Modelling and mapping soil pH in Andalusia (Spain) using phenological products as predictor features

Francisco M. Canero¹, Victor Rodriguez-Galiano¹, Aaron Cardenas-Martinez², and Juan Antonio Luque-Espinar²

¹Physical Geography and Regional Geographic Analysis, University of Seville, 41004 Seville, Spain
²Spanish Geological Survey (IGME), Granada, Spain.

Materials and methodology

Study area, sampling and target feature

3215 samples for Andalusia (Fig. 1) were obtained from Geochemical Atlas of Spain, made by Spanish Geological Survey (IGME).

Fig 1. Location map of Andalusia (NUTS2 region)

pH histogram (Fig 2) showed that target feature was bimodal: values have a two peak concentration centered around 5.7 and 8.3 values

Fig 2. Histogram of pH sampling

The aim of this work is two-folded:
1. Mapping of pH over Andalusia, Spain
2. Evaluate new features derived from remotely sensed time-series.

Methodology

Generation of the regression matrix

- Data filtering
  - Dropping predictive features
  - Exclusion of predictive features values for each sampling point

Data matrix split
- Calibration data (55% of regression matrix/5121 samples)
- Validation data (for testing): a random selection matrix/1024 samples

Prediction
- Modelling using multiple linear regression
- Modelling using Random Forest (RF) algorithm (Fig 4)

Testing
- RMSE of the out-of-bag (OOB) error for hyperparameter setting purpose on RF, and RMSE and RF of the independent fixed test as validation data

Predictive features

79 different features were used in the modelling, which can be summarised in 3 broader groups: climatic features, terrain features, and land surface phenology (LSP) features.

Raster layers were resampled to the LSP grid (derived from MODIS products)

Climatic features

- Daily rainfall
- Mean temperature
- Minimum temperature
- Maximum temperature
- Total rainfall
- Average daily rainfall
- Summer rainfall
- Winter rainfall
- Total rainfall for the 1971-2000 period

Terrain features

- Location
- Slope
- Aspect
- General Curvature
- Relief Curvature
- Topographic Wetness Index
- Convergence Index
- LS Factor
- Multiresolution Valley Bottom Flatness Index
- Multiresolution Ridge Top Flatness Index
- Terrain ruggedness index

Climatic grid obtained from thousands of ground stations at a 200 m resolution.

• Annual and monthly total rainfall for the 1971-2000 period.
• Average maximum and minimum temperatures in the 1971-2000 period.

Climatic grid obtained from a digital elevation model (DEM) originally from the 20m Spanish 15m DEM and obtained using SAGA software.

• Elevation
• Slope
• Aspect
• General Curvature
• Relief Curvature
• Topographic Wetness Index
• Convergence Index
• LS Factor
• Multiresolution Valley Bottom Flatness Index
• Multiresolution Ridge Top Flatness Index
• Terrain ruggedness index

3) Results

Random Forest MLR

<table>
<thead>
<tr>
<th>R²</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.66</td>
<td>0.76</td>
</tr>
<tr>
<td>0.58</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Fig 8. Validation results (with independent test)

Fig 9. Most important features in RF modelling. (IncNodePurty > 100). From right to left, features are median of large integral, maximum temperatures in September, valley depth, median of maximum value of NDVI, maximum temperature in June, rainfall in July, and median of date of end of season

Fig 10. MLR map

4) Conclusions

- RF outperformed MLR modelling, due to advantages of ML modelling against traditional statistic approach (non-linear modelling, overfitting reduction…), especially when target feature statistical distribution is not Gaussian.
- LSP features and rainfall were found as the most important features related to soil pH, with an inverse relation. ML feature selection also considered maximum temperatures (in September and June) as an important predictive feature.
- Large integral (LINT) was found as best predictor feature in both feature pairwise correlations (0.85) and RF feature importance measurement: this could be on account of LINT as gross primary production (GPP) proxy, and the trend of soils to acidification because of an increased presence of the organic complex; so, the greater the value of LINT, pH value was lower.
- Improvements could be done in such many ways: incorporation of geological and other predictor features, using feature selection algorithms to reduce data dimensionality and Hughes effect, comparison between different ML algorithms, analysis of the geographical distribution of error measures…

References