Open source approaches to integration of geo-spatial data at the water catchment scale

Peter Metcalfe, Keith Beven

Lancaster University Environment Centre,
Bailrigg, Lancaster, LA1 4YQ, UK
p.metcalfe@lancs.ac.uk

Introduction

Vast quantities of topographic, land cover and other geo-spatial data have become available in recent decades, along with numerous GIS systems and spatial databases for their storage and analysis.

The emerging research framework of Integrated Catchment Management (ICM), the water quality and ecologically-driven European Water Framework Directive (WFD) and corresponding need for environmental "models of everywhere" (Beven and Alcock 2007) present an unprecedented opportunity to apply these resources.

Significant challenges arise in the management of the large, heterogeneous geo-spatial data sets, however. Conventional approaches using of GIS to integrate and handle such data may discourage users without access to or unfamiliar with these tools. This has the potential to inhibit the cross-disciplinary collaboration necessary, in particular, to maximise stakeholder engagement with research findings and model predictions.

We will argue that approaches with open-source, lightweight environments and tools can often provide sufficient geo-spatial handling capabilities for research within ICM and encourage collaboration across a broad range of research fields. Some techniques then will be demonstrated through a digital terrain analysis of an upland UK catchment.

1. Data assimilation, stake holder engagement and deployment challenges in catchment research

In the UK, geo-referenced spatial data are now readily available from data sources such as the 25m Countryside Survey Land Cover Map 2000 and 2007, the DEFRA Agricultural Land Classification (ALC), BGS Geological Maps and NextMap 5m DEMs.

This has made feasible the integration of multiple layers of landscape data at the various scales relevant to the water catchment. The "traditional" approach would be to assimilate these data as layers within one of the many commercial or freeware GIS systems.

GIS are powerful tools, however to be used effectively they require considerable experience, training and resources that are not available to all researchers.

Moreover, there are significant issues with deployment of GIS data to a wider audience, both non-specialist catchment "stakeholders" and specialists in fields such as economics, ecology, and political and social science.

Stakeholder engagement is seen as key to the effective application of ICM, and inter-disciplinary collaboration will be important in achieving this aim (see Macleod, Scholefield et al. (2007). Many stakeholders will be unfamiliar with GIS, potentially hindering collaborative work using digital geo-spatial data.

For other users interoperability with other application and platforms may be an issue. Most GIS architectures allow installation of custom toolboxes of hydrological and other geo-spatial analyses. However their methodologies are not always transparent and use can lead to an over-reliance on "magic" functions that produce results without providing insight.

We will demonstrate an approach to catchment research which emphases open data formats and small-footprint, low or zero-configuration open-source tools and platforms and that can provide sufficient GIS functionality for most applications.

It minimises installation, training and licensing requirements for non-specialist users and may encourage both cross-disciplinary collaboration and stakeholder confidence in model predictions. Scripting capabilities and access to the underlying code provide a means of replicating, debugging and verifying others' analyses, providing insight into their model structures and confidence in their algorithmic correctness.

2. R and spatial data

We will in particular promote the use of the free, popular, R language and environment (http://www.R-project.org/), as both a lightweight tool for interactive, data-driven exploration of models and data and as an environment for analysis and visualisation of spatial datasets.

Versions are available for most OS, and installation-free binaries allow its use in "locked-down" environments. In addition - as there is no need for a licencing server (unlike Matlab) the environment can be used whilst offline.

R is primarily known as a statistical language, which may make it familiar to users from the social sciences, for example. Increasingly however it is being used to handle geo-spatial and data from other areas using the numerous add-on libraries (packages) found in the CRAN archive (http://cran.r-project.org/).

Geo-spatial data can be handled in an object-oriented fashion using the extensive class hierarchy implemented in `sp` (Pebesma and Bivand, 2005) These and other CRAN packages provide extensive operations for loading, re-projecting and geo-processing data can in many industry-standard formats.

Researchers are free to choose the subset of methods and packages most applicable to their interests and research

A small selection of common GIS / geo-processing operations along with R equivalents and the packages in which they are implemented is given below

Table 1: GIS operations and R equivalents

Operation	R equivalent: package::method
Load raster layer data	`raster::raster`[1]
Load vector layer data	`rgdal`[2]`::readOGR`
Load ASCII grid data	`maptools`[3]`::readAscii`
Reproject raster data to new spatial reference system (SRID)	`raster::projectRaster`
Reproject vector data to new SRID	`sp: spTransform.`
Apply operation to all multi-band raster	`raster::stackapply`
Create polygons (vector) layer from raster	`raster::rasterToPolygons`
Rasterize vector data	`raster::rasterize`

In addition, visualisation methods such as `plot` , `persp` and `contour` can be applied to R spatial data of different classes, the correct operation being selected according to the type of parameter supplied.

Analysis of spatial data as native R objects combined with the broad range of CRAN packages will encourage the reuse of these data in tools from outside the hydrological and GIS worlds, enabling cross-disciplinary research.

3. Lightweight and web-based GIS

For more sophisticated analysis and camera ready cartographical presentations, a full-GIS system is likely to be used. There are a number of free and open source GIS such as QGIS, GRASS and SAGA whose operation will be familiar to professionals in this area.

Most catchment stakeholders will, however, be more familiar with Google Maps and Google Earth. We will demonstrate how to leverage these popular applications effectively to provide both source and visualisation tools for spatial data. The R packages `dismo`, `Rgooglemaps` and `plotKML,` amongst others, provide facilities for retrieving and manipulating geo-referenced raster data via the Google Web API.

---

[1] Hijmans and van Etten, 2012
[2] Keitt et al., 2012
[3] Lewin-Koh, Bivand. et al, 2012

Complex GIS operations can be achieved through the `RSAGA` package, which allows interoperability between the R environment and the SAGA GIS run-time, and `spgrass6` with GRASS . Similar tools exist for the Python language.

4. Case study – data integration for a hydrological model

The Dynamic TOPMODEL (Beven and Freer 2001) is a semi-distributed grid-based, model developed to predict a catchment's discharge from input precipitation, evapotranspiration and antecedent wetness conditions.

As with the original TOPMODEL (Beven and Kirkby 1979) computational efficiency is achieved by identifying groups of DEM elements that share landscape characteristics – geological, topographical or land cover - relevant to the hydrological response. In the physical model these are then treated as effectively discrete units within an iterative time-stepping scheme, with a contribution to the catchment response proportional to their combined plan areas.

A parameter of interest is the upslope contributing area, a, the effective upslope area per unit contour draining through each cell.. Other features of the catchment, both surface and subsurface, will also affect the hydrological response. Topographically derived data, e.g slope aspect, or aspects of surface cover such as vegetation or land use may also be used to group similar elements with each other.

In order to promote the use of the model to a wider audience an implementation, based on existing Fortran code, was undertaken using R to be distributed via the CRAN archive. It was considered, that an approach a static GIS analysis would make too time-consuming investigation of different strategies for grouping the DEM elements (catchment "discretisation").

Data were obtained for the Upper Wye, a small (c. 10.5km²), heavily instrumented upland catchment in mid Wales. It has been monitored since the late 1960s, originally as part of the Plynlimon research catchment study undertaken by The Institute of Hydrology (now the Centre for Ecology and Hydrology based at Lancaster, see Robinson et al., 2012).

A 15m DEM (ASCII grid) of the catchments derived from an aerial survey commissioned for the original study was retrieved from the CEH Information Gateway along with a vector data set comprising the catchment boundaries. This was processed into a multiband raster comprising the Wye catchment DEM and soils was using GRASS GIS and raster operations, resulting in the equivalent R object, a raster `stack`. Code to load and visualise this native data source in a number of ways is given below.
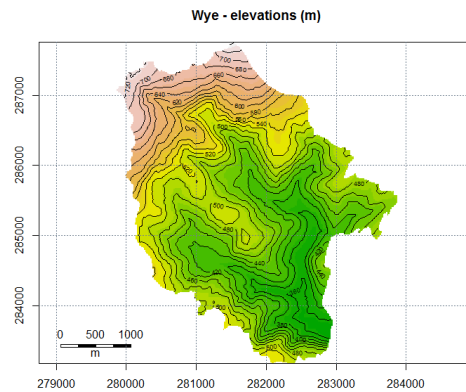
It is then straightforward to apply various analyses, such as the terrain methods in `raster` and hydrological functions `flowlength` and `topidx` methods included in the `topmodel` packages (Buytaert 2011) to our data source. Information on the soils and land cover of the catchment (e.g. the report of Bell, 1971) can also easily be overlaid onto the topography.

1. Load and display elevation raster of target catchment

R code:

```
> load elevation raster from disk
> elevs <- raster("elevs.tif")

# heights (metres abms)
> plot(elevs, col=terrain.colors(25),
main="Wye – elevations (m)",
legend=F))
> scalebar(d=1000,type="bar",divs=5,
below="m")
# add a grid and some contour lines
> grid(col="slategrey")
> contour(elevs, add=T, nlevels=20)
```


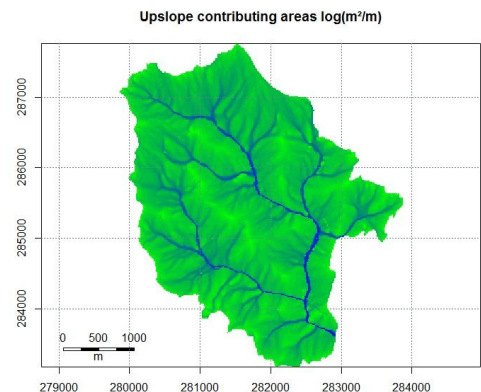2. Retrieve background and overview rasters from Google:Maps server

R code:

```
> require(dismo) # implements a gmap
method
> # retrieve an overview map of Wales from
> #Google Maps
> wales <- gmap("Wales", type="roadmap")
> # transform bounds of catchment raster
to > # Mercator projection used by
> # Google maps rasters
> # then use to define bounds for
background > # map retrieved from Google
> bounds <- projectExtent(elevs,
crs=wales@crs)
> wye<-gmap(bounds, type="hybrid")
> plot(wye)
```

3. Calculate effective upslope area a draining through each grid cell. Estimate location of channel network from areas exceeding a threshold upslope area *a*
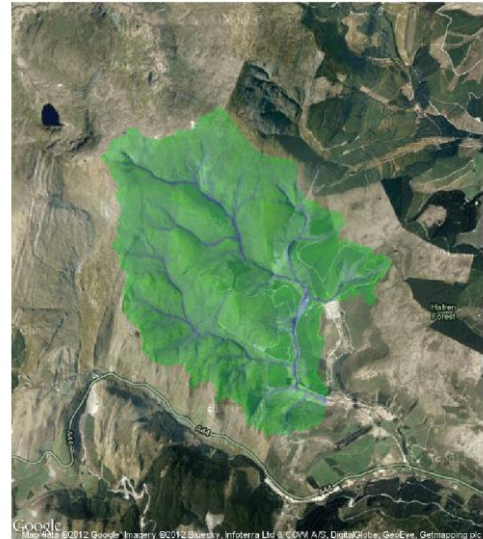
R code:

```
> require(grDevices)
> require(topmodel)
# topmodel implementation:
> areas<-
topidx(as.matrix(elevs),resolution=15)$area
# pick out the wettst areas in blue
> cols <- colorRampPalette(c("green", "blue"))
> areas <- raster(elevs)
> areas<- setValues(areas,
topmodel::topidx(as.matrix(elevs),resolution=15)$a
rea)
> plot(log(areas),col=cols(20))
> scalebar(d=1000,type="bar",divs=5, below="m")
> grid()
```
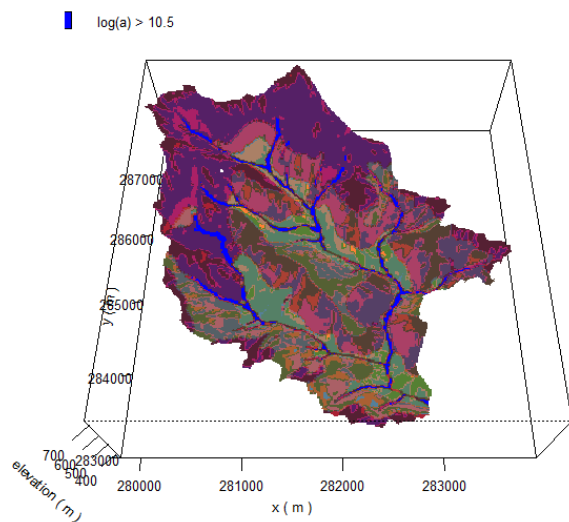

5

Plot an overview map with the upslope areas overlain – this can be cross referenced with the topography and river channel network seen for example from the satellite image retrieved from Google

R code:

```
> plot(wye)
> reproject the upslope areas raster to
> the Mercator CRS used by Google Maps
> bound <- projectExtent(elevs, crs=wye@crs)
> areasMerc <- projectRaster(areas, bound)
> Plot the upslope areas on top of the
> overview – alpha parameter controls
> transparency
> plot(areasMerc, add=T, col=cols(20),
alpha=0.35,legend=F)
>
```


Display a 3-D model of catchment with areal groupings derived from *a* and hydrological soils classification

R code:

```
  # Scale by the grid size to get the
actual coords in m from the SW corner of
the DEM
  x <- ((0:(nrow(elevs)-1))*dx)+xmin
  y <- ((0:(ncol(elevs)-1))*dx)+ymin
  # persp plot with index for facets
colour taken from the HRU id of the
corresponding element
  # class 1 is for elements classed as
"river" due to upslope areas exceeded a
threshold and should be shown in blue
  xLab <- paste("x (", unit, ")")
  yLab <- paste("y (", unit, ")")
  zLab <- paste("elev (", unit, ")")
#  par("font.sub"=2)
#  par("crt"=90),

 # org <- par(mai=c(0.5,0.3,0.75,0.3))
  par(cex.main=0.75)
  drape.plot(x,y, elevs, cm, col=col,
zlab=zLab, xlab=xLab, ylab=yLab,
main=title, sub=sub,
    add.legend=F, phi=phi, theta=theta,
cex.axis = 0.5, cex.lab=0.75,
    expand=expand, ticktype="detailed",
border=border, box=box)
  # legend at top L
  cap <- paste(log(a/tan(b)) >", thresh)
```

5. Conclusions and further development

The heterogeneous data sets required for future catchment research demand a data-driven, exploratory approach, combined with a solidly ground spatial context in order to integrate the data used .

Such capabilities can be provided by existing GIS tools and databases, but we argue that in many cases these cannot be used due to training requirements or lack of resources on the part of catchment stakeholders or researchers.

We have given a brief overview of an alternative approach that makes use of open source and lightweight tools for handling such large, varied spatial data sets.

In the particular case of catchment research such an approach clearly has the potential to maximise stakeholder engagement and encourage productive cross-disciplinary research. These techniques can, however, can be shown to have applications in many areas of the environmental sciences.

Biography

Keith Beven is Distinguished Professor of Hydrology and Fluid Dynamics at Lancaster Environment Centre. His main research interests are in hydrological modelling and understanding the predictive uncertainties associated with environmental models. He is the originator of the Generalised Likelihood Uncertainty Estimation (GLUE) methodology which has been applied to a wide variety of fields.

Peter Metcalfe graduated with an MSc. in computing and physics from the University of Newcastle upon Tyne. He worked in the IT sector for almost 15 years, developing novel design and rating tools for the process engineering and aerospace industries. He is currently studying for a PhD at the Lancaster Environment Centre.

References

(2012). "The Comprehensive R Archive Network." from http://cran.r-project.org/.

Beven, K. J. and R. E. Alcock (2007). "Modelling everything everywhere: a new approach to decision-making for water management under uncertainty." Freshwater Biology **57**: 124-132.

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

Robert J. Hijmans & Jacob van Etten (2012). raster: Geographic analysis and modeling with raster data. R package version 1.9-92. http://CRAN.R-project.org/package=raster

Macleod, C. J. A., D. Scholefield, et al. (2007). "Integration for sustainable catchment management." Science of The Total Environment**373**(2–3): 591-602.

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), http://cran.r-project.org/doc/Rnews/.

Quinn, P., K. Beven, et al. (1991). "The Prediction of Hillslope Flow Paths for Distributed Hydrological Modeling Using Digital Terrain Models." Hydrological Processes**5**(1): 59-79.

Robinson, M, Rooda, J and Sutcliffe, J V, (2012) Long-term environmental monitoring in the UK:origins and achievements of the Plynlimoncatchment study, Trans. Instn. Brit. Geogr., doi: 10.1111/j.1475-5661.2012.00534.x

Sørensen, R. and J. Seibert (2007). "Effects of DEM resolution on the calculation of topographical indices: TWI and its components." Journal of Hydrology**347**(1–2): 79-89.

Tomislav Hengl, Contributions by: Pierre Roudier, Dylan Beaudette and Daniel Nuest (2012). plotKML: Visualization of spatial and spatio-temporal objects in Google Earth. R package version 0.2-5. http://CRAN.R-project.org/package=plotKML

R core Team (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria.