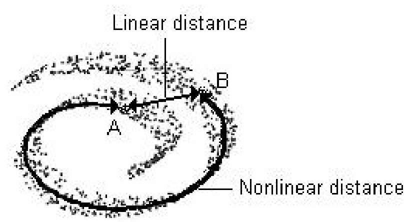


Figure 2: Linear and nonlinear distance in a nonlinear manifold

Linear exploratory methods base their representation of data structure on linear distance between vectors in the data space. If the manifold diverges significantly from linearity, linear distance measures can give distorted results.

The classic response to the discovery of nonlinearity in data is to remove it using well established methods like log-transformation (i.e. Clarke and Cooke, 1998:571-4), and then to analyze the linearized data using a linear method. This risks throwing the proverbial baby out with the bathwater. Nonlinearity is not always just a nuisance to be eliminated, but may reflect a fundamental aspect of the thing being studied; in fact, the study of nonlinearity in natural systems is now well established across a range of disciplines (Scott, 2004). If nonlinearity is found in natural language corpus data, the default should be to retain it on the grounds that it might reflect a scientifically interesting aspect of corpus structure. If it is retained, however, linear analytical methods become inapplicable in principle, and nonlinear ones which measure distance along the curvature of the manifold must be used.

3. Exploratory analysis of the NECTE data

3.1 The NECTE data

The *Newcastle Electronic Corpus of Tyneside English* (NECTE) is a corpus of dialect speech from Tyneside in North-East England (Allen et al., 2005). It includes phonetic transcriptions of 63 interviews together with social data about the speakers, and as such offers an opportunity to study the sociophonetics of Tyneside speech of the late 1960s. Moisl et al. (2006) and Moisl and Maguire (2007) have begun that study using exploratory analysis of the transcriptions with the aim of generating hypotheses about phonetic variation among speakers in the Tyneside dialect area. These studies were based on comparison of profiles associated with each of the informants. A profile for any speaker S is the number of times S uses each of the phonetic segments in the NECTE transcription scheme in his or her interview. More specifically, the profile P associated with S is a vector having as many elements as there are segments such that each vector element P_j represents the *j*'th segment, where *j* is in the range 1..number of segments in the NECTE phonetic transcription scheme, and the value stored at P_j is an integer representing the number of times S uses the *j*'th segment. There are 156 segments, and so a speaker profile is a length-156 vector. There are 63 TLS speakers, and their profiles are represented in a matrix M having 63 rows, one for each profile.

3.2 Identifying nonlinearity

Where the data dimensionality is 3 or less, nonlinearity can be identified by creating a scatterplot of the manifold and looking for curvature. Visual interpretation is subjective, however. It can be unreliable when the shape of the manifold is not as clear cut as, say, in figure 2, and needs to be supplemented with some quantitative measure of nonlinearity; for high-dimensional data direct graphical representation is impossible (Andrienko and Andrienko, 2005, ch. 4), and quantitative measurement is the only alternative. The most straightforward measures are based the residuals in linear and nonlinear regression: the sum of squares of residuals, or SS_R, gives the total divergence of the data variables from the line of best fit, and the standard error their average dispersion around the line in a way analogous to univariate standard deviation.

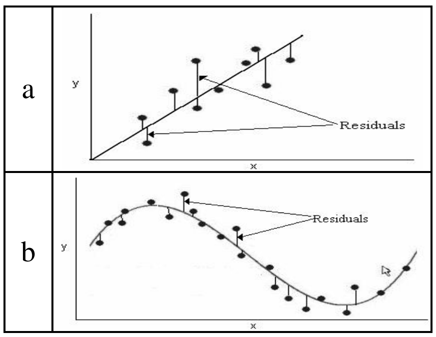


Figure 3: Lines of best fit in linear and nonlinear regression

For a given pair of variables, if the SS_R and standard error from a nonlinear regression are less than those from a linear one, then a curve fits the data better than a straight line and the relationship of the two variables is nonlinear.

In applications where the dimensionality of the data can be in the hundreds or even thousands, pairwise regression-based testing of nonlinearity can quickly become onerous since, for any given dimensionality n,

$$p_n = \frac{n(n-1)}{2}$$

For $n = 100$, there would be 4950 different variable pairs to consider. The situation can be salvaged in cases where some variables are more important than others relative to the research question by examining only a tractable subset of important variables. Several criteria for variable importance are available, such as variance, term frequency / inverse document frequency (Robertson, 2004) and Poisson distribution (Church and Gale, 1995a, 1995b); the use of variance for this purpose is exemplified below.

With a dimensionality of 156, 12090 variable pairs would have to be tested for nonlinearity, which is not impossible but certainly onerous. The number of pairs to be considered was therefore reduced to a manageable level using the relative variances of the 156 variables as a selection criterion. The justification for using variance for this purpose is as follows. Classification of objects in any domain of study depends on there being variation in their characteristics. When the objects to be classified are described by variables, then a variable is only useful for the purpose if there is significant variation in the values that it takes; those with little or no variation can be disregarded. The variances of the column vectors of M were calculated, sorted in descending order of magnitude, and plotted in figure 4.

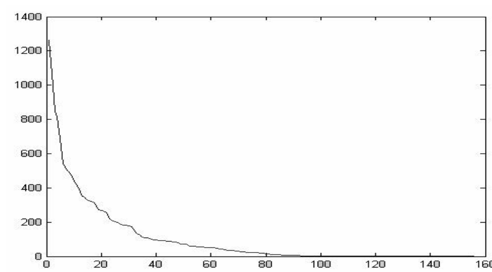


Figure 4: Variances of column vectors of N

The highest-variance dozen variables were selected and linear, quadratic, and cubic regression were applied to all 66 distinct pairings of them, in each case calculating SS_R and standard error. Three examples are given: figure 5a is representative of the linearly-related pairs, figure 5b of moderately nonlinear pairs, and figure 5c of strongly nonlinear ones. The frequencies of these are 12 linear, 25 moderately nonlinear, and 29 strongly nonlinear.

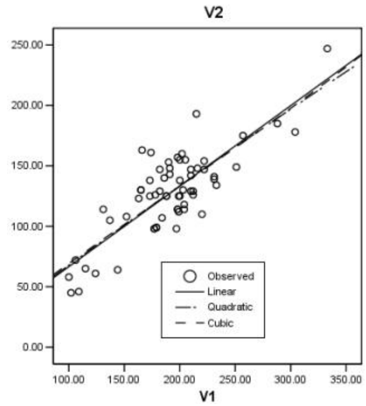
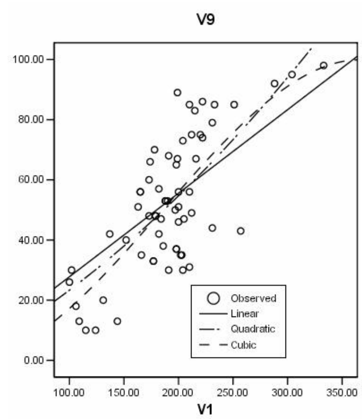
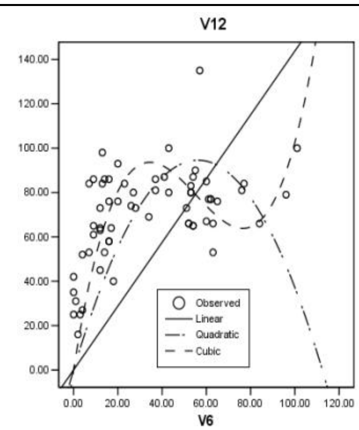
<table>
<tr><th></th><th>Regression plots</th><th>Quantifications</th></tr>
<tr><td>a</td><td>
</td><td>
<table>
<tr><th>Linear</th><th></th></tr>
<tr><td>SS_R</td><td>26682.00</td></tr>
<tr><td>Standard Error</td><td>20.74</td></tr>
<tr><td>Quadratic</td><td></td></tr>
<tr><td>SS_R</td><td>26622.40</td></tr>
<tr><td>Standard Error</td><td>20.89</td></tr>
<tr><td>Cubic</td><td></td></tr>
<tr><td>SS_R</td><td>26598.20</td></tr>
<tr><td>Standard error</td><td>21.05</td></tr>
</table>
</td></tr>
<tr><td>b</td><td>
</td><td>
<table>
<tr><th>Linear</th><th></th></tr>
<tr><td>SS_R</td><td>53703.43</td></tr>
<tr><td>Standard Error</td><td>29.67</td></tr>
<tr><td>Quadratic</td><td></td></tr>
<tr><td>SS_R</td><td>53496.58</td></tr>
<tr><td>Standard Error</td><td>29.86</td></tr>
<tr><td>Cubic</td><td></td></tr>
<tr><td>SS_R</td><td>38880.40</td></tr>
<tr><td>Standard error</td><td>25.67</td></tr>
</table>
</td></tr>
<tr><td>c</td><td>
</td><td>
<table>
<tr><th>Linear</th><th></th></tr>
<tr><td>SS_R</td><td>95071.21</td></tr>
<tr><td>Standard error</td><td>39.16</td></tr>
<tr><td>Quadratic</td><td></td></tr>
<tr><td>SS_R</td><td>49281.20</td></tr>
<tr><td>Standard error</td><td>28.42</td></tr>
<tr><td>Cubic</td><td></td></tr>
<tr><td>SS_R</td><td>22206.88</td></tr>
<tr><td>Standard error</td><td>19.24</td></tr>
</table>
</td></tr>
</table>

Figure 5: Sample regressions of variable pairs from data matrix M

The essentially linear relationship of v1 and v2 is clear both visually and in the uniformity of SS_R and standard error measures, where the nonlinear regressions yield no meaningful improvement over the linear. For v1 and v9 cubic regression shows some improvement over linear and quadratic both visually and quantitatively. For v6 and v12 the quadratic regression line is visually a much better fit to the data than the linear one, and the cubic one is even better; correspondingly, the quadratic quantifications show a substantial improvement over the linear ones, and the cubic ones even more so. The relationships between the highest-variance variables in M can, therefore, be said to range from linear to strongly nonlinear.

3.3 Linear and nonlinear analysis of the NECTE data

Moisl et al. (2006) analyzed the NECTE data with what is probably the most widely used of the linear exploratory methods: hierarchical cluster analysis (Everitt et al., 2001). This is actually a class of methods each of which defines clusters differently, but all of which represent cluster structure as nested constituency trees. Infamously --and unsurprisingly, given that each uses a different definition of what constitutes a cluster-- the variant methods can and often do assign different tree structures to the same data, and it is not usually clear which is to be preferred (Everitt et al., 2001, ch. 4). In the NECTE case, however, a range of variants (single link, complete link, average link, Ward's Method) converged on a stable structure of four main clusters exemplified by the Ward tree shown in figure 6.

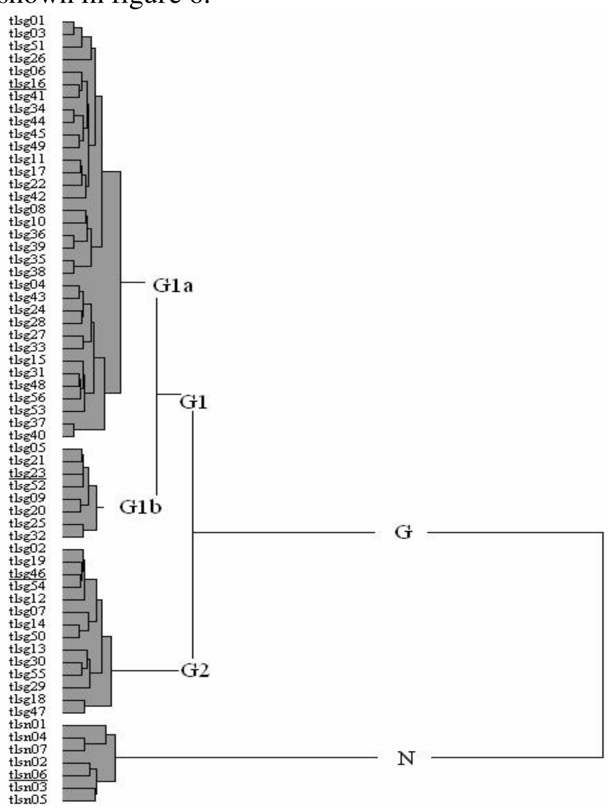


Figure 6: Ward's Method cluster tree for data matrix M

When interpreted in terms of the social data that NECTE provides for the speakers, a clear correlation between phonetic usage and social factors emerged. The main distinction is between middle class, well educated speakers from Newcastle on the north side of the river Tyne, labelled N, and working class, less well educated speakers from Gateshead on the south side of the Tyne, labelled G. The Gateshead speakers are categorized into G2 (exclusively male), and G1 (mainly through not exclusively female); G1 is subcategorized into G1a (working class males and females) and G1b (males and females with relatively higher socioeconomic status). Moisl and Maguire (2007) subsequently used the centroids of these clusters to identify the phonetic features most characteristic of each. Three sets of vowels were found to be of particular importance. Although all of these had been commented on before, their relative (and cumulative) sociolinguistic importance had hitherto escaped attention. They are:

- various types of [ə].
- [ɔː] and [ɑː], which correspond to RP [əʊ], and are found in words of the GOAT lexical set as defined by Wells (1982:146-7).
- [aɪ], [ɑː], and [eɪ], which correspond to RP [aɪ], and are found in words belonging to the PRICE lexical set as defined by Wells (1982:149-50).

For nonlinear analysis the self-organizing map, or SOM, was selected from among the various available nonlinear exploratory methods because it has been successfully used in a very wide range of applications (Kaski et al., 1998; Oja et al., 2001). The standard SOM (Kohonen, 2001) projects the topology of a data manifold in a space of arbitrary dimensionality *n* onto a two-dimensional lattice, where the structure of the manifold can be visually inspected. It does this by partitioning the vectors on the manifold surface into a Voronoi tesselation (Aurenhammer and Klein, 2000), thereby assigning all the data vectors within a defined topological neighborhood to the same cell of the tesselation, as shown in figure 7.

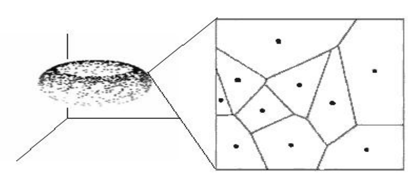


Figure 7: Voronoi tesselation of a manifold surface

For example, the doughnut shape on the left of figure 7 is a manifold in 3-dimensional space, and the square on the right represents the way in which a SOM partitions its surface: each dot represents a quantized vector and the lines enclosing a dot represent the boundaries of the area of the tesselation cell containing the *k* vectors within the specified topological neighborhood. All the vectors in a given cell are mapped to the same lattice unit, and the vectors in adjoining cells are mapped to adjacent lattice units. The result of this topology preservation is that all vectors close to one another in the input space in the sense that they are in the same or adjoining topological neighbourhoods will be close on the SOM output lattice (for further discussion see Ritter et al., (1992), ch. 4). The topology preservation is, moreover, nonlinear because the tesselation is based not on a global distance measure between vectors on the manifold but on local neighborhood distance, and as such the tesselation follows the manifold surface: if the surface is nonlinear, so is the topology-preserving representation of it.

The NECTE data was analyzed using a range of SOM parameters for output lattice size and shape and various initializations such as starting neighborhood, learning rate, and rate of neighborhood decrease. The results converged on a stable analysis of which the following map is representative.

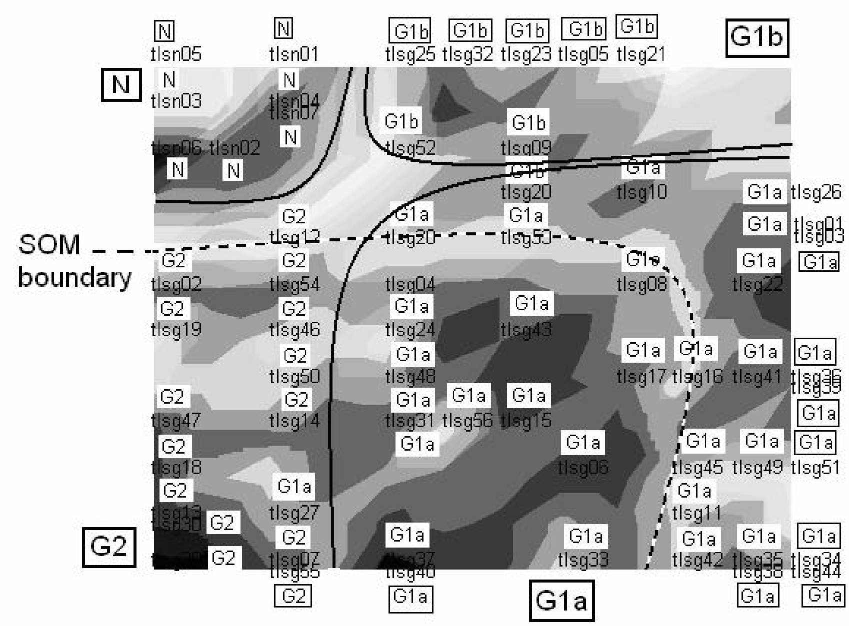


Figure 8: SOM analysis of the NECTE data

The speaker labels were positioned automatically on the lattice by the SOM's input-to-lattice mapping function, and the shading was generated using the U-matrix method (Ultsch, 1993). This shading must be understood in order to interpret the above SOM correctly, so a brief explanation is given here. It has already been noted that the SOM preserves the topology of the *n*-dimensional input manifold in the sense that vectors which are close in the input space are also close in the two-dimensional output space. The converse is not true, however: just because vectors are close in the output space does not necessarily mean that they are close on the input manifold. This apparently-paradoxical situation arises because the SOM mapping function does not use a global distance measure but only local neighborhood distance, and it consequently cannot and does not represent proportionality of distance between vector pairs in the input space. Instead, is squeezes its representation of the input topology onto the lattice in such a way that closely adjacent lattice cells may represent vectors which are far apart on the input manifold. Because, therefore, spatial distance is a delphic guide to interpretation of the SOM, some way must be found of demarcating the shape of the manifold representation given by the lattice. The U-matrix is a way of doing this. How it works can only be explained in terms of the details of SOM architecture, which cannot be given here on account of space constraints. It is, however, important to understand that lighter regions of the map represent manifold boundaries and darker ones the manifold surface; metaphorically, the darker areas are islands representing the shape of the manifold, and the lighter areas the sea separating them. The remaining annotations in figure 8, finally, were added by hand to facilitate discussion in the next subsection, and are explained there.

3.4 Discussion

Associated with each speaker on the SOM is a label which shows that speaker's place in the hierarchical cluster tree --tlsg08 on the SOM is in cluster G1a in the tree, for example. In addition, solid-line curves have been added to the SOM which show the approximate areas of the map that correspond to the main hierarchical clusters and, for each region, the relevant hierarchical cluster label has been shown surrounded by a square --the upper left corner of the SOM, for example, is bounded by a solid curve and labelled N to show that the speaker vectors found there correspond to those in the N hierarchical cluster. Using these annotations, it might appear that the hierarchical and SOM analyses are similar: the hierarchical analysis shows four main clusters, and the SOM has four disjoint regions corresponding to those clusters. This perception of correspondence is, however, based on spatial placement of the speaker vectors on the SOM, and, as we have seen, relative spatial distance on a SOM can be misleading. If one looks instead at the U-matrix shading that demarcates the manifold boundaries, the Newcastle group is as clearly distinguished from the Gateshead speakers by the SOM as by the hierarchical analysis, but the Gateshead speakers are grouped in a way that differs subtly from the hierarchical analysis. The hierarchical analysis says that there are three distinct Gateshead groups: G1a consists of working class men and women, G1b of lower middle class men and women, and G2 of working class men. The SOM, on the other hand, says that the Gateshead speakers fall into only two main groups the boundary between which is shown in figure 8 as a dotted-line curve. The one above and to the right of the dotted line (and excluding the Newcastle group) consists of lower middle class men and women and working class women. The other, below and to the left of the dotted line, comprises working class men together with two women (tlsg37 and tlsg40) who are classified with men both here and in the hierarchical analysis.

The linear and nonlinear methods, therefore, offer results that differ substantively. From a methodological point of view, the SOM result must be preferred because the data contains nonlinearity, and a nonlinear method can be expected to give a more accurate analysis of nonlinear data than a linear one. A sociolinguist might find the SOM analysis preferable on grounds of simplicity: there is no obvious distinction in the social data between the working class men that the hierarchical analysis assigns to separate clusters. The present paper is, however, a methodological one, and no further comment is ventured on this.

5. Conclusion

The discussion began with the observation that existing work on exploratory analysis of linguistic corpora does not take the possibility of data nonlinearity into account, and claimed that the presence of nonlinearity in data has a fundamental bearing on the conduct of exploratory analysis. The first part of the discussion explained why this is so in principle, and the second exemplified the explanation via exploratory analysis of the *Newcastle Electronic Corpus of Tyneside English* using both linear and nonlinear methods. That the two types of method gave substantively different results supports the case in principle that data should be screened for nonlinearity prior to exploratory analysis and that, if substantial degree of it is found, a nonlinear analytical method should be used.

References

Will Allen, Joan Beal, Karen Corrigan, Warren Maguire, and Hermann Moisl. 2007. A Linguistic Time Capsule: the Newcastle Electronic Corpus of Tyneside English. In: Joan Beal, Karen Corrigan, and Hermann Moisl (eds.). 2007. *Using Unconventional Digital Language Corpora: Diachronic Corpora*. Palgrave Macmillan, Basingstoke, 16-48.

Natalie Andrienko and Gennady Andrienko. 2005. *Exploratory Analysis of Spatial and Temporal Data: a Systematic Approach*. Springer-Verlag, Berlin.

Franz Aurenhammer and R. Klein. 2000. Voronoi Diagrams. In: J.-R. Sack and J. Urrutia (eds.) *Handbook of Computational Geometry*. North-Holland, Amsterdam, 201-290.

Cristoforo Bertuglia and Franco Vaio. 2005. *Nonlinearity, Chaos, and Complexity: The Dynamics of Natural and Social Systems*. Oxford University Press, Oxford.

Kenneth Church and William Gale. 1995a. Poisson mixtures. *Natural Language Engineering*, 1: 163-190.

Kenneth Church and William Gale. 1995b. Inverse Document Frequency (IDF): A Measure of Deviation from Poisson. *Proceedings of the Third Workshop on Very Large Corpora*. Association for Computational Linguistics. Reed Elsevier, 121-130.

G. Clarke and D. Cooke. 1998. *A Basic Course in Statistics*. 4th ed. Arnold, London.

Brian Everitt, Sabine Landau, and Morven Leese. 2001. *Cluster Analysis.* 4th ed. Arnold, London.

Samuel Kaski, J. Kangas, and Teuvo Kohonen. 1998. Bibliography of Self-Organizing Map (SOM) Papers: 1981-1997. *Neural Computing Surveys*, 1:102-350.

Teuvo Kohonen. 2001. *Self-Organizing Maps.* 3rd ed. Springer-Verlag, Berlin.

Christopher Manning and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass.

Hermann Moisl, Warren Maguire, and Will Allen. 2006. Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English. In: Frans Hinskens (ed.). *Language Variation-European Perspectives*. John Benjamins Publishing, Amsterdam, 127-141.

Hermann Moisl and Warren Maguire. 2007. Identifying the Main Determinants of Phonetic Variation in the Newcastle Electronic Corpus of Tyneside English. *Journal of Quantitative Linguistics* 2007, 14: to appear.

M. Oja, Samuel Kaski, and Teuvo Kohonen. 2001. Bibliography of self-organizing map (SOM) papers: 1998-2001. *Neural Computing Surveys*, 3:1-156.

Helge Ritter, T. Martinetz, and K. Schulten. 1992. *Neural computation and self -organizing maps.* Addison-Wesley, Wokingham, UK.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation,* 60:503-520.

Alwyn Scott. 2004. *Encyclopedia of Nonlinear Science*. Fitzroy Dearborn Publishers. .

Josh Tenenbaum, V. de Silva, and John Langford. 2000. A global framework for nonlinear dimensionality reduction. *Science*, 290:2319--2323.

Alfred Ultsch. 1993. Self-organizing neural networks for visualization and classification. In: O.Opitz, B. Lausen, and R. Klar, (eds.), *Information and classification :concepts, methods, and applications*. Springer-Verlag, Berlin, 307-313.

John Wells. 1982. *Accents of English*. Cambridge: Cambridge University Press, Cambridge, UK.