Machine learning techniques to derive bioclimatic classifications for Colombia ☆

Richard Rios^{a,b,*}, Elkin A. Noguera-Urbano^b, Jairo Espinosa^a, Jose Manuel Ochoa^b

^a Universidad Nacional de Colombia, Facultad de Minas, GAUNAL, Medellín, Colombia

^b Programa de Evaluación y Monitoreo, Instituto Alexander von Humboldt, Bogotá, Colombia

Abstract

Bioclimatic classifications seek to divide a study region into geographic areas with similar bioclimatic characteristics. In this study we proposed two bioclimatic classifications for Colombia using machine learning techniques. We firstly characterized the precipitation space of Colombia using principal component analysis. Based on Lang classification, we then projected all background sites in the precipitation space with their corresponding categories. We sequentially fit logistic regression models to re-classify all background sites in the precipitation space with six redefined Lang categories. New categories were the used to define a new modified Lang and Caldas-Lang classifications.

Keywords: Bioclimatic classification, Caldas classification, Lang classification, logistic regression models, dimensionality reduction

1. Introduction

Bioclimatic classifications seek to divide a study region into geographic areas with similar bioclimatic characteristics (Medina Bermudez and Aldana Buitrago, 2019; IDEAM, 2015). Empiric methods are the most common approaches used to derive bioclimatic classifications by describing relationships between bioclimatic variables, such as temperature or precipitation (Medina Bermudez and Aldana Buitrago, 2019; IDEAM, 2015). Caldas, Lang, and Caldas-Lang classifications are three of the most common classifications derived for Colombia using mean annual values of temperature and precipitation (Medina Bermudez and Aldana Buitrago, 2019; IDEAM, 2015, 2011). For example, Lang classification defines a factor to describe the humidity characteristics of background sites and classify the climate into six categories. This factor is derived as the ratio between the annual precipitation and mean annual temperature. However, this ratio of bioclimatic variables could lead that background sites with low temperatures have similar humidity characteristics than background sites with high

☆This document is a collaborative effort between the Instituto Alexander von Humboldt (IAvH) and GAUNAL research group from the Universidad Nacional de Colombia.

*Corresponding author: Richard Rios (rriospa@unal.edu.co)

15 level of annual precipitation. As a result, geographical areas like Sierra Nevada de Santa Marta and Chocó could have the same bioclimatic category. In this document we proposed two bioclimatic classifications for Colombia using machine learning techniques. We initially obtained a modified Lang classification using dimensionality reduction and classification techniques. We then integrated the modified Lang and Caldas classifications to derive a second bioclimatic classification for Colombia.

2. Material and methods

2.1. Data

25 The study area encompassed all the continental territory of Colombia. We used five bioclimatic variables at $1 - km^2$ spatial resolution, which were: annual mean temperature (Bio_1), annual precipitation (Bio_{12}), precipitation of wettest month (Bio_{13}), precipitation of driest month (Bio_{14}), precipitation seasonality (Bio_{15}) (Fick and Hijmans, 2017).

2.2. Caldas Classification

30 The Caldas classification relates the variability of the mean annual temperature with the altitude, making it only valid for regions located in the tropical zones (IDEAM, 2015). Figure B.4 of supplemental material depicts the Caldas classification for Colombia. Bioclimatic categories are provided in Table 2.

Table 1: Caldas classification. Abbreviations: T, temperature

Thermal floor	Altitude range	Temperature range
Warm	0 to 1000	$T \geq 24^\circ C$
Temperate	1001 to 2000	$24 > T \geq 17.5^\circ C$
Cold	2001 to 3000	$17.5 > T \geq 12^\circ C$
Highly cold	3001 to 3700	$12 > T^\circ C$

2.3. Lang Classification

35 Lang classification uses the ratio between the annual precipitation [mm] and mean annual temperature [$^\circ C$] to describe the humidity characteristics of background sites (IDEAM, 2015). This ratio called Lang factor is then used to classify the climate into six categories, see Table 3. Figure B.5 of supplemental material depicts the Lang classification for Colombia.

2.4. Modified Lang Classification

40 We firstly obtained a characterization of the Colombia's precipitation space by applying a principal component analysis (PCA) to the set of precipitation variables, using standard procedures (Abdi and Williams, 2010). We then calculated the Lang factor for each background site and determined its corresponding bioclimatic category.

Table 2: Lang classification. Abbreviations: P, precipitation; T, temperature

Humidity factor	Lang factor (P/T)
Desert	0 to 20
Arid	20 to 40
Semi-arid	40 to 60
Semi-wet	60 to 100
Wet	100 to 160
Highly wet	> 160

Later, we projected all background sites with their bioclimatic categories into the precipitation space, and based on these categories, we sequentially fit logistic regression models to redefine the Lang classification's bioclimatic categories (Bishop, 2006; Hastie et al., 2017). We assumed that two adjacent or closer background sites in the precipitation space should share similar bioclimatic characteristics.

Characterization of precipitation space

For the set of precipitation variables, the first two principal components (PC)s contributed 96.8% of the variability. Fig 1 depicts the loadings of the precipitation variables projected in the latent space spanned by the first two PCs, hereby name precipitation space. It can be observed, for instance, that negative and positive values of the first PC are related to areas with high and low levels of precipitation, respectively. Meanwhile, positive values of the second PC are related to areas with high variability of monthly precipitation over the course of the year. Fig 2 depicts the projection of all background sites in the precipitation space with their corresponding bioclimatic categories. Then, it can be observed that background sites classified as highly wet (blue dot points) were distributed along the first PC: some highly wet sites were even close to dry and desert categories, see green and red colors in Figure 2, respectively.

Classification of precipitation space

We sequentially fit a set of classifiers to divide the precipitation space in six categories. We used the Lang classification's categories as labels to define a multi-class classification problem. All classifiers were fit by using regularized logistic regression, see details in the supplement material. We described as follows the step performed to derive a modified Lang classification :

1. Based on the Lang classification, we fit a logistic model to classify dry (class 0) and desert (class 1) background sites. Then, this model was used to classify all background sites between 0 or 1. We redefined the bioclimatic category of desert with all background sites assigned to the class 1.
2. Based on the semi-dry, dry, and redefined desert categories, we fit a logistic model to separate semi-dry background sites (class 0) from dry and desert background sites (class 2). Then, this model was used to classify all background sites between 0 or 2. We redefined the bioclimatic category of dry with all background sites assigned to the class 2 and class 1 (from step 1).

Figure 1: Projection of the precipitation variables in the first two PCs. Abbreviations: P, precipitation.

Figure 2: Projection of all background sites in the first two PCs with their corresponding categories in the Lang classification. This projection represent the characterization of the Colombia's precipitation space, whose axes are given by the first two PCs. Abbreviations: P, precipitation.

Figure 3: Projection of all background sites in the first two PCs with their corresponding redefined categories. Logistic regression models were used to classify all background sites into the six categories. Abbreviations: P, precipitation.

3. We then fit two logistic regression models to separate semi-wet background sites (class 0) from semi-dry, dry, and desert background sites (class 3). Then, these models were used to assign all background sites to class 0 or 3. We redefined the bioclimatic category of semi-dry with all background sites assigned to the class 3 and class 2 (from step 2).
4. Similarly, we fit two logistic regression models to separate wet background sites (class 0) from semi-wet, semi-dry, dry, and desert background sites (class 4). Then, these models were used to assign all background sites to class 0 or 4. We redefined the bioclimatic category of semi-wet with all background sites assigned to the class 4 and class 3 (from step 3).
5. Finally, we fit a logistic model to separate highly wet background sites (class 0) from wet and semi-wet background sites (class 5). Then, this model was used to assign all background sites to class 0 or 5. We redefined the bioclimatic category of wet with all background sites assigned to the class 5 and class 4 (from step 4). Similarly, we redefined the bioclimatic category of highly-wet with all background sites assigned to the class 0.

Parameters tuned for the classifiers are provided in Table S1. Meanwhile, Figure 3 depicts the projections of all background sites in the precipitation space with their corresponding new redefined bioclimatic category. Similarly, Figure B.6 in the supplemental material depicts the modified Lang classification for Colombia by re-projecting the background sites to the geographical space. As a result, it can be observed that highly-wet background sites were mostly located in the Pacific region of Colombia,

which has high-level of annual precipitations. Meanwhile, for instance, areas like
 100 Sierra Nevada de Santa Marta or Parque Nacional Natural el Cocuy were recatego-
 rized as semi-wet, semi-arid, and arid.

Table 3: Lang classification. Abbreviations: P, precipitation; T, temperature

Humidity factor	Lang factor (P/T)
Desert	0 to 20
Arid	20 to 40
Semi-arid	40 to 60
Semi-wet	60 to 100
Wet	100 to 160
Highly wet	> 160

2.5. Modified Caldas-Lang Classification

Temperature and humidity factor are two important covariates used to describe the
 climate in tropical regions (IDEAM, 2015; Medina Bermudez and Aldana Buitrago,
 105 2019). Caldas-Lang classification is derived by combining both Caldas and Lang
 Classification. Categories are defined by taking a first label from the Caldas clas-
 sification and a second label from the Lang Classification. We followed the same
 procedure to derive a modified Lang-Caldas classification by integrating the Caldas
 and modified Lang classification. Figure B.7 in the supplemental material depicts the
 110 modified Caldas-Lang classification for Colombia.

3. Declaration of competing interest

The authors declare no conflict of interest associated with this study.

4. Acknowledgements

This research was supported by the Departamento Administrativo de Ciencia, Tec-
 115 nología e Innovación (COLCIENCIAS) – Colombia with the “ Programa de Estancias
 Postdoctorales Beneficiarios COLCIENCIAS 2017 ” grant.

References

- Medina Bermudez AE, Aldana Buitrago AE. Anlisis comparativo de las zonifica-
 ciones climticas de Caldas-Lang y Holdridge, con las zonificaciones del clima ed-
 120 ficio del estudio semidetallado de suelos, en la cuenca del rio Cauca, Departamento
 del Valle del Cauca. Tesis de Maestria, Universidad Santo Tomas, Bogot; 2019.
- IDEAM. Clasificacion de los climas. 2015. URL:
<http://atlas.ideam.gov.co/basefiles/clima-text.pdf>.

IDEAM . Clasificaciones climaticas Colombia. 2011.

125 Fick SE, Hijmans RJ. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 2017;37(12):4302–4315.

Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2010;2(4):433–459. doi:10.1002/wics.101.

130 Bishop CM. Pattern Recognition and Machine Learning. D ed.; USA: Springer New York; 2006. URL: <http://www.springer.com/computer/image+processing/book/978-0-387-31073-2>.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2d edition ed.; Springer Series in statistics; 2017.

Appendix A. Regularized logistic regression model

135 Considering a set of n records represented in a M -dimensional feature space as w_i ($i = 1, \dots, n$), the goal was to determine a logistic regression model that optimally separate records between class 1 and 0. This model is described as follows:

$$f(w) = \frac{1}{1 + e^{-\theta^T w}} \quad (\text{A.1})$$

where $\theta = [\theta_1, \dots, \theta_m]^T$ was the vector of model parameters, and $w = [w_0, \dots, w_m]^T$ was the vector of features, holding that $w_0 = 1$. The function f then predicted the probability that $y = 1$ on the input vector w , i.e., $f(w) = P(y = 1/\theta, w)$. As a result, 140 given a feature vector w , the function f assigned it to the class 1 if $\theta^T w \geq 0.5$ while to class 0 otherwise.

For a training set $(w^{(i)}, y^{(i)})$, where $i = 1, \dots, n$, we sought to find the vector of parameters θ and λ that minimize the following cost function:

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n y^{(i)} \log(h_{\theta}(w^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(w^{(i)})) + \frac{\lambda}{2n} \sum_{j=1}^m \theta_j^2 \quad (\text{A.2})$$

145 where λ is the regularization parameter. The following table provides the parameters tuned for the logistic regression models.

Table A.4: Parameters of the logistic regression models used to derive the modified Lang classification.

Parameter	Step 1 model 1	Step 2 model 2	Step 3 model 3 model 4		Step 4 model 5 model 6		Step 5 model 7
θ_0	-18.1	-22	1.7	-2.6	-0.45	0.4	5.8
θ_1	4.9	-7.8	0	0.88	1.4	0	2.1
θ_2	1.6	0.7	0	1	1	0	1.0
λ	1	1	0	1	1	0	1

Appendix B. Bioclimatic classifications

All bioclimatic classification are available at the Zenodo repository <https://doi.org/10.5281/zenodo.5396516>

Figure B.4: Caldas classification

Figure B.5: Lang classification

Figure B.6: Modified Lang classification

Figure B.7: Modified Caldas-Lang classification