**Methods For Developing Landscape-Scale Ecological Units In Southern California**

22 January 2016

Allan D. Hollander

**Aim and Rationale:**

In undertaking a study of Southern California on ecosystem services we desired to have a physiographically-based classification and mapping of the landscape into a relatively limited number of ecologically-relevant types. Such a classification would improve geographical understanding of the magnitude of these ecological services and allow for simple statistical summaries of their values by their ecological unit type. Prior work on classifying this landscape included an ecological unit inventory analysis of the four National Forests in Southern California, published in 2001 (Gallegos et al., 2001) We opted to develop our own classification rather than use this work for a number of reasons. First, much better raster data is available now for building this classification, especially fine-scale raster climate layers at a 270-meter resolution. Next, the existing landscape units were confined to the areas of the National Forests, hindering analyses on the lands buffering the forests. Finally, this prior work was more focused on land facet mapping rather than developing a simple classification, resulting in it having a high number of classificatory types in the system (e.g. 301 different ecological land units are mapped).

**Methods**:

1) The first step was gathering a stack of GIS layers for the Southern California study using a variety of sources and resolutions. These layers ran in their original resolution from 10 meters to 270 meters. All analyses were performed at a 30 meter raster resolution, with resampling done as necessary. The following datasets were used. We decided not to include vegetation in the data stack as the aim is to capture biophysical characteristics and vegetation will reflect current landscape history and land use patterns (e.g. fire history, type conversion from shrubland, agricultural use). Lakes and reservoirs were omitted from the subsequent analysis. Table 1 below provides a summary of the data layers.

a) Soil suborders. This was a discretely-classified raster layer with 22 soil suborder classes included in the Southern California region. This was derived from the gridded Soil Survey Geographic Database (gSSURGO, available at http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2\_053628). This product is a rasterization of the county-scale SSURGO data published by the USDA Natural Resources Conservation Service.

b) Terrain geomorphons. This raster layer derives from a DEM surface and classifies the landscape into 10 discrete landform types, examples being ridges, slopes, hollows, and valleys. The algorithm for geomorphon classification uses a pattern recognition approach based on line of sight analysis (Jasiewisc and Stepinski 2013). This layer was created from a 30 meter DEM in GRASS 7.0.0, using the extension *r.geomorphon* (https://grass.osgeo.org/grass70/manuals/addons/r.geomorphon.html).

c) Annualized solar irradiation. This layer uses the *r.sun* model available in GRASS 7.0.0 (https://grass.osgeo.org/grass70/manuals/r.sun.html) which calculates direct, diffuse, and reflected solar irradiation for a given day, location, topography, and atmospheric conditions. This layer was created from a 30 meter DEM and assumes clear-sky conditions. To estimate the total annual irradiation, the model was run for every 15th day and these values were integrated over the year.

d) Flow accumulation. This layer is another product of 30 meter DEM data and measures the upslope area in pixel count that conceivably drains into a given pixel. This was calculated using the accumulation option in the GRASS 7.0.0 command *r.watershed* (https://grass.osgeo.org/grass70/manuals/r.watershed.html)

e) Slope. This was derived from 30 meter DEM data using the GRASS 7.0.0 command r.slope.aspect, and is measured in degrees.

f) Annual precipitation. This layer came from the 2014 Basin Characterization Model for California (Flint et al. 2013) and gives the average annual precipitation between 1981 and 2010 at a 270-meter resolution.

g) Annual minimum temperature.This layer came from the 2014 Basin Characterization Model for California (Flint et al. 2013) and gives the average annual minimum temperature between 1981 and 2010 at a 270-meter resolution. Minimum temperature was included in the set of climate variables to represent montane winter conditions.

h) Climatic water deficit. This layer came from the 2014 Basin Characterization Model for California (Flint et al. 2013) and gives the average climatic water deficit between 1981 and 2010 at a 270-meter resolution. The two evapotranspiration variables (climatic water deficit and actual evapotranspiration) are included in this set because they are strong drivers of vegetation distribution (Stephenson 1998).

i) Actual evapotranspiration. This layer came from the 2014 Basin Characterization Model for California (Flint et al. 2013) and gives the average actual evapotranspiration between 1981 and 2010 at a 270-meter resolution.

**Table 1 - Summary of GIS data stack**

| Layer | Original Source | Original Resolution | Theme |
| --- | --- | --- | --- |
| Soil suborders | gSSURGO | 10 meters | Soil type |
| Terrain geomorphons | Digital elevation model | 30 meters | Geomorphometry |
| Solar irradiation | Digital elevation model | 30 meters | Energy balance |
| Flow accumulation | Digital elevation model | 30 meters | Geomorphometry |
| Slope | Digital elevation model | 30 meters | Geomorphometry |
| Annual precipitation | Basin Characterization Model | 270 meters | Climate |
| Annual minimum temperature | Basin Characterization Model | 270 meters | Climate |
| Climatic water deficit | Basin Characterization Model | 270 meters | Climate |
| Actual evapotransiration | Basin Characterization Model | 270 meters | Climate |

2) The next step was overlaying 10,000 random points on this data stack within a mask covering the geographical extent of the study area, which was 35,158 square kilometers. This generated a data table giving the values of the above GIS layers at each of the random points. The role of this data table is to provide a basis for sorting the random points into a limited number of clustered types. The first step in doing this is calculating in multivariate space the distance with respect to these environmental variables each random point is from every other point, in other words creating a dissimilarity matrix.

3) The need to create a dissimilarity matrix leads to a subanalysis. Because the 9 environmental variables use completely different metrics and are a combination of numerical and categorical types, calculating an environmental distance between any two of these random points requires some weighting to be assigned to each of the environmental variables to sum up their relative distances. How do we assign this weighting?

4) This subanalysis to determine the weighting focused on the smaller Santa Clara River watershed study region. Since these ecological units are intended to summarize a diverse set of ecological services, we chose three different proxy variables from the GIS data available for this region to represent biomass, hydrological response, and biodiversity. These proxies included mean annual MODIS EVI value for biomass, recharge for hydrological response, and habitat type in the CWHR classification for biodiversity.

a) The MODIS EVI data was derived by averaging over the 2000-2014 period the maximum EVI value in a single year. The MODIS index used was MOD13Q1 (https://lpdaac.usgs.gov/dataset\_discovery/modis/modis\_products\_table/mod13q1) at a 250 meter resolution, available at 16-day intervals.

b) The hydrological recharge was taken from the 2014 Basin Characterization Model (Flint et al. 2013) at 270 meter resolution.

c) The habitat type came from the 2015 FRAP vegetation layer (FVEG15\_1, from http://frap.fire.ca.gov/data/frapgisdata-sw-fveg\_download), available at a 30 meter resolution.

5) We used random forest regression and classification (Hastie et al. 2009) to determine a ranking of a ranking of importance values of these predictor variables using random forest regression for EVI and recharge and random forest classification for the habitat type. These were calculated using the *randomForest* package in R (Liaw and Wiener 2002).

6) We then averaged these three sets of importance values to create an overall set of weightings for these variables to enter into the dissimilarity matrix. The weights for these variables, reflecting their relative importance for the ecological units, were the following:

|  |  |
| --- | --- |
| Variable Name | Weight |
| Precipitation | 1.00 |
| Annual minimum temperature | 0.600 |
| Slope | 0.507 |
| Climatic water deficit | 0.413 |
| Annualized solar radiation | 0.404 |
| Soil suborder type | 0.374 |
| Actual evapotranspiration | 0.367 |
| Flow accumulation | 0.355 |
| Geomorphon type | 0.226 |

7) Once this subanalysis was finished, we then applied these weights to create a dissimilarity matrix for the10,000 random points in the Southern California study area. This was calculated using the *daisy* function in the R package *cluster* (Maechler et al. 2015).The “gower” metric was used in the *daisy* function: this distance metric (Gower 1971) allows one to combine numerical and categorical variables in a single measure.

8) To group these random points into discrete cluster types we used a clustering technique called Partitioning Around Medoids, or PAM (Kaufman and Rousseeuw 2005, Hollander 2012). This is a partitional cluster algorithm which is one where the set of clusters is established all at once, usually requiring assignment of the desired number of clusters beforehand. The PAM technique is similar to the more commonly used k-means algorithm but differs from the latter in that a) cluster centroids are assigned to actual data observations rather than means of the data values and b) the PAM algorithm accepts categorical as well as numerical data. We used the *pam* function in the R *cluster* package to carry out this clustering.

9) Because the PAM algorithm calls for requesting a particular number of clusters ahead of time, we need to establish the optimum number of clusters to portray the landscape variability. An approach for doing this is to calculate sets of provisional clusters over a range of requested numbers of clusters, and to evaluate the performance of each of the sets using a combination of internal and external homogeneity metrics. We created a collection of cluster sets ranging from 23 to 50 clusters in each.

10) The measure used for internal homogeneity was the average silhouette width of the clustering (Rousseeuw 1987). This value, also accompanied by a graphical display called a silhouette plot, helps illustrate which observation points lie near the center of each respective cluster, and which observation points fall in between two or more clusters. Averaging this measure of cluster validity across every observation point gives the average silhouette width of the cluster analysis, a higher value meaning a better partitioned cluster.

11) The metric used for external homogeneity was a measure of within-cluster variability for the three proxy variables. When this number is minimized, the clusters were best separated with respect to that particular variable. To take an example, for a given cluster set (say one with 40 clusters in it), the standard deviation of the MODIS EVI values was taken for the observation points grouped by cluster assignment. The mean value of these standard deviations was then calculated to give an overall value of within-cluster variability for that particular cluster set. These overall values were then compared across the collection of cluster sets. A low overall value means that the cluster assignment does a good job of separating observation points on the basis of their MODIS EVI value ranges.

12) Based on this evaluation, looking at the average silhouette width for a measure of internal homogeneity and the three within-cluster variability measures for the proxy variables, we chose the clustering with 37 classes in it across the Southern California study area.

13) We then mapped these clusters across the Southern California landscape. To do this, we needed a method to take the cluster assignments for the 10,000 random sample points and generalize these assignments across a raster surface covering the study area. To do this, we constructed a random forest model to predict class assignments given environmental variables from the raster stack, constructing the model by using the data from the original sample point table as independent variables. We then used the *raster* library in R (Hijmans 2015) to run the random forest model across the GIS stack of 9 input rasters. Finally, we smoothed the GIS cluster map slightly by running a 3x3 majority filter over it.

14) We then filled in areas of the map with missing soil data by creating a supplemental random forest classifier with soils variables removed from the list of environmental variables, and ran this classifier to map clusters in areas with missing soils data.

**References**:

Flint, Lorraine E., Alan L. Flint, James H. Thorne, and Ryan Boynton. 2013. “Fine-Scale Hydrologic Modeling for Regional Landscape Applications: The California Basin Characterization Model Development and Performance.” *Ecological Processes* 2 (1): 25. doi:10.1186/2192-1709-2-25.

Gallegos, Alan, Fred Levitan, Catherine Phillips, and Brent Roath. 2001. “Southern California Province Landtype Association Ecological Unit Inventory.” United States Forest Service.

Gower, John C. 1971. “A General Coefficient of Similarity and Some of Its Properties.” *Biometrics* 27: 623–37.

Hollander, Allan D. 2012. “Using GRASS and R for Landscape Regionalization through PAM Cluster Analysis.” *OSGeo Journal* 10 (1): 6.

Jasiewicz, Jarosław, and Tomasz F. Stepinski. 2013. “Geomorphons — a Pattern Recognition Approach to Classification and Mapping of Landforms.” *Geomorphology* 182 (January): 147–56. doi:10.1016/j.geomorph.2012.11.005.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. Berlin: Springer.

Hijmans, Robert J.. 2015. “raster: Geographic Data Analysis and Modeling”, R package version 2.3-40.

Liaw, Andy and Matthew Wiener. 2002 . “Classification and Regression by randomForest”. *R News* 2(3), 18–22.

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

Maechler. Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2015. “cluster: Cluster Analysis Basics and Extensions” R package version 2.0.3.

Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20 (November): 53–65. doi:10.1016/0377-0427(87)90125-7.