Machine learning model for the classification of municipalities by illicit crops in Colombia from 2010 to 2020

Modelo de machine learning para la clasificación de municipios por cultivos ilicitos en Colombia de 2010 a 2020

DOI: http://doi.org/10.17981/ingecuc.19.1.2023.05

Scientific Research Article. Fecha de Recepción: 05/12/2022. Fecha de Aceptación: 14/12/2022.

Andrés Eduardo Narváez Figueroa
Universidad Pedagógica y Tecnológica de Colombia. Tunja (Colombia)
andres.narvaezfigueroa@un.org

Gustavo Cáceres Castellano
Universidad Pedagógica y Tecnológica de Colombia. Tunja (Colombia)
gustavo.caceres@uptc.edu.co

Juan Sebastián González Sanabria
Universidad Pedagógica y Tecnológica de Colombia. Tunja (Colombia)
juansebastian.gonzalez@uptc.edu.co

To cite this paper
Narváez Figueroa, G. Cáceres Castellano & J. González Sanabria, "Machine learning model for the classification of municipalities by illicit crops in colombia from 2010 to 2020", *INGE CUC*, vol. 18, no. 2, pp. 47–60, 2023. DOI: http://doi.org/10.17981/ingecuc.19.1.2023.05

Abstract

Introduction— The United Nations Office on Drugs and Crime (UNODC) classifies Colombia as one of the countries where drug trafficking and crime threaten the security, peace and development opportunities of its citizens.

Objective— This article presents the application of the unsupervised K-means classification algorithm to categorize municipalities with coca cultivation presence in Colombia. Methodology- The CRISP-DM methodology was used for data mining, and the PCA (Principal Component Analysis) algorithm was used for the correlation of variables.

Results— Multiple sources of information were used, such as: the number of hectares of coca per municipality, seizures, laboratories destroyed, manual eradication and fumigation, monitored by national institutions, in order to make crosses with socioeconomic and performance variables of the municipalities with coca crops in the period from 2010 to 2020. Based on the classification, the scenarios of each category were analyzed to find scenarios that allow elucidating the dynamics of the territories suffering from this scourge.

Conclusions— It was found that the behavior of coca-producing municipalities responds mainly to 4 groups. It was also found that the municipality of Tumaco in Nariño does not fit into any category since it exceeds the production with respect to the other municipalities.

Keywords— Unsupervised classification; illicit crops; data mining; fight against drugs; cocaine; Colombia

Resumen

Introducción— La Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC) clasifica a Colombia como uno de los países donde el narcotráfico y el delito ponen en riesgo la seguridad, la paz y las oportunidades de desarrollo de los ciudadanos.

Objetivo— Este artículo presenta la aplicación del algoritmo de clasificación no supervisado K-means para categorizar los municipios que tienen presencia de cultivos de coca en Colombia. Metodología: Se hizo uso de la metodología CRISP-DM, propia de la minería de datos, y para la correlación de variables se utilizó el algoritmo PCA (Análisis de Componentes Principales).

Resultados— Se utilizaron múltiples fuentes de información como: el número de hectáreas de coca por municipio, incautaciones, laboratorios destruidos, erradicación manual y fumigación, monitoreadas por la institucionalidad nacional, con el fin de realizar cruces con variables socioeconómicas y de desempeño de los municipios que tienen cultivos de coca en el periodo de 2010 a 2020. Partiendo de la clasificación, se analizaron los escenarios de cada categoría para hallar escenarios que permitan dilucidar las dinámicas de los territorios que sufren este flagelo.

Conclusiones— Se encontró que el comportamiento de los municipios productores de coca responde principalmente a 4 grupos. También se encuentra que el municipio de Tumaco en Nariño no encaja en ninguna categoría ya que excede la producción respecto a los demás municipios.

Palabras clave— Clasificación no supervisada; cultivos ilícitos; minería de datos; lucha contra la droga, Cocaína, Colombia

I. Introduction

Colombia did not begin as a coca growing power. It was initially a marimba (Colombian slang word for marihuana) growing power. In the 1960s, taking advantage of the hippie boom in the United States, marijuana crops appeared, mainly in the Serranía del Perijá and the Sierra Nevada de Santa Marta [1]. This bonanza did not last long due to the appearance of Californian marijuana in the narcotics market.

Between the 1970s and 1980s, the so-called coca bonanza phenomenon appeared, during which coca pasta base, a cocaine derivative, was being exported from Peru and Bolivia to Colombia. Then, it was processed and sent as cocaine to the United States. The high profits encouraged the growth of the business, allowing a rapid increase and expansion of production in far away regions such as those in the departments of Caquetá, Guaviare and Putumayo [2].

In the late 1980s and early 1990s drug trafficking cartels emerged as the main exponents in Cali, Medellín and Coastal areas of the country while implementing the entire coca production chain in Colombia, starting from the cultivation to the production of cocaine hydrochloride, including even money laundering.

The Colombian state response to the massive appearance of drug trafficking cartels [3], was the creation of the National Council of Narcotic Drugs (In Spanish, the CNE stands for the abbreviation of Consejo Nacional de Estupefacientes) through Law 30 of 1986 [4]. The CNE, among other things, regulates the areas where plants are cultivated in order to be later processed into drugs. Since 1987, the CNE has issued punitive decrees to control the processing of coca leaves. Law 30, in its article 7°, allows indigenous people to grow it for their own consumption, according to their cultural patterns, and establishes that the national government must promote crop substitution programs in areas where indigenous people and settlers have started to grow coca leaves for commercial purposes, before the Law was enacted [4].

At that time, the guerrillas used to exhibit an increasingly energetic presence in areas with oil, mining, illicit crops, border areas and important agricultural and livestock activity [5]. In many of these regions, large economic interests, whether or not linked to the world market, used to finance the proliferation of illegal security army groups and paramilitaries as a way to put an end to the guerrillas. The war increased the displacement of the rural civilian population, which has been historically affected by the territorial dispute between the different armed actors, as well as by the agrarian crisis. However, in Colombia, the drug problem and, within this, the problem of illicit crops, is relevant in the conflict discussion because it was once a financing source for both the insurgency and the development of paramilitary groups. Consequently, forced eradication actions have become part of the national security policy [5].

After more than two decades of implementing forced eradication strategies, the net area of coca cultivation has decreased from its peak by more than 50% [6]; but the area affected by coca cultivation has decreased by only 17% during the period from 2001 through 2010 [7]. Despite advances in surveillance, monitoring and even the implementation of manual eradication in many regions, as of 2010, 23 out of 32 departments reported coca plantations, reaching a total of 62,000 cultivated hectares of farmland [8]. Two decades of illicit crops attest that eradication as an isolated practice does not consolidate areas free of illicit crops [9].

Regarding the coca problem in Colombia, studies conducted by UCB (USA) correlate the presence of coca crops through a linear regression with the variables inherent to this problem, such as manual eradication and aerial spraying, to mention a few [6]. In addition, it also takes into account economic and social variables such as the number of human rights violations by illegal armed actors and the economic capacity of the vulnerable communities that suffer from this scourge in their territories. This research concludes that eradication was parallel to and even surpassed coca cultivation and, nevertheless, eradication has not managed to change Colombia's status as a producer of almost half of the world's coca leaf production [10], not to forget the investment made by the United States government in aerial spraying, and where the findings tell us that social investment, in addition to generating social welfare, emerges as a complementary response for controlling illicit crops.

In the study conducted by Uniandes (CO) propose a theoretical model to better understand the relationship between coca and the armed conflict in Colombia, based on the Ramsey dynamic optimization model [3]. The model depends on variables that are not easy to establish with accuracy, such as: the footprint of the illegal armed groups, territorial control of the groups, the salary earned by the portion of the population that works in coca production, the salary and cost of the military equipment used by the active guerrillas and the cost of maintaining control over the territory, all of which could possibly make the model difficult to implement.

They also look for a relationship between coca hectares and the behavior of these in neighboring towns with a coefficient of spatial correlation. To determine the presence of armed groups in the territories, they use a Boolean variable, where 1 shows if there are armed groups in a town and 0 if there aren't any. This would show the action of the groups, but not their intensity [10].

Using GIS tools, UnManizales made a comparison of how coca cultivation coverage has changed in the municipalities of Cáceres and Tarazá in the department of Antioquia (Colombia), and how the landscape has been affected by coca cultivation [11].

Ujaveriana researchs compare the changes in forest coverage in the town of Tibú [12], as a result of illicit crops, in the period between 2000 and 2014 taken from the Colombian Environmental Information System (In Spanish, the SIAC stands for the abbreviation of Sistema de Información Ambiental de Colombia) and the extension of coca crops from the Integrated Illicit Crops Monitoring System (In Spanish, the SIMCI stands for the abbreviation of Sistema Integrado de Monitoreo de Cultivos Ilícitos), between 2005 and 2014 in order to build a timeline and find the relationship between these two problems and thus determine the environmental impact caused by these illegal crops.

In the two previous investigations [11], [12], it is found that GIS comparisons are made in different time periods, taking into account changes in the territories, but it is also found that other phenomena related to the dynamics of illicit crops are not included in the studies.

The studies mentioned above are mostly related to coca crops and the map analysis of illegal crops in Colombia using methods and tools from the field of systems engineering. Of course, there are many more studies referring to the problem of illicit crops and especially coca crops, both in Colombia and in Latin America, but mainly from an economic and social point of view.

UNED shows a repertoire of data mining techniques to be used in conjunction with GIS in order to deepen spatial analysis [13]. Within the repertoire, it is found that descriptive data mining techniques are the most appropriate for the proposed analysis, especially clustering, which, together with geovisualization, would be a powerful alternative to understand the dynamics of the illicit coca cultivation business.

Correspondingly, using data mining techniques, MARA carries out a project which is applied to agriculture in Argentina [14]. Using a combination of hierarchical clustering and the K-Means algorithm to obtain homogeneous groups of climate and soil, showing as a result that generalizations by region are not adequate, but by homogeneous environmental conditions. This research shows the possibility of performing data mining processes, and specifically the use of descriptive techniques such as clustering, specifically the K-Means algorithm in agriculture, showing the possibility of following this path in the proposed research.

The driving force of this research is to serve as a tool for decision makers to confront the problems generated by the drug market, starting with an initial descriptive analysis to look at the behavior of the variables and their distribution and a categorization of the towns where there is a presence of coca cultivation, based on the results of the exploratory analysis.

The objectives in terms of data mining are:

- To establish the correlation of variables taking as the main variable the number of hectares of coca leaf crop per town.
- To categorize the towns with a presence of coca cultivation, using the K-means classification algorithm.

II. Methodology

This work was performed using the K-means algorithm, which refers to a clustering algorithm, where k refers to the number of centroids from which the grouping is performed. Also, the algorithm places each of the measurements within each of the k groups, taking into account the distance of each measurement to the centroid assigned to its group. Depending on the number of iterations that the algorithm repeats, the data subgroups are debugged (Fig. 1).

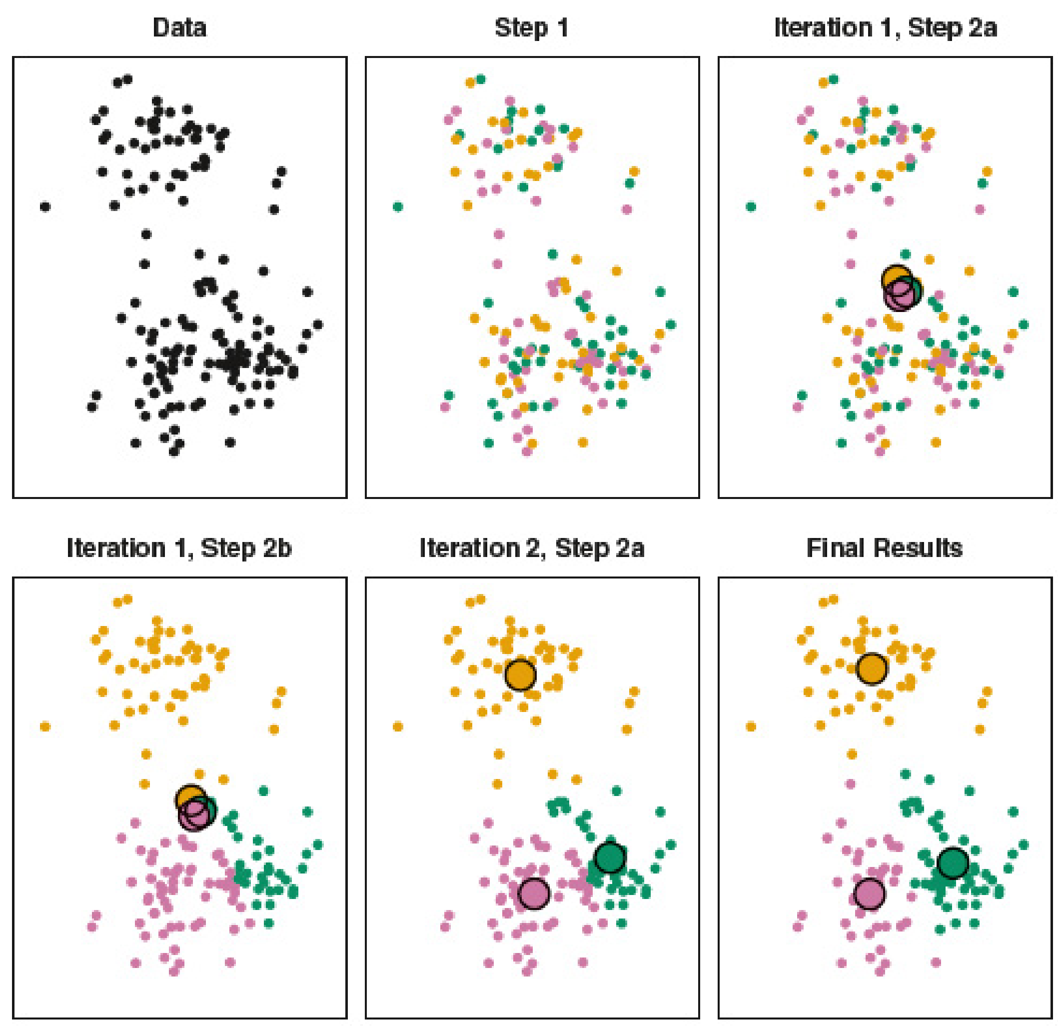

Fig. 1. Progress of K-means algorithm after several iterations.
Source: [15].

In order to observe the relationship between the variables, the PCA algorithm in R Studio PCA was used during the development. "PCA is one of the unsupervised learning techniques, which are usually applied as part of the exploratory analysis of the data" [16, par. 1]. In order to observe the correlation between variables, (the PCA) the Principal Component Analysis algorithm is used. The PCA is part of a multivariate analysis, allowing the analysis of high-dimensional data sets [16].

The CRISP-DM methodology was used specifically for data mining. It is divided into phases. In the first phase, understanding the problem, the objectives of the problem are determined, the situation is assessed including the resources, the objectives of the data mining are determined and the project plan is made. The objective of the applied data mining was to make an unsupervised classification of the towns in Colombia where coca crops are historically found. Based on this classification, scenarios were determined to evaluate which of the selected categories a town could belong to.

The databases of the coca leaf crop control monitoring system were considered for this research. The data on the monitoring of the presence of coca crops at a town level show the destruction of laboratories (primary production infrastructure and cocaine hydrochloride), manual eradication, seizures (coca leaf, coca paste and cocaine hydrochloride). Also, information regarding the market of this illegal economy, such as prices of coca leaf, coca paste and cocaine hydrochloride in pesos and dollars, as well as the average price of gold and the dollar as a source to keep track in case that the fluctuation of these two economic indicators involves any relation with the demand, the price and the number of cultivated coca leaf hectares in Colombia.

The second phase was called data comprehension. In this phase data were collected and the sources were explained focusing on the way they were coded while their quality was being verified. In order to carry out this task, a track of the official Colombian data infrastructure in the area of illicit crop was kept.

In terms of monitoring coca cultivation, the official data infrastructure in Colombia has two organizations that monitor the government's fight against this illegal market. The first is the United Nations Office on Drugs and Crime (UNODC), which through the International Illicit Crop Monitoring System (SIMCI) tracks the census of coca cultivation in Colombia, and where coca prices throughout the production chain (coca leaf, coca paste and cocaine hydrochloride) can also be found. The second organization is the Colombian Drug Observatory (ODC), which has the Colombian Drug Information System (SIDCO) with the following databases of interest for the problem to be characterized: manual eradication, cocaine hydrochloride seizures, coca leaf seizures, cocaine base paste seizures, destruction of cocaine hydrochloride laboratories and destruction of primary production infrastructure laboratories.

It was also found that in order to categorize the towns, it is important to consider the variables that show the socioeconomic status of the towns, such as the Unsatisfied Basic Needs Index (UBN), the integral performance index and the category of the town.

The selected data include information related to coca cultivation crops as such, data on the institutional fight against the coca chain, economic indicators such as the price of the dollar and gold, socioeconomic information as well as a characterization of the towns. All information was disaggregated to a town level for the period between 2010 and 2020.

The third phase was called data preparation. In this phase, the data were processed and prepared for the application of data mining techniques. This phase was intended to be segmented in order to obtain the subsets, standardize and clean the data, calculate new data from existing data and format the data looking forward to facilitating their processing.

- *Data selection*: Initially, the National Administrative Department of Statistics (In Spanish DANE stands for the abbreviation of Departamento Administrativo Nacional de Estadísticas) database of Colombia's 1 122 municipalities with its DIVIPOLA coding was used allowing it to cross all the databases. The databases exported from the ODC only export data annually showing the towns that record events or measurements. It means that at the time of crossing the databases there were some towns without information. In order to balance this situation, those towns that at the time of the crossing of databases are left without information were given a value of 0. This processing of data was carried out for the following databases: hectares of coca, manual eradication, cocaine hydrochloride seizures, coca leaf seizures, coca paste seizures, destruction of cocaine hydrochloride laboratories and destruction of primary production laboratories.

  Regarding the variables of coca leaf prices, coca paste prices and cocaine hydrochloride prices, used as constants every year, it is good to mention it was not possible to find information on these prices categorized by town. Finally, for the prices of gold and dollar, the average values per month for each year were used as constants since there is no variation at a town level.

- *Data cleaning*: As a first step, the dataset is consolidated keeping in mind that the study is based on the behavior of towns where coca cultivation takes place. Those towns with no register or measurements during the period of time when the study was conducted are excluded from this research.

  Price variables were also excluded since they are seen as constants and have no variation. For this reason, they are not ideal for the unsupervised classification process.

  In addition to the final exclusion of the variables of town classification, DANE coding and municipal performance index class because of their nature of being text variables. It happened mainly because the Kmeans algorithm had problems processing this type of variables.

- *Building the data*: Initially, the Dataset used to consider both the town and the department fields, since there are several towns with the same name but belonging to different departments in Colombia. However, in the Kmeans algorithm it is only possible to have a text type domain field. In order to sort out this requirement of the algorithm, the two fields separated by a hyphen (-) are concatenated in the format as follows: 'Antioquia-Medellín'.
- *Integrating the data*: As described in the Data selection stage, the different databases selected are integrated through the DANE code linked to each town, which is a unique town identifier that can be found in all the merged databases.

III. Results and Discussion

To start with the unsupervised classification process, the first thing to check is that the dataset is consistent in terms of the relationship between the variables that make it up.

This type of graphs are interpreted as follows "it indicates the % of variance explained by the first (Dim1) and second component (Dim2), positively correlated variables are grouped together or close together, while negatively correlated variables are plotted on opposite sides of the origin or opposite quadrants" [13, p. 1].

According to this, the first thing that can be observed is that the sum of the two axes Dim1 and Dim2 explain 50.13% of the total variance of the dataset, secondly, the behavior of the vectors corresponding to the variables that form the dataset shows some sort of stacking in their direction.

The stacking highlighted in red implies a relationship between the variables that describe the behavior of the crops and the dynamics concerning the illegal business such as hectares of coca, seizures, destruction of laboratories. The group categories highlighted in blue show a correlation between the variables that describe the behavior of the town and the variables corresponding to the integral performance index (Fig. 2). They are grouped in opposite parts to those that describe the BNI, which implies that the higher the performance index of the town, the lower the BNI.

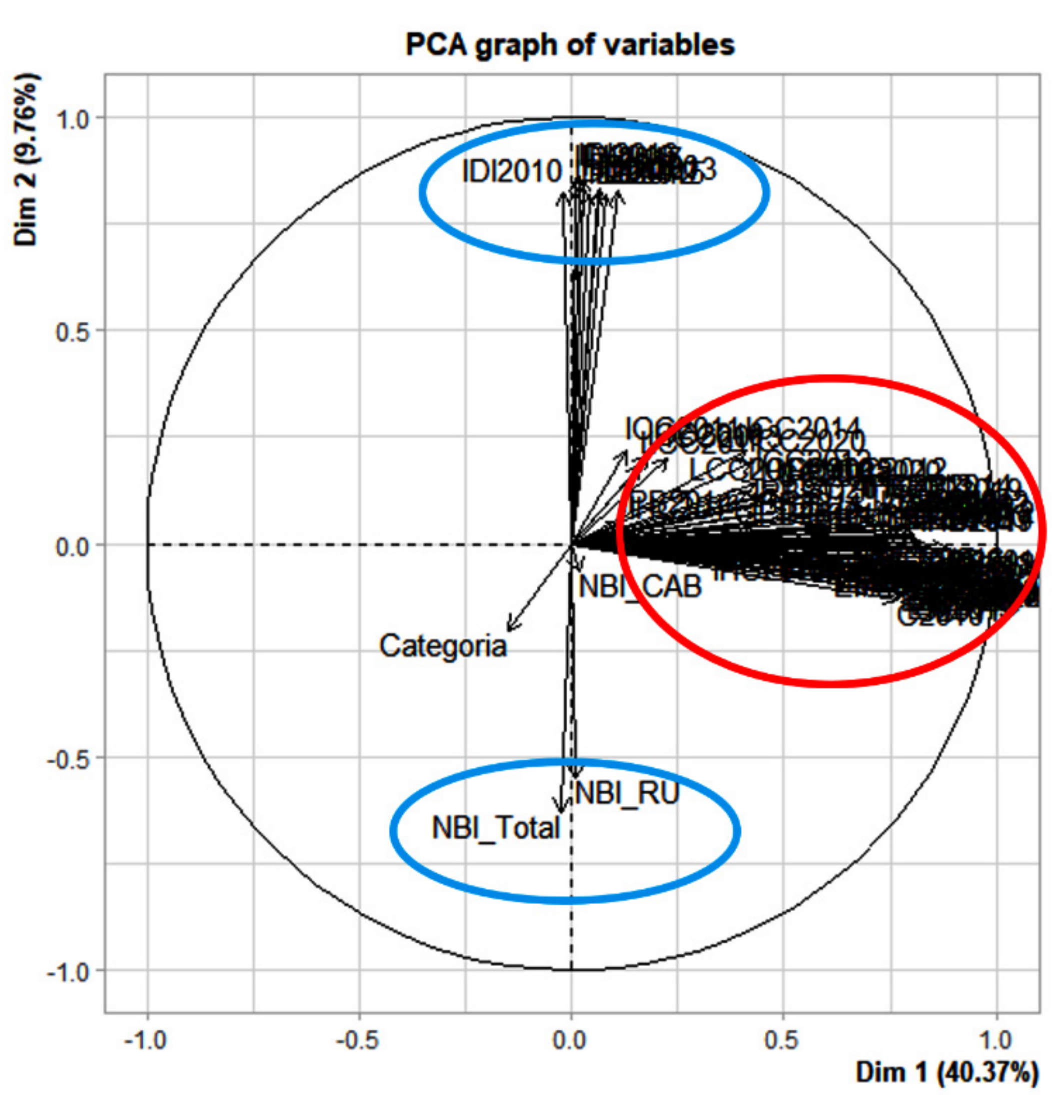

Fig. 2. Correlation of variables PCA algorithm, first iteration.
Source: Own preparation.

There are 28 out of 92 variables that would explain 60% of the data and approximately 70 variables would be necessary to explain 99% of the data (Fig. 3). As can be seen in Fig. 5, the variables are related in general terms and the ones that have the greatest value in the axes in the PCA algorithm are those that refer to the hectares of coca per town, which indicates that the central variables in the dataset are the hectares of coca. This assures the consistency and relevance of the data for analyzing the problem of illicit coca cultivation in Colombia.

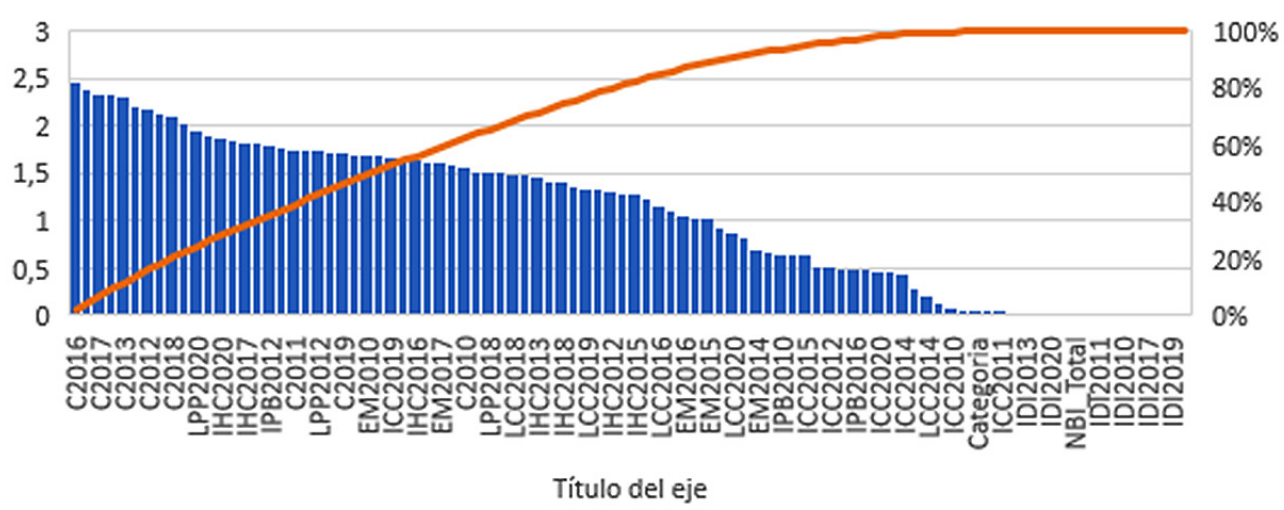

Fig. 3. Value of the variables according to PCA
Source: Own preparation.

It was decided to use 4 centroids to do the classification which will result in the classification of the data into four categories.

Having implemented the K-Means algorithm, the towns were grouped into four categories (Fig. 4). The algorithm classifies as follows: 15 towns are grouped in category 1; 229 are grouped in category 2; 24 are grouped in category 3; and in category 4 is only grouped 1 town out of the 269 towns with coca cultivation within the time period of this study.

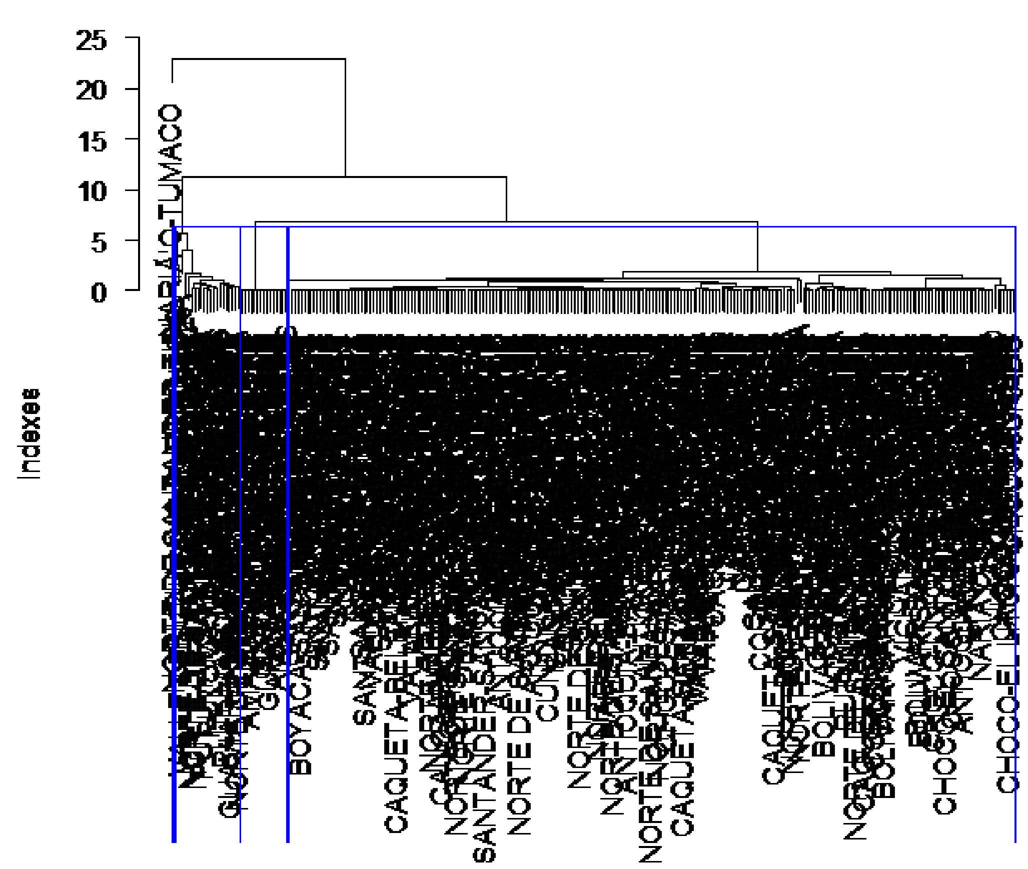

Fig. 4. Group classification of towns using the K-Means algorithm.
Source: Own preparation.

Since the purpose of this type of algorithm is to group based on similar behavior among the variables that make up a group of data, finding a category composed of only one element is not ideal for classification projects.

Tumaco's city performance mainly exceeds the performance of other cities or towns considering the number of hectares manually eradicated, seizures of cocaine hydrochloride and coca leaf, and a high number of cultivated hectares. As the city with the largest number of cultivated hectares, it was also the focus of efforts in manual eradication programs.

In order to obtain a more precise group classification, it was decided to iterate the implementation again, while temporarily excluding Tumaco. This was intended to observe the behavior of the rest of the towns and then add it to a category which best fits or consider it as a special case. In the second iteration, the PCA algorithm is re-evaluated to observe the correlation of variables in the dataset (Fig. 5).

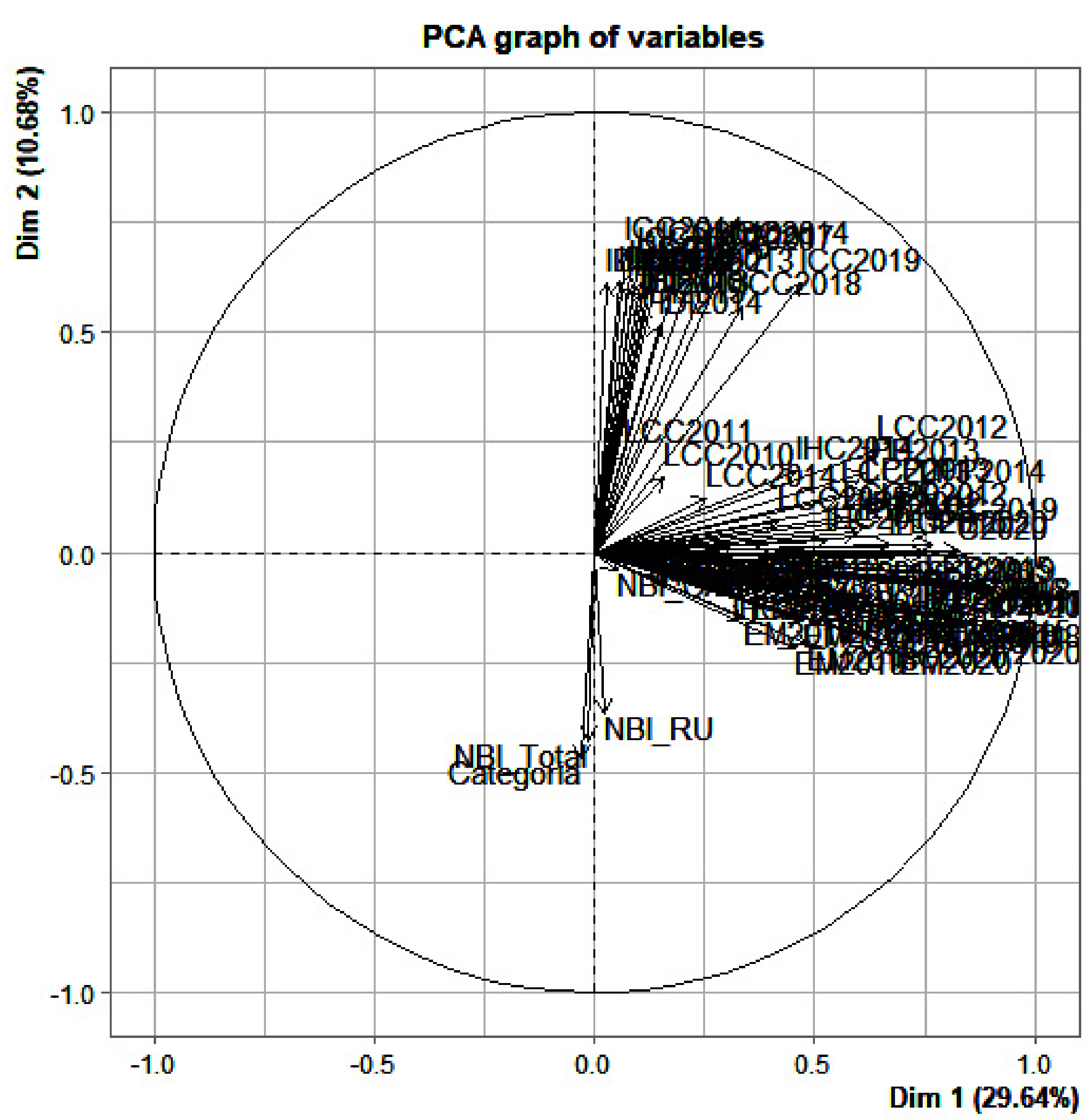

Fig. 5. Correlation of variables PCA algorithm, second iteration.
Source: Own preparation.

The percentage of variance compared to that of the first iteration is lower, described by dimension one and two, while in the first iteration the sum of the two dimensions is close to 50%; in the second iteration it is close to 40%. In the case of the correlation of variables, they are observed to be closer, which implies a greater relationship between the variables, both those with a direct relationship and those with an opposite relationship.

There are 23 variables out of 92 that would explain 60% of the data and approximately 65 variables would be necessary to explain 99% of the data. A decrease in the number of variables explaining the behavior of the data can be observed, from 28 to 23 to explain 60% of the data, and from 70 to 65 to explain 99% of the data.

As shown in Fig. 6, the variables are related in general terms and the ones that have the greatest value in the axes in the PCA algorithm are those that refer to the hectares of coca per town, which indicates that the central variables in the dataset are the hectares of coca. Other variables that are important in the data are the destruction of primary production and cocaine hydrochloride laboratories. Having observed the correlation of variables, we proceeded to look for the appropriate number of centroids to develop the classification and then defined four centroids as the adequate number.

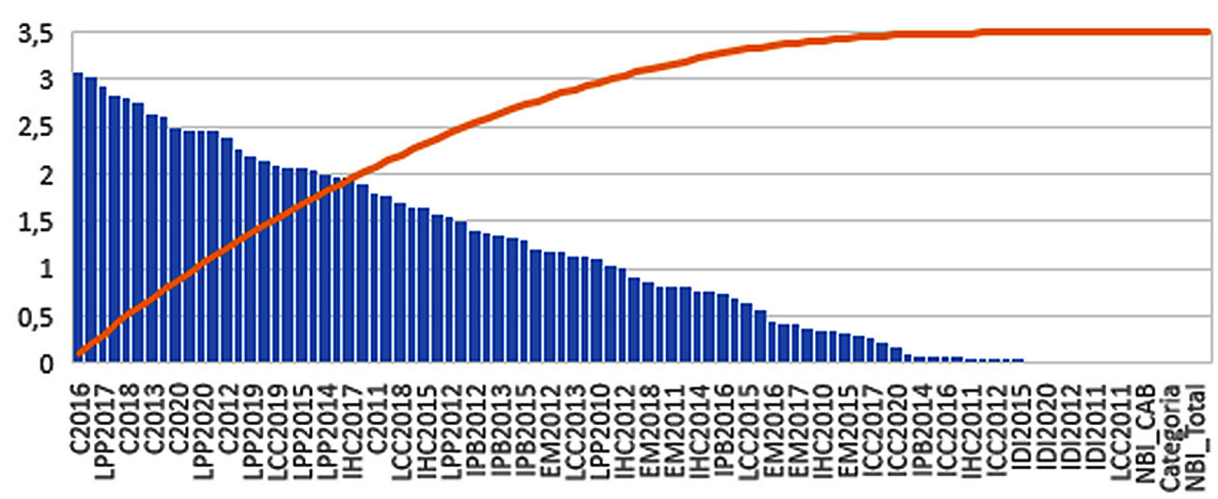

Fig. 6. Value of variables according to PCA second iteration.
Source: Own preparation.

The algorithm classifies this scenario as follows: 15 towns are grouped in category 1; 225 towns are grouped in category 2; 26 towns are grouped in category 3; and category 4 is made up of 2 towns (Fig. 7).

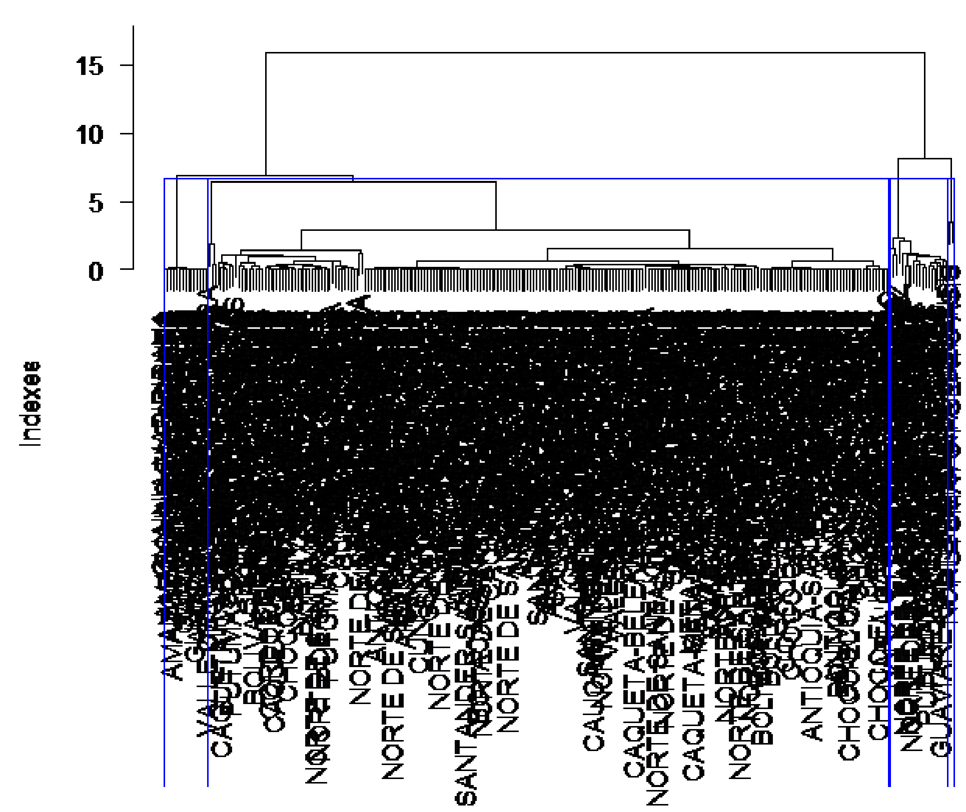

Fig. 7. Group classification of towns using the K-Means algorithm, second iteration.
Source: Own preparation.

As no categories formed by a single town were found, the results of the second iteration were taken as the final result of the classification and it was decided to classify the city of Tumaco for its particular behavior, as a special case, not belonging to any other category.

To analyze the results of the classification, we first analyzed the variables that define each category looking forward to understanding the dynamics that mark each group category (Table 1; Table 2; Table 3; Table 4).

Table 1.
Variables defining category 1 in the second iteration.

Variable	Test Value	Average Variable	Frequency	Global Average
Unsatisfied Basic Needs (UBN) Total	6.097	69.05	15	36.544
Unsatisfied Basic Needs (UBN) Rural	5.045	69.154	15	42.632
Category	–0.274	5.733	15	5.787
Cocaine hydrochloride seizures in 2015	–0.56	0	15	320.674
Coca leaf seizures in 2011	–0.571	14.8	15	3615.762
Cocaine hydrochloride seizures in 2010	–0.606	60.4	15	328.288
Cocaine hydrochloride seizures in 2018	–0.747	26.6	15	563.715
Cocaine hydrochloride seizures in 2016	–0.763	0	15	372.704
Cocaine hydrochloride seizures in 2014	–0.784	0	15	198.281
Cocaine hydrochloride seizures in 2011	–0.8	0	15	285.696
Coca Base Paste Seizures in 2014	–0.812	1.645	15	130.586
Cocaine hydrochloride seizures in 2012	–0.818	0	15	276.711
Cocaine hydrochloride laboratories in 2014	–0.847	0	15	0.433
Cocaine hydrochloride seizures in 2020	–0.855	0.007	15	493.673
Cocaine hydrochloride seizures in 2017	–0.892	0	15	597.163
Cocaine hydrochloride seizures in 2019	–0.939	0	15	384.31
Cocaine hydrochloride seizures in 2013	–1.004	0	15	270.053
Coca leaf seizures in 2010	–1.015	30.133	15	2953.333
Cocaine hydrochloride laboratories in 2010	–1.015	0	15	0.403
Cocaine hydrochloride laboratories in 2019	–1.018	0	15	0.94
Manual eradication in 2013	–1.063	3.953	15	68.147
Manual eradication in 2010	–1.073	7.867	15	102.971
Cocaine hydrochloride laboratories in 2012	–1.095	0	15	0.679

Source: Own preparation.

Table 2.
Variables defining category 2 in the second iteration.

Variable	Test Value	Average Variable	Frequency	Global Average
Integrated performance index in 2019	8.512	51.692	225	48.281
Integrated performance index in 2020	8.093	47.868	225	44.8
Integrated performance index in 2018	8.09	46.7	225	43.772
Integrated performance index in 2016	7.972	44.12	225	41.345
Integrated performance index in 2017	7.773	44.502	225	41.785
Integrated performance index in 2011	7.566	57.471	225	53.632
Integrated performance index in 2014	7.394	67.777	225	63.955
Integrated performance index in 2013	7.36	64.901	225	61.072
Integrated performance index in 2012	7.35	60.201	225	56.493
Integrated performance index in 2015	6.904	68.47	225	64.814
Integrated performance index in 2010	6.779	58.84	225	55.279
NBI Municipal capital	1.877	25.962	225	24.858
Cocaine hydrochloride laboratories in 2011	0.451	0.342	225	0.328
Category	0.393	5.796	225	5.787
Cocaine hydrochloride laboratories in 2010	–0.281	0.391	225	0.403
Cocaine hydrochloride seizures in 2013	–0.556	254.122	225	270.053
Cocaine hydrochloride seizures in 2011	–0.764	256.632	225	285.696
Coca Base Paste Seizures in 2014	–0.766	117.631	225	130.586
Coca leaf seizures in 2011	–1.003	2942.002	225	3615.762
Cocaine hydrochloride seizures in 2010	–1.066	278.15	225	328.288
Cocaine hydrochloride seizures in 2012	–1.164	234.773	225	276.711
Cocaine hydrochloride seizures in 2016	–1.219	309.288	225	372.704
Cocaine hydrochloride seizures in 2014	–1.672	153.258	225	198.281

Source: Own preparation.

Table 3.
Variables that define category 3 in the second iteration

Variable	Test value	Average Variable	Frequency	Global Average
Coca 2012	10.545	769.5	26	159.414
Coca 2013	10.327	803.692	26	155.138
Coca 2011	10.138	1100.692	26	217.06
Coca 2015	9.588	1565.104	26	295.24
Laboratories for primary production in 2015	9.576	59.5	26	12.31
Coca 2014	9.442	1137.692	26	224.511
Coca 2010	9.042	956.808	26	210.25
Coca leaf seizures in 2015	8.8	13827.39	26	2653.722
Coca 2017	8.513	2759.607	26	567.082
Coca leaf seizures in 2019	8.473	7208.682	26	1285.028
Coca 2018	8.389	2316.045	26	458.926
Coca 2016	8.219	2809.635	26	570.789
Coca leaf seizures in 2014	8.137	10284.76	26	1837.224
Laboratories for primary production in 2012	8.074	29.192	26	7.112
Laboratories for primary production in 2019	7.99	84.308	26	16.03
Coca leaf seizures in 2017	7.611	9815.185	26	1991.945
Coca leaf seizures in 2016	7.523	17731.408	26	3328.278
Coca leaf seizures in 2020	7.455	8817.058	26	1718.989
Manual Eradication in 2017	7.349	755.973	26	133.626
Laboratories for primary production in 2010	7.218	38.385	26	8.041
Coca Base Paste Seizures in 2018	7.186	617.083	26	123.93
Coca leaf seizures in 2012	6.964	6920.806	26	1478.528
Coca 2019	6.938	2611.581	26	532.259

Source: Own preparation.

Table 4.
Variables that define category 4 in the second iteration.

Variable	Test Value	Average Variable	Frequency	Global Average
Cocaine hydrochloride laboratories in 2019	13.132	35	2	0.94
Laboratories for primary production in 2017	12.417	335	2	13.194
Cocaine hydrochloride laboratories in 2017	12.045	26	2	0.896
Laboratories for primary production in 2018	11.886	356	2	14.679
Coca 2016	11.546	10120.41	2	458.926
Coca Base Paste Seizures in 2019	11.423	4565.017	2	170.705
Coca 2017	11.409	11675.145	2	567.082
Coca 2019	11.315	13351.225	2	532.259
Laboratories for primary production in 2020	11.301	379	2	16.672
Coca 2020	11.1	12517.715	2	499.817
Coca 2018	10.98	11877.31	2	570.789
Cocaine hydrochloride laboratories in 2012	10.237	18.5	2	0.679
Coca Base Paste Seizures in 2019	10.034	5164.034	2	219.591
Coca Base Paste Seizures in 2019	9.88	4000.757	2	162.148
Coca 2015	9.828	5215.555	2	295.24
Cocaine hydrochloride laboratories in 2018	9.79	21	2	0.963
Manual Eradication in 2020	9.737	8680.32	2	435.371
Coca leaf seizures 2018	9.707	35042.338	2	1569.84
Coca 2014	9.553	3717	2	224.511
Coca Base Paste Seizures in 2019	9.309	2538.96	2	123.93
Laboratories for primary production in 2014	9.183	118	2	7.299
Laboratories for primary production in 2016	9.158	242.5	2	15.381
Laboratories for primary production in 2019	8.59	293.5	2	16.03

Source: Own preparation.

The first category which is made up of 15 towns highlights most of the values for the total UBN and the rural UBN, which stands for unsatisfied basic needs. For DANE [17, par. 1] "the UBN seeks to determine, with the help of some simple indicators, whether the population's basic needs are covered. Groups that do not reach a fixed minimum threshold are classified as poor". This implies that they are economically and socially vulnerable towns with a small number of coca leaf, coca paste and cocaine hydrochloride seizures. In many situations, these seizures are found to be zero.

The second category which is made up of 225 towns, remarks heavily the values in the integral performance index, which, according to the Dictionary of Public Administration [18, par. 1] is "The Integral Performance Index (IDI), seeks to evaluate public management (in its programming, execution and follow-up stages) and decision-making in the use of municipal resources". Another variable involved in the description of category 2 is the NBI for municipal capitals.

The third category, made up of 26 towns, shows that coca cultivation is the most important factor, however, the destruction of primary production laboratories and coca leaf seizures also define this category.

The fourth category, made up of two towns, refers to a heavy presence of cocaine hydrochloride laboratories and primary production. It is also important to take into account coca cultivation, which, unlike the third category, is much higher in this one (Fig. 8).

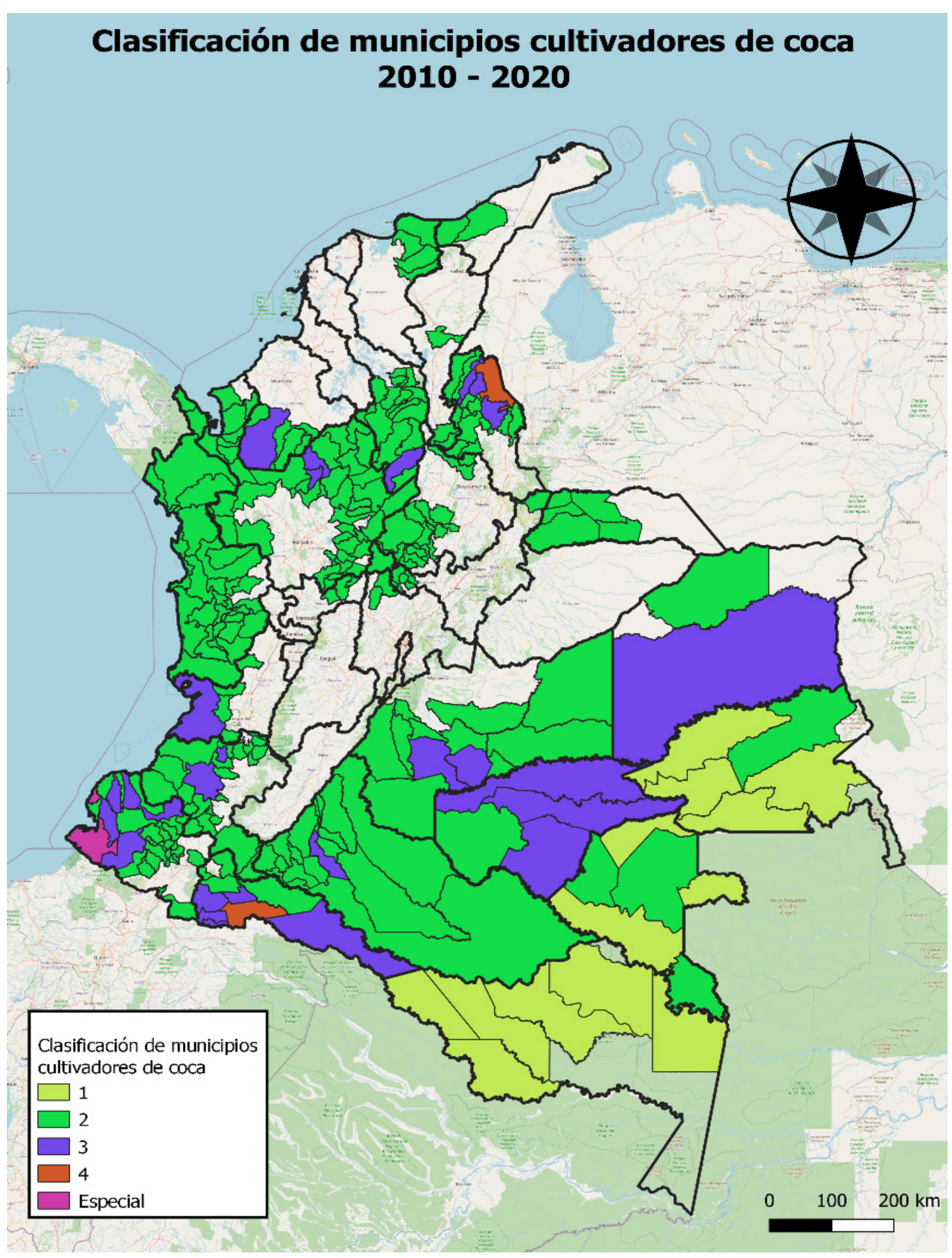

Fig. 8. Classification of coca-growing towns 2010-2020
Source: Own preparation

IV. Conclusions

- Unsupervised classification algorithms are an alternative to find clusters that have similar behavior in terms of the data that describe them, allowing to find relationships that are not easy to find with the naked eye, even if you are an expert in a certain subject.
- This research manages to classify the coca-producing towns in Colombia into four categories, depending on the behavior of the variables that characterize them with respect to this phenomenon. In addition, this research finds that in the department of Tumaco, in Nariño, there is an atypical behavior that does not fit into any of the categories found, which indicates that it is advisable to apply a special treatment in this region.
- In each category there are variables that represent a high value in each group classification, which would allow a categorization of the towns in a certain category with the same action plan to achieve crop reduction, not only by confronting the problems taking military action, but most importantly, in an integral way that provides options to the inhabitants to explore and find alternatives to get away of this illegal economy.

References

[1] E. Garzón, A. Lopez, A. Reyes Posada, R. Rocha y S. Uribe, *Drogas ilícitas en Colombia: su impacto económico, político y social.* BO, CO: Ariel, 1997.

[2] E. Ciro, "Cultivando coca en el Caquetá: vidas y legitimidades en la actividad cocalera", *Tesis doctoral*, FCPyS, UNAM, CDMX, MX, 2016. Disponible en https://ru.dgb.unam.mx/handle/DGB_UNAM/TES01000751075

[3] A. Díaz y F. Sánchez, "Geografía de los cultivos ilícitos y conflicto armado en Colombia", *Trabajo grado*, Fac Econ, UNIANDES, BO, CO, 2004. Disponible en http://hdl.handle.net/1992/7865

[4] M. Ramírez, Entre el Estado y la guerrilla: identidad y ciudadanía en el movimiento de los campesinos cocaleros del Putumayo. BO, CO: ICANH, 2001. Disponible en https://babel.banrepcultural.org/digital/collection/p17054coll10/id/2898/

[5] P. Sacipa, "Desplazamiento forzado y política de erradicación de cultivos ilícitos," *Scripta Nova*, vol. 94, no. 39, Ago. 2001. Disponible en https://revistes.ub.edu/index.php/ScriptaNova/article/view/364

[6] E. Davalos, "New answers to an old problem: Social investment and coca crops in Colombia," *Int. J. Drug Policy*, vol. 31, pp. 121–130. May 2016. https://doi.org/10.1016/j.drugpo.2016.02.002

[7] UNODC, "Informe Mundial sobre las Drogas 2014", UN, NYC, NY, USA, *V.14-04627 (S)*, 2014. Available: https://www.unodc.org/wdr2014/

[8] UNODC, *Informe Mundial sobre las Drogas 2011.* NYC, NY, USA: UN, 2011. Available: https://www.unodc.org/unodc/en/data-and-analysis/WDR-2011.html

[9] G. Buzai, C. Baxendale, N. Principi, M. Cruz, G. Cacace, N. Caloni, L. Humucata y J. Mora, *Sistemas de Información Geográfica (SIG): teoría y aplicación.* LUJ, AR: UNLU, 2013.

[10] UNODC, *Informe Mundial sobre las Drogas 2013.* NYC, NY, USA: UN, 2013. Available: https://www.unodc.org/lpo-brazil/en/drogas/relatorio-mundial-sobre-drogas.html

[11] D. Ortiz, "Influencia de los cultivos ilícitos como dinámica territorial de los municipios de Cáceres y Tazará (Antioquia) en los períodos 2007 – 2010, mediante herramientas de sistemas de información geográfica – SIG," *Trabajo Final*, Fac Cienc Ing, UManizales, MZL, CO, 2018. Disponible en https://ridum.umanizales.edu.co/xmlui/handle/20.500.12746/4174

[12] L. Camargo, "Lineamientos para la gestión de la deforestación, generada por los cultivos ilícitos asociados al conflicto armado, en el Municipio de Tibú, en el contexto del posconflicto", *Trabajo de Grado*, FEAR, UJaveriana, BO, CO, 2022. https://doi.org/10.11144/Javeriana.10554.38071

[13] J. Gil, *Minería de texto con R. Aplicaciones y técnicas estadísticas de apoyo.* MA, ES: UNED, 2021.

[14] MARA. "Las tres ideas de Sarmiento sobre la minería". *Proyecto Mara.* Disponible en https://www.infomara.com.ar/ (consultado 2010 Sep. 28).

[15] C. Gil. "Análisis de datos scRNA-Seq con bioconductor". *Rpubs*, 28 Feb. 2020. Disponible en https://rpubs.com/Cristina_Gil/scRNS-Seq_Bioconductor

[16] J. Hair, R. Anderson, R. Tatham & W. Black, *Multivariate Data Analysis*, 6th ed. TTN, NJ, USA: Prentice Hall, 2006.

[17] DANE, "Información Censo nacional de población y vivienda 2018". *Gov.co*, 2018. Disponible en https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/necesidades-basicas-insatisfechas-nbi#:~:text=La%20metodolog%C3%ADa%20de%20NBI%20busca,fijado%2C%20son%20clasificados%20como%20pobres

[18] Gov.co, "Índice de Desempeño Integral (IDI)". *Servicio al ciudadano-Glosario*, 2022. Disponible en https://www.funcionpublica.gov.co/glosario/-/wiki/Glosario+2/%C3%8Dndice+de+Desempe%C3%B1o+Integral+%3COPEN_PARENTHESIS%3EIDI%3CCLOSE_PARENTHESIS%3E

Andrés Eduardo Narváez Figueroa. Universidad Pedagógica y Tecnológica de Colombia. Tunja (Colombia).

Gustavo Cáceres Castellano. Universidad Pedagógica y Tecnológica de Colombia. Tunja (Colombia). https://orcid.org/0000-0001-9621-3585

Juan Sebastián González Sanabria. Universidad Pedagógica y Tecnológica de Colombia. Tunja (Colombia). https://orcid.org/0000-0002-1024-6077