Multidimensional Scaling for Location Dependent Anisotropy

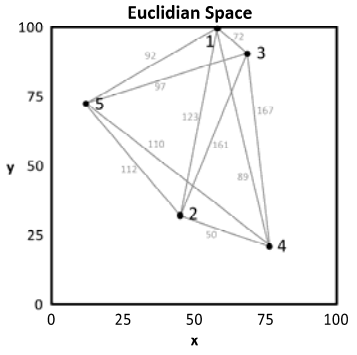
Jeff Boisvert and Clayton V. Deutsch

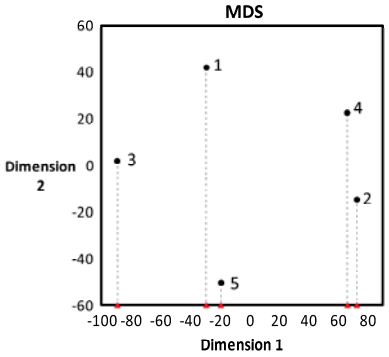
This paper discusses the application of multidimensional scaling (MDS) to kriging with covariance structures resulting from a non-Euclidian distance function. The run time of the Dijkstra algorithm is excessive and there is no guarantee of positive definite covariance matrices resulting from the non-Euclidian distance measure. This research investigates the use of MDS for non-Euclidian distance matrices. Nodes are mapped from the original xyz coordinates to a higher dimension Euclidian space where inter-point distances are approximately preserved and positive definiteness is guaranteed. Classical MDS is computationally expensive because it requires the Eigen solution to an NxN matrix, where N is the number of nodes. Mapping all nodes in a geostatistical model is a large problem. Algorithms recently presented have reduced the computational demand of MDS (i.e. the entire NxN eigen solution is not required). There are two MDS algorithms that will be applied and compared, ISOMAP and LLE. It was found that LLE produces unstable results while ISOMAP generates acceptable kriged maps. While MDS solves the problem of the indefinite kriging systems; the computational cost remains high.

Introduction

The methodology to generate geostatistical models with locally varying anisotropy, as presented in previous CCG papers, has two main flaws (1) lack of positive definite covariance matrices and (2) prohibitive run time. Positive definite matrices are required for the existence of a single, plausible solution to the kriging systems. The lack of a positive definite guarantee in the proposed methodology will be addressed with MDS and the run time concerns will be left for future work.

MDS is a statistical technique used to explain/map any type of similarities (i.e. covariances) or dissimilarities (i.e. distances) between data observations. Consider a road map example, the problem is to represent the travel time between cities on a map (Figure 1). This travel time depends on a number of factors: distance between cities; speed limit on roads; conditions of roads; traffic; and road construction. Assuming accesses to the travel times (Figure 1), such a measure of the difference between cities would be highly variable. Two geographically close cities may require a large travel time, whereas two distant cities separated by a freeway may have a short travel time. The problem is how to represent the dissimilarity (time) between cities? One option would be to rearrange the cities in two or more dimensions until the Euclidian distance between them is equivalent to the travel time. MDS can be applied to accomplish this rearrangement (Figure 1). The coordinates of the cities are mapped to a new coordinate system such that dissimilarities are approximately honored. Moreover, MDS allows the user to specify the number of dimensions in the new coordinate system. The inter-point differences can be honored in a high dimensional system. Using fewer dimensions results in an inferior reproduction of the inter-point differences.





Cities	Euclidian (m)	Time (hr)	(MDS)/(% difference to time) 1D MDS (hr)	2D MDS (hr)	3D MDS (hr)
1-2	69	123	102 / 17%	116 / 6%	123 / 0%
1-3	14	72	60 / 51%	72 / 41%	79 / 36%
1-4	81	89	95 / 22%	97 / 21%	97 / 21%
1-5	53	92	10 / 92%	92 / 25%	93 / 24%
2-3	63	161	162 / 32%	163 / 33%	163 / 32%
2-4	33	50	6 / 95%	38 / 69%	59 / 52%
2-5	52	112	92 / 25%	98 / 20%	112 / 9%
3-4	70	167	155 / 26%	157 / 28%	161 / 31%
3-5	59	97	70 / 43%	87 / 29%	99 / 20%
4-5	82	110	86 / 30%	112 / 9%	113 / 8%
Average % Difference			43%	28%	23%

Figure 1: Left – Locations of 5 cities in Euclidian space with travel times indicated (hours). Middle – Cities in the first two MDS dimensions. Note that the cities can also be represented in 1D as a projection onto the first MDS dimension (triangles on the x-axis). Right – Table of distances/times between cities. The MDS representation applies to the *time* between cities rather than the distance.

Using time as the dissimilarity measure between cities can result in an implausible difference matrix. Using an isotropic Gaussian variogram model with range=300hr the covariance matrix for Figure 1 (left) can be generated; this covariance matrix must have positive eigenvalues. The eigenvalues in this example are (3.72, 0.66, 0.54, 0.17, -0.10). This illustrates the power of MDS as the coordinates can be mapped to an Euclidian space (Figure 1 middle) where positive definiteness is guaranteed.

In the context of kriging with a non-Euclidian distance measure a similar situation occurs, i.e. the resulting covariance matrix may not be positive definite. Using MDS the grid locations or the city locations (as in Figure 1) can be mapped to an Euclidian coordinate system where the distance between points is approximately honored.

Classical MDS requires the eigenvalues/vectors of the difference matrix of all points that must transformed. In the above example this is a 5x5 matrix of the travel times between 5 cities. In the locally varying anisotropy context, this would be the distance matrix between ALL nodes in the model grid. This poses two problems (1) solving the eigen problem for a large *dense* system is not feasible, a million grid cell model would require the eigen solution of a 1,000,000 x 1,000,000 dense matrix and (2) generating the difference matrix between ALL nodes in a grid using the Dijkstra algorithm would be computationally infeasible. There are solutions to these problems which involve using MDS of a subset of the problem. Two main types of algorithms exist, local and global MDS. Local MDS requires a sparse difference matrix between nodes in the grid and local neighboring nodes only. This reduces the *dense* eigen problem to a *sparse* eigen problem, which is solvable even for large 1,000,000 x 1,000,000 matrices. Global MDS algorithms reduce the number of points that are used in the distance matrix such that the eigen solution can be feasibly solved.

Both global and local MDS algorithms are attractive in the context of locally varying anisotropy. Global algorithms will maintain the complex non-linear characteristics of large scale geological features while local algorithms would maintain the small scale details found within the larger scale structures. This paper will explore one global and one local MDS algorithm. Locally Linear Embedding (LLE) uses the distance between points found within a local search neighborhood and ignores the distances between points outside the search, generating a sparse distance matrix. ISOMAP is a global MDS algorithm that uses a series of landmark points in calculating the distance matrix, reducing the size of the distance matrix.

This paper is organized as follows. The next section contains the background for classical MDS, LLE and ISOMAP. These MDS algorithms are then applied to a small synthetic example, the same example used in the previous paper. This is followed by a discussion of the techniques and potential future work.

Background: Classical MDS

Consider N data observations with $M=N(N-1)/2$ (dis)similarities between points. These (dis)similarities are often the covariance between points or the inter-point distances. For this discussion we will assume the inter-point distances provide a dissimilarity measure to be used in MDS; therefore, the inputs to the MDS algorithm are M distances in a symmetric matrix (**D**). MDS maps the input data, **D**, to an Euclidian space where the inter-point distances approximately match the input distances. There will be some mismatch in distance (Figure 1) because of the impossible geometric configuration of the input distance matrix. MDS minimizes the mismatch between the original differences (d_{ik}^q) and the mapped differences ($\hat{d}_{ik}^q$) in q dimensions between all points, defined by a stress measure:

$$Stress(q) = \left\{ \frac{\sum_{k=1}^{N} \sum_{i=1}^{k-1} (d_{ik}^q - \hat{d}_{ik}^q)^2}{\sum_{k=1}^{N} \sum_{i=1}^{k-1} (d_{ik}^q)^2} \right\}^{1/2} \quad (1)$$

The stress value also provides a measure of how well the MDS mapping honors the input distances (this will be used to determine how many dimensions q to use in MDS). Applying MDS to a distance matrix involves (Mardia et al., 1976 or Loland and Host 2003):

a) Construct matrix $\mathbf{A} = (-0.5 d_{ik}^2)$. Note that in **D**, $d_{ii} = 0$ and $d_{ik} \geq 0$

b) Construct matrix **B**. Elements of **B** are: $b_{ik} = a_{ik} - \bar{a}_{i\bullet} - \bar{a}_{\bullet k} + \bar{a}_{\bullet\bullet}$, $\bar{a}_{i\bullet} = \sum_{k=1}^{N} \frac{a_{ik}}{N}$,

$\bar{a}_{\bullet k} = \sum_{i=1}^{N} \frac{a_{ik}}{N}$, $\bar{a}_{\bullet\bullet} = \sum_{k=1}^{N} \sum_{i=1}^{N} \frac{a_{ik}}{N^2}$.

c) Find q normalized eigenvectors of **B** that correspond to positive eigenvalues:

$$\mathbf{V} = (v_1, ..., v_q). \tag{2}$$

d) The coordinates of the input data are the q rows of **V**.

Note that the value of q is a *choice* and is often selected to be 2 or 3 so that the results can be visualized; however, the size of q can be as large as the number of positive eigenvalues of **B**. Calculating the stress resulting from different values of q can help in the selection of the number of dimensions to retain. Using more dimensions will always improve inter-point distance reproduction (i.e. lower stress value).

Background: LLE and ISOMAP

LLE (Saul and Roweis, 2003) is a local MDS technique. As discussed previously, LLE attempts to reduce the computational demand of MDS by reducing the number of entries in **D**. This reduces the above step *c* to finding the eigen solution of a sparse matrix, which is possible for very large matrices (Watkins, 2007). LLE maps the input data to a set of coordinates where the neighborhood points in the original space are still nearby in the mapped space. 'Nearby' is defined by a search radius and all pairs separated by a distance greater than the search are ignored and thus, not explicitly reproduced in the mapped coordinate system. Alternatively, a fixed number of nearest neighbors could be used to determine the local neighborhood. The 'locally linear ' in Locally Linear Embedding is a result of this search. LLE assumes that each point and its neighbors lie on a locally linear patch of the manifold defined by the inter-point distances of the points in the neighborhood. Points are embedded in the new coordinate space using a set of reconstruction weights which optimize the squared error between the input coordinates of the neighborhood points ($\vec{X}_i$) and the reconstructed coordinates:

$$Error(W) = \sum_i \left(\vec{X}_i - \sum W_{ij} \vec{X}_i \right)^2 \tag{3}$$

The LLE procedure is summarized in Figure 2. Saul and Roweis (2003) give a detailed description of the LLE algorithm.

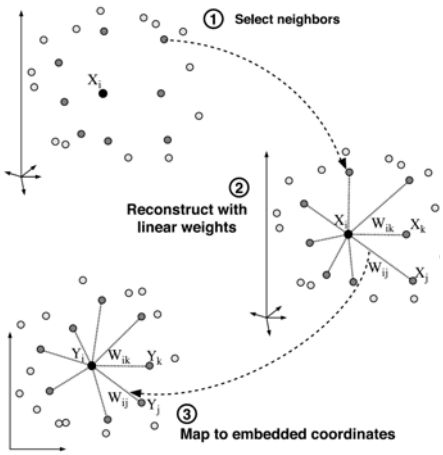


Figure 2: (as shown in Roweis and Saul 2000) Steps of LLE: (1) Assign neighbors to each data point $\vec{X}_i$ (for example by using the K nearest neighbors). (2) Compute the weights W_{ij} that best linearly reconstruct $\vec{X}_i$ from its neighbors, solving the least-squares problem in Equation 3. (3) Compute the low-dimensional embedding vectors $\vec{Y}_i$ best reconstructed by W_{ij}, minimizing Equation 3.

ISOMAP is a global MDS technique, the major references for ISOMAP are Silva and Tenenbaum (2003), Balasubramanian, Schwartz, Tenenbaum, Silva, and Langford (2002) and Tenenbaum, Silva and Langford (2000). ISOMAP uses a set of L landmark points to reduce the computational demand of classical MDS. Consider the geostatistical grid in Figure 3. A classical MDS of these coordinates would require the eigen solution of a 121x121 matrix (121 grid nodes in the model). Consider only calculating the mapping of the coarser landmark grid points. This would require the eigen solution of a 9x9 matrix. Moreover there are fewer distances to calculate, only the $M=L(L-1)/2$ distances between landmark points are needed rather than the full $N(N-1)/2$. Note that this is a classical MDS mapping of the landmark points, to locate the remaining points in the new coordinate system their distances to the landmark points in the original coordinate system are used, requiring NL distances (a *total* of NL inter-point distances are required). This locates all grid nodes in the new Euclidian coordinate system defined by the MDS of the landmark points. This is analogous to trilateration (i.e. determining location based on distance from a known set of satellites as done with GPS).

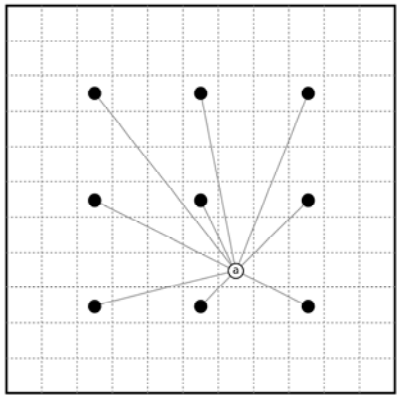


Figure 3: Locating point *a* using distance to landmark points (black).

From a computational perspective, ISOMAP requires the eigen solution of an LxL matrix and NL distance calculations.

Examining only LLE and ISOMAP falls short of a reasonable cross-section of the algorithms available for mapping coordinates; however, they do provide a local and global MDS and they are often used in the literature as benchmarks.

Background: MDS Previous Work

Previous work using MDS in geostatistics focuses on using stream distances for pollution or fish migration studies. The major disadvantage to the techniques used in this area is the classical application of MDS. For these studies large grids are not required, 1D grids oriented along the streams are often used. In the case of geostatistical modeling for reserve prediction, large 3D models containing millions of cells are required; classical MDS would be computationally infeasible.

Curriero (2006) provides a summary of the work that has been done with non-Euclidian distance measures in geostatistics, the interested reader is referred to this summary or the following references: Cressie and Majure (1997) Curriero (1996) Kern and Higdon (1999) Krivoruchko and Gribov (2004) Little, Edward and Porter (1997) Loland and Host (2003) Rathburn (1998), Ver Hoef, Peterson and Theobald (2006) or Sampson and Guttorp (1992). The majority of these works are environmental applications where the distance measure is a stream distance with the waterway assigned a width of 0 (i.e. embed the original stream coordinates into a 1D coordinate system along the stream) or repeated contaminate measurements in time are used to access to the locally varying covariance structure (Sampson and Guttorp, 1992). MDS could be used in all these applications to map a space where the resulting covariance function (variogram) would be valid (Curriero, 2006); otherwise, a kernel method could be used to generate valid covariance models (Ver Hoef, Peterson and Theobald, 2006).

Case Study: Classical MDS

The application of MDS to a possibly indefinite distance matrix has been explained above (Figure 1). The two algorithms that will be used have also been discussed. This section will apply LLE and ISOMAP to the case study explored in the previous paper (Figure 4) where the direction of anisotropy varies locally and the anisotropy ratio is a constant 10:1.

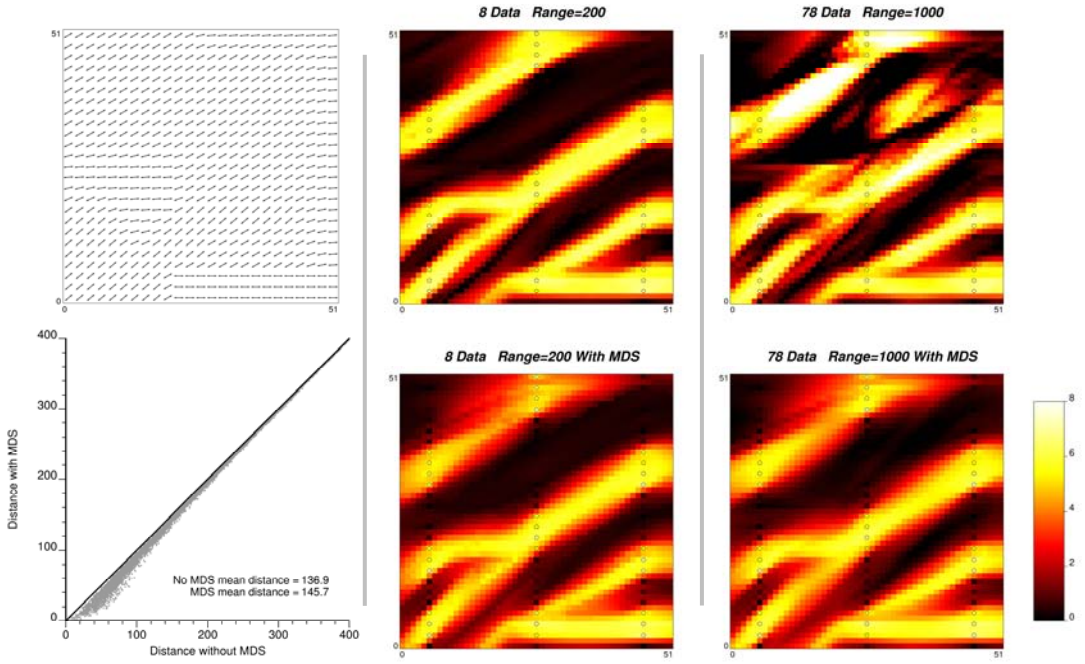


Figure 4: Above left: a channel/vein deposit LVA field. Below left: plot of the difference in inter-point distances when using MDS, 6740 of the 6765201 inter-point distances are plotted. Above center: kriging with 8 samples, range=200 and no MDS. Above right: kriging with 78 samples, range=1000 and no MDS. Below center: kriging with 8 samples, range=200 and classical MDS. Below right: kriging with 78 samples, range=1000 and classical MDS.

The kriging result in this case study (Figure 4 middle) has no indefinite systems because of the implementation parameters (local search for 8 data and an exponential variogram range of 200). Indefinite systems are rare but can still occur; increasing the omni-directional variogram range or increasing the number of data used in kriging may cause indefinite systems. If the number of data are increased to 78 (global kriging) and the variogram range to 1000, indefinite systems occur (Figure 4). Indefinite systems are rare, the range must be increased to almost 20 times the size of the domain in order to cause indefinite systems in this example; however, MDS can be used to ensure positive definiteness (Figure 4 right). In this example classical MDS is used to transform the coordinates to an Euclidean coordinate system in 1007 dimensions where the distances are approximately honored (Figure 4 lower left). In general, the mapping of coordinates by MDS has the effect of increasing the distance between points slightly and smoothing the resulting kriging map (Figure 4 right). This is the price of embedding the coordinates in an Euclidian space.

In the case study, the maximum number of dimensions that could be used would be 2601 (i.e. the number of coordinates transformed); however, the distance matrix between these 2601 points is indefinite and only 1007 dimensions can be kept to ensure positive definiteness (q in Equation 2).

Classical MDS is computationally expensive (Figure 5) as the eigen solution is required for the entire distance matrix (i.e. 2601x2601 in the case study). Extrapolating the fit in Figure 5, the eigen solution for a 100,000 cell model would require 45 days. LLE and ISOMAP will be applied to the case study to reduce the run time of MDS.

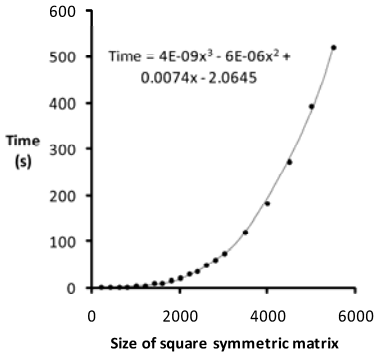


Figure 5: Time to calculate eigenvalues/ vectors using *eig* in *Matlab*.

Case Study: LLE and ISOMAP

Both LLE and ISOMAP are MDS algorithms that reduce the computational demand of MDS. Both will embed the coordinates of the grid into a higher dimensional Euclidean space, guaranteeing positive

definiteness. The dimension of the new coordinate system is determined by the user (as discussed above and shown in Figure 1). Using more dimensions results in a better reproduction of the inter-point distances after MDS. There are two main reasons for considering fewer dimensions (1) it is more computational demanding to calculate the distance between points in a higher Euclidian space and (2) considering 2 or 3 dimensions allows the user to use all the common covariance functions (or variogram structures) that are positive definite in 2 or 3 dimensions. Neither of these arguments are strong. Firstly, the computational demand of calculating distance in higher dimensions is not great, it requires more subtraction and multiplication operations but is only a fraction of the overall cost of kriging. Secondly, the exponential variogram (Equation 4) is conditionally negative definite in all dimensions and is fairly flexible as the power (α) can be varied to obtain the desired fit:

$$\gamma(h) = 1 - \exp\left[-3\left(\frac{h}{a}\right)^{\alpha}\right] \qquad 0 < \alpha \leq 2 \tag{4}$$

Therefore, it is recommended that the maximum number of dimensions (q in Equation 2) be used for MDS. The effect of using fewer dimensions on run time and inter-point distance reproduction will be shown below.

Beyond the number of dimensions both LLE and ISOMAP require a single input parameter; note that classical MDS only requires the number of dimensions. LLE requires a local search radius (Figure 2) while ISOMAP requires the spacing of the landmark points (Figure 3). The effect of these parameters are shown on the case study (Figure 8).

Discussion

The results of LLE are too erratic (Figure 8) to be considered further. In this application, the mapped coordinates from LLE generates unstable kriged maps. There is too much sensitivity to the input search range and number of dimensions. This is undesirable as it would be infeasible in practice to generate multiple mappings for various *epsilon* and *d* values. For this reason, LLE will not be developed. ISOMAP generates more predictable results (Figure 8), the maps transition from poor results (few dimensions/few landmark points) to better results (many dimensions/many landmark points) in a logical manner.

Of interest when considering ISOMAP is run time. Ideally, all possible dimensions would be retained and the number of landmark points would equal the number of grid blocks (classical MDS); this is too computationally expensive (Figure 6). Moreover, classical MDS is often not required and approximations (ISOMAP) can generate acceptable results (Figure 8). In this case study selecting 25 (10x10) landmark points is sufficient. Also note that increasing the number of dimensions does not significantly increase computational demand.

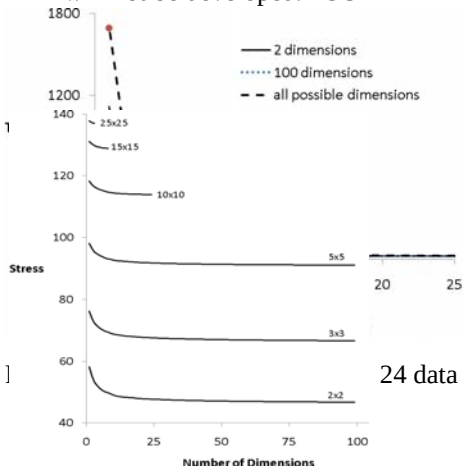


Figure 7: Stress plot for MDS. Each line represents a different configuration of landmark points. Dimensions higher than 100 are not shown.

Parameter selection is critical for a reasonable implementation of ISOMAP. In addition to all parameters that are required in LVA kriging two additional parameters are required with ISOMAP (1) the number of dimensions to map the coordinates and (2) the number and location of landmark points. Selecting the number of dimensions is reasonably straightforward and a traditional variance plot of the principle components could be used, although the stress plot (Figure 7) would be more appropriate. Stress (Equation 1) is a direct measure of how well the inter-point distances are reproduced in the mapped coordinates. When selecting the appropriate number of dimensions it should be noted that the maximum number of dimensions that can be selected with ISOMAP is *L-1* rather than *q* in Equation 2 (Balasubramanian et al 2002).

Selecting the location of the landmark points is more involved. Short of generating many maps and examining acceptable configurations (Figure 8), this is an open research topic. The current implementation

uses regularly spaced landmark points but this is not required. It is anticipated that superior results could be generated by increasing the density of landmark points in model areas where the LVA field is more variable. Areas where the LVA field is smooth would likely require fewer landmark points. Selection of landmark points will be considered in future work.

Conclusions

LVA kriging requires development in two areas (1) guaranteeing positive definiteness and (2) run time. This paper presents a mapping of the grid nodes to a higher dimensional space based on MDS. This mapping attempts to honor the inter-point distances in a high dimensional Euclidian space. In this new coordinate system the distance between points is calculated along a straight line path and approximately honors the inter-point distance in the original space. Positive definiteness is guaranteed. Classical MDS is computationally expensive and infeasible for large geostatistical grids; therefore, two MDS variants were considered (1) LLE was found to be too sensitive to input parameters and (2) ISOMAP was shown to generate acceptable results. Selection of the landmark points will be considered in future work, as will alternative MDS algorithms. Future work will also be directed at improving the run time of the MDS/Dijkstra algorithms so that large geostatistical models can be considered. Moreover, alternative MDS algorithms should be explored. The final area of future work will be to explore the LVA field generation from hard data, soft data, geological interpretations, etc. Incorporating LVA field uncertainty is also required.

References

Balasubramanian, M, Schwartz, E L, Tenenbaum, J B, Silva, V and Langford, J C, 2002. *The Isomap algorithm and topological stability*, Science 295, 7 p.

Cressie, N, and Majure, J J, 1997. *Spatio-temporal statistical modeling of livestock waste in streams.* Journal of Agric. Biol. Envir. Statistics 2:24–47.

Curriero, F C, 1996. *The use of non-Euclidean distances in geostatistics.* PhD Dissertation, Department of Statistics, Kansas State University, 213 p.

Kern, J C and Higdon, D M, 1999. *A distance metric to account for edge effects in spatial data analysis*, *in* Proceeding of the American Statistical Society, Biometrics Section, Alexandria, VA, pp 49–52.

Krivoruchko, K and Gribov, A, 2004. *Geostatistical interpolation and simulation in the presence of barriers*, *in* Geostatistics for Environmental Applications IV, Barcelona, Spain, pp 331–342.

Little, L S, Edwards, D and Porter, E E, 1997. *Kriging in estuaries: As the crow flies, or as the fish swims?* Journal of Exp. Mar. Biol. Ecol. 2:131–11.

Loland, A and Host, G, 2003. *Spatial covariance modelling in a complex costal domain by multidimensional scaling.* Environmetrics 14:307-321.

Mardia, K V, Kent, J T and Bibby, J M, 1979. *Multivariate Analysis.* Academic Press, 521 p.

Rathbun, S L, 1998. *Spatial modelling in irregularly shaped regions: kriging estuaries.* Environmetrics 9:109–129.

Roweis, S and Saul, L, 2000. *Nonlinear dimensionality reduction by locally linear embedding*, Science 290:2323-2326.

Saul, L and Roweis S, 2003. *Think globally, fit locally: unsupervised learning of low dimensional manifolds*. Journal of Machine Learning Research 4:119-155.

Silva, V and Tenenbaum, J B, 2003. *Global versus local methods for nonlinear dimensionality reduction*, in *Advances in Neural Information Processing Systems 15*, (Eds. S Becker, S Thrun and K Obermayer) MIT Press, Cambridge, pp 721–728.

Tenenbaum, J B, Silva, V. and Langford J, 2000. *A global geometric framework for nonlinear dimensionality reduction,* Science 290:2319–2323.

Ver Hoef, J M, Peterson, E and Theobald, D, 2006. *Spatial statistical models that use flow and stream distance.* Environmental and Ecological Statistics 13:449-464.

Watkins, S D, 2007. *The matrix eigenvalue problem: GR and Krylov subspace methods*: Society for Industrial and Applied Mathematics, 452 p.

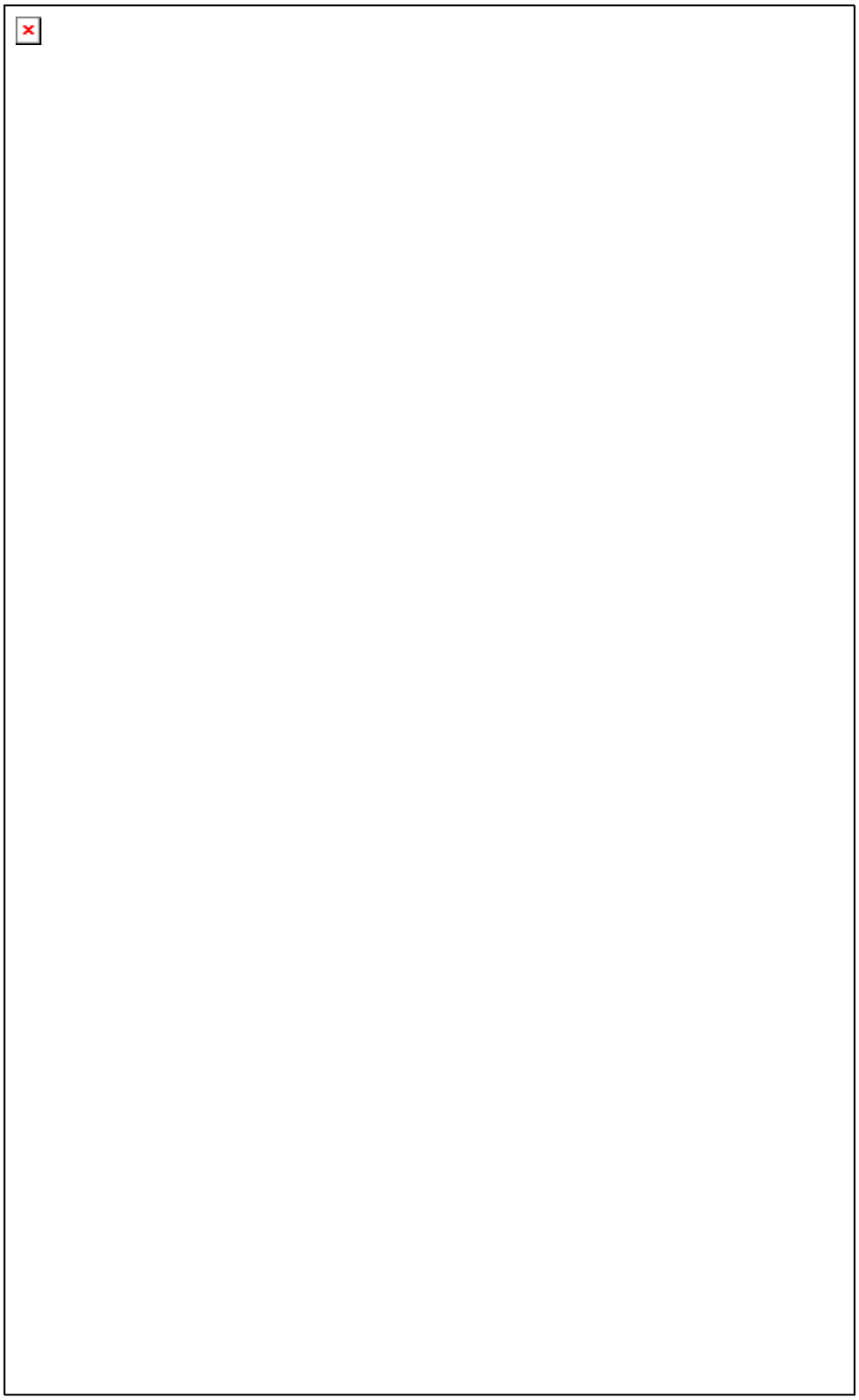


Figure 8: Above plots: ISOMAP kriging with 24 data. Number of dimensions and landmark points are varied (i.e. 25x25 = a landmark point every 25 blocks in the *x* and *y* directions). Lower plots: LLE kriging with 24 data. Neighborhoods are defined by a search 'radius'. If the anisotropic inter-point distance is less than the search, the points are considered neighbors.