

## Spatial analysis of academic performance in the areas of systems and telematics in the Saber Pro tests in the Colombian Caribbean

Análisis espacial del rendimiento académico en las áreas de sistemas y telemática de las pruebas Saber Pro en el caribe colombiano

Gabriel Elías Chanchí Golondrino 

Manuel Alejandro Ospina Alarcón 

Javier Antonio Pinedo Cabarcas 

Universidad de Cartagena, Colombia

### OPEN ACCESS

Recibido:

14/10/2024

Aceptado:

9/12/2024

Publicado:

13/01/2025

Correspondencia:

gchanchig@unicartagena.edu.co

DOI:

<https://doi.org/10.17081/invinno.13.1.7561>



Copyright 2025 by  
Investigación e Innovación en  
Ingenierías

#### Abstract

**Objective:** To conduct a study based on spatial data analysis to characterize academic performance in the five competencies of the Saber Pro tests, carried out by students of Systems Engineering and related programs in the departments of the Colombian Caribbean. **Methodology:** For the development of this research, a four-phase adaptation of the CRISP-DM methodology was utilized: F1. Business and data understanding, F2. Data preparation, F3. Modeling, F4. Evaluation and analysis. **Results:** As a result, the analysis of quantiles identified the departments in the Caribbean coast with the highest and lowest performance across the five areas of the test. Additionally, spatial regression analysis was used to determine the correlation among the five competencies of the test. Finally, spatial clustering analysis identified groups of departments that collectively achieved the best and worst results. **Conclusions:** the results obtained in this study are intended to guide strategic decision-making aimed at improving the quality of education in the Colombian Caribbean coast. Furthermore, the proposed study serves as a foundation for extrapolating spatial data analysis to other regions and specific levels of the Saber tests.

**Keywords:** Spatial analysis, spatial correlation, spatial data, saber pro tests, academic performance.

#### Resumen

**Objetivo:** Conducir un estudio basado en análisis espacial de los datos para la caracterización del rendimiento académico en las cinco competencias de las pruebas saber Pro, por parte de los estudiantes de carreras de Ingeniería de Sistemas y afines en los departamentos del caribe colombiano. **Metodología:** Para el desarrollo de la presente investigación se hizo uso de una adaptación a 4 fases de la metodología CRISP-DM: F1. Entendimiento del negocio y de los datos, F2. Preparación de los datos, F3. modelado, F4. Evaluación y análisis. **Resultados:** Como resultado, se obtuvo mediante análisis de cuantiles los departamentos de la costa caribe con mejor y peor rendimientos en las 5 áreas de la prueba. Así mismo, mediante análisis de regresión espacial se determinó la correlación entre las 5 competencias de la prueba. Finalmente, mediante análisis de clustering espacial se determinaron los grupos de departamentos que obtuvieron en conjunto los mejores y peores resultados. **Conclusiones:** Los resultados obtenidos en este trabajo pretenden servir de guía para la toma de decisiones estratégicas que contribuyan a mejorar la calidad de la educación en la costa caribe de Colombia. Así mismo, el estudio propuesto sirve de base para extrapolar el análisis espacial de los datos en otras regiones y niveles específicos de las pruebas saber.

**Palabras claves:** Análisis espacial, correlación espacial, datos espaciales, pruebas saber pro, rendimiento académico.

**Como citar (IEEE):** G.E. Chanchí Golondrino, M.A. Ospina Alarcón, J.A. Pinedo Cabarcas "Spatial analysis of academic performance in the areas of systems and telematics in the Saber Pro tests in the Colombian Caribbean", Investigación e Innovación en Ingenierías, vol. 13, no. 1, pp. 1-14, 2025, doi: <https://doi.org/10.17081/invinno.13.1.7561>

## Introduction

Educational policies outlined by the Colombian Ministry of National Education and the ICFES (Colombian Institute for the Evaluation of Education) have as their primary purpose the development of academic competencies at different educational levels, from basic to higher education [1]. This is based on the understanding that the development of these competencies contributes to a successful professional life and the progress of society. In accordance with the above, one of the key indicators in Colombia for evaluating the development of academic competencies in different disciplines and consequently the quality of education are the results of students from different academic levels on the Saber Pro exams [2]. The results obtained on the Saber Pro exams serve as a diagnosis to identify the strengths and weaknesses of students in different areas, so that educational institutions can obtain feedback on curricular aspects that allow them to make strategic decisions focused on improving the effectiveness of the teaching-learning process [3,4].

In Colombia, the ICFES is the government institute responsible for the design, administration, and evaluation of the Saber Pro exams at the basic education level (grades three and five), the middle education level (grades nine and eleven), and the higher education level [5]. The Saber Pro exams assess the generic and specific competencies of each undergraduate program and are applied when university students have completed at least 75% of the program credits and are a graduation requirement [6,7,8]. In this sense, according to [9], the Saber Pro exams are a set of instruments that the Colombian state uses to evaluate the quality of formal education received by students who complete the different professional academic programs in Higher Education Institutions. In addition to promoting the improvement of higher education [10], the importance of the Saber Pro exams lies in the fact that they are a fundamental tool for evaluating the key competencies that affect the preparation of future professionals for their insertion into the labor market, so that by evaluating academic, cognitive, and professional skills, these evaluations provide a panoramic view of the students' ability to face the challenges of the workplace [11].

Several studies have been conducted from the field of data science for the characterization of academic performance, specifically regarding the analysis of the results of the different variants of the Saber tests. In [12], a study based on spatial analysis of data on the Saber 5 test dataset from 2019 at the national level is carried out, obtaining as results the areas with the highest spatial correlation and the departments with the best overall results in the different competencies. In [13], a study based on data mining and machine learning is carried out in order to determine the socioeconomic factors that affect the performance of students in the Saber Pro tests in the reading and writing skills. In [14], a supervised learning model supported by decision tree algorithms is applied in order to detect the factors associated with the academic performance of 11th grade students at the national level who took the Saber 11 tests between 2015 and 2016. Similarly, in [15] a decision tree model is applied for the characterization and obtaining of the factors that affect the performance of Colombian students in the engineering faculties in the Saber Pro tests between the years 2012 and 2014. In the same sense, in [16] the decision tree model was used to determine the factors that affect the performance of students in the 11th grade at the national level in the reading comprehension skill, obtaining as main attributes age, stratum, school day and ICT conditions of the student. In [10] a study is developed on the Saber Pro tests of engineering students in Antioquia from a dataset with 108 academic, economic and sociodemographic variables, which had 3 fundamental objectives: to group the types of students, to select the factors that have the most impact on performance in the tests and predict performance in the tests from the selected variables. In [17]

the decision tree model was used to determine the patterns associated with the academic performance of students at the Javeriana University of Cali in the years 2017 and 2018 at the level of generic competencies. In [18] data mining techniques and specifically the decision tree model were applied in order to determine the factors that affect the interaction of students in the virtual English distance course of the National University of Loja, between the years 2013 and 2014. In [19] a logistic decision tree model was used to characterize the factors that affect the performance of students entering computer science careers in the admission exams of the Autonomous University of Yucatan, in order to contribute to the early detection of academic risk. In [20] data mining techniques and specifically decision tree models are applied in order to characterize and determine the main factors that have an impact on student dropout in higher education institutions in Bogotá.

Upon reviewing the previous studies, it is evident that most of them have focused on the characterization and analysis of the academic performance of students of different educational levels using data mining and machine learning techniques, without addressing the spatial data analysis approach. Likewise, regarding the studies that analyze the results of the Saber tests, no research has been identified that focuses on the analysis of student performance in different levels for the context of the Colombian Caribbean coast. Finally, the work that analyzes the Saber tests through spatial data analysis focuses on the national context and on the Saber 5 tests. Therefore, the main contribution of this work is the conduction of a study based on spatial data analysis on the results of the Saber Pro tests for the year 2022, focusing on students whose undergraduate degree is in the common core of systems engineering, telematics, and related fields in the context of the Colombian Caribbean coast.

Spatial data analysis encompasses a diverse set of techniques and models that rely on the geospatial location of data, with the focus on understanding and leveraging the spatial relationships and interactions between cases. This type of analysis involves the use of tools such as spatial correlation to identify patterns of association between variables in space, quantile analysis to examine the spatial distribution of values, and spatial clustering to group geographic areas with similar characteristics [21,22,23].

The proposed study was developed based on the creation of a dataset derived from the national Saber Pro test dataset. Using the advantages of the Python pandas library, each competency was categorized into four levels and a percentage count was made of the students from each department in the Colombian Caribbean coast who fell into each category. Similarly, geospatial data were linked to the dataset using the free QGIS tool. Spatial analysis was developed using the GeoDa tool and allowed to obtain, in the first instance, the spatial correlation between the five competencies evaluated in the test, as well as the quantile analysis for these five areas in their best and worst performances. In the same way, the study included the application of spatial clustering models relating different areas in their best and worst performances, in order to determine the departments that together had the best results in the test competencies. The results obtained are intended to serve as a reference for decision-making by government entities and higher education institutions in the Colombian Caribbean coast, with a view to improving the quality of education in this region. Likewise, the methodological phases used in this research can be extrapolated to other application contexts, in order to expand the number of research projects in Colombia from the perspective of spatial data analysis.

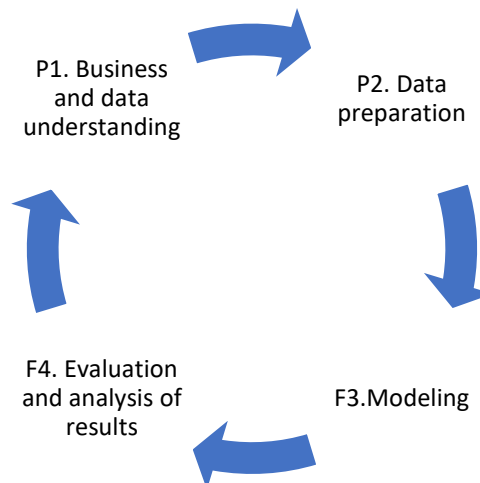
The remainder of the article is structured as follows: The next section outlines the methodology employed in this research. Subsequently, the results and discussion

of the study are presented, including spatial regression analysis among the five competencies evaluated in the test, as well as quantile analysis for each competency and the developed clustering analysis. In this context, the discussion of the results is also compared with findings from other studies in the field. Finally, the concluding section presents the conclusions and future research directions stemming from this study.

## Methodology

For the development of this research, an adaptation of the CRISP-DM (Cross-Industry Standard Process for Data Mining) data mining methodology was made into four phases: P1. Business and data understanding, P2. Data preparation, P3. Modeling, P4. Evaluation and analysis of results (see Figure 1). The CRISP-DM methodology represents a robust and structured framework that allows guiding the data mining process from business understanding to the implementation of solutions in different application domains. Thus, it has been widely adopted in the data science community, providing a coherent and replicable structure to address the complex challenges inherent in data analysis and informed decision-making [24,25,26, 27].

Figure 1. Methodology considered



Source: own elaboration

In Phase 1 of the methodology, the dataset with the national results of the Saber Pro tests from the year 2018 to the year 2022 was downloaded from the open data portal, and the exploration and identification of the instances and attributes that made up the dataset were carried out. This dataset is made up of a total of 1,217,482 instances and 57 attributes, which, in addition to the scores corresponding to the 5 academic competencies (quantitative reasoning, written communication, English, critical reading, and civic competencies), include a set of social, demographic, and economic variables associated with each student (see Table 1). In Phase 2, from the explored dataset, the instances belonging to the year 2022 and corresponding to the 7 departments of the Colombian Caribbean coast were first filtered using the Python Pandas library.

Table 1. Columns of the resulting dataset

Columns	Description
State	Corresponds to one of the seven states considered in the study: Atlántico, Bolívar, Cesar, Cordoba, La Guajira, Magdalena and Sucre.

English_A1, English_A2, English_B1, English_B2	Corresponds to the percentage of students in each department classified in categories A1, A2, B1 and B2 in the area of English. These categories are specified by the international exams.
Raz_Cuan_B, Raz_Cuan_M1, Raz_Cuan_M2, Raz_Cuan_A	Corresponds to the percentage of students in each department classified in the low (B), medium-low (M1), medium-high (M2) and high (A) categories in the area of quantitative reasoning.
Com_Es_B, Com_Es_M1, Com_Es_M2, Com_Es_A	Corresponds to the percentage of students in each department classified in the low (B), medium-low (M1), medium-high (M2) and high (A) categories in the area of written communication.
Lec_Crit_B, Lec_Crit_M1, Lec_Crit_M2, Lec_Crit_A	Corresponds to the percentage of students in each department classified in the low (B), medium-low (M1), medium-high (M2) and high (A) categories in the critical reading area.
Comp_C_B, Comp_C_M1, Comp_C_M2, Comp_C_A	Corresponds to the percentage of students from each department classified in the categories low (B), lower-middle (M1), upper-middle (M2), and high (A) in the civic competencies area.
Global_B, Global_M1, Global_M2, Global_B	Corresponds to the percentage of students from each department classified in the categories low (B), lower-middle (M1), upper-middle (M2), and high (A) in the global score.

Source: own elaboration

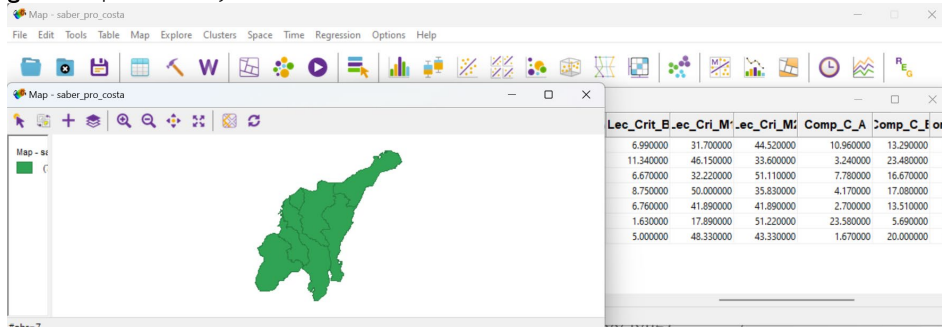
Likewise, the instances corresponding to students belonging to the common core of Systems Engineering, Telematics, and Related Fields were filtered, so that a total of 1,263 instances were obtained in total. Once the instances of interest were filtered, the grouping or counting of the instances for the 7 departments of interest was carried out. Subsequently, the columns or attributes of the dataset that do not correspond to the scores in the 5 evaluated competencies and the global score were eliminated. Thus, with the score columns, a categorization process to 4 levels is carried out (low (B), lower-middle (M1), upper-middle (M2), and high (A)) and the One Hot Encoding coding method is applied. In this way, each of the 4 levels corresponding to the 5 competencies and the global score of the test will have its own column (see Table 1). It is worth mentioning that for the English area, the categorization that the dataset included by default was used, which corresponds to the levels used by international exams. Finally, using the free QGIS tool, the data is georeferenced and exported to the .shp format, which is compatible with the GeoDa data spatial analysis tool.

In phase 3, based on the dataset formed, the following analyses were conducted: spatial correlation analysis among the variables of the dataset at high and low levels, analysis by quantiles of each competency at high and low levels, and clustering analysis relating the variables of the dataset, considering high and low levels. Finally, in phase 4, the analysis of the results obtained in the evaluated competencies was performed, taking into account, in the case of spatial correlation, the coefficient of self-determination.

## Results

After the spatial dataset for the Caribbean coast was created with the percentage level of each of the levels of the 5 evaluated competencies in the students of the common core of telematics and related fields, the dataset was loaded into the free Geoda tool (see Figure 2) and the correlation analysis was carried out in the first instance between the variables for the high and low levels of the different competencies.

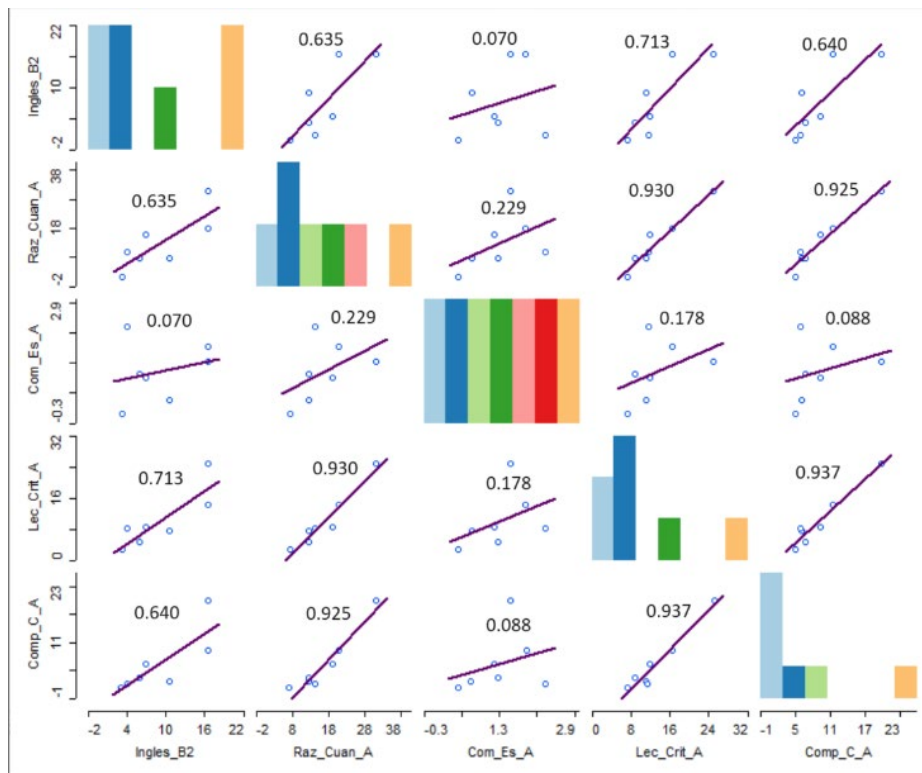
**Figure 2.** Special analysis dataset of the data loaded into the GeoDa tool



Source: own elaboration

Thus, Figure 3 shows the scatter plots that allow identifying the spatial correlation between the 5 evaluated competencies in their high level (A), as well as the determination coefficient obtained by applying the linear regression model between the different areas.

**Figure 3.** Spatial correlation between high levels of the 5 competencies



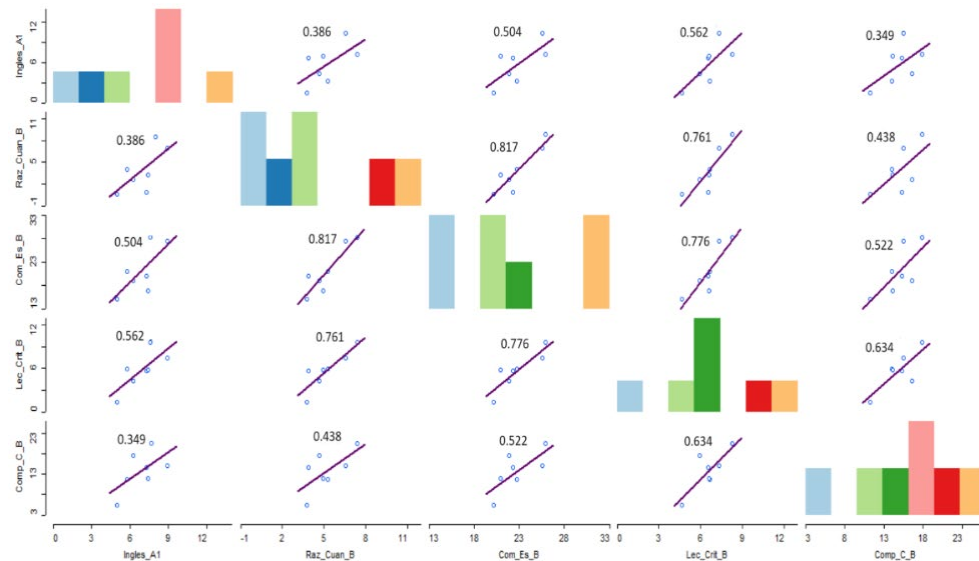
Source: own elaboration

As observed in Figure 3, the English (B2) area exhibits the best correlation with the critical reading area, with a determination coefficient of 0.713. Similarly, the quantitative reasoning area presents the best correlations with the civic competencies and critical reading areas, with respective determination coefficients of 0.925 and 0.930. On the other hand, the written communication area does not present a representative correlation with any of the areas, with the highest determination coefficient being 0.229, corresponding to the quantitative reasoning area. Similarly, at the level of the critical reading area, it presents the best correlations with the civic competencies and quantitative reasoning areas, with respective determination coefficients of 0.937 and 0.930. Finally, regarding the civic

competencies area, it presents the best correlations with the critical reading and quantitative reasoning areas, with respective values in the determination coefficient of 0.937 and 0.930. Thus, it can be concluded that the best correlations occur between the quantitative reasoning and critical reading areas, quantitative reasoning and civic competencies, and civic competencies and critical reading, in all these cases with values greater than 0.925 in the determination coefficient. This indicates that there is a direct relationship between the mentioned areas at the high-performance level, such that they can be used to predict performance.

Furthermore, Figure 4 presents the scatter plots that allow identifying the correlation between the 5 evaluated competencies in their low level (B), as well as the determination coefficient obtained by applying the linear regression model between the different areas.

**Figure 4.** Spatial correlation between low levels of the 5 competencies

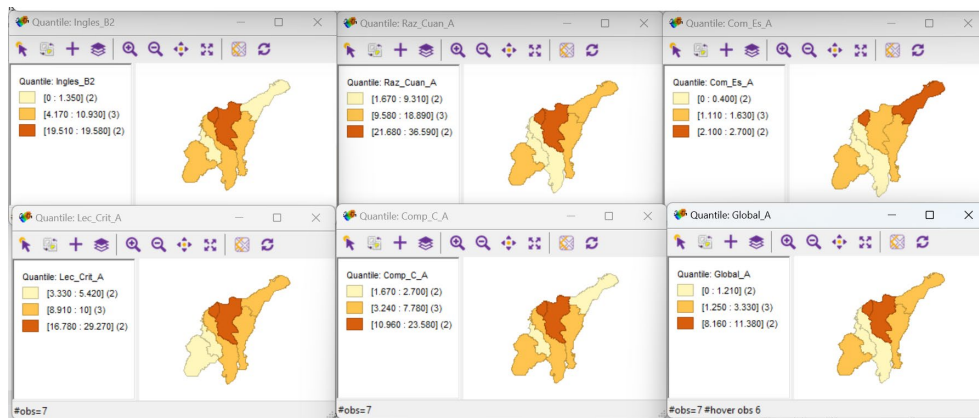


**Source:** own elaboration

As can be observed in Figure 4, the English area (A1) does not exhibit a significant correlation with any of the remaining four areas, with the highest coefficient of determination being 0.562 for critical reading. Similarly, at the quantitative reasoning level, the best correlation is found with the written communication area, with a coefficient of determination of 0.817. Likewise, for the written communication area, the best correlation is also found with the quantitative reasoning area, with a coefficient of determination of 0.817. On the other hand, for the critical reading area, the best correlation is found with the written communication area, with a coefficient of determination of 0.776. Finally, the civic competence area does not exhibit a significant correlation with the remaining four areas, with the best correlation being found with the critical reading area, with a coefficient of determination of 0.634. It can be observed that the best correlation for the low level is found between the written communication and quantitative reasoning areas, with a coefficient of determination of 0.817. Thus, it can be stated that students who had low performance in written communication also had low performance in quantitative reasoning.

Furthermore, Figure 5 presents the analysis by three quantiles for the high level in each of the five competencies evaluated, as well as for the overall score obtained on the test.

**Figure 5.** Analysis of 3 quantiles for the high level of the 5 competencies



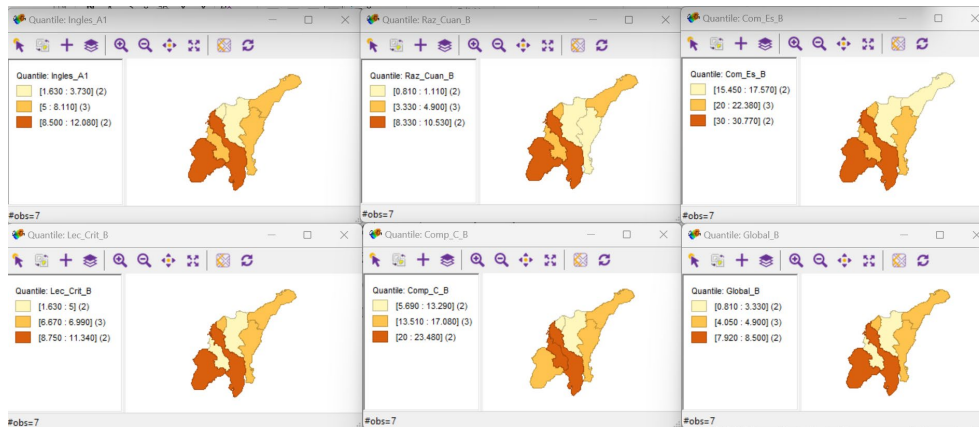
**Source:** own elaboration

For the English area at the high level (B2), it can be observed that the departments with the highest percentage of students in this category are Atlántico and Magdalena, with a percentage around 19%. Likewise, the departments with the lowest percentage of students in this category are Sucre and Guajira, with a percentage lower than 1.3%. The remaining three departments (Bolívar, Cesar, and Córdoba) are in an intermediate quantile, with a percentage ranging from 4.170% to 10.930%. With regard to the quantitative reasoning area, the departments with the highest percentage of students at the high level were Atlántico and Magdalena, with a percentage between 21.7% and 36.6%, while the departments with the lowest percentage of students in this category are Sucre and Bolívar, with a percentage lower than 9.3%. The remaining three departments (Guajira, Cesar, and Córdoba) are in the intermediate quantile in this area, with a percentage ranging from 9.6% to 18.9%. Similarly, with regard to the written communication area, it can be observed that the departments with the highest percentage of students in this category are Atlántico and Guajira, with percentages between 2.1% and 2.7%, while the departments with the lowest percentage of students in this category are Bolívar and Sucre, with a percentage lower than 0.4%. Likewise, the remaining three departments (Cesar, Magdalena, and Córdoba) are in an intermediate quantile in this area, with a percentage ranging from 1.1% to 1.6%. Similarly, in the critical reading area, it can be observed that the departments with the highest percentage of students in this category are Atlántico and Magdalena, with percentages between 16.8% and 29.27%, while the departments with the lowest percentage of students in this category are Sucre and Córdoba, with a percentage lower than 5.4%. The remaining three departments (Guajira, Cesar, and Bolívar) are in an intermediate quantile in this area, with a percentage between 8.9% and 10%. On the other hand, in the civic competence area, it can be observed that the departments with the highest percentage of students in this category are Atlántico and Magdalena, with percentages between 10.9% and 23.6%, while the departments with the lowest percentage of students in this category are Guajira and Sucre, with a percentage lower than 2.7%. Likewise, the remaining three departments (Bolívar, Córdoba, and Cesar) are in an intermediate quantile in this area, with percentages between 3.2% and 7.8%. Finally, the overall score has a distribution equal to the results of quantitative reasoning, so that the departments with the best overall percentages in this category are Atlántico and Magdalena, while the departments with the lowest overall percentage in this category are Bolívar and Sucre. It is worth mentioning that the department of Atlántico appears in all 5 skills with the best percentages, while the department of Magdalena appears in 4 of 5 skills with the best percentages. Finally, the department of Sucre appears in all 5 skills with the lowest percentage in the analyzed category.



In the same vein, Figure 6 presents the analysis at 3 quantiles for the low level in each of the 5 skills assessed in the test, as well as for the overall score obtained in the test.

**Figure 6.** Analysis of 3 quantiles for the low level of the 5 competencies



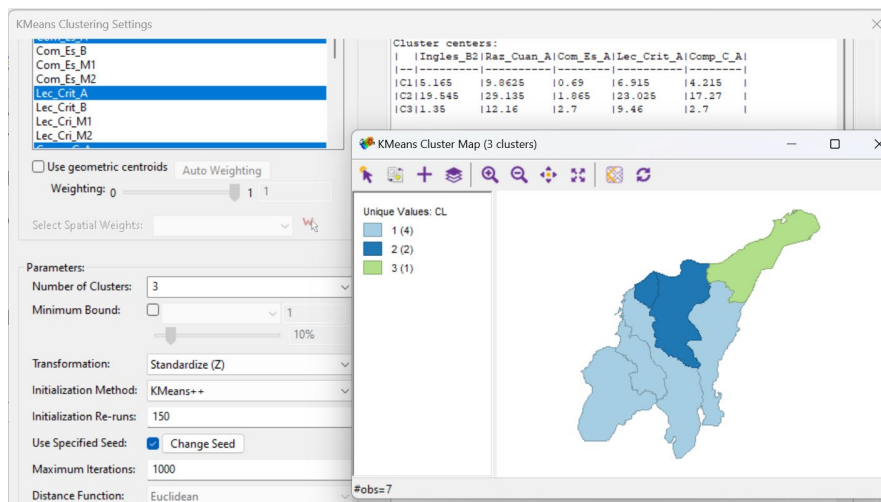
**Source:** own elaboration

In the case of the English area at the low level (A1), it can be observed that the departments with the highest percentage in this category are Bolívar and Córdoba, with a percentage between 8.5% and 12.08%, which are in the intermediate quantile at the high level. Similarly, with regard to the quantitative reasoning area at the low level, the departments with the lowest percentage in this level are also Bolívar and Córdoba, with a percentage between 8.3% and 10.5%, which are in the lowest quantile and in the intermediate quantile at the high level, respectively. Thus, the department of Bolívar is the one that has the highest percentage of students at the low level and the lowest percentage of students at the high level for the quantitative reasoning area. On the other hand, regarding the written communication area, the departments of Bolívar and Córdoba are the ones that have the highest percentage of students in this category, with a percentage between 30% and 30.8%, which are in the lowest quantile and in the intermediate quantile at the high level, respectively. This indicates that the department of Bolívar is the one that has the highest percentage of students at the low level and the lowest percentage of students at the high level for the written communication area. Thus, the department of Bolívar is the one that has the highest percentage of students at the low level and the lowest percentage of students at the high level for the written communication area. Regarding the critical reading area, the departments of Bolívar and Córdoba also have the highest percentage of students in this category, with a percentage ranging from 8.75% to 11.34%, which are in the intermediate and low quantiles at the high level, respectively. This indicates that the department of Córdoba has the highest percentage of students at the low level and the lowest percentage of students at the high level for the critical reading area. Similarly, for the civic competence area, the departments with the highest percentage of students at the low level are Bolívar and Sucre, with percentages between 20% and 23.5%, which correspond to the intermediate and low quantiles at the high level, respectively. Finally, regarding the overall score, it is important to mention that it has a distribution equal to the results of the critical reading area, so that the departments with the best overall percentages at the low level are Bolívar and Córdoba. It is worth mentioning that the department of Bolívar appears in all 5 skills with the highest percentages at the lowest level, while the department of Córdoba appears in 4 of 5 skills with the best percentages at the lowest level.

On the other hand, when performing the spatial clustering analysis based on the KMeans model with 3 centroids for the high level of the dataset (see Fig. 7), it is

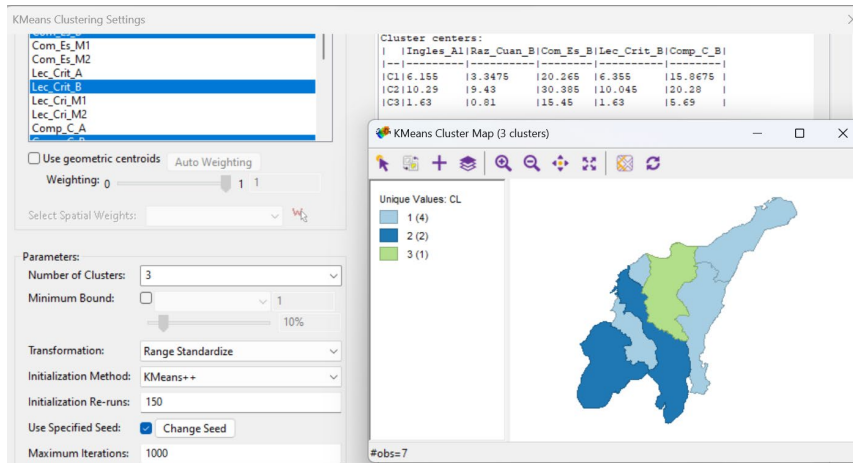
obtained that cluster 1 includes in its centroid the lowest percentages in 3 (quantitative reasoning, written communication, and critical reading) of the 5 skills and the intermediate percentage in civic skills and english, being associated with 4 departments namely: Bolívar, Sucre, Cesar, and Córdoba. Likewise, cluster 2 includes in its centroid the highest percentages in 4 (english, quantitative reasoning, critical reading, and civic skills) of the 5 skills and the intermediate percentage in the written communication area, being associated with the departments of Atlántico and Magdalena. Finally, cluster 3 includes the lowest percentages in 2 (English and civic skills) of the 5 skills, the intermediate percentage in 2 of the 5 skills (quantitative reasoning and critical reading), and the highest percentage in the written communication skill, being associated with the department of La Guajira.

Figure 7. Analysis of 3 clusters for the 5 competencies evaluated at high level



Source: own elaboration

On the other hand, when performing the spatial clustering analysis based on the KMeans model with 3 centroids for the low level of the dataset (see Figure 8), it is obtained that cluster 2 is the one that includes in its centroid the highest percentages in the low level for the 5 skills assessed and corresponds to the departments of Bolivar and Córdoba, which is consistent with the results obtained in the quantile analysis. On the other hand, it is important that cluster 3 includes in its centroid the lowest values in the percentages of the low level and corresponds to the department of Atlántico, which is consistent with the quantile analysis carried out.

**Figure 8.** Analysis of 3 clusters for the 5 competencies evaluated at the low level

**Source:** own elaboration

As for the discussion of the results, it is worth mentioning that the results obtained in this study are of great relevance compared to those proposed in [12]. This study specifically focuses on the characterization of the Saber Pro test results for the Colombian Caribbean coast, considering the percentage of students at different levels of the skills instead of the student count. This approach provides a more accurate analysis that is independent of the demographic density of each department. Moreover, unlike the studies proposed in [13] and [15], which use data mining and decision tree models to identify the factors that influence student performance in the Saber Pro tests, this article offers a new approach based on spatial analysis. This approach aims to identify the spatial correlation existing by areas in the different departments of the Caribbean coast, as well as the best and worst performances for each evaluated skill.

## Conclusion

Considering that in the field of data science, various studies have focused on identifying the factors that influence student performance on the Saber tests, this article presents a new approach based on spatial data analysis. This approach aims to determine the Caribbean coast departments that exhibit a higher correlation in the different performance levels for each skill. This work intends to serve as a reference for making strategic decisions aimed at improving the quality of higher education in the Caribbean coast.

This study demonstrated the relevance of using free and open-source software tools for conducting spatial data analysis studies. The pandas library proved to be suitable for filtering, counting, and grouping data by department from the original Saber Pro test dataset. Likewise, the QGIS tool proved to be suitable for georeferencing the data and creating the dataset in .shp format, which is compatible with the GeoDa tool. Finally, the GeoDa tool proved to be useful in applying different spatial methods such as correlation analysis, quantile analysis, and clustering analysis.

Regarding the results obtained in the spatial study for students in the Caribbean coast of Colombia in the area of systems engineering, telematics, and related fields, it is noteworthy at the high level that the best correlations occur between the areas of quantitative reasoning and critical reading, quantitative reasoning and civic skills, and civic skills and critical reading, in all these cases with values exceeding 0.93 in the coefficient of determination. Similarly, regarding the quantile analysis for the high level, it is important to highlight that the department of Atlántico appears in the 6 skills with the best percentages, while the department of

Magdalena appears in 5 of 6 skills with the best percentages. Finally, regarding the clustering analysis, it is important to conclude that the departments of Bolívar and Córdoba appear in the cluster with the highest percentages of students at the low level.

As a derivative work of this research, it is intended to extrapolate the present study to other regions of the country and at the national level, in order to contrast the percentages of student distribution obtained in each of the 5 evaluated competencies. Similarly, following the methodology and tools considered, it is intended to extrapolate the present study to the other variants of the Saber tests in the different educational levels.

## Bibliographic References

- [1] N. Palacios Mena, «El currículo de ciencias sociales y las pruebas Saber 11 en Colombia: consonancias y disonancias», *Voces Silenc. Rev. Latinoam. Educ.*, vol. 9, n.o 2, pp. 80-106, dic. 2018. DOI: <https://10.18175/vys9.2.2018.06>
- [2] J. R. García-González, P. A. Sánchez-Sánchez, M. Orozco, y S. Obredor, «Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia», *Form. Univ.*, vol. 12, n.o 4, pp. 55-62, ago. 2019. DOI: <https://10.4067/S0718-50062019000400055>
- [3] W. Acero, J. F. Sánchez, D. Suárez, y C. Téllez, «Modelo de recalificación para la prueba Saber 11», *Comun. En Estad.*, vol. 9, n.o 1, pp. 43-54, 2016. DOI: <https://10.15332/s2027-3355.2016.0001.02>
- [4] L. A. Sanabria James, M. C. Pérez Almagro, y L. E. Riascos Hinestroza, «Pruebas de evaluación Saber y PISA en la Educación Obligatoria de Colombia», *Educ. Siglo XXI*, vol. 38, n.o 3 Nov-Feb, pp. 231-254, oct. 2020. DOI: <https://10.6018/educatio.452891>
- [5] J. C. Morales-Pinero, M. C. Cote-Sánchez, y I. A. Molina-Bernal, «Incidencia de las TIC en el mejoramiento de las pruebas saber 11 a partir del modelo TPACK», presentado en Encuentro Internacional de Educación En Ingeniería 2019, Cartagena, Colombia. [En línea]. Disponible en: <https://acofipapers.org/index.php/eiei/article/view/40/35>
- [6] J. Pérez Rave y F. González Echavarría, «Classification Trees vs. Logistic Regression in the Generic Skill Development in Engineering», *Comput. Sist.*, vol. 22, n.o 4, dic. 2018. DOI: <https://10.13053/cys-22-4-2804>
- [7] J. Herrera-Cardozo, «Prueba Saber Pro y módulo de comunicación escrita 2016: un análisis estadístico descriptivo», *Rev. Neuronum*, vol. 4, n.o 1, 2018, [En línea]. Disponible en: <https://eduneuro.com/revista/index.php/revistaneuronum/article/view/102>
- [8] S. Castro-Casadiago, D. Guevara-Ibarra, L. Acevedo-Jaimes, y B. Medina-Delgado, «Análisis descriptivo de los factores de impacto en las pruebas Saber Pro de estudiantes de Ingeniería Electrónica», *Rev. Educ. En Ing.*, vol. 15, n.o 30, pp. 1-8, 2020. DOI: <https://10.26507/rei.v15n30.1135>
- [9] D. F. Poveda Pineda, J. E. Cifuentes Medina, y J. A. Chacón Benavides, «Apreciación en los resultados de las pruebas Saber Pro», *Rev. Bol.*

Redipe, vol. 10, n.o 12, pp. 271-284, dic. 2021. DOI: <https://10.36260/rbr.v10i12.1587>

[10]A. Carrascal, I. Oviedo, y J. Jiménez-Giraldo, «Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO», Rev. Politécnica, vol. 15, n.o 29, pp. 128-139, 2019. DOI: <https://10.33571/rpolitec.v15n29a10>

[11]J. I. Silva-Ortega, «Implementación de nueva herramienta de seguimiento académico que valida la evaluación por competencias genéricas dentro de la facultad de ingeniería de la Universidad de la Costa (CUC)», Rev. Educ. En Ing., vol. 9, n.o 18, 2014. DOI: <https://10.26507/rei.v9n18.427> .

[12]G.-E. Chanchí-Golondrino, M.-E. Ospino-Pineno, y L.-F. Muñoz-Sanabria, «Application of Spatial Data Science on Results of the Saber 5 Test», Rev. Fac. Ing., vol. 30, n.o 58, 2021. DOI: <https://10.19053/01211129.v30.n58.2021.13823> .

[13]D. A. Solís Flórez, D. F. Alegría-Castrillón, E. Gutiérrez-Vidal, V. Zapata-Bedoya, F. Vidal-Alegría, y R. Timarán-Pereira, «Identificación de Patrones de Rendimiento Académico en las Pruebas Saber Pro entre 2012-2014, en las Competencias Lectura Crítica y Comunicación Escrita con Técnicas Predictivas de Minería de Datos», Cuad. Act., n.o 11, pp. 51-64, 2019. DOI: <https://10.53995/20278101.581>

[14]R. Timarán-Pereira, J. Caicedo-Zambrano, y A. Hidalgo-Troya, «Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11º», Rev. Investig. Desarro. E Innov., vol. 9, n.o 2, pp. 363-378, feb. 2019. DOI: <https://10.19053/20278306.v9.n2.2019.9184> .

[15]R. Timarán-Pereira, A. Hidalgo-Troya, y F. Vidal-Alegría, «Una mirada al desempeño académico en las pruebas saber pro de los estudiantes de ingeniería desde la minería de datos educativa», RISTI Rev. Ibérica Sist. E Tecnol. Informação, n.o E28, pp. 29-42, 2020.

[16]R. Timarán Pereira, A. Hidalgo Troya, y J. Caicedo Zambrano, «Factores asociados al desempeño académico en Lectura Crítica en las pruebas Saber 11º con árboles de decisión», Investig. E Innov. En Ing., vol. 8, n.o 3, pp. 29-37, nov. 2020. DOI: <https://10.17081/invinno.8.3.4701>

[17]A. Timarán Buchely y R. Timarán Pereira, «Minería de datos educativa para descubrir patrones asociados al desempeño académico en competencias genéricas», Rev. Colomb. Technol. Av. RCTA, vol. 2, n.o 38, pp. 87-95, jul. 2023. DOI: <https://10.24054/rcta.v2i38.1282> .

[18]A. Jaramillo y H. Paz-Arias, «Aplicación de técnicas de minería de datos para determinar las interacciones de los estudiantes en un entorno virtual de aprendizaje», Rev. Tecnológica, vol. 28, n.o 1, 2015, [En línea]. Disponible en: <https://dspace.unl.edu.ec/jspui/handle/123456789/11457>

[19]E. Ayala Franco, R. E. López Martínez, y V. H. Menéndez Domínguez, «Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos», Rev. Educ. Distancia RED, vol. 21, n.o 66, abr. 2021. DOI: <https://10.6018/red.463561>

[20]J. E. Sotomonte-Castro, C. C. Rodríguez-Rodríguez, C. E. Montenegro-Marín, P. A. Gaona-García, y J. G. Castellanos, «Hacia la construcción de

un modelo predictivo de deserción académica basado en técnicas de minería de datos - Towards the construction of a predictive model of academic desertion based on data mining techniques», *Rev. Científica*, vol. 3, n.o 26, p. 35, oct. 2016. DOI: <https://10.14483/23448350.11089>

[21] M. Goodchild y R. Haining, «SIG y análisis espacial de datos: perspectivas convergentes», *Investig. Reg.*, n.o 6, pp. 175-201, 2005.

[22] M. Sánchez-Rivero, «Análisis espacial de datos y turismo: nuevas técnicas para el análisis turístico. Una aplicación al caso extremeño», *Rev. Estud. Empres.*, n.o 2, pp. 48-66, 2008.

[23] A. Hernández-Vásquez et al., «Análisis espacial del sobrepeso y la obesidad infantil en el Perú, 2014», *Rev. Perú. Med. Exp. Salud Pública*, vol. 33, n.o 3, p. 489, jul. 2016. DOI: <https://10.17843/rpmesp.2016.333.2298>

[24] J. J. Espinosa Zúñiga, «Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública», *Ing. Investig. Tecnol.*, vol. 21, n.o 1, pp. 1-13, ene. 2020. DOI: <https://10.22201/fi.25940732e.2020.21n1.008>

[25] C. Schröer, F. Kruse, y J. M. Gómez, «A Systematic Literature Review on Applying CRISP-DM Process Model», *Procedia Comput. Sci.*, vol. 181, pp. 526-534, 2021. DOI: <https://10.1016/j.procs.2021.01.199>

[26] C. G. Hidalgo Suarez, V. A. Bucheli Guerrero, F. Restrepo Calle, y F. A. González Osorio, "Estrategia de enseñanza basada en la colaboración y la evaluación automática de código fuente en un curso de programación CS1", *Investigación e Innovación en Ingenierías*, vol. 9, n.o 1, pp. 50-60, ene. 2021. DOI: <https://doi.org/10.17081/invinno.9.1.4185>

[27] F. Martínez-Plumed et al., «CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories», *IEEE Trans. Knowl. Data Eng.*, vol. 33, n.o 8, pp. 3048-3061, ago. 2021. DOI: <https://10.1109/TKDE.2019.2962680>