18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

An optimisation of the survey gap analysis technique to minimise computational complexity and memory resources in order to accommodate fine grain environmental and site data

Manion, G.[1], Ridges, M.[2]

[1] *Landscape modeling & Decision Support Section, NSW Department of Environment and Climate Change*
[2] *Regional Assessment Unit, Culture & Heritage Division, NSW Department of Environment and Climate Change*
Email: glenn.manion@environment.nsw.gov.au

Abstract: Survey Gap Analysis is a valuable and widely used technique to determine the adequacy of a set of sample collection points to span an environmental ordination space. However, when fine grain environmental data is required to be used over relatively large spatial domains, the unavoidable 'greedy' nature of this algorithm can easily consume computational memory resources, especially desktop computers which are the primary hardware platforms used by researchers for this type of analysis.

This paper outlines a technique of optimising the speed of the Survey-Gap technique by dividing key components of the execution of the algorithm between RAM and hard-drive space. This technique exploits the speed of physical memory for fine grain calculations and the capacity of hard-drive space for storing essential environmental distance metrics to facilitate iteration of the algorithm without recalculation.

While the size of a binary file under a 32-bit operating system is usually limited to
2^31 – 1, (2147483647) bytes to facilitate random access, the exploitation of sequential reading methods to allow for the file size to greatly exceed the aforementioned limit imposed under 32 bit operating systems.

Survey-Gap analysis utilises the concept of Environmental Distance. That is, the sum of Manhattan distances between any pair of sites in a geographic domain. This is done by 'drilling down' through a stack of environmental and climatic grid surfaces and collecting a pair of vectors of grid values from which the Manhattan distance is extracted.

Survey-Gap analysis is more often than not applied in an iterative manner for the selection of multiple sites, with key lookup sections of the algorithm assessing data that remains unchanged between iterations. This provides the opportunity to store much of the information required on a hard drive. This method has most of the initial calculation work being done in the first iteration and successive runs using this data in lookup form.

The solution used has enabled the algorithm to be applied to web-based processing and broader applications such as estimating the reliability of spatial predictive modelling.

Keywords: *Survey-Gap Analysis (SGA), Environmental Diversity (ED).*

Manion G, Ridges M. An optimisation of the survey gap analysis technique to minimise computational complexity and memory resources in order to accommodate fine grain environmental and site data.

1. INTRODUCTION

A vital facet of understanding biodiversity within any bio-geographical region is the ability to determine species and assemblage distribution within a particular area. This is also just as important for understanding distribution of cultural and anthropogenic artifacts. Both these application require knowledge of the spatial distribution of elements from collection data.

However, it is rare to have collection data that can be considered complete enough to make reasonable predictions as to the spatial distribution of the target element across under-sampled regions. In terms of biodiversity, several international initiatives to link museum collections online, such as the Global Biodiversity Information Facility (GBIF), have highlighted the shortcomings and shear costs of compiling databases from data that suffers from spatial inaccuracy and to a lesser extent, misidentification. Because all the target elements in a region can never be directly observed and counted, practical identification of the relative biodiversity or cultural assemblages depends on surrogate information.

Undertaking expeditions to identify and map cultural and biological distributions has always and will always been extremely expensive. Competent and experienced professionals who are willing to travel to remote areas for months and even years are difficult to find and largely underappreciated by administrators and collection agencies.

Compounding this problem are the issues of limited access to potential survey sites because of differing land tenure, ever-changing sovereign government policies and attitudes to scientific collection and the shear tyranny of distance between suitable survey sites.

As a result, novel techniques have evolved using a combination of relatively cheap computing power, geographic information systems, museum collection data and access to increasingly accurate environmental data. This enables 'virtual' surveys to be done using computer software, and then, if the resources are available, areas that have shown to be under-sampled can be visited for directed mini-surveys. Survey-design methods such as gap analysis have become the tools of choice for improving the quality and quantity of biological and cultural data.

2. SURVEY-GAP ANALYSIS

2.1. Overview

Survey-Gap Analysis is a technique based on the finding that sampling different parts of the environmental space will yield a good representation of the biological diversity of a region (Faith & Walker, 1996), and that this technique can equally be applied to the problem of selecting survey sites. The survey-gap analysis tool developed by the NSW Department of Environment and Climate Change in Armidale, NSW (NSW NPWS, 1998; Ferrier, 2002) analyses the survey coverage of a region in relation to the underlying continuous environmental and geographical space. The survey-gap analysis tool adapted Faith & Walker's (1996) environmental diversity (ED) measure and was developed for selecting sets of sites that represent regional biodiversity by providing the best possible coverage of regional environmental variation (Funk et al 2005).

In the context of the survey gap analysis tool, the objects in the ED analysis are the geographical location of the sites, which are represented as a matrix of pair-wise distances. The matrix can be assumed to indicate relative underlying feature relationships. The pattern exhibited by these relationships provides predictions to the degree of complementarity between a particular site to any given set of sites. That is, the overall ED value, the p-median, will decrease at a larger rate for sites with higher ED complementarity. Using this property, sites that will contribute to the greatest reduction in p-median could be considered candidates for survey.

The sort of environmental data used in conjunction with the site data are predictors such as temperature, rainfall and geology. For cultural artifact analyses, other surrogate layers that are used are distance to streams and digital elevation models. These continuous data sets are represented as GIS rasters which in this context can be considered as matrices.

Manion G, Ridges M. An optimisation of the survey gap analysis technique to minimise computational complexity and memory resources in order to accommodate fine grain environmental and site data.

2.2. Description of the algorithm

All distance measures are Manhattan distances . That is, for any two sites or any two grid-cells in the stack of environmental raster grids, the distance between the two sites is expressed as the sum of the absolute values between the environmental values for the two sites in each grid.

$$D_{i,j} = \sum_{g=1}^{g=n} | g_i - g_j |$$

where i and j are sites and g refer to a stack of n environmental grids

The input grids should be range standardized and weighted if necessary prior to running SGA to avoid numeric overflows if too many values are summed when calculating the p-median. Create an evenly spaced sub-set set of data points that act as a covering mesh across the grid data. These points are referred to as the **Demand Points**. The set of current collection sites used in the SGA are referred to as the **Survey Points**.

For each demand point, store the manhattan distance between it and its nearest survey point i.e. site that has the smallest manhattan distance between it and each demand point in an array ***nearest_survey_site*** in the following pseudo code.

set function ***Manhattan(i,j)*** to be the manhattan distance between site i and site j
allocate an array named ***tmpPMvalues*** of size = number of demand points
create a grid ***OutputGrid*** to hold p-median values from SGA calculations

```
for each grid cell  GC
{
   Stage 1: initialising the  gridcell distance to each demand point
   for each demand point  DP
   {
         set variable dist  = Manhattan (GC, demandPoints[DP])

         set variable diff  = nearest_survey_site[DP]  -  dist

         ( only add  to tmpPMvalues if  environmental distance has improved)
         if diff  >  0 then tmpPMvalues[DP] += diff

   }

   Stage 2: calculating the p-median…
   set variable p-median = 0
   for each demand point  DP
   {
      set variable diff =  nearest_survey_site[DP] -  tmpPMvalues[DP]

      (only add to p-median if environmental distance has improved)
      if diff  >  0 then         p-median += diff

   }

   ( set the output for this cell to the average p-median)
   set OutputGrid[GC] to p-median /  number of demand points
}
```

The output grid will now contain a set of continuous values from 0 to the maximum p-median. The cell with the largest p-median value will be the best candidate for a new survey site. This chosen site is then added to the current list of survey sites and the algorithm is re-run as many times as desired to produce a collection of new survey sites.

Manion G, Ridges M. An optimisation of the survey gap analysis technique to minimise computational complexity and memory resources in order to accommodate fine grain environmental and site data.

3. AN APPLICATION OF THE NON-OPTIMISED METHOD

One of challenges for interpreting the predictions of a model describing the expected spatial occurrence of a species or cultural feature through space is to understand the reliability of those predictions (Fielding 2002). SGA can be used as an accompanying method to predictive modeling where generating spatial surfaces describing the distance to the p-median can be re-interpreted as a function of the representativeness of the input sample sites for the domain of the predictive model. As the distance to the p-median increases for any given cell, the interpreted reliability of the model at that cell can be inferred to fall as the model effectively has a poorer sample of the environmental ordination space at that site from which to construct the model.

An application of this approach was applied to a study constructing predictive models for the occurrence of Aboriginal features across NSW. The predictive models provide a baseline for developing conservation strategies to protect Aboriginal Heritage across the state. The models were developed using a logistic regression technique implemented in S-Plus 7.0 using the GRASP tool (Lehmann 2000) that implements a non-linear generalized additive model (GAM). The variables used were a suite of environmental layers commonly employed as surrogates for factors affecting the distribution of Aboriginal features in the landscape (Ridges 2006). These included various distance to water measures, various measures of terrain, pre-European settlement vegetation, soils, geology, and a variable describing the visibility of prominent points in the landscape. Distance to water measures were derived using a cost-distance function, where the origin was streamlines, and the cost was slope adjusted for walking speed using Tobler's (1993) equations. The cost-distance function was performed separately for each stream order, and then combined with the stream order value as a weighting (see Ridges 2006, 128). Terrain measures included elevation, slope, aspect, and curvature derived with standard GIS functions.

The input datasets for the SGA analysis were quite large, covering NSW at a resolution of 100m (80,225,671 data cells). Parameters for the SGA run included the use of 50,000 demand points, 17 environmental layers, and variable weights derived from the relative contribution of each variable to variance explained within the GAM model derived with GRASP. The run was performed on an average desktop PC (Windows XP SP2, Pentium 4 (HT) 3.2 GHz- 512KB L2 cache, System bus 800 MHz, physical memory 1024MB, virtual memory 2048 MB), and took 118 hours to complete.

The application of an SGA technique to mapping model reliability spatially is significant in the context of other model testing approaches. For instance, independent data can be used to quantify model performance, monte-carlo type tests can quantify model consistency, and goodness-of fit measures like the ROC statistic (Fielding 2002, 276) quantify the ability of the model to discriminate the phenomenon of interest effectively. Although these approaches are commonly used to report the fitted accuracy of GAM models, they do not provide a spatial representation of that accuracy. In on-the-ground applications, having a spatial map of both the models predictions and their reliability is a powerful tool to assist more informed decision making.

4. POTENTIAL PROBLEMS ON DESKTOP COMPUTERS

SGA is a 'greedy' algorithm. There is no getting away from having to visit all the data cells in the input grids. The collection of demand points needs to be visited twice for each grid cell, once to 'train' the demand points (S*tage1*), and again to calculate the p-median for the grid cell (*Stage 2).*

For most applications, the grid data will contain thousands, if not millions of data cells. A suitable number of demand points to adequately cover the grid data could be say 1000. The nature of the SGA technique will involve running the algorithm not just once but many times. The cell that best reduces the p-median is added to the list of survey sites and the SGA is run again. It is quite normal to run this algorithm 10 or more times to accumulate a set of potential survey sites.

Thus, an estimate of the atomic operations for SGA would be $10 * 2 * 1000 * \alpha *$ grid cells, allowing for 10 iterations with 1000 demand points and α being the number of environmental grids involved in the analysis. To add to this complexity, the number of grid cells will grow at a quadratic rate as

the resolution of the input grids increase. The data type used for such calculations would be at least a single-precision floating point type (4 byte) up to a double-precision floating point type (8 bytes).

Also, under the windows operating system there is the added constraint of a 2 gigabyte process space for any application so a survey-gap analysis on a desktop computer could easily overflow available memory, thrash hard-drives and consume all CPU resources. This situation could occur with low numbers of relatively course data grids, and as such could make SGA somewhat unusable will all but 'toy' data.

4.1 Solution strategy

While there is no real solution to the greedy nature of the SGA algorithm, the following strategies have made analysis with reasonable sized data possible on desktop computers. This has involves a trade-off between data in memory and data on permanent hard-drive storage and the minimization of reads and writes to and from the hard-drive into memory.

Given that the results of the calculations for *Stage1* remain unchanged over any number of iterations of the SGA, the first optimization is to implement *Stage1* of the algorithm as a binary file lookup table located on the hard drive. This is done by sequentially writing the environmental distance between each grid-cell and each of the demand points to a binary file located on the computer hard drive. Given that these distances don't change under any iterations of the SGA algorithm this file becomes a lookup table for selecting each successive potential new survey site.

It is useful to note that the access to each respective data grid cell remains unchanged under iteration as grids cells with no-data are skipped in the same order. Binary files on 32 bit operating systems using the common FAT32 format have a limit of 2147483647 bytes in size but, this is only for random access where the seek offset must be a signed 4 byte integer. However because this lookup table is accessed sequentially from the start to the end of the file for each new survey point selection the file can far exceed this limit. It is only limited by the physical hard-drive space on the computer running the analysis.

Say that one is using 1000 demand sites**,** the lookup table would be a sequential binary file of single precision values (4 bytes) where the manhattan distance between each grid cell and the demand points are stored as records. For each grid cell, the record contains 4000 bytes and therefore, a data set with 20 million data cells would require a binary file of around 80 gigabytes. For the entire SGA, this binary file is kept open. When selecting each new potential survey point, the file pointer is simply returned to the start of the file and the SGA begins again.

To optimise the reading of this binary file, the data could be read for an entire grid row at once. So for a grid of 5000 columns by 4000 rows and 1000 demand points, the data is read into memory in chunks of 5000 * 1000 * 4 bytes, only 20meg. This would need to be done only 4000 times for selection of each new survey point with a grid of this size.

Such a strategy would enable a normal laptop computer with standard resources to run a fine grain SGA easily. Of course the CPU speed will be the determining factor in how long the whole analysis will take. The advantage to this strategy is that the lookup table only needs to be created once and is reused for each iteration of the SGA.

The second stage of the SGA uses the data read from the lookup table and is basically the same as described above in the pseudo code. Thus getting the first survey site will take the longest but successive new survey sites will be derived some magnitudes faster. If the user intended to use the lookup table for some future survey site selection then the binary file could be compressed and stored away or simply deleted if no longer required.

4. AN APPLICATION OF THE OPTIMISED METHOD

An opportunity to use this optimized method arose from the task of designing a web-based version of the SGA tool for the GBIF-MAPA application. This application used species site data derived from museum datasets that were linked to the GBIF site and used environmental data in the form of raster files. These raster files comprised 35 ANUCLIM predictors encompassing temperature, precipitation, moisture and radiation indices and were in 5 km and 20 km resolutions. The geographic area for the data was limited to continental Australia and as such the 5km grids had 298664 data cells and the 20km grid had

19312 data cells. To use datasets of the resolution described above in the non-optimised example would have been untenable what with a single run of the SGA taking 188 hours to complete and requiring a lookup table some 16TerraBytes in size, clearly not a good idea for a web application.

This application enables a user to utilise locations of existing specimen records and mapped environmental variables to create a mapped complementarity surface indicating the relative priority for additional survey or collection effort throughout the region of interest. The priority being based on the potential for an area (based on climatic conditions) to compliment existing survey effort in the region of interest. The choice of environmental rasters with 5km or 20km resolution was the first way to reduce the complexity of the algorithm. Using 1000 demand points would therefore require a lookup file of 298664 * 1000 * 4 bytes resulting in a 1.2Gig lookup file for the 5km dataset and 19312 * 1000 * 4 bytes resulting in a 77Meg lookup file for the 20km dataset.

Using a 5km resolution grid set for continental Australia with 298664 data cells, 1000 demand points and site data for genus Macropus from the GBIF data set, a 10 iteration SGA took 60 seconds to create the lookup table and each iteration to extract the p-median grid and choose the optimal site took 5 seconds. Thus the whole operation took under two minutes as opposed to approximately ten minutes if the non-optimised version was used for this purpose.

5. DISSCUSSION AND CONCLUSIONS

As can be seen by the two examples described above, the SGA algorithm can be applied to widely varying applications. Neither the optimized or the non-optimised methods serve to replace each other but serve to complement the application of SGA to different problems. The non-optimised method traded computation time for increased spatial resolution and had the advantage of only requiring one iteration whereas the optimized version was specifically targeted to multiple iteration and traded spatial resolution for computational efficiency.

A .NET version of SGA application that allows for the use of the optimized and non-optimised has been produced for PCs running Windows XP or Vista. It will be available for download from the DECC in the middle of 2009. It allows the user to provide their own site and environmental data rather than the set provided on the GBIF site. It also allows the user to use the lookup table method for multiple iterations of the SGA or to use fine grain data and just do a single iteration with no lookup table. When using multiple iterations on the .NET version, a priority raster surface is produced for each iteration. The best site derived each iteration is then set as a new survey site. The SGA is then run for the next iteration. This for N iterations, their will be N priority surfaces created and a table containing the N selected best potential survey sites.

In conclusion, the ‘greedy‘ nature of the Survey Gap Analysis algorithm can not be universally solved for all applications. The choice of methods will be determined by the user’s requirements for spatial resolution (the choice of grid resolution), the predictive accuracy (the number of demand points) and the choice of single or multiple iterations of the SGA (the type of output product). Thus Survey Gap Analysis cannot be considered a ‘one size fits all’ application but one that is contingent on the application user’s research requirements.

Manion G, Ridges M. An optimisation of the survey gap analysis technique to minimise computational complexity and memory resources in order to accommodate fine grain environmental and site data.

REFERENCES

Faith, D.P, Walker, P.A., (1996) Environmental Diversity: on the best-possible use of surrogate data for accessing the relative biodiversity of sets of areas. *Biodiversity and Conservation,* **5:** 399-415.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology* **51:** 331-363.

Ferrier, S., Smith A,P., (1990). Using geographical information systems for biological survey design, analysis and extrapolation. *Australian Biologist* **3:** 105-116.

Fielding, A.H. 2002. What are the appropriate characteristics of an accuracy measure? In Predicting species occurrences. Issues of accuracy and scale. (eds) J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall & F.B. Samson. Washington: Island Press.

Funk, V.A., Richardson, K.S., Ferrier, S., (2005) Survey-gap analysis in expeditionary research: where do we go from here? *Biological Journal of the Linnean Society,* **85:** 549-567.

Lehmann, A., J.R. Leathwick & J.M. Overton. 2002. GRASP: Generalized regression analysis and spatial predictions. Ecological Modelling 157, 189-207.

Ridges, M. 2006. Regional dynamics of hunting and gathering. An Australian case study using archaeological predictive modelling. In GIS and archaeological site location modelling. (eds) K. Westcott & M. Mehre. Boca Raton, Florida: Taylor and Francis.

W. Tobler. 1993. Three Presentations on Geographical Analysis and Modeling. Technical Report TR-93-1, National Center for Geographic Information and Analysis, University of California, Santa Barbara.