Management of Gridded Climate Data for National Scale Integrated Assessment Models

G. C. Sims, R. A. Warrick, G. J. Kenny and W. Ye.
Centre for Environmental and Resource Studies (CEARS)
University of Waikato, PB 3105, Hamilton, New Zealand
g.sims@waikato.ac.nz

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

Abstract

Computer model analyses of climate change impacts are data intensive due to the spatial and temporal dimensions over which climate operates. Data intensity proves a major constraint in the design of such climate models. For policy oriented climate models this constraint proves critical,given the lower specification computer hardware readily available to decision makers.This paper discusses the use of spatial data orderings in combination with run-length encoding to spatially compress climate data. Experiments have been conducted which test the application of various data ordering schemes to the storage of climate data for New Zealand,Australia and Bangladesh.The results of these experiments are presented.

1. Introduction

Under the Framework Convention on Climate Change, signatory parties have an obligation to report to the Conference of the Parties regarding their vulnerability and adaptive capacity to climate change.This places reporting countries in an awkward position; policy makers are advised that the greenhouse effect is real, and probably already occurring, but they often have little quantitative information on the impacts of global warming on which to base their assessments.To make informed decisions,policy makers need tools which enable them to estimate the implications of climate change over a wide range of policy options, and which can provide a concise overview of the uncertainties surrounding global climate change (Hulme *et al.*, 1994; Dowlatabadi and Morgan, 1993). Importantly, this requires the consideration of the spatio-temporal impacts of climate variability and change.

The most efficient way of dealing with climate impacts over time and space is through the use of computer models. However, the development of computer models for climate impact assessment is fraught with difficulty. Due to the spatial and temporal nature of the analyses, such climate impact models usually process data over at least two dimensional space,and thus tend to be data intensive.Typically, information systems store and manipulate one dimensional data. Data,therefore,proves to be a bottle-neck in many climate impact models, requiring significant amounts of storage space and fast computer hardware (notably disk drives and processors). As such, research is necessary to improve the design and implementation of data structures and algorithms for the management of spatially referenced climate data. This paper examines techniques for the storage of spatial climate data.Attention is focused on the use of various data orderings in combination with run-length encoding (RLE) to reduce storage requirements.

2. Integrated Assessments Models — the context

As noted, there is an immediate need for policy decisions on how to prevent or adapt to climate change. For this, information on climate change is fundamental. However, the most scientifically advanced climate models, general circulation models (GCMs), are too computationally demanding for such purposes, generally requiring large amounts of computer processing power and time. Essentially, the complexity of such models makes them more suited to scientific analyses, rather than for direct use in policy or impact analysis which tend to require multiple model runs.Additionally, the spatial resolution of GCMs is often too low to prove of any real benefit for national or local scale policy or impact analysis. Finally, GCMs themselves say little about biophysical and socio-economic impacts or mitigation and adapation options.

To overcome this methodological gap, a new class of *integrated assessment models* (IAM) has evolved (Weyant *et al*., 1996). Such systems combine climate, environmental and socio-economic impact models in order to provide the flexibility to evaluate the effects of climate change and variability. Often, these systems integrate subjective expert judgement about poorly understood parts of the problem, with formal analytical treatment of the well understood parts (Dowlatabadi and Morgan, 1993).These IAMs typically attempt to capture the most salient features of more advanced climate models in a reduced-form, or as results generated off-line and used as model input data.Modularity, inherent in IAMs, ensures the software is readily updated to reflect scientific advances.The most comprehensive and complex versions of IAMs are the highly aggregated global-scale IAMs (eg. IMAGE;Alcamo *et al*., 1994)

At a national scale, simpler integrated models are being developed for New Zealand (CLIMPACTS), Bangladesh (BD-CLIM), and Australia (OzClim) (Kenny *et al*., 1995; Warrick *et al*., 1996).The purpose of these models is to examine the spatial impacts and sensitivities of various sectors (in New Zealand, for example, pastoral, horticultural and arable cropping sectors are examined) to climate variability and change.The models can be viewed as a graphic user interface that provides a structured route through a collection of climate and sectoral impact models. In essence, they operate by coupling a simple global climate model (MAGICC - Model for the Assessment of Greenhouse gas Induced Climate Change, see Osborn and Wigley (1994);Wigley and Raper, (1992);Wigley and Raper, (1993);Wigley, (1993); Hulme *et al*., (1994)) with a regional climate change model to generate scenarios (raster images of climate variables) in five year increments to the year 2100.These images are, in turn, used as input to the sectoral impact models. For a more complete description of this methodology see Kenny *et al*. (1995) and Warrick *et al.* (1996).

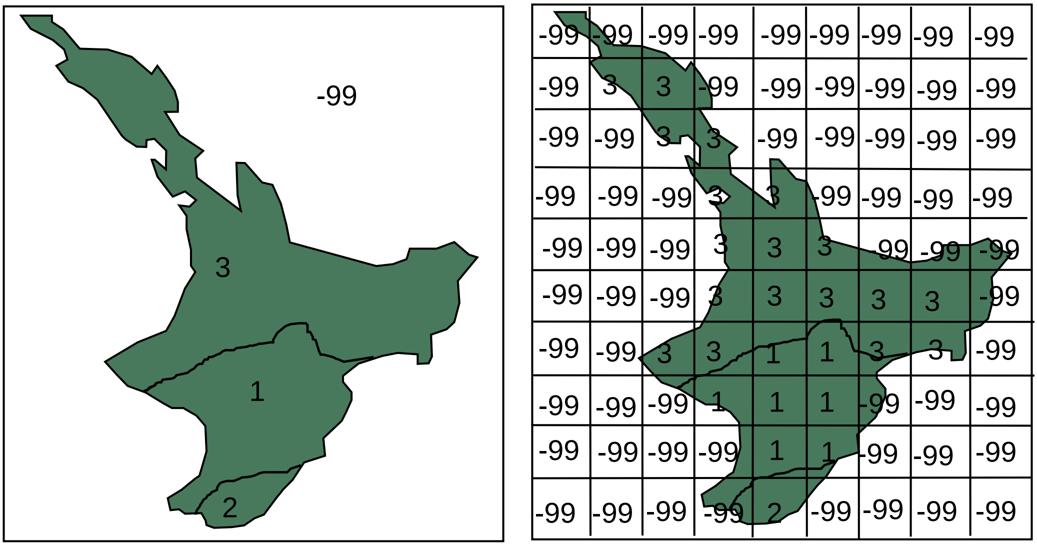
In designing such national-scale integrated models, it is important to consider computational efficiency. A policy-oriented tool should allow multiple experiments to be undertaken quickly, thus allowing sensitivity analysis of various model inputs and assumptions. However, many integrated models are designed to run on desktop computers. The reduced processing power, memory and secondary storage (disk space) of desk top computers is a determinant of the spatial and temporal resolutions at which the software operates, as well as of the scientific complexity of the model.Thus, there is a need to increase computational efficiency. Some techniques, researched in the context of national-scale IAM development, are discussed below.

3. Spatial data-structures

3.1 Run-Length Encoding

Many features which are mapped change gradually over space. If such a feature is mapped in a raster format, there is a probability that neighbouring cells will have the same attribute value.As such, raster maps and images generally have some degree of homogeneity (Bell *et al*., 1988).The degree of homogeneity depends on important factors such as the spatial variability of the feature and the resolution it is being mapped at. Figure 1 illustrates a simple map, and a possible raster representation of this map.Although this is an example, it illustrates that often cells in a raster image have the same value as a neighbouring cell.

If the grid is read from left to right (row order) it is evident that there is repetition of data.Table 1 illustrates the one dimensional row order representation of the above grid within a file. As can be seen, there is considerable redundancy in the file due to repetition of values.This is common in spatial data with some degree of homogeneity. Run-length encoding (RLE) takes advantage of homogene-



-99	-99	-99	-99	-99	-99	-99	-99	-99
-99	3	3	-99	-99	-99	-99	-99	-99
-99	-99	3	3	-99	-99	-99	-99	-99
-99	-99	-99	3	3	-99	-99	-99	-99
-99	-99	-99	3	3	3	-99	-99	-99
-99	-99	-99	3	3	3	3	3	-99
-99	-99	3	3	1	1	3	3	-99
-99	-99	-99	1	1	1	-99	-99	-99
-99	-99	-99	-99	1	1	-99	-99	-99
-99	-99	-99	-99	2	-99	-99	-99	-99

Figure 1, Raster Representation

-99 -99 -99 -99 -99 -99 -99 -99 -99 -99 3 3 -99 -99 -99 -99 -99 -99 -99 -99 3 3 -99 -99 -99.....

Table 1, Row Order File Structure

-99:10 3:2 -99:8 3:2 -99:8 3:2.....

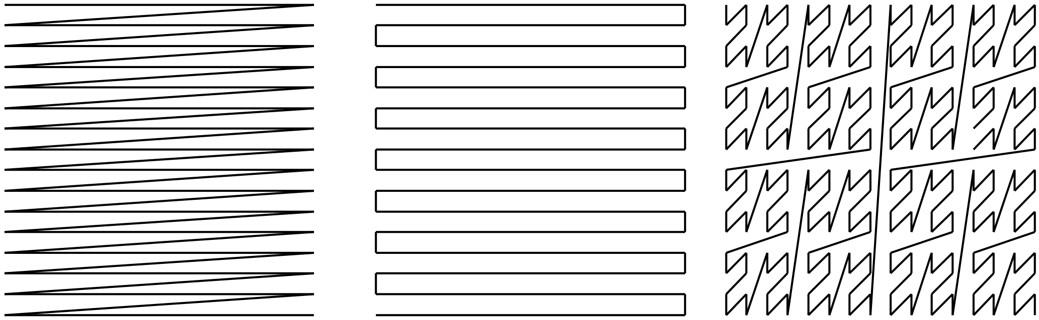
Table 2, Row Order File Structure with Run-Length Encoding

ity in data to reduce the amount of disk space necessary to store the data (Eastman, 1992a; Eastman, 1992b; Holroyd and Bell, 1992; Goodchild and Grandfield, 1983; Abel and Mark, 1990). When writing a raster to file using RLE, repetitions of values are recognised. Instead of writing each individual value to the file (as in the above example), as each repetition of values is encountered, the value is written once along with the number of repetitions of it (the run-length). A record can be eliminated from the file whenever a cell has the same value as the cell previously processed. Table 2 illustrates the row order representation of the above grid using RLE.

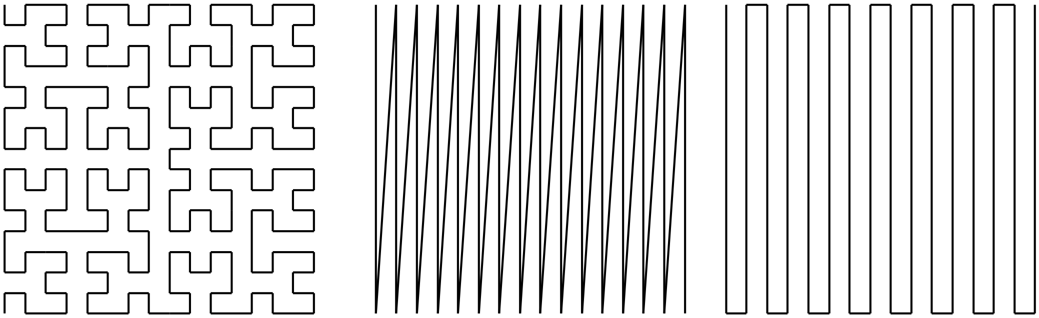
3.2 Data Ordering

A raster image can occupy different amounts of storage depending on how it is structured and ordered. To benefit more from RLE and reduce storage requirements, homogeneity can be increased by using different data orderings. Geographic data are essentially two (or more) dimensional, whereas computer storage and processing are essentially one dimensional (Mark, 1986). No linear (one dimensional) sequence can preserve, and therefore benefit from, all spatial properties of geographic data (Mark, 1986). Using RLE, longer run-lengths will result in less storage requirements. Intuitively then, one would expect orderings which attempt to best preserve spatial relationships to increase run-lengths and reduce storage requirements.

Experimentation with data orderings is often credited to somewhat obscure work carried out by Morton in the mid 1960's for the Canada Geographic Information System (cited in Mark, 1986; Goodchild and Grandfield, 1983; and Lauzon *et al.*, 1985). Morton's order, which was published in an internal report for IBM Canada, allowed cells which are close together in two dimensional space to be placed in similar positions in the linear sequence of the file (see Figure 2c). Further research into data orderings was undertaken by Goodchild and Grandfield (1983) and Abel and Mark (1990). In Goodchild and Grandfield's experiment, four data orders were empirically tested to determine their compression capability. Goodchild and



(a) Row Order (b) Row-prime Order (c) Morton Order



(d) Hilbert Order (e) Column Order (f) Column-prime Order

Figure 2, Spatial data orderings (after Goodchild and Grandfield, 1983)

Grandfield tested row order, row-prime order, Morton order, and Hilbert order (Figures 2a through 2d).

From Figure 2, it would appear that both Hilbert and Morton orders help to preserve the spatial relationships of the two dimensional raster in the translation to a one dimensional sequence, and, as such, longer run-lengths can be expected. In the experiments conducted by Goodchild and Grandfield (1983) boolean images with varying degrees of spatial homogeneity were used to test the compression capability of the various orders. Their results indicated that for images with a high degree of local spatial homogeneity, storage could be reduced by up to 60% using Hilbert order, and 25% using Morton order, as opposed to using row order. For images with little spatial homogeneity their tests resulted in approximately a 5% reduction from Hilbert order and a 5% increase from Morton order, over row order. A comparative analysis, undertaken by Abel and Mark (1990), found row, row prime and Hilbert orderings to be of equivalent performance, and suggested that storage could by reduced by approximately 40% when used in combination with RLE.

4. Application to climate data

Following the work outlined above, experiments were designed to test the relative merits of six different orders for the storage of climate data. The experiments differ from those described above; rather than encoding boolean images, multicoloured images were tested. Raster images corresponding to annual precipitation (total) and annual mean monthly temperature for the North and South Islands of New Zealand, Australia, Queensland, and Bangladesh were used to test the orders. Each record (cell) in the input images contains a four-byte floating point value, recording either its total rainfall or mean annual temperature.

These test images chosen were selected from the databases of IAM currently under development, and therefore the results of the experiment are directly relevant to ongoing research for national-scale IAM development. Summary data pertaining to each the test images can be seen in Table 3.

Dataset	Spatial Resolution	Cells	Background cells
North Island	0.05	22,701	18,085
South Island	0.05	29,141	22,423
Australia	0.2	24,249	12,916
Queensland	0.06	22,185	3,418
Bangladesh	0.05	12,500	7,736

Table 3, The sample images

The record structure for the RLE file consists of a sequence of five byte records; a four byte colour value, followed by one byte recording the run-length. Due to the overhead of storing run-length, it is possible to actually increase the storage requirement for raster images with little or no homogeneity.The decision to allocate one byte to the run-length variable involves a trade-off; images with a low degree of homogeneity will often not reach the upper run-length limit (255),and therefore increasing the size of this variable would increase the size of each record in the file. On the other hand, images with a high degree of homogeneity will often reach the run-length limit, and require an extra record to take the overflow.

The four orders discussed above were tested, as well as two others; column and column prime ordering (Figures 2e and 2f). Column and column prime ordering operate similarly to row and row prime ordering except that the traversal is from top to bottom rather than left to right. Column and column prime orderings were included as it was expected that they should perform best when the climate data has some degree of longitudinal gradient (ie. the values vary less over latitude than they do over longitude).

5. Results

The results from the experiment are illustrated in Table 4. The values in this table represent the percent reduction in the run-length encoded file from its original size.The graphs in Figures 3 and 4 aid interpretation of the results. From these, we can see that the differences in the comparative performance of the orderings are actually very small. Generally, for both climate variables, row and row prime orders provide the best compression, followed by column and column prime, Hilbert, and lastly, Morton order. It is interesting to note that, in most cases, the two dimensional orderings (Morton and Hilbert) are outperformed by the other orders. This is most likely due in part to the fact that the two dimensional orderings are quadrant recursive, and therefore each image needs to be transposed onto a grid with x and y dimensions of 2^n, thus increasing the number of cells that need to be encoded.

The Bangladesh images show the largest range of compression over the six orders.The most effective orderings are row and row prime for temperature, and conversely for precipitation the most effective orderings are column and column prime. This would seem to indicate the flat topography of Bangladesh has less influence effect on the climate and therefore slight latitudinal and longitudinal gradients exist for temperature and rainfall respectively.

Studying the results further, a lack of difference between row and row prime orderings, and column and column prime orderings is apparent. This can be explained by the fact that all the images, except Queensland, have a land mass surrounded by background values (usually a coastal or country boundary), and therefore the colour values in the images are completely surrounded by background values. In effect, this diminishes the differences between the row and column orderings and their prime variants. In the

	Row	Row Prime	Column	Column Prime	Morton	Hilbert
Temperature						
North Island	73.50	73.50	73.47	73.43	71.83	72.34
South Island	70.57	70.57	70.42	70.42	69.35	69.69
Australia	40.95	40.90	40.60	40.60	39.04	39.78
Queensland	-2.67	-2.23	-2.65	-2.47	-4.88	-3.57
Bangladesh	65.40	65.39	59.26	59.26	55.46	62.47
Precipitation						
North Island	73.50	73.50	73.44	73.40	71.80	72.33
South Island	70.51	70.51	70.32	70.32	69.24	69.58
Australia	41.76	41.71	41.38	41.38	39.70	40.62
Queensland	-3.18	-2.67	0.88	1.05	-3.43	-2.27
Bangladesh	51.16	51.15	58.71	58.71	53.61	54.87

Table 4, Percentage compression of individual images

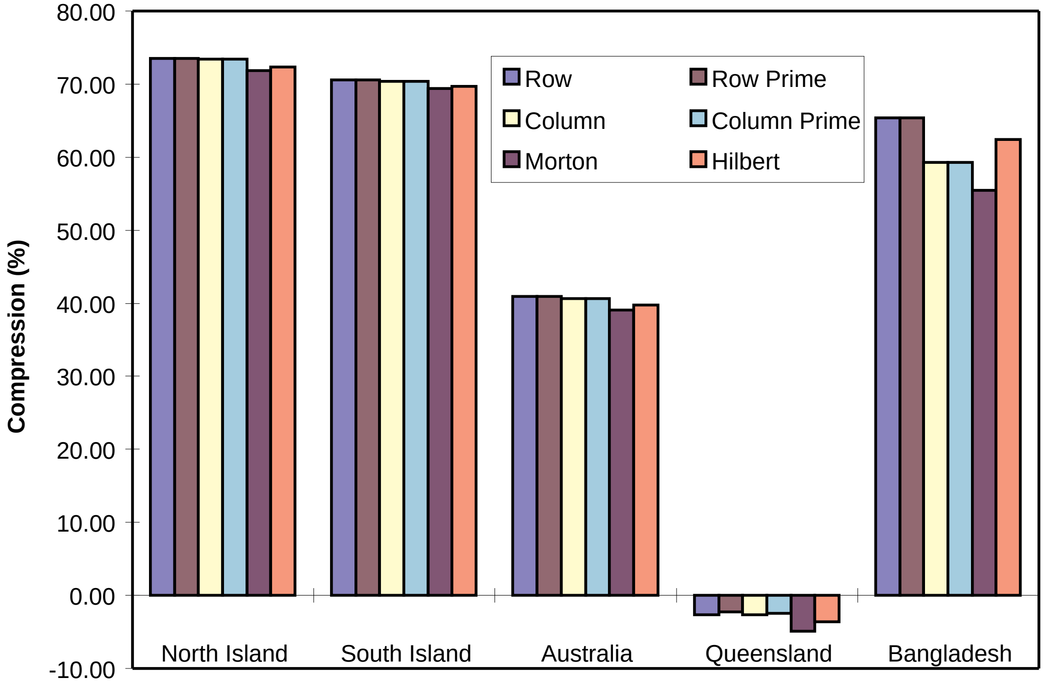


Figure 3, Percentage compression of temperature images

case of the Queensland images, the differences are more pronounced, and for three of the four cases the row and column orderings are less effective than the prime orderings.

In all cases the Queensland images fail to compress to a size smaller than the original file. This could be due to particularly low homogeneity of the colour values for the particular images. However, it is more likely an indication that, in terms of reduction of actual colour values, little is to be gained through any form of data ordering and run-length encoding. The graph in Figure 5 illustrates the relationship between compression (based on the average compression of all the orders) of a given raster and the number

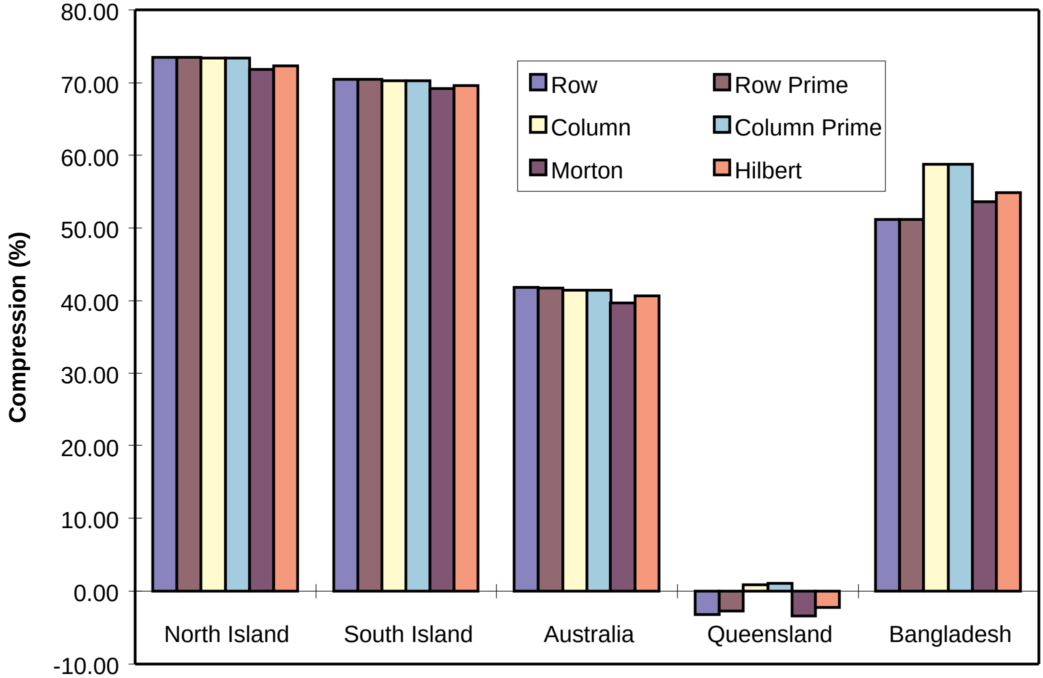


Figure 4, Percentage Compression of Precipitation images

of background cells. As would be expected, a relationship exists, as background cells are often contiguous and of the same value, therefore large run-lengths should result. However, the relationship is very strong, and in effect, the lack of variation from the fitted line indicates that for the experimental images there is very little spatial homogeneity (and therefore compression) of the coloured cells.

6. Discussion

Climate is driven by solar energy. It is well known that the amount of solar energy received varies latitudinally; the closer one is positioned towards the poles the less energy is received. From this, we could perhaps expect a low degree of latitudinal homogeneity with climate data, and therefore a row order traversal (longitudinal) would result in longer run-lengths. This could perhaps explain why the row and row prime orders marginally outperformed the other tested orders. However, in reality, the problem is not that simple. Climate is highly variable over both latitudinal and longitudinal dimensions, due to the influence of factors such as topography, orography, continentality, and oceans, and this variability undoubtedly influences the poor run-lengths of colour values. Additionally, factors such as spatial resolution of the encoded images will often affect homogeneity — a high resolution image will have a greater degree of spatial auto-correlation than the same image gridded to a coarser resolution.

Another possible contributor to the poor run-lengths for the coloured values lies in the nature of the tested data. Most images of climate variables are interpolated using mathematical procedures from meteorological station weather records. The nature of some of these interpolation algorithms tends to produce images which vary over space between the original site data, sometimes in an unrealistic manner, and fail to adequately represent regional or local scale climates. This is most evident in interpolation algorithms which treat the climate parameter as an independent variable (such as inverse distance weighted algorithms). More advanced techniques, such as co-kriging (Bogaert *et al.*, 1995) or partial thin-plate smoothing splines (Hutchinson, 1995) include the influence of variables, such as elevation, in the interpolation. Often, this type of ap-

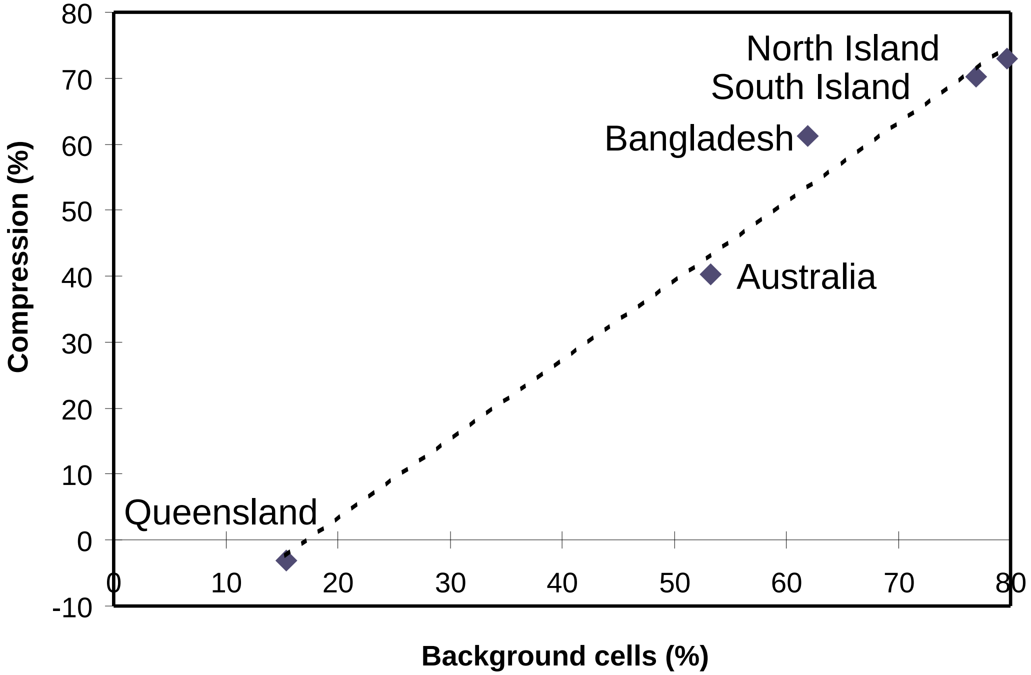


Figure 5, Relationship between compression and number of background cells

proach better captures the spatial variability of climate, and one could expect a higher degree of spatial auto-correlation in the interpolated image. Further to this, interpolated images are only as good as the quality of the point data they are interpolated from. Low density station networks and erroneous site records produce poor quality images with more spatial variability.

7. Conclusions

Overall, the results suggest that, due to the factors outlined in the discussion, it cannot be assumed that any particular data ordering scheme will perform better than any other, or indeed result in any compression at all, with gridded climate data. As such, it would appear that it is not possible to develop a generic algorithm based on one particular data ordering. However, this does not necessarily mean that RLE is an unsuitable technique for application to IAMs. In almost all cases, the size of the original files were substantially reduced through the effective compression of the background values. If it is important that geographic referencing is retained implicitly in the images used by the climate model (for example, if the climate model makes use of images with differing resolutions, projections, or geographic windows), and the images contain a reasonable number of background values, then the use of data orderings and RLE are worthy of consideration.

Acknowledgements

This research was funded by the New Zealand Foundation for Research, Science and Technology (contract C01612).

References

Abel, D.J., Mark, D.M. (1990). A Comparative Analysis of Some Two-Dimensional Orderings. *International Journal of Geographical Information Systems*. 4 (1) pp 21-31

Alcamo, J. (ed.) (1994). *IMAGE 2.0: Integrated Modelling of Global Climate Change*. Kulwer Academic Publishers, Dordrecht.

Bell, S.B.M., Diaz, B.M., Holroyd., F.C. (1988). Capturing Image Syntax Using Tesseral Addressing and Arithmetic. In: Muller J.P. (ed). *Digital Image Processing in Remote Sensing*. Taylor and Francis, London. pp 135-152

Bogaert, P., Mahau, P., Beckers, F. (1995). *The Spatial Interpolation of Agro-Climatic Data: Cokriging Software and Source Code User's Manual*. Agrometeorology Serries Working Paper Number 12. Food and Agriculture Organisation, Rome.

Dowlatabadi, H., Morgan, M.G. (1993). Integrated Assessment of Climate Change. *Science*. (259) pp 1813-1932

Eastman, J.R. (1992a). *IDRISI Users Guide*. Graduate School of Geography, Clark University. Worchester, Massachusetts.

Eastman, J.R. (1992b). *IDRISI Technical Reference*. Graduate School of Geography, Clark University. Worchester, Massachusetts.

Goodchild, M.F., Grandfield, A.W. (1983). Optimising Raster Storage: An Examination of Four Alternatives. *Proceedings, Auto Carto 6*, pp 400-407

Holroyd, F., Bell, S.B.M. (1992). Raster GIS: Models of Raster Encoding. *Computers and Geosciences*. 18 (4) pp 419-426

Hulme, M., Raper, S.C.B., Wigley, T.M.L. (1994). An Integrated Framework to Address Climate Change (ESCAPE) and Further Developments of the Global and Regional Climate Modules (MAGICC). *Energy Policy*. 23 (4/5) pp 347-355

Hutchinson, M.F. (1995). Interpolating Mean Rainfall Using Thin Plate Smoothng Splines. *International Journal of Geographical Information Systems*. 9 (4) pp 385-403

Kenny, G.J., Warrick, R.A., Mitchell, N.D., Mullan, A.B., Salinger, M.J. (1995). CLIMPACTS: An Integrated Model for Assessment of the Effects of Climate Change on the New Zealand Environment. *Journal of Biogeography*. 22 (4/5) pp 883-895

Lauzon, J. P., Mark, D. M., Kikuchi, L., Guevara, J. A. (1985). Two-Dimensional Run-Encoding for Quadtree Representation. *Computer Vision, Graphics, and Image Processing*. (30) pp 56-69

Mark, D.M., Lauzon, J.P. (1984). Linear Quadtrees for Geographic Information Systems. Proceedings, *International Synopsium on Spatial Data Handling*. Vol. 11, Zurich, Switzerland. pp 412-430

Mark, D.M. (1986). The use of Quadtrees in Geographic Information Systems and Spatial Data Handling. *Proceedings, Auto Carto 8*, pp 517-527

Osborn, T.J. and Wigley, T.M.L. (1994). A simple model for estimating methane concentration and lifetime variations. *Climate Dynamics*. (9) pp 181-193

Warrick, R. W., Kenny, G. J., Sims, G.C., Ericksen, N.E., Ahmad, Q.K., Mirza, M.Q. (1996). Integrated Model Systems for National Assessments of the Effects of Climate Change: Applications in New Zealand and Bangladesh. *Journal of Water, Air, and Soil Pollution*. 92 (1/2) pp 215-227

Weyant, J., Davidson, O., Dowlatabadi, H., Edmonds, J., Grubb, M., Parson, E.A., Richels, R., Rotmans, J., Shulka, P.R., Tol, R.S.J., Cline, W., Fankhauser, S. (1996). Integrated Assessment of Climate Change: An Overview and Comparison of Approached and Results. In: Brunc, J.P., Lee, H., Haites, E.F. (eds). Climate Change 1995: Economic and Social Dimensions of Climate Change. Cambridge University Press, Cambridge. pp 364-396

Wigley, T.M.L. and Raper, S.C.B. (1992). Implications for climate and sea level of revised IPCC emissions scenarios. *Nature*. (357) pp 293-300

Wigley, T.M.L. (1993). Balancing the carbon budget. Implications for projections of future carbon dioxide concentration changes. *Tellus*. (45B) pp 409-425

Wigley, T.M.L. and Raper, S.C.B. (1993). Future changes in global-mean temperature and sea level. In: Warrick, R.A., Barrow, E.M. and Wigley, T.M.L. (eds). *Climate and Sea Level Change: Observations, Projections and Implications*. Cambridge University Press, Cambridge. pp 111-133