FOSS4G

DENVER 2011

Conference Proceedings

Free and Open Source Software for Geospatial Conference

OSGeo

Your Open Source Compass

Volume 10 Contents

Welcome from the Conference Chair

Welcome to this special edition of the OSGeo Journal, featuring selected papers from the academic track that were presented at the FOSS4G (Free and Open Source Software for Geospatial) 2011 conference in Denver.[1] The conference was the largest FOSS4G yet, with 914 attendees from 42 countries. Feedback from attendees was very positive, with the post-conference survey giving it an overall rating of 4.32 out 5. The attendance reflects the strong growth in interest in open source software that we are currently seeing in the geospatial industry.

We made a conscious effort in 2011 to enhance the academic track at the conference by providing improved publishing opportunities. We did this through publishing papers both in "Transactions in GIS" and in this edition of the OSGeo Journal. I would like to thank Rafael Moreno for leading this effort, as well as the rest of the organizers of the academic track who Rafael recognizes below.

Peter Batty, Ubisense
FOSS4G 2011 Conference Chair

---

[1]FOSS4G: http://foss4g.org

FOSS4G 2011 Conference Proceedings

Using GRASS and R for Landscape Regionalization through PAM Cluster Analysis

Allan D. Hollander

Abstract

Landscape regionalization is a frequently encountered need in the geographical sciences, having applications ranging from sampling design to conservation prioritization. One technique for partitioning the landscape is to use cluster analysis of GIS layers describing the area under study. Here I present a GIS technique that uses partitioning around medoids as its clustering algorithm. Partitioning around medoids (PAM) is a non-hierarchical clustering algorithm that is related to the commonly-used k-means clustering technique. PAM differs from the k-means algorithm in that a) PAM assigns cluster centroids to actual data observations, rather than using values averaged over subsets of the entire dataset and b) PAM accepts categorical data as input in addition to numerical data. These properties of PAM make it useful for landscape regionalization because often one wishes to incorporate categorical variables such as vegetation class or soil types in the regionalization. I illustrate the PAM technique with an example of sampling design for a local-scale analysis of agroecosystems in Northern California. For this work I use GRASS and R, generating in GRASS a set of random points covering the study area and attributing these points with values from raster and vector layers of interest, importing this data table into R for the PAM cluster analysis, and exporting the resulting clusters back to GRASS for geographic visualization.

Background

A long-standing concern in geography has been regionalization, or how to define areas in space that share similar characteristics according to some theme of interest. Quantitative work on regionalization methods has been undertaken for a considerable period of time (e.g. Spence and Taylor (1970), Johnston (1970) and continues to the present (see Duque et al. (2007) for a recent review). In the earth and environmental sciences, regionalization finds application in areas as diverse as hydrology (Wiltshire (1986)), biogeography (e.g. Patten and Smith-Patten (2008), Proches̨ (2005)), and climatology (Stooksbury and Michaels (1991)). Although no quantitative method can provide a single definitive regionalization suitable for all applications, it is useful to have a suite of formal methods available when a problem calling for defining a new set of regions arises. In what follows I present a simple method for generating regions for use in a GIS analysis, working with the open source GIS software GRASS (Team (2009a)) and the statistical package R (Team (2009b)).

The particular focus of this approach is to use cluster analysis methods to support studies in landscape ecology and in conservation. In conservation, one class of problems that can be addressed by cluster analyses is reserve design, where a classical problem is to ensure that a reserve network contains sufficient representation of all elements of conservation interest (e.g. Vane-Wright et al. (1991)). A cluster analysis approach can serve to capture dissimilarities in the environment by placing these in different clusters and hence provide a means to adequately represent the biological diversity of a landscape (e.g. Trakhtenbrot and Kadmon (2005)). Another class of problems is sampling design. In ecological inventories, the number of sampling plots one can place is always quite limited. A usual strategy to increase the efficiency of one's sampling process is stratified design: that is, rather than placing sampling points randomly across the study region, one partitions the landscape across environmental variables of interest, and subsequently places similar numbers of sampling points within each partition. Cluster analysis lends itself well to defining the strata, since it provides a formal method to separate the variability in the environment into different groups.

Cluster analysis algorithms can be divided into hierarchical algorithms and partitional algorithms (Kaufman and Rousseeuw (2005)). In hierarchical algorithms, clusters are constructed sequentially to form a nested hierarchy of cluster types. In partitional algorithms, the set of clusters is constructed all at once, often requiring establishment of the desired number of clusters beforehand. In general, cluster analysis proceeds in two steps, the first being creation of a dissimilarity matrix, where one uses a metric to compute the distance in attribute space between a pair of observations and assigns this value to the element of the matrix where its row and column is indexed by the id numbers of the observation pair. The second step is to take the distances in the dissimilarity matrix and apply an iterative algorithm to sort the observations into different clusters.

In landscape analysis, often the environmental properties of interest are categorical variables. Examples of these include habitat types or soil classifications. The fact that some of the variables are categorical poses problems for many clustering algorithms, for instance the most commonly-used partitional clustering algorithm, k-means clustering. In k-means clustering, the number of clusters is set before running the algorithm. and the algorithm is initialized by randomly placing that number of cluster seeds among the multidimensional environmental attribute space. The algorithm then adjusts the position of these cluster seeds so as to minimize the dissimilarities within a cluster and maximize the dissimilarities between clusters. The k-means algorithm works only with continuous variables, and the cluster centroids are defined to be the mean values of the attributes of all the observations in the cluster. Since this algorithm relies upon computation of mean values, it fails to work with categorical variables, and an alternative algorithmic strategy needs to be taken if one wishes to incorporate categorical variables.

There are two elements in this strategy. The first is to use a dissimilarity metric that works with both categorical and numerical variables. One such metric is Gower's similarity coefficient (Gower (1971)). This coefficient computes the distance between two observations by averaging a difference metric over both the numerical variables and the categorical variables. This difference metric ranges from 0 to 1 for both types of variables. For the numerical variables, the difference metric is the absolute value of the ratio of the difference of the values of the two observations and the range of the variable across all observations. For categorical variables, the difference metric is 0 if the two observations share the same categorical value and 1 if they differ. The second element in this strategy is using a clustering algorithm that works with mixed variables. The algorithm used here is partitioning around medoids (PAM) (Kaufman and Rousseeuw (2005)). This algorithm is similar to k-means clustering but differs from it in that the cluster centers (the "medoids") are actual observation points rather than centroids averaged over a number of observations. The medoids therefore provide the sample point that is the exemplar for each cluster.

This approach can easily be carried out using the open source software programs R and GRASS. R is excellent for testing novel statistical techniques since academic statisticians nowadays develop many new algorithms in the R statistical environment. GRASS provides excellent integration of raster and vector GIS capabilities and scripting abilities.

Methodology

Figure 1 illustrates the workflow in this analysis. The diagram at upper left shows the stack of GIS layers together with the random points to be overlaid on the GIS stack. This creates a data table which is fed into the R statistical environment at upper right. At lower right the clusters are constructed, which are then mapped in the GIS at lower right. Below I detail the steps one takes to carry out this workflow.

1. The first step is to assemble the data layers to be used in the analysis in GRASS. Both raster and vector data can be used. It is important to recognize which datasets are categorical (e.g. land cover) and which are numeric (e.g. soil percent organic matter content), since treatment of these will vary in the statistical analysis. For vector data there may be several fields of interest within the attribute table of a single data layer.

2. The sampling density needs to be determined. There are no fast rules for doing this, but one needs to pay attention to the spatial resolution of the datasets being queried. Given a range of spatial resolutions of the datasets, there is no point in sampling at the finest resolution, but sampling only at the coarsest resolution will by definition lose a lot of the detail in the rest of the datasets. A choice somewhere in the middle of the range may be the most satisfying choice. For instance, in an application where many of the data variables come from a 1-kilometer gridded soils map, one might choose a sampling density of 4 points per square kilometer.

3. The routine for generating random points in GRASS, v.random, asks for a fixed number of points to be generated within the current geographic region. Computing

that value from a desired sampling density requires calculating the area of the region from its bounding box and dividing through by the sampling density.

4. Often one wishes to restrict sampling to a subregion within the rectangular geographic region. In GRASS this may be accomplished by defining the subregion as a vector polygon layer, and then selecting the points that fall within that polygon using the `v.select` command.

5. Next one makes a list of the variables to be sampled and adds these as columns to the attribute table of the sample point layer using the command `v.addtable`, paying attention to the data types of the variables when one creates the columns.

6. These columns then get instantiated with data from the layers of interest, using the commands `v.what.rast` and `v.what.vect`. Running these commands over a large stack of layers can get tedious but this routine can easily be scripted.

7. The attribute table for the sample points is then exported so that it can be read into the R statistical environment. There are several options for export formats; of these a good choice is dbf since it directly preserves data type information of the data table. To export dbf, one selects the dbf option from the `db.out.ogr` command.

8. One then switches over to the the R environment. Though support for spatial datasets in R is well-developed,[34] which includes a direct interface to GRASS through the spgrass6 package), the steps that follow do not depend on this interface, and hence one does not need to understand how R treats spatial data in this workflow. The first step in the R environment is to read the sample point data using the `read.dbf()` function which is part of the foreign package.

9. One then checks to see that the variables in the data frame have the correct data type. Categorical variables may have been represented as integers in this table, especially if they have been sampled from a raster, for example one displaying discrete land cover classes. Such variables need to be reassigned to categorical type using the `factor()` function in R. Conversely, sometimes string variables in the data table actually represent ordinal values, one example being soil drainage being labelled as "poorly drained", "well-drained", or "excessively drained". For analysis, these variables should be manually reassigned to integer ranking.

10. When one reads in the sample point data table into R with the `read.dbf()` function, the first column of the data frame contains the category id value from the vector sample point layer. This id value obviously is not useful in multivariate cluster analysis and needs to be omitted. This is easily accomplished by constructing a new data frame from the original data frame (function `data.frame()`), leaving out the category id variable as well as any other variables that one now wishes to omit. One then assigns the category id values to the row names with the `row.names()` function to preserve identification of the sample points through the cluster analysis.

11. There are two steps to the cluster analysis, creating a dissimilarity matrix and then running a clustering algorithm on the resulting matrix. In R, the library that does both functions is `cluster`.

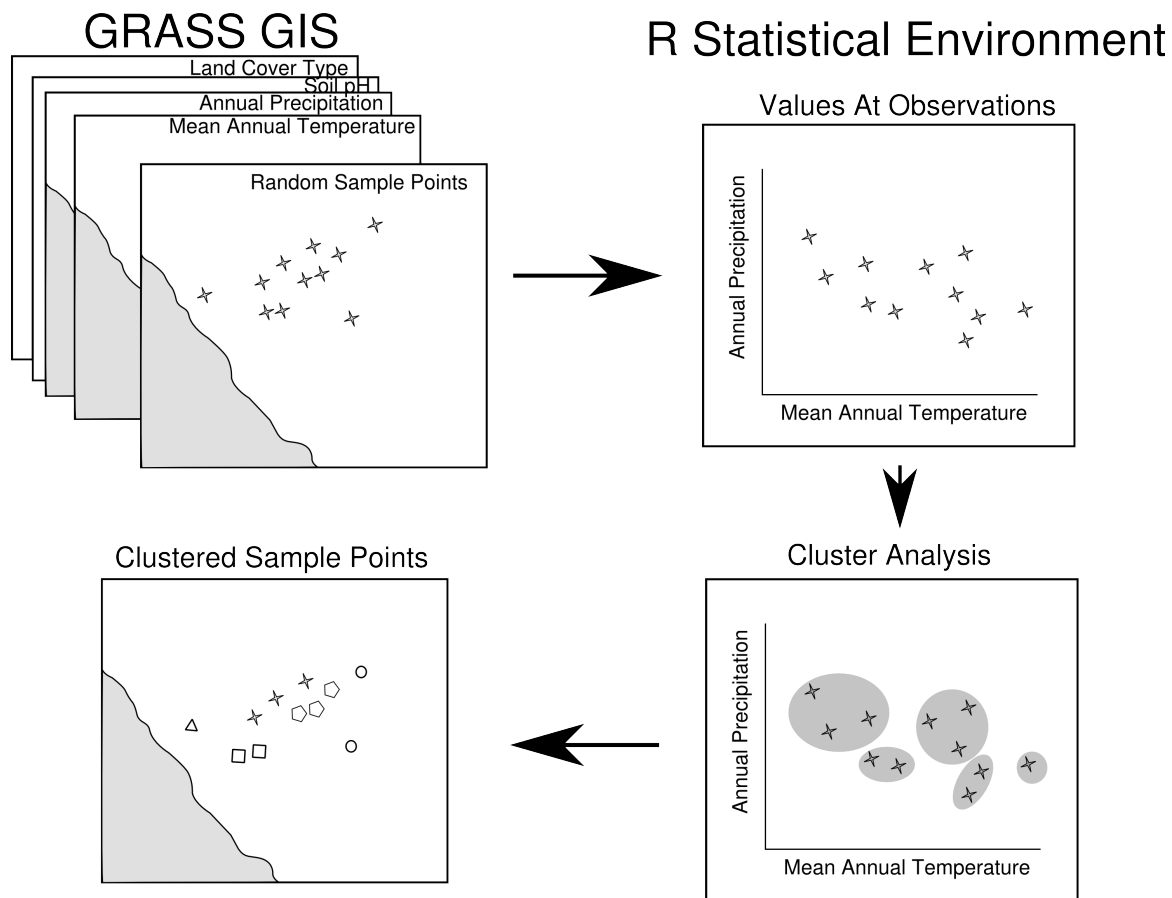[34]See http://cran.r-project.org/web/views/Spatial.html

Figure 1: PAM cluster workflow

12. The routine to construct the dissimilarity matrix is called `daisy()`. It has options for several dissimilarity metrics but defaults to using Gower's similarity metric if the inputted data frame has both numeric and categorical variables. Applying `daisy()` to the data frame above creates a dissimilarity matrix where each entry represents the environmental attribute dissimilarity between a pair of sample points.

13. This dissimilarity matrix is then fed into the `pam()` clustering routine. A key parameter that is also input into `pam()` is the number of clusters to be generated by the algorithm. Because in this set of studies the clusters were to be used in a sampling design containing relatively few plots, few clusters were desired, so the starting point for this parameter was 5 to 7.

14. The R cluster library provides a visualization aid called silhouette plotting (Rousseeuw (1987)) in the function `silhouette()`. This tool helps identify clusters which are poorly defined in the sense that many of their members could easily be placed in another cluster on the basis of dissimilarity distance. By using this tool while iterating through a broad range of numbers of clusters in a set of `pam()` runs, one can identify the choice of the number of clusters that provides for best cluster discrimination. This choice of the parameter is then used in selecting the final cluster analysis.

15. Because the PAM algorithm selects particular sample points that are most representative of each cluster (the

"medoids") it is easy to inspect the values at these sample points by extracting rows from the original data frame by the medoid id number. One can then give the different clusters interpretive labels e.g. "Mixed hardwood upland".

16. Next one brings the result of the cluster analysis back into GRASS. There are several steps to this process. First, one exports a table from R giving the id values for the sample points in one column and the cluster analysis assignments in the second column. (The function `write.csv()` is useful for this). Next one needs to assign the cluster labels to the sample points. This is most easily done by adding a column to the sample point vector layer with the command `v.db.addcol` to contain the integer cluster labels. This column can then be filled with the cluster labels by running a series of update SQL commands that assign an integer cluster label based upon the category id value of the sample points. Shell commands such as awk and sed can be used to generate a file containing the update SQL commands from the R export table.

17. The original sample points can then be plotted on a map labeled with the cluster id values. Depending upon the application, it may be desirable to assign cluster values to the entire surface of the map region, and not just refer to the labeled sample points. One way to accomplish this is to construct a Voronoi tessellation using the command `v.voronoi` and assign each of the demarcated polygons the cluster value of its enclosed sample point.

Case Study

As an illustration of this method I present an example from a study of carbon storage in a mixed agricultural and natural landscape in Northern California (Williams et al. (2011)). Estimation of carbon stocks on the landscape has become an important concern in light of a desire to mitigate greenhouse gas emissions through carbon sequestration (e.g. van Kooten (2009)). Methodologies for estimating carbon stocks are only just now being developed, an example being the California Climate Action Registry Forest Protocol (Reserve (2009)), and these methodologies need to be expanded to a broader suite of landscapes. This study examined a set of vine tracts and the adjoining woodland ecosystems from an organic vineyard in Mendocino County (the Bonterra label of the Fetzer Vineyards). The aim of the study was to estimate carbon stocks across both the vineyards and wildlands using field sampling and model building, to compare the aboveground stocks with the soil carbon stocks, and to develop a methodology that can use these models across a broader landscape for accurate carbon accounting.

Over the 1149 hectare landscape of the study area, it was only feasible to place 93 vegetation plots and dig 44 soil pits, and to maximize the efficiency of placing these samples, we followed the PAM cluster analysis methodology outlined above to classify the landscape for sampling. In this procedure, we used four different data layers in the GIS data stack. These included vegetation type — a categorical variable describing the land cover using one of nine types in the California Wildlife Habitat Relationships system, elevation, slope, and clear-sky solar radiation summed over the entire year as modeled from the elevation map. (calculated using the `r.sun` routine in GRASS). We sampled these four layers using 1013 points distributed randomly over the Fetzer lands. With the data matrix from this random sampling, we produced a matrix giving the dissimilarity between each of the points. This dissimilarity matrix, computed using Gower's coefficient, was used as the input into the PAM clustering routine. Based on the distinctiveness of the groupings produced when several different numbers of clusters were requested, we selected seven as the final number of clusters. Table 1 presents the values of these environmental variables for the 7 medoid points. These clusters could be characterized as following:two vineyard types, one on flat ground at low elevations and one upland type; an annual grassland type at moderate elevation; a mixed hardwood-conifer type at relatively high elevation and steep slopes; and two mixed hardwood types, one at relatively high elevation, steep slopes, and lower solar radiation, and the other at more moderate slopes and elevation. Figure 2 plots the location of the different clusters on the 5 tracts of the Fetzer lands.

Table 1 also shows the number of sample points that were assigned to each cluster. As is evident, the clustering was only used as a rough guide to distributing cluster points. No vegetation plots were assigned to vineyards or annual grasslands, since the aim was to focus on woody vegetation types. The soil pits were divided roughly equally into pits on vineyards and pits on wildlands, but the wildland pits were predominantly assigned to the Mixed Hardwoods II class. The vegetation plots and soil pits were assigned to the cluster types in the field by reference to a map of the cluster points similar to Figure 2.
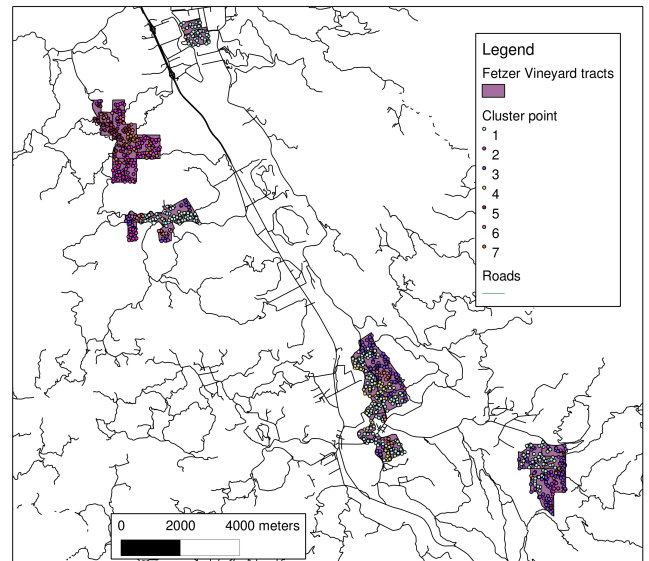

Figure 2: Labeled Cluster Points for Fetzer Study Area

Rather than estimating by eye what cluster point is nearest a given location, it is possible to create a surface from the cluster points and assign a cluster to a location directly. One method for doing this is the Voronoi tessellation technique outlined. This was not done in the case study, but Figure 3 shows the result of doing this for the cluster points in the Fetzer study region.
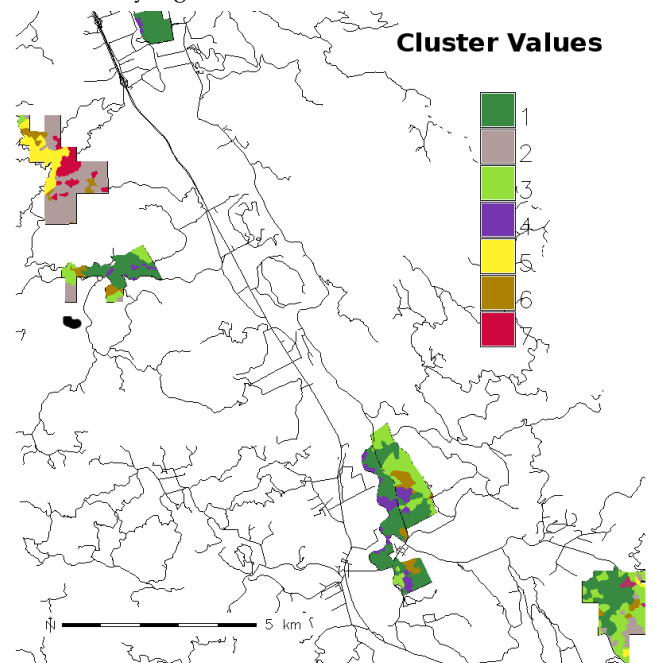

Figure 3: Voronoi Tessellation of Fetzer Cluster Points

Conclusion

This example illustrates how one can use the PAM cluster analysis technique with a set of GIS layers to guide field sampling for research in landscape ecology. We have used this technique to sample other agroecosystems, for example in a study of soil ecology along streams and canals of an agricultural landscape in Yolo County, California (Culman et al.

Number of Vegetation Plots	Number of Soil Pits	Habitat Type	Elevation (meters)	Slope (degrees)	Solar Radiation ($w\ hr/m^2/yr$)
0	19	Vineyard I	173	0.7	2.18×10^6
0	6	Vineyard II	572	10.7	1.99×10^6
0	2	Annual Grassland	213	5.0	2.27×10^6
17	1	Valley Riparian	152	0.4	2.18×10^6
39	2	Mixed Hardwoods I	534	21.3	1.97×10^6
26	13	Mixed Hardwoods II	225	10.7	2.16×10^6
11	1	Mixed Conifer-Hardwood	407	25.8	2.06×10^6

Table 1: Values of Environmental Variables for Cluster Medoids in Fetzer Study

(2010), Young-Mathews, Culman, Sánchez-Moreno, O'Geen, Ferris, Hollander, and Jackson (2010)), and in ongoing work examining soil ecology and gaseous exchange in vineyards in Napa and San Joaquin counties in California. Many more sophisticated variations on cluster analysis could be used: for example, there are a number of algorithms that use spatial constraints directly as part of the clustering process (e.g. Guo and Wang (2011)). But as the example shows, in many cases the results of the clustering are only used as a rough guide for later work, and adopting a simple algorithm is advantageous. An aim for later work will be to automate this procedure. This is made possible by using the GRASS Python scripting API to facilitate the interaction with R in performing the PAM analyses.

Acknowledgements

Bibliography

S. Culman, A. Young-Mathews, A. Hollander, H. Ferris, S. Sánchez-Moreno, A. O'Geen, and L. Jackson. Biodiversity is associated with indicators of soil ecosystem functions over a landscape gradient of agricultural intensification. *Landscape Ecology*, 25(9):1333–1348, November 2010. ISSN 0921-2973. URL http://dx.doi.org/10.1007/s10980-010-9511-0.

J. C. Duque, R. Ramos, and J. Suriñach. Supervised regionalization methods: A survey. *International Regional Science Review*, 30(3):195–220, July 2007. ISSN 0160-0176. doi: 10.1177/0160017607301605. URL http://irx.sagepub.com/cgi/doi/10.1177/0160017607301605.

J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, December 1971. ISSN 0006-341X. doi: 10.2307/2528823. URL http://www.jstor.org/stable/2528823.

D. Guo and H. Wang. Automatic region building for spatial analysis. *Transactions in GIS*, 15:29–45, July 2011. ISSN 13611682. doi: 10.1111/j.1467-9671.2011.01269.x. URL http://doi.wiley.com/10.1111/j.1467-9671.2011.01269.x.

R. J. Johnston. Grouping and regionalizing: some methodological and technical observations. *Economic Geography*, 46(supplement): 293–305, 1970.

L. Kaufman and P Rousseeuw. *Finding groups in data an introduction to cluster analysis.* Wiley,, Hoboken, N.J. :, 2005. ISBN 9780471735786.

M. A. Patten and B. D. Smith-Patten. Biogeographical boundaries and Monmonier's algorithm: a case study in the northern Neotropics. *Journal of Biogeography*, 35(3):407–416, March 2008. ISSN 0305-0270. doi: 10.1111/j.1365-2699.2007.01831.x. URL http://doi.wiley.com/10.1111/j.1365-2699.2007.01831.x.

Ş Procheş. The world's biogeographical regions: cluster analyses based on bat distributions. *Journal of Biogeography*, 32(4):607–614, March 2005. ISSN 03050270. doi: 10.1111/j.1365-2699.2004.01186.x. URL http://doi.wiley.com/10.1111/j.1365-2699.2004.01186.x.

Climate Action Reserve. Forest project protocol, version 3.1. Technical report, Climate Action Reserve, Sacramento, CA, 2009.

P Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, November 1987. ISSN 03770427. doi: 10.1016/0377-0427(87)90125-7. URL http://linkinghub.elsevier.com/retrieve/pii/0377042787901257.

N. A. Spence and P. J. Taylor. Quantitative methods in regional taxonomy. *Progress in Geography*, 2:1–63, 1970.

D. E. Stooksbury and P. J. Michaels. Cluster analysis of southeastern U.S. climate stations. *Theoretical and Applied Climatology*, 44(3-4): 143–150, 1991. ISSN 0177-798X. doi: 10.1007/BF00868169. URL http://www.springerlink.com/index/10.1007/BF00868169.

GRASS Development Team. Geographic resources analysis support system (GRASS) software., 2009a. URL http://grass.osgeo.org/.

R Development Core Team. R: A language and environment for statistical computing, 2009b. URL http://www.r-project.org/.

A. Trakhtenbrot and R. Kadmon. Environmental cluster analysis as a tool for selecting complementary networks of conservation sites. *Ecological Applications*, 15(1):335–345, February 2005. ISSN 1051-0761. doi: 10.1890/04-0077. URL http://www.esajournals.org/doi/abs/10.1890/04-0077.

G. C. van Kooten. Biological carbon sequestration and carbon trading re-visited. *Climatic Change*, 95(3-4):449–463, April 2009. ISSN 0165-0009. doi: 10.1007/s10584-009-9572-8. URL http://www.springerlink.com/content/8372834132127434/.

R.I. Vane-Wright, C.J. Humphries, and P.H. Williams. What to protect? — systematics and the agony of choice. *Biological Conservation*, 55(3):235–254, January 1991. ISSN 00063207. doi: 10.1016/0006-3207(91)90030-D. URL http://linkinghub.elsevier.com/retrieve/pii/000632079190030D.

J. N. Williams, A. D. Hollander, A. T. O'Geen, L. A. Thrupp, R. Hanifin, K. Steenwerth, G. McGourty, and L. E. Jackson. Assessment of carbon in woody plants and soil across a vineyard-woodland landscape. *Carbon Balance and Management*, 6(11), 2011.

S.E. Wiltshire. Identification of homogeneous regions for flood frequency analysis. *Journal of Hydrology*, 84(3-4):287–302, May 1986. ISSN 00221694. doi: 10.1016/0022-1694(86)90128-9. URL http://

linkinghub.elsevier.com/retrieve/pii/0022169486901289.

A. Young-Mathews, S. W. Culman, S. Sánchez-Moreno, A. O'Geen, H. Ferris, A. D. Hollander, and L. E. Jackson. Plant-soil biodiversity relationships and nutrient retention in agricultural riparian zones of the sacramento valley, california. *Agroforestry Systems*, 80(1):41–60, July 2010. ISSN 0167-4366. doi: 10.1007/s10457-010-9332-9. URL http://www.springerlink.com/content/m57j684058401536/.

Allan D. Hollander
Information Center for the Environment
University of California, Davis
adhollander@ucdavis.edu

This PDF article file is a sub-set from the larger OSGeo Journal. For a complete set of articles please the Journal web-site at:

and ready for the future. TRLIB was only recently released under an open source license (and made available through `https://bitbucket.org/KMS/trlib`), but in the near future we hope to implement means for better interoperability with the more well established libraries in the open source geomatics field.

Acknowledgements: We thank Willy Lehmann Weng and Knud Poder for commenting on the draft of this paper.

Bibliography

C. W. Clenshaw. A note on the summation of chebyshev series. *Math. Tables Aids Comput.*, 9(51):118–120, 1955. URL `http://www.jstor.org/stable/2002068`.

Karsten E. Engsager and Knud Poder. A highly accurate world wide algorithm for the transverse mercator mapping (almost). In *Proc. XXIII Intl. Cartographic Conf. (ICC2007), Moscow*, page 2.1.2, 2007.

Gerald I. Evenden. Cartographic projection procedures for the unix environment—a user's manual, 1990. US Geological Survey Open-File Report 90–284.

C. Gram, O. Hestvik, H. Isakson, P.T. Jacobsen, J. Jensen, P. Naur, B.S. Petersen, and B. Svejgaard. Gier - a danish computer of medium size. *IEEE Transactions on Electronic Computers*, EC-12(5):629–650, December 1963. URL `http://www.datamuseum.dk/site_dk/rc/gierdoc/ieeeartikel.pdf`.

Charles Karney. Transverse mercator with an accuracy of a few nanometers. *Journal of Geodesy*, 2011(1):1–11, 2011. doi: http://dx.doi.org/10.1007/s00190-011-0445-3.

R. König and K. H. Weise. *Mathematische Grundlagen der Höheren Geodäsie und Kartographie, Erster Band*. Springer, Berlin/Göttingen/Heidelberg, 1951.

Knud Poder and Karsten Engsager. Some conformal mappings and transformations for geodesy and topographic cartography, 1998. National Survey and Cadastre, Denmark, Publications, 4. series vol. 6.

Carl Christian Tscherning and Knud Poder. Some geodetic applications of clenshaw summation. *Bolletino di Geodesia e Scienze Affini*, XLI(4):349–375, 1982.

Thomas Knudsen
National Survey and Cadastre
`kms.dk`
`thokn AT kms DOT dk`