Proceedings of the *Open source GIS - GRASS users conference 2002* - Trento, Italy, 11-13 September 2002

Advantages of open-source GIS to improve spatial environmental modelling

Scott Mitchell*, Ferko Csillag*, Christina Tague**

* Department of Geography, University of Toronto at Mississauga, 3359 Mississauga Road, Mississauga, Ontario L5L 1C6 Canada, tel. ++19058283862, fax ++19058285273, e-mail smitch@geog.utoronto.ca / fcs@geog.utoronto.ca
** Department of Geography, San Diego State University, San Diego, California, 92182-4493 USA, tel. ++16195943230 fax ++16195944938 e-mail ctague@mail.sdsu.edu

1 Introduction

Spatial heterogeneity is a source of complexity in environmental models, but can also be an important part of landscape dynamics and our understanding of environmental processes. The majority of our knowledge of environmental processes to date has been acquired from relatively small plot-based experiments. Recently there has been a growing recognition that in order to apply this knowledge to regional management issues, we need to understand how the processes scale and interact across landscapes of varying sizes and complexities. GIS can be important in this research, qualified by the caution that many programs constrain landscape representation to a particular view, whereas complex environmental models require flexible combinations of representations.

Since direct environmental measurements are usually difficult or impossible to make at the temporal and spatial scales dealt with in landscape management, modelling has become an important tool. It is not obvious, however, how different environmental functions scale from local plots to regional applications and beyond, and significant bias can be introduced by naively applying relationships derived from experimental plots to larger areas [1]. The amount and character of this bias depends on the nature of the relationship between driving variables and the process being predicted, and the heterogeneity of these variables across the landscape. Improvements in computing capacity notwithstanding, it is impossible to have a believable representation for every square metre of a management region. Further, a focus on many small local variations quickly becomes intractable from a management perspective and is often not relevant to the problems being addressed. Aggregation is, therefore, necessary and desirable. The balance of important local differences versus noise is key to developing an appropriate model. Therefore in spatial problems, the treatment of heterogeneity is critical in the design of environmental models, and in the management and estimation of prediction uncertainty in the model application.

These issues make GIS a natural tool for environmental modelling since it allows a flexible and generic base representation of space (e.g. a raster map), on top of which different conceptual representations of space can be explored. Open software such as GRASS allows us to quickly modify code to experiment with different representations, as well as to invent new tools without having to develop all of the underpinnings required for basic spatial processing (e.g. functions provided by the GRASS libraries).

We have been using a combination of existing and new programs to handle issues of spatial heterogeneity for the purposes of environmental modelling, including translating between GIS data constructs and the statistics used to measure spatial heterogeneity, the preparation of model data structures based on these statistics and data stored in GIS data layers, and the mapping of model predictions. In this paper, we present some examples of these tools.

2 Methods for representing heterogeneity within a GIS

There is a wide spectrum of spatial data analysis tools to create 'appropriate' database representations for certain environmental models. Figure 1 describes our view of how GIS tools are used to help move between fine and coarse levels of aggregation in the preparation of model representations of the environment. Moving from a fine resolution to more aggregated units can be accomplished with a bottom-up approach such as watershed delineation, or by a top-down approach such as quadtree partitioning. Sampling can characterize changes in variance as a function of aggregation at any location.

To explore these different aspects of the effects of data aggregation for the purposes of characterizing heterogeneity in environmental modelling, we have used a variety of tools to translate between distributed GIS data and environmental data models. GRASS provides both standard GIS tools for managing databases to support modelling projects, and an environment for programming new tools[1].

[1] For purposes of brevity, details of program development status are omitted from this paper. The modules discussed range from undocumented code that can access GRASS data, through GRASS programs that could be considered beta quality but with little documentation, to one that has been relatively well tested in GRASS 5, but now there is another very similar program in the default GRASS distribution. In October 2002, an information web page will be created about our software, taking information gathered at the GRASS 2002 conference into account. The address will be: `http://eos.geog.utoronto.ca/~smitch/grassprogs`

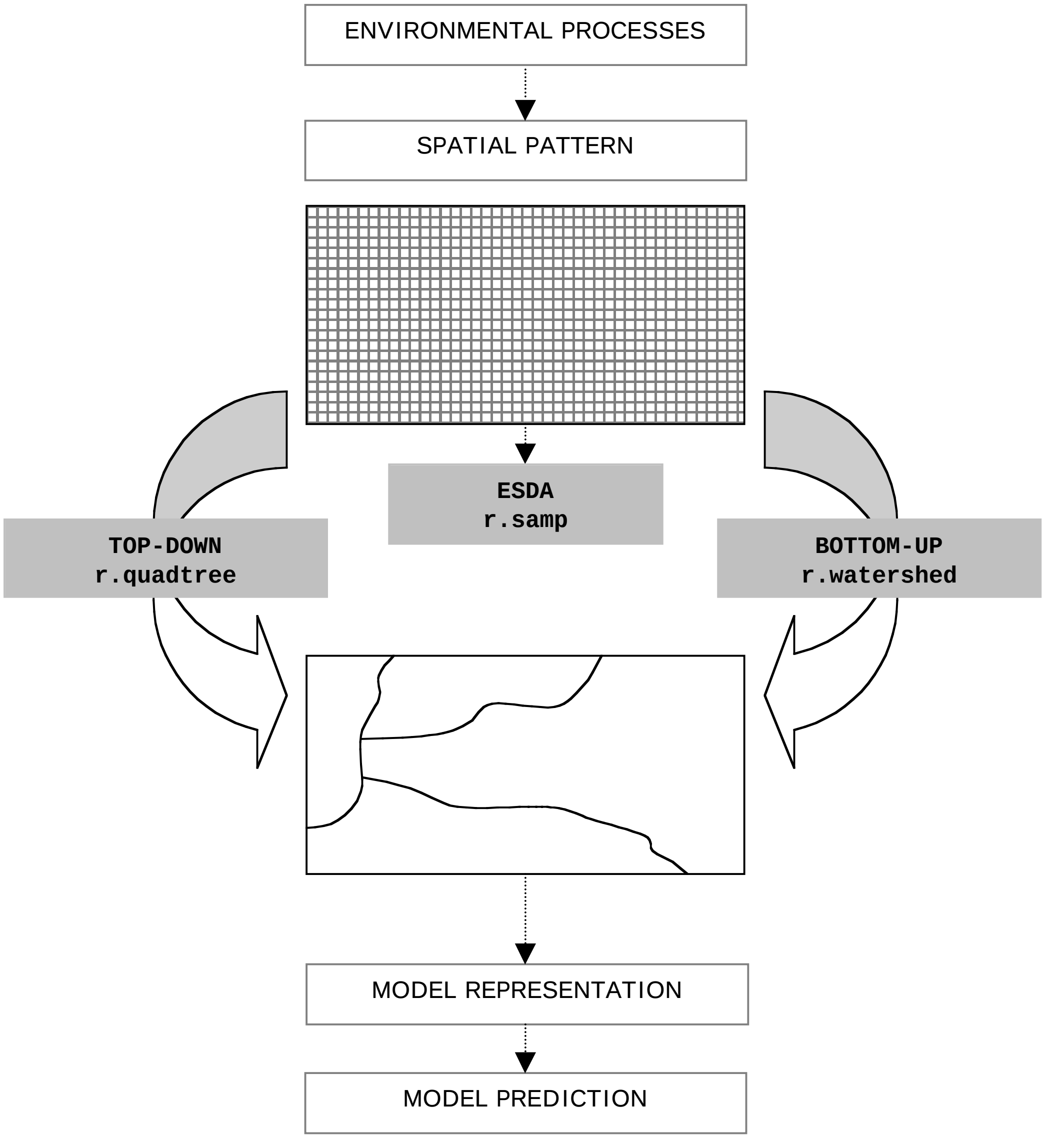

Figure 1 - Framework describing the relationships and tools used to move between patterns and processes in the environment and in models**.**

a) Sampling

Many spatial statistics are based on point sampling across a landscape. There are several tools in GRASS which can be combined to do exploratory data analysis (ESDA) using sampled points on raster layers, but we have developed our own ESDA tool `r.samp` in order to represent specific spatial statistical expertise. The basic operation of the program is summarized in Figure 2, and uses a combination of the following sets of user-controlled options:

- the number of samples,
- the spatial distribution of the sites,
- the sampling resolution,
- the number of layers to be sampled, and
- the types of statistics to be computed.

The number of samples can be expressed as either an absolute count, or a percentage of all cells in a raster layer. The spatial distribution can be random, regular, hierarchical distance- or layer-controlled, or user-supplied. The random option uses `r.random`. The regular option sets a random, or optionally user-supplied starting point and then calculates where samples should be located to cover a raster map at a user-supplied interval. The hierarchical distance-controlled samples for nested ANOVA[2, 3], requiring the number of levels, the number of samples for each level and the required distances between levels. The hierarchical-layer-controlled option implements stratified random sampling according to categories supplied in a separate layer. Finally, there is an option for a user to supply a GRASS site-file for the sampling locations.

The sampling resolution is varied by sampling from the single sites, or by aggregating across a local neighbourhood, defined by a circular or square window of a specified radius. A unique feature of this sampler is that it can sample multiple layers simultaneously. If more than one layer is specified, the samples will be aggregated across all layers – this is particularly useful for environmental simulation time series. Finally, the types of statistics calculated can be selected from mean, mode, minimum, maxumum, standard deviation, variance, number of unique values, and number of values different from the centre. In the case of multiple layers, correlation or covariance can also be computed. The results are presented in a text file which we read directly in to the R statistics package[4] for further processing.

This `r.samp` module can be used, for example, to assess various aggregation effects on spatial variability (e.g., increased smoothing with coarser sampling resolution). This information is useful as a preparatory step before partitioning or in testing the appropriateness of an aggregation (e.g., the homogeneity of smoothing with all paritioning units having less than a prescribed threshold). Spatial statistics generated by this tool can also be used directly by stochastic environmental models (e.g. TOPMODEL[5]) that are designed to simulate across distributions of input parameters.

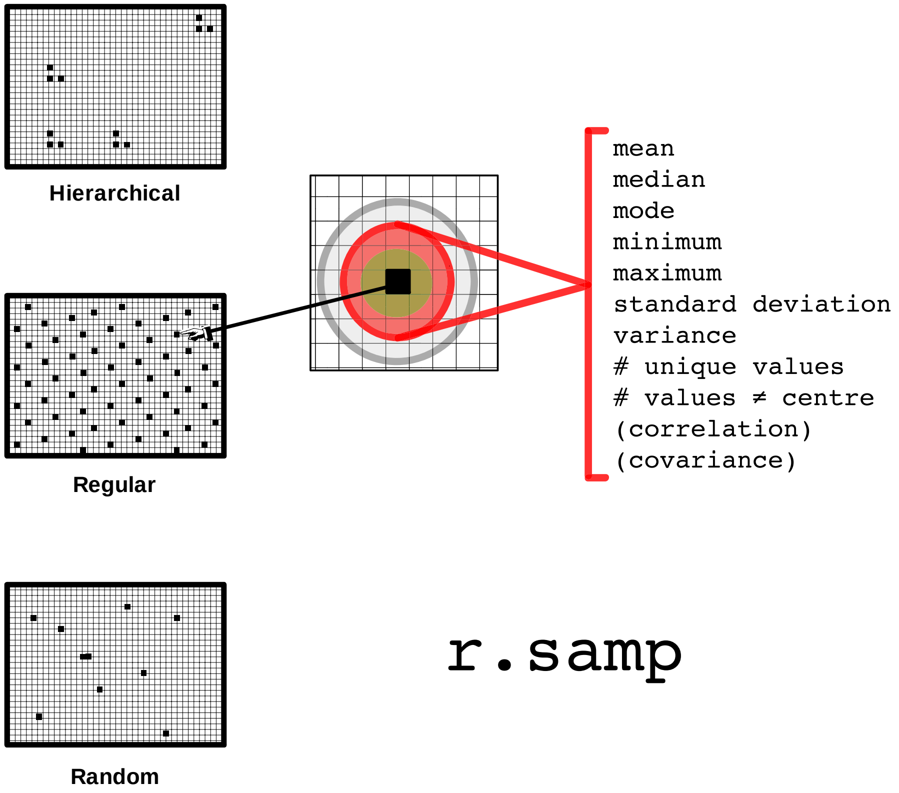

Figure 2 – Schematic of `r.samp`, a GRASS module for exploratory spatial data analysis using point sampling.

b) Partitioning

Partitioning a landscape according to local heterogeneity is a statistically sound method to manage or minimize prediction uncertainty when scaling relationships in environmental models[1, 6]. The entire range of possible realizations of a partitioned landscape are typically unmanageable (i.e. as many as $M^{(N-1)}$-1 where N=number of original units (pixels) and M=number of partitioning units)[7], but some basis for choosing an appropriate method is far preferred over arbitrary choices [6, 8-12]. Just as in developing model algorithms, one needs to find the right balance between enough detail to capture the important processes that make critical differences to the overall behaviour, but not include excessive detail to the point of making the problem intractable, or too complex to extract answers to the questions at hand. Increasing the detail of a representation may better capture data for which we have fine resolution sources, but this is severely constrained by the requirement to know a large number of other model parameters for each model unit.

We have worked primarily with “hydroecological” models, those dealing with some combination of hydrologic and vegetation growth processes, and the associated processes of carbon and nitrogen allocation, transport, and sequestration. Numerous examples have shown that there are often key locations in landscapes which although small, have important local conditions (e.g. pockets of saturated soil) which have major impacts on the aggregate behaviour of the larger region (e.g. areas of high primary productivity or nitrogen transformation). If the input conditions are averaged out by excessive aggregation, the overall predicted response of the aggregate unit can under or over estimate associated behaviour [9, 10, 13-18].

We have performed partitioning experiments using GRASS to develop a range of different input partitioning methods for use with various environmental models. These partitioning methods include regular grids, watersheds, quadtrees, and site surveys (Figure 3).

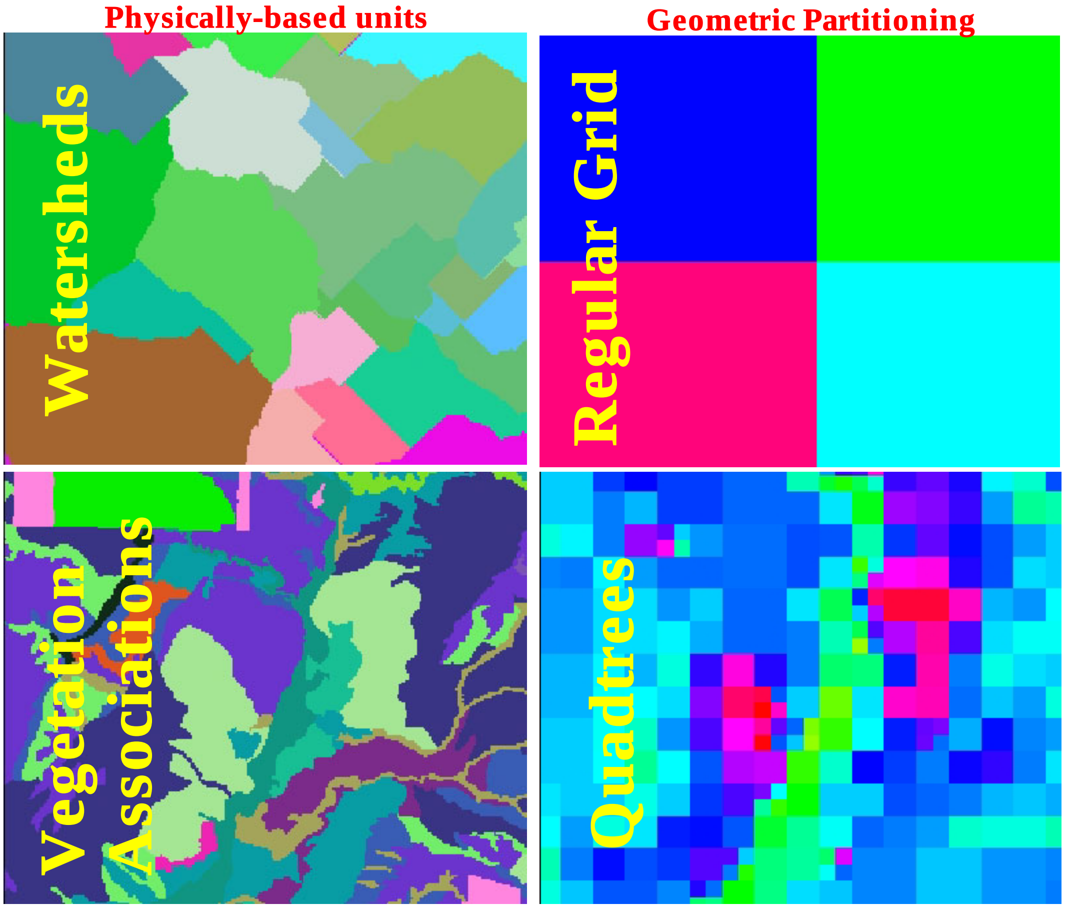

Figure 3 – Examples of spatial partitioning for environmental modelling

Regular grids are often used for distributing a model across space because of the ease of implementation, and indeed it is trivial to create new grids in GRASS with arbitrary origins and resolutions. Watersheds are an example of a process based method of partitioning, where apriori knowledge of the relevant system dynamics (i.e. the flow of water to follow topography) is used to define appropriate units. Watersheds of various sizes can be built in GRASS using `r.watershed`. Quadtrees are similar to a regular grid in that they use rectangular elements to represent a surface, but have a variable resolution, so that more data “resources” (resolution and therefore parameterisation and computing effort) can be allocated where they are needed the most [19] (areas of high variability lead to to finer units). Quadtrees can be built using our own program `r.quadtree`, which partitions an image into a selectable number of units such that smaller units are used in areas of high variability[20].

The quadtree algorithm starts with a fine resolution version of a dataset, and builds a pyramid of raster resolutions from the starting image to the most coarse possibility, using 4-to-1 aggregation at each step, up to a single pixel representation of the area. Each level of the pyramid stores the mean of the four pixels (called "children") that fill the same space in the next finer layer, as well as a measure of the variance of the children's values. Thus, at the coarsest level we store the scene global mean, and at the finest level we have the original dataset.

This pyramid is then used to build possible representations using either a maximum number of units to characterize the data, or a maximum residual variance. An iterative steepest descent algorithm starts at the coarse end of the pyramid, searches for the pixel (or "leaf", in quadtree terminology) in the underlying (finer) layer which aggregates the greatest amount of variability in the data – this leaf is chosen to be split into its four child leaves, and the process is repeated until the threshold conditions are met, building the quadtree representation.

To analyse the effect of each partitioning possibility on the modelling effort, it is useful to be able to characterize the aggregation of the input data, which involves describing the distribution of values of a raster surface under a polygon map. We have built a program called `r.polystats` to do this, which collects statistics to describe the distribution of raster values under a polygon coverage, writing the results to a text file and / or a new raster map[2].

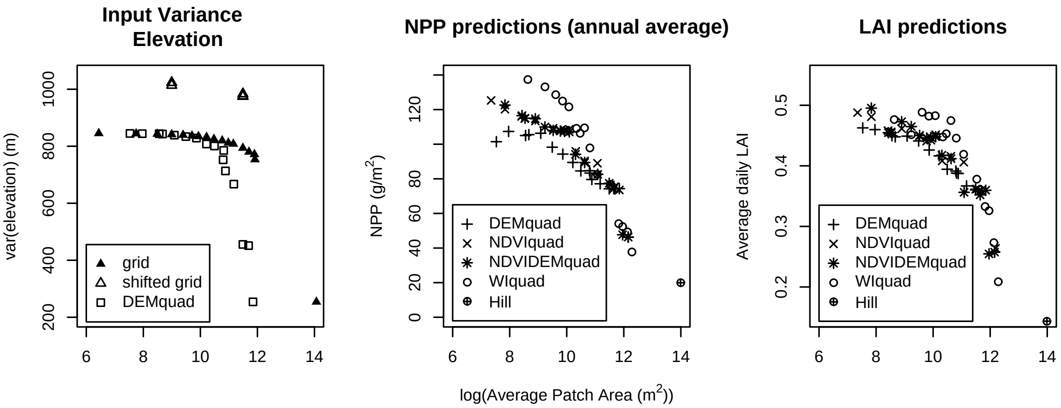

Figure 4 – Examples of plots used to analyse landscape partitions in terms of input aggregation and model response, using GRASS and R. The left-most figure shows **within**-unit variability of elevation under regular grids (with a common origin as the quadtree, plus small shifts) and a quadtree built based on elevation. Quadtrees were also built based on other inputs, or combinations of inputs; the plots of predicted net primary productivity (NPP) and leaf area index (LAI) show the average prediction across all units under quadtrees built based on elevation (DEMquad), Normalized Difference Vegetation Index (NDVIquad), wetness index [5] (WIquad), and a combination of NDVI and elevation (NDVIDEMquad). "Hill" represents a maximum aggregation.

[2] Recent versions of GRASS have included a very similar program, r.statistics; our program has a few differences, such as the ability to simultaneously work with multiple cover layers, which have convinced us to keep working on it.

The aggregation of input data can be used to understand differences in predictions, and to help select appropriate partitions. For example, in studies of grassland productivity comparing regular grid and quadtree representations, it was found that the variance-based strategy of quadtrees allowed us to preserve inter-patch (and minimize within-patch) variability of key inputs such as elevation, and therefore predict important pockets of variable primary productivity with a lower number of modelling units and less uncertainty (Figure 4) [9]. These tools are useful after a partitioning scheme is chosen, as well - evaluating the uncertainty in predictions requires keeping track of the residual variance (after partitioning or aggregation of a fine resolution data set), when analytical methods for computing prediction error estimates (e.g. Taylor-series expansion [1, 21]) are not feasible.

c) Preparing / building model constructs

Once data layers have been prepared for use in an environmental model, there is usually an intermediate data processing step to properly format the data for the model. The complexity of this step depends on the design of the model. RHESSys (Regional Hydro-Ecological Simulation System) [22] is a model we often use for simulating dynamics in forested and grassland ecosystems. It uses a hierarchical spatial framework which divides up the world into basins, hillslopes, zones, and patches, which correspond to the scales at which the relevant processes are thought to operate (Figure 5).

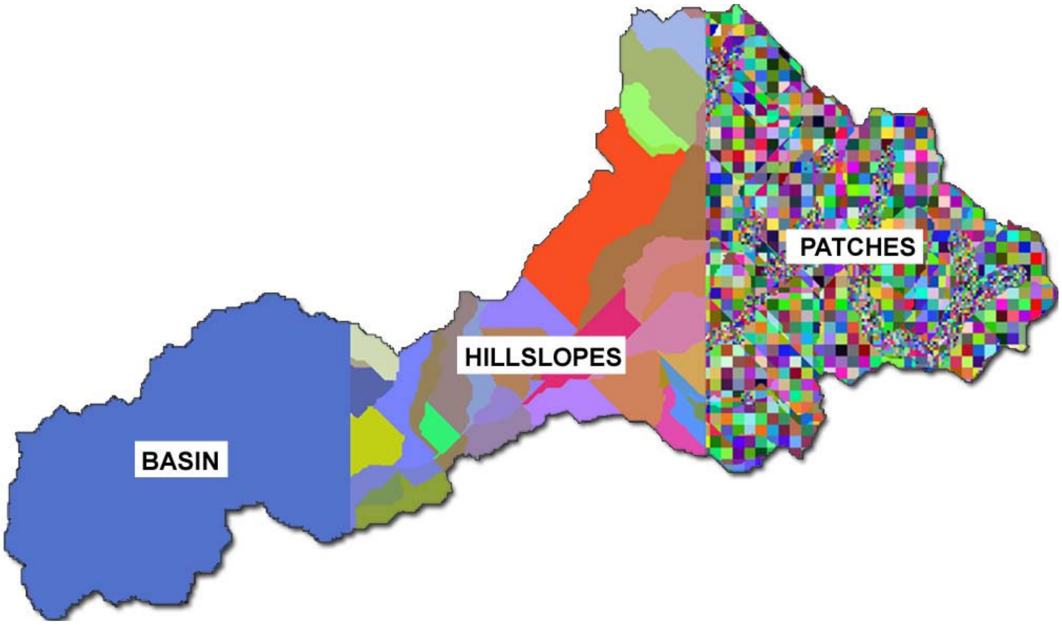

Figure 5 – The RHESSys "world" is a nested division into basins, hillslopes, and patches.

This flexible spatial model allows the integration of numerous spatial data layers, which creates a need for a spatial database manager such as the GRASS GIS. Further, each spatial data layer is an instance of aggregation and thus can take advantage of the various techniques described above. For example, since hillslopes are designed to organize drainage patterns, these spatial units are typically defined using the processed based partitioning strategy, `r.watershed`. Patches on the other hand correspond to units of similar soil moisture and biogeochemical cycling characteristics. Consequently, approaches such as quadtrees, that focus on variance, can be used to generate the patch spatial layers based on variance in multiple input layers (i.e. soil, land-use and topographic maps).

Although approaches like quadtrees tend to minimize the number of patches , the use of RHESSys in large, highly complex landscape often results in a large number of patches and can produce an unacceptably long computation time. An alternative approach being considered is to move to a sampling based stochastic approach (rather than an explicit partitioning).

Developing the data layers for a model such as RHESSys also requires fairly sophisticated data management. The modelled “world” is initialised using a series of text files holding default values for various parameters that do not vary across space or time for a given ecosystem type, and a large “worldfile” which describes the layout and state of every spatial unit and its dynamic state variables. Since there is a large number of state variables in the model, creating this worldfile manually for anything but a simplistic test case would be an overly onerous task. A program called “grass2world” was developed, allowing the user to create a template file to select GIS layers which define each level of the spatial hierarchy, as well as default starting values for many variables, or formulae to calculate them based on statistical properties of GIS input layers as appropriate (`r.samp` can assist the design process). This tool was adapted from another GRASS / environmental modelling programming project called Macaque (F. Watson, unpublished); the sharing of source code has allowed additional benefit out of the original programming effort, and the common GRASS code facilitated portability between projects.

3 Conclusions

The open nature of GRASS and associated projects has allowed rapid, collaborative development of the tools described above, using the accumulated expertise in the GRASS library and existing modules to reduce the need to develop spatial processing algorithms. The flexible framework and collaborative culture of open source programs like GRASS are both important characteristics in research oriented projects, where the application in question is constantly evolving, therefore the data preparation tools must frequently be altered in turn. The command-line orientation of GRASS also allows valuable batch execution possibilities, enabling the automated development of hundreds of possible spatial definitions or modelling scenarios. The combination of these features has allowed us to design a GRASS-GIS based system to explore a range of options including statistical sampling and spatial partitioning for dealing with heterogeneity in complex environmental models such as RHESSys.

References

[1] Rastetter, E.B., A.W. King, B.J. Cosby, G.M. Hornberger, R.V. O'Neill, and J.E. Hobbie, Aggregating fine-scale ecological knowledge to model coarser-scale attributes of ecosystems. *Ecological Applications*. **2**(1): p. 55-70, 1992.

[2] Webster, R. and M.A. Oliver, *Statistical methods in soil and land resource survey*. Oxford: Oxford Press. 316, 1990.

[3] Davidson, A. and F. Csillag, A comparison of nested ANOVA and geostatistics for characterizing the spatial structure of patchy grassland under a limited sampling budget: a study using simulated landscapes. *Canadian J of Remote Sensing*. **In Press**, 2002.

[4] Ihaka, R. and R. Gentleman, R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. **5**(3): p. 299-314, 1996.

[5] Beven, K.J. and M.J. Kirkby, A physically based, variable contributing area model of basin hydrology. *Hyd Sci - Bull*. **24**(1): p. 43-69, 1979.

[6] Handcock, R.N., S.W. Mitchell, and F. Csillag, Monte-carlo Sensitivity Analysis of Spatial Partitioning Schemes: Regional Predictions of Nitrogen-Loss, in K. Lowell and A. Jaton, editors, *Spatial accuracy assessment: land information uncertainty in natural resources*. Ann Arbor Press: Chelsea, MI. p. 247-254, 1999.

[7] Csillag, F. and S. Kabos. How many regions ? Toward a definition of regionalization efficiency. in Proceedings of *Auto-Carto 13*. Seattle, 1997.

[8] Mitchell, S.W. and F. Csillag. Does Pattern Matter ? Handling bias, uncertainty and stability of predicted vegetation growth in Grasslands National Park, Saskatchewan. in Proceedings of *4th International Conference on Integrating GIS and Environmental Modelling (GIS/EM4): Problems, Prospects and Research Needs*. Banff, Alberta, Canada, 2000.

[9] Mitchell, S.W., Evaluating the impacts of scale and variability of inputs in predictions of grassland productivity dynamics, 2002 (In Preparation).

[10] Lammers, R.B., Extending hydro-ecological simulation models from local to regional scales Ph.D. thesis, Department of Geography, University of Toronto, Toronto. 163 pp,1998.

[11] Lammers, R.B., L.E. Band, and F. Csillag, Partitioning landscapes for hydro-ecological modeling, Toronto, 1996 (unpublished manuscript).

[12] White, J.D. and S.W. Running, Testing scale dependent assumptions in regional ecosystem simulations. *Journal of Vegetation Science*. **5**: p. 687-702, 1994.

[13] Pierce, L.L. and S.W. Running, The effects of aggregating sub-grid land surface variation on large-scale estimates of net primary production. *Landscape Ecology*. **10**(4): p. 239-253, 1995.

[14] Tague, C. and L. Band, Simulating the impact of road construction and forest harvesting on hydrologic response. *Earth Surface Processes and Landforms*. **26**: p. 135-151, 2001.

[15] Tague, C.L. and L.E. Band, Evaluating explicit and implicit routing for watershed hydro-ecological models of forest hydrology at the small catchment scale. *Hydrological Processes*. **15**(8): p. 1415-1439, 2001.

[16] Mitchell, S.W. and F. Csillag, Assessing the stability and uncertainty of predicted vegetation growth under climatic variability: northern mixed grass prairie. *Ecological Modelling*. **139**(2-3): p. 101 - 121, 2001.

[17] Mackay, D.S. and L.E. Band, Forest ecosystem processes at the watershed scale: Dynamic coupling of distributed hydrology and canopy growth. *Hydrological Processes*. **11**(9): p. 1197-1217, 1997.

[18] Creed, I.F., L.E. Band, N.W. Foster, I.K. Morrison, J.A. Nicolson, R.S. Semkin, and D.S. Jeffries, Regulation of nitrate-N release from temperate forests: A test of the N flushing hypothesis. *Water Resources Research*. **32**(11): p. 3337-3354, 1996.

[19] Csillag, F., Variations on hierarchies: toward linking and integrating structures, in M.F. Goodchild, *et al.*, editors, *GIS and Environmental Modeling: progress and research issues*. GIS World Books. p. 433-437, 1996.

[20] Kertész, M., F. Csillag, and Á. Kummert, Optimal tiling of heterogeneous images. *International Journal of Remote Sensing*. **16**(8): p. 1397-1415, 1995.

[21] Heuvelink, G.B.M., P.A. Burrough, and A. Stein, Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*. **3**(4): p. 303-322, 1989.

[22] Band, L.E., C.L. Tague, S.E. Brun, D.E. Tenenbaum, and R.A. Fernandes, Modelling watersheds as spatial object hierarchies: structure and dynamics. *Transactions in GIS*. **4**(3): p. 181-196, 2000.