# Geospatial Analysis with Python and R for Land Suitability Assessment in Agriculture. Case study: La Unión - Valle del Cauca

**Cristian Fabian Gallego-Martinez, Jorge Andrés Jola-Hernández, Mario Alejandro Celis-Torres**

## 1. Introduction

The municipality of La Unión is located in the department of Valle del Cauca, Colombia. It borders Toro to the north, Roldanillo to the south, the Cauca River and the municipalities of La Victoria and Obando to the east, and El Dovio and Versalles to the west. Its total area is 125 km², of which 2.81 km² corresponds to the urban area and 122.19 km² to the rural area. Agriculture is the main economic activity in the municipality, with notable cultivation of grapes, passion fruit, various fruit crops, and sugarcane (Quillas, 2024).

Land-use conflict is a significant issue in La Unión, where the current use of land does not align with its environmental, ecological, cultural, social, and economic potential and constraints. The Regional Autonomous Corporation of Valle del Cauca (CVC), in collaboration with the municipal government (Villaquiran G, 2024), conducted a study to identify areas where vegetation cover or land use diverges from the physical conditions of the soil. This study revealed that 5,700 hectares, equivalent to 47% of the municipal area, show a high degree of land-use conflict, while 2,048 hectares, representing 17% of the municipal area, exhibit moderate conflict. In total, over 50% of the municipality faces land-use issues, with cases where conservation-oriented lands are being used for agricultural or forestry activities, increasing the risk of contamination due to land-use conflicts (IGAC, 2012).

Geographic Information Systems (GIS) using python has become an essential tool for managing and analyzing spatial information (Bolstad P, 2016). The rapid development of GIS, along with the emergence of methodologies based on artificial intelligence and deep learning, as well as increased access to spatial data, has positively impacted problems involving land use and land cover. In a Bayesian characterization study of urban land-use configurations using very high-resolution (VHR) remote sensing images, Li et al. (2020) integrated spatial disposition and composition variables for urban land-use extraction. The results showed that the proposed method produced urban land-use extractions with comparable or better accuracy than existing methods, achieving 86% and 93% precision, surpassing existing methods' 83% and 88%.

Zoungrana et al. (2023) developed a methodology to leverage optical and radar time-series images for estimating wheat cultivated areas before the harvest period with an accuracy of 84%. In coastal areas, the application of convolutional neural networks combined with object-based image analysis improved classification accuracy, producing final maps for regional and national decision-making with overall accuracy values of 93.5% using Pléiades satellite images (Zaabar N et al., 2022).

Additionally, Pachón et al. (2018) developed a raster geographic viewer and profile for the Regional Autonomous Corporation of Valle del Cauca (CVC). Using digital elevation models and other GIS tools, the viewer facilitates geospatial data visualization and analysis, supporting environmental planning and management. The implementation used Python,

ArcGIS, and geographic web services to enhance accessibility and utility for decision-making. Jaramillo, F. (2024) created a geographic web portal to support water resource planning in La Unión by combining geospatial information from various sources with a multi-criteria analysis methodology. This initiative identified strategic areas for water conservation. Salazar, A. (2021) developed a monitoring and management tool for soils and zoning for the Roldanillo, La Unión, and Toro irrigation district in Valle del Cauca, integrating 208 spatial entities into a spatial database.

Spatial analysis within GIS enables the collection, management, analysis, and presentation of geospatial data, including sampling methods, visualization, representation, and spatial component analysis (Pu Hao, 2019). The location and attributes of spatial objects are critical in such analyses, as results depend on the positions and relationships of the analyzed objects. Data serve as the starting point for spatial analysis, and their quality depends on the tools and methods used for collection. In the past, obtaining data was time-consuming and costly, involving extensive fieldwork. Today, satellite images from platforms like Sentinel 1 and 2, processed in the Google Earth Engine geospatial analysis platform with 10 m spatial resolution, are readily available online.

Final data analyses have been conducted using programming languages like Python due to its versatility and libraries such as *sklearn*, *matplotlib*, *pandas*, *geopandas*, *keras, rasterio*, *pyidw*, and *geometry*. Panyadee P. (2004) used these tools to generate a flood risk map in northern Thailand.

This project aims to conduct an advanced spatial analysis using technological tools, including specialized programming languages like Python. The primary objective is to develop well-grounded recommendations to promote sustainable land management, contributing to the optimization of land productivity based on variables derived from diverse data sources.

## 2. Area of interest

The eastern sector of La Unión lies in the alluvial plain of the lower Cauca River valley, with soils primarily used for agriculture, supported by the RUT irrigation district (1958-1966), covering 10,300 hectares (Figure 1). Moving westward towards the Pacific, the landscape transitions into foothills with a dry warm climate (IGAC, 2012), progressively increasing in slope and reaching mountainous terrain (1,600-1,700 m.a.s.l.). The climate becomes moderately humid, and the steeper slopes (70%-100%), along with a greater presence of forest trees, shift the land's suitability towards conservation (Villaquirán G., 2024).

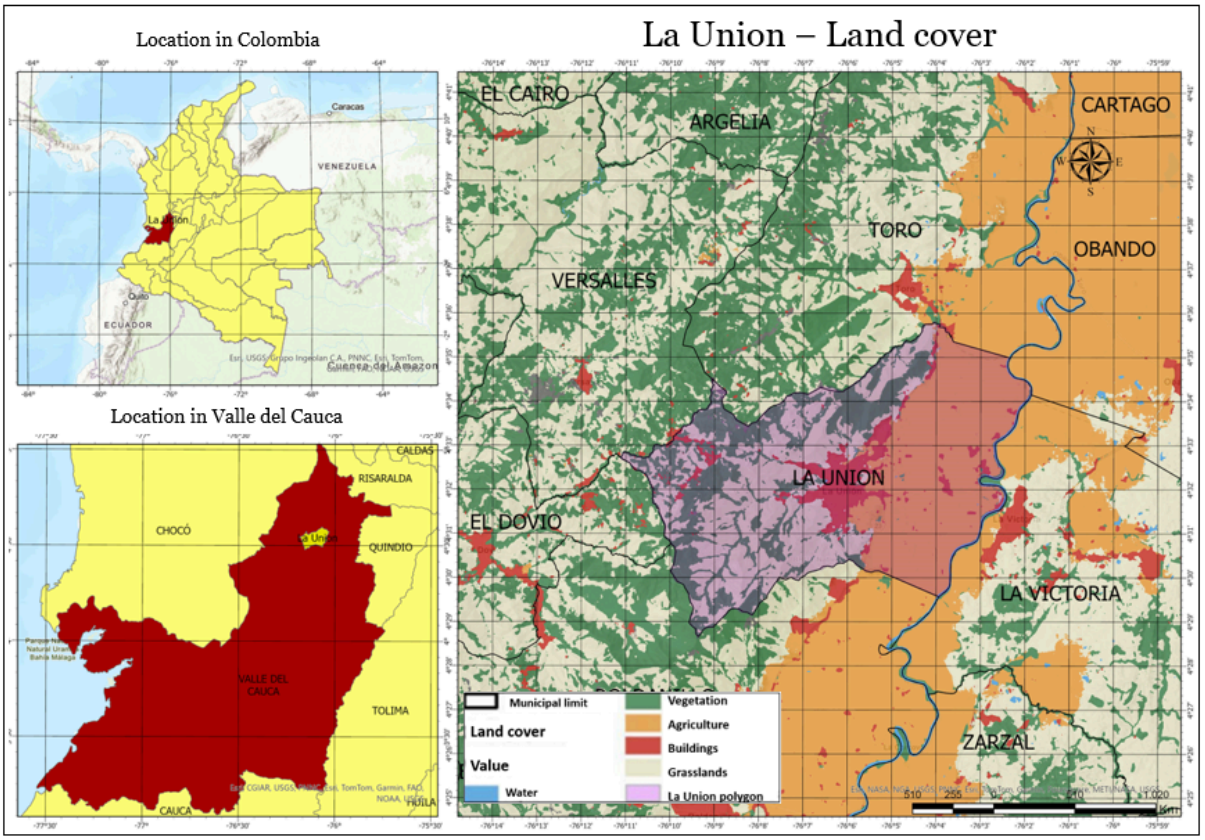



**Figure 1.** Research area.

## 3. Objectives

### General Objective:

- Implementing Operations and Algorithms Using Python for Geospatial and Edaphoclimatic Analysis to Estimate Agricultural Land Suitability in the Municipality of La Unión.

### Specific Objectives:

- Execute an algorithm to identify conflict areas between land use and land suitability through vocation and coverage maps.
- Develop and implement processes in Python to estimate Agricultural Land Suitability in plots, considering factors such as edaphoclimatic information, slopes, water access, and proximity to main roads, using GIS programming techniques.
- Generate code to identify properties affected by mining areas and calculate the distance of the properties to the main roads.
- Automate the estimation of agricultural activity viability on properties in the municipality of La Unión using a Python and R workflow, integrating it into an accessible information system (Web Interface).

## 4. Methodology

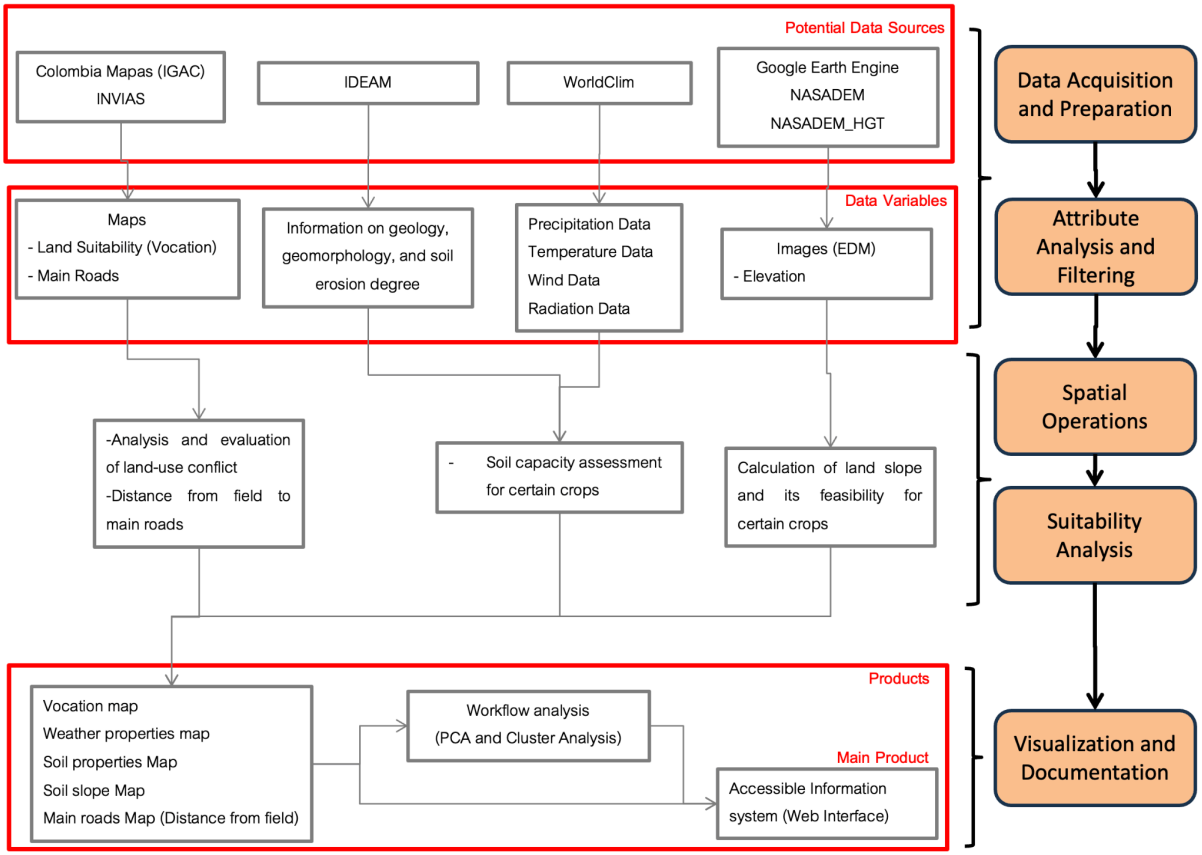



**Figure 2.** Methodology flux diagram.

### 4.1 Datasets

**IGAC:** Land use zoning maps from the national level in the "Colombia Mapas" web geographic service.

**INVIAS:** Cartographic information of the National Road Network of Non-Concessioned Highways under INVIAS, concessioned roads under ANI administration and secondary roads (Regional Road Plan Program - PVR)

**IDEAM:** Data from monitoring and tracking of soils and lands, with coverage of the National Land Cover Legend and CORINE Land Cover methodology adapted for Colombia, scale 1:100.000.

**WorldClim:** Climatic information with a spatial resolution of 30 meters for the year 2020.

**Google Earth Engine:** Extraction of a digital elevation model with a spatial resolution of 30 meters.

### 4.2 Operations

#### 4.2.1 Mining impact of each plot.

To estimate properties affected by mining, a vector file containing points of mining activity areas was used. A 500-meter buffer was generated ('miner.buffer(500)'), and this information was saved as a GeoPandas object. Then, properties intersecting with these "mining areas" were selected and saved in a new .geojson file.

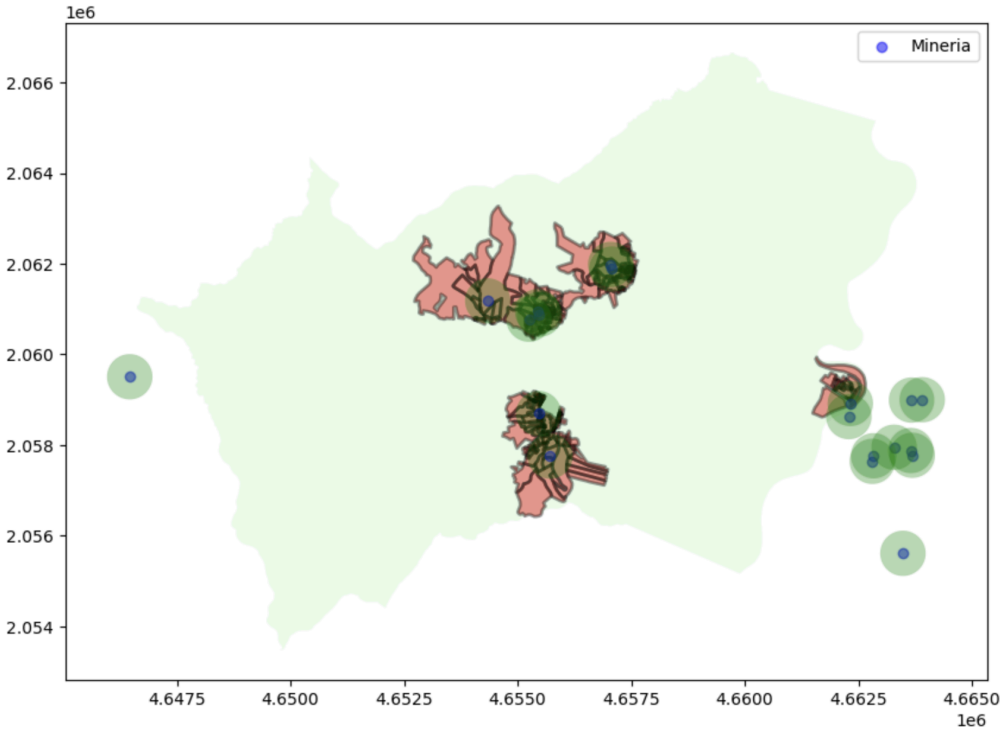



**Figure 3.** Mining impact in La Union municipality.

### 4.2.2 Slope of each plot.

To estimate the slope of each property, the elevation raster file, shown in Figure 4, was opened using rasterio. The elevation data was read and the slope was calculated by applying the gradient to the elevation values. The slope was then computed using the arctangent of the gradient and converting it to degrees.

Next, the properties were checked to ensure their coordinate reference system (CRS) matched the CRS of the raster file. If necessary, the properties were transformed to the raster's CRS. For each property in the dataset, the geometry was extracted, and a mask was applied to the elevation raster using a mask to focus on the area corresponding to the property. The valid (non-NaN) values within the masked area were used to calculate the average slope. This value was appended to a list.

Finally, the calculated slopes were added to the properties dataset, and the results were saved to a new .geojson file.

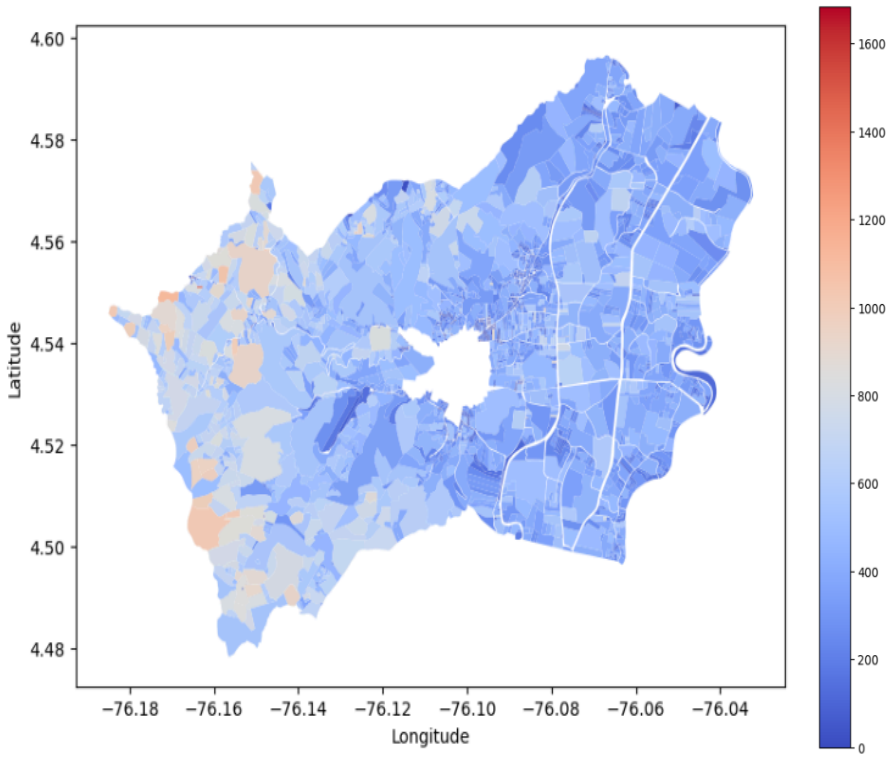



**Figure 4.** Properties elevations.

### 4.2.3 Land vocation of each plot

To estimate the land vocation for each property, the coordinate reference systems (CRS) of the property dataset and the land vocation dataset were checked. If the CRS of the two datasets didn't match, the vocation dataset was transformed to match the CRS of the property dataset.

Then, a spatial join was performed to assign the land vocation value from the vocation dataset to the properties in the property dataset that intersected with the land vocation areas. The result was saved in a new .geojson file, as shown in Figure 5

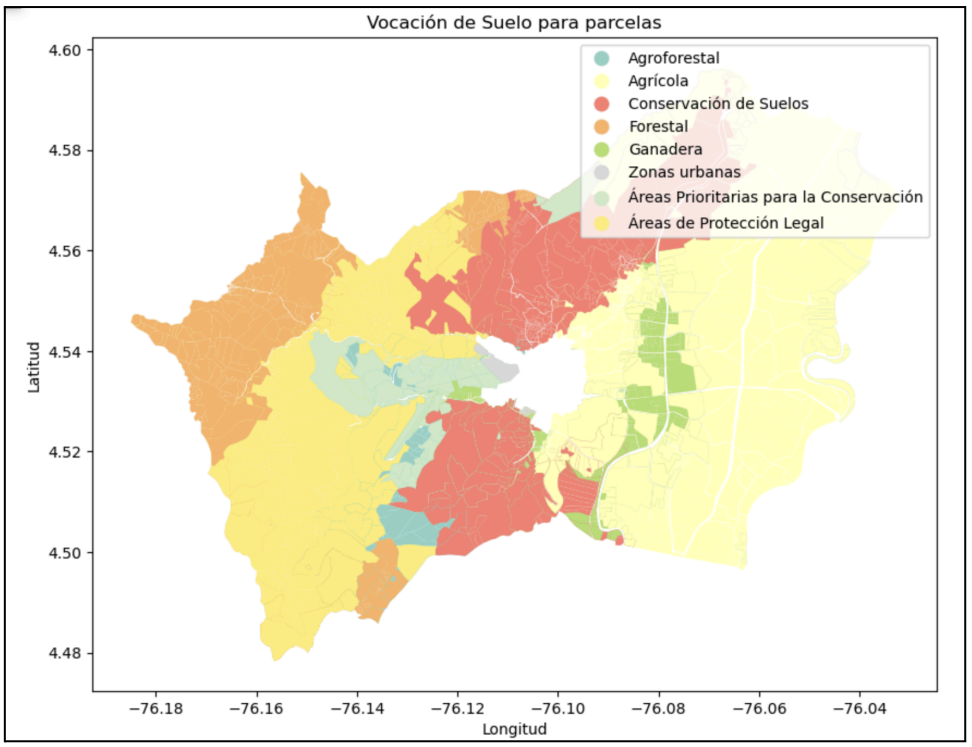
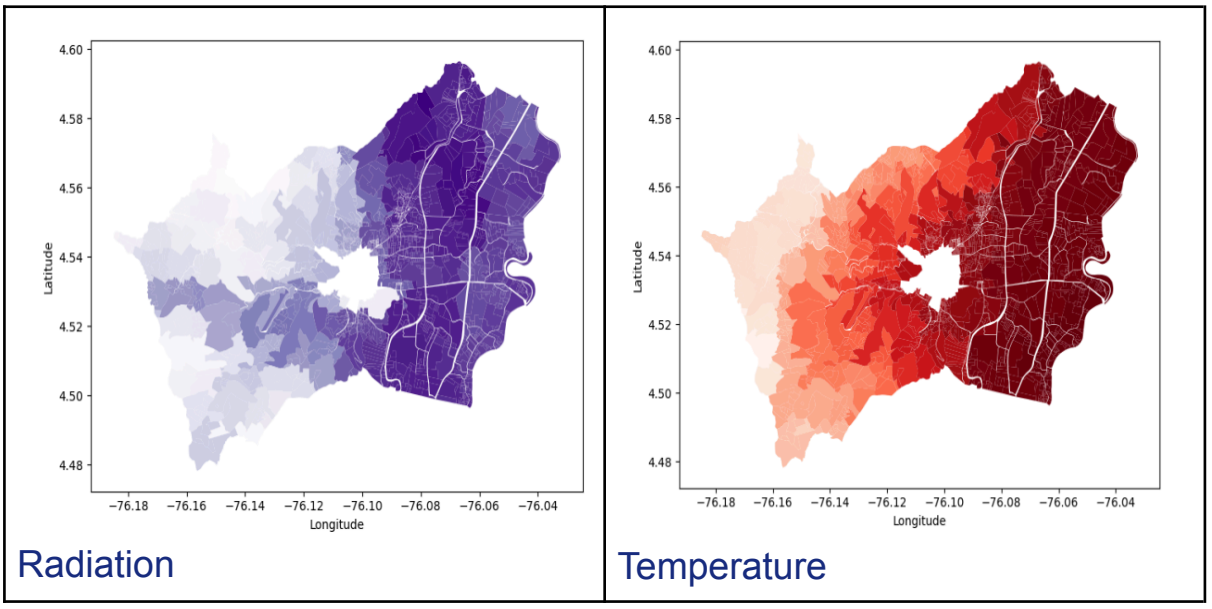



**Figure 5.** Properties Land vocation.

#### 4.2.4 Edaphoclimatic properties of each plot.

The centroids of the properties were used to extract the values of climatic variables (temperature, precipitation, wind speed, vapor pressure, and radiation) from the corresponding rasters. These values were obtained at the centroid coordinates and added to a new field in the properties' GeoDataFrame. The result was saved in a GeoJSON file for further analysis and visualization.



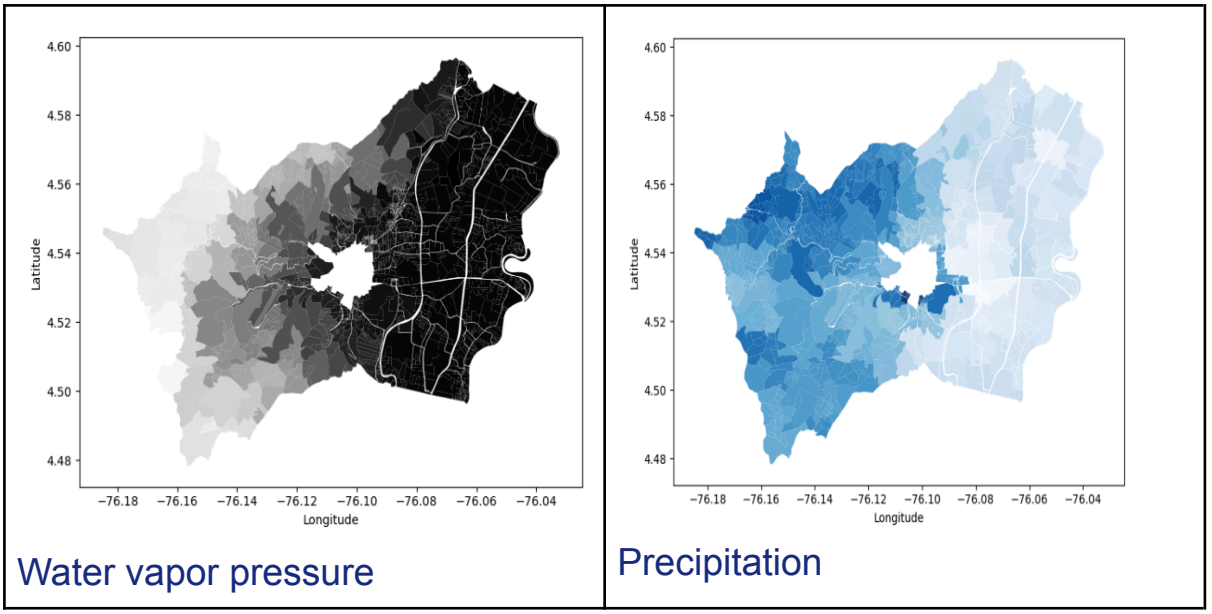



**Figure 6.** Edaphoclimatic properties of each plot.

#### 4.2.5 Distance to main roads of each plot.

Determining the distance from each parcel to the nearest road is essential in agriculture to optimize product transport, reduce logistics costs, improve access to inputs and machinery, and improve connectivity to markets and distribution centers. To calculate these distances, parcel and road datasets were loaded as GeoDataFrames using the Python GeoPandas library, both datasets were projected to the Magna-SIRGAS / UTM zone 18N (EPSG:3116) coordinate reference system to ensure accuracy of the distance measurements.

The centroid of each plot was calculated and the minimum distance from each centroid to the nearest road was determined. These distance values were added to the parcel dataset and a new shapefile was saved with the updated properties. Finally, a map was generated to visualize the distance distribution, with plots color-coded according to their proximity to the nearest road and roads highlighted in yellow.

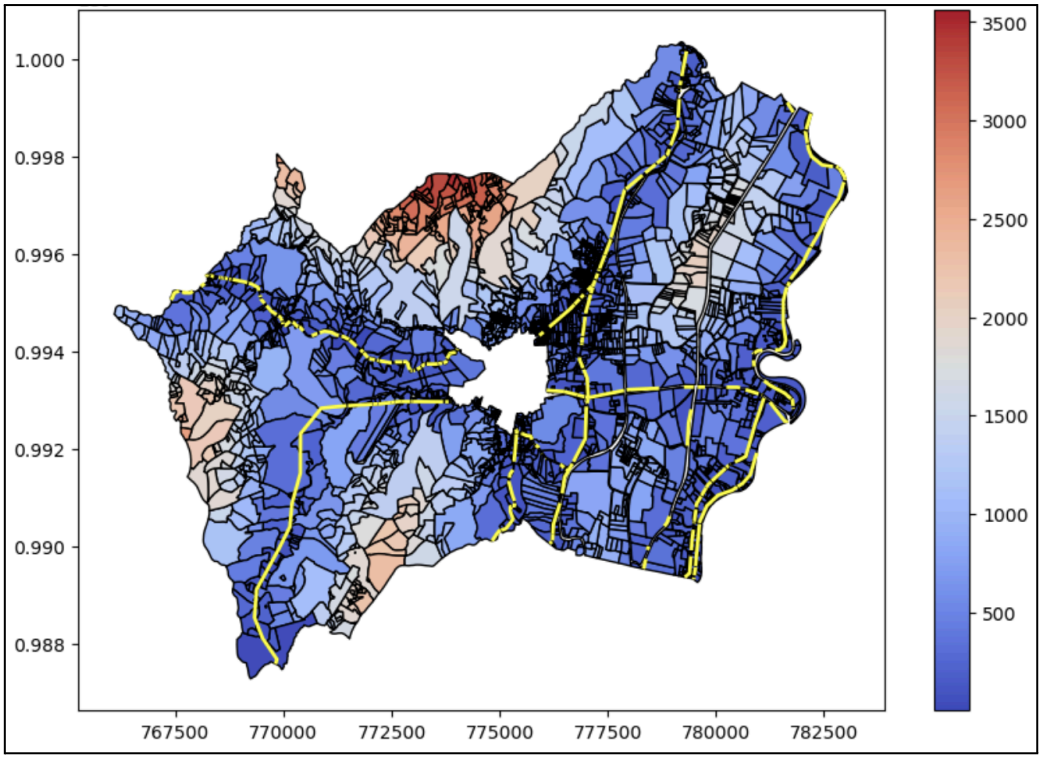



**Figure 7.** Main roads distances.

## 5. Results and Discussion

To reduce dimensionality, a Principal Component Analysis (PCA) was conducted using the edaphoclimatic variables from the study. These variables include pH, slope, radiation, temperature, water vapor, effective depth, wind speed, and precipitation. To ensure comparability and eliminate scale differences, all variables were standardized using the Z-score method, which transforms data to have a mean of zero and a standard deviation of one. This standardization allows for a more accurate assessment of variable contributions in the principal components, improving the interpretation of underlying patterns in the dataset.

### 5.1 Principal Components Analysis

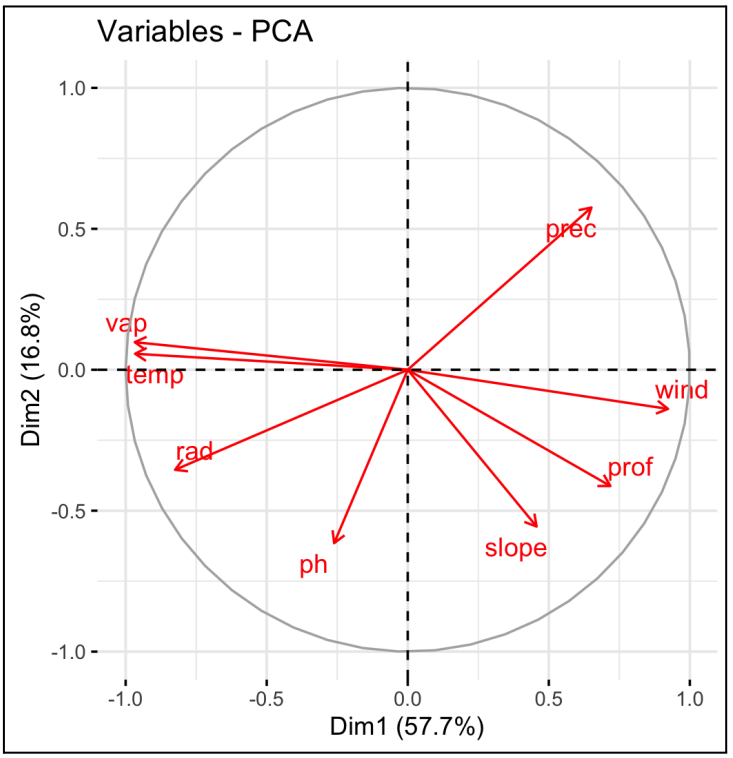



**Figure 8.** Principal Components Analysis.

This analysis explains how different environmental variables contribute to the variation in the dataset. The first dimension (Dim1) accounts for 57.7% of the total variance, making it the most significant factor in distinguishing data points. The second dimension (Dim2) explains an additional 16.8% of the variance, providing further insight into secondary patterns. Together, these two dimensions summarize a large portion of the variability in the dataset.

The variables precipitation, wind, depth, and slope strongly influence Dim1, indicating that this dimension captures differences related to topographic and climatic conditions. Meanwhile, temperature and water vapor are more aligned with Dim2, suggesting that this component is linked to atmospheric factors. pH and radiation also contribute but have a lesser impact on the primary dimension. Variables pointing in similar directions are positively correlated, while those in opposite directions have a negative relationship.

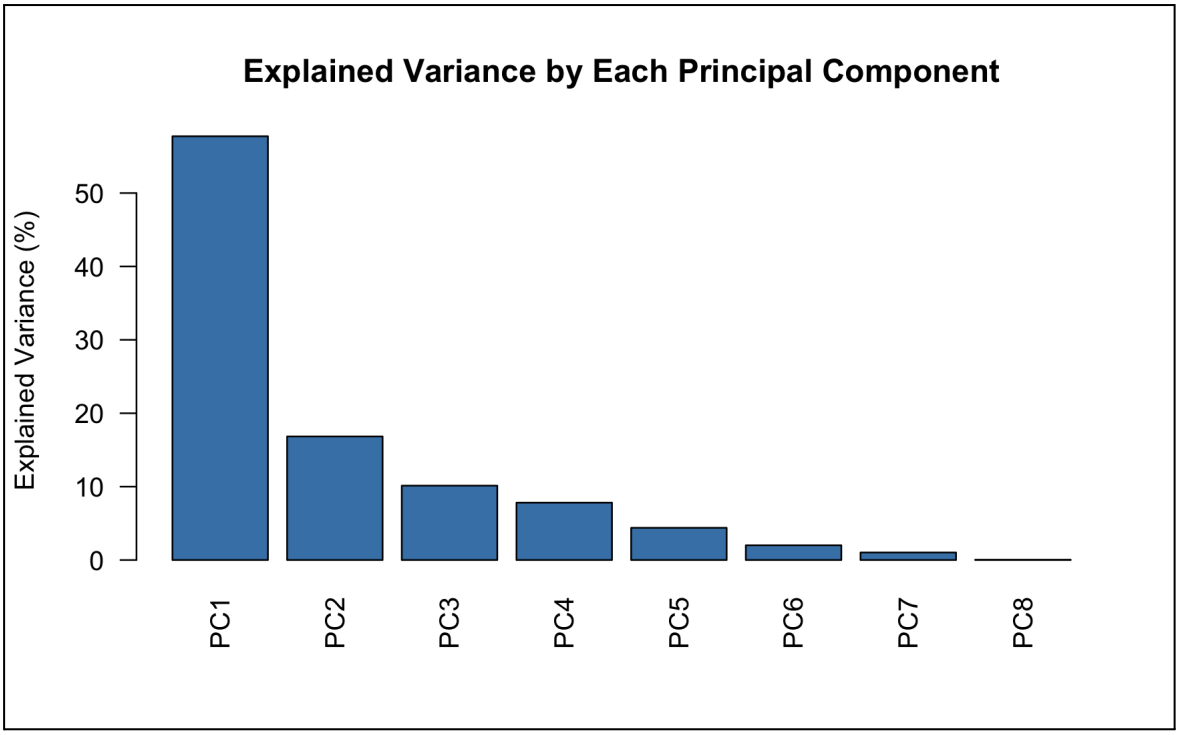
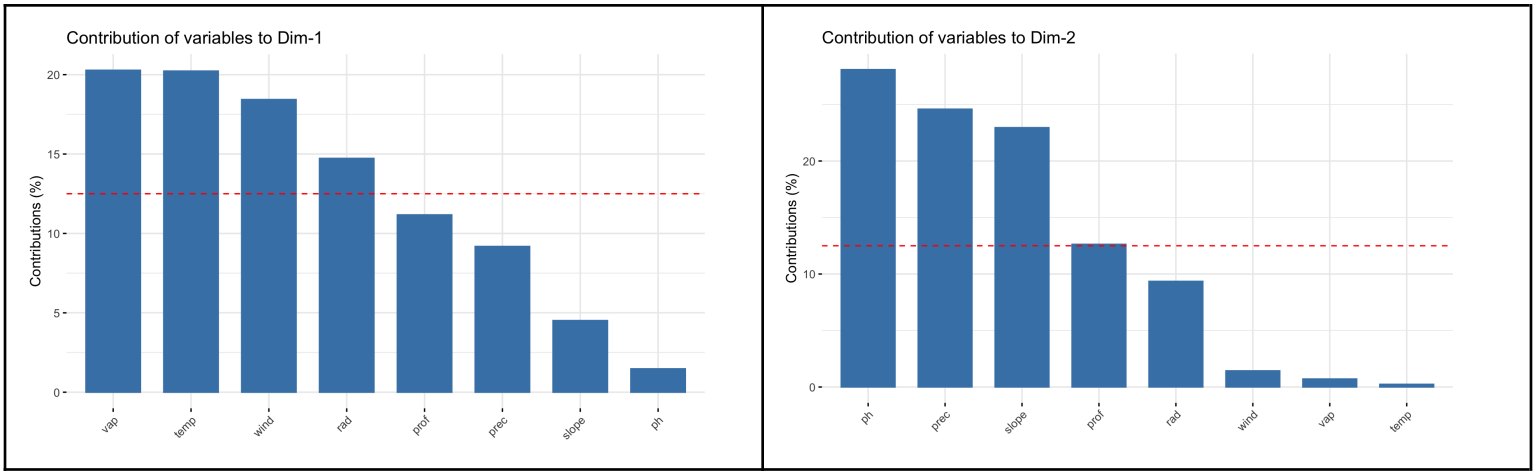



**Figure 9.** Explained variance by each principal component**.**

The first three principal components were selected, explaining 84.71% of the total variation in the data. This selection allows for a significant reduction in dimensionality while retaining most of the original information, facilitating data interpretation and further analysis in agricultural suitability assessment.



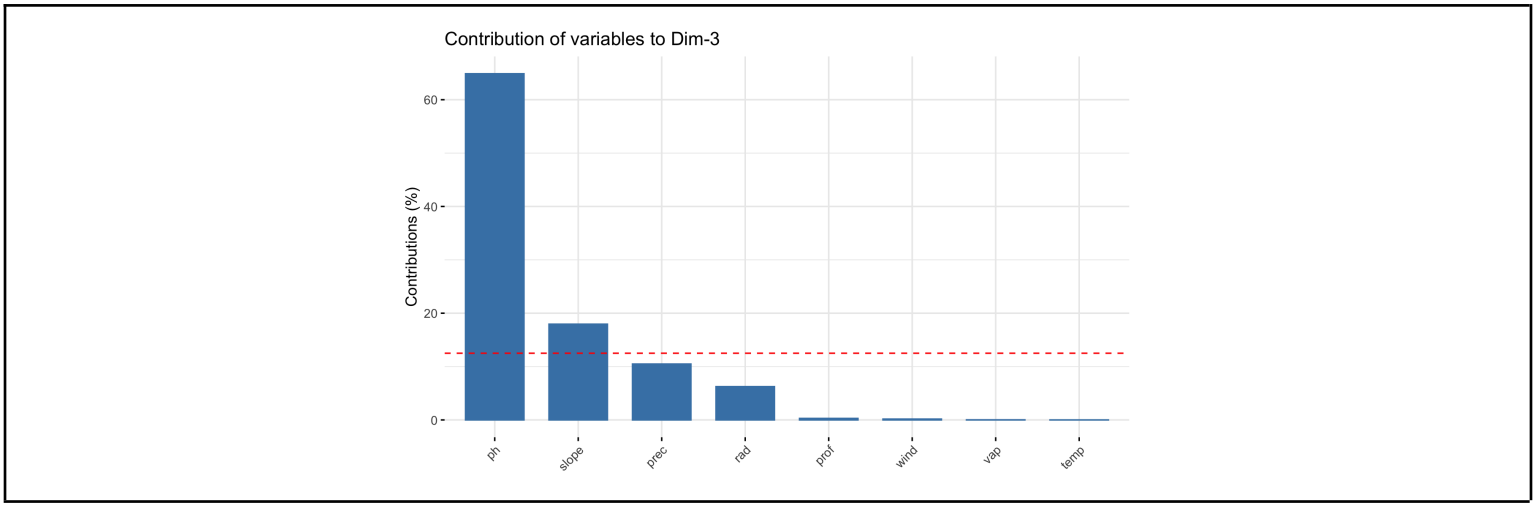



**Figure 10.** Contribution of variables.

The principal component analysis (PCA) results reveal the relative contributions of different edaphoclimatic variables to the main dimensions of variation. The first principal component (Dim-1) is primarily influenced by water vapor, temperature, and wind speed, followed by radiation and effective soil depth. These variables have the highest percentage contributions, indicating their significant role in explaining the overall variability in the dataset.

The second principal component (Dim-2) is mainly associated with soil pH, precipitation, and slope, which contribute the most to this dimension. Effective depth and radiation also play a role, albeit to a lesser extent. Meanwhile, variables such as wind speed, water vapor, and temperature have minimal influence on this component, suggesting that Dim-2 captures variability related more to soil and precipitation characteristics rather than atmospheric factors.

The third principal component (Dim-3) is strongly dominated by soil pH, which has an overwhelmingly high contribution compared to other variables. Slope and precipitation also contribute notably, while radiation, effective depth, and atmospheric variables such as wind speed and temperature have negligible impact. This indicates that Dim-3 largely reflects variations in soil chemical and topographic properties rather than climatic influences.

### 5.2 Spatial Non Supervised Classification

The mixing parameter alpha was established to determine its influence on the clustering process, specifically in weighting the importance of geographic coordinates. In this case, the geographic component is represented by the distances between the centroids of all plots. To achieve this, two matrices were used: the matrix of explanatory variables (D0), which consists of the first three principal components obtained from PCA, and the distance matrix (D1), which accounts for spatial distances.

A balance point between D0 and D1 was set at approximately 0.8 to ensure that both matrices contribute similarly to the explained inertia. This balance helps maintain an optimal proportion between environmental and spatial factors, leading to a well-calibrated alpha

value for clustering. By selecting this value, the clustering process integrates both edaphoclimatic variables and spatial distribution effectively.

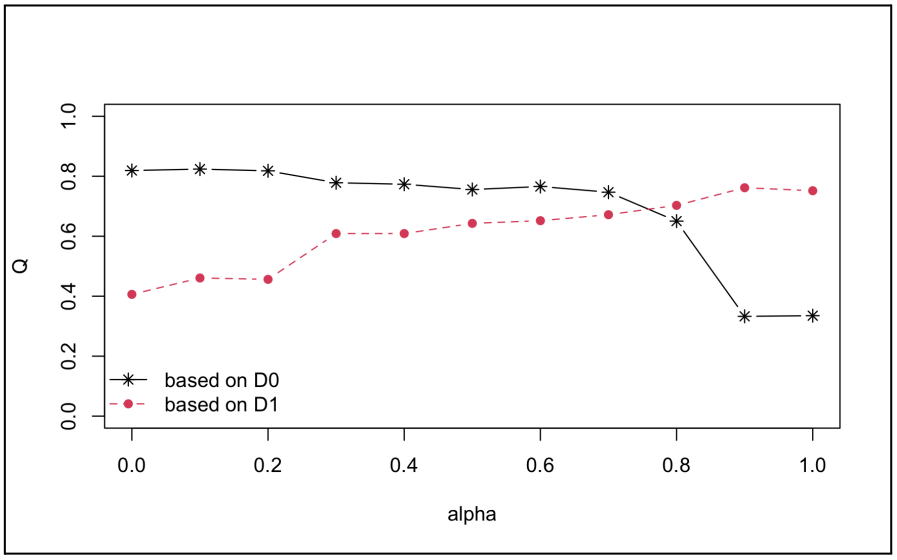



**Figure 11.** Non Supervised Classification.

With an alpha of 0.8, five plot clusters were created using the Ward clustering algorithm. Figure 12 illustrates the resulting clusters. A graphical web interface was developed to allow users to test different numbers of clusters and visualize additional variables, such as plots affected by landmines and the proximity of plots to major roads.

Link: Web page link

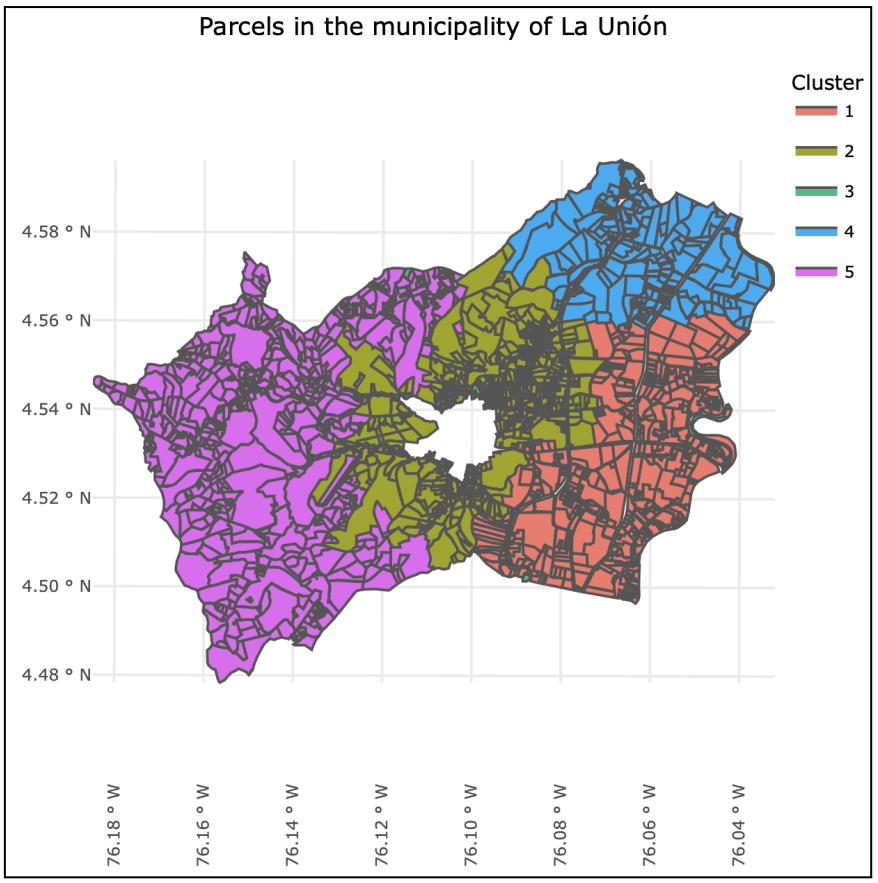



**Figure 12.** La Unión clustered plots

### 5.3 Parcel suitability

Table 1 presents the soil and climate characteristics, according to each identified cluster in La Union municipality as shown in Figure 12. There are 5 clusters, each of them having specific values regarding the 6 edaphoclimatic variables. Soil pH is an important indicator of the suitability of a soil for plant growth. For most crops, pH of 6 to 7.5 is optimal. For all clusters the pH is suitable for agriculture, except for cluster 3. Effective depth refers to the soil conditions that allow good growth and development of the roots in depth, ensuring that roots don't obstruct their growth and have a good formation (Martínez L, 2006).

Climate variables influence agricultural productivity and agricultural management decisions. For clusters 1 to 4, average temperature values are similar, showing that there is not a high elevation gradient. In cluster 5, temperature is lower, indicating altitude on this sector is higher. Total precipitation has a more heterogeneous behaviour in La Union municipality. The highest rainfall occurs in cluster 5, located in the west of the municipality, where slopes are the highest. On the other hand, clusters 1 to 4 registered lower rainfall values, where slopes are also lower.

**Table 1.** Clusters climate and soil characteristics.

| Cluster | pH | Effective depth (m) | Radiation | Slope | Temperature | Precipitation |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| **1** | 7.068815 | 0.351220 | 8.652698 | 18.192388 | 24.260627 | 1542.648084 |
| **2** | 6.508405 | 0.122478 | 8.591060 | 8.617913 | 23.764361 | 1778.917387 |
| **3** | 0.000000 | 0.000000 | 8.551484 | 6.088423 | 23.732143 | 1910.423469 |
| **4** | 6.716599 | 0.502834 | 8.672764 | 16.456989 | 24.024697 | 1542.469636 |
| **5** | 5.502706 | 0.902165 | 8.342840 | 26.968732 | 20.398376 | 2090.193505 |

The soil and climate information presented in the table above is useful for identifying the suitability of each cluster for agricultural activities. In order to know in which cluster it is possible to establish a specific crop, a relationship can be traced between the environmental offer (Table 1), and the requirements of 5 selected crops from a list of crops present in La Union municipality. According to the Municipal Agricultural Assessments carried out by the Ministry of Agriculture and Rural Development, there are 28 crops present in the municipality of La Unión in 2018, of which 5 were selected based on their economic importance for the municipality. Table 2 shows these crops and their soil and climate requirements.

**Table 2.** Edaphoclimatic requirements of the 5 selected crops.

| Variable | Sugarcane | Pineapple | Pawpaw | Tomato | Melon |
|---|---|---|---|---|---|
| **pH** | 5.5 - 7.5 | 5 - 6 | 5.5 - 7.5 | 6 - 6.5 | 6.0 - 7.5 |
| **Effective depth (m)** | 1 | 0.6 | 0.5 | 0.4 | 1.5 |
| **Elevation (amsl)** | 400 - 1.300 | 0 - 1200 | 0 - 1000 | 100 - 1500 | 0 - 1.200 |
| **Slope (%)** | 5 - 10 | 5 - 10 | 3 - 7 | 1 - 2 | < 25 |
| **Temperature (°C)** | 25 - 30 | 22 - 30 | 18 - 38 | 18 - 30 | 18 - 25 |
| **Precipitation (mm/year)** | > 1500 | 1500 - 3500 | 1500 - 2000 | 1500 - 2000 | 500 - 1.500 |
| **Relative humidity (%)** | 80 - 85 | N/A | 60 - 85 | 60 - 80 | 55 - 75 |
| **Light (h/day)** | 6 a 9 | 4 | 6 | 6 | 6-8 |

According to the Colombian Ministry of Agriculture (Ministerio de Agricultura y Desarrollo Rural), in 2018, 188,002 hectares of sugarcane were planted in the Valle del Cauca department, which represents 78.5% of national production. The municipality of La Unión is one of the 31 municipalities where sugarcane is grown, so it is important to establish which areas are suitable for its cultivation. Sugarcane requirements were taken from CONADESUCA (2015). Regarding soil pH, all clusters are suitable, whereas Effective depth is not. The closest is cluster 5 with 0.9 m. However, cluster 5 is not suitable for temperature and for slope. Cluster 2 seems to be the most suitable for sugarcane cultivation, however Effective depth should be improved.

Pineapple crop in Valle del Cauca represents 18% of total nation production (MADR, 2018). Thanks to the development and progress that this agricultural chain has in this region, the

cultivated area has increased in the last decade (MADR, 2018). Temperature is the most important factor in its production, ranging from 22 to 30 °C (Garzón J, 2016). All clusters, except for the number 5, are suitable for its cultivation regarding temperature. Cluster 4 has the best suitability for its cultivation.

Pawpaw crop has a large temperature range from 18 to 38 °C (García M, 2010), hence it can be cultivated in all clusters. The same applies for pH values. Regarding temperature, cluster 5 can be discarded, as it is more than 2000 mm. If slope is considered, just clusters 2 and 3 can be included. For tomato crop, the most suitable clusters are 1 and 4, as they fulfil all the requirements. For melon rainfall is a limiting factor, as it is very susceptible to excess moisture, and therefore requires well-distributed rainfall of 500 to 1,500 mm per year (). According to the rainfall values for all clusters, none of them are suitable for melon crop. For this reason it should be evaluated its presence in the municipality.

## 6. Conclusions

- The use of Python modules such as GeoPandas, Rasterio, and Scikit-learn facilitates the management and analysis of spatial data. It enables the automation of key tasks, such as slope estimation, land-use conflict identification, the assessment of parcel accessibility to main roads and the assessment of parcel. The use of GIS programming optimizes these processes, reducing analysis time and improving decision-making accuracy.
- Dimensionality reduction methods, such as Principal Component Analysis (PCA) and spatial clustering analysis, are powerful tools that enhance data interpretation and enable territorial segmentation based on objective and reproducible criteria. While these tools are freely available in R and Python libraries, it is essential to understand their logic and functionality to determine their suitability for specific use cases.
- The results of this study provide key information for land-use decision-making in La Unión. The combination of GIS, machine learning algorithms, and edaphoclimatic data enables the optimization of agricultural planning and helps reduce environmental impact.
- The comparison of crop requirements with the environmental supply is essential to establish the productive potential of a crop. In the present study, the majority of crops have productive potential in the municipality of La Union, however, there are limiting edaphoclimatic variables that reduce the suitability of the soils, hence management tasks are necessary to allow a better establishment of the crops.

## 7. References


- Bolstad, P. (2016). *GIS Fundamentals: A First Text on Geographic Information Systems*. Eider (Press Minnesota).
- CONADESUCA. (2015). Ficha técnica del cultivo de caña de azúcar (*Saccharum officinarum L.*).
- García M. (2010). Guía técnica del cultivo de la papaya. CENTRO NACIONAL DE TECNOLOGIA AGROPECUARIA Y FORESTAL “Enrique Alvarez Córdova”.

- Garzón J. (2016). Establecimiento y manejo de un cultivo de piña en la sede de la asociación de ingenieros agrónomos del llano en Villavicencio. Universidad de los Llanos.
- IGAC. (2012). *Estudio suelos del Territorio Colombiano a escala 1:100.000. Departamento: Valle Del Cauca.* Departamento de agrología. https://geoportal.igac.gov.co/contenido/datos-abiertos-agrologia
- IGAC. (2012). *Estudio de los conflictos de uso del territorio Colombiano a escala 1:100.000*. https://geoportal.igac.gov.co/contenido/datos-abiertos-agrologia
- Jaramillo Urdinola, F. (2024). *Portal Web geográfico para apoyar la planificación del recurso hídrico en La Unión, Valle del Cauca*.
- Li, Mengmeng & Stein, Alfred & de Beurs, Kirsten. (2020). *A Bayesian characterization of urban land use configurations from VHR remote sensing images*. International Journal of Applied Earth Observation and Geoinformation. 92. 102175. 10.1016/j.jag.2020.102175.
- Ministerio de Agricultura y Desarrollo Rural. (2018). Cadena de azúcar 2018.
- Ministerio de Agricultura y Desarrollo Rural. (2019). Cadena de la piña. Dirección de Cadenas Agrícolas y Forestales Junio 2019.
- Pachón, L. E., Enrique, J., & Ojeda, H. (2018). *VISOR GEOGRÁFICO RASTER Y PERFIL PARA LA CORPORACIÓN AUTÓNOMA REGIONAL DEL VALLE DEL CAUCA*.
- Panyadee, P., & Champrasert, P. (2024). *Spatiotemporal Flood Hazard Map Prediction Using Machine Learning for a Flood Early Warning Case Study: Chiang Mai Province, Thailand*. Sustainability, 16(11), 4433. https://doi.org/10.3390/su16114433
- Hao, P. (2019). Spatial Analysis. The Wiley Blackwell Encyclopedia of Urban and Regional Studies.
- Quillas, C. I. L., Rodríguez, H. F. R., & Sabogal, J. R. (2024). Instituciones y desempeño institucional en la región del centro del Valle del Cauca. https://doi.org/10.25100/peu.913
- Salazar, A. (2021). *IMPLEMENTACIÓN DE CUADRO DE MANDO PARA LA GESTIÓN DEL DISTRITO DE RIEGO ROLDANILLO, LA UNIÓN Y TORO, VALLE DEL CAUCA*.
- Villaquiran G. (2024). Plan de desarrollo 2024-2027 municipio La Unión- Valle. Alcaldía municipal.
- Zoungrana, L.E., Barbouchi, M., Toukabri, W. *et al.* Sentinel SAR-optical fusion for improving in-season wheat crop mapping at a large scale using machine learning and the Google Earth engine platform. *Appl Geomat* **16**, 147–160 (2024). https://doi.org/10.1007/s12518-023-00545-4