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

Francisco M. Canero[1], Victor Rodriguez-Galiano[1], Aaron Cardenas-Martinez[1], and Juan Antonio Luque-Espinar[2]

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

Materials and methodology

Study area, sampling and target feature

Fig 1. Location map of Andalusia (NUTS2 region)

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

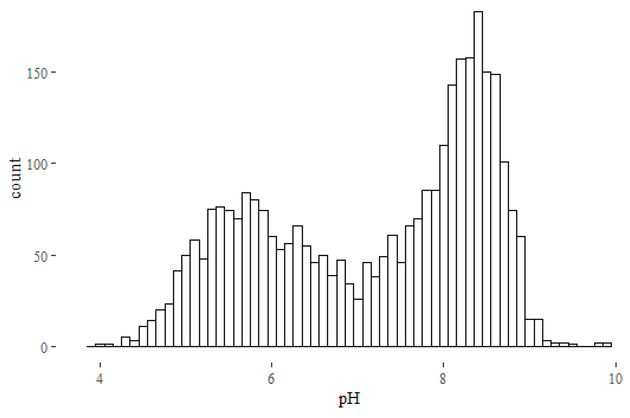


Fig 2. Histogram of pH sampling

pH histogram (fig 2) showed that target feature was bimodal: values have a two-peak concentration centered around 5,7 and 8,3 values

Objectives

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

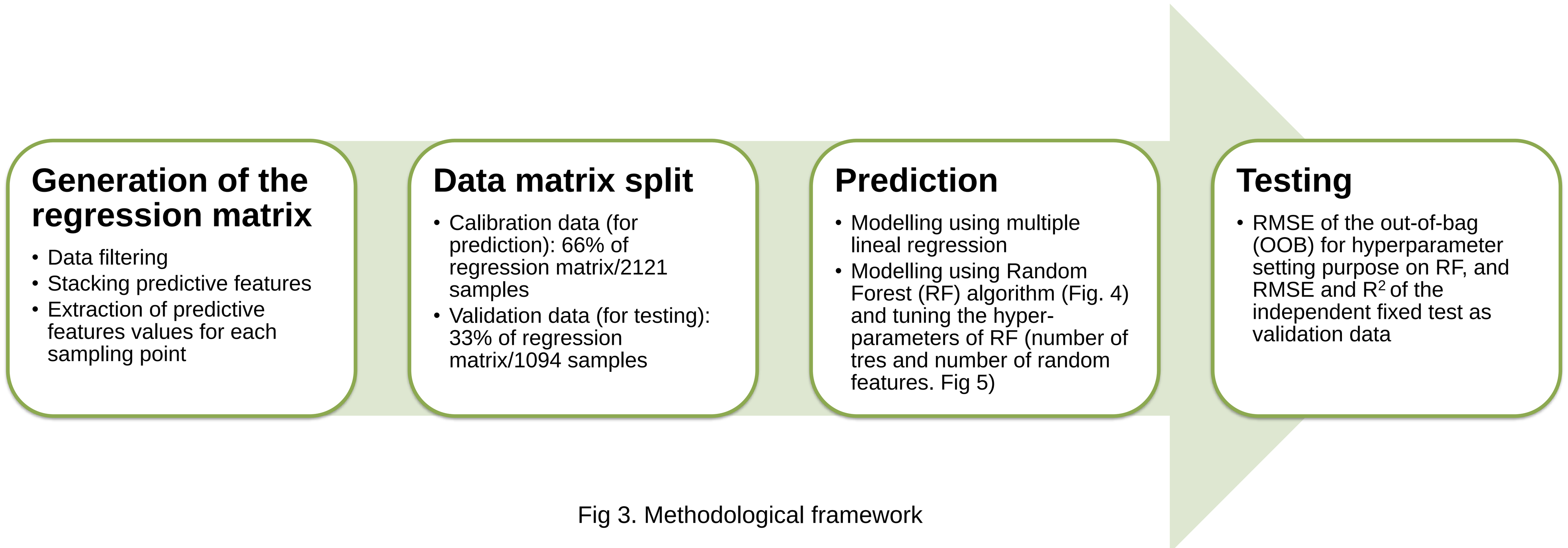


Fig 3. Methodological framework

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	Land Surface Phenology	Terrain features
Climatic grids obtained from thousands of ground stations at a 250-m resolution. • Annual and monthly total rainfall for the 1971-2000 period. • Average maximum and minimum temperatures in the 1971-2000 period.	Median and standard deviation for the 2003-2007 period of the following LSP features: • Date for the start of the season (SOS). • Date for the end of the season (EOS). • Length of the season (LOS). • Base level (BV). • Time for the mid of the season (MOS). • Largest data during the season (MAX). • Seasonal amplitude (AMP). • Rate of increase at the beginning of the season (LDER). • Rate of decrease at the end of the season (RDER). • Large seasonal integral (LINT). • Small seasonal integral (SINT). • Value for the start of the season (VSOS). • Value for the end of the season (VEOS).	Terrain attributes were derived from a 250-m aggregated DEM (originally from the 25m Spanish IGN DEM) and obtained using SAGA software. • Elevation. • Slope. • Aspect. • General Curvature. • Plan Curvature. • Profile Curvature. • Topographic Wetness Index. • Convergence Index. • LS Factor. • Multirresolution Valley Bottom Flatness Index. • Multirresolution Ridge Top Flatness Index. • Valley Depth. • Terrain Ruggedness Index.

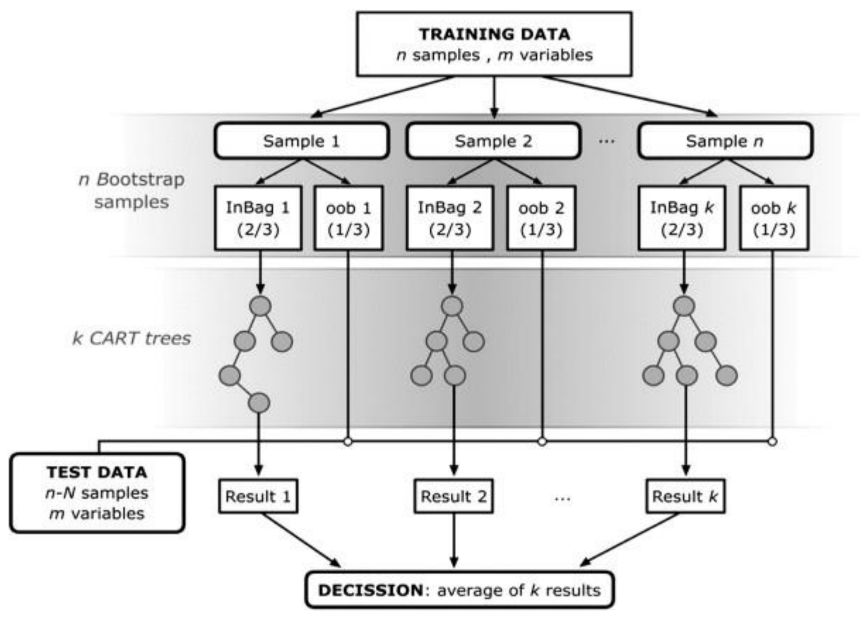


Fig 4. Performance on Random Forest. Based on Rodriguez-Galiano et al, 2012

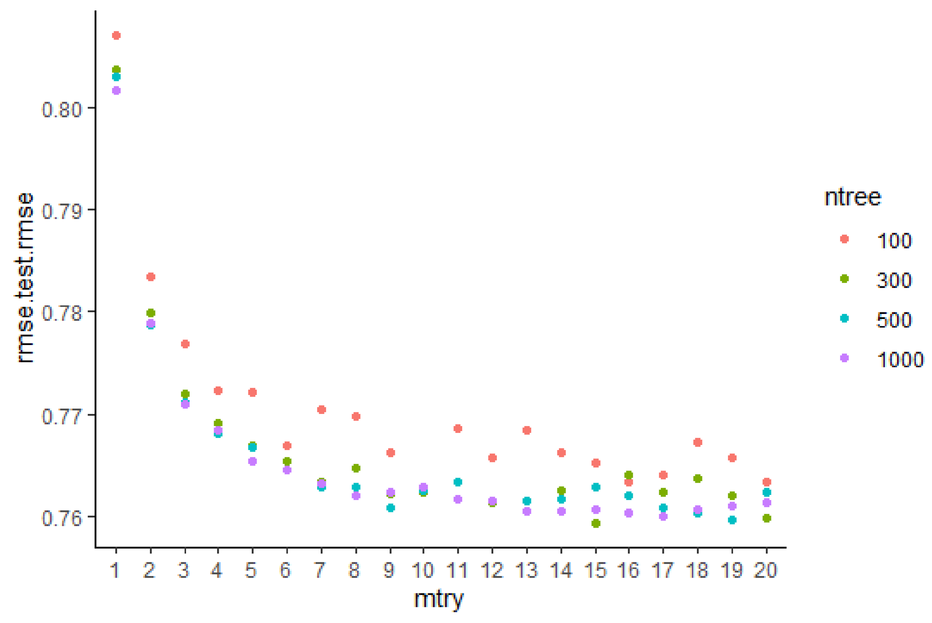


Fig 5. Performance of hyperparameter iteration on RF. Selected combination was 300 ntree, 15 mtry.

3) Results

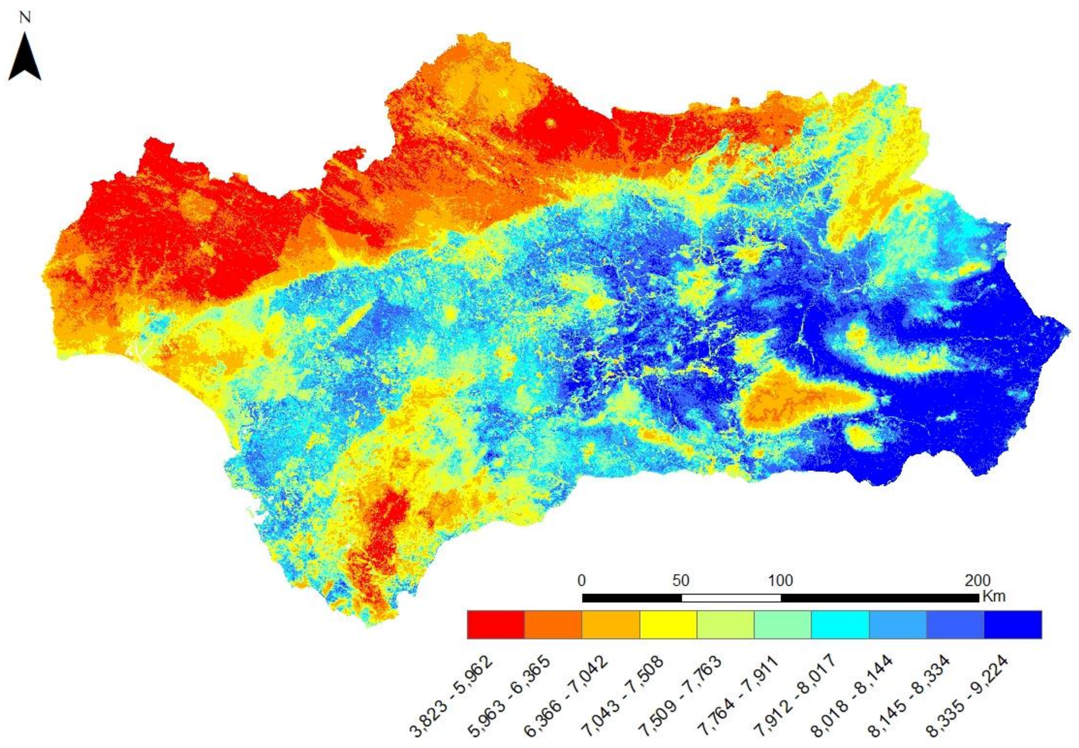


Fig 6. Random Forest map

	Random Forest	MLR
R^2	0.66	0.58
RMSE	0.76	0.83

Fig 8. Validation results (with independent test)

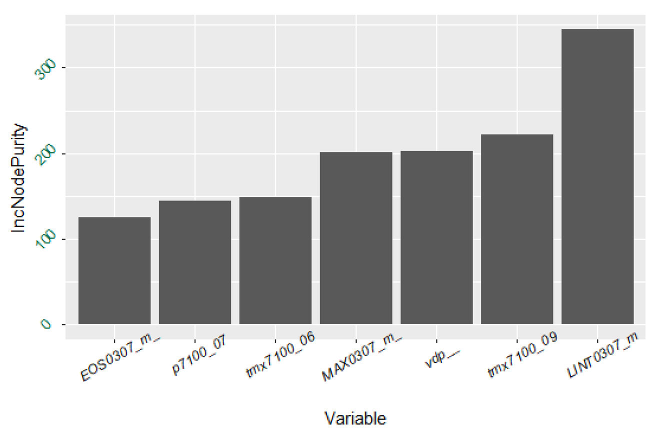


Fig 8. Most important features in RF modelling (IncNodePurity > 100). From right to left, features are median of large integral, maximum temperatures in September, valley depth, median of maximum value of NDVI, maximum temperatures in June, rainfall in July, and median of date of end of season

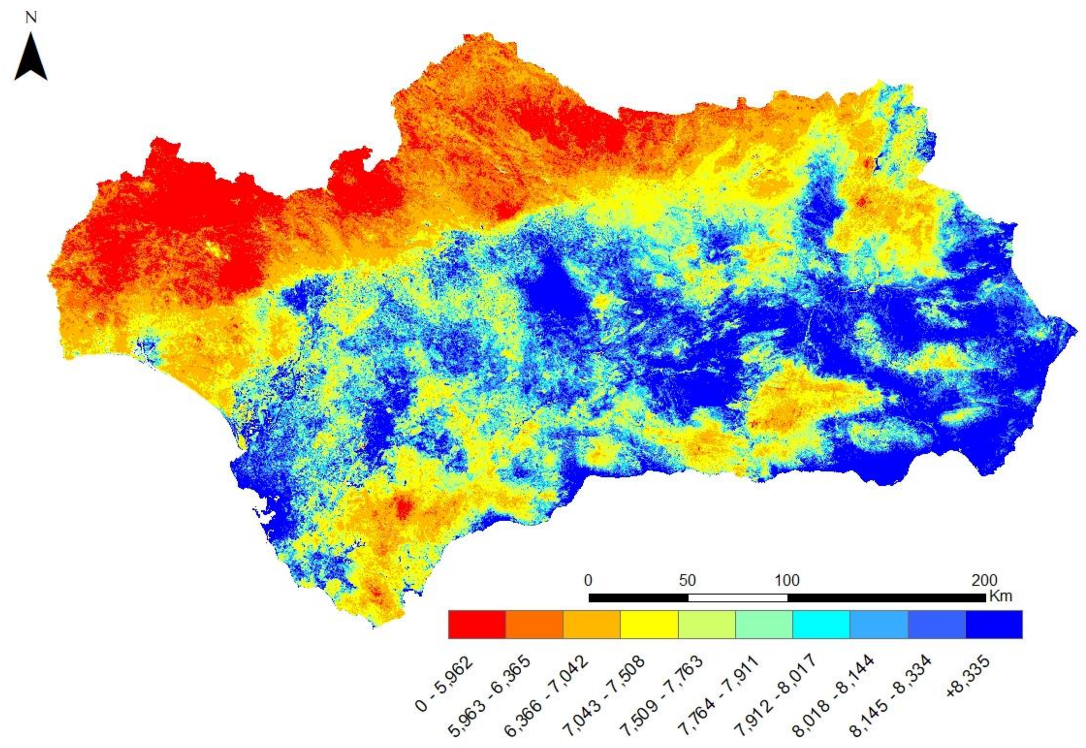


Fig 10. MLR map

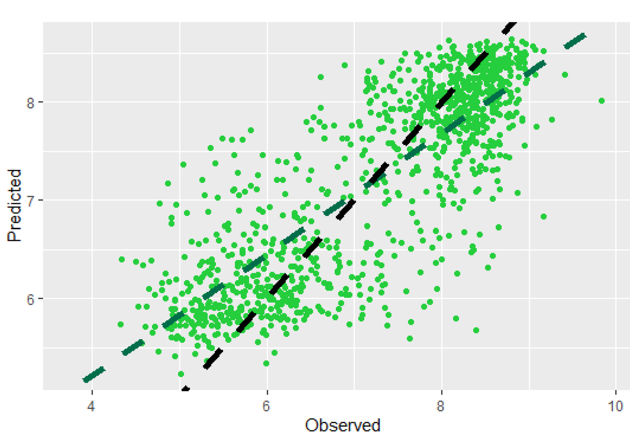
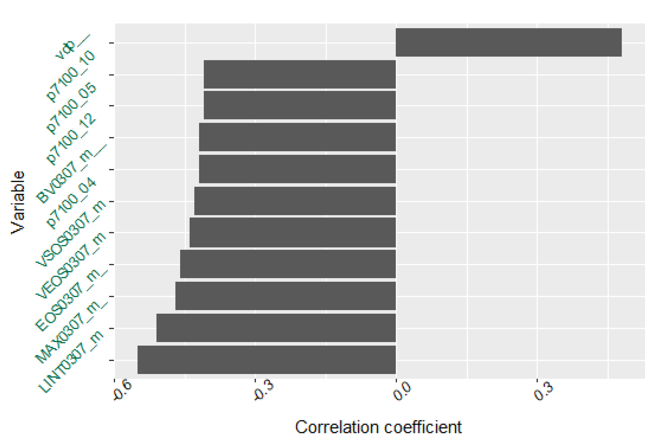


Fig 7. Observed and predicted RF values for pH

Fig 9. Most important features in pairwise correlation. From above to below: valley depth (positive correlation), rainfall in October, May and December, median of base level, rainfall in April, median of value at start of season, median of value at end of season, median date of end of season, median maximum value of NDVI series, and median of large integral



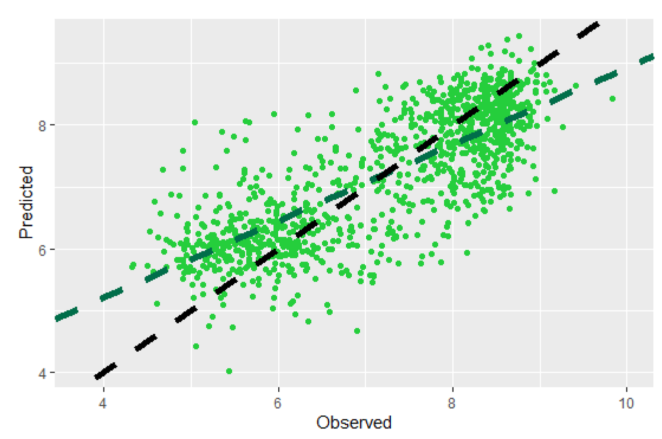


Fig 11. Observed and predicted MLR values for pH

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 (-0,55) 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

[1] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93-104.

CONTACT:
fcanero@us.es
vrgaliano@us.es