Data Questions in GeoComputation

Brian G Lees
Department of Geography,
Australian National University
ACT 0200,
e-mail Brian.Lees@anu.edu.au

AIR NEW ZEALAND

Air New Zealand Guest Keynote Speaker

*Presented at the second annual conference of GeoComputation '97 & SIRC '97,
University of Otago, New Zealand, 26-29 August 1997*

1: Introduction

The new computational solutions which have developed in parallel with the widespread uptake of computer power in geography tend to represent a distinct move away from the more traditional, parametric statistical, methods introduced to the discipline during the so-called quantitative revolution. The implications of adopting these new methods have not yet been fully appreciated by many researchers, indeed the retreat by large sections of the discipline from any serious engagement with quantitative approaches to geographical problems means that there is a growing potential for misuse and abuse of these solutions as they become more accessible.

It is well worth remembering that our new computational solutions comprise several critical components. Firstly, the new hardware configurations without which they could not be implemented. Secondly, the new algorithms themselves. Thirdly, the data, and fourthly, the problem. Perhaps this fourth element could be better described as the problem statement. Success depends on the adequacy of all of these, and on their correct integration. The data, the problem statement and the links between the algorithm and the data, between the data and the problem statement, and the problem statement and the algorithm have all received much less attention than the matching of algorithms and hardware. As we become increasingly engaged in harnessing our new computer power to geocomputation it is perhaps worth casting our eyes over these other, equally important, facets of quantitative investigation, analysis and prediction in the geosciences.

In this discussion I will refer to only two of the large suite of data-driven modelling techniques rather than range across the field. There is simply neither time, nor space, to do otherwise. Most of the points I want to make are relevant to a much wider group of techniques and these are merely convenient examples for discussion. The two I will use are decision trees and so-called back-propagation. Both travel under fancier names on occasion, but these labels are broadly understood and will suffice.

2: Data

Any examination of data in geocomputation needs to consider at least the data distribution, the data model and the way in which the data has been sampled. The precision and accuracy of the data, both in the spatial and attribute domains, also need to be considered. These latter aspects of data have been extensively dealt with elsewhere (Goodchild & Gopal, 1989).

2.1: Data Distribution

After decades of accepting the questionable proposition that most phenomena in the natural world are normally distributed we are now adopting non-parametric methodologies with enthusiasm. The fact that many of these can be parallelised has, perhaps, added to this enthusiasm. There are some costs in this enthusiasm. It is safe to say that, given a data distribution approximating a normal distribution, parametric methods tend to produce a better result than non-parametric methods unless large, or carefully chosen, samples are selected.

To produce results of matching quality in the analysis of non-normally distributed data using non-parametric methods, much larger and more carefully structured samples are needed. By definition, non-parametric, supervised inductive learning systems have no information on the distribution of the data other than that which can be inferred from the learning sample. Few of the studies which have appeared in the literature indicate that this has been recognised. That this is such a problem is perhaps due to the fact that many of those involved in this branch of geocomputation are not data gatherers, but data processors. There is a real temptation to use 'legacy' data sets for experiments in geocomputation and this leads to the use of proportional samples. Given the error minimisation rule on which many of these systems are based, the use of a proportional sample as a learning sample will bias the system towards the largest categories.

2.2: Data Models

Much of the published work on data models focuses on the data model as the rationale for organising data in the computer. In computer science it is a means of capturing the semantics of the data through definitions of the operations related to classes, describing which combinations of operations are legal, which combinations of operations are equivalent, and consistency constraints among data. This bias towards the computer science view of data models is quite understandable as it is a necessary tool to deal with the data, but many phenomena have not been carefully scrutinised by domain experts in the same way and I suspect that, when this happens, the whole concept of data model will become considerably more complex and critical.

When we start to consider whether the measure used to code the data is appropriate to the phenomena we wish to examine we need to remember that many disciplines, including geography, routinely classify data as part of their collection protocols. This pre-analysis processing is often not recognised as such but can be a major limitation to accurate prediction based on such sampling. All too often phenomena distributed as a continuum are discretised into gaussians on the assumption that this is an appropriate data model for the phenomenon. The type of measure used is also critical. The use of nominal measures, rather than ratio or interval measures, increases the requirement for an unbiased sample significantly. Whilst ordinal measures are not as difficult to deal with as nominal measures, they are considerably less informative than, say, an interval measure.

The problem outlined above is the natural consequence of a habit widespread through many disciplines. The classification of data prior to analysis is almost an unconscious act for many field scientists. That this is unnecessary now that we are no longer bound by the cartographic model of spatial data has not really penetrated the consciousness, and standard procedures, of many disciplines. Indeed, in many cases, the data collection itself imposes this structure. The step between observation, and the recording of that observation is often one in which some form of classification takes place. The value of each observation, as a unique data point, is then immediately degraded.

All other things being equal, if one can provide a learning system with some indication of how values in an attribute relate one to another, then the system will do a better job. Humans like to simplify these relationships as we are unable to deal very effectively with high frequency variability in data. By coding data to suit human perceptions, we degrade it and remove information a non-human learning system may be able to interpret. For example, in many natural systems tasks geology is an important variable. The taxonomy in geology being what it is, the relationship between a granite, an Essexite and a Monzonite, and the lack of a close relationship between those and a Sandstone are not apparent (to an algorithm) from the class numbers used to represent these in a GIS. It is necessary to recode these categories using some appropriate interval or ordinal scale, in the case of an erosion study 'K' values would be appropriate. The 'K' value is a ratio value with a direct relationship to erodibilty. In the case of vegetation modelling, geology can be recoded according to some interval or ordinal scale of nutrient status. Deriving appropriate

measures requires a knowledge of both the attribute, and the interactions of attributes relating to the phenomenon being modelled.

These simple pre-processing stages are needed overcome the knowledge gap which exists between human and algorithmic 'intelligence'. Most natural scientists understand the relative difference in nutrient status between weathered granite and sandstone. The names communicate a suite of attributes to the expert human listener. Unfortunately, there is no inherent information in the terminology to inform either the non-expert human or algorithm. Even worse, because of the necessity of labelling attribute classes with numerical identifiers when data is imported to a GIS, there is sometimes a tendency to carry out analyses which improperly utilise the mathematical relationships between identifiers, when no such relationship is implied. This is a common trap for non-expert users, but it is also a trap for expert users working with data from domains in which they are not expert.

2.3: Data Sampling

I have already mentioned the importance of sample characteristics briefly. In the use of optimising, or error minimisation, techniques, it is important that each case one wishes to predict or classify is equally well represented in the learning sample. Proportional sampling techniques will not produce this. One must resort to quite structured, stratified methods to achieve this sort of sample. One must also attend closely to the scale at which one samples. Now that we can move away from the restrictions of the cartographic model, many disciplines have not yet understood that data scale and display scale are no longer synonymous and need to be considered separately. For our purposes, the display scale is much less important than the scale at which the data was measured. This is particularly true when one is looking at context, spatial or temporal.

Both spatial and temporal variability are strongly scale dependant. There is a general trend in most land cover data for spatial autocorrelation to be low at fine scale, to rise to a maximum at an intermediate scale and then to decline. One can see a similar pattern in many forms of temporal data. The diurnal range of bio-activity, illumination, temperature and pressure is often nearly as great as the annual range (based on daily observations), and much greater than the inter-annual range. We need to move to epochal time scales to see the diurnal range exceeded. We filter out the fine scale variations when we make observations, but we tend do this informally. To reduce data based error in analyses it is important that we exercise more conscious control of input data scale. If we cannot control it, then we need to be aware of the consequent errors.

Spatial and temporal variability also depends on the data space, or domain, in which one views the data. Spatial data exists in a number of discrete domains (Lees, 1994; Aspinall & Lees, 1995). In each of these there exist topological relationships, but these relationships vary from domain to domain. We are most familiar with spatial data existing in a geographic space defined by latitude, longitude and elevation. Movement from point to point in this space is a vector. It is not possible to move from one point to another without transiting intermediate points. Each point is unique.

In the other, conceptual, domains or data spaces topological relationships are different. These data spaces can be spectral space, environmental data space, even socio-economic data space. The fundamental, and shared, characteristic of these spaces is that movement through the space has a logical meaning. Spectral space, for example, forms the basis for most analysis of remotely sensed data. Proximity suggests similar colour. Trajectories of reflectance values for developing crops on different soils form the basis for the common Kauth-Thomas, or Tasseled Cap, transformation. Trajectories in spectral space form the basis for sub-pixel modelling of vegetation structure. In these analyses vectors represent changes in the reflectance at a point, through time. No motion in geographic space is envisioned. A large number of points in geographic space can occupy a single location in spectral space. The converse is not true.

In environmental data space, the basis for environmental domain analysis, topological relationships are linked directly to environmental gradients. Vectors in this space drive the continuum of change in vegetation composition observed in nature. The conflict in ecological literature between those who favour a community view of vegetation and those who view it as a continuum lies squarely on the fact that community is a spatial concept in geographic space, whilst the continuum is a spatial concept in environmental data space (Austin and Smith, 1989). Both are common representations, but fundamentally different in the way they can be analysed. In geographic space one can move from one point to another along a vector. This same motion in environmental data space may result in no motion, if the environments along this vector in geographic space are the same, or a jump from point to point if say, a soil boundary is crossed. As before, a large number of points in geographic space can occupy a single location in environmental data space and, once again, the converse is not true.

This particular dichotemy, between representation of vegetation distribution in geographic space and environmental data space, is a dichotomy between data models. The 'mapping' school reduce observations of vegetation to a series of vegetation classes, even forest types. In some ecosystems, particularly Australian eucalypt forests, these class boundaries are cultural (statistical) artefacts. Slight changes in contribution to the canopy can lead to a change in class. In such cases, there is often more variation within the class than between classes. Nevertheless, the fundamental structure of choropleth mapping requires this reduction of variance to permit the mapping of polygons. This mismatch between the phenomenology of the data and the data model, excusable in the days where choropleth mapping was the only means of representation, has been carried forward to the present.

Domain knowledge is fundamental to constructing the necessary spaces for analysis, and for understanding the relationships between the spaces. In many problems different parts of the analysis need to be carried out in different data spaces. Importantly, a sample which can be considered to be representative in one domain may not be representative in another. Sampling strategies therefore need to consider the data distributions in all of the relevant domains.

3: Interactions Between the Algorithm and Data

In parametric statistics a classifier is an algorithm, in non-parametric, data-driven analyses the classifier results from the interaction between an algorithm and a learning sample. The characteristics of the learning sample determine, to a large degree, the behaviour of the classifier. Careful design of learning samples is vital for good performance in this area. The behaviours of the different algorithms in the way they use the learning sample is also very important in the design of analyses.

3.1: Decision Trees

The recursive partitioning which is the basis for decision tree algorithms seemed to be an ideal strategy for dealing with the data domain problem. Each split, or decision rule, is made in only one data domain. The tree building (learning) procedure moves from data domain to data domain as it searches for optimum splits and makes only minimal assumptions about the relationships between variables. This sort of inductive learning produces clear and explicit results. Careful monitoring of the derived rules is necessary to identify rules based on statistical artefacts rather than process relationships. This monitoring, preferably by a domain expert, is vital to weed out nonsensical relationships which would induce error when the tree was used as a classifier. High correlations between independent variables often confuse this sort of system. For example, in the modelling of vegetation distribution around Kioloa a decision tree may indicate that elevation is an important variable. Examination of the tree will show that geology is an alternate split at that point. The high correlation between geology and elevation in the Kioloa learning set is a statistical artefact of the data set. The area is predominantly Sydney Basin sediments which are flat lying. Changes in geology correlate with changes in elevation for much of

the data set and the digital elevation model is the higher resolution variable. It therefore comes up as being more significantly related to change in tree species than does geology. However, as the elevations concerned are not extreme enough to generate significant climatic gradients, it is clear that the process driving the change in species is the slight change in nutrient status associated with the different geology types. A domain expert would be able to identify this quite readily and change the variable at that point accordingly. Slope also acts as a useful correlate for changes in geology, often at scales well below that at which geological information is available. This explicit nature of decision trees is very attractive but in many applications does not offset their hunger for huge learning samples.

3.2: Artificial Neural Nets

Having experimented with the decision tree approach for some time with good results (Moore et al.,. 1991; Lees & Ritman, 1991) it became clear that, for some applications, the amount of learning data required to produce the required level of discrimination (number of classes) was impractically high. This is particularly true where some classes are poorly represented in the learning sample as, with a stopping point of 25 or 30 points, many classes simply have no chance of being predicted. It is possible to plot probability surfaces, or fuzzy set membership, using the membership of the populations at each terminal node to overcome this, but these problems prompted a further search for methods less hungry for data. After a short search, several types of Artificial Neural Net appeared to offer attractive solutions to the problem (Fitzgerald & Lees, 1993; 1994). It is useful to think of some of these algorithms as doing in parallel what decision trees do in series.

Artificial Neural Nets are a field, rather than a group, of quite unrelated algorithms. Many originated as projects to understand human information processing and were never intended as the analytical tools they are now sometimes seen as being. Neural Nets are part of a suite of data-driven modelling techniques which are useful when the processes underlying a phenomenon are either unknown, only partially known, or would necessitate the generation of an impracticable level (scale, volume or cost) of input data. Within the suite of data-driven techniques they are useful for dealing with non-parametric data when there is insufficient data to use a more explicit technique such as decision trees. The sigmoid and hyperbolic tangent transfer functions used mean that neural nets are rather better at dealing with fuzzy data than the crisp logic of decision trees. Two types of approach are of particular interest in this context. One can be roughly typed as an unsupervised approach, the other as a supervised approached. In some network configurations these can be combined.

The unsupervised approach is exemplified by the Kohonen network or by Self Organising Maps (SOMs) (Kohonen, 1984). A Kohonen network is a single layer of neurodes. Their initial values are set randomly. As each input (training) vector is fed to the layer the neurode with a value closest to the input vector fires. This 'win' by the successful neurode is 'rewarded' by the neurode being allowed to migrate its value closer to that of the input value. Its neighbours are similarly rewarded by being allowed to migrate their values towards the input value, but by a smaller amount. This procedure continues until the Kohonen layer has developed a pattern where similar values are closely adjacent. in the layer. This behaviour is similar to that of a decision tree with the neurodes at the end of training being roughly equivalent to the terminal nodes of a tree. procedure is organising the layer in response to similarities in the input vectors. Like decision trees it can result in a number of neurodes, often widely separated, being used to produce a single class in the final thematic map. The problem with this is that no information on the level of discrimination required is being supplied to the training procedure. This is where understanding the link between the problem and the data is very important. The algorithm is grouping the input vectors and has no information on how this relates to a useful output. In some projects this is not a problem. However, if one is trying to produce a thematic map with classes representing the sea, non-forest areas and , say, ten forest types there is a level of

imbalance in the level of discrimination being sought. In order to produce the ten forest types, one would have to produce perhaps as many grassland and non-forest land cover types, probably many more, and as many shallow/deep water classes. This makes the number of neurodes required in the Kohonen layer quite large and consequently, the learning time considerably longer. If this is not done, then one can suppress variability which is needed for subtle discrimination between closely allied classes.

Supervised procedures can avoid this and the commonly used Back Propagation Network is a good model to discuss in this context (Rumelhart & McClelland, 1986). The input vectors are passed down through a multi-layered network. In the training phase, the output layer is compared to the known (or desired) output value or class associated with the input vector. If the output is in error the network weights are altered slightly to reduce the chance of this path being followed next time. If you were a Skinnerian dealing with rats, this could be described as punishing the network for its mistake. Samples are randomly drawn from the training data for as many iterations as are necessary. After a while, the network error rate will tend to stabilise and training can cease. This ability to use the training sample for as many iterations as are necessary is one of the most attractive features of neural nets.

Neural nets of the type discussed here (BPN) work best with a representative learning sample which is made up of vectors which are modal to the desired output classes. If this is done the learning sample size can be kept small. This keeps the degrees of freedom low and increases the level of confidence in the final result.

3.3: Pushing Things to the Limit

Unlike decision trees, BPN can be remarkably tolerant of noisy data if handled carefully. If much of what has gone before sounds like an impossible string of motherhood statements about how we need to clean up our data for these systems, then it is heartening to have a technique which, if used carefully, can cope quite nicely with the realities of data. Indeed, one can even structure investigations which take advantage of this characteristic and are probably not achievable using any other method.

This might best be illustrated using the example of an exercise we carried out across the Liverpool Plains in the Murray Darling Basin. They form part of a highly productive agricultural area, increasingly affected by dryland salinity, which is estimated to cost $10 million per annum in lost agricultural production. Cropping in the area is highly variable, temporally and spatially, as a result of opportunity, summer and winter cropping cycles, and strip and broadacre paddocks. The Liverpool Plains cover an area of 1.2 million ha.

Hydrologically, the Plains are considered as an evaporative basin with a small leak, rather than a fluvial system. Groundwater movement through the basin is complex, and dominated by salinity gradients, microtopographic features and subtle lithological heterogeneities rather than topographic slope. Accurate modelling of this movement would require detailed, and expensive, sub-surface data. An alternative was to attempt to identify empirical evidence of the groundwater movement on the surface and infer its behaviour from that. A first step in doing this was to try to use remotely sensed data. The Liverpool Plains region have a dominant pattern of intensive agriculture. Slight variation in cropping responses due, in the main, to the geochemistry of the soils, is detectable in some places. This is a classic signal detection problem. We are looking for a change in signal on which we can base management strategies. The dominant pattern/signal does not relate to salinity and tends to overwhelm the pattern/signal which may do. In order to provide a more useful management tool we set out to teach a neural network to discriminate the dominant spatial pattern of agriculture, using GIS, and to process the remotely sensed data as though there were no field boundaries and only one crop present.

With an optimising technique to work on this data the number of 'hit' cells in the presence data must be greater than the number of 'miss' cells. Conversely, the number of 'miss' cells in the absence data must be greater than the number of 'hit' cells. If these differences are great, then

the network will converge on an optimum solution without a great deal of trouble. However, the less significant the differences the more care must be taken in setting the learning rate to achieve some sort of convergence.

In order to achieve this we classified SPOT imagery over the area and constructed a polygon coverage of field pattern. Using a modal filter we then labelled these polygons with the modal spectral class within the polygon. This created a simplified image of the land cover, one which 'tends' to be true. There is no assumption that these classes correspond to any particular crop or land cover. We then selected a class which was well represented and was adjacent, at some location or other, to most of the other classes to be the reference class. Using the questionable principle that soil characteristics will not change dramatically over short distances, we then labelled points in each field class as being equivalent to the reflectance value of a neighbouring point in the adjacent reference class field. Because of the necessity to avoid mixels along the field boundaries these two locations were spaced about four cells apart. We then trained a network to learn that the correct reflectance for these points tended to be that of their neighbours across the fence. If this had been true, then all that would have been necessary to do would have been to construct a simple look-up table. Because it only 'tended' to be true, we needed to structure a network learning exercise as though we were dealing with a very poor, or noisy, learning sample. This involved setting a very low learning rate, over a large number of iterations.

The network extracted patterns which appear to represent real geomorphic features, We are now carrying out chemical tests on soil samples to identify the characteristics which are identifiable by the network. This is necessary because the Liverpool Plains are covered by one of the most visually monotonous and homogeneous surfaces it has been my misfortune to deal with. If results from such tests are promising, the network can be further developed and field tested over a larger area. The advantages of this particular methodology, if proven to be a successful predictive tool that can be replicated on scenes from different dates, are that it requires limited input and is independent of vegetation and therefore of growing conditions and cropping cycle, year and stage in season.

Perhaps its naughty to use the tolerance of the algorithm to bad data in this way, but it does illustrate that a good understanding of the interactions between the algorithm and the data can pay off in unexpected ways.

4: Conclusion

In such a sweeping review as this its difficult to point to a single, tight conclusion. It is however possible to say that times and techniques are changing rapidly and that it is very important not to be distracted from the necessary houskeeping tasks of data management by the fascinating range of new techniques becoming available to us. Indeed, given the characteristics of many of these new techniques in geocomputation, these are perhaps more important than ever.

Bibliography

Aspinall, R. & Lees, B.G. 1995. 'Sampling and analysis of spatial environmental data.' in Waugh, T.C. & Healey, R.G. (eds) Advances in GIS Research, Taylor and Francis, Southhampton, 1086-1099.

Austin, M.P. & Smith, T.M. 1989. 'A new model for the continuum concept.' Vegetatio, 83: 35-47.

Fitzgerald, R.W. & Lees, B.G., 1993. Assessing the classification accuracy of multisource remote sensing data. REMOTE SENSING OF THE ENVIRONMENT, 41: 1-25.

Fitzgerald R.W., & Lees, B.G. 1994. 'Spatial context and scale relationships in raster data for thematic mapping in natural systems.' in Waugh, T.C. & Healey, R.G. (eds) Advances in GIS Research, Taylor and Francis, Southhampton, 462-475.

Kohonen, T., 1984. Self-Organisation and Associative Memory. Springer Verlag, New York.

Lees, B.G. 1994. 'Decision trees, artificial Neural Networks and Genetic Algorithms for classification of remotely sensed and ancillary data.' Proceedings 7th Australa-

sian Remote Sensing Conference, Remote Sensing and Photogrammetry Association, Australia Ltd, Floreat, W.A. v1: 51-60.

Lees, B.G. & Ritman, K. 1991. A decision tree and rule induction approach to the integration of remotely sensed and GIS data in the mapping of vegetation in disturbed or hilly environments. ENVIRONMENTAL MANAGEMENT, 15, 823-831.

Moore, D.M., Lees, B.G. & Davey, S. 1991 'A new method for predicting vegetation distributions using Decision Tree Analyses in a Geographic Information System.' ENVIRONMENTAL MANAGEMENT, 15, 59-71.

Rumelhart, D. & McClelland, J., 1986. Parallel Distributed Processing, Vols 1 & 2. MIT Press.