



Universidad del Norte

Faculty of Engineering

Industrial engineering department

Intervention Framework Based On Geostatistical Models With Language
And Location Models To Focus Intervention Initiatives Associated With
The Consumption Of Psychoactive Substances

Submitted by:

Kevin Rafael Palomino Pacheco

A dissertation submitted in fulfillment of
the requirements for being awarded for the
Doctor degree in Industrial Engineering

Directed by

PhD. Carmen Regina Berdugo Correa

Barranquilla, 2023

Contents

Declaration.....	8
Acknowledgements	9
Publications and presentations.....	10
Abstract.....	12
Chapter 1	14
1. Introduction	14
1.1. Problem Statement.....	14
1.2. Objectives	20
1.3. Significance	21
1.4. Conceptual Framework.....	24
1.5. Research Design	26
1.5.1. First Stage.....	26
1.5.2. Second Stage	26
1.5.3. Third Stage	27
1.5.4. Fourth Stage	27
1.6. Structure of the thesis	29
References.....	30
Chapter 2	35
2. Statistical Analysis and Machine Learning in Psychoactive Substance Use: A Bibliometric Analysis.....	35
2.1. Abstract.....	35
2.2. Introduction.....	35
2.3. Methodology	38
2.4. Results.....	40
2.5. Discussion and conclusion.....	48
References.....	50
Chapter 3	59
3. Leading Consumption Patterns of Psychoactive Substances in Colombia: A Deep Neural Network-Based Clustering-Oriented Embedding Approach.....	59
3.1. Abstract.....	59

3.2. Introduction.....	60
3.3. Materials and Methods.....	63
3.3.1. Study Area.....	63
3.3.2. Data Sources.....	63
3.3.3. Convolutional Auto-Encoder-Deep Embedded Clustering Algorithm.....	64
3.3.4. Convolutional Auto-Encoder (CAE).....	66
3.3.5. Clustering Layer	67
3.3.6. The CAE-DEC Model.....	68
3.3.7. Framework Evaluation	69
3.4. Results.....	70
3.4.1. Model Comparison for Identifying Drug Consumption Patterns	70
3.4.2. Identification of Psychoactive Drugs Consumption Clusters	71
3.4.3. Spatial Analysis of Psychoactive Drugs Consumption.....	75
3.3.4. Regionalization of Clusters	80
3.5. Discussion and Conclusion	82
References.....	87
Chapter 4	96
4. Bi-Objective Location-Allocation Model of Interventions in High Drug Consumption Areas Incorporating Twitter Topic Modeling	96
4.1 Abstract.....	96
4.2 Introduction.....	96
4.3 Model formulation	98
4.4 Epsilon constraint method	101
4.5 Twitter topic modeling.....	102
4.6 Case study	104
4.6.1 Model parameters.....	105
4.6.2 Computational results and insights	109
4.6.3 Topic modelling results.....	116
4.6 Conclusion	122
References.....	124
Chapter 5	129
5. Data Analytics and Mental Health: Would Ethics Be the Only Safeguard Against the Risks of Identifying "Potential Patients"?	129
5.1. Abstract.....	129

5.2. Introduction.....	129
5.3. Data Analytics in Mental Healthcare	131
5.3.1. Some Emerging Trends	131
5.3.2. Medication Selection.....	131
5.3.3. Suicide Prevention.....	132
5.3.4. Outcome Monitoring and Treatment.....	132
5.3.5. Ethical Data Usage	133
5.4. Distinctive Ethical Questions.....	134
5.4.1. How Devastating Could a Risk Assessment be for a Medical Patient?	134
5.4.2. Can Mental Health Risk Assessments Affect Insurance Costs?	135
5.4.3. How Should Social Media be Used to Promote Mental Health?	135
5.4.4. How Much Informed Consent Do Doctors Require for Data Analytics Work?	136
5.5. Ethics, the Only Safeguard?.....	136
5.5.1. Balancing Precision and Ethics in Data Analytics	137
5.5.2. Who Is Responsible for Considering Treatment Decisions and Ethics?.....	138
5.5.3. Ethical Imperatives in Healthcare Institutions	139
5.5.4. Discussion and Conclusion	140
References.....	142
Chapter 6	145
6. Conclusions and Further Directions	145
6.1. Conclusions.....	145
6.2. Limitations and Future Research	149
Appendix A	153
Appendix B.....	153
Appendix C.....	153

Figures

Figure 1. Total number of global deaths due to psychoactive substances consumption.....	15
Figure 2. The number of cases of persons hospitalized for mental disorders.....	17
Figure 3. GEOTELO Framework.....	28
Figure 4. Bibliometric analysis flow chart.....	39
Figure 5. Publications per year.....	41
Figure 6. Performance of authors.....	42
Figure 7. Journal performance.....	43
Figure 8. Network visualization map of author keywords.....	45
Figure 9. Performance of countries.....	46
Figure 10. Architecture of the proposed CAE-DEC model.....	65
Figure 11. Derived Latent Feature Space based on the (a) CAE and (b) CAE-DEC models.....	71
Figure 12. Resulting clusters for individuals.....	72
Figure 13. Absolute deviation around class medians (ADCM).....	75
Figure 14. Cluster maps.....	76
Figure 15. Maps of Local Indicators of Spatial Association (LISA).....	77
Figure 16. Frequency of consumption of (a) legal and (b) illegal drugs in Colombia.....	78
Figure 17. Consumption of illegal drugs by department.....	79
Figure 18. Consumption of legal drugs by department.....	80
Figure 19. Cluster map of drug use after regionalization.....	81
Figure 20. Definition of the optimal number of clusters based on a Dendrogram.....	81
Figure 21. Health center location.....	105
Figure 22. Distribution of positive, neutral, and negative comments per location.....	108
Figure 23. Sentiment algorithms results.....	109
Figure 24. Pareto frontier.....	110
Figure 25. Flow network.....	114
Figure 26. Sensitivity analysis of pi and alpha.....	115
Figure 27. Coherence score over the number of topics.....	117
Figure 28. Top nine topic word clouds.....	121

Tables

Table 1. Total number of publications	40
Table 2. Top 10 most cited publications.	47
Table 3. Pseudo code for the CAE-DEC training process.	69
Table 4. Distribution of demographic and social variables across clusters.....	72
Table 5. Adjusted residuals comparing the observed and expected frequencies.	73
Table 6. Performance metrics for different models.	74
Table 7. Feature coherence measurements (goodness of fit).	82
Table 8. Level of development, urbanity, rurality, and drug production.	84
Table 9. Distribution of tweets per location.....	109
Table 10. Prevention hour distribution in each health center.....	111
Table 11. Mitigation hour distribution in each health center	112
Table 12. Demand coverage (< 40 km)	113
Table 13. Health center capacity.....	114
Table 14. Sample of the collected tweets.....	116
Table 15. Topic Contribution.....	122

There is no health without mental health.

Declaration

This work entitled “Intervention Framework Based On Geostatistical Models With Language And Location Models To Focus Intervention Initiatives Associated With The Consumption Of Psychoactive Substances” was conducted between January 2019 and December 2022 at the Universidad del Norte in Colombia. This thesis represents my original research for the purpose of obtaining the PhD degree from the Universidad del Norte. Unless explicitly cited, all ideas and findings presented in this document are my own. Furthermore, I confirm that this thesis has not been submitted for any other academic qualification at any other institution.

Kevin Rafael Palomino Pacheco
Barranquilla, Colombia
Abril 2023

Acknowledgements

I would like to express my deepest gratitude and appreciation to my loving family, who have been my unwavering source of support and strength throughout this journey. Their unconditional love, encouragement, and understanding have played an instrumental role in my accomplishments, and for that, I am eternally grateful. To my parents, thank you for your unwavering belief in me and for instilling in me the values of hard work, perseverance, and determination. Your sacrifices and tireless efforts have paved the way for my success, and I am indebted to you both.

First and foremost, I express my sincere thanks to the Universidad del Norte, Colombia, for awarding me a doctoral fellowship and providing me with the necessary resources and facilities to pursue my research. I extend my heartfelt gratitude to the Department of Industrial Engineering for providing me with the opportunity to work on this project and for their support throughout the years. I am indebted to Dr. Carmen Regina Berdugo Correa for her invaluable guidance, insightful feedback, and unwavering encouragement during the course of this research.

I am also grateful to Dr. Carmen Berdugo for her assistance and support, which was critical to the successful completion of my work. Their expert advice and guidance helped me navigate some of the most challenging aspects of my research.

I would like to extend my heartfelt appreciation to the Department of Industrial and Management Systems Engineering at the University of South Florida (USF) for their warm hospitality and unwavering support during my research internship. Their generous guidance, insights, and suggestions have significantly enhanced the quality of my research. I am profoundly grateful to Dr. Tapas, Dr. Zayas, Dr. Acuna, and Dr. Silva, as well as all those who have contributed to my research internship. Their exceptional expertise, guidance, and support have been invaluable, and I consider myself fortunate to have had the privilege of working alongside such esteemed professionals.

Finally, I am grateful to the people who have contributed to my journey toward earning a PhD degree. Without their support, this accomplishment would not have been possible.

Thank you all from the bottom of my heart.

Publications and presentations

Published Papers

- 1. Palomino, K., Garcia, D., & Berdugo, C. (2022).** "A MILP facility location model with distance value adjustments for demand fulfillment using Google Maps". *Journal of Engineering Research*, 10(2A), 270–291. <https://doi.org/10.36909/JER.10473>. This manuscript is based on integrating georeferenced data using Google Maps with location-allocation model presented in **Chapter 4**.
- 2. Palomino, K., & Berdugo, C. (2023).** "Statistical analysis and machine learning in psychoactive substance use: a bibliometric analysis," *Nexo Rev. Científica*, vol. 36, no. 02, pp. 96–109, Mar. 2023, doi: 10.5377/NEXO.V36I02.16017. This manuscript is based on the results presented in **Chapter 2**.

Papers Under Review

- 1. Palomino, K., Berdugo, C., Velez, J. I.** "Leading Consumption Patterns of Psychoactive Substances in Colombia: A Deep Neural Network-based Clustering-oriented Embedding Approach". This manuscript is based on the results presented in **Chapter 3**, and it is under review in PLOS ONE.
- 2. Palomino, K., Berdugo, C.** "Big Data Analytics and Mental Health: Would Ethics be the Only Safeguard Against the Risks of Identifying Potential Patients?" This manuscript is based on the results presented in **Chapter 5**, and it is under review in IEEE Intelligent Systems.

Papers in preparation

1. **Palomino, K.,** Berdugo, C., Acuna, J., & Zayas, J. "Bi-objective location-allocation model of interventions in high drug consumption areas incorporating Twitter topic modeling". This manuscript is based on the results presented in **Chapter 4**.
2. **Palomino, K.,** Jafaripakzad, M., Berdugo, C., Acuna, J., & Zayas, J. "Identifying patterns of psychoactive substances consumption for high dimensional datasets through deep learning algorithms: A comparison of clustering methods based on experimental designs". This manuscript is based on the model proposed presented in **Chapter 3** and the social media analysis model in **Chapter 4**.

Presentations

1. **Palomino, K.,** Berdugo, C. "Actitudes Sexistas Como Predictores De Violencia: Modelo Estructural Confirmatorio". 4° Congreso Internacional De Investigación Multidisciplinaria. Instituto Tecnológico Superior De La Sierra Negra De AJALPAN. May 2021.
2. **Palomino, K.,** Berdugo, C. "A mixed integer linear programming model for facility location in disaster relief operations using georeferenced data". Expo Ingenieria. Universidad de Antioquia. October 2022.
3. **Palomino, K.,** Berdugo, C. "Cooperative game study of airlines based on flight price optimization in times of Covid-19". International congress of industrial and mechanical engineering. Universidad Pontificia Bolivariana. October 2022.

Abstract

Mental health issues are becoming more widespread among the global population, posing a substantial burden on public health and resulting in significant societal costs. These costs arise from factors such as lost productivity, disability, early death, increased healthcare expenses, as well as criminal justice and social welfare expenditures, among other social ramifications. Over the past twenty years, the swiftly growing body of research, along with its broad geographical scope, has underscored the escalating worldwide apprehension surrounding psychoactive substance use (PSU). Despite these advancements, a prominent gap remains in the literature regarding the utilization of intervention frameworks that combine geostatistical models, natural language processing, and location-allocation models, which could facilitate the identification and implementation of preventive and mitigating measures concerning psychoactive substances. In this context, the primary aim of this study was to develop a comprehensive intervention framework, grounded in a geostatistical model that incorporates linguistic and geographical components (referred to as GEOTELO). This framework was specifically designed to detect and pinpoint statistically significant clusters of spatial and unstructured data. By doing so, it enables targeted intervention strategies to be implemented, focusing on addressing the use of psychoactive substances within these specific areas.

A research framework was developed to achieve our objectives, comprising four distinct stages. In the initial stage, sociodemographic and spatial patterns are examined utilizing a Deep Neural Network-based Clustering-oriented Embedding Algorithm. Two databases are leveraged to discern drug consumption patterns in Colombia. The primary database is sourced from the 2019 National Survey of Psychoactive Substance Consumption in the General Population, conducted by Colombia's National Statistical System (DANE). This survey encompasses data from 49,600 households, detailing information on housing, location, individual characteristics, consumption of legal and illegal psychoactive substances (PAS), and applied treatments. The secondary database is derived from the Colombian Drug Observatory and provides information on PAS production by area during 2019. By executing this stage, spatial consumption patterns of PAS are pinpointed,

and construct an ensemble algorithm that integrates an autoencoder, a clustering algorithm, and a spatial model to address the feature space and clusters. As a subsequent step, data from Twitter was extracted to analyze consumers' opinions about psychoactive substances, using the information acquired from the previous stage as input. First, an algorithm was developed to retrieve posts over a specific period and store them in a Postgres SQL database. After gathering all the posts from areas (states or departments) with high drug consumption, topics related to psychoactive substances were identified within the Twitter data. To extract these subjects, unsupervised text modeling was employed using Latent Dirichlet Allocation (LDA).

In the third stage, a location-allocation framework was constructed to optimize intervention policies under resource constraints, with the aim of enhancing population health outcomes. The model is founded on a bi-objective integer programming structure for the location and allocation of health centers and consumers. It considers the reduction of overall risk associated with drug consumption and the minimization of the distance between patients and facilities while ensuring an equitable distribution of facilities among the population. Lastly, a communication tool was developed for presenting results in a way that facilitates monitoring, visualization, and informed decision-making regarding the issue of psychoactive substances.

Chapter 1

1. Introduction

The proposed doctoral thesis aims to develop an intervention framework that leverages the power of geostatistical, clustering, natural language processing, and location-allocation models. This integrated approach would be used to analyze structured and unstructured georeferenced data related to psychoactive substance consumption. By identifying statistically significant patterns in such data, the framework could provide valuable insights into substance use patterns and related factors. The problem statement driving this research is the increasing incidence of substance use disorders, which pose significant public health challenges. Current approaches to studying substance use patterns often lack granularity and do not account for the complexity and diversity of factors influencing substance use behaviors. By integrating various modeling approaches, the proposed framework could offer a more comprehensive understanding of substance use patterns and associated factors. The proposed framework's justification lies in its potential to inform public health policies and interventions. By providing a more nuanced understanding of substance use patterns, the framework could help policymakers and healthcare professionals design targeted interventions and prevention strategies. The problem statement and justification are presented as follows:

1.1. Problem Statement

Mental health problems are increasingly prevalent in the world population, presenting one of the highest disease burdens [1]–[3], and represent a high cost to society due to lost productivity, disability, premature mortality, increased health care spending, criminal justice and social welfare costs, and other social consequences [4]–[6]. According to the mental health action plan issued by the World Health Organization (WHO), there are concerns about caring for the population affected by problems associated with mental health [7]. This report shows that 76% and 85% of people with severe mental disorders do not receive treatment in low-income countries and between 35% and 50% in high-

income countries. Furthermore, according to the World Economic Forum "*The Global Economic Burden of Non-communicable Diseases*" conducted by the Harvard School of Public Health¹, the cumulative global impact of mental disorders in terms of economic losses will be US\$ 16.3 trillion between 2011 and 2030. On the other hand, the WHO reports that mental disorders influenced by psychoactive substances represent one of the highest global morbidity burdens (13%). For this reason, there is a need to promote health policies or plans that guarantee priority care for people with mental disorders, especially regarding psychoactive substances. In this sense, WHO proposed the "Universal Health Coverage for Mental Health" initiative in 2019, seeking to contribute to poor health outcomes, reducing premature deaths, human rights violations, and global economic losses, with sustained funding for large-scale health mental services [8].

The Institute for Health Metrics and Evaluation (IHME) has reported an alarming trend of increasing deaths caused by psychoactive substances in recent years, with a record high of around 300,000 deaths in 2019 alone (as shown in Figure 1).

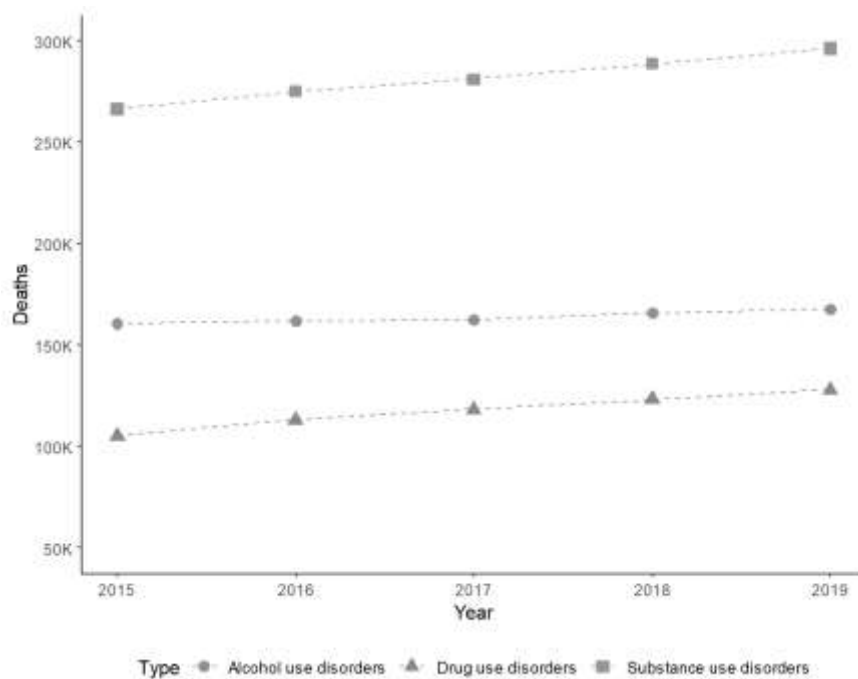


Figure 1. Total number of global deaths due to psychoactive substances consumption.

Source: Author's elaboration taken from Global Burden of Disease Collaborative Network [10].

¹ World Economic Forum, the Harvard School of Public Health. The global economic burden of non-communicable diseases. Ginebra, Foro Económico Mundial, 2011

Notably, the number of deaths associated with alcohol use is higher than those related to drug use, including opioids, cocaine, amphetamine, and cannabis, among others. These trends indicate a significant public health concern that warrants urgent attention. Moreover, the World Health Organization (WHO) conducted a survey across 130 countries and found that the COVID-19 pandemic has resulted in increased demand for mental health services. The isolation, loss of income, loss of a family member, and fear caused by the pandemic have all contributed to an increased need for mental health support. However, the pandemic has also led to a significant reduction in access to hospital mental health services, with 93% of countries experiencing such reductions. These findings suggest that addressing the issue of substance use disorders and mental health during the pandemic is crucial. Adequate access to mental health services is essential, particularly for vulnerable populations, such as those with substance use disorders. Effective policies and interventions should be designed to reduce the negative impact of substance use disorders and ensure adequate access to mental health services, especially during challenging times like the COVID-19 pandemic. Consequently, it is inferred that there is a high risk for people facing higher levels of alcohol and drug use [9].

In Colombia, despite efforts to adopt guidelines for the promotion of health policies or plans, the country is still lagging in complying with the objectives of mental health regulations; this is evidenced in a study conducted by Rojas-Bernal [11], which shows: "shortcomings in the mandatory health plan in 1998 with the national mental health policy", "exclusion of psychotherapy in Law 100 of 1993", "failure to guarantee comprehensive care in Decree 3039 of 2007" and "limited health access in the ten-year public health plan.". The inadequate implementation and interpretation of mental health policies and regulations in Colombia have resulted in a lack of coherence between what is proposed and what is happening in reality. This discrepancy is evident in the increasing number of cases associated with mental disorders due to psychoactive substances in recent years, as reported by the Colombian Drug Observatory (ODC). The trend has continued to rise since 2019, with the COVID-19 pandemic exacerbating the situation (See Figure 2). The increasing numbers of mental health cases underscore the urgency of addressing the shortcomings in mental health policies and regulations. By adopting a

comprehensive and coordinated approach to mental health care, Colombia can help prevent further harm to its citizens and ensure their overall well-being.

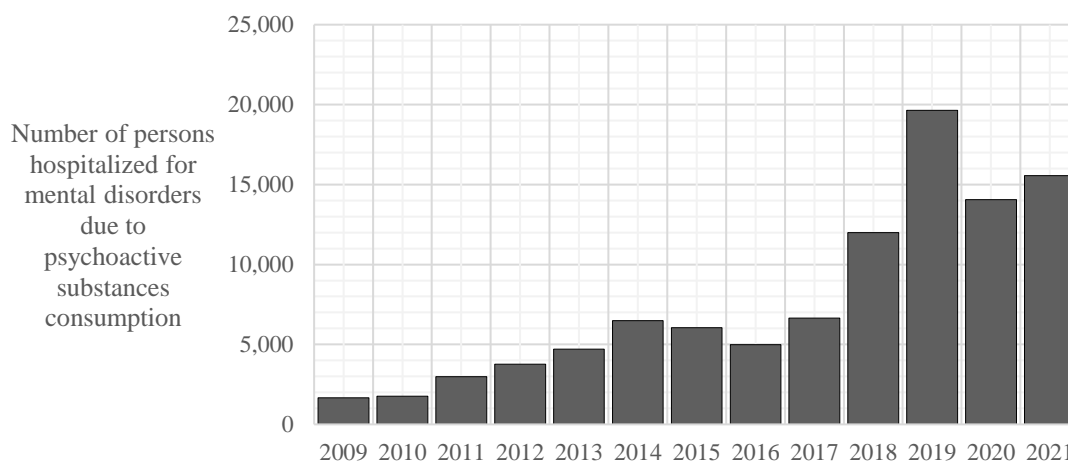


Figure 2. The number of cases of persons hospitalized for mental disorders.

Source: Author's elaboration taken from: <<http://rssvr2.sispro.gov.co/ObsSaludMental/>>

Other findings are those found by Gómez Restrepo [12] through the application of the National Survey of Mental Health 2015, revealing that the consumption of tobacco, alcohol, marijuana, cocaine, heroin, or other psychoactive substances, has increased in children between the ages of 7 and 11 years in Colombia, highlighting some associated risk factors such as genetic vulnerability, family environment (parents consumers), lack of information and documentation, limited risk perception of initiating consumption, a victim of physical or psychological mistreatment, sexual abuse, influences by friends, availability and easy access to substances, psychiatric disorders (depression or suicide attempt and neurodevelopmental alterations). In other words, the local context in which children, adolescents, and young people develop leads to a significantly higher risk of consuming psychoactive substances [13]. The National Statistical System (DANE) conducted the latest National Survey on Consumption of Psychoactive Substances in 2019, revealing that 33.3% of individuals between the ages of 12 and 65 have reported ever using tobacco or cigarettes in their lifetimes. The survey also showed that the highest percentage of consumption was reported in Cundinamarca (45.3%) and Boyacá (42.3%), while the lowest proportions were found in Archipelago the San Andrés y Providencia (7.9%) and Sucre (15.2%). These findings highlight the need to address the prevalence of tobacco use in Colombia and to implement targeted interventions and prevention strategies in regions with the highest consumption rates.

Likewise, 84.0% of persons between 12 and 65 years of age reported having consumed alcohol, with Boyacá (92.9%) and Risaralda (92.5%) registering the highest prevalence of consumption of this type of drug. In comparison, the Archipelago of San Andrés y Providencia showed the lowest rates (54.7%), followed by Guainía (65.7%). Meanwhile, 9.7% of the population reported having consumed some illegal psychoactive substance (inhalants, Methylene Chloride, popper, marijuana, cocaine, basuco, ecstasy, heroin, methamphetamine, LSD, mushrooms, yagé, ketamine, GHB or 2CB), having a Putumayo (25.6%) and Risaralda (20.6%) with the highest prevalence rates; In contrast, the lowest percentages were recorded in the Archipelago the San Andrés y Providencia (0.3%) and Chocó (1.6%) [14]. Considering that mental health refers mainly to all those risks associated with a mental health disorder such as epilepsy, gender-based violence, domestic violence, sexual abuse, psychoactive substances consumption (PAS), schizophrenia, mental retardation, attention deficit hyperactivity disorder (ADHD), this research intends to address only the PAS dimension. However, this term in the literature is equivalent to: "drug abuse", "drug use", "psychotropic substances consumption", "abuse of psychotropic substances," or "abuse of psychoactive substances". Thus, (WHO) defines psychoactive substances as the consumption of substances that affect mental processes so that, if not controlled, they can provoke personal, family, social, educational, and other problems. In summary, mental health disorders are due to psychotropic substances such as alcohol; opioids; cannabinoids; sedatives or hypnotics; cocaine; other stimulants, including caffeine; hallucinogens, tobacco, and volatile solvents, among others [15].

The evolution of research on psychoactive substance consumption has been increasing but not constant and has been oriented towards different approaches. Just to mention a few examples, some studies focus on adolescent drug abuse [16]–[22], others are directly associated with psychological treatments based on social and emotional learning (SEL) or pharmacological treatments [23]–[26], and the last ones are related to the social and geographic implications of drug trafficking [27]–[32]. In turn, some researchers states that, in studying a given disease, it is essential to know when and where it develops and how it is distributed in the geographic space [33]. In this way, it is possible to establish hypotheses about its causes, helping to understand its etiology better and its risk factors.

As a complement to the above, identifying spatial clustering patterns in diseases is very useful for Public Health policies and planning. This knowledge can support decisions on the location and allocation of new resources, the management of existing resources, the definition of actions aimed at priority diseases, and the design of prevention, vigilance, and control programs [34]. In the academic literature, in recent years, some studies show the application of spatial analysis to locate disease patterns [35]–[39]; others focus on machine learning tools, especially in text processing language and big data, to analyze text patterns in social networks [41]–[46]. In contrast, others focus on locating health centers to mitigate psychoactive substance consumption [40]–[44]. However, there is no framework that integrates a geostatistical model, a text-processing language model, and an optimization model that responds with topics and intervention points to address global health problems associated with psychoactive substances.

A comprehensive and effective solution for addressing the increasing incidence of substance use disorders and their impact on public health, requires an integrated approach. This can be achieved through the incorporation of geostatistical modeling, text-processing language modeling, and optimization modeling. Such an approach would enable the identification of significant patterns of substance use and the factors associated with it. Moreover, it would facilitate the development of targeted interventions and prevention strategies to tackle the problem effectively. Given the significance of addressing substance abuse and its associated problems, the development of an integrated framework that combines these different models is crucial. Consequently, the research question that needs to be directed is,

How to design an intervention framework based on a geostatistical, language, and location-allocation model that allow focusing intervention initiatives associated with psychoactive substance use?

1.2. Objectives

Design an intervention framework based on geostatistical models with language and location models (GEOTELO) identifying and locating statistically significant clusters of spatial data and unstructured data to target intervention initiatives associated with the consumption of psychoactive substances.

- Conduct a review of the existing literature, including as many relevant studies as possible, to synthesize their findings and offer a comprehensive perspective on the development of the research field over the past few decades.
- Identify psychoactive substance consumption patterns using machine learning and geospatial tools with geo-referenced data to gain valuable insights into the demographics, behaviors, and trends associated with substance use.
- Delimit the different topics geographically through language models in social networks (Twitter) about psychoactive substances to enable a comprehensive understanding of local discussions, trends, and perceptions related to drug use.
- Design a location-allocation model to optimize intervention policies under resource constraints using a bi-objective integer programming to simultaneously minimize the distance between patients and facilities and maximize health outcomes considering an equitable distribution of the facilities among citizens.
- Design an information tool for communicating and controlling data on spatial groupings and geo-referenced topics associated with psychoactive substances.

1.3. Significance

As defined by the World Health Organization (WHO), mental health encompasses promoting well-being, preventing mental disorders, and treating and rehabilitating individuals affected by such disorders. It is regarded as an essential component of overall health, to the extent that "*there can be no health without mental health*" [45]. In this sense, WHO defines health as "a state of complete physical, mental and social well-being, and not merely the absence of disease. However, mental health remains a neglected part of global efforts to improve health, with people with mental health problems suffering widespread human rights violations, discrimination, and stigmatization, which is supported by global statistics, i.e., more than 80% of people with mental health problems, including neurological and substance use disorders, lack any quality and affordable mental health care.[8]. In this way, the need arises to analyze this problem from different sciences and disciplines. Industrial engineering is a field of applied knowledge as defined by the Institute of Industrial Engineering (IIE):

"It is a field of knowledge and professional performance concerning designing, improving, and installing integrated systems of people, materials, information, equipment, and energy. It draws upon specialized knowledge and skill in the mathematical, physical, and social sciences, together with engineering analysis and design principles and methods, to specify, predict, and evaluate the results obtained from such systems". [46]

Following this same line, one of the functions of industrial engineering is to improve the conditions of a system by controlling the variables involved in that system. In this sense, applied research has been implemented to control variables affecting human health in all disciplines [47]. Hence, general health can be considered an industrial engineering problem. To exemplify, The Royal Academy of Engineering argues that modern medicine and health depend heavily on engineering to improve disease prevention, diagnosis, and treatment, using efficient technologies in medical imaging, cardiac implants, neuro-engineering, artificial joints, telemedicine, and regenerative medicine, among others [48], [49]. Recently, engineering has played a vital role in serving and advancing healthcare, so many scientists and professionals have adopted a new term called healthcare

engineering, defined as *"all aspects of the prevention, diagnosis, treatment, and management of disease, as well as the preservation and improvement of physical and mental health and well-being, through the services offered to human beings by the medical and allied professions."*[50].

The development of scientific knowledge requires the emergence of new and diverse proposals in any field of research. Therefore, it is necessary to design intervention framework to address mental health problems associated with psychoactive substances. Currently, there is a lack of intervention framework that integrate geostatistical models with natural processing language and location-allocation models, which can help locate and mitigate psychoactive substance-related programs. This research project will contribute to constructing a theoretical framework on this topic, facilitating future research on drug use and consumption. Furthermore, it aims to explore a new approach or methodology that may be more effective in supporting national mental health objectives. The study will provide reliable and specific information for the formulation of mental health plans and policies, thus facilitating intervention among the target population and strengthening health epistemological indicators in Colombia. For this reason, it is of great interest to apply spatial data analysis tools to detect possible statistically significant spatial concentrations (disease clustering) that indicate the possible existence of common risk factors for a given disease or transmission processes by proximity. [51]. In sum, the integration of language models is intended to obtain fundamental topics related to drug use through information obtained from social networks (Twitter). [52], [53]. Based on the above, a mathematical model will be designed to locate intervention programs under resource constraints to improve population health outcomes under the characteristics extracted from the geostatistical and language model.

Furthermore, this study highlights the potential to undertake research that establishes a comprehensive system for addressing mental health risks in any geographical location. It also aims to enhance our understanding of the spatial behavior of the variables studied in PAS, which is the primary focus of this investigation. By undertaking this approach, a more comprehensive understanding of the factors contributing to mental health risks can be gained and develop effective strategies to address them. All of the above is of

enormous value for the current Colombian context if it takes into account that, at the moment, there have been many interventions for the treatment of psychoactive substance use in the country, many of them without evidence of effectiveness since they are designed for a single treatment or a single objective and do not respond to the needs of each one of the existing patients [54]. Finally, this research is noteworthy for its contribution to the attainment of national and international following objectives in mental health:

- Colombia's National Development Plan (2018-2022) 3005-III. Pact for Equity: a modern social policy that is family-centered, efficient, high-quality, and connected to markets. Strategic line: 300503-2: Health for all with quality and efficiency, sustainable by all.
- Sustainable Development Goals, SDG 3: Ensure healthy lives and promote well-being for all at all ages; SDG 17: Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development.

1.4. Conceptual Framework

- **Mental health:** Mental health is defined as an integrated system between health and well-being, which is established in one of the principles of the WHO constitution -*Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity*- and that health is a fundamental right that every individual should enjoy without distinction of race, religion, political ideology, or economic and social condition [7]. Therefore, there is no health without mental health.
- **Psychoactive Substance:** Any substance introduced into the body by any route of administration (ingested, smoked, inhaled, injected, among others) produces an alteration in the functioning of the central nervous system of the individual, which modifies the consciousness, mood, or thought processes, leading to dependence if its consumption is frequent [55].
- **Geostatistics:** It is a quantitative study of phenomena located in space that brings together a set of techniques for analyzing topological, geometric, and geographical properties of spatial data. In others words, this statistical model is used to describe, quantify and explain geographic variations in disease; assess the relationship between incidence, prevalence, and morbidity of disease and potential risk factors; identify geographic clusters of disease[56].
- **Artificial Intelligence (AI):** It is the study and analysis of human behavior in the areas of comprehension, perception, problem-solving, and decision-making to reproduce them with the help of a computer. Thus, the applications of AI are mainly in the simulation of human intellectual activities [58].
- **Language models:** Language models are models that are based on a probabilistic description of language phenomena. In this sense, language models have many uses, such as part-of-speech labeling, parsing, machine translation, handwriting recognition, speech recognition, and information retrieval. [59], [60].

- **Optimization:** Optimization is the process of selecting the best solution among the set of candidate solutions, the degree of goodness of the solution being the objective function to be minimized or maximized. The search process is carried out based on the system model and constraints. Thus, the objective of the optimization is to maximize (or minimize) the value of a function subject to a set of constraints. These constraints take the form of equality and inequality expressions [61].

1.5. Research Design

This research aims to develop an intervention framework that combines geostatistical models with language and location models in order to detect and categorize statistically significant clusters of spatial and unstructured (text) data. The GEOTELO framework has been assessed through a case study focused on the factors of consumption and production associated with mental disorders linked to the use of psychoactive substances in Colombia. The data employed originates from a DANE database, obtained from the National Survey of Psychoactive Substances in 2019, Colombian Drug Observatory (ODC) drug production data, and tweets. The framework primarily consists of four separate stages, which are outlined as follows:

1.5.1. First Stage

In this stage, sociodemographic and spatial patterns were analyzed using a Deep Neural Network-based Clustering-oriented Embedding Algorithm. Two databases to identify drug consumption patterns in Colombia were used. The first database was retrieved from the 2019 National Survey of Psychoactive Substance Consumption in the General Population conducted by Colombia's National Statistical System (DANE) [64]. This survey includes observations of 49,600 households, where information on housing, location, general characteristics of individuals, consumption of legal and illegal PAS, and implemented treatments is registered. The second database comes from the Colombian Drug Observatory and contains information on the production of PAS per area during 2019. The implementation of this stage allows us to *(i)* identify spatial consumption patterns of PAS; and *(ii)* build an ensemble algorithm integrating an autoencoder with a clustering algorithm and a spatial model to deal with the feature space and clusters.

1.5.2. Second Stage

Social networks such as Facebook, LinkedIn, and Twitter have been crucial sources of information for a broad spectrum of users. In this study, data from Twitter was extracted to analyze the consumers' opinions about psychoactive substances using as input the information obtained from the previous stage. First, an algorithm was built to pull posts over some time and store them in a database in Postgres SQL. Once all the posts were

obtained from areas (States or departments) with high drug consumption, it was identified the topics available on Twitter data related to psychoactive substances. Then, an unsupervised text modeling with Latent Dirichlet Allocation (LDA) was used to extract the different subjects. Also, a scheme was implemented to evaluate the data quality which examines each user's post in the data set to ensure it includes related terms from the ontology proposed by Nasrallah [53].

1.5.3. Third Stage

The Comprehensive Policy for the Prevention and Care of Psychoactive Substance Use in Colombia considers prevention and mitigation programs that aim to improve the care of individuals, families, and communities at risk or with problematic use of psychoactive substances. The effectiveness of these programs depends on the participation level and user accessibility to the facilities that provide the services. Factors influencing access include facility type, location, and the number of patients assigned to the facility. Therefore, in this stage a location-allocation structure was built to optimize intervention policies under resource constraints to improve population health outcomes. The model is based on a bi-objective integer programming structure for the location and allocation of health centers and consumers. The objectives considered are I) to reduce the overall risk of drug consumption, and II) to minimize the distance between patients and facilities, considering an equitable distribution of the facilities among citizens.

1.5.4. Fourth Stage

Finally, in this stage, a tool for communicating results was implemented to help monitor, visualize, and support decision-making on the problem of psychoactive substances. For this purpose, a dashboard is created using the Power Bi Desktop, a business analytics and data visualization software developed by Microsoft. This platform allows users to create interactive and engaging visualizations with a wide range of chart types, tables, matrices, maps, and custom visuals.

The source code is available on GitHub at the following path:



`<https://github.com/Mental-Health-Framework/Stage-4.git>`

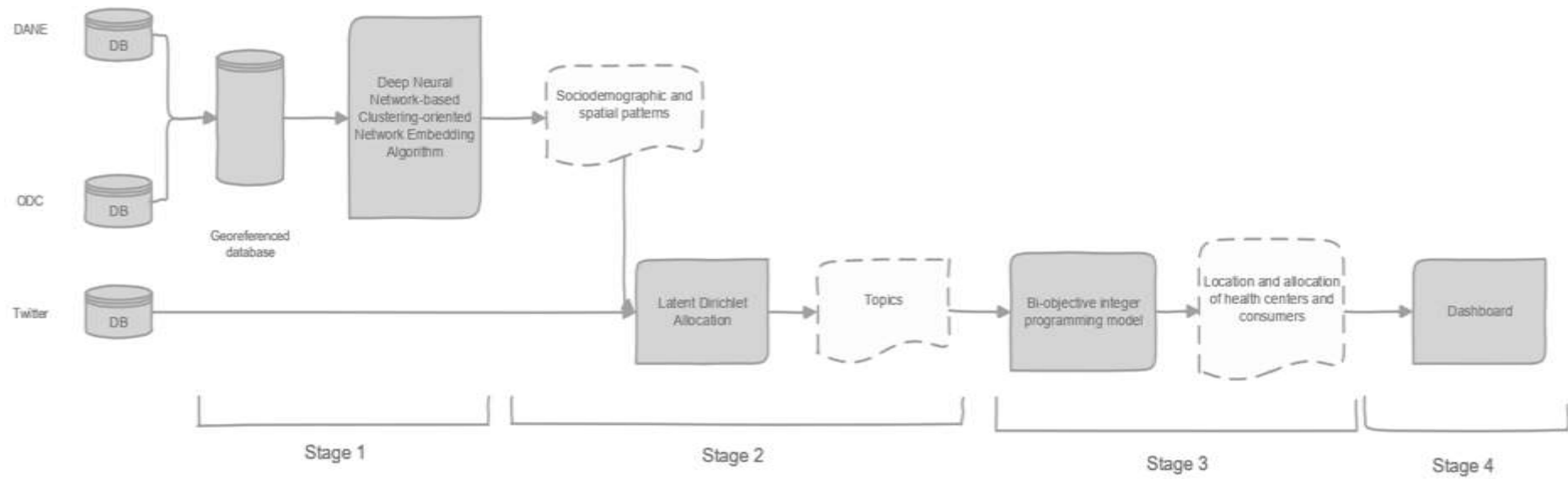


Figure 3. GEOTELO Framework.

1.6. Structure of the thesis

This section provides an overview of the dissertation's structure, which comprises six chapters primarily based on articles. A brief description of the chapters is shown below:

- Chapter 1 (Introduction): presents the problem statement, research objectives, significance, conceptual framework, and research design. Its purpose is to establish the context and importance of the study while providing a clear roadmap for the research.
- Chapter 2 (Statistical Analysis and Machine Learning in Psychoactive Substance Use: A Bibliometric Analysis): provides a comprehensive literature review of Bayesian Analysis, Spatial Analysis, Multivariate or Univariate Statistical Analysis, and Machine Learning techniques applied to model problems related to psychoactive substance use.
- Chapter 3 (Leading Consumption Patterns of Psychoactive Substances in Colombia: A Deep Neural Network-Based Clustering-Oriented Embedding Approach): presents the design of an ensemble model that integrates an autoencoder with clustering and spatial models to find sociodemographic and spatial patterns of georeferenced data. This section corresponds to the GEOTELO framework stage 1.
- Chapter 4 (Bi-Objective Location-Allocation Model of Interventions in High Drug Consumption Areas Incorporating Twitter Topic Modeling): shows the design of an optimization model to locate and allocate health centers and consumers. In addition, geo-referenced topics extracted from Twitter data related to psychoactive substances are presented. This section corresponds to the GEOTELO framework stage 2 and stage 3.
- Chapter 5 (Data Analytics and Mental Health: Would Ethics Be the Only Safeguard Against the Risks of Identifying "Potential Patients"?): explores the use of data analytics in mental health and the ethical implications of identifying "potential patients." It argues that ethics should serve as the primary safeguard against such risks.
- Chapter 6 (Conclusion): concludes with an overview of findings and implications, as well as suggestions for future research directions.

References

- [1] B. F. Grant *et al.*, “Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders - Results from the national epidemiologic survey on alcohol and related conditions,” *Arch. Gen. Psychiatry*, vol. 61, no. 8, pp. 807–816, 2004, doi: 10.1001/archpsyc.61.8.807.
- [2] R. C. Kessler, C. B. Nelson, K. A. McGonagle, M. J. Edlund, R. G. Frank, and P. J. Leaf, “The epidemiology of co-occurring addictive and mental disorders: Implications for prevention and service utilization,” *Am. J. Orthopsychiatry*, vol. 66, no. 1, pp. 17–31, 1996, doi: 10.1037/h0080151.
- [3] D. A. Regier *et al.*, “Comorbidity of Mental Disorders With Alcohol and Other Drug Abuse: Results From the Epidemiologic Catchment Area (ECA) Study,” *JAMA*, vol. 264, no. 19, pp. 2511–2518, 1990, doi: 10.1001/jama.1990.03450190043026.
- [4] R. Z. Goetzel, K. Hawkins, R. J. Ozminkowski, and S. H. Wang, “The health and productivity cost burden of the ‘top 10’ physical and mental health conditions affecting six large US employers in 1999,” *J. Occup. Environ. Med.*, vol. 45, no. 1, pp. 5–14, 2003, doi: 10.1097/00043764-200301000-00007.
- [5] K. Sanderson and G. Andrews, “Prevalence and severity of mental health-related disability and relationship to diagnosis,” *Psychiatr. Serv.*, vol. 53, no. 1, pp. 80–86, 2002, doi: 10.1176/appi.ps.53.1.80.
- [6] W. F. Stewart, J. A. Ricci, E. Chee, S. R. Hahn, and D. Morganstein, “Cost of lost productive work time among US workers with depression,” *JAMA-JOURNAL Am. Med. Assoc.*, vol. 289, no. 23, pp. 3135–3144, Jun. 2003, doi: 10.1001/jama.289.23.3135.
- [7] WHO, “Mental Health Action Plan 2013-2020.,” Ginebra, Suiza, 2013. Accessed: Oct. 10, 2019. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/97488/9789243506029_spa.pdf;jsessionid=0E501BE5A5A6368F961A1179340D38B3?sequence=1
- [8] WHO, “The WHO Special Initiative for Mental Health (2019-2023): Universal Health Coverage for Mental Health,” 2019. Accessed: Apr. 23, 2021. [Online]. Available: <http://www.who.int/iris/handle/10665/89966>
- [9] WHO, “COVID-19 disrupting mental health services in most countries,” 2020. Accessed: Apr. 23, 2021. [Online]. Available: <https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey>
- [10] IHME, “Global Burden of Disease Study 2019 (GBD 2019) Results,” Seattle, United States, 2020. [Online]. Available: <http://ghdx.healthdata.org/gbd-results-tool>
- [11] L. Á. Rojas-Bernal, G. A. Castaño-Pérez, and D. P. Restrepo-Bernal, “Salud mental en Colombia. Un análisis crítico,” *Medicina (B. Aires)*, vol. 32, no. 2, pp. 129–140, 2018, Accessed: Oct. 10, 2019. [Online]. Available: <http://www.scielo.org.co/pdf/cesm/v32n2/0120-8705-cesm-32-02-129.pdf>
- [12] C. Gómez Restrepo, “La Encuesta Nacional de Salud Mental–ENSM 2015,” *COLCIENCIAS*, pp.

- 45–60, 2015. doi: 10.1016/j.rcp.2016.09.006.
- [13] P. Balladelli, “El rol de la salud mental en el contexto americano y global: hacia un enfoque de derecho y equidad,” 2013.
- [14] DANE, “Encuesta Nacional de Consumo de Sustancias Psicoactivas,” 2020. Accessed: Apr. 23, 2021. [Online]. Available: <https://www.dane.gov.co/files/investigaciones/boletines/encspa/comunicado-encspa-2019.pdf>
- [15] WHO, “International Classification of Diseases (ICD-11),” 2018.
- [16] B. F. Grant, D. A. Dawson, F. S. Stinson, P. S. Chou, W. Kay, and R. Pickering, “The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample,” *Drug Alcohol Depend.*, vol. 71, no. 1, pp. 7–16, Jul. 2003, doi: 10.1016/S0376-8716(03)00070-X.
- [17] R. Ali *et al.*, “The alcohol, smoking and substance involvement screening test (ASSIST): development, reliability and feasibility,” *ADDICTION*, vol. 97, no. 9, pp. 1183–1194, Sep. 2002.
- [18] J. E. Schulenberg and J. L. Maggs, “A developmental perspective on alcohol use and heavy drinking during adolescence and the transition to young adulthood,” *J. Stud. Alcohol*, no. 14, pp. 54–70, Mar. 2002, doi: 10.15288/jsas.2002.s14.54.
- [19] H. W. Perkins, “Social norms and the prevention of alcohol misuse in collegiate contexts,” *J. Stud. Alcohol*, no. 14, pp. 164–172, Mar. 2002, doi: 10.15288/jsas.2002.s14.164.
- [20] M. Dennis *et al.*, “The Cannabis Youth Treatment (CYT) Study: Main findings from two randomized trials,” *J. Subst. Abuse Treat.*, vol. 27, no. 3, pp. 197–213, Oct. 2004, doi: 10.1016/j.jsat.2003.09.005.
- [21] R. W. Hingson, T. Heeren, R. C. Zakocs, A. Kopstein, and H. Wechsler, “Magnitude of alcohol-related mortality and morbidity among US college students ages 18-24,” *J. Stud. Alcohol*, vol. 63, no. 2, pp. 136–144, Mar. 2002, doi: 10.15288/jsa.2002.63.136.
- [22] J. R. Greenmyer, M. G. Klug, C. Kambeitz, S. Popova, and L. Burd, “A Multicountry Updated Assessment of the Economic Impact of Fetal Alcohol Spectrum Disorder: Costs for Children and Adults,” *J. Addict. Med.*, vol. 12, no. 6, pp. 466–473, 2018, doi: 10.1097/ADM.0000000000000438.
- [23] J. A. Durlak, R. P. Weissberg, A. B. Dymnicki, R. D. Taylor, and K. B. Schellinger, “The Impact of Enhancing Students’ Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions,” *CHILD Dev.*, vol. 82, no. 1, pp. 405–432, 2011, doi: 10.1111/j.1467-8624.2010.01564.x.
- [24] C. E. Hostinar, R. Nusslock, and G. E. Miller, “Future Directions in the Study of Early-Life Stress and Physical and Emotional Health: Implications of the Neuroimmune Network Hypothesis,” *J. Clin. CHILD Adolesc. Psychol.*, vol. 47, no. 1, pp. 142–156, 2018, doi: 10.1080/15374416.2016.1266647.
- [25] B. G. Gudmundsdottir, L. Weyandt, and G. B. Erudottir, “Prescription Stimulant Misuse and ADHD Symptomatology Among College Students in Iceland,” *J. Atten. Disord.*, vol. 24, no. 3, SI, pp. 384–401, Feb. 2020, doi: 10.1177/1087054716684379.

- [26] M. Isorna, L. Fernández-Ríos, and A. Souto, "Treatment of drug addiction and psychopathology: A field study," *Eur. J. Psychol. Appl. to Leg. Context*, vol. 2, no. 1, pp. 3–18, 2010.
- [27] C. Potier, V. Laprevote, F. Dubois-Arber, O. Cottencin, and B. Rolland, "Supervised injection services: What has been demonstrated? A systematic literature review," *Drug Alcohol Depend.*, vol. 145, pp. 48–68, 2014, doi: 10.1016/j.drugalcdep.2014.10.012.
- [28] L. Lu, Y. Fang, and X. Wang, "Drug abuse in China: Past, present and future," *Cell. Mol. Neurobiol.*, vol. 28, no. 4, pp. 479–490, Jun. 2008, doi: 10.1007/s10571-007-9225-2.
- [29] R. B. Felson and J. Staff, "Committing Economic Crime for Drug Money," *CRIME Delinq.*, vol. 63, no. 4, pp. 375–390, 2017, doi: 10.1177/0011128715591696.
- [30] S. Metternich, S. Zoerntlein, T. Schoenberger, and C. Huhn, "Ion mobility spectrometry as a fast screening tool for synthetic cannabinoids to uncover drug trafficking in jail via herbal mixtures, paper, food, and cosmetics," *DRUG Test. Anal.*, vol. 11, no. 6, pp. 833–846, Jun. 2019, doi: 10.1002/dta.2565.
- [31] D. S. Dolliver, S. P. Ericson, and K. L. Love, "A Geographic Analysis of Drug Trafficking Patterns on the TOR Network," *Geogr. Rev.*, vol. 108, no. 1, pp. 45–68, 2018, doi: 10.1111/gere.12241.
- [32] E.-U. Nelson and I. Obot, "Beyond prohibition: responses to illicit drugs in West Africa in an evolving policy context," *DRUGS AND ALCOHOL TODAY*, 2020, doi: 10.1108/DAT-07-2019-0033.
- [33] M. Ward, "Spatial Epidemiology: Where Have We Come in 150 Years?," 2008, pp. 257–282. doi: 10.1007/978-1-4020-8507-9_13.
- [34] A. B. Lawson *et al.*, "Disease mapping models: An empirical evaluation: Disease mapping collaborative group," in *Statistics in Medicine*, Sep. 2000, vol. 19, no. 17–18, pp. 2217–2241. doi: 10.1002/1097-0258(20000915/30)19:17/18<2217::AID-SIM565<3.0.CO;2-E.
- [35] Y. Mizuno, D. H. Higa, C. A. Leighton, M. Mullins, and N. Crepaz, "Is co-location of services with HIV care associated with improved HIV care outcomes? A systematic review," *AIDS Care*, vol. 31, no. 11, pp. 1323–1331, Nov. 2019, doi: 10.1080/09540121.2019.1576847.
- [36] D. F. Cuadros, A. Tomita, A. Vandormael, R. Slotow, J. K. Burns, and F. Tanser, "Spatial structure of depression in South Africa: A longitudinal panel survey of a nationally representative sample of households," *Sci. Rep.*, vol. 9, no. 1, p. 979, Dec. 2019, doi: 10.1038/s41598-018-37791-1.
- [37] P. Barros *et al.*, "Social consequences and mental health outcomes of living in high-rise residential buildings and the influence of planning, urban design and architectural decisions: A systematic review," *Cities*, vol. 93, pp. 263–272, Oct. 2019, doi: 10.1016/j.cities.2019.05.015.
- [38] B. Kim, D. Callander, R. DiClemente, C. Trinh-Shevrin, L. E. Thorpe, and D. T. Duncan, "Location of Pre-exposure Prophylaxis Services Across New York City Neighborhoods: Do Neighborhood Socio-demographic Characteristics and HIV Incidence Matter?," Jul. 2019. doi: 10.1007/s10461-019-02609-2.
- [39] R. Dannefer *et al.*, "The Neighborhood as a Unit of Change for Health: Early Findings from the East Harlem Neighborhood Health Action Center.," *J. Community Health*, Aug. 2019, doi:

- 10.1007/s10900-019-00712-y.
- [40] K. Hwang, T. B. Asif, and T. Lee, “Choice-driven location-allocation model for healthcare facility location problem,” *Flex. Serv. Manuf. J.*, vol. 34, no. 4, pp. 1040–1065, Dec. 2022, doi: 10.1007/S10696-021-09441-8/FIGURES/7.
- [41] T. de M. Sathler, J. F. Almeida, S. V. Conceição, L. R. Pinto, and F. C. de Campos, “Integration of Facility Location and Equipment Allocation in Health Care Management,” *Brazilian J. Oper. Prod. Manag.*, vol. 16, no. 3, pp. 513–527, Aug. 2019, doi: 10.14488/BJOPM.2019.V16.N3.A13.
- [42] R. Rezaee, F. Rahimi, and A. Goli, “Urban Growth and urban need to fair distribution of healthcare service: a case study on Shiraz Metropolitan area,” *BMC Res. Notes*, vol. 14, no. 1, pp. 1–6, Dec. 2021, doi: 10.1186/S13104-021-05490-2/FIGURES/2.
- [43] N. Vidarthi and O. Kuzgunkaya, “The impact of directed choice on the design of preventive healthcare facility network under congestion,” *Health Care Manag. Sci.*, vol. 18, no. 4, pp. 459–474, Dec. 2015, doi: 10.1007/S10729-014-9274-2.
- [44] P. Mitropoulos, I. Mitropoulos, I. Giannikos, and A. Sissouras, “A biobjective model for the locational planning of hospitals and health centers,” *Health Care Manag. Sci.*, vol. 9, no. 2, pp. 171–179, May 2006, doi: 10.1007/S10729-006-7664-9/METRICS.
- [45] WHO, “Mental health: strengthening our response,” Mar. 2018. <https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response> (accessed Apr. 22, 2021).
- [46] IIE, “Industrial and Systems Engineering body of knowledge,” 2021. <https://www.iise.org/Details.aspx?id=43631> (accessed Apr. 22, 2021).
- [47] K. Ochoa, “Aportes de la ingeniería a la salud y la calidad de vida: una revisión,” *Rev. Tecnol.*, vol. 12, no. 3, pp. 88–98, 2013, doi: 10.18270/rt.v12i3.1832.
- [48] RAE, “Engineering for health,” 2012. Accessed: Apr. 22, 2021. [Online]. Available: <https://www.raeng.org.uk/publications/reports/engineering-for-health#:~:text=Engineers are working with biologists,patient’s ability to self-heal.>
- [49] M. Pavel *et al.*, “The role of technology and engineering models in transforming healthcare,” *IEEE Rev. Biomed. Eng.*, vol. 6, pp. 156–177, 2013, doi: 10.1109/RBME.2012.2222636.
- [50] M. C. Chyu *et al.*, “Healthcare engineering defined: A white paper,” *J. Healthc. Eng.*, vol. 6, no. 4, pp. 635–648, Dec. 2015, doi: 10.1260/2040-2295.6.4.635.
- [51] P. Elliott, J. Wakefield, N. Best, and D. Briggs, “Clustering, cluster detection, and spatial variation in risk,” in *Spatial Epidemiology Methods and Applications*, New York: Oxford University Press, 2001.
- [52] A. Sarker *et al.*, “Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from twitter,” *Drug Saf.*, vol. 39, no. 3, pp. 231–240, Jan. 2016, doi: 10.1007/s40264-015-0379-4.
- [53] T. Nasrallah, O. El-Gayar, and Y. Wang, “Social media text mining framework for drug abuse: Development and validation study with an opioid crisis case analysis,” *J. Med. Internet Res.*, vol. 22, no. 8, p. e18350, Aug. 2020, doi: 10.2196/18350.

- [54] MINSALUD, “Actualización de la guía practica de atención integral en farmacodependencia,” 2004.
- [55] MINSALUD, “ABECÉ de la prevención y atención al consumo de sustancias psicoactivas.” pp. 1–6, 2016. [Online]. Available: <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/Abece-salud-mental-psicoactivas-octubre-2016-minsalud.pdf>
- [56] M. De pina, S. Ferreira, A. Correia, and A. Castro, “Epidemiología espacial: nuevos enfoques para viejas preguntas,” *universitasodontologica*, vol. 29, no. 63, pp. 47–65, 2010, Accessed: Oct. 14, 2019. [Online]. Available: <https://dialnet.unirioja.es/descarga/articulo/3986944.pdf>
- [57] T. Bailey and A. Gatrell, *Interactive spatial data analysis*. 1995. Accessed: Oct. 23, 2019. [Online]. Available: <http://www.personal.psu.edu/faculty/f/k/fkw/rsoc597/Introduction.pdf>
- [58] T. Hardy, “IA: Inteligencia Artificial,” *Rev. la Univ. Boliv.*, vol. 1, no. 2, pp. 1–24, 2001.
- [59] M. J. Hofmann, C. Biemann, and S. Remus, “Benchmarking n-grams, Topic Models and Recurrent Neural Networks by Cloze Completions, EEGs and Eye Movements,” in *Cognitive Approach to Natural Language Processing*, Elsevier Inc., 2017, pp. 197–215. doi: 10.1016/B978-1-78548-253-3.50010-X.
- [60] V. N. Gudivada, “Natural Language Core Tasks and Applications,” in *Handbook of Statistics*, vol. 38, Elsevier B.V., 2018, pp. 403–428. doi: 10.1016/bs.host.2018.07.010.
- [61] G. Alonso, E. del Valle, and J. R. Ramirez, “Optimization methods,” *Desalin. Nucl. Power Plants*, pp. 67–76, Jan. 2020, doi: 10.1016/B978-0-12-820021-6.00005-3.

Chapter 2

2. Statistical Analysis and Machine Learning in Psychoactive Substance Use: A Bibliometric Analysis

2.1. Abstract

Because psychoactive substance use is a topic that has received worldwide attention, this area has added several scientific outcomes. It is essential to conduct a comprehensive analysis comprising as many studies as are available to summarize the separate studies and provide an overall view of how the research field has been evolving over the last few decades. This study performs a bibliometric analysis using a large dataset of published papers from 2000 to 2021. The study examined 1022 publications from those 20 years. About 79% used statistical analyses, and machine learning techniques were utilized by almost 21%. It is worth mentioning that the publications related to statistical analysis were grouped in the following way: multivariate or univariate statistical analysis (52.4%), Bayesian analysis (21.7%), and spatial analysis (50.5%). There were several key points regarding the overall results of the research. Results illustrated that publications had grown significantly during the last two decades. The majority of the publications come from the United States. In addition, the most prolific authors and journals were identified. Over the last decade, due to advanced technological methods, more research has been focused on enhancing and designing Bayesian techniques for using psychoactive substances.

2.2. Introduction

Mental health refers mainly to the risks associated with mental problems and disorders, including epilepsy, gender-based violence, domestic abuse, sexual abuse, psychoactive substance use (PSU), schizophrenia, mental retardation, Attention Deficit and Hyperactivity Disorder, along with others. Mental health issues and illnesses are

increasingly frequent in the world's population. They are becoming one of the highest disease burdens [1]–[3]. In addition, they constitute a substantial societal cost in terms of losses in productivity, early mortality, rising healthcare spending, criminal justice, well-being costs, and other societal consequences [4]–[6]. The plan entitled ‘Comprehensive Mental Health Action 2013-2030’ by the World Health Organization (WHO) estimates that the accumulated worldwide impact in terms of economic losses from mental health disorders from 2011 to 2030 will be \$16.3 trillion [7]. Regarding psychoactive substance use (PSU), this term in the literature is equivalent to: ‘drug abuse’, ‘drug use’, ‘psychotropic substances use’, ‘psychotropic substances abuse’, or ‘psychoactive substances abuse’. The WHO has defined Psychoactive Substances Use to mean the consumption of substances that affect mental processes in such a way that if they are not controlled, they can trigger personal, family, social, and educational problems. In brief, PSU is considered a mental disorder caused by the abuse of substances of psychotropic origin. Some commonly abused substances are alcohol, opioids, cannabinoids, sedatives, and hypnotics. Stimulants are also frequently abused [8].

The development of research on psychoactive substance use has been oriented towards different approaches. For example, some studies focus on psychoactive substance abuse in teenagers [9]–[15], psychological treatments based on social and emotional learning (SEL) or pharmacological treatments [16]–[18], and social and geographical implications of drug trafficking [19]–[24]. As the PSU has received worldwide interest, this field has produced many scientific results, including previous publications. Therefore, it is crucial to have a comprehensive analysis covering the most significant number of investigations available to consolidate individual studies and show what has been happening in the field over recent decades. Bibliometrics, understood simply as the use of statistical techniques on books and other forms of media, offers a comprehensive information-based analysis of publications with an objective and applied approach [25], [26]. In addition, a bibliometric analysis can give insights into how research is advancing through different procedures, such as performance analysis, which performs a quantitative analysis using citation data [27], [28].

Significant studies conduct a conceptualization, literature review, or bibliometric analysis of psychoactive substance use [29]–[32]. One of the first articles referring to bibliometric analysis in the use of psychoactive substances is the work of Herrán et al. (1996), who

developed a descriptive study based on the review of the journal *Atención Primaria* on mental health between 1984 and 1995 [33]. Later, some authors studied scientific productivity using bibliometric analysis. For instance, López et al. (2008) conducted a bibliometric analysis of ADHD-related scientific publications from 1980-2005 [34]; Bramness et al. (2014) compared citation rates within substance abuse research in Europe with those in the United States from 2001 to 2011 [35]; and Sweileh et al. (2014) assessed the productivity of research in the field of substance abuse in Arabic countries employing bibliometric measures [36]. Additionally, researchers are using analytical software packages more often in qualitative studies. In recent work, for instance, Zyoud et al. (2017) analyzed publications regarding cocaine intoxication research trends during 1975-2015. They used VOSviewer to show high-frequency terms associated with cocaine toxicity [37]. Thus, despite recent advances in research on psychoactive substance use and the valuable contributions of several authors, there still needs to be more literature regarding statistical and machine learning methods. Also, some of these previous publications need to be updated.

Considering the abundant bibliometric approaches, this study aims to present a visual overview of the PSU and carry out a comprehensive study of the state of research and advances in this field. In particular, this study presents a bibliometric analysis of scientific publications associated with statistical analysis and machine learning over the last 20 years (2000-2021). The analysis includes publications per year, publications per country, the number of citations per year, most cited publication ranked, most published journal ranked, and most published authors ranked by the total number of publications, among others. A bibliometric analysis of this kind can summarize academic proposals for healthcare policymakers and decision-makers. Likewise, this analysis contributes to the scientific community by establishing a starting point for new collaborations since researchers can learn about the current state of psychoactive substance use research proposals. Regarding its importance for the mental health field, a bibliometric analysis gives an adequate understanding of the different approaches developed and their research directions, identifying different characteristics or aspects of scientific productivity and providing information on the current state of the research agenda as their main gaps.

This chapter is structured into four main sections, which include this introduction. The second presents methods focusing on how the data have been prepared for analysis.

Subsequently, the third part shows the results of this review. Then, the fourth part discusses these results, including opportunities for future work.

2.3. Methodology

This investigation aimed to conduct a quantitative and visualized analysis of the representative studies related to Statistical Analysis (Bayesian analysis, spatial analysis, or multivariate or univariate statistical analysis) and Machine Learning in PSU through a bibliometric approach. Clarivate Analytics' Web of Science is an electronic indexing service for scientific references that offer comprehensive citation searching. It also provides multiple database access and contains nearly 1.9 billion references cited from more than 171 million records; WoS was therefore selected to obtain the data. The keywords used for the data collection included combinations of “psychoactive substances” and their variant, such as ‘drug abuse’, ‘psychotropic drug’, ‘psychoactive drug’ or ‘psychotropic substances’. Multiple wildcards and searching operators were applied to enhance the precision of the retrieval outputs. Search queries were repeatedly tested and adjusted to find relevant papers. Having found a suitable publication, other search terms were also identified by mapping topic headings and reviewing keywords based on the following queries:

- i. *(("Drug Abuse") OR ("Psychoactive Substances") OR ("Psychotropic Drug") OR ("Psychoactive Drug") OR ("Psychotropic Substances") OR (Psychotropic*) OR (Psychoactive*) OR ("Drug Use") OR ("Substance Abuse")) AND (("Machine Learning") OR ("Deep Learning"))).*
- ii. *(("Drug Abuse") OR ("Psychoactive Substances") OR ("Psychotropic Drug") OR ("Psychoactive Drug") OR ("Psychotropic Substances") OR (Psychotropic*) OR (Psychoactive*) OR ("Drug Use") OR ("Substance Abuse")) AND (("Spatial Analysis") OR ("Spatial Statistic") OR ("Spatial Statistical") OR (Geostatistical*) OR (Geostatistic*) OR (Geoanalysis*)).*
- iii. *(("Drug Abuse") OR ("Psychoactive Substances") OR ("Psychotropic Drug") OR ("Psychoactive Drug") OR ("Psychotropic Substances") OR (Psychotropic*) OR (Psychoactive*) OR ("Drug Use") OR ("Substance Abuse")) AND (("Bayesian Analysis") OR ("Bayesian Statistic") OR ("Bayesian Statistical") OR (Bayesian*) NOT (("Spatial Analysis") OR ("Spatial Statistic") OR ("Spatial Statistical") OR (Geostatistical*) OR (Geostatistic*) OR (Geoanalysis*))).*
- iv. *(("Drug Abuse") OR ("Psychoactive Substances") OR ("Psychotropic Drug") OR ("Psychoactive Drug") OR ("Psychotropic Substances") OR (Psychotropic*) OR (Psychoactive*) OR ("Drug Use") OR ("Substance Abuse")) AND (("Statistical Analysis") OR ("Statistic Analysis") NOT (("Bayesian Analysis") OR ("Bayesian*

Statistic") OR ("Bayesian Statistical") OR (Bayesian) OR ("Spatial Analysis") OR ("Spatial Statistic") OR ("Spatial Statistical") OR (geostatistical*) OR (geostatistic*) OR (Geoanalysis*)*)).

Our search was strictly limited to peer-reviewed articles from journals published in English from 2000 to November 2021. References were excluded based on title if they were not in the relevant subject area: Published articles in which statistical analysis (Bayesian analysis, spatial analysis, or multivariate or univariate statistical analysis) or machine learning was used in Psychoactive Substances. For each publication, both authors completed the data collection. The discussion continued if the authors disagreed until a consensus was reached. To encompass as much information as possible, we exported complete records for each retrieval item in ".csv" file format in WoS. RStudio software was used to verify and delete duplicate and non-related items by hand. Different data types were extracted, including total publications per year, total publications per country, number of citations per year for each publication, the most cited publication ranked, most published journal ranked, and most published authors ranked, among others.

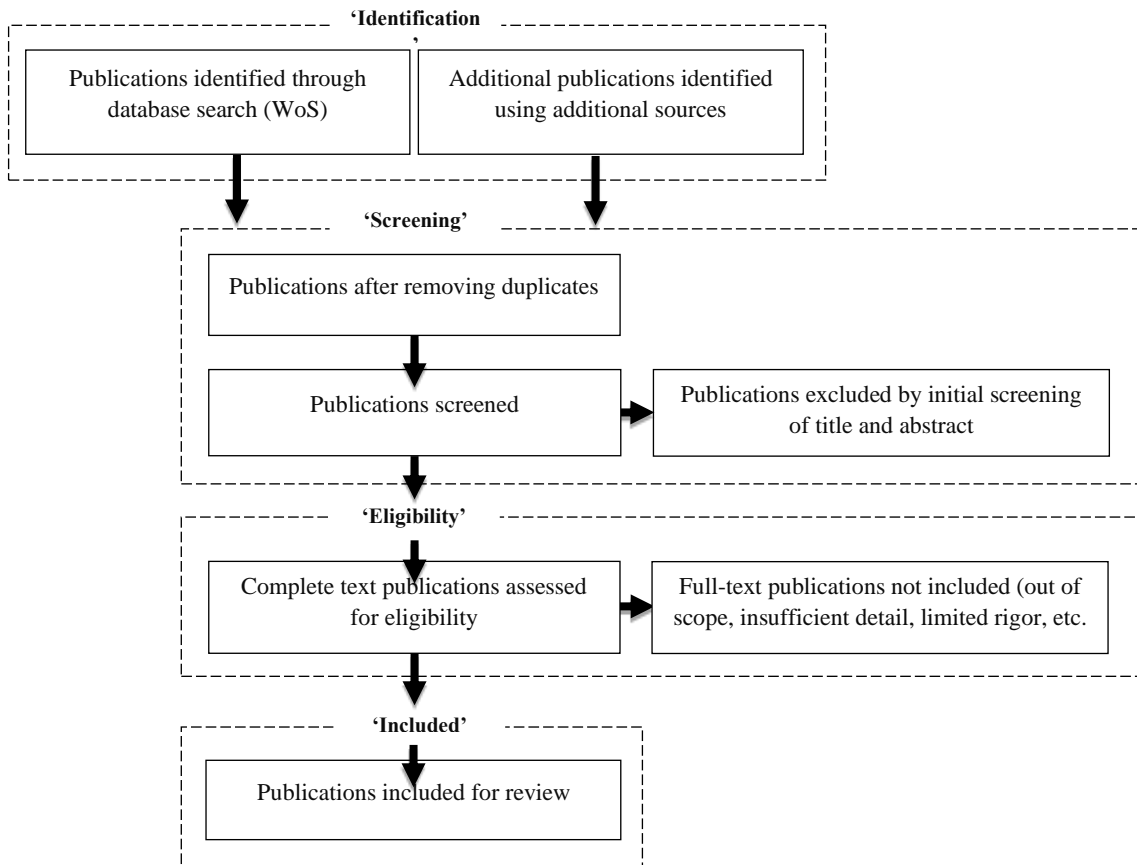


Figure 4. Bibliometric analysis flow chart.

A detailed description of the publication selection process can be found in the flow chart in Figure 4. Using the filtering and search scheme, 1,022 publications were identified for review. Technically, numerous tools and computer software packages are available to perform the bibliometric process. Among these tools, RStudio is an open-source software for data science developed by Joseph J. Allaire. RStudio was robust enough to conduct approximately all the bibliometric analyses. In addition, VOSviewer executed visualized and interactive features for easy understanding of patterns, including composing automated labeling of clusters based on cited publication terms. Further criteria adopted were: Where available, additional references were searched through a second search of the reference lists in key publications. Experts in this field were also asked to confirm that all key articles had been included. This aspect was especially critical considering the search terms involved in this subject field: these were commonly used words unrelated to the subject matter, which resulted in a high volume of irrelevant articles. All articles of the relevant subject area were then reviewed using a data extraction sheet, and key findings were reviewed. Finally, papers and book chapters that were distinctly publication reviews were not included; those that were relevant are discussed in the introductory section.

2.4. Results

Table 1 indicates the number of academic publications in English between 2000 and November 2021 using four analytical techniques. 1,022 publications were examined in those twenty years (all references are provided as supplementary material). About 79% (n = 808) used statistical analyses. Machine learning techniques were utilized by almost 21% (n = 214). It is worth mentioning that the publications related to Statistical Analysis were grouped as follows: Multivariate or Univariate Statistical Analysis - MUSA (n = 536, 52.4%), Bayesian Analysis (n = 222, 21.7%), and Spatial Analysis (n = 50,5%).

Table 1. Total number of publications

Field	Number of publications (N = 1,022)
1. Statistical Analysis	808 (79.1%)
1.1. Bayesian Analysis	222
1.2. Spatial Analysis	50
1.3. Multivariate or Univariate Statistical Analysis	536
2. Machine Learning	214 (20.9%)

In this way, in the last twenty years, publications related to Bayesian analysis on psychoactive substances have grown steadily, peaking in 2019 with 25 papers. The annual number of publications on psychoactive substances using Spatial Analysis averages 7.17 publications per year, with a peak of 7 in 2019. The number of publications using multivariate or univariate statistical analysis (MUSA) has also increased. The annual number of publications on psychoactive substances using MUSA was 19.34, with a slight increase from 2008 onwards and a peak of 63 publications in 2020. Publications using Machine Learning show an uneven skew to the distribution. From 2001 to 2014, the number of publications was relatively flat; from 2016-2021, the number had a higher output. Thus, publications present an average annual growth rate of 33.3% and have risen considerably since 2014, with a peak of 54 in 2019. In general, although the use of statistical analysis increased over the years, machine learning was the technique that showed the highest annual growth rate. The increase in the total of publications reflects an expansion of global scientific inquiry into this field. Figure 1 presents the number of publications per year.

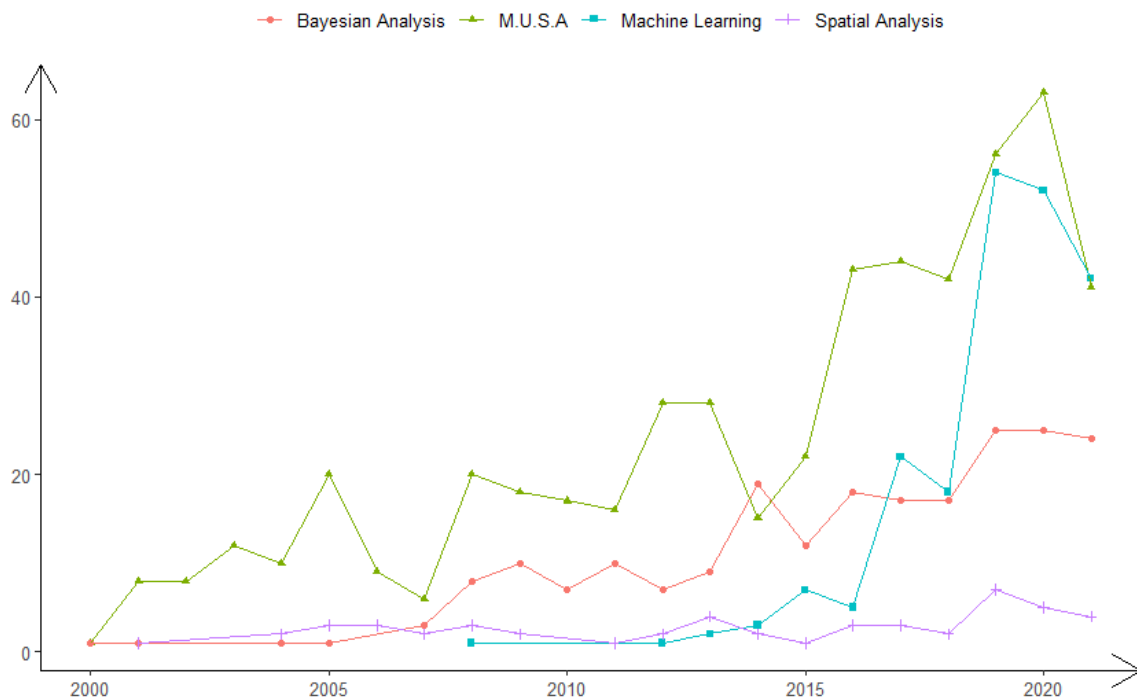


Figure 5. Publications per year.

Bibliometrix R-package was used to calculate and plot the performance of authors (in terms of the total publications) in descending order. Figure 6 below outlines each approach's top ten contributing authors and the number of publications.

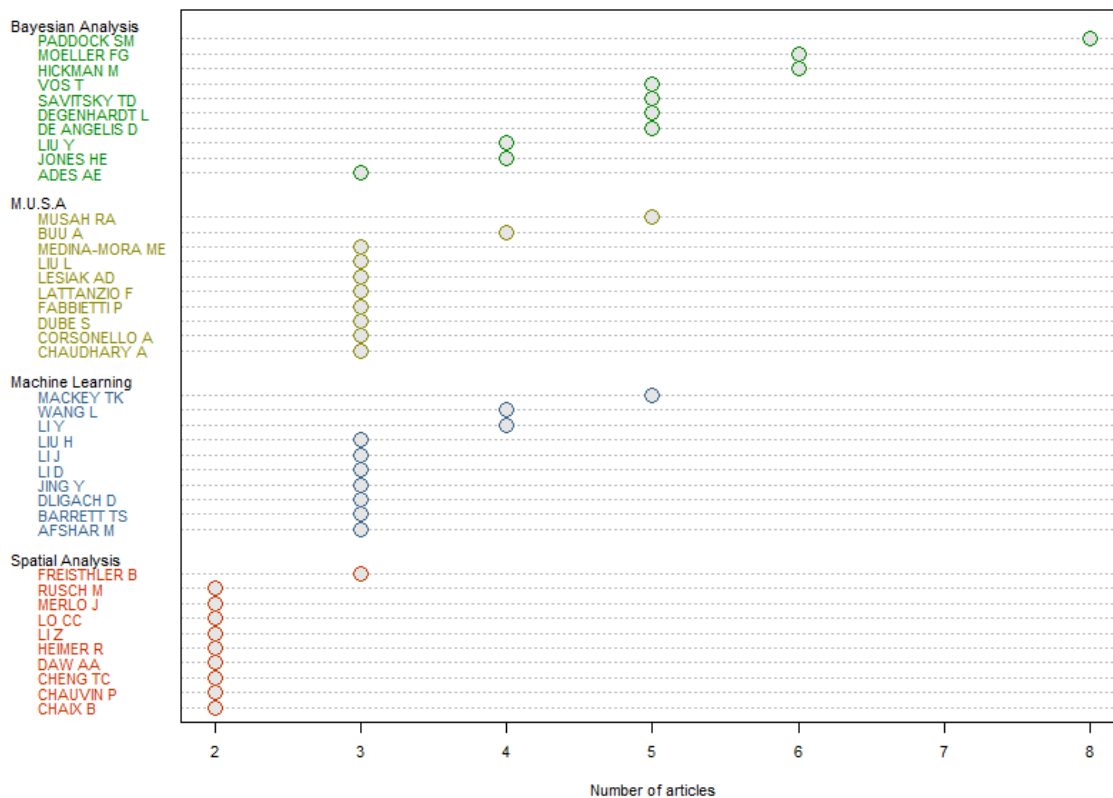


Figure 6. Performance of authors.

The authors with the highest publication numbers were Paddock, Moeller, and Hickman (8, 6, and 6 articles published, respectively) using Bayesian techniques on psychoactive substances research; their research background is related to substance abuse treatment in therapy groups and neuroimaging and genetic studies concerning drug dependence. As for Spatial analysis, Freisthler was the author with more publications (3 scientific articles in total); their research is related to drug market operations and child abuse geography; for this approach, there were fewer than three publications by other authors. It appears that Musah and Buu dominated the list of contributing authors with the most significant number of publications using multivariate or univariate statistical analysis in psychoactive substance use with 5 and 4 publications, respectively; their research interests are in mass spectrometry for the detection and identification of psychoactive plants, and genome-wide association studies for drug dependence. Regarding the authors leading the list in machine learning, Mackey and Wang have published the most, having a total of 5 and 4 publications, respectively; their research interests are associated with predicting the severity of substance use severity from childhood through adulthood. Comparing the results, it should be noted that the size of the active window necessary to be considered the primary active author is different in each case.

Analyzing journal performance results can help determine which journals have contributed the most scientific papers, generated significant interest, and which journal a manuscript should be submitted to. Figure 7 shows the 10 most fruitful journals in decreasing order of the number of publications. ‘Plos One’ ranks first with 11 publications, followed by ‘Addiction’ with 10 publications using Bayesian analysis approaches. Regarding spatial analysis, the most productive journals are the ‘International Journal of Drug Policy’ with 6 articles, and the ‘International Journal of Environmental Research and Public Health’ with only 3 in total. As to the multivariate or univariate statistical analysis, the journal ‘Revista de Saúde Pública’ ranked first with 11 registered publications, followed by the ‘Journal of Evolution of Medical and Dental Sciences’ with 8 published papers. In Machine Learning, once again ‘Plos One’ is the leader of the ranking with 12 publications, with the ‘Journal of Medical Internet Research’ following with 9 publications. From a broad point of view, ‘Plos One’ is the most productive journal, reaching a total of 24 publications in all the approaches included in this review, followed by the ‘Drug and Alcohol Dependence’, and the journal ‘Addiction’, with 19 and 12 publications, respectively. It is relevant to highlight that the journal ‘Plos One’ has the highest H-index (300), followed by Revista de Saúde Pública’ with 182.

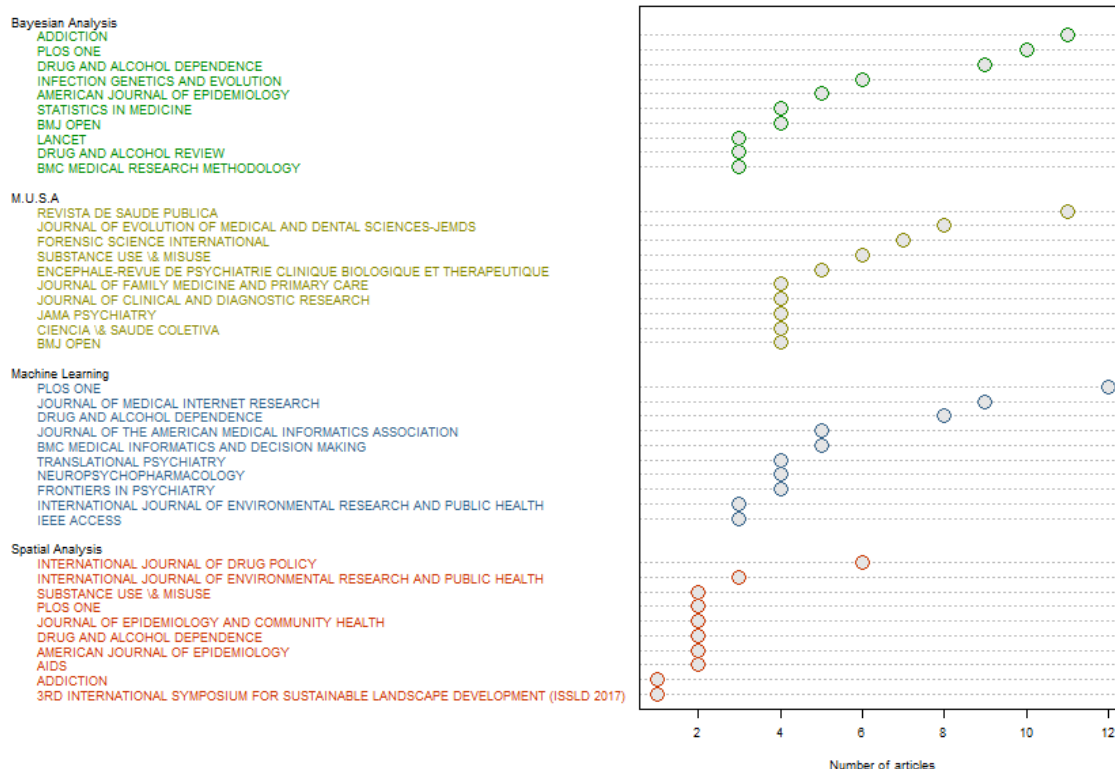


Figure 7. Journal performance.

Figure 8 is the network analysis graph focusing on the keyword combination. The map circles indicate the keywords used in the publications extracted. The wider the circle, the higher the level of keyword frequency. Fig.8(A) shows some clusters of keyword occurrences associated with Bayesian analysis. Overall, Bayesian techniques applied in studies of psychoactive substance use aim to find epidemiological patterns through models to analyze prevalence, comorbidity, and mortality related to drugs in the population; Researchers seem to use longitudinal data to try to understand within-sample changes over time and how those changes lead to outcomes detected in later waves of the data. Furthermore, in Fig. 8(B), the network map for trends based on keyword analysis in multivariate or univariate statistical analysis is presented. As can be seen, these techniques are most associated with drug use and abuse, addiction, and prevalence in adolescents. Alcohol consumption, the risk of developing schizophrenia, and the use of predictive models are highlighted. Similarly, Fig. 8(C) presents some clusters of keywords associated with spatial analysis. The research emphasizes the risks and patterns of PSU and the presence of crime, violence, and child abuse in urban or suburban environments. Also critical is the use of Geographic Information Systems and the analysis of epidemiological prevalence factors. As shown in Fig. 8(D), techniques such as Social Media Monitoring, Natural Language Processing, Random Forest, and Deep learning are highlighted in the field of machine learning. The prediction and prevalence associated with risk factors and mental health problems, including depression and schizophrenia, are also essential. The United States seems to be where these studies are most widely applied.

Another important aspect was the mapping of the most productive countries. Figure 9 presents the total number of publications and the most contributing countries in the past 20 years. Spatial calculations were executed using the RStudio packages "tmap", "rgeos", and "rgdal". Fig. 9(A) shows the most productive countries publishing the most significant articles associated with Bayesian analysis on psychoactive substance use from 2000 to 2021. As can be seen, the United States generates the greatest number of publications (73; 42%), followed by the United Kingdom and Australia, with 17 and 14 articles published, respectively. In the same way, the most productive country in the publishing of articles associated with multivariate or univariate analysis of psychoactive substance use is the United States, with 95 publications, which represents 22% of the total number of articles published, followed by Brazil (39; 8%) and India (24; 5%), as shown in Fig. 9(B). In Fig. 9(C), the United States also leads the ranking of countries with the

most publications on psychoactive substance research using spatial analysis, with 22 publications representing 54% of the total number of articles published in this field. The list is preceded by Canada (4; 9%) and China (3; 7%). Last but not least, as seen from Fig. 9(D), the United States is again the country that contributes with the most significant number of scientific publications (86;66%), followed by China (73; 42%) and India (5;4%) on psychoactive substances research using machine learning.

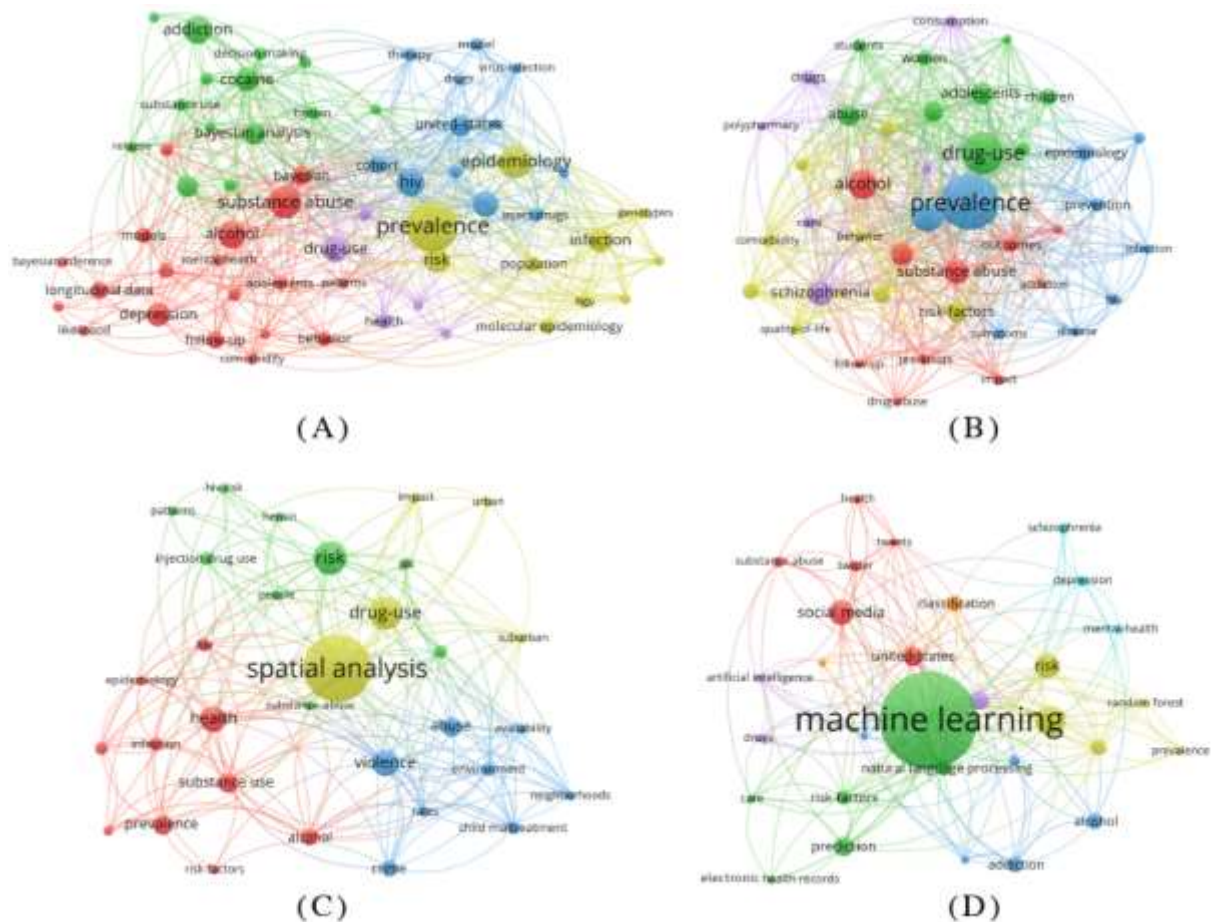


Figure 8. Network visualization map of author keywords.

In addition, the 10 references with the highest bursts of citations from 2000 to 2021 were also exported, as shown in Table 2. The academic performance of a particular area of research can be described by the publications that received the most outstanding increment of references, i.e., the bursts of citations. Citation burst suggests the probability that the academic community has paid or is paying particular attention to the underlying contribution. In the case of Machine learning, Wager et al. (2013) led the ranking with their work entitled "An fMRI-Based Neurologic Signature of Physical Pain", which registered 534 citations with 66.7 citations on average per year. It is important to mention

that the most cited techniques in this group of top publications are ‘support vector machine’, ‘random forest’, ‘natural language processing’, and ‘social media monitoring’. Regarding Bayesian analysis, the most prominent publication is "Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study" by Whiteford et al (2010). This work records a total of 2337 citations, averaging 292 per year. From a general point of view, Bayesian meta-regression and Bayesian hierarchical models are the most attractive approaches.

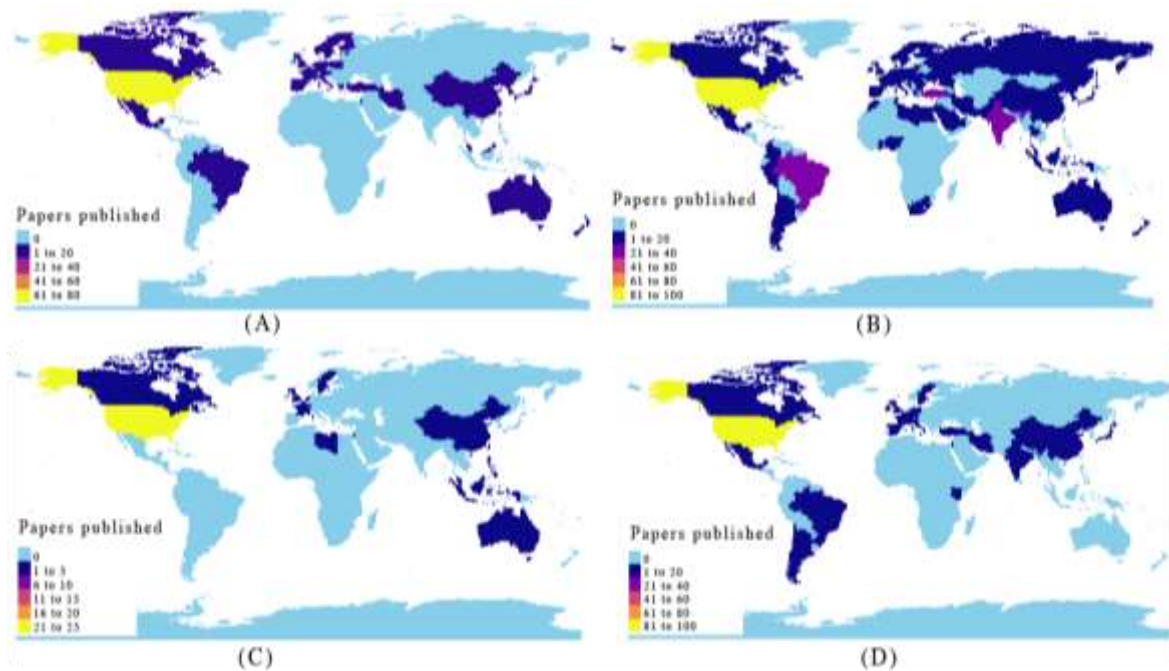


Figure 9. Performance of countries.

For Spatial Analysis, the work of Shannon et al. (2008) entitled "Mapping violence and policing as an environmental-structural barrier to health service and syringe availability among substance-using women in street-level sex work" is the most cited (162) with a mean of 12.5 citations per year. It should be noted that spatial distributions, spatial regression models, spatial scan statistics, variograms, and social mapping are the most widely used techniques to address the problem of psychoactive substance use. Finally, as far as Multivariate or Univariate Statistical Analysis is concerned, the most striking article is "Structural abnormalities in the brains of human subjects who use methamphetamine", which was carried out by Thompson et al. (2004). It is worth mentioning that techniques such as independent group T-Test, descriptive analysis correlations, logistic regression, and confirmatory factor analysis are associated with the most cited papers. In this way, academic interest in these highly cited articles led to forming the main representative

approaches. Knowledge of these papers with solid bursts of citations can provide a quick insight into developing new research interests and the continuing evolution of treatment for psychoactive substance use.

Table 2. Top 10 most cited publications.

	Ranking	Approach	Times cited	Times cited per year
Machine learning	1	[38] (Wager et al. 2013) LASSO-PCR (Least Absolute Shrinkage and Selection Operator-Regularized Principal Components Regression)	534	66.75
	2	[39] (Lee 2013) Model-Based Reinforcement Learning	63	7.88
	3	[40] (Henriksson et al. 2015) Natural Language Processing	41	6.83
	4	[41] (Conway et al. 2016) Social Media, Big Data and Natural Language Processing	38	7.60
	5	[42] (Squeglia et al. 2017) Random Forest	37	9.25
	6	[43] (Arnold et al. 2012) Ensemble Machine Learning with Random Forests	35	3.89
	7	[44] (Bedi et al. 2014) Natural Language Toolkit and Support Vector Machines	31	4.43
	8	[45] (Katsuki et al. 2015) Support Vector Machines	27	4.50
	9	[46] (Struyf et al. 2008) Support Vector, Nearest Shrunken Centroids, Decision Trees, Nave Bayes, and Nearest Neighbor	25	1.92
	10	[47] (Alvaro et al. 2015) NPL, SVM, GLM, C50, Naive Bayes, Bayesian Generalized Linear Model, and Multi-Layer Perceptron	24	4.00
Bayesian Analysis	1	[48] (Whiteford et al. 2013) Disease Modeling – Meta-Regression, Bayesian Meta-Regression	2337	292.12
	2	[49] (Woolcott et al. 2009) Updated Bayesian Odds Ratios	591	49.25
	3	[50] (Degenhardt et al. 2013) Bayesian Meta-Regression Technique (DisMod-MR)	392	49.00
	4	[51] (Collins et al. 2015) Bayesian Model Averaging	140	23.33
	5	[52] (Sung et al. 2004) Hierarchical Bayesian Framework, Gibbs Sampling Approach	118	6.94
	6	[53] (Schwan et al. 2010) Bayesian Data-Mining Algorithm	88	8.00
	7	[54] (Rivkees et al. 2010) Gamma-Poisson Shrinker (MGPS) Data Mining Algorithm, Bayes Geometric Mean	84	7.64
	8	[55] (Degenhardt et al. 2014) Bayesian Meta-Regression Technique (DisMod-MR)	74	10.57
	9	[56] (Bowman et al. 2008) Bayesian Extension of Voxel-Level Analyses, Markov Chain Monte Carlo, Bayesian Hierarchical Model	70	5.38
	10	[57] (O'Brien et al. 2014) Bayesian Inference Models	64	9.14
Spatial Analysis	1	[58] (Shannon et al. 2008) Social Mapping	162	12.46
	2	[59] (Chaix et al. 2005) Spatial Analytical Perspective, Geo-additive Models, Multilevel Approach	123	7.69
	3	[60] (Freisthler 2004) Spatial Regression Models	123	7.24
	4	[61] (Prosser et al. 2006) Visual-Spatial Analysis	99	6.60
	5	[62] (Banta-Green et al. 2009) Spatial Epidemiology, Spatial Distribution	88	7.33
	6	[63] (Freisthler et al. 2005) Spatial Regression Techniques	87	5.44
	7	[64] (Dovey et al. 2001) Socio-Spatial Analysis	84	4.20
	8	[65] (Chaix et al. 2006) Spatial Distributions, Spatial Scan Statistic	52	3.47
	9	[66] (Heimer et al. 2008) Spatial Distribution	36	2.77
	10	[67] (Bass et al. 2004) Variograms (Spatial Dependence)	24	1.41
MUSA	1	[68] (Thompson et al. 2004) Student's T Test, Descriptive Analysis, Confidence Intervals	440	25.88
	2	[69] (Mowbray et al. 2003) Internal Consistency Reliability, CFA, Cluster Analysis, Pearson Correlations, Coefficient Kappa, Confidence Interval, Cochran's Q Test, P Statistic	414	23.00
	3	[70] (Fazel et al. 2008) Logistic Regression Analysis, Descriptive Analysis	308	23.69
	4	[71] (Shoji et al. 2004) Proportional Odds Model	269	15.82
	5	[72] (Maas et al. 2006) Correlation	165	11.00
	6	[73] (Peet et al. 2005) Descriptive Analysis	163	10.19
	7	[74] (Gilbert et al. 2006) Fisher's Exact Test, Chi-Quadrat Test, Descriptive Analysis	149	9.93
	8	[75] (Neubert et al. 2004) Descriptive Analysis	106	6.24
	9	[76] (Wazaify et al. 2005) Descriptive Analysis	98	6.12
	10	[77] (Merson et al. 2000) Independent Group T-Test and The Paired T-Test.	92	4.38

2.5. Discussion and conclusion

Based on the data set, including 1,022 publications from Web of Science from 2000 to 2021, this paper conducted a complete bibliometric analysis of statistical analysis and machine learning in psychoactive substance use. Several important facts about research performance were observed. First, results illustrated that annual articles grew considerably in the past two decades. The United States notably contributed to the most significant number of publications, followed by China, Brazil, and India. Furthermore, the most prolific authors and journals were identified. In the last decade, due to advanced technological methods, most research has steered toward attempts to enrich and design Bayesian techniques on psychoactive substance use. Few spatial analysis papers appear in journals, likely due to researchers' perceptions concerning spatial data access. Additionally, it can be concluded that machine learning and multivariate or univariate statistical analysis will remain a critical part of understanding psychoactive substances. Thus, the psychoactive substance use study analysis is meaningful to researchers; for example, deciding which researcher to follow or co-author with, determining which journal to focus on or submit manuscripts to, and selecting which country to enrich collaboration or exchanges with. Also, the keyword network analysis showed can help researchers understand the landscape of psychoactive substance use and set up future research directions. In brief, the bibliometric analysis provides solid evidence that PSU is becoming gradually accepted, validated, and embraced in broader geographical regions, research fields, and periods.

The majority of publications related to Bayesian analysis were related to epidemiological models, whereas publications related to multivariate or univariate statistical analysis were associated with predictive models and techniques involving drug use and abuse, addiction, and prevalence, mainly in adolescents. Spatial analysis research was most commonly conducted on psychoactive substance use, risks and patterns, and the presence of crime, violence, and child abuse in urban or suburban environments. Machine learning techniques such as social media monitoring, natural language processing, random forest, and deep learning were commonly used in these studies. The study acknowledged that there were limitations to its approach. One of the main limitations was that it only focused on papers published in the WoS database, which may not be fully representative of all relevant publications. This means that there may be other important studies on the topic

of psychoactive substance use that were not included in the analysis. Likewise, the study also noted that there are other methods of reviewing publications, such as systematic review and meta-analysis, that may yield different outcomes. Comparing the results of different methods could provide more comprehensive guidance on the topic. The study emphasized the need for further research to clarify the design and approach of the publications reviewed, as well as their outcomes and potential for follow-up research. Additionally, the sparse number of publications in this field makes it difficult to generalize results from one study to another, given the heterogeneity of treatment duration, exclusion factors, and control groups used. Therefore, future efforts should focus on more homogeneous groups and address results in greater depth to draw more specific conclusions. Overall, while the study provided valuable insights into the publications on psychoactive substance use, further research is needed to fully understand the topic and draw more definitive conclusions.

In this way, the rapidly expanding body of research and its extensive geographic coverage have highlighted the growing global concern surrounding psychoactive substance use (PSU) over the past two decades. Despite this, a noticeable gap persists in the literature regarding the application of spatial analysis, possibly due to challenges in obtaining spatial data. Implementing a GEOTELO framework could address this deficiency by employing geostatistical modeling to identify patterns and hotspots related to PSU, thereby facilitating targeted intervention strategies. Incorporating language and location models would enable the system to process unstructured data, enriching and broadening the scope of its analyses. As a result, this could lead to more sophisticated insights into PSU patterns and associated risks, especially in urban and suburban areas where problems like criminal activities, aggression, and child mistreatment coincide with substance misuse. Additionally, a GEOTELO framework could help mitigate the methodological heterogeneity in existing research, which encompasses diverse techniques from Bayesian analysis to machine learning. By standardizing data analysis and interpretation, GEOTELO could generate more consistent and reliable findings, allowing for easier comparisons across studies and promoting stronger conclusions. The capacity to process vast amounts of data would also empower the system to manage multiple variables and address potential confounding factors. This could help overcome limitations arising from the use of varying samples, treatment durations, and exclusion factors in individual studies, ultimately enhancing the generalizability of the results. To sum up, developing a

GEOTELO-based intervention framework could significantly improve our understanding of PSU, offering a valuable tool for researchers and policymakers alike. This innovation could revolutionize the field by providing new, data-driven approaches to tackle this intricate and multifaceted issue—an essential step considering the rising prevalence and acceptance of PSU across diverse geographical regions and research disciplines.

References

- [1] B. F. Grant *et al.*, “Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders - Results from the national epidemiologic survey on alcohol and related conditions,” *Arch. Gen. Psychiatry*, vol. 61, no. 8, pp. 807–816, 2004, doi: 10.1001/archpsyc.61.8.807.
- [2] R. C. Kessler, C. B. Nelson, K. A. McGonagle, M. J. Edlund, R. G. Frank, and P. J. Leaf, “The epidemiology of co-occurring addictive and mental disorders: Implications for prevention and service utilization,” *Am. J. Orthopsychiatry*, vol. 66, no. 1, pp. 17–31, 1996, doi: 10.1037/h0080151.
- [3] D. A. Regier *et al.*, “Comorbidity of Mental Disorders With Alcohol and Other Drug Abuse: Results From the Epidemiologic Catchment Area (ECA) Study,” *JAMA*, vol. 264, no. 19, pp. 2511–2518, 1990, doi: 10.1001/jama.1990.03450190043026.
- [4] R. Z. Goetzel, K. Hawkins, R. J. Ozminkowski, and S. H. Wang, “The health and productivity cost burden of the ‘top 10’ physical and mental health conditions affecting six large US employers in 1999,” *J. Occup. Environ. Med.*, vol. 45, no. 1, pp. 5–14, 2003, doi: 10.1097/00043764-200301000-00007.
- [5] K. Sanderson and G. Andrews, “Prevalence and severity of mental health-related disability and relationship to diagnosis,” *Psychiatr. Serv.*, vol. 53, no. 1, pp. 80–86, 2002, doi: 10.1176/appi.ps.53.1.80.
- [6] W. F. Stewart, J. A. Ricci, E. Chee, S. R. Hahn, and D. Morganstein, “Cost of lost productive work time among US workers with depression,” *JAMA-JOURNAL Am. Med. Assoc.*, vol. 289, no. 23, pp. 3135–3144, Jun. 2003, doi: 10.1001/jama.289.23.3135.
- [7] WHO, “Mental Health Action Plan 2013-2020.” Ginebra, Suiza, 2013. Accessed: Oct. 10, 2019. [Online]. Available: <https://apps.who.int/iris/bitstream/handle/>

10665/97488/9789243506029_spa.pdf;jsessionid=0E501BE5A5A6368F961A1179340D38B3?sequence=1

- [8] WHO, “International Classification of Diseases (ICD-11),” 2018.
- [9] R. Ali *et al.*, “The alcohol, smoking and substance involvement screening test (ASSIST): development, reliability and feasibility,” *ADDICTION*, vol. 97, no. 9, pp. 1183–1194, Sep. 2002.
- [10] M. Dennis *et al.*, “The Cannabis Youth Treatment (CYT) Study: Main findings from two randomized trials,” *J. Subst. Abuse Treat.*, vol. 27, no. 3, pp. 197–213, Oct. 2004, doi: 10.1016/j.jsat.2003.09.005.
- [11] B. F. Grant, D. A. Dawson, F. S. Stinson, P. S. Chou, W. Kay, and R. Pickering, “The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample,” *Drug Alcohol Depend.*, vol. 71, no. 1, pp. 7–16, Jul. 2003, doi: 10.1016/S0376-8716(03)00070-X.
- [12] J. R. Greenmyer, M. G. Klug, C. Kambeitz, S. Popova, and L. Burd, “A Multicountry Updated Assessment of the Economic Impact of Fetal Alcohol Spectrum Disorder: Costs for Children and Adults,” *J. Addict. Med.*, vol. 12, no. 6, pp. 466–473, 2018, doi: 10.1097/ADM.0000000000000438.
- [13] R. W. Hingson, T. Heeren, R. C. Zakocs, A. Kopstein, and H. Wechsler, “Magnitude of alcohol-related mortality and morbidity among US college students ages 18-24,” *J. Stud. Alcohol*, vol. 63, no. 2, pp. 136–144, Mar. 2002, doi: 10.15288/jsa.2002.63.136.
- [14] H. W. Perkins, “Social norms and the prevention of alcohol misuse in collegiate contexts,” *J. Stud. Alcohol*, no. 14, pp. 164–172, Mar. 2002, doi: 10.15288/jsas.2002.s14.164.
- [15] J. E. Schulenberg and J. L. Maggs, “A developmental perspective on alcohol use and heavy drinking during adolescence and the transition to young adulthood,” *J. Stud. Alcohol*, no. 14, pp. 54–70, Mar. 2002, doi: 10.15288/jsas.2002.s14.54.
- [16] J. A. Durlak, R. P. Weissberg, A. B. Dymnicki, R. D. Taylor, and K. B. Schellinger, “The Impact of Enhancing Students’ Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions,” *CHILD Dev.*, vol. 82, no. 1, pp. 405–432, 2011, doi: 10.1111/j.1467-8624.2010.01564.x.
- [17] B. G. Gudmundsdottir, L. Weyandt, and G. B. Ernudottir, “Prescription Stimulant

- Misuse and ADHD Symptomatology Among College Students in Iceland,” *J. Atten. Disord.*, vol. 24, no. 3, SI, pp. 384–401, Feb. 2020, doi: 10.1177/1087054716684379.
- [18] C. E. Hostinar, R. Nusslock, and G. E. Miller, “Future Directions in the Study of Early-Life Stress and Physical and Emotional Health: Implications of the Neuroimmune Network Hypothesis,” *J. Clin. CHILD Adolesc. Psychol.*, vol. 47, no. 1, pp. 142–156, 2018, doi: 10.1080/15374416.2016.1266647.
- [19] D. S. Dolliver, S. P. Ericson, and K. L. Love, “A Geographic Analysis of Drug Trafficking Patterns on the TOR Network,” *Geogr. Rev.*, vol. 108, no. 1, pp. 45–68, 2018, doi: 10.1111/gere.12241.
- [20] R. B. Felson and J. Staff, “Committing Economic Crime for Drug Money,” *CRIME Delinq.*, vol. 63, no. 4, pp. 375–390, 2017, doi: 10.1177/0011128715591696.
- [21] L. Lu, Y. Fang, and X. Wang, “Drug abuse in China: Past, present and future,” *Cell. Mol. Neurobiol.*, vol. 28, no. 4, pp. 479–490, Jun. 2008, doi: 10.1007/s10571-007-9225-2.
- [22] S. Metternich, S. Zoerntlein, T. Schoenberger, and C. Huhn, “Ion mobility spectrometry as a fast screening tool for synthetic cannabinoids to uncover drug trafficking in jail via herbal mixtures, paper, food, and cosmetics,” *DRUG Test. Anal.*, vol. 11, no. 6, pp. 833–846, Jun. 2019, doi: 10.1002/dta.2565.
- [23] E.-U. Nelson and I. Obot, “Beyond prohibition: responses to illicit drugs in West Africa in an evolving policy context,” *DRUGS AND ALCOHOL TODAY*, 2020, doi: 10.1108/DAT-07-2019-0033.
- [24] C. Potier, V. Laprevote, F. Dubois-Arber, O. Cottencin, and B. Rolland, “Supervised injection services: What has been demonstrated? A systematic literature review,” *Drug Alcohol Depend.*, vol. 145, pp. 48–68, 2014, doi: 10.1016/j.drugalcdep.2014.10.012.
- [25] N. De Bellis, *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Scarecrow Press, 2009.
- [26] A. Pritchard, “Statistical Bibliography or Bibliometrics?,” *J. Doc.*, vol. 25, pp. 348–349, Jan. 1969.
- [27] A. F. J. van Raan, “Measuring Science BT - Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems,” H. F. Moed, W. Glänzel, and U. Schmoch, Eds. Dordrecht: Springer Netherlands, 2005, pp. 19–50. doi: 10.1007/1-4020-2755-9_2.

- [28] E. C. M. Noyons, H. F. Moed, and M. Luwel, "Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study," *J. Am. Soc. Inf. Sci.*, vol. 50, no. 2, pp. 115–131, Jan. 1999, doi: [https://doi.org/10.1002/\(SICI\)1097-4571\(1999\)50:2<115::AID-ASI3>3.0.CO;2-J](https://doi.org/10.1002/(SICI)1097-4571(1999)50:2<115::AID-ASI3>3.0.CO;2-J).
- [29] F. Guo, W. Lv, L. Liu, T. Wang, and V. G. Duffy, "Bibliometric analysis of simulated driving research from 1997 to 2016," *Traffic Inj. Prev.*, vol. 20, no. 1, pp. 64–71, 2019, doi: 10.1080/15389588.2018.1511896.
- [30] E. Pallari, T. Soukup, A. Kyriacou, and G. Lewison, "Assessing the European impact of alcohol misuse and illicit drug dependence research: Clinical practice guidelines and evidence-base policy," *Evid. Based. Ment. Health*, vol. 23, no. 2, pp. 67–76, 2020, doi: 10.1136/ebmental-2019-300124.
- [31] A. Schäfer, C. Hiemke, and P. Baumann, "Consensus guideline for therapeutic drug monitoring in psychiatry (2004): Bibliometric analysis of citations for the period 2004-2011," *Nord. J. Psychiatry*, vol. 70, no. 3, pp. 202–207, 2016, doi: 10.3109/08039488.2015.1080296.
- [32] D. R. Schneider, A. Vidal-Infer, M. Bolaños-Pizarro, R. Alexandre-Benavent, F. J. B. Cañigral, and J. C. Valderrama-Zurián, "Scientific collaboration between Latin America and the European Union (2001-2010) on drug abuse from the ISI Web of Science," *Salud Ment.*, vol. 37, no. 3, pp. 199–210, 2014, [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84903882830&partnerID=40&md5=c88ca9859b3f76b0d11e2ff39896b6f6>
- [33] A. Herrán, J. Artal, and J. L. Vázquez-Barquero, "Mental health in primary care: a bibliometric study," *Aten. Primaria*, vol. 18, no. 9, pp. 502–506, 1996.
- [34] F. Lopez-Munoz, C. Alamo, F. J. Quintero-Gutierrez, and P. Garcia-Garcia, "A bibliometric study of international scientific productivity in attention-deficit hyperactivity disorder covering the period 1980-2005," *Eur. Child Adolesc. Psychiatry*, vol. 17, no. 6, pp. 381–391, Sep. 2008, doi: 10.1007/s00787-008-0680-1.
- [35] J. G. Bramness, B. Henriksen, O. Person, and K. Mann, "A Bibliometric Analysis of European versus USA Research in the Field of Addiction. Research on Alcohol, Narcotics, Prescription Drug Abuse, Tobacco and Steroids 2001-2011," *Eur. Addict. Res.*, vol. 20, no. 1, pp. 16–22, 2014, doi: 10.1159/000348260.
- [36] W. M. Sweileh, S. H. Zyoud, S. W. Al-Jabi, and A. F. Sawalha, "Substance use disorders in Arab countries: research activity and bibliometric analysis," *Subst.*

- Abuse Treat. Prev. Policy*, vol. 9, p. 33, 2014, doi: 10.1186/1747-597X-9-33.
- [37] S. H. Zyoud, W. S. Waring, S. W. Al-Jabi, and W. M. Sweileh, “Global cocaine intoxication research trends during 1975-2015: a bibliometric analysis of Web of Science publications,” *Subst. Abus. Treat. Prev. POLICY*, vol. 12, Feb. 2017, doi: 10.1186/s13011-017-0090-9.
- [38] T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross, “An fMRI-Based Neurologic Signature of Physical Pain,” *N. Engl. J. Med.*, vol. 368, no. 15, pp. 1388–1397, 2013, doi: 10.1056/NEJMoa1204471.
- [39] D. Lee, “Decision Making: From Neuroscience to Psychiatry,” *Neuron*, vol. 78, no. 2, pp. 233–248, 2013, doi: 10.1016/j.neuron.2013.04.008.
- [40] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, “Identifying adverse drug event information in clinical notes with distributional semantic representations of context,” *J. Biomed. Inform.*, vol. 57, pp. 333–349, Oct. 2015, doi: 10.1016/j.jbi.2015.08.013.
- [41] M. Conway and D. O’Connor, “Social media, big data, and mental health: current advances and ethical implications,” *Curr. Opin. Psychol.*, vol. 9, pp. 77–82, Jun. 2016, doi: 10.1016/j.copsyc.2016.01.004.
- [42] L. M. Squeglia *et al.*, “Neural Predictors of Initiating Alcohol Use During Adolescence,” *Am. J. Psychiatry*, vol. 174, no. 2, pp. 172–185, Feb. 2017, doi: 10.1176/appi.ajp.2016.15121587.
- [43] S. E. Arnold *et al.*, “Plasma biomarkers of depressive symptoms in older adults,” *Transl. Psychiatry*, vol. 2, 2012, doi: 10.1038/tp.2011.63.
- [44] G. Bedi, G. A. Cecchi, D. F. Slezak, F. Carrillo, M. Sigman, and H. de Wit, “A Window into the Intoxicated Mind? Speech as an Index of Psychoactive Drug Effects,” *NEUROPSYCHOPHARMACOLOGY*, vol. 39, no. 10, pp. 2340–2348, Sep. 2014, doi: 10.1038/npp.2014.80.
- [45] T. Katsuki, T. K. Mackey, and R. Cuomo, “Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data,” *J. Med. INTERNET Res.*, vol. 17, no. 12, 2015, doi: 10.2196/jmir.5144.
- [46] J. Struyf, S. Dobrin, and D. Page, “Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia,” *BMC Genomics*, vol. 9, Nov. 2008, doi: 10.1186/1471-2164-9-531.
- [47] N. Alvaro, M. Conway, S. Doan, C. Lofi, J. Overington, and N. Collier,

- “Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use,” *J. Biomed. Inform.*, vol. 58, pp. 280–287, 2015, doi: 10.1016/j.jbi.2015.11.004.
- [48] H. A. Whiteford *et al.*, “Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010,” *Lancet*, vol. 382, no. 9904, pp. 1575–1586, Nov. 2013, doi: 10.1016/S0140-6736(13)61611-6.
- [49] J. C. Woolcott *et al.*, “Meta-analysis of the Impact of 9 Medication Classes on Falls in Elderly Persons,” *Arch. Intern. Med.*, vol. 169, no. 21, pp. 1952–1960, Nov. 2009, doi: 10.1001/archinternmed.2009.357.
- [50] L. Degenhardt *et al.*, “Global burden of disease attributable to illicit drug use and dependence: findings from the Global Burden of Disease Study 2010,” *Lancet*, vol. 382, no. 9904, pp. 1564–1574, Nov. 2013, doi: 10.1016/S0140-6736(13)61530-5.
- [51] A. J. Collins, R. N. Foley, D. T. Gilbertson, and S.-C. Chen, “United States Renal Data System public health surveillance of chronic kidney disease and end-stage renal disease,” *KIDNEY Int. Suppl.*, vol. 5, no. 1, pp. 2–7, Jun. 2015, doi: 10.1038/kisup.2015.2.
- [52] M. J. Sung, A. Erkanli, A. Angold, and E. J. Costello, “Effects of age at first substance use and psychiatric comorbidity on the development of substance use disorders,” *Drug Alcohol Depend.*, vol. 75, no. 3, pp. 287–299, Sep. 2004, doi: 10.1016/j.drugalcdep.2004.03.013.
- [53] S. Schwan, A. Sundstrom, E. Stjernberg, E. Hallberg, and P. Hallberg, “A signal for an abuse liability for pregabalin-results from the Swedish spontaneous adverse drug reaction reporting system,” *Eur. J. Clin. Pharmacol.*, vol. 66, no. 9, pp. 947–953, Sep. 2010, doi: 10.1007/s00228-010-0853-y.
- [54] S. A. Rivkees and A. Szarfman, “Dissimilar Hepatotoxicity Profiles of Propylthiouracil and Methimazole in Children,” *J. Clin. Endocrinol. Metab.*, vol. 95, no. 7, pp. 3260–3267, Jul. 2010, doi: 10.1210/jc.2009-2546.
- [55] L. Degenhardt *et al.*, “The global epidemiology and burden of psychostimulant dependence: Findings from the Global Burden of Disease Study 2010,” *Drug Alcohol Depend.*, vol. 137, pp. 36–47, 2014, doi: 10.1016/j.drugalcdep.2013.12.025.
- [56] F. D. Bowman, B. Caffo, S. S. Bassett, and C. Kilts, “A Bayesian hierarchical framework for spatial modeling of fMRI data,” *Neuroimage*, vol. 39, no. 1, pp.

- 146–156, 2008, doi: 10.1016/j.neuroimage.2007.08.012.
- [57] J. W. O’Brien *et al.*, “A Model to Estimate the Population Contributing to the Wastewater Using Samples Collected on Census Day,” *Environ. Sci. Technol.*, vol. 48, no. 1, pp. 517–525, 2014, doi: 10.1021/es403251g.
- [58] K. Shannon, M. Rusch, J. Shoveller, D. Alexson, K. Gibson, and M. W. Tyndall, “Mapping violence and policing as an environmental-structural barrier to health service and syringe availability among substance-using women in street-level sex work,” *Int. J. DRUG POLICY*, vol. 19, no. 2, pp. 140–147, 2008, doi: 10.1016/j.drugpo.2007.11.024.
- [59] B. Chaix, J. Merlo, S. V Subramanian, J. Lynch, and P. Chauvin, “Comparison of a spatial perspective with the multilevel analytical approach in neighborhood studies: The case of mental and behavioral disorders due to psychoactive substance use in Malmo, Sweden, 2001,” *Am. J. Epidemiol.*, vol. 162, no. 2, pp. 171–182, Jul. 2005, doi: 10.1093/aje/kwi175.
- [60] B. Freisthler, “A spatial analysis of social disorganization, alcohol access, and rates of child maltreatment in neighborhoods,” *Child. Youth Serv. Rev.*, vol. 26, no. 9, pp. 803–819, Sep. 2004, doi: 10.1016/j.childyouth.2004.02.022.
- [61] J. Prosser, L. J. Cohen, M. Steinfeld, D. Eisenberg, E. D. London, and I. I. Galynker, “Neuropsychological functioning in opiate-dependent subjects receiving and following methadone maintenance treatment,” *Drug Alcohol Depend.*, vol. 84, no. 3, pp. 240–247, Oct. 2006, doi: 10.1016/j.drugalcdep.2006.02.006.
- [62] C. J. Banta-Green, J. A. Field, A. C. Chiaia, D. L. Sudakin, L. Power, and L. de Montigny, “The spatial epidemiology of cocaine, methamphetamine and 3,4-methylenedioxymethamphetamine (MDMA) use: a demonstration using a population measure of community drug load derived from municipal wastewater,” *ADDICTION*, vol. 104, no. 11, pp. 1874–1880, Nov. 2009, doi: 10.1111/j.1360-0443.2009.02678.x.
- [63] B. Freisthler, B. Needell, and P. J. Gruenewald, “Is the physical availability of alcohol and illicit drugs related to neighborhood rates of child maltreatment?,” *Child Abuse Negl.*, vol. 29, no. 9, pp. 1049–1060, Sep. 2005, doi: 10.1016/j.chiabu.2004.12.014.
- [64] K. Dovey, J. Fitzgerald, and Y. J. Choi, “Safety becomes danger: dilemmas of drug-use in public space,” *Health Place*, vol. 7, no. 4, pp. 319–331, 2001, doi:

10.1016/S1353-8292(01)00024-7.

- [65] B. Chaix *et al.*, “Spatial clustering of mental disorders and associated characteristics of the neighbourhood context in Malmo, Sweden, in 2001,” *J. Epidemiol. Community Health*, vol. 60, no. 5, pp. 427–435, May 2006, doi: 10.1136/jech.2005.040360.
- [66] R. Heimer, R. Barbour, A. V Shabolts, I. F. Hoffman, and A. P. Kozlov, “Spatial distribution of HIV prevalence and incidence among injection drugs users in St Petersburg: implications for HIV transmission,” *AIDS*, vol. 22, no. 1, pp. 123–130, 2008, doi: 10.1097/QAD.0b013e3282f244ef.
- [67] J. K. Bass and S. F. Lambert, “Urban adolescents’ perceptions of their neighborhoods: An examination of spatial dependence,” *J. Community Psychol.*, vol. 32, no. 3, pp. 277–293, May 2004, doi: 10.1002/jcop.20005.
- [68] P. M. Thompson *et al.*, “Structural abnormalities in the brains of human subjects who use methamphetamine,” *J. Neurosci.*, vol. 24, no. 26, pp. 6028–6036, Jun. 2004, doi: 10.1523/JNEUROSCI.0713-04.2004.
- [69] C. T. Mowbray, M. C. Holter, G. B. Teague, and D. Bybee, “Fidelity criteria: Development, measurement, and validation,” *Am. J. Eval.*, vol. 24, no. 3, pp. 315–340, 2003, doi: 10.1177/109821400302400303.
- [70] S. Fazel, H. Doll, and N. Langstrom, “Mental disorders among adolescents in juvenile detention and correctional facilities: A systematic review and metaregression analysis of 25 surveys,” *J. Am. Acad. CHILD Adolesc. PSYCHIATRY*, vol. 47, no. 9, pp. 1010–1019, Sep. 2008, doi: 10.1097/CHI.0b013e31817eeef3.
- [71] A. Shoji, H. Yamanaka, and N. Kamatani, “A retrospective study of the relationship between serum urate level and recurrent attacks of gouty arthritis: Evidence for reduction of recurrent gouty arthritis with antihyperuricemic therapy,” *ARTHRITIS Rheum. CARE Res.*, vol. 51, no. 3, pp. 321–325, Jun. 2004, doi: 10.1002/art.20405.
- [72] A. I. R. Maas *et al.*, “Efficacy and safety of dexamethasone in severe traumatic brain injury: results of a phase III randomised, placebo-controlled, clinical trial,” *LANCET Neurol.*, vol. 5, no. 1, pp. 38–45, 2006, doi: 10.1016/S1474-4422(05)70253-2.
- [73] M. Peet and C. Stokes, “Omega-3 fatty acids in the treatment of psychiatric disorders,” *Drugs*, vol. 65, no. 8, pp. 1051–1059, 2005, doi: 10.2165/00003495-

200565080-00002.

- [74] M. Gilbert *et al.*, “Outbreak in Alberta of community-acquired (USA300) methicillin-resistant *Staphylococcus aureus* in people with a history of drug use, homelessness or incarceration,” *Can. Med. Assoc. J.*, vol. 175, no. 2, pp. 149–154, Jul. 2006, doi: 10.1503/cmaj.051565.
- [75] A. Neubert *et al.*, “The impact of unlicensed and off-label drug use on adverse drug reactions in paediatric patients,” *DRUG Saf.*, vol. 27, no. 13, pp. 1059–1067, 2004, doi: 10.2165/00002018-200427130-00006.
- [76] M. Wazaify, E. Shields, C. M. Hughes, and J. C. McElnay, “Societal perspectives on over-the-counter (OTC) medicines,” *Fam. Pract.*, vol. 22, no. 2, pp. 170–176, 2005, doi: 10.1093/fampra/cmh723.
- [77] M. H. Merson, J. M. Dayton, and K. O’Reilly, “Effectiveness of HIV prevention interventions in developing countries,” *AIDS*, vol. 14, no. 2, pp. S68–S84, Sep. 2000.

Chapter 3

3. Leading Consumption Patterns of Psychoactive Substances in Colombia: A Deep Neural Network-Based Clustering-Oriented Embedding Approach

3.1. Abstract

The number of health-related incidents caused using illegal and legal psychoactive substances (PAS) has dramatically increased over two decades worldwide. In Colombia, the use of illicit substances has increased up to 10.3%, while the consumption alcohol and tobacco has increased to 84% and 12%, respectively. It is well-known that identifying drug consumption patterns in the general population is essential in reducing overall drug consumption. However, existing approaches do not incorporate Machine Learning and/or Deep Data Mining methods in combination with spatial techniques. To enhance our understanding of mental health issues related to PAS and assist in the development of national policies, here we present a novel Deep Neural Network-based Clustering-oriented Embedding Algorithm that incorporates an autoencoder and spatial techniques. The primary goal of our model is to identify general and spatial patterns of drug consumption and abuse, while also extracting relevant features from the input data and identifying clusters during the learning process. As a test case, we used the largest publicly available database of legal and illegal PAS consumption comprising 49,600 Colombian households. We estimated and geographically represented the prevalence of consumption and/or abuse of both PAS and non-PAS, while achieving statistically significant goodness-of-fit values. Our results indicate that region, sex, housing type, socioeconomic status, age, and variables related to household finances contribute to explaining the patterns of consumption and/or abuse of PAS. Additionally, we identified three distinct patterns of PAS consumption and/or abuse. At the spatial level, these

patterns indicate concentrations of drug consumption in specific regions of the country, which are closely related to specific geographic locations and the prevailing social and environmental contexts. These findings can provide valuable insights to facilitate decision-making and develop national policies targeting specific groups given their cultural, geographic, and social conditions.

3.2. Introduction

Psychoactive substances (PAS) are chemical substances that change the function of the nervous system and cause alterations in people's perception, mood, consciousness, cognition, or behavior [1]. PAS can be grouped according to their chemical structure as synthetic cannabinoids, synthetic cathinones, phenethylamines, arylcyclohexylamines, tryptamines, indolalkylamines, new synthetic opioids, piperazines, ketamine, and designer benzodiazepines. They can also be grouped according to their origin as a natural origin or synthetic molecules [2–4]. The increased numbers of drug use among young people are drawing the attention of national governments [5]. Because the number of health-related incidents caused by using legal and illegal PAS worldwide has dramatically increased over the last two decades [2, 6], this phenomenon has become some of the largest burdens of disease [7, 8]. Drug use constitutes a high cost to society due to premature mortality, increased health expenditure, criminal justice (drug and micro-trafficking), social welfare costs, and other social consequences [9, 10].

Colombia is ranked as one of the largest drug producers in the world [11]. Unfortunately, the production and commercialization of drugs through drug- and micro-trafficking, constantly expands in locations with high levels of poverty and limited government presence [12]. Statistics and indicators of drug consumption, production, and distribution, as well as reports from the National Statistical System (DANE) of Colombia, highlight a dramatic increase in (1) drug production [13]; (2) intern consumption of PAS at early ages; and (3) the prevalence in use and/or abuse of drugs have dramatically increased over the last 20 years [14, 15]. Furthermore, the country also has the highest prevalence of drug use among school students in recent years compared to other Latin American countries [16]. Thus, there is an urgent need to develop effective interventions to prevent the use and/or abuse of PAS. The first step towards reducing this consumption is to identify drug consumption patterns in the general population [17]. Several studies have identified patterns associated with drug use and consumption [18, 19]. According to the Center for

Disease Control and Prevention, individuals who do not have their own homes and live in rented accommodations are more likely to use drugs [20]. Other research studies suggest that neighborhood contextual characteristics may increase the risk of substance abuse [21–26]. Additionally, population density may also influence substance use and overdose risk through a higher level of socialization in densely populated urban areas [27–29]. Other authors have identified that anxiety, sleep disorders, suicide, depression, and other mental illnesses are risk factors for the consumption and abuse of PAS [30,31]. Furthermore, early marijuana use has been shown to increase the risk of consuming other PAS [32]. Furthermore, people involved in sports and artistic activities perceive drugs as enhancers element for improving their performance [33].

In Colombia, drug consumption patterns and risk factors have also been identified. For instance, Kalyanam et al. [34] analyzed the social impact of basuco and inhalant use among street youths. Narvaez-Chicaiza [35] assessed the social factors that lead to the adoption of harm reduction policies and how these factors influence treatments for substance abuse disorders. Additionally, Restrepo-Escobar & Cardona [36] demonstrated that university students with low satisfaction in their studies tend to be heavy users of alcohol, tobacco, and marijuana. However, these approaches do not consider the use of Machine Learning (ML) and/or Deep Data Mining techniques in combination with spatial models to analyze drug consumption data from the general population. To our knowledge, we have not found any robust models integrating ML and spatial models to identify drug consumption patterns using publicly available Colombian databases.

Although several techniques for analyzing drug consumption patterns are currently available (i.e., ML, Bayesian, spatial, traditional multivariate, or univariate statistical models, or, in some cases, a combination of these), new trends in pattern identification and analysis techniques focus on hybrid and ensemble models [37]. Currently, the most widely used ML techniques are Support Vector Machines (SVMs), Random Forest (RF), and Natural Language Processing (NLP)[38–42]. Among Bayesian models, Bayesian meta-regression (DisMod-MR), Bayesian hierarchical models, and Markov Chain Monte Carlo are the most attractive methods [43–45]. Regarding spatial models, Spatial Distributions, Spatial Regression Models, Spatial Scan Statistics, Variograms, and Social Mapping are the most frequently used techniques [46–49]. On the other hand, logistic regression, confirmatory factor analysis, and correlational analysis are the most employed traditional statistical models to identify drug-associated patterns [50–53].

Fraley and Raftery [54] suggest separating clustering approaches into hierarchical and partitioning techniques. Partitioning techniques are divided into density-, model-, and grid-based methods, the most popular of which are K -means, PAM, CLARA, DBSCAN and CLIQUE. On the other hand, hierarchical techniques are divided into agglomerative and divisive methods. Of these, the best-known methods are BRICH, CURE, ROCK, and CHAMELEON (see [55] for further reading). Although these techniques have been shown to perform well when relevant features are removed *a priori*, it is well-known that in clustering algorithms, irrelevant and redundant features in the data may degrade the quality of clusters and lead to high computational cost. Therefore, removing such features may alleviate these issues. Thus, we focus on identifying patterns of PAS consumption using an ensemble model integrating an autoencoder with both a clustering algorithm and a spatial model. As part of our approach, we used the most recent and representative works for data clustering, and different dimensionality reduction and feature selection methodologies proposed in the literature.

Feature selection approaches in clustering can be split into filter, wrapper, embedded, and hybrid approaches [37]. While wrappers depend on the clustering algorithms to evaluate the clustering quality of a selected feature subset, filters are independent of the clustering algorithm. Embedded approaches also work with a clustering algorithm and, unlike wrappers, incorporate knowledge about the clustering structure. Another type of method is hybrid approaches, which combine filter and wrapper approaches into a single strategy. However, studies on embedded and hybrid feature selection approaches in clustering are limited [37]. Other feature learning-based approaches using Deep Neural Networks have been shown to work well for linear and nonlinear models [56]. For instance, Xie et al. in (2016) [57] propose to work on feature extraction and clustering using pre-trained Auto-Encoders simultaneously. However, these are mainly used to work and process images. In general, deep clustering models use Auto-Encoders since they can learn input features without labels on the data; performance measures show that this approach is reliable for different data types [58]. Thus, deep clustering methods have become a growing field of research for feature selection [58]. In this regard, the use of convolutional networks in autoencoders and the application of feature selection for clustering are open questions that have not been fully addressed yet, especially when dealing with data from different statistical distributions [37].

Here, we propose a Deep Neural Network-based Clustering-oriented Embedding Algorithm that allows us to (i) identify consumption patterns of PAS; and (ii) build an ensemble algorithm integrating an autoencoder with a clustering algorithm and a spatial model to deal with the feature space and cluster memberships. Our approach is based on the model proposed by Xie et al. [57] and B. Li et al. [59], and expands their work by creating an autoencoder from a convolutional network to represent high-order interactions in the data accurately and simultaneously incorporate a spatial analysis to describe drug consumption patterns properly. Our main hypothesis is that incorporating these two critical elements in our proposal will help to identify and better understand drug consumption patterns and support national policy development processes.

3.3. Materials and Methods

3.3.1. Study Area

Located in South America, the Republic of Colombia is a diverse country with a population of over 50 million people distributed over a territory of 440,831 square miles [60], encompassing jungles, highlands, grasslands, deserts, coasts, and islands, distributed in six regions and 32 departments (states) [61]. It is worth noting that, unfortunately, Colombia has been a major producer of illegal drugs for a long time, which has had a significant impact on drug consumption and abuse. According to the United Nations Office on Drugs and Crime, Colombia is the first cocaine-producing country and the eighth country with the highest production of cannabis [62]. In addition, the Colombian Drug Observatory indicates that the use of illicit substances in the territory has increased to 10.3%, with men between the ages of 18 and 24 being the heaviest consumers of these types of drugs. Reports also indicate that consumption of licit substances such as alcohol and tobacco has recently increased dramatically [63].

3.3.2. Data Sources

We used two databases to identify drug consumption patterns in Colombia. The first database was retrieved from the 2019 National Survey of Psychoactive Substance Consumption in the General Population (DANE-DIMPE-ENCSPA-2019; URL: <https://microdatos.dane.gov.co/index.php/catalog/680/data->

dictionary> conducted by the National Statistical System (DANE) of Colombia [64]. This survey includes observations of 49,600 households, where information on housing, location, general characteristics of individuals, consumption of legal and illegal PAS, and implemented treatments is registered. The second database comes from the Colombian Drug Observatory and contains information on the production of PAS per area during 2019. All these databases are fully available and completely anonymized. In this study, we used departments (states) as georeferenced areas using polygons (i.e., a shapefile) as implemented in ArcGIS Hub [65]. Thus, an ethics statement approved by an ethics committee is not required since we are using public information without the identification or individual information of the people involved.

3.3.3. Convolutional Auto-Encoder-Deep Embedded Clustering Algorithm

Figure 10 presents the proposed Convolutional Auto-Encoder- Deep Embedded Clustering (CAE-DEC) framework based on the implementation presented by Xie et al. [57]. However, unlike the Xie et al. model, our structure is developed by applying convolutional layers for the deep autoencoder (DA) architecture instead of a linear one to represent high-order interactions in the data. In addition, a spectral clustering-based centroid estimation is proposed to achieve an improved initial centroid calculation. We chose the CAE-DEC framework based on its ability to reduce both the number of model parameters and the dimensionality, while creating clusters simultaneously.

In this approach, the input data is first mapped to a lower-dimensional feature space called the latent feature space (LFS) using an encoder structure. The encoder tries to extract relevant information from the input data and compresses it into a lower-dimensional representation. Next, the LFS is passed through a decoder structure to reconstruct the original input data. The decoder tries to reconstruct the original input data as accurately as possible from the lower-dimensional representation. Simultaneously, the LFS is also passed through a clustering layer that aims to perform an improved clustering assignment. The clustering layer computes the probability distribution of each data point belonging to different clusters based on the distance between the data point and the cluster centroids. The clustering layer seeks to minimize the divergence between the target distribution and the centroid-based probability distribution. This means that the clustering layer tries to find the best possible assignment of data points to clusters that are consistent with the

distances between the data points and the cluster centroids. Overall, the encoder-decoder combination and the clustering layer work together to achieve an efficient and accurate clustering framework that preserves relevant information from the original input data.

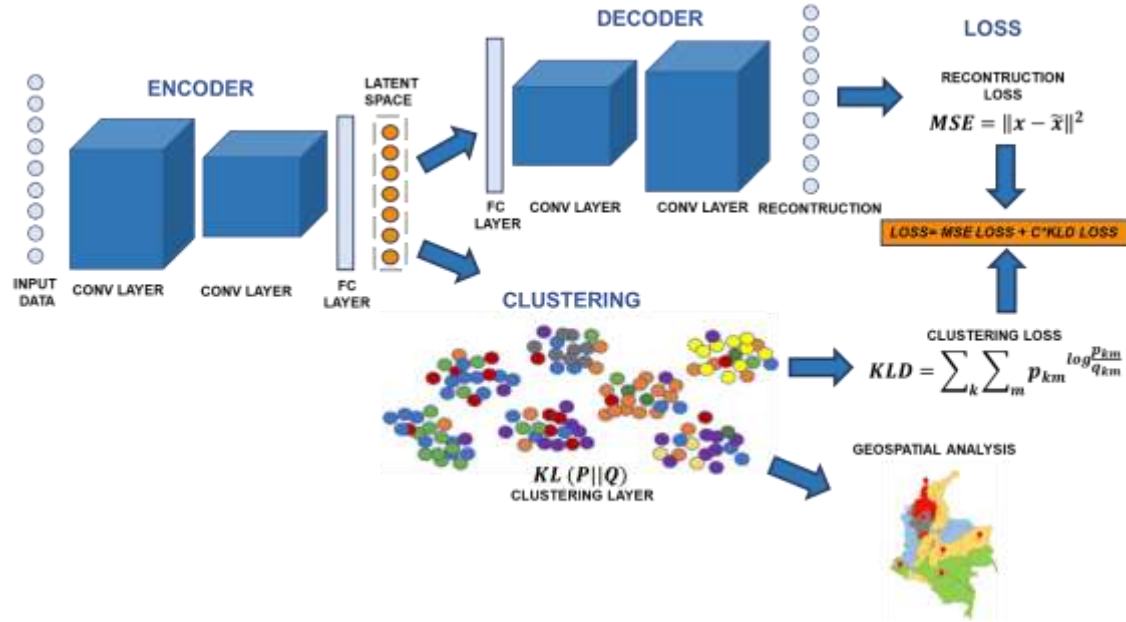


Figure 10. Architecture of the proposed CAE-DEC model.

In the last stage of the framework, a spatial analysis was performed using the feature space generated from the autoencoder as input. Here, the spatial data exploration is initially performed using Global Spatial Autocorrelation to determine to which level the similarity between observations in a dataset relates to the similarity of the locations of such observations [66]. To assess GSA, the Moran's I [67], Geary's C [68], and Getis and Ord's G [69] statistics are estimated. We also measure the Local Spatial Autocorrelation, which focuses on the relationships between each observation and its surroundings, rather than providing a single-number summary of these relationships across the map [70]. This is estimated based on the ability to determine whether spatial autocorrelation is present in a geographically referenced data set. Finally, we perform regionalization, which corresponds to a special kind of clustering where the objective is to group similar observations based on their statistical attributes and spatial location [71]. In this sense, regionalization embeds the same logic as standard clustering techniques while applying a series of geographical constraints [72]. The Python version 3.11 was utilized along with the libraries TensorFlow <<https://www.tensorflow.org/>> and PyTorch <<https://pytorch.org/>> to construct this framework [73].

3.3.4. Convolutional Auto-Encoder (CAE)

The DA is a deep neural network architecture capable of learning unsupervised representations of an input data set. Typically, DA networks are used for dimensionality reduction or denoising tasks. The structure of a DA is based on two deep networks: a network to transform the original input data into a latent feature space, and a network trained to reconstruct the original input data using the extracted latent space as input. The first network, used to extract the latent space, is called the encoder, while the second is called the decoder. Rather than using fully connected layers, the implemented DA architecture incorporates convolutional (CONV) layers and fully connected (FC) layers for LFS extraction and reconstruction (See Figure 1). Integrating convolutional layers in a DA is also called CAE [74]. Compared to a DA, which is built with only fully connected layers, the CAE structure can reduce the number of parameters compared to a DA [75].

3.3.4.1. Convolutional Layer

The proposed CAE structure is designed using four convolutional layers, two CONV layers during the encoder stage, two CONV layers during the decoder stage, and two fully connected layers (Figure 1). The convolution operation can be denoted as:

$$z_{i,j,k}^l = W_k^{lT} X_{i,j}^l + b_k^l \quad (1)$$

where $z_{i,j,k}^l$ is the value of each feature at the (i, j) location in the k -th feature map of the l -th layer, W_k^l and b_k^l represent the weight and bias of the k -th filter of the l -th layer, and $X_{i,j}^l$ denotes the input value at location (i, j) of the l -th layer. For non-linear mapping, an activation function $g(\cdot)$ is applied over the convolutional feature $z_{i,j,k}^l$ as follows:

$$a_{i,j,k}^l = g(z_{i,j,k}^l) \quad (2)$$

where $a_{i,j,k}^l$ is the activation value resulting from applying the activation function $g(\cdot)$. The Rectified Linear Unit (ReLU) function is set as the activation function on each CONV, except in the final decoder CONV layer where a sigmoidal activation is applied.

3.3.5. Clustering Layer

The clustering layer is inspired by Xie et al.[57]. Initially, a soft assignment is computed between the latent space, also known as embedded space, and the cluster centroids. Then, update steps are repeated to define the final cluster centroids and embedded space. The Kullback–Leibler (KL) divergence is used as loss function during the optimization procedure. The objective is to minimize the KL divergence between a soft clustering distribution Q and an auxiliary target distribution P . The KL loss is calculated as:

$$L_c = KL(P\|Q) = \sum_k \sum_m p_{km} \log \left(\frac{p_{km}}{q_{km}} \right) \quad (3)$$

where L_c is the clustering loss. To measure the similarity between embedded point z_k and the cluster centroid c_m , the t Student's distribution is used as a kernel:

$$q_{km} = \frac{(1 + \|z_k - c_m\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_i (1 + \|z_k - c_i\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \quad (4)$$

with α the degrees of freedom of the t Student's distribution and q_{km} is a soft clustering assignment distribution of each embedded point (i.e., probability of assigning point k to cluster m). As in Xie et al., when setting $\alpha=1$ the similarity function q_{km} can be calculated as:

$$q_{km} = \frac{(1 + \|z_k - c_m\|^2)^{-1}}{\sum_i (1 + \|z_k - c_i\|^2)^{-1}} \quad (5)$$

To compute the target distribution p_{km} , the second power of q_{km} is calculated, and a cluster normalization is applied as follows:

$$p_{km} = \frac{q_{km}^2 / \sum_k q_{km}}{\sum_i (q_{ki}^2 / \sum_k q_{ki})} \quad (6)$$

Then, by minimizing the divergence between P and Q , the embedding learning is achieved through highly confident assignments.

3.3.5.1. Center Initialization

As previously mentioned, the cluster centroids are initialized using a spectral clustering-based approach. The spectral clustering allows flexible distance metrics and provides better cluster estimations than K -means [57]. However, most spectral clustering algorithms have high computational requirements. To overcome these computational requirements, random samples are taken to estimate the cluster centroids. As spectral clustering does not estimate any centroid during the learning process, once the clusters are defined, the mean of each cluster is used as the centroid estimator.

3.3.6. The CAE-DEC Model

Initially, the input data is normalized within the interval $[0, 1]$. This normalization allows the network to use the most advanced learning rate and avoid the vanishing gradient problems, as well as alleviate overfitting. Further, to achieve a better learning process, the last CONV layer in the decoder structure is activated by a sigmoid activation function. Then, two training steps will be executed during the CAE-DEC learning process. Firstly, a CAE model will be trained to minimize the reconstruction loss L_r computed as

$$L_r = \|x - \tilde{x}\|^2 \quad (7)$$

where x is the normalized input and \tilde{x} is the reconstructed output. This pretrained CAE model is then used as the DA structure in the CAE-DEC model.

In the second step, the CAE-DEC model is trained to simultaneously minimize reconstruction loss and clustering loss. The total loss during this training step will be set as:

$$L_t = L_r + C \cdot L_c \quad (8)$$

where L_r is the CAE-DEC reconstruction loss, L_c is the CAE-DEC clustering loss, and C is a coefficient to control the loss balance. The training process is shown in Table 3. The goal is to obtain a latent space that minimizes the total loss. Finally, the label of each embedded point is established as:

$$Label_j = \arg \max_m q_{jm} \quad (9)$$

where q_{jm} is the probability that point j belongs to a specific cluster center m . On the other hand, the maximum number of iterations $Mint$ and the target distribution P update condition P_change was chosen based on multiple experiments. The final $Mint$ and P_change values were 3000 and 5, respectively. This final P_change improves stability during the training process.

Table 3. Pseudo code for the CAE-DEC training process.

Pseudo code: The CAE-DEC training process

Input data: Number of clusters n ; Normalized input data x ; Maximum number of iterations $Mint$; Balance coefficient C ; Pretrained CAE; Stop condition $Stop$; Target distribution P update condition P_change .

Training process:

1. Generate an initial latent space (Z) through the pre-trained CAE
2. Run spectral clustering with Z to generate the initial cluster centers (C)
3. Initialize the CAE-DEC model with the pretrained CAE.
4. Calculate soft assignment distribution Q and target distribution P based on Z and C

for epoch < $Mint$ do:

if epoch % P_change == 0 then:

Calculate soft assignment distribution Q and target distribution P based on Z and C

end if

Feed the CAE-DEC with the normalized input data x

Calculate the reconstruction loss and the clustering loss

Update CAE-DEC parameters. Weight, Bias, and Centers.

if $Stop$ == True then:

Break

end if

end for

Obtain the label for each data point from the las optimized Q .

Output: Latent space, labels

3.3.7. Framework Evaluation

We trained the CAE-DEC method using data retrieved from the National Survey of Psychoactive Substance Consumption (DANE-DIMPE-ENCSPA-2019), which contains 49,600 observations. A second database with PAS production figures, was used in the spatial analysis stage to correlate the PSA consumption and production. In order to evaluate the framework, we compared our CAE-DEC approach with other approaches, including CAE, and Principal Component Analysis integrated with clustering (PCA-DEC). For evaluation and comparison purposes, we use the Calinski-Harabasz [76], Davies-Bouldin [77], and Silhouette [78] index as intrinsic clustering metrics. In addition, we used the χ^2 statistic to investigate potential associations and differences among the patterns (clusters) identified using our approach.

3.4. Results

3.4.1. Model Comparison for Identifying Drug Consumption Patterns

Figure 11 depicts the LFS resulting after applying the CAE and CAE-DEC models to the data. Among all individuals, we identified three different clusters; 14935 (30.19%) individuals belong to cluster 0, 11528 (23.30%) individuals belong to cluster 1, and 23005 (46.50%) individuals belong to cluster 2. Interestingly, the LFS generated with the CAE-DEC has more defined clusters than the CAE model. The goal of a CAE is to learn an encoding function that maps the input data to a lower-dimensional latent feature space (LFS), while still preserving the essential characteristics of the input data. However, the proposed CAE-DEC model goes beyond the traditional CAE by introducing a clustering component to the encoder structure. This clustering component forces the encoder to generate representative clusters in the latent feature space, which can then be used for tasks such as unsupervised classification or clustering. In other words, the CAE-DEC model not only preserves the important characteristics of the input data, but it also forces the encoder to learn a more meaningful and structured latent feature space that can be used for downstream tasks. By doing so, the model can extract more informative features that can lead to better performance on tasks such as classification or clustering.

Furthermore, it's observed that the reconstruction loss incurred through the Convolutional Autoencoder (CAE) model is more substantial compared to the CAE-Deep Embedded Clustering (CAE-DEC) model. This outcome might be associated with the CAE-DEC model incorporating a pre-trained CAE model during its formation, which could have helped in reducing the reconstruction loss. The CAE model, in its standalone form, does not possess the capability to identify the labels of individual data points or to categorize them into specific clusters. This limitation necessitates the use of additional techniques to facilitate such classifications. For instance, spectral clustering was employed to derive the clusters visible in Figure 11. These spectral clusters, in turn, served as foundational elements for initializing the centroids in the CAE-DEC model. This means that the initial positioning of the centroids within the CAE-DEC model was guided by the cluster data

obtained through spectral clustering, thus providing a starting point for the iterative process of refining the clusters in the CAE-DEC model.

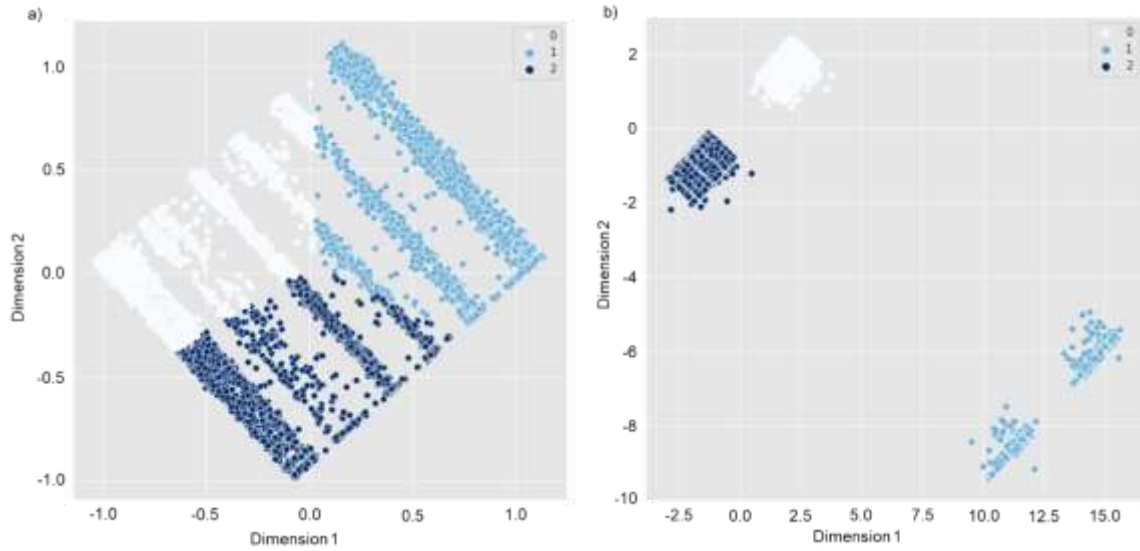


Figure 11. Derived Latent Feature Space based on the (a) CAE and (b) CAE-DEC models.

3.4.2. Identification of Psychoactive Drugs Consumption Clusters

Here we analyse the patterns in each cluster obtained using the CAE-DEC model. We defined a priority dummy variable Y_{ij} quantifying whether the i th person in household j th has consumed PAS; $Y_{ij} = 1$ when an individual has never consumed PAS and $Y_{ij} = 2$ otherwise. Out of the 49468 individuals in the sample, only 5514 (11.15%) consume PAS. Figure 12a and Figure 12b depict, respectively, the derived cluster structure for individuals consuming PAS and those who reported not consuming, derived from the CAE-DEC model. Our results indicate that individuals in clusters 0 and 2 are more likely to consume some PAS (Figure 12a), while most individuals in cluster 1 do not (Table 4). In particular, 1726 individuals (11.56%) in cluster 0, 392 (3.4%) individuals in cluster 1, and 3396 (14.76%) individuals in cluster 2 have used PAS (Table 4). A χ^2 -based test of independence reveals that the region where individuals are located, age (years), the type of household they live in, their socioeconomic status (SES), and whether they contribute to the household finances are statistically significantly associated with the cluster they belong to (Table 4).

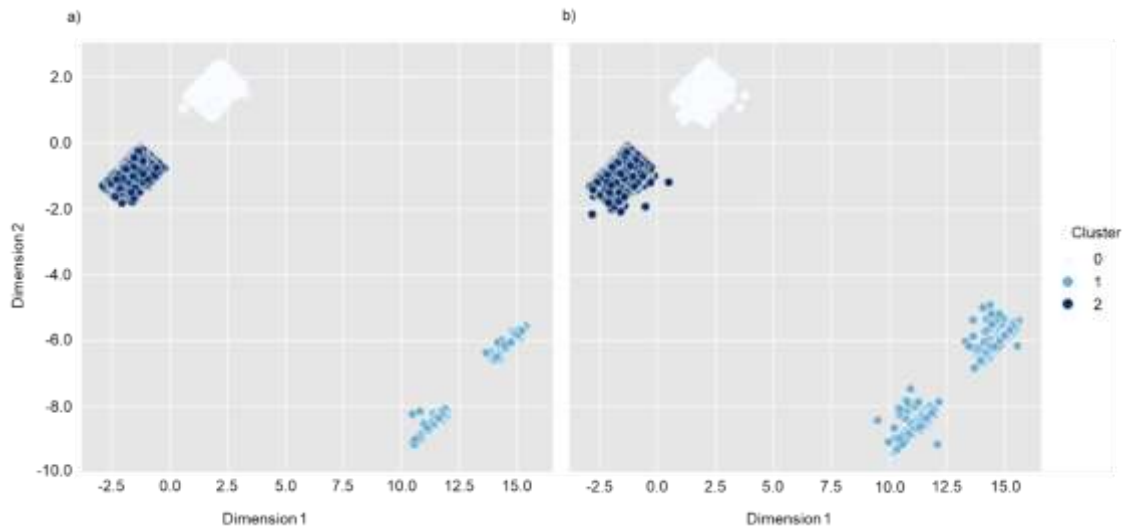


Figure 12. Resulting clusters for individuals.
 (a) consuming and (b) not consuming psychoactive substances based on the CAE-DEC model.

Table 4. Distribution of demographic and social variables across clusters.

Variables	Cluster 0 (n=14935)	Cluster 1 (n=11528)	Cluster 2 (n=23005)	χ^2	df	P-value	
Region	Caribbean	3004	2991	4075	1472.9	10	<.0001
	Central-Eastern	2526	2299	6175			
	Central-Southern	1843	1299	1702			
	Eje Cafetero – Antioquia	3902	2354	6371			
	Llanos Orientales	1687	1305	1496			
	Pacific	1973	1280	3186			
Gender	Male	6606	3927	10233	386.72	2	<.0001
	Female	8329	7601	12772			
Housing type	House	8250	6538	11834	114.18	6	<.0001
	Apartment	6275	4724	10595			
	Room	395	256	543			
	Indigenous dwelling	15	10	33			
Socioeconomic status	1	3889	3626	6945	843.92	10	<.0001
	2	4627	3902	8851			
	3	4284	2867	5683			
	4	1341	718	979			
	5	510	256	362			
	6	284	159	185			

	(0, 20]	2105	1576	2998			
Age (years)	(20, 40]	6814	4329	10889	345.87	4	<.0001
	(40, 68]	6016	5623	9118			
Contribute to the household finances	Yes	10128	7473	16041	85.24	2	<.0001
	No	4807	4055	6964			

Table 5 shows the adjusted residuals for our model. According to our results, the Central-Eastern region significantly contributes to the Region variable. In this region, the observed value is higher than the expected value in cluster 2, while the observed value is lower than the expected value for cluster 0. Although to a lesser extent, the Llanos Orientales region also significantly contributes the χ^2 statistic. Indeed, this region shows fewer observed individuals than the expected number of individuals in cluster 2 and a higher number observed than expected individuals in clusters 0 and 1. On the other hand, Gender has a higher-than-expected value of males in clusters 0 and 2, while it is lower in cluster 1. For females, the opposite occurs in cluster 1, and lower values are observed in clusters 0 and 2. Similarly, Housing Type has a higher-than-expected value of individuals living at houses in cluster 1 and a lower-than-expected in cluster 2. Conversely, cluster 2 has more individuals living in apartments, and cluster 1 has the lowest.

Table 5. Adjusted residuals comparing the observed and expected frequencies.

	Variables	Cluster 0 (n=14935)	Cluster 1 (n=11528)	Cluster 2 (n=23005)
Region	Caribbean	-0.88	17.02	-13.61
	Central-Eastern	-18.72	-6.76	22.97
	Central-Southern	12.54	6.09	-16.7
	Eje Cafetero - Antioquia	2.02	-14.36	10.31
	Llanos Orientales	11.32	9.59	-18.55
	Pacific	0.84	-6.97	5.13
Gender	Male	6.68	-19.66	10.52
	Female	-6.68	19.66	-10.52
Housing type	House	4.17	7.13	-9.88
	Apartment	-4.83	-6.61	10.05
	Room	2.2	-1.54	-0.72
	Indigenous dwelling	-0.72	-1.09	1.59

Socioeconomic status	1	-10.26	5.99	4.37
	2	-12.72	-3.3	14.51
	3	9.14	-3	-5.87
	4	17.29	0.44	-16.29
	5	11.12	-0.49	-9.82
	6	8.26	1.2	-8.62
Age (years)	(0, 20]	2.54	0.61	-2.85
	(20, 40]	3.2	-17.23	11.66
	(40, 68]	-4.98	16.93	-9.77
Contribute to the household finances	Yes	-0.61	-8.37	7.65
	No	0.61	8.37	-7.65

Regarding SES, a higher-than-expected number of individuals in strata 3, 4, 5, and 6 in cluster 0 were found (Table 5). We also observed a lower-than-expected number of individuals in strata 3, 4, 5, and 6 in cluster 2 and a higher-than-expected number in strata 1, 4 and 6 in cluster 1 (Table 5). Moreover, the age variable shows a higher-than-expected observed value for the (0,20] range in cluster 0. For ages between (20,40] years, cluster 2 has a higher-than-expected number of individuals. Conversely, there is a lower number of individuals in cluster 1. Finally, the household economy variable results show that cluster 2 has a higher-than-expected value of individuals contributing to the household finances, and cluster 1 has a lower-than-expected value of individuals not contributing to it. Comparison of Calinski-Harabasz, Davies-Bouldin, and silhouette metrics between a principal component analysis (PCA)-based deep autoencoder (PCA-DEC) and our proposed CAE-DEC model indicates the superiority of the latter (Table 6)

Table 6. Performance metrics for different models.

Performance metric	CAE-DEC	PCA- <i>K</i> -means	CAE-Spectral
Calinski-Harabasz	775992.45	128651.83	22468.26
Davies-Bouldin	0.2898	0.567	0.63
Silhouette	0.786	0.6061	0.62

3.4.3. Spatial Analysis of Psychoactive Drugs Consumption

Different alternative classification algorithms were used to determine the number of choropleth class limits (i.e., Equal Intervals, Quantiles, Maximum Breaks, Box plot, Head-Tail Breaks, Jenks-Caspall, Fisher-Jenks, and Max-p) and compared using the absolute deviation around class medians optimization criterion (Figure 13). According to our results, the Fisher-Jenks classifier performed better and hence was selected.

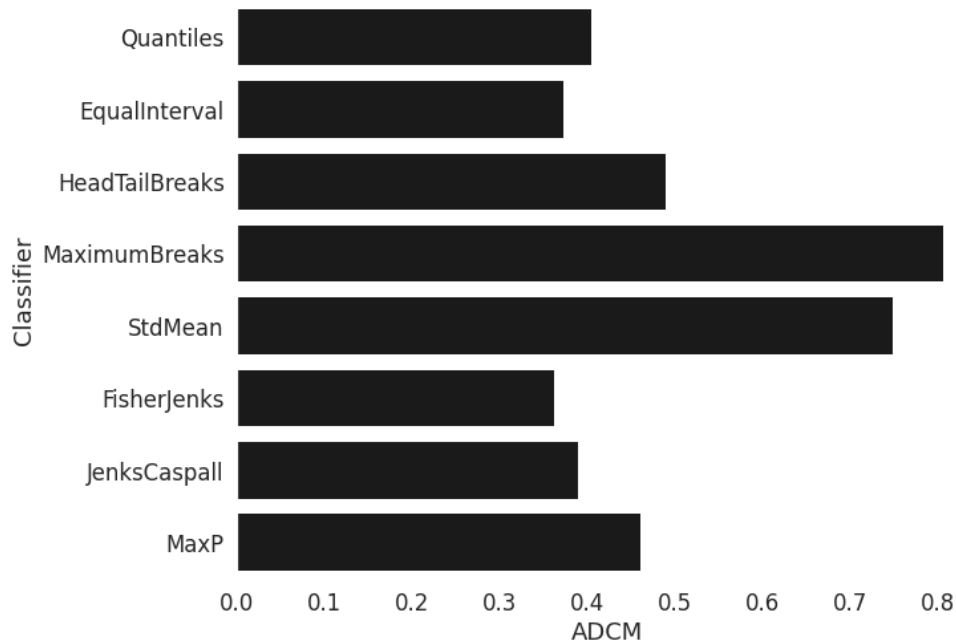


Figure 13. Absolute deviation around class medians (ADCM).

After conducting an exploratory spatial analysis, we proceeded to create a choropleth map that displays the percentage of PAS usage in all of the 32 departments in Colombia (refer to Figure 14a). The results indicated that there are certain departments such as Arauca, Vichada, Caquetá, Chocó, Magdalena, Cesar, Bolivar, Sucre, Cordoba, and Norte de Santander where drug usage rates were comparatively low. However, it is important to note that some of these departments like Cordoba and Guaviare are known to be major drug-producing regions, as highlighted by the data from the Drug Observatory of Colombia [79]. Similarly, we observed that Putumayo has the highest percentage of PAS usage when compared to other departments, as illustrated by Figure 14a. The global Moran's I results show the presence of a statistically significant positive global spatial autocorrelation ($I = 0.2005$, $P < 0.01$). Thus, the null hypothesis that the map is random (i.e., that the map shows more spatial patterns than we would expect if the values had been randomly assigned to a location) is rejected. In addition, other global indices such

as Geary's C ($C = 0.693, P = 0.003$) and Getis and Ord's G ($G = 0.800, P = 0.049$) confirm the presence of statistically significant global spatial autocorrelation.

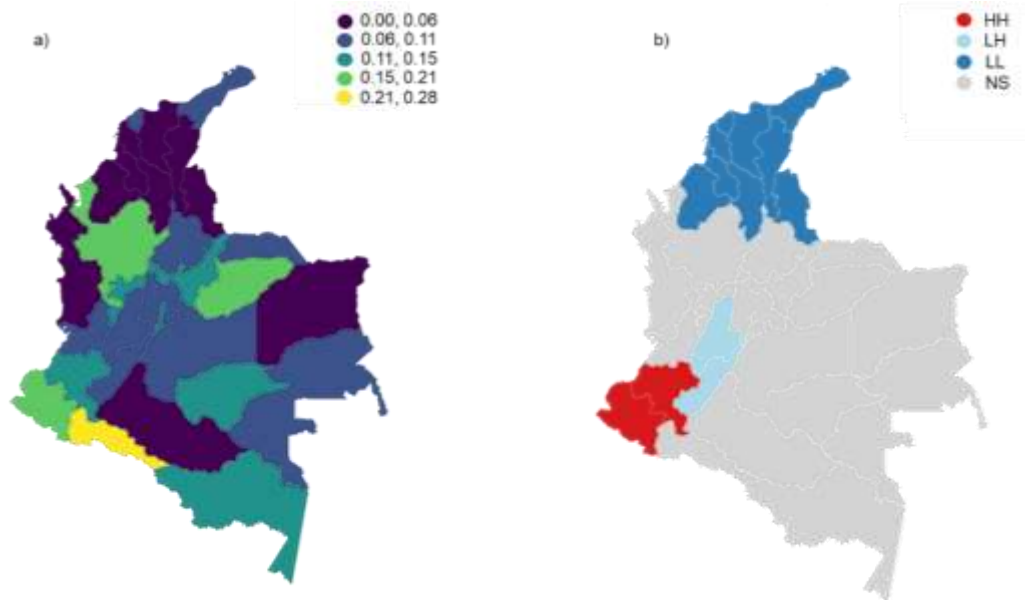


Figure 14. Cluster maps.

(a) Consumption percentage of psychoactive substances according to the CAE-DEC model; (b) Moran's I statistic; (c) Moran's cluster map. Here, HH, LH, LL and ns represent high-high, low-high, low-low, and not statistically significant quadrants, respectively. This clustering pattern leads to a statistically significant Moran's I statistic of 0.2 (P -value < 0.01).

To further explore the relationships between each observation and its environment, the Local Indicators of Spatial Association (LISA) were estimated (Figure 15). Figure 14b depicts the Moran diagram, indicating each quadrant's positive (or negative) association. Specifically, the high-high (HH) and low-low (LL) quadrants indicate a positive association between high and low drug use. On the other hand, the low-high (LH) and high-low (HL) quadrants indicate negative associations with drug use (Figure 14b). Following our results, we found that departments such as Nariño and Cauca belong to the HH cluster. In contrast, la Guajira, Atlántico, Magdalena, Cesar, Norte de Santander, Sucre, and Cordoba belong to the LL. This clustering pattern leads to a statistically significant Moran's I statistic (P -value < 0.01). Thus, a little over 39.4% of the departments are considered, by this analysis, to be part of a spatial cluster (i.e., statistically significant with a P -value $< 5\%$). We also identified that, among legal drugs, alcohol and tobacco are the most frequently consumed in the national territory (Figure 16a). At the same time, marijuana, followed by non-prescription tranquilizers and Yagé, and a slight

consumption of opioids and Poppers, are the most frequently consumed illegal drugs (Figure 16b).

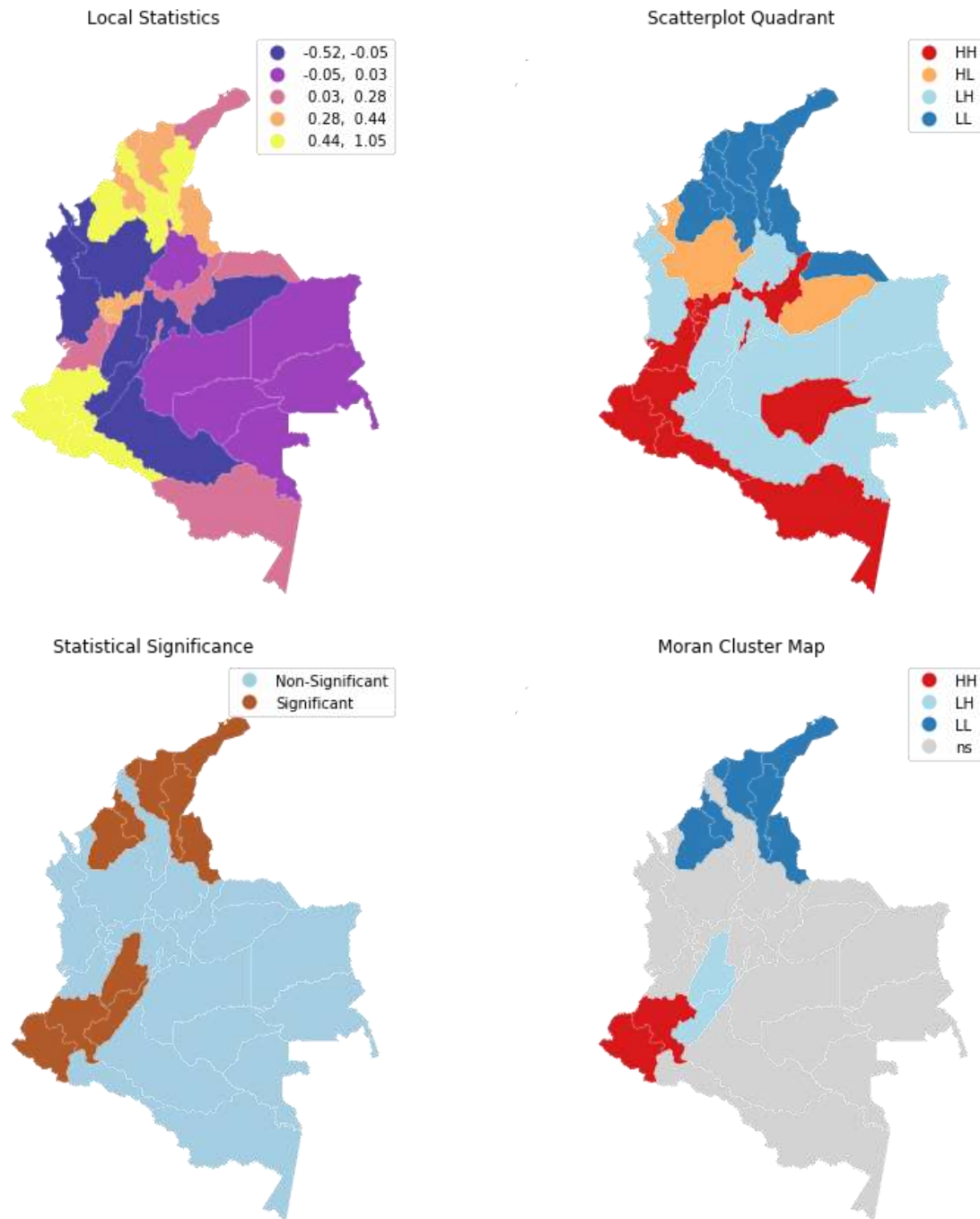


Figure 15. Maps of Local Indicators of Spatial Association (LISA).

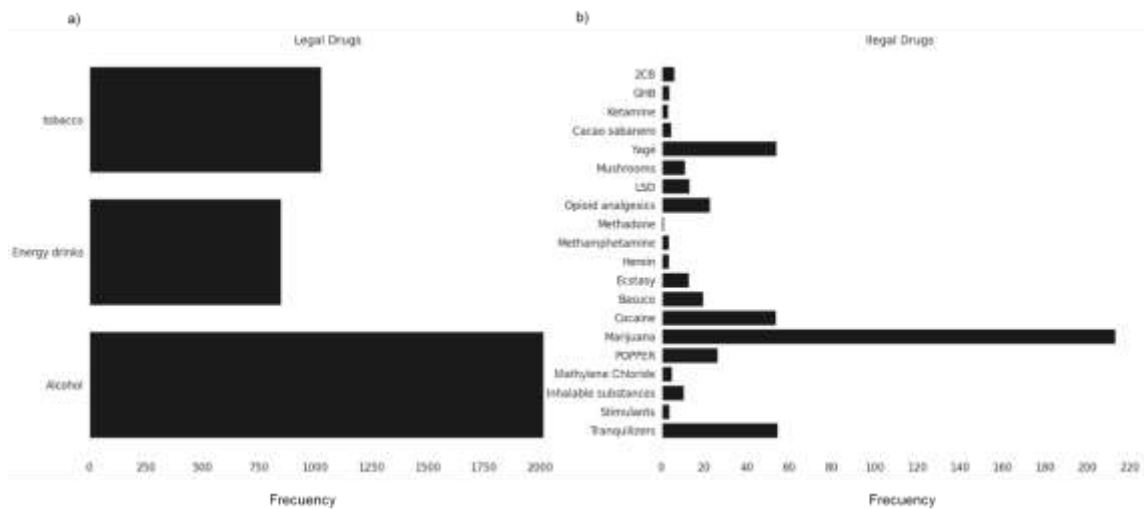


Figure 16. Frequency of consumption of (a) legal and (b) illegal drugs in Colombia.

In terms of legal drug consumption, alcohol emerges as the most widely consumed substance in certain regions of Colombia. Notably, the highest consumption rates are observed in Bogotá, Cundinamarca, and Chocó (Figure 17). Additionally, alcohol usage is moderately high in other departments, such as Vaupés, Nariño, Bolívar, Magdalena, La Guajira, and Atlántico. Energy drinks, another type of legal drug, show the highest consumption levels in Casanare and Guaviare. Moreover, these drinks have a slightly elevated usage rate in Boyacá, Nariño, Risaralda, and Arauca. This indicates that energy drinks are popular in these regions, though not to the same extent as alcohol. Regarding tobacco consumption, Cundinamarca has the highest usage rates in the country. However, moderately high consumption is also observed in various other departments, such as Bogotá, Boyacá, Nariño, Casanare, Tolima, Quindío, Risaralda, Guainía, Caldas, and Vaupés. While these specific regions show higher consumption rates for alcohol, energy drinks, and tobacco, it is important to note that usage of these legal drugs is not limited to these areas alone. In fact, the consumption of these substances can be found throughout the entire country, albeit with lower incidence rates (Figure 17). This widespread usage highlights the overall prevalence of legal drug consumption across Colombia.

Concerning illegal drugs, non-prescription tranquilizers and stimulants are most prevalent in Casanare (Figure 18). However, the consumption of tranquilizers is slightly higher in Nariño, while inhalants have the highest consumption in Quindío, followed by Cauca, Caldas, and Nariño. Methylene Chloride has the highest consumption in Cauca and a high consumption in Quindío and Nariño; Antioquia, followed by Caldas and Risaralda, shows the highest consumption of popper. On the contrary, marijuana has its highest

consumption in Risaralda and moderately high consumption in Caldas, Bogotá, Antioquia, and Quindío. As for cocaine, its consumption is the highest in Risaralda and moderately high in Antioquia (Figure 18).

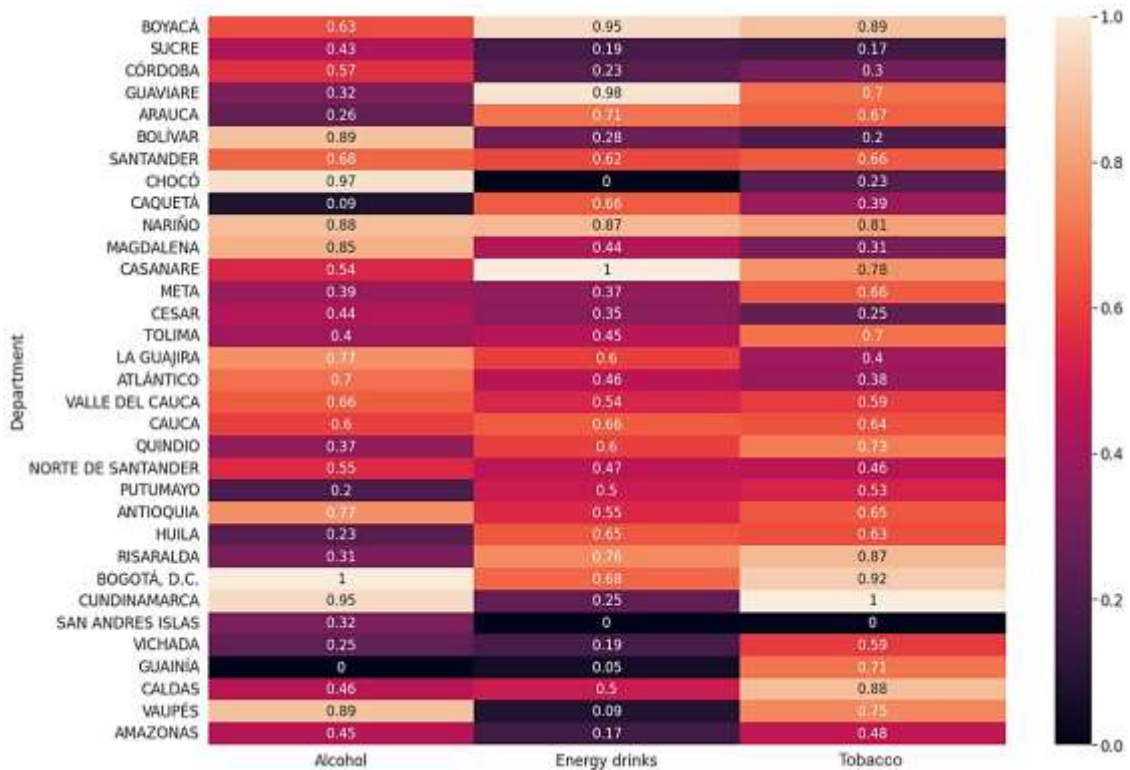


Figure 17. Consumption of illegal drugs by department.

Note: For interpretation purposes, number represents values scaled on a range of 0 to 1. For instance, Bogotá D.C. has the highest LSD consumption and Putumayo has the lowest.

On the other hand, basuco (i.e., cocaine paste) has the highest consumption rate in Guaviare, and critical consumption in Nariño, Cauca, Quindío, Antioquia, and Amazonas; ecstasy has its highest consumption in Risaralda, followed by Bogotá and Caldas; heroin consumption is highest in Vaupes, Huila, Cauca, Quindío, and Arauca, and is slightly higher in Casanare; methamphetamine consumption is highest in Casanare and is moderately high in Boyacá; methadone is most widely used in Quindío, but has slightly high levels of use in Valle del Cauca and Caquetá; opioids are most prevalent in Casanare, followed by Sucre; LSD is most prevalent in Bogotá, but has high levels of use in Caldas, Risaralda, Quindío, and Nariño; mushrooms have their highest consumption in Boyacá and have moderately high uses in Quindío, Risaralda, Bogotá, Cauca, and Casanare; Yagé has a higher incidence in Putumayo; cacao sabanero has its highest consumption in Caldas, and has moderate consumption in Cundinamarca, Bogotá, Antioquia, and Quindío; ketamine has the highest consumption in Casanare, followed by Antioquia; and

GHB has the highest consumption in Risaralda, followed by Santander, Valle del Cauca, and Norte de Santander. Finally, 2CB has the highest consumption rate in Risaralda, followed by Caldas. Although the consumption pattern of some departments is not mentioned, there is low and moderate consumption for certain drugs in some of them (Figure 18).

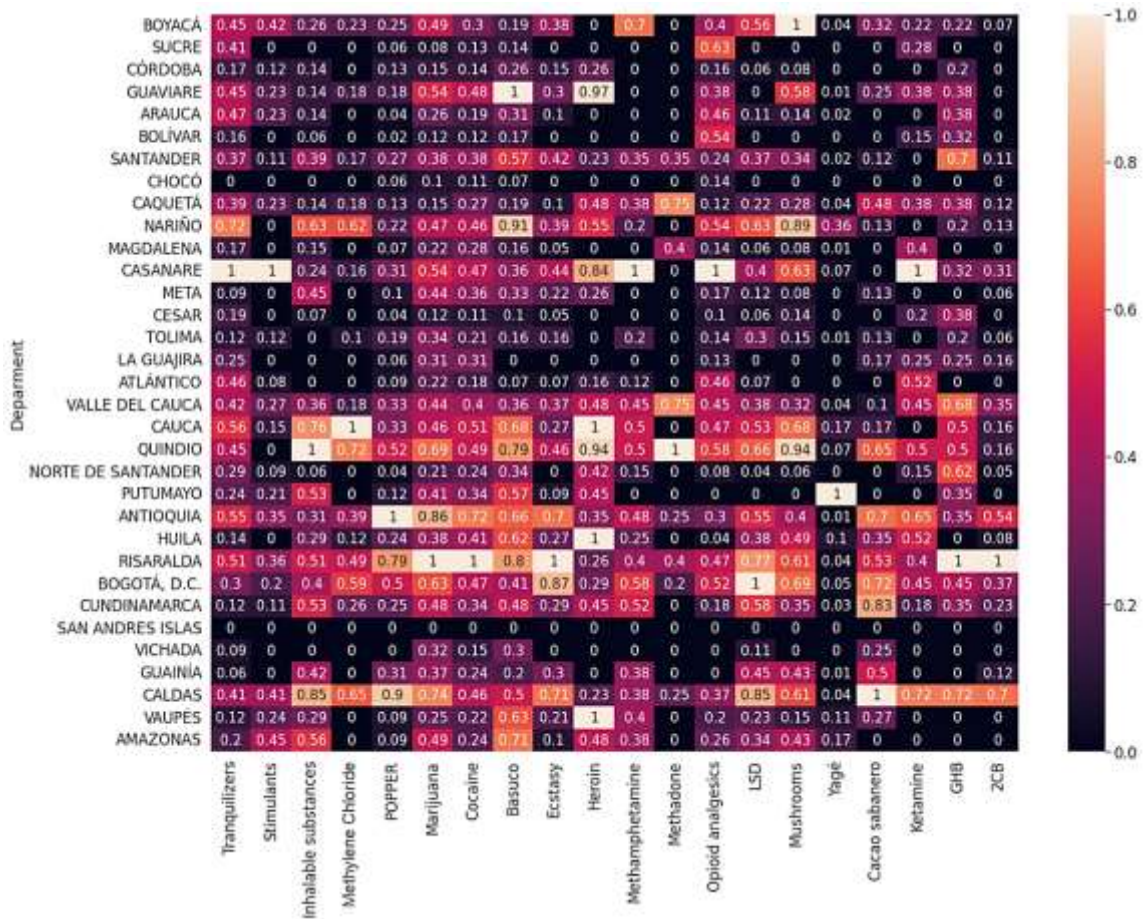


Figure 18. Consumption of legal drugs by department.

Note: Number represents values scaled on a range of 0 to 1 for psychoactive substance use. Conventions as in Fig 17.

3.3.4. Regionalization of Clusters

We applied a regionalization method as a grouping technique for imposing a spatial restriction, i.e., the result of a regionalization algorithm contains clusters with geographically coherent areas and coherent data profiles. Our approach uses a spatially constrained hierarchical clustering algorithm, which identified three clusters representing the consumption of PAS in the country (Figure 19). The number of clusters was estimated based on the average silhouette indexes, the total intra-cluster variance, and dendrograms (Figure 20). Following our results, cluster 0 is comprised of departments such as La

Guajira, Cesar, Atlántico, Magdalena, Norte de Santander, Bolivar, Sucre, and Cordoba, all of them located in the Northern region of the country; cluster 1 is comprised of Antioquia, Santander, Boyacá, Caldas, Risaralda, and Quindío; and cluster 2 is integrated by the remaining departments (Figure 19). When testing geographical coherence, which is the measure that assesses the “compactness” of a given shape, our results indicate that the clusters derived using the regionalization model represent moderately compact regions. In addition, the feature coherence test using different metrics showed that our 3-cluster regionalization structure properly fits the data (Table 7).



Figure 19. Cluster map of drug use after regionalization

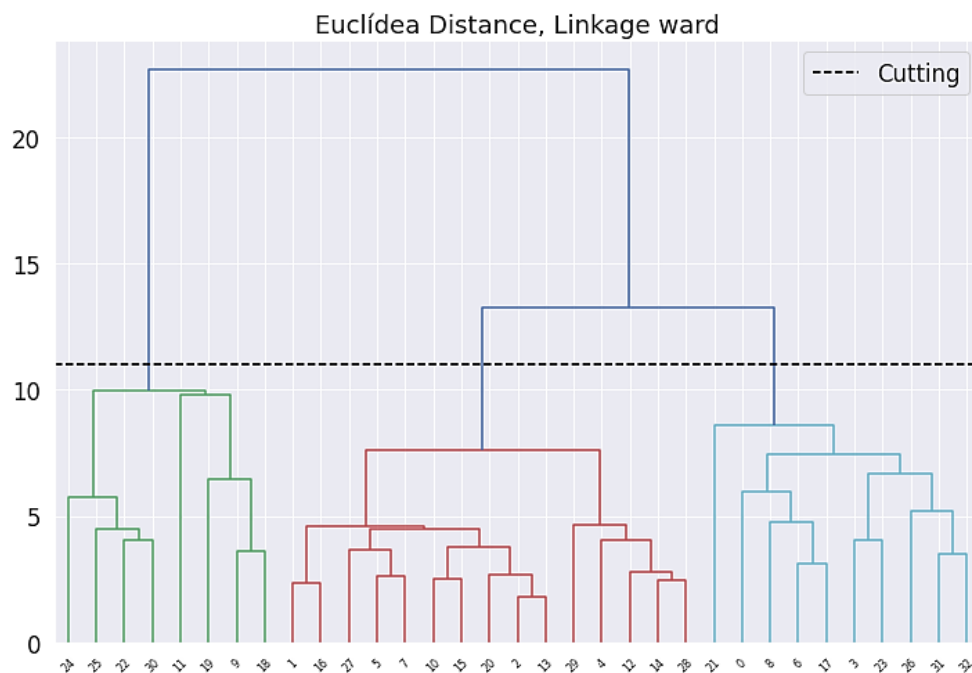


Figure 20. Definition of the optimal number of clusters based on a Dendrogram.

Table 7. Feature coherence measurements (goodness of fit).

	Cluster	KNN	AHC	RQ	RKNN
Geographical coherence (cluster)	0	0.035	0.033	0.058	0.078
	1	0.030	0.040	0.085	0.178
	2	0.178	0.178	0.145	0.173
CH Score		18.424	18.445	8.536	10.240

CH: Calinski and Harabasz score; KNN: k -means clustering; AHC: Hierarchical Clustering; RQ: Regionalization with QEEN constraint; RKNN: Regionalization with KNN constraint.

3.5. Discussion and Conclusion

In this study, we propose and test a Deep Neural Network-based Clustering-oriented Embedding algorithm (i.e., a ML-based model) for identifying psychoactive substance (PAS) use and abuse patterns in Colombia. This model allows the automatic extraction of features from the input data (such as sex, age, socioeconomic status, and housing type) to determine whether an individual has consumed PAS. It then creates clusters in the new data space generated during the learning process, following the methods outlined in [57, 59]. After the training process, a latent feature space (LFS) is generated, and the results are subsequently analyzed.

We have identified clearly marked clusters where the prevalence of individuals who use or do not use PAS is notable. Additionally, we found that region, sex, housing type, socioeconomic strata, age, and whether individuals contribute to household finances have a statistically significant impact on the clustering structure. These findings are consistent with previous studies aimed at identifying PAS consumption patterns [19, 80, 81]. Interestingly, when comparing the CAE-DEC model proposed in this study and the CAE-Spectral model using different metrics (i.e., Silhouette statistic, which measures the internal density of each cluster and the distance that separates them from each other, the Calinski-Harabasz index and the Davies-Bouldin index [DBI]), we found that our model performs better (Silhouette: 0.62 vs. 0.786; Calinski-Harabasz: 22468.26 vs. 775992.45; DBI: 0.2898 vs. 0.63; Table 6).

Based on our findings, individuals more likely to consume PAS are grouped in cluster 2, while cluster 1 consisted of individuals who did not consume PAS (Table 12). Not

surprisingly, a significant proportion of females characterizes cluster 1. In addition, most individuals belong to socioeconomic strata 1, are 40 years old or older, and do not contribute economically to support their household. In contrast, cluster 2 is characterized by a higher proportion of males aged between 20 and 40 in socioeconomic strata 1 and 2, who do not contribute to the household finances (Table 12). Finally, cluster 0 is characterized by a small proportion of males, a higher proportion of individuals in strata 3, 4, 5, and 6, and individuals are more likely to contribute to the household economy (Table 12). At the level of spatial statistics, we identified that legal drugs such as alcohol have a high prevalence in all regions of Colombia, with a slight tendency to more consumption in coastal areas (Figure 17). In our country, the coastal areas are often popular tourist destinations, and many tourists come to these areas looking for a relaxing experience, which can increase alcohol consumption. Coastal areas typically have warmer temperatures and more sunshine, increasing thirst and making people more likely to consume alcohol. Additionally, bars, clubs, and restaurants serve alcoholic beverage due to the high demand from tourists and locals [82, 83]. Another characteristic of this area is the fishing and maritime culture.

This culture is often associated with hard work and long working hours, and alcohol may be seen as a way to relax and unwind after a tough day at the sea [84]. Finally, this region has 68% urban and 32% rural zones [60]. The level of development, as measured by gross domestic product (GDP), is the third region with significant economic development in the country [85] (Table 8). Interestingly, the consumption of illegal drugs is lower in the Northern region than in other regions of the country. However, there is a more representative consumption of non-prescription tranquilizers, opioids, ketamine, GHB, and heroin. In particular, the Atlántico department has the highest consumption proportion within this region (Figure 18). Tobacco consumption is present in all regions, with a higher proportion in the Central region, where climate conditions resemble temperate weather. Also, this region has a diverse consumption pattern, where drugs such as marijuana, popper, cocaine, ecstasy, inhalants, methadone, heroin, LSD, GHB, 2CB, and mushrooms prevail. This region has Colombia's largest cities (i.e., Bogotá and Medellín); Bogotá has the highest population density and is a hub for drug trafficking routes, while Medellín has an unfortunate history of drug cartels and gang violence. Ultimately, this region is comprised of 77% urban areas, and the most developed cities in the country are located there [60, 85] (Table 8).

Table 8. Level of development, urbanity, rurality, and drug production.

	Department	Urban	Rural	% Urban	% Rural	GDP	Drug Production
Northern region (coastal areas)	Atlántico	2404831	130686	94,8%	5,2%	52311	-
	Bolívar	1549063	521047	74,8%	25,2%	41698	-
	Cesar	903411	297163	75,2%	24,8%	23037	38
	Córdoba	937319	847464	52,5%	47,5%	20570	2882
	La Guajira	410636	469924	46,6%	53,4%	14466	-
	Magdalena	938320	403426	69,9%	30,1%	15996	7
	San Andrés	44893	16387	73,3%	26,7%	1761	-
	Sucre	569089	335774	62,9%	37,1%	9659	-
				68,8%*	31,2%*	179498**	2927**
Central region	Antioquia	4972941	143416	77,6%	22,4%	176451	-
	Caldas	740865	257390	74,2%	25,8%	19745	-
	Caquetá	258280	143569	64,3%	35,7%	4670	4511,46
	Huila	669697	430689	60,9%	39,1%	19473	-
	Quindío	471910	67994	87,4%	12,6%	9733	-
	Risaralda	736164	207237	78,0%	22,0%	19263	-
	Tolima	907506	422681	68,2%	31,8%	25221	-
	Bogotá, D.C.	7387400	25166	99,7%	0,3%	298268	-
				76,3%*	23,7%*	572824**	4511**
Western region	Cauca	545902	918586	37,3%	62,7%	21107	17356
	Chocó	243194	291632	45,5%	54,5%	4856	1248
	Nariño	716592	914000	43,9%	56,1%	17971	36964
	Valle del Cauca	3809542	666344	85,1%	14,9%	114864	2329
				53,0%*	47,0%*	158798**	57897**
Eastern region	Boyacá	708006	509370	58,2%	41,8%	30803	-
	Cundinamarca	2090845	828215	71,6%	28,4%	72481	-
	Meta	795061	244661	76,5%	23,5%	39459	1466
	Norte de Santander	1173712	317977	78,7%	21,3%	18360	41711
	Santander	1655627	529210	75,8%	24,2%	74025	2
				72,1%*	27,9%*	235128**	43179**
Southern region	Amazonas	37047	39542	48,4%	51,6%	880	125
	Arauca	172634	89540	65,8%	34,2%	6310	-
	Casanare	295434	125070	70,3%	29,7%	17236	-
	Guainía	20279	27835	42,1%	57,9%	431	22
	Guaviare	45991	36776	55,6%	44,4%	931	3119
	Putumayo	174539	173643	50,1%	49,9%	4091	24973
	Vaupés	12090	28707	29,6%	70,4%	323	26
	Vichada	25833	81975	24,0%	76,0%	775	245
				48,2%*	51,8%*	30977**	28385**

Note: * Average, ** Sum. The urban and rural areas represent the total population in each department. The production drug is measured in hectares. All records are for 2019.

Energy drinks are more frequently used in the Eastern region, characterized by a continental climate surrounded by flat territory. Our results are in line with the scientific literature suggesting that the location of regions within countries is directly associated with the consumption of PAS [26, [86–88]. The consumption of heroin, basuco, non-prescription tranquilizers, stimulants, methamphetamines, opioids, and ketamine characterizes this region. This zone is the second most developed region in the country, and 72% of urban areas [60], [85](Table 8).

Our findings also show that the Southern region is more likely to consume illegal drugs, including basuco, heroin, and Yagé (Figure 18). One of the main reasons for this result is that, unfortunately, this region has favorable environmental characteristics (i.e., majority rainforest) for their consumption and production, being the second largest illegal drug-producing region in Colombia [79]. Furthermore, this region has the highest percentage of rurality (52%) compared to the other regions, and its level of development is low as measured by the GDP [60, 85] (Table 8). In the Western region, also known as the Pacific region, consumption mostly mainly includes Methylene Chloride, GHB, heroin, opioids, and methamphetamines. This region (Pacific) is mainly characterized known for its geographical isolation, poverty, and ongoing conflict, which have contributed to the growth of drug production and trafficking in the area. Poverty is one of the main factors driving drug production in the Pacific region, which has led many people to turn to drug cultivation and trafficking for survival. Additionally, the region's rugged terrain and limited infrastructure have made it difficult for the Colombian government to establish a strong presence, allowing drug traffickers to operate with relative impunity [89]. This region has a similar percentage of urban (53%) and rural (47%) populations than the Southern region and ranks second among the regions with the lowest levels of development (Table 8).

In the Western region, also known as the Pacific region, consumption mainly includes methylene chloride, GHB, heroin, opioids, and methamphetamines. This region is mainly characterized for its geographical isolation, poverty, and ongoing conflict, which have contributed to the growth of drug production and trafficking in the area. Poverty is one of the main factors driving drug production in the Pacific region, which has led many people to turn to drug cultivation and trafficking for survival. Additionally, the region's rugged

terrain and limited infrastructure have made it difficult for the Colombian government to establish a strong presence, allowing drug traffickers to operate with relative impunity [89]. This region has a similar percentage of urban (53%) and rural (47%) populations than the Southern region, and ranks second among the regions with the lowest levels of development (Table 8).[89]. This region has a similar percentage of urban (53%) and rural (47%) populations than the Southern region. On the other hand, this region ranks second among the regions with the lowest levels of development (Table 8).

In summary, the proposed CAE-DEC model simultaneously integrates a feature extraction process within the clustering design, prioritizing features that improve the separation between groups, thus avoiding the manual extraction of features, which is a frequent process in traditional models. Additionally, a geospatial component is sequentially included to expand the resulting insights by considering geographic constraints. Currently, these types of architectures are scarce in understanding mental health problems. As part of future work, the architecture of the proposed model could be improved to integrate the automatic extraction of features while optimizing a geospatial loss. Following our experience with the proposed CAE-DEC in PAS consumption, the application of this model to other mental health problems, such as suicide, depression, and domestic violence, among other pathologies, could be explored. Based on these results, effective interventions and/or government policies to prevent and/or mitigate their impact could be promoted and evaluated, for example, by developing regional interventions based on the types of drugs most prevalent in the area and the cultural and socio-economic characteristics. This can include education, treatment, and harm reduction programs. Also, this information can be used to develop public health campaigns to raise awareness about the risks of drug use and reduce their negative impact. Furthermore, this information can be used to crack down on drug trafficking and distribution networks. On the other hand, this information can be used to alert healthcare providers and regulatory bodies to take appropriate action to prevent their use and discover new drugs.

References

- [1] OPS, “Abuso de sustancias ,” *Organización Panamericana de la Salud*, 2022. <https://www.paho.org/es/temas/abuso-sustancias> (accessed Jan. 25, 2022).
- [2] C. Heesun, L. Jaesin, and K. Eunmi, “Trends of novel psychoactive substances (NPSs) and their fatal cases,” *Forensic Toxicol.*, vol. 34, no. 1, pp. 1–11, 2016, Accessed: Jan. 20, 2022. [Online]. Available: https://jglobal.jst.go.jp/en/detail?JGLOBAL_ID=201602283742633549
- [3] A. L. Riley *et al.*, “Abuse potential and toxicity of the synthetic cathinones (i.e., ‘Bath salts’),” *Neurosci. Biobehav. Rev.*, vol. 110, pp. 150–173, Mar. 2020, doi: 10.1016/J.NEUBIOREV.2018.07.015.
- [4] S. Assi, N. Gulyamova, P. Kneller, and D. Osselton, “The effects and toxicity of cathinones from the users’ perspectives: A qualitative study,” *Hum. Psychopharmacol. Clin. Exp.*, vol. 32, no. 3, p. e2610, May 2017, doi: 10.1002/HUP.2610.
- [5] V. Lukić, R. Micić, B. Arsić, B. Nedović, and Ž. Radosavljević, “Overview of the major classes of new psychoactive substances, psychoactive effects, analytical determination and conformational analysis of selected illegal drugs,” *Open Chem.*, vol. 19, no. 1, pp. 60–106, Jan. 2021, doi: 10.1515/CHEM-2021-0196/XML.
- [6] N. Uchiyama, S. Matsuda, M. Kawamura, R. Kikura-Hanajiri, and Y. Goda, “Two new-type cannabimimetic quinolinyl carboxylates, QUPIC and QUCHIC, two new cannabimimetic carboxamide derivatives, ADB-FUBINACA and ADBICA, and 5 synthetic cannabinoids detected with a thiophene derivative α -PVT and an opioid receptor agonist AH-7921 identified in illegal products,” *Forensic Toxicol. 2013 312*, vol. 31, no. 2, pp. 223–240, Mar. 2013, doi: 10.1007/S11419-013-0182-9.
- [7] B. F. Grant *et al.*, “Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders - Results from the national epidemiologic survey on alcohol and related conditions,” *Arch. Gen. Psychiatry*, vol. 61, no. 8, pp. 807–816, 2004, doi: 10.1001/archpsyc.61.8.807.
- [8] CDC, “Understanding the Epidemic,” 2020. <https://www.cdc.gov/opioids/basics/epidemic.html> (accessed Jan. 21, 2022).
- [9] R. Z. Goetzel, K. Hawkins, R. J. Ozminkowski, and S. H. Wang, “The health and productivity cost burden of the ‘top 10’ physical and mental health conditions

- affecting six large US employers in 1999,” *J. Occup. Environ. Med.*, vol. 45, no. 1, pp. 5–14, 2003, doi: 10.1097/00043764-200301000-00007.
- [10] W. F. Stewart, J. A. Ricci, E. Chee, S. R. Hahn, and D. Morganstein, “Cost of lost productive work time among US workers with depression,” *JAMA-JOURNAL Am. Med. Assoc.*, vol. 289, no. 23, pp. 3135–3144, Jun. 2003, doi: 10.1001/jama.289.23.3135.
- [11] F. L. G. Garcia and J. C. A. Murillo, “The United Nations and 21st century security challenges in Colombia,” *Rev. Cient. Gen. Jose Maria Cordova*, vol. 19, no. 36, pp. 929–940, 2021, doi: 10.21830/19006586.875.
- [12] J. P. Aschner and J. C. Montero, “Architectures, spaces, and territories of illicit drug trafficking in Colombia and Mexico:,” <https://doi.org/10.1177/1741659020910212>, vol. 17, no. 3, pp. 327–351, Mar. 2020, doi: 10.1177/1741659020910212.
- [13] ODC, “Observatorio de drogas de Colombia,” 2022. <https://www.minjusticia.gov.co/programas-co/ODC/Paginas/SIDCO-departamento-municipio.aspx> (accessed Jun. 09, 2022).
- [14] DANE, “Encuesta Nacional de Consumo de Sustancias Psicoactivas,” 2020. Accessed: Apr. 23, 2021. [Online]. Available: <https://www.dane.gov.co/files/investigaciones/boletines/encspa/comunicado-encspa-2019.pdf>
- [15] DANE, “Estudio nacional de consumo de sustancias psicoactivas en Colombia,” Bogotá, 2014. Accessed: Jan. 17, 2022. [Online]. Available: https://www.unodc.org/documents/colombia/2014/Julio/Estudio_de_Consumo_UNODC.pdf
- [16] UNODC, “Drogas sintéticas y nuevas sustancias psicoactivas en América Latina y el Caribe 2021,” Viena, 2021. Accessed: Jan. 21, 2022. [Online]. Available: [https://www.minjusticia.gov.co/programas-co/ODC/Documents/Publicaciones/GlobalSmartLA\(1\).pdf?csf=1&e=MH9EHg](https://www.minjusticia.gov.co/programas-co/ODC/Documents/Publicaciones/GlobalSmartLA(1).pdf?csf=1&e=MH9EHg)
- [17] P. Griffiths and R. Mcketin, “Developing a global perspective on drug consumption patterns and trends-the challenge for drug epidemiology,” *Bull. Narcotics*, vol. 5, no. 1, 2003.
- [18] W. A. Lanier, E. M. Johnson, R. T. Rolfs, M. D. Friedrichs, and T. C. Grey, “Risk factors for prescription opioid-related death, Utah, 2008-2009,” *Pain Med.*, vol. 13, no. 12, pp. 1580–1589, 2012, doi: 10.1111/J.1526-4637.2012.01518.X.
- [19] S. S. Martins, L. Sampson, M. Cerdá, and S. Galea, “Worldwide Prevalence and

- Trends in Unintentional Drug Overdose: A Systematic Review of the Literature,” *Am. J. Public Health*, vol. 105, no. 11, pp. e29–e49, Nov. 2015, doi: 10.2105/AJPH.2015.302843.
- [20] CDC, “Today’s Heroin Epidemic,” *Centers for Disease Control and Prevention*, 2015. <https://www.cdc.gov/vitalsigns/heroin/index.html> (accessed Jan. 16, 2022).
- [21] C. M. Fuller *et al.*, “Effects of race, neighborhood, and social network on age at initiation of injection drug use,” *Am. J. Public Health*, vol. 95, no. 4, pp. 689–695, Apr. 2005, doi: 10.2105/AJPH.2003.02178.
- [22] P. J. Fite, P. Wynn, J. E. Lochman, and K. C. Wells, “The Influence of Neighborhood Disadvantage and Perceived Disapproval on Early Substance Use Initiation,” *Addict. Behav.*, vol. 34, no. 9, p. 769, Sep. 2009, doi: 10.1016/J.ADDBEH.2009.05.002.
- [23] S. R. Friedman *et al.*, “Income inequality, drug-related arrests, and the health of people who inject drugs: Reflections on seventeen years of research,” *Int. J. Drug Policy*, vol. 32, pp. 11–16, Jun. 2016, doi: 10.1016/J.DRUGPO.2016.03.003.
- [24] M. Jensen, L. Chassin, and N. A. Gonzales, “Neighborhood Moderation of Sensation Seeking Effects on Adolescent Substance Use Initiation,” *J. Youth Adolesc.*, vol. 46, no. 9, p. 1953, Sep. 2017, doi: 10.1007/S10964-017-0647-Y.
- [25] C. Sarah and J. Leonard A, “Contextual Perspectives on Heroin Addiction and Recovery: Classic and Contemporary Theories,” *Int. Arch. Public Heal. Community Med.*, vol. 2, no. 1, Dec. 2018, doi: 10.23937/IAPHCM-2017/1710009.
- [26] P. Bozorgi, D. E. Porter, J. M. Eberth, J. P. Eidson, and A. Karami, “The leading neighborhood-level predictors of drug overdose: A mixed machine learning and spatial approach,” *Drug Alcohol Depend.*, vol. 229, p. 109143, Dec. 2021, doi: 10.1016/J.DRUGALCDEP.2021.109143.
- [27] S. Galea, S. Rudenstine, and D. Vlahov, “Drug use, misuse, and the urban environment,” *Drug Alcohol Rev.*, vol. 24, no. 2, pp. 127–136, Mar. 2005, doi: 10.1080/09595230500102509.
- [28] C. A. Latkin, V. Forman, A. Knowlton, and S. Sherman, “Norms, social networks, and HIV-related risk behaviors among urban disadvantaged drug users,” *Soc. Sci. Med.*, vol. 56, no. 3, pp. 465–476, Feb. 2003, doi: 10.1016/S0277-9536(02)00047-3.
- [29] J. R. Schroeder, C. A. Latkin, D. R. Hoover, A. D. Curry, A. R. Knowlton, and D.

- D. Celentano, “Illicit drug use in one’s social network and in one’s neighborhood predicts individual heroin and cocaine use,” *Ann. Epidemiol.*, vol. 11, no. 6, pp. 389–394, 2001, doi: 10.1016/S1047-2797(01)00225-3.
- [30] A. Campo-Arias, Y. P. Suárez-Colorado, and C. C. Caballero- Domínguez, “Asociación entre el consumo de Cannabis y el riesgo de suicidio en adolescentes escolarizados de Santa Marta, Colombia,” *Biomédica*, vol. 40, no. 3, p. 569, 2020, doi: 10.7705/BIOMEDICA.4988.
- [31] A. L. Fajardo, “Consumption of psychopharmaceuticals in the city of Bogota (Colombia): a new reality,” *Arch. Med.*, vol. 18, no. 2, 2018, doi: 10.30554/archmed.18.2.2743.2018.
- [32] O. Scoppetta and G. A. Castaño, “Early drug consumption and subsequent risk of illicit drug use in Colombia,” *Addict. Disord. their Treat.*, vol. 18, no. 1, pp. 10–14, Mar. 2019, doi: 10.1097/ADT.000000000000144.
- [33] C. Scheuer *et al.*, “El consumo de sustancias psicoactivas en jóvenes estudiantes de una institución educativa del municipio de Neira (Caldas): un estudio de caso desde la mirada de la educación inclusiva,” *Cult. y Drog.*, vol. 23, no. 26, pp. 343–354, Jul. 2018, doi: 10.2/JQUERY.MIN.JS.
- [34] J. Kalyanam, T. Katsuki, G. R.G. Lanckriet, and T. K. Mackey, “Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning,” *Addict. Behav.*, vol. 65, pp. 289–295, Feb. 2017, doi: 10.1016/j.addbeh.2016.08.019.
- [35] M. A. Narvaez-Chicaiza, “Harm Reduction Policies Where Drugs Constitute a Security Issue,” *Heal. Care Anal.*, vol. 28, no. 4, pp. 382–390, Dec. 2020, doi: 10.1007/S10728-020-00415-9.
- [36] S. M. Restrepo-Escobar and E. A. S. Cardona, “Educational and prevention campaigns. A review on the use of psychoactive substances in Colombian university students,” *Interdisciplinaria*, vol. 38, no. 2, pp. 199–208, 2021, doi: 10.16888/INTERD.2021.38.2.13.
- [37] E. Hancer, B. Xue, and M. Zhang, “A survey on feature selection approaches for clustering,” *Artif. Intell. Rev. 2020 536*, vol. 53, no. 6, pp. 4519–4545, Jan. 2020, doi: 10.1007/S10462-019-09800-W.
- [38] T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross, “An fMRI-Based Neurologic Signature of Physical Pain,” *N. Engl. J. Med.*, vol. 368, no. 15, pp. 1388–1397, 2013, doi: 10.1056/NEJMoa1204471.

- [39] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, "Identifying adverse drug event information in clinical notes with distributional semantic representations of context," *J. Biomed. Inform.*, vol. 57, pp. 333–349, Oct. 2015, doi: 10.1016/j.jbi.2015.08.013.
- [40] L. M. Squeglia *et al.*, "Neural Predictors of Initiating Alcohol Use During Adolescence," *Am. J. Psychiatry*, vol. 174, no. 2, pp. 172–185, Feb. 2017, doi: 10.1176/appi.ajp.2016.15121587.
- [41] M. Conway and D. O'Connor, "Social media, big data, and mental health: current advances and ethical implications," *Curr. Opin. Psychol.*, vol. 9, pp. 77–82, Jun. 2016, doi: 10.1016/j.copsyc.2016.01.004.
- [42] T. Katsuki, T. K. Mackey, and R. Cuomo, "Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data," *J. Med. INTERNET Res.*, vol. 17, no. 12, 2015, doi: 10.2196/jmir.5144.
- [43] L. Degenhardt *et al.*, "The global epidemiology and burden of psychostimulant dependence: Findings from the Global Burden of Disease Study 2010," *Drug Alcohol Depend.*, vol. 137, pp. 36–47, 2014, doi: 10.1016/j.drugalcdep.2013.12.025.
- [44] H. A. Whiteford *et al.*, "Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010," *Lancet*, vol. 382, no. 9904, pp. 1575–1586, Nov. 2013, doi: 10.1016/S0140-6736(13)61611-6.
- [45] F. D. Bowman, B. Caffo, S. S. Bassett, and C. Kilts, "A Bayesian hierarchical framework for spatial modeling of fMRI data," *Neuroimage*, vol. 39, no. 1, pp. 146–156, 2008, doi: 10.1016/j.neuroimage.2007.08.012.
- [46] K. Shannon, M. Rusch, J. Shoveller, D. Alexson, K. Gibson, and M. W. Tyndall, "Mapping violence and policing as an environmental-structural barrier to health service and syringe availability among substance-using women in street-level sex work," *Int. J. DRUG POLICY*, vol. 19, no. 2, pp. 140–147, 2008, doi: 10.1016/j.drugpo.2007.11.024.
- [47] B. Freisthler, B. Needell, and P. J. Gruenewald, "Is the physical availability of alcohol and illicit drugs related to neighborhood rates of child maltreatment?," *Child Abuse Negl.*, vol. 29, no. 9, pp. 1049–1060, Sep. 2005, doi: 10.1016/j.chiabu.2004.12.014.
- [48] J. K. Bass and S. F. Lambert, "Urban adolescents' perceptions of their

- neighborhoods: An examination of spatial dependence,” *J. Community Psychol.*, vol. 32, no. 3, pp. 277–293, May 2004, doi: 10.1002/jcop.20005.
- [49] B. Chaix *et al.*, “Spatial clustering of mental disorders and associated characteristics of the neighbourhood context in Malmo, Sweden, in 2001,” *J. Epidemiol. Community Health*, vol. 60, no. 5, pp. 427–435, May 2006, doi: 10.1136/jech.2005.040360.
- [50] C. T. Mowbray, M. C. Holter, G. B. Teague, and D. Bybee, “Fidelity criteria: Development, measurement, and validation,” *Am. J. Eval.*, vol. 24, no. 3, pp. 315–340, 2003, doi: 10.1177/109821400302400303.
- [51] M. Peet and C. Stokes, “Omega-3 fatty acids in the treatment of psychiatric disorders,” *Drugs*, vol. 65, no. 8, pp. 1051–1059, 2005, doi: 10.2165/00003495-200565080-00002.
- [52] K. Chichester *et al.*, “Pharmacies and features of the built environment associated with opioid overdose: A geospatial comparison of rural and urban regions in Alabama, USA,” *Int. J. Drug Policy*, vol. 79, May 2020, doi: 10.1016/J.DRUGPO.2020.102736.
- [53] P. Geissert *et al.*, “High-risk prescribing and opioid overdose: prospects for prescription drug monitoring program-based proactive alerts,” *Pain*, vol. 159, no. 1, pp. 150–156, Jan. 2018, doi: 10.1097/J.PAIN.0000000000001078.
- [54] C. Fraley and A. E. Raftery, “How many clusters? Which clustering method? Answers via model-based cluster analysis,” *Comput. J.*, vol. 41, no. 8, pp. 586–588, 1998, doi: 10.1093/COMJNL/41.8.578.
- [55] A. Saxena *et al.*, “A review of clustering techniques and developments,” *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017, doi: 10.1016/J.NEUCOM.2017.06.053.
- [56] B. Li, D. Pi, Y. Lin, and L. Cui, “DNC: A Deep Neural Network-based Clustering-oriented Network Embedding Algorithm,” *J. Netw. Comput. Appl.*, vol. 173, Jan. 2021, doi: 10.1016/J.JNCA.2020.102854.
- [57] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised Deep Embedding for Clustering Analysis,” *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 1, pp. 740–749, Nov. 2016, Accessed: Jan. 15, 2022. [Online]. Available: <https://arxiv.org/abs/1511.06335v2>
- [58] S. Sharifipour, H. Fayyazi, and M. Sabokro, “Unsupervised Feature Selection using Encoder-Decoder Networks,” *6th Iran. Conf. Signal Process. Intell. Syst.*

- ICSPIS 2020*, Dec. 2020, doi: 10.1109/ICSPIS51611.2020.9349608.
- [59] B. Li, D. Pi, Y. Lin, and L. Cui, “DNC: A Deep Neural Network-based Clustering-oriented Network Embedding Algorithm,” *J. Netw. Comput. Appl.*, vol. 173, no. May 2020, p. 102854, 2021, doi: 10.1016/j.jnca.2020.102854.
- [60] DANE, “Departamento Administrativo Nacional de Estadística. Censo Nacional de Población y Vivienda 2018. Proyecciones de Población 2018-2020, total municipal por área Junio 30.” Bogotá D.C, Colombia, 2018.
- [61] DNP, “Avances y complementariedades estratégicas de los Distritos en el marco de los esquemas asociativos territoriales,” Bogotá D.C, 2018. [Online]. Available: [https://colaboracion.dnp.gov.co/CDT/Desarrollo Territorial/Conversatorio Distrito Cali 04_10_2018 - Santiago Arroyo.pdf](https://colaboracion.dnp.gov.co/CDT/Desarrollo_Territorial/Conversatorio_Distrito_Cali_04_10_2018_-_Santiago_Arroyo.pdf)
- [62] UNODC, “Monitoreo de territorios afectados por cultivos ilícitos 2020,” Bogotá, 2021. Accessed: Jan. 14, 2022. [Online]. Available: https://www.unodc.org/documents/crop-monitoring/Colombia/Colombia_Monitoreo_de_territorios_afectados_por_cultivos_ilicitos_2020.pdf
- [63] ODC, “Estudio nacional de consumo de sustancias psicoactivas,” Bogotá, 2019. Accessed: Jan. 14, 2022. [Online]. Available: [https://www.odc.gov.co/Portals/1/publicaciones/pdf/estudio Nacional de consumo 2019.pdf](https://www.odc.gov.co/Portals/1/publicaciones/pdf/estudio_Nacional_de_consumo_2019.pdf)
- [64] DANE, “Encuesta Nacional de Consumo de Sustancias Psicoactivas en Población General 2019,” 2020. https://microdatos.dane.gov.co/index.php/catalog/680/get_microdata (accessed Jan. 14, 2022).
- [65] J. Espinosa, “Shapefile,” 2022. <https://hub.arcgis.com/datasets/de0e829ddb743c895ba6dcee1b74fae/about> (accessed Jun. 09, 2022).
- [66] L. Anselin, *Spatial Econometrics: Methods and Models*, Springer Netherlands. 1988. doi: 10.1007/978-94-015-7799-1_1.
- [67] P. Moran, *The Interpretation of Statistical Maps*, 2nd ed., vol. 10. Journal of the Royal Statistical Society, 1948. Accessed: Jan. 23, 2022. [Online]. Available: <https://www.jstor.org/stable/2983777>
- [68] R. C. Geary, “The Contiguity Ratio and Statistical Mapping,” *Inc. Stat.*, vol. 5, no. 3, p. 115, Nov. 1954, doi: 10.2307/2986645.

- [69] A. Getis and J. K. Ord, “The Analysis of Spatial Association by Use of Distance Statistics,” *Geogr. Anal.*, vol. 24, no. 3, pp. 189–206, Jul. 1992, doi: 10.1111/J.1538-4632.1992.TB00261.X.
- [70] L. Anselin, “Local Indicators of Spatial Association—LISA,” *Geogr. Anal.*, vol. 27, no. 2, pp. 93–115, Apr. 1995, doi: 10.1111/J.1538-4632.1995.TB00338.X.
- [71] J. C. Duque, R. Ramos, and J. Suriñach, “Supervised Regionalization Methods: A Survey:,” <http://dx.doi.org/10.1177/0160017607301605>, vol. 30, no. 3, pp. 195–220, Jul. 2016, doi: 10.1177/0160017607301605.
- [72] S. Rey, D. Arribas-Bel, and L. Wolf, *Geographic Data Science with Python*. 2020. Accessed: Jan. 23, 2022. [Online]. Available: <https://geographicdata.science/book/intro.html>
- [73] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
- [74] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, “Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6791 LNCS, no. PART 1, pp. 52–59, 2011, doi: 10.1007/978-3-642-21735-7_7.
- [75] A. Shrestha and A. Mahmood, “Review of deep learning algorithms and architectures,” *IEEE Access*, vol. 7, pp. 53040–53065, 2019, doi: 10.1109/ACCESS.2019.2912200.
- [76] T. Caliński and J. Harabasz, “A Dendrite Method For Cluster Analysis,” *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.
- [77] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.
- [78] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.
- [79] ODC, “Density of drug production in Colombia,” 2021. <https://www.datos.gov.co/d/acs4-3wgp/visualization> (accessed Jul. 05, 2022).
- [80] H. Clarke, N. Soneji, D. T. Ko, L. Yun, and D. N. Wijeyesundera, “Rates and risk factors for prolonged opioid use after major surgery: population based cohort study,” *BMJ*, vol. 348, Feb. 2014, doi: 10.1136/BMJ.G1251.
- [81] Y. F. Kuo, M. A. Raji, N. W. Chen, H. Hasan, and J. S. Goodwin, “Trends in

- Opioid Prescriptions Among Part D Medicare Recipients From 2007 to 2012,” *Am. J. Med.*, vol. 129, no. 2, pp. 221.e21-221.e30, Feb. 2016, doi: 10.1016/J.AMJMED.2015.10.002.
- [82] S. Puigcorbé *et al.*, “Assessing the association between tourism and the alcohol urban environment in Barcelona: a cross-sectional study,” *BMJ Open*, vol. 10, no. 9, p. e037569, Sep. 2020, doi: 10.1136/BMJOPEN-2020-037569.
- [83] M. Easwaran, J. Bazroy, V. Jayaseelan, and Z. Singh, “Prevalence and determinants of alcohol consumption among adult men in a coastal area of south India,” *Int. J. Med. Sci. Public Heal.*, vol. 4, no. 3, p. 360, 2015, doi: 10.5455/IJMSPH.2015.1010201479.
- [84] P. Chinnakali, P. Thekkur, A. Manoj Kumar, G. Ramaswamy, B. Bharadwaj, and G. Roy, “Alarming high level of alcohol use among fishermen: A community based survey from a coastal area of south India,” *J. Forensic Leg. Med.*, vol. 42, pp. 41–44, Aug. 2016, doi: 10.1016/J.JFLM.2016.05.006.
- [85] DANE, “Producto Interno Bruto por departamento,” 2021.
- [86] M. C. García *et al.*, “Opioid Prescribing Rates in Nonmetropolitan and Metropolitan Counties Among Primary Care Providers Using an Electronic Health Record System — United States, 2014–2017,” *MMWR. Morb. Mortal. Wkly. Rep.*, vol. 68, no. 2, pp. 25–30, Jan. 2019, doi: 10.15585/MMWR.MM6802A1.
- [87] K. M. Keyes, M. Cerdá, J. E. Brady, J. R. Havens, and S. Galea, “Understanding the Rural–Urban Differences in Nonmedical Prescription Opioid Use and Abuse in the United States,” *Am. J. Public Health*, vol. 104, no. 2, p. e52, Feb. 2014, doi: 10.2105/AJPH.2013.301709.
- [88] N. B. King, V. Fraser, C. Boikos, R. Richardson, and S. Harper, “Determinants of Increased Opioid-Related Mortality in the United States and Canada, 1990–2013: A Systematic Review,” *Am. J. Public Health*, vol. 104, no. 8, p. e32, 2014, doi: 10.2105/AJPH.2014.301966.
- [89] UNODC, “Persistencia de los cultivos de coca en la Región Pacífica,” 2010.

Chapter 4

4. Bi-Objective Location-Allocation Model of Interventions in High Drug Consumption Areas Incorporating Twitter Topic Modeling

4.1 Abstract

The Comprehensive Policy for the Prevention and Care of Psychoactive Substance Use in Colombia considers prevention and mitigation programs that aim to improve the care of individuals, families, and communities at risk or with problematic use of psychoactive substances. The effectiveness of these programs depends on the participation level and user accessibility to the facilities that provide the services. Factors influencing access include facility type, location, and the number of patients assigned to the facility. This paper proposes a location-allocation structure to optimize intervention policies under resource constraints to improve population health outcomes. The model is based on a bi-objective integer programming structure for the location and allocation of health centers and consumers. The objectives considered are I) to reduce the overall risk of drug consumption and II) to minimize the distance between patients and facilities, considering an equitable distribution of the facilities among citizens. Additionally, we identified the drug-related interests of the target population's consumers through social network analysis (Twitter), in an effort to design more effective interventions.

4.2 Introduction

During the last few years, substantial progress has been made in the research and development of methods for the efficient location of prevention and mitigation intervention for reducing drug consumption. This theoretical framework addresses the planning issues of intervention locations and service allocations, offering several relevant

implications for planners, hospital managers, and regulators. This model combines two key factors, geographically equitable consumption risk and topical relevance to drug addiction, to optimize the allocation of drug addiction treatment interventions in high-drug consumption areas.

Drug addiction is a complex and chronic condition that requires comprehensive treatment approaches to address its physiological, psychological, and social aspects. Over the years, various interventions have been developed and implemented to effectively manage drug addiction and support individuals in their journey toward recovery. The case of Pharmacological Interventions involve the use of medications to manage drug addiction [1]; Behavioral Therapies which focus on identifying and modifying maladaptive thoughts and behaviors related to drug use [2]; Residential Treatment Programs which provide a structured and supportive environment for individuals to receive intensive treatment for drug addiction [3]; Peer Support and Mutual Aid Group which offer a supportive community, opportunities for sharing experiences, and guidance in maintaining abstinence [4]; and Integrated Approaches combine pharmacological, behavioral, and psychosocial interventions to address the complex needs of individuals with drug addiction.

The availability and accessibility of appropriate treatment locations play a critical role in addressing the challenges associated with drug addiction. Accessible treatment locations ensure that individuals with substance use disorders can receive timely and effective interventions. This proposed model has the potential to significantly improve the effectiveness of drug addiction treatment interventions, particularly in areas where the drug epidemic is most prevalent. By incorporating Twitter topic modeling, the model can identify the most relevant topics related to drug addiction in a given area and allocate interventions accordingly. This approach to drug addiction treatment interventions has been the subject of several research studies. For example, some studies seek to minimize the distances between consumers and care centers [5], [6]. Others, simultaneously determine the location and the size of the facilities as well as the allocation of patients to the Preventive healthcare facilities under congestion [7]–[9]. Additionally, some studies incorporate the minimization of distance between patients and facilities, and equitable distribution of the facilities among citizens [10], [11]. On the other hand, other studies focused on addressing the health care facility location to maximize the demand satisfaction by improving the center service quality [12], [13]. There are also several

articles where the location of the health center is planned, taking into account variables such as total hospital charges, discharge destination, and length of hospital stay, among others [14]–[16].

Furthermore, in the literature review highlights the importance of diverse drug addiction treatment locations in addressing the needs of individuals with substance use disorders. Outpatient treatment centers [17], [18], residential/inpatient facilities [19], community-based programs [20], telemedicine [21], and integration with primary care settings offer different levels of care and accessibility options. Tailoring treatment location to individual needs and preferences can enhance treatment engagement, retention, and overall outcomes.

Considering the abundant location-allocation approaches and treatment types, this paper aims to present a bi-objective location-allocation model of interventions in high drug consumption areas incorporating Twitter topic modeling to design a network that first identifies where to place an intervention point (health centers), taking into account distance proximity and equitable distribution, then allocate demand to each established center. In addition, this model seeks to provide information about the different opinions about drugs present in social network interactions to provide policymakers and public health officials with relevant information for the design of prevention and mitigation treatments. In this way, it could improve health outcomes for individuals and communities affected by drug addiction.

4.3 Model formulation

The healthcare facility network design problem determines the location and capacity of establishments that serve diverse patients (population zones) in need of their services [6]. An important feature of the location-allocation of health care facilities is their multi-objective nature due to the need for an effective system with good service quality and the endeavor to retain a balanced spatial distribution of services with people having access to the facilities, even in the most distant areas (equity) [7]. For the optimal design of the health care system, we consider two critical factors, efficiency, and equity as applied to the location of facilities. An effective system maximizes social welfare (Minimizes the risk) in high-consumption areas under defined budgets. Unfortunately, the unequal distribution of accessibility to HC facilities causes a troublesome situation. Patients far

from HC facilities have less opportunity to be served due to increased travel costs and travel time. The principle of equity imposes that accessibility costs to equal welfare opportunities should be evenly distributed [8]. Therefore, the optimal health provision system must be spatially fair and efficient.

The model's basic assumptions are: *i*) health centers are capacitated; *ii*) the time and distance between the demand node and the health center nodes are constant. *iii*) serving more patients decreases the consumption risk. The proposed bi-objective integer programming (BI-IP) location-allocation model is formulated as follows:

The first objective function is to maximize the weighted sum of total patient welfare and minimum total distance (Equation 1)

$$\text{Max OF1} = \delta \left(\sum_{i \in I} \sum_{j \in J} \sum_{k \in K} R_i^k * Y_{ij}^k \right) + (1 - \delta) \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} \left[\frac{-d_{ij}}{\text{Max}(d_{ij})} \right] * Y_{ij}^k \quad (1)$$

The total patient welfare is defined as the number of patients at high risk served at the demand node i . The risk index related to mitigation is calculated as the average of the SPA consumer demand percentage, the negative posts percentage on social networks (Twitter) associated with the SPA consumption, and the risk factors in node i (Equation 2). The factor risk is defined as those personal, environmental, or substance-related circumstances or characteristics (Violence, crime, robbery, etc.) that increase the likelihood that a person will use drugs [9]. In order to estimate the risk index related to prevention, we assume for every two people at mitigation risk, there is one person in prevention i.e., $R_i^2 = \frac{1}{2} R_i^1$.

$$R_i^1 = \frac{\bar{D}_i + \bar{\Psi}_i + \bar{\gamma}_i}{3} \quad (2)$$

The distance is defined as the shortest driving distance from demand node i to health center j . An algorithm was designed that calculates the distance between nodes using the GeoJSON API of the web mapping service developed by Google Maps based on [10] (See appendices). This algorithm was based on the possibility of accessing public cartography and accurate road network data for specific study problems and then exporting these data remotely to a database. Users can easily access the application from any computer connected to the Internet using a standard browser and an API Key. API

keys are generated in the Google Cloud console and are unique identifiers that authenticate calls to the Google Maps Platform. Similarly, the second objective function maximizes the weighted sum of equity and minimum total distance (Equation 3).

$$\text{Max OF2} = \pi \left(\sum_{i \in I} \sum_{j \in J} \sum_{k \in K} \sigma_i^k * Y_{ij}^k \right) + (1 - \pi) \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} \left[\frac{-d_{ij}}{\text{Max}(d_{ij})} \right] * Y_{ij}^k \quad (3)$$

The equity index corresponds to the product of the multidimensional poverty index (MPI) and the rurality proportion in node i (Equation 4). To estimate the risk index related to prevention, we implemented $\sigma_i^2 = \frac{1}{2} \sigma_i^1$.

$$\sigma_i^1 = \text{Rural}_i * \text{MPI}_i \quad (4)$$

A demand constraint was required (Equation 5). It Ensures that all flows from node i to HC should be as close as possible to demand D_i . We use MPI as the index to represent equity due to MPI measures the complexities of poor people's lives, individually and collectively [11]. Generally, the poorest people live in areas of difficult access (non-urban) and relatively low living conditions.

$$\sum_{j \in J} Y_{ij}^k \leq D_{i,k}, \quad \forall i \in I \text{ and } k \in K \quad (5)$$

In the same way, Equation 6 is the budget constraint. It ensures that all financial resources must not exceed the allocated funds for preventing and mitigating psychoactive substance consumption.

$$F^k \sum_{j \in J} C_j^k \leq B^k, \quad \forall k \in K \quad (6)$$

Similarly, two constraints were introduced in the model; these define the capacity of each facility. First, one sets a minimum number of working-hour to retain the accreditation R_{\min} in the HC if it is decided to open the center in node j (Equation 7).

$$\sum_{k \in K} C_j^k \geq R_{\min}^k * X_j, \quad \forall j \in J \quad (7)$$

The other constraint ensures that the estimated capacity of a health center is not exceeded (Equation 8). L is the number of hours a patient can be assigned.

$$L \sum_{i \in I} Y_{ij}^k \leq C_j^k, \quad \forall j \in J, k \in K \quad (8)$$

In addition, link constraints were considered. It is impossible to define a health center's capacity unless it is open. In Equation 9, M is a big number, and the binary variable X_j can take two values, either 0 or 1. Considering the options and choosing this M , it is not recommended to make it excessively big because it slows down some of the solution time. Instead, it was proposed to set it equal to the sum of all the demand because the flow on each arc can never be greater than the total demand required.

$$\sum_{i \in I} \sum_{k \in K} Y_{ij}^k \leq M * X_j, \quad \forall j \in J \quad (9)$$

Finally, equity constraints were considered. Based on the assumption derived from the public mental health policy, where the benefits should be distributed fairly among the population. It was established that at least φ patients should be satisfied in each demand node D_i .

$$\sum_{j \in J} \sum_{k \in K} Y_{ij}^k \geq \varphi, \forall i \in I \quad (10)$$

4.4 Epsilon constraint method

To solve the bi-objective model, we implemented the epsilon constraint method, a mathematical procedure for solving multi-objective optimization problems that involve multiple conflicting objectives. It works by introducing additional constraints into a single-objective optimization problem that limits the value of some of the objectives [27]. The method can be formulated as follows:

Given a multi-objective optimization problem with n decision variables and m objectives:

Minimize or maximize $f_1(x)$, subject to:

$$g_1(x) \leq 0, \dots, g_k(x) \leq 0, \dots, g_m(x) \leq \varepsilon_m \quad (11)$$

where x is the vector of decision variables, $f_1(x)$ is the primary objective function to be optimized, $g_1(x), \dots, g_k(x)$ are the constraints that define the feasible region of the problem, and ε_m is a positive scalar that determines the maximum value of the m -th objective. The epsilon constraint method transforms the multi-objective optimization

problem into a set of single-objective optimization problems, each with a different value of epsilon. By solving each of these problems, the method produces a set of Pareto optimal solutions, which represent the trade-off between the objectives. The mathematical procedure of the epsilon constraint method can be summarized as follows:

- Step 1: Choose one objective as the primary objective and set it as the objective function of the optimization problem.
- Step 2: For each of the other objectives, add an epsilon constraint that limits its maximum value. If the objective function is max, the constraint is $g_m(x) \geq \varepsilon_m$. The selected epsilon must be in range of $f_m(x)^{min} \leq \varepsilon_m \leq f_m(x)^{max}$.
- Step 3: Solve the optimization problem with the epsilon constraints for a given value of epsilon.
- Step 4: Repeat steps 2 and 3 for different values of epsilon.
- Step 5: Identify the set of Pareto optimal solutions by analyzing the solutions obtained for each value of epsilon.
- Step 6: Choose the value with the less distance (Euclidean distance) between the ideal point (objective value without using the epsilon constraint) and pareto frontier.

4.5 Twitter topic modeling

Social networks such as Facebook, LinkedIn, and Twitter have been crucial sources of information for a broad spectrum of users. In this study, we extracted data from Twitter to analyze the consumers' opinions about psychoactive substances. First, we built an algorithm to pull posts over some time and store them in a database in Postgres SQL (See appendices). Once we had all the posts from areas with high drug consumption, we identified the topics available on the Twitter data related to psychoactive substances. Unfortunately, the language used in tweets is often informal, containing grammatically creative text, slang, emoticons, and abbreviations, making it more challenging to extract topics than from standard text [28]. So, we used a framework to evaluate the data quality [29]. This framework examines each user's post in the data set to ensure it includes related terms from the ontology, where a tweet is selected if it has at least two terms from the ontology. We used the ontology proposed by [29].

Regarding topic modeling, it is defined as a statistical method used to identify topics or themes in a large collection of text documents. It is commonly used in natural language processing and machine learning applications, including social media analysis. These probabilistic models usually present topics as multinomial distributions over words, assuming that each document in a collection can be described as a mixture of topics [30]. The following are the general steps involved in Twitter topic modeling:

- **Step 1 (Data Collection):** Collect Twitter data using the Twitter API importing Tweepy package. Then, we used the URL <https://api.twitter.com/2/tweets/search/all> under the research academic access to extract tweets based on the following query:

```
query_params = {'query': '(droga OR drogadicto OR marihuana OR hierba OR cannabis OR cannabis OR opioide OR opioides OR opiaceo OR opiaceos OR morfina OR oxicodona OR oxicotin OR fentanilo OR durogesic OR hidromorfona OR meperidina OR tramadol OR tramal OR hidrocodona OR vicodin OR sinalgen OR heroína) lang:es -is:retweet point_radius: [-74.98342 10.71750 25mi]', 'tweet.fields': 'author_id, created_at, public_metrics', 'start_time': '2019/01/01', 'end_time': '2019/12/31', 'max_results': '100', 'place.fields': 'country, country_code', 'expansions': 'geo.place_id'}
```

Then, it is stored all the tweets collected in a database in Postgres SQL.

- **Step 2 (Data Cleaning):** Preprocess the collected tweets by removing URLs, special characters, stop words, and irrelevant content such as retweets or replies. Also, it is performed tokenization, stemming, and lemmatization to standardize the text.
- **Step 3 (Feature Extraction):** Convert the preprocessed tweets into a numerical representation using a vectorization technique such as TF-IDF, TF or bag-of-words. This step involves creating a term-document matrix where each row represents a tweet, and each column represents a unique term or word.
- **Step 4 (Topic Modeling):** Apply a clustering or decomposition algorithm to the term-document matrix to identify the underlying topics in the tweets. In this research, we used Latent Dirichlet Allocation (LDA), a probabilistic topic modeling algorithm to identify the underlying topics in a collection of documents [31]. The mathematical procedure for LDA can be described as follows:
 1. **Input:** A corpus of N documents, where each document is represented as a bag-of-words of M terms. Let W_{ij} be the frequency of term j in document i , and let K be the number of topics to be identified.
 2. **Initialization:** Assign each term j in each document i to a random topic k between 1 and K .

3. Iteration: Repeat the following steps until convergence: (a) For each topic k , estimate the probability distribution over the terms in the vocabulary, denoted by β_k . (b) For each document i , estimate the probability distribution over the topics, denoted by θ_i . (c) For each term j in each document i , estimate the probability distribution over the topics, denoted by φ_{ij} . (d) Sample new topic assignments for each term j in each document i based on the estimated probability distributions β_k , θ_i , and φ_{ij} .
4. Output: The final topic assignments for each term j in each document i , as well as the estimated probability distributions β_k and θ_i .

In LDA, each topic k is represented as a probability distribution over the terms in the vocabulary, denoted by β_k . Each document i is represented as a probability distribution over the topics, denoted by θ_i . Each term j in each document i is represented as a probability distribution over the topics, denoted by φ_{ij} . The goal of LDA is to find the probability distributions β_k , θ_i , and φ_{ij} that best explain the observed data. This is achieved by maximizing the likelihood of the data given the model parameters, using the Expectation-Maximization (EM) algorithm. The EM algorithm iteratively updates the probability distributions β_k , θ_i , and φ_{ij} based on the observed data, and the resulting topic assignments are used to update the model parameters.

- Step 5 (**Topic Evaluation**): Evaluate the quality and coherence of the identified topics using metrics such as perplexity and coherence.
- Step 6 (**Interpretation**): Interpret the identified topics by examining the top terms and their co-occurrence patterns.

4.6 Case study

The model was implemented in Atlántico, Colombia. In general terms, the Department of Atlántico is located in the north of the national territory, in the Caribbean region; located between 10°15'36" and 11° 06'37" north latitude, and 74°42'47" and 75°16'34" west longitude. It has an area of 3,386 km², representing 0.29 % of the national territory. This state has 22 towns, including the district of Barranquilla as its capital, and has a population of approximately 2'535,517 inhabitants distributed in 94.8% urban and 5.2% rural [32].

In terms of the level of development, as measured by gross domestic product (GDP), Atlántico is the department with the highest level of development in the Caribbean region, maintaining its economy above the national average [33]. Regarding psychoactive substance use, there is a more representative consumption of non-prescription tranquilizers, opioids, ketamine, GHB, alcohol, and heroin [34]. In particular, this research only considers the implementation of the model for opioids since these are the drug most frequently consumed in this department [34]. To represent the demand of an area, the centroid of the area was calculated, as shown in Figure 21.

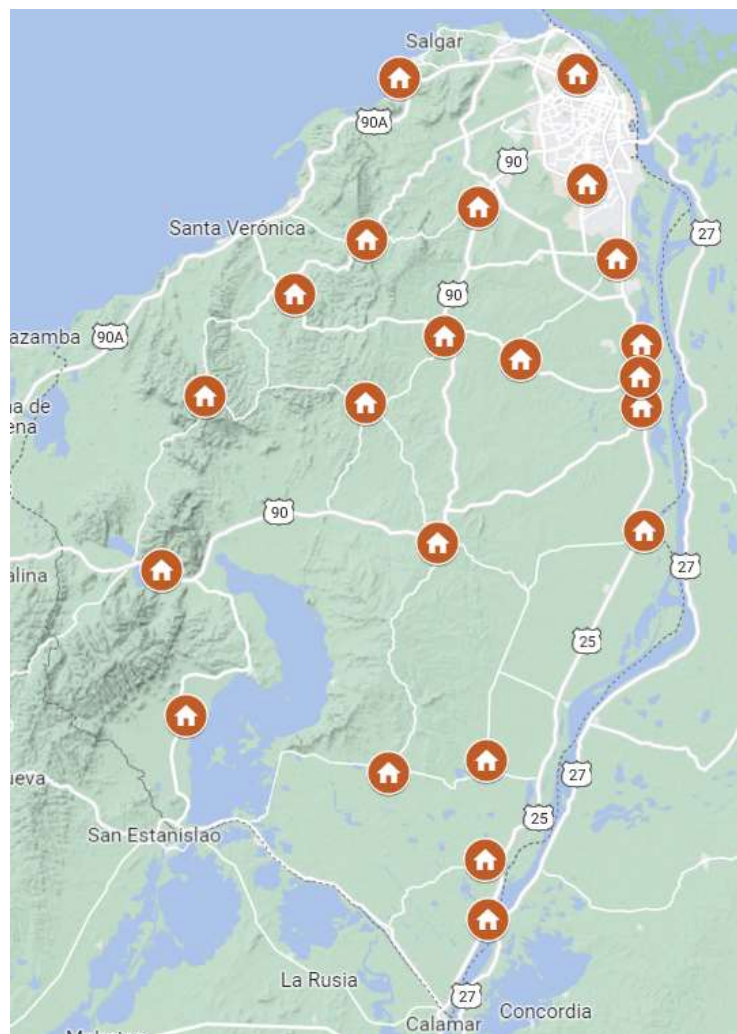


Figure 21. Health center location

4.6.1 Model parameters

- **Demand (D_i):** We used information retrieved from the 2019 National Survey of Psychoactive Substance Consumption in the General Population (DANE-DIMPE-ENCSPA-2019) conducted by the National Statistical System (DANE) of Colombia to estimate the consumption demand of each location. In this sense, the

PAS consumption demand was calculated as the count of people using drugs in each town in the Department of Atlántico during 2019. We consider at-risk people who are older than 15 years old. According to DANE, 68.26% of people in Colombia are between 15 and 64 years old.

- **Distance (d_{ij}):** Computing the distance between combinations of supply and demand nodes is an input required by the model. In addition, a network structure defined on maps and routing optimization algorithms is needed. However, the availability of this data and the price of good mapping can be challenging. Therefore, a Python code was developed and integrated with the GeoJSON API to overcome these limitations. With this code, it is possible to collect real distance data from the road network for any case study and record this information in a database. In this case, the API distance matrix created by Google is a service that provides the distance and travel time for a matrix of origins and destinations according to the recommended route between the start and endpoints. Anyone can access the API through an HTTP interface, with requests built as a URL string, using sources and destinations, along with an API key [18]. By default, distances are calculated for driving mode using the road network. Also, distance values may be subject to certain restrictions. Restrictions are indicated by choosing what Google should avoid when calculating the travel time (tolls, highways, ferries, indoor, or default: null). No restrictions were included in this code. In addition, units specify either metric or imperial units when displaying distances in the results. If units are not specified, the origin country of the query determines the units to use. The Python code is shown in Appendices.
- **Budget (B^k):** We used the information the Ministry of Health in Colombia presented to establish the funds available for reducing psychoactive substance use [19]. Then, to estimate the total budget for preventing and mitigating psychoactive substance use, we used the distribution of financial resources by the department [20]. In total, \$3,089,239,825.73 Colombian pesos were allocated for mitigation campaigns and \$654,162,455.84 Colombian pesos for prevention.
- **Multidimensional poverty index (MPI_i):** The MPI values for each department were retrieved from the multidimensional poverty report presented by the National Statistical System (DANE) [11].

- **Cost of service per hour (F^k):** The monthly pay for a psychiatrist in Colombia is \$6,127,416 COP, whereas a psychologist's average salary is \$1,800,000 COP. The role of a psychiatrist involves administering mitigation treatments, while a psychologist focuses on providing prevention treatments. It is assumed that the professionals work for eight hours a day, twenty days a month.
- **Weights in the objective functions (π, δ):** The aim of using weighted weights is to balance the different objective functions. In the case of objective function 1, it considers the number of people served at risk, and the total distance traveled, with δ being the controlling parameter. The research uses a value of δ : 0.5, but this value can vary based on how important distance or risk is relative to each other. Similarly, objective function two is treated analogously, with the parameter π being used.
- **Minimum number of working-hour to retain the accreditation (R_{\min}):** This value represents the number of hours per year in minimum capacity. Thus, it is estimated based on a shift of eight hours per day for 240 days per month. A distribution of equal hours for mitigation and prevention is assumed.
- **Number of persons per hour of therapy (L^k):** After conducting interviews with various healthcare organizations that offer treatment for patients with psychoactive substance use, it has been concluded that mitigation therapy typically lasts for one hour and is conducted on an individual basis. Conversely, prevention therapy also lasts for one hour, but it is administered to a group of 25 people.
- **Factor risk ($\overline{\psi}_l$):** A risk factor refers to any variable or characteristic that increases the likelihood of an individual developing a substance use disorder or experiencing negative consequences as a result of substance use. In this study, we estimate this value based on the proportion of homicides, thefts, terrorism, sexual crimes, domestic violence, and threats in each department town. This information was retrieved from the databases of the Colombian National Police; URL: <https://www.policia.gov.co/grupo-informacion-criminalidad/estadistica-delictiva>.
- **Negative posts risk on social media ($\overline{\gamma}_l$):** For this study, we utilized the database obtained from the Twitter Topic Modeling section in a previous chapter. To identify negative tweets according to location, we employed four Pre-training of

Deep Bidirectional Transformers, including Robustly Optimized BERT Pretraining Approach (Roberta) [21], [22]; Valence Aware Dictionary for sEntiment Reasonin VADER [23]; Bidirectional Encoder Representations from Transformers model trained on a Spanish corpus BETO [24]; and *roBERTa-base-bne* [25], which is a transformer-based masked language model designed for the Spanish language. This last model was built on top of the *RoBERTa* base model and was pre-trained using the largest known Spanish corpus to date, which consists of 570GB of clean, deduplicated text collected from web crawls conducted by the National Library of Spain between 2009 and 2019. The distribution of positive, neutral, and negative comments per location is shown in Figure 22. On the other hand, Figure 23 shows the results of the applied sentiment algorithms. It is noted that *roBERTa-base-bne* has the highest probability values for rating negative tweets compared to the rest of the algorithms. It is essential to mention that these results should be calculated considering Table 9, which indicates the number of tweets per town. For example, in the case of Ponedera, only one single tweet is registered; however, in Figure 23, for all algorithms, this location seems to have a high percentage of negative tweets.

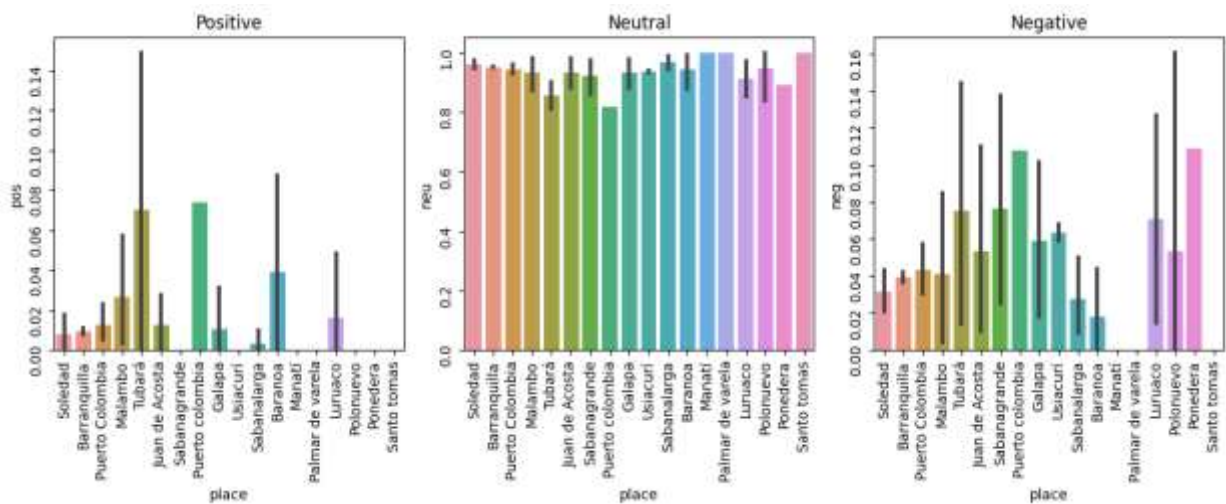


Figure 22. Distribution of positive, neutral, and negative comments per location.

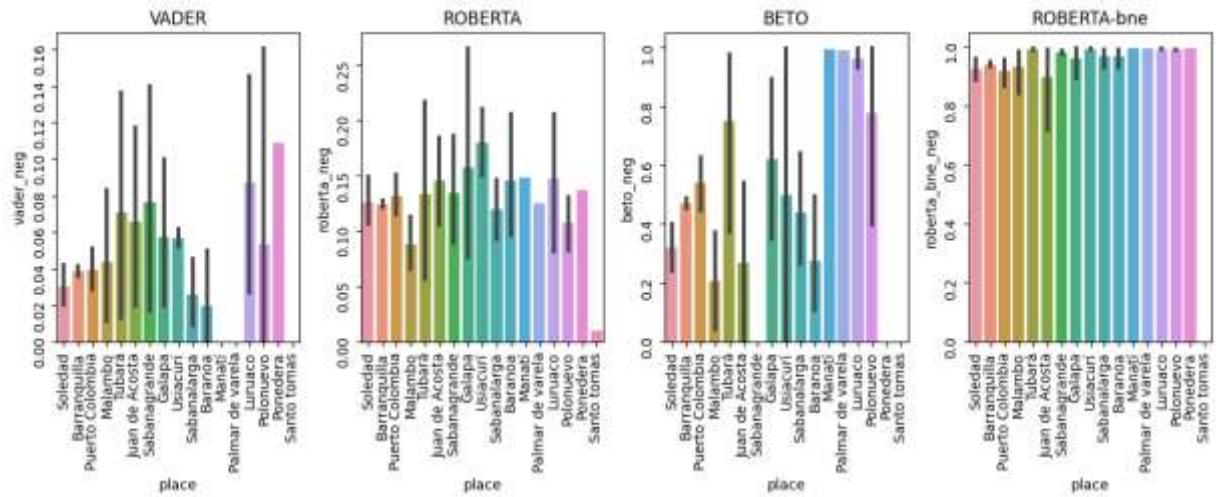


Figure 23. Sentiment algorithms results.

Table 9. Distribution of tweets per location.

Department	Tweets
Barranquilla	2741
Baranoa	18
Campo de La Cruz	-
Candelaria	-
Galapa	11
Juan de Acosta	11
Luruaco	-
Malambo	24
Manatí	1
Palmar de Varela	1
Piojón	-
Polonuevo	5
Ponedera	1
Puerto Colombia	108
Repelón	-
Sabanagrande	8
Sabanalarga	24
Santa Lucía	-
Santo Tomás	1
Soledad	124
Suan	-
Tubará	5
Usiacurí	2

4.6.2 Computational results and insights

In this section, we present the case study results to demonstrate the tradeoff in the model and understand the interaction among facility location, capacity selection, and allocation of patients to facilities in the design of a healthcare facility network under resource

constraints. The computational experiments were performed using Python V.3.10 and the Gurobi Optimizer V 9.5.2. The proposed algorithm converged to a solution with an optimality gap close to 10^{-4} . A gap refers to the difference between the objective value of the best-known solution and the optimal value of the problem being solved. It measures how far the current solution is from the best possible solution. A small gap indicates that the algorithm is close to finding the optimal solution, while a large gap suggests that further improvements are needed.

As a result, the integer programming model presented in Section 4.3 has 1127 integers and 23 binary variables. The model was solved to exact optimality on a Core I5 machine with a 2.9 GHz processor and 16 GB RAM in 0.01s computational time for each run. Adopting the epsilon constraint method, we optimize the first objective. At the same time, the other one is constrained to values that vary through a range of feasible values; then, we repeat the process until we have the Pareto frontier. The tradeoff between the risk and equity objectives produces efficient solutions, as we observe diagrammatically in Figure 24. For these solutions, it is not possible to improve an objective without deteriorating the other function. The ideal point (29492.86, 6223.28) corresponds to solving the model for each objective function; meanwhile, the Optimal point (26665.56, 470.04) corresponds to the point with the minimal distance between the ideal point and the Pareto frontier. We added two new constraints to the model based on the optimal point with a lower bound of 26665.56 and 470.04 for objectives one and two, respectively.

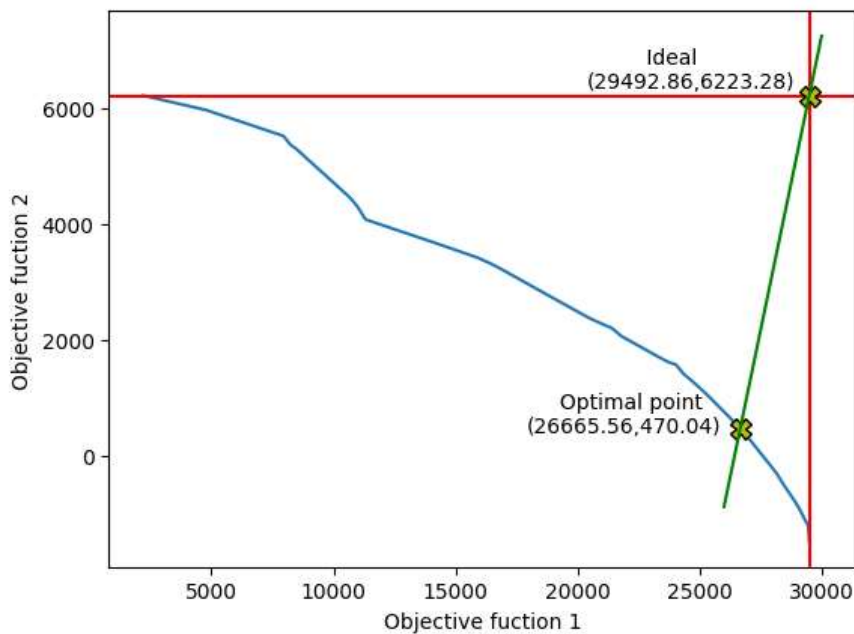


Figure 24. Pareto frontier

Table 10 shows the total therapies of each demand node to the health center for prevention. Considering only the first objective function, the solution indicates that two health centers (Barranquilla, Baranoa) must be established to serve people under preventive treatment. Meanwhile, the second objective function shows that the centers to be opened are Juan de Acosta and Sabanalarga. These independent solutions show a conflict between where to locate the intervention center and how to assign people to each center, the first trying to put intervention in areas with high drug consumption, and the second trying to set a center closer to rural zones. In this sense, the solution obtained by the bi-objective model presents an equilibrium between the previous solutions, i.e., to consider setting up a health center in Barranquilla that assist zones with high drug risk and other in Sabanalarga serving rural zones. The center in Barranquilla would serve Malambo, Puerto Colombia, Soledad, Barranquilla, and Tubará, while the center in Sabanalarga would serve the remaining towns.

Table 10. Prevention hour distribution in each health center

	Bi model		Objective 1		Objective 2	
	Center 1: Barranquilla	Center 2: Sabanalarga	Center 1: Barranquilla	Center 2: Baranoa	Center 1: Juan de acosta	Center 2: Sabanalarga
Barranquilla	62387	-	67340	-	2000	-
Baranoa	-	2000	-	2000	2000	-
Campo de La Cruz	-	2000	-	2000	-	2000
Candelaria	-	2000	-	2000	-	2000
Galapa	-	2000	-	2000	2000	-
Juan de Acosta	-	2000	-	2000	14334	-
Luruaco	-	2000	-	2000	-	19233
Malambo	1353	647	-	2000	1993	7
Manatí	-	2000	-	2000	-	2000
Palmar de Varela	-	2000	-	2000	-	2000
Piojó	-	2000	-	2000	4511	-
Polonuevo	-	2000	-	2000	2000	-
Ponedera	-	2000	-	2000	-	16279
Puerto Colombia	2000	-	2000	33628	2000	-
Repelón	-	2000	-	2000	-	2000
Sabanagrande	-	2000	-	2000	-	2000
Sabanalarga	-	6953	-	2000	-	11121
Santa Lucía	-	2000	-	2000	-	2000
Santo Tomás	-	2000	-	2000	-	2000
Soledad	2000	-	400	1600	2000	-
Suan	-	2000	-	2000	-	2000
Tubará	2000	-	-	2000	11862	-
Usiacurí	-	2000	-	2000	-	2000

Note: Prevention treatment serves 20 people, and each therapy takes one hour. The total number of therapies represents the annual distribution.

Concerning the mitigation treatments, Table 11 shows the total therapies of each demand node to the health center. This way, the center in Barranquilla would serve the population from Malambo, Barranquilla Galapa, Sabanagrande, Puerto Colombia, Soledad, and Tubará. In contrast, the center in Sabanalarga would serve the remaining towns. It is essential to mention that this solution seeks to reduce travel distances from the demand node to each center and reduce the risk of consumption in the target population. As stated in section 4.3, for prevention and mitigation, a minimum number of hours of therapy (100) was estimated to be distributed to each town (equity).

Table 11. Mitigation hour distribution in each health center

	Bi model		Objective 1		Objective 2	
	Center 1: Barranquilla	Center 2: Sabanalarga	Center 1: Barranquilla	Center 2: Baranoa	Center 1: Juan de acosta	Center 2: Sabanalarga
Barranquilla	18855	-	20137	-	100	-
Baranoa	-	100	-	100	3057	-
Campo de La Cruz	-	100	-	100	-	100
Candelaria	-	100	-	100	-	784
Galapa	100	-	-	100	100	-
Juan de Acosta	-	100	-	100	1029	-
Luruaco	-	1380	-	100	-	1381
Malambo	100	-	-	100	100	-
Manatí	-	100	-	100	-	100
Palmar de Varela	-	100	-	100	-	100
Piojó	-	100	-	100	324	-
Polonuevo	-	100	-	100	892	-
Ponedera	-	100	-	100	-	1169
Puerto Colombia	100	-	100	-	100	-
Repelón	-	100	-	100	-	1284
Sabanagrande	100	-	-	100	-	100
Sabanalarga	-	101	-	100	-	4570
Santa Lucía	-	100	-	100	-	100
Santo Tomás	-	100	-	100	-	100
Soledad	100	-	20	80	100	-
Suan	-	100	-	100	-	100
Tubará	101	-	-	100	852	-
Usiacurí	-	100	-	100	-	100

On the other hand, Table 12 presents the percentage of demand covered within a 40 km range. Under objective function 1, Piojó demonstrates coverage values of 44.34% and 30.86%, while Barranquilla registers 8.18% and 34.07% for prevention and mitigation, respectively. Suan follows with 24.54% and 17.06% coverage for prevention and mitigation. These municipalities exhibit the highest coverage percentages within the 40 km radius. With objective function 2, Piojó, Juan de Acosta, Luruaco, Ponedera, and Tubará achieve full coverage (100%) within the 40 km range. Consequently, the entire population served in these towns will need to travel no more than 40 km, equivalent to

40–50 minutes of public transport or car travel. In general, implementing the Bi-model results in an average coverage of 78.26%. However, using objective function 1 (risk) would cover only 73.91% of the demand, while relying solely on objective function 2 (equity) would cover 69.57% within less than 40 km.

Table 12. Demand coverage (< 40 km)

	Bi model		Objective 1		Objective 2	
	Prevention	Mitigation	Prevention	Mitigation	Prevention	Mitigation
Barranquilla	7.58	31.9	8.18	34.07	0.24	0.17
Baranoa	4.7	3.27	4.7	3.27	4.7	100
Campo de La Cruz	12.81	8.91	12.81	8.91	12.81	8.91
Candelaria	18.33	12.76	18.33	12.76	18.33	100
Galapa	4.83	3.36	4.83	3.36	4.83	3.36
Juan de Acosta	13.95	9.72	13.95	9.72	100	100
Luruaco	10.4	99.93	10.4	7.24	100	100
Malambo	2.29	1.59	2.29	1.59	2.29	1.59
Manatí	14.79	10.3	14.79	10.3	14.79	10.3
Palmar de Varela	10.13	7.05	10.13	7.05	10.13	7.05
Piojó	44.34	30.86	44.34	30.86	100	100
Polonuevo	16.1	11.21	16.1	11.21	16.1	100
Ponedera	12.29	8.55	12.29	8.55	100	100
Puerto Colombia	5.95	4.14	5.95	4.14	5.95	4.14
Repelón	11.18	7.79	11.18	7.79	11.18	100
Sabanagrande	9.06	6.31	9.06	6.31	9.06	6.31
Sabanalarga	10.92	2.21	3.14	2.19	17.47	100
Santa Lucía	18.28	12.72	18.28	12.72	18.28	12.72
Santo Tomás	9.82	6.84	9.82	6.84	9.82	6.84
Soledad	0.49	0.34	0.49	0.34	0.49	0.34
Suan	24.54	17.06	24.54	17.06	24.54	17.06
Tubará	16.86	11.85	16.86	11.74	100	100

Table 13 presents the comprehensive capacity totals from implementing three models to assess health centers' capacity. When considering objective function 1, the health center in Barranquilla would be capable of accommodating 23,744 therapies, with 2,487 designated for prevention and 20,257 for mitigation purposes. Similarly, the health center in Baranoa would provide 4,160 therapies, evenly distributed between prevention and mitigation (2,080 each). Applying objective function 2, the health center in Luruaco would handle 8,889 therapies, with 2,235 allocated for prevention and 6,654 for mitigation. In contrast, the health center in Sabanalarga would accommodate 13,320 therapies, with 3,332 intended for prevention and 9,988 for mitigation. With the implementation of the proposed BI model, the health center in Barranquilla would administer 22,943 therapies, dividing them into 3,487 for prevention and 19,456 for

mitigation. Additionally, the health center in Sabanalarga would provide 4,961 therapies, with 2,080 allocated for prevention and 2,081 for mitigation. Notably, this allocation ensures an equitable distribution of services among different demand zones. It is essential to mention that Barranquilla would account for 82% of the treatments, while Sabanalarga would cover only 18% of the total treatments available in a year within the budget constraints. The flow path to each health center is shown in Figure 25.



Figure 25. Flow network

Table 13. Health center capacity

	Bi model		Objective 1		Objective 2	
	Prevention	Mitigation	Prevention	Mitigation	Prevention	Mitigation
Barranquilla	3487	19456	3487	20257	-	-
Baranoa	-	-	2080	2080	-	-
Campo de La Cruz	-	-	-	-	-	-
Candelaria	-	-	-	-	-	-
Galapa	-	-	-	-	-	-
Juan de Acosta	-	-	-	-	2235	6654

Luruaco	-	-	-	-	-	-
Malambo	-	-	-	-	-	-
Manatí	-	-	-	-	-	-
Palmar de Varela	-	-	-	-	-	-
Piojó	-	-	-	-	-	-
Polonuevo	-	-	-	-	-	-
Ponedera	-	-	-	-	-	-
Puerto Colombia	-	-	-	-	-	-
Repelón	-	-	-	-	-	-
Sabanagrande	-	-	-	-	-	-
Sabanalarga	2080	2881	-	-	3332	9988
Santa Lucía	-	-	-	-	-	-
Santo Tomás	-	-	-	-	-	-
Soledad	-	-	-	-	-	-
Suan	-	-	-	-	-	-
Tubará	-	-	-	-	-	-
Utiacurí	-	-	-	-	-	-

The proposed objective functions are a linear combination of distance, risk, and equity factors. To implement the model, we use π and δ value of 0.7, meaning there is more weight for the risk and equity factor than for the distance between the demand node and the health center node. However, a sensitivity analysis was performed by varying π and δ . The results are presented in Figure 26, showing that for values between 0.1 and 0.4, the highest coverage is achieved using objective function 2 (Equity). While for objective function 1, the values for the highest coverage are between 0.1 and 0.3. It should be noted that having a value of 0.9 would achieve a coverage greater than 85%.

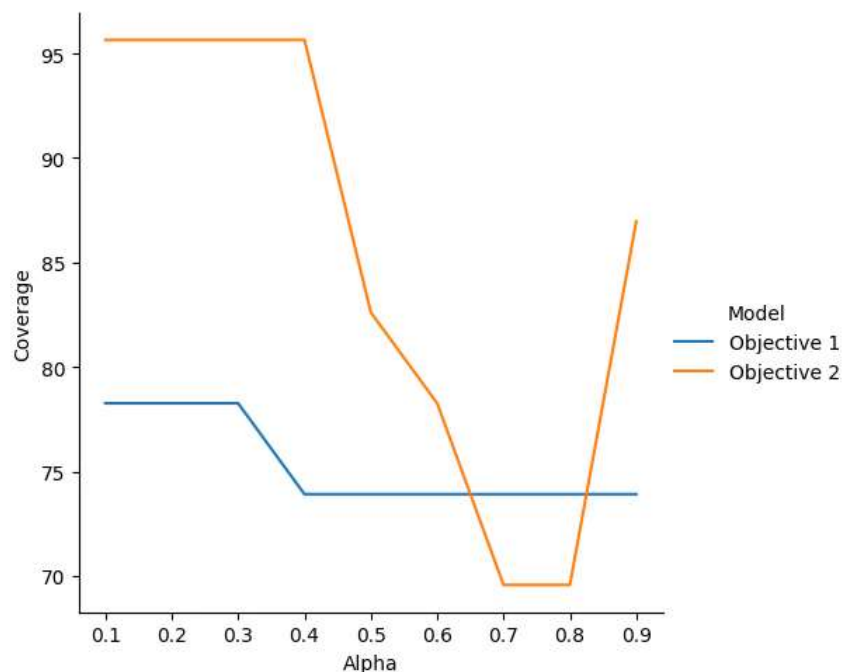





Figure 26. Sensitivity analysis of pi and alpha.

4.6.3 Topic modelling results

Once the intervention points were defined, the topics related to psychoactive substances were identified. To extract the topics, a database of 4,363 tweets belonging to the Department of Atlántico in Colombia was used. Using the evaluation matrix by [14]; 1,274 tweets were eliminated, qualifying 3,089 tweets as relevant (good quality). An example of the tweets collected can be seen in Table 14.

Table 14. Sample of the collected tweets

	El consumo intensivo de cannabis por parte de jóvenes con del estado de puede estar relacionado con un mayor riesgo de autolesión intentos de suicidio y muerte según estudios
	134 desempleados esquineros y vagos fumadores de marihuana y que seguramente perdieron el ICFES no queda de otra
	Falta que legalicen la marihuana ya logramos con el aborto

We utilized the LDA topic-mining algorithm as per the methodology. To assess the quality of topics produced by the topic model, we relied on two measures, namely the Topic Coherence score and Perplexity. The Topic Coherence score gauges the semantic coherence of the generated topics by LDA and is based on the similarity between words within a topic. A higher coherence score indicates the topic is more meaningful and interpretable, whereas a lower score implies randomness or a lack of coherence [26]. On the other hand, Perplexity is a commonly used measure to evaluate the performance of language models, including topic models. It assesses how accurately a language model can predict a held-out test set of documents, and a lower score indicates higher accuracy [27]. We discovered that coherence score peaked at nine topics, while Perplexity decreased with an increase in the number of topics but ultimately converged at around nine topics (Perplexity: -40,814). Therefore, we set the number of topics to nine (kindly see Figure 27).

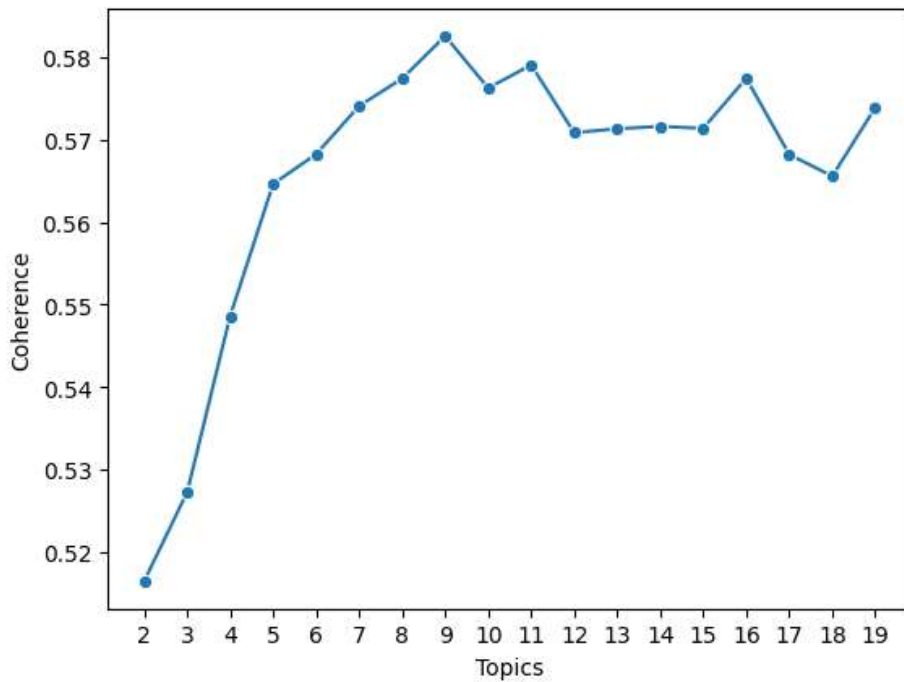


Figure 27. Coherence score over the number of topics

Figure 28 displays the results of the LDA model, which exhibits the word clouds for the top nine topics, with the size of each word indicating its unigram TF-IDF score. Since we do not have information (Tweets) for each municipality in the Department of Atlántico, we will use the model to represent the department's overall opinions. The description of the nine topics is presented below:

- T1 (Drug cartels and their impact on the world): Colombian former president Duque, was likely to be involved in efforts to combat drug cartels, which are known to produce and distribute illegal drugs. “*Fumar*”, or smoking, can also be linked to this issue, as tobacco is a legal drug that can have harmful effects on health, and some cartels may also be involved in the illegal tobacco trade. The impact of drug cartels can extend beyond Colombia to the rest of the world, as they can contribute to issues such as drug addiction, crime, and political instability. Overall, these words can be connected to the theme of drug-related issues and their global impact.
- T2 (Drug use and its consequences in different countries): “*Droga*”, or drugs, are substances that can have negative effects on health and well-being. “*Meter*”, or to use, can be linked to drug use, as it refers to the act of taking drugs. Parque, or park, may represent a public space where drug use can occur, highlighting the issue of drug use in public areas. “*Países*”, or countries, suggests that this issue

can have different manifestations and consequences in different parts of the world. Finally, “*decision*”, or decision, could represent the need for policymakers and individuals to make choices regarding drug use and its impact on society. Overall, these words can be connected to the theme of drug use and its impact on public spaces, individual health, and society as a whole, emphasizing the importance of making informed decisions and taking appropriate actions to address this issue.

- T3 (Personal choice and responsibility regarding cannabis consumption): Cannabis is a plant that can be used for recreational or medicinal purposes, and its consumption (“*consumo*”) can have both positive and negative effects on the health and well-being of individuals. The term “*persona*”, or person, emphasizes the importance of considering the individual and their unique circumstances when it comes to cannabis consumption. Finally, “*produce*”, or to produce, can refer to the cultivation and production of cannabis, highlighting the need for responsible and sustainable practices in this industry. Overall, these words can be connected to the theme of cannabis production and consumption and its impact on individuals and society, emphasizing the need for informed decision-making and responsible practices to ensure the safety and well-being of all stakeholders involved.
- T4 (Responsible and ethical behavior in the context of drug use and drug-related activities): “*Ser*”, or to be, can refer to the innate characteristics and values that individuals possess, highlighting the importance of personal responsibility in making decisions related to drug use. “*Hacer*”, or to do, can refer to the actions and behaviors individuals take regarding drugs, such as being responsible in the use and distribution of drugs. “*Gente*”, or people, represents the wider community in which individuals live and interact with, emphasizing the importance of responsible and ethical behavior towards others, particularly in the context of drug use. “*Favor*”, or favor, can represent the act of helping others, emphasizing the importance of supporting those who may be struggling with drug addiction or abuse. Legal, or legal, highlights the importance of following laws and regulations related to drug use and drug-related activities, emphasizing the importance of preventing harm and promoting public safety. Finally, the word drug itself emphasizes the importance of responsible and ethical decision-making when it comes to drug use, as the consequences of drug use can have significant impacts on both individuals and society as a whole. Overall, these words can be connected

to the theme of responsible and ethical behavior in the context of drug use and drug-related activities, emphasizing the importance of personal responsibility, empathy, and public safety.

- T5 (Controversy surrounding the legalization of marijuana and the impact it has on individuals and society): “*Legalización*”, or legalization, refers to the process of making marijuana legal for consumption, sale, and distribution. “*Millones*”, or millions, refers to the large number of people who may be affected by the legalization of marijuana, both positively and negatively. “*Vida*”, or life, highlights the potential impact of marijuana use on individuals' health and well-being, both physically and mentally. “*Hierba*”, or herb, refers to marijuana as a natural plant-based substance, which may be viewed positively or negatively depending on cultural and societal attitudes towards drug use. “*Droga*”, or drug, is a more general term that can be used to refer to any substance that has the potential to be abused or cause harm. “*Quiere*”, or wants, suggests the desire or interest that some individuals or groups may have in legalizing marijuana. Consume, or consume, refers to the act of using marijuana, emphasizing the potential impact of legalization on individuals' behavior and consumption patterns. Overall, these words can be connected to the ongoing debate and controversy surrounding the legalization of marijuana and the impact it may have on individuals and society, including issues related to health, legality, and cultural attitudes towards drug use.
- T6 (Drug addiction impacts individuals, society, criminal organizations, and vulnerable populations, such as minors): “*Adicción*”, or addiction, refers to the compulsive and harmful use of drugs that can lead to physical and mental health issues. “*Amigo*”, or friend, may highlight the role that social networks and peer pressure can play in drug use and addiction. “*Droga*”, or drug, refers to the substances that can lead to addiction and other harmful effects. Cartel and “*FARC-Santos*” refer to criminal organizations that are involved in drug trafficking and may exacerbate issues related to drug addiction and violence. “*Menores*”, or minors, refers to young people who may be particularly vulnerable to the negative effects of drug use and addiction. Together, these words suggest the complex and interconnected nature of the issue of drug addiction, which involves individual behavior, social networks, criminal organizations, and societal factors such as access to healthcare and education. They also highlight the need for

comprehensive and evidence-based approaches to address the issue of drug addiction and its impact on individuals and society.

- T7 (Pervasive and destructive impact of drug-related violence and corruption on individuals, communities, and society): “*Droga*”, or drug, refers to the substances that can lead to addiction and violence, and the devastating consequences of drug-related deaths. “*Muerte*”, or death, highlights the tragic toll of drug-related violence on individuals and communities, including innocent bystanders and those caught in the crossfire. “*Corrupción*”, or corruption, emphasizes the corrosive impact of illicit drug trade and the criminal enterprises it fuels on governance, law enforcement, and society at large. “*Poder*”, or power, highlights the ways in which drug-related violence and corruption are perpetuated by those who hold political, economic, or social power, and the challenges of addressing these entrenched interests. “*Grande*”, or big, underscores the scale and complexity of the drug trade and its impact on society, including the significant resources required to address it. “*Problema*”, or problem, underscores the urgency of addressing the root causes of drug-related violence and corruption, including poverty, inequality, and social exclusion. “*Corrupción*” and “*Partido*” both suggest the need for comprehensive and coordinated efforts to combat corruption and address the political and institutional factors that allow it to thrive. Together, these words highlight the need for a multi-faceted approach to address the complex and interconnected issues of drug-related violence and corruption, including a focus on prevention, law enforcement, and systemic reform.
- T8 (Potential negative impact of alcohol, sex, and addiction on individuals and communities, particularly in urban area): “Alcohol” represents the consumption of alcohol, which can lead to addiction and negative health consequences. “Sexo” refers to sexual activity, which can also have negative consequences when not practiced safely and consensually. “Barrio”, or neighborhood, suggests that these issues may be particularly acute in urban areas with high levels of poverty, crime, and social inequality. “Video” may represent the role of media and technology in promoting and normalizing these behaviors, as well as potential negative consequences such as addiction to video games or pornography. “Vicio”, or addiction, highlights the risk of developing problematic and potentially harmful behaviors associated with alcohol, sex, or other addictive substances or activities. Together, these words suggest the need for education, awareness, and support

services to address the negative consequences of these behaviors and promote healthier, safer lifestyles in urban communities.

- T9 (Drugs and organized crime): The words "*droga*", "*muerte*", "*social*", "*dosis*", "*guerra*", "*violadores*", "*homosexuales*", "*ladrones*", and "*secuestradores*" could be related to the negative effects of drug use on society. Drug use can lead to death, either from overdose or from involvement in drug-related violence. Drug addiction can also have social consequences, such as alienation from family and friends and a decline in productivity. The term "*dosis*" refers to the amount of drugs taken, and a high dose can be particularly dangerous. The term "*guerra*" (war) could be used to refer to the ongoing battle against drug trafficking and drug-related crime. The inclusion of "*violadores*" (rapists), "*homosexuales*" (homosexuals), "*ladrones*" (thieves), and "*secuestradores*" (kidnappers) may suggest that drug use can be associated with criminal behavior and contribute to a culture of violence. Overall, these words highlight the complex and multifaceted nature of the drug problem, which requires a comprehensive approach that addresses not only the physical and psychological aspects of drug use but also the social, economic, and political factors that contribute to drug-related violence and insecurity.

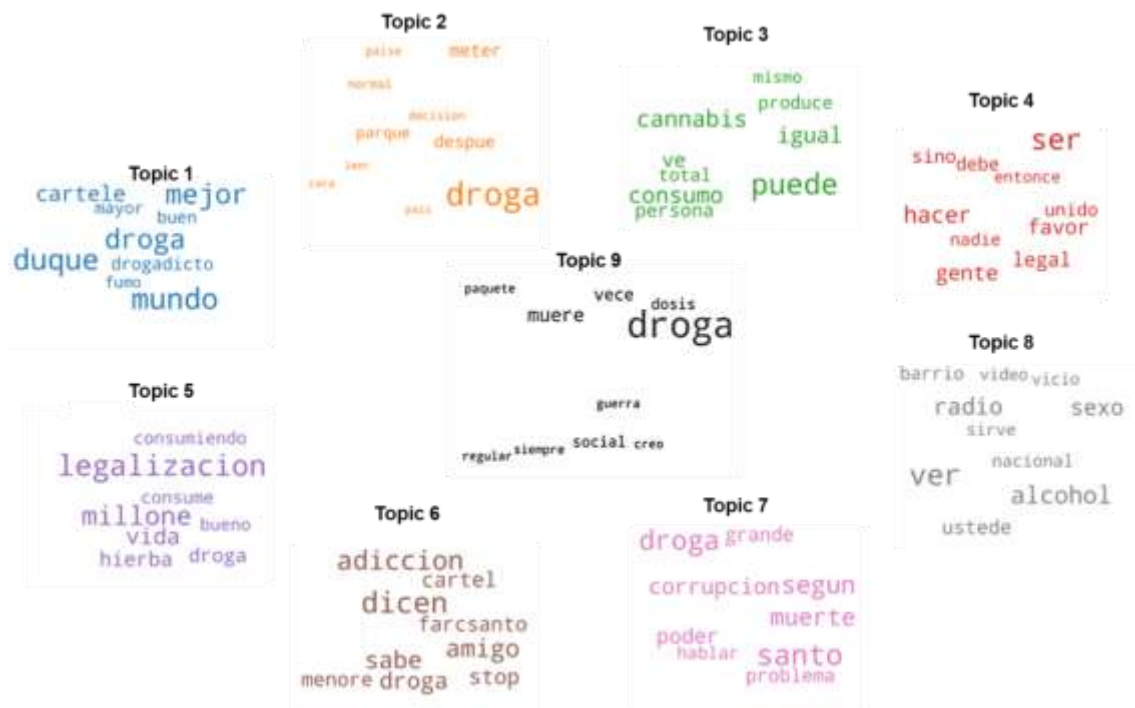


Figure 28. Top nine topic word clouds

Table 15 shows the distribution of public opioid tweets over the topics. The most prevalent topics were related to the opioid crisis. Many posts were related to topics, such as Drug cartels and their impact on the world; Drug use and its consequences in different countries; Personal choice and responsibility regarding cannabis consumption; Responsible and ethical behavior in the context of drug use and drug-related activities; Controversy surrounding the legalization of marijuana and the impact it has on individuals and society; Drug addiction impacts individuals, society, criminal organizations, and vulnerable populations, such as minors; Pervasive and destructive impact of drug-related violence and corruption on individuals, communities, and society; Potential negative impact of alcohol, sex, and addiction on individuals and communities, particularly in urban area; and drug and organized crime.

Table 15. Topic Contribution

Topic Number	Topic Contribution	Keywords	Representative Text
1	0.4869	mundo, duque, mejor, droga, carteles, dio, drogadicto, buen, mayor, fumo	['salmo', 'impaciente', 'causa', 'malignos', 'envidia',]
2	0.5941	droga, meter, parque, normal, países, decisión, cara, país, leer	['hacen', 'lugares', 'rumbear', 'diferente', 'ambiente', 'meter', 'droga', 'diferente', 'ambiente']
3	0.5448	puede, cannabis, consumo, igual, ve, do, persona, produce, total, mismo	['total', 'igual', 'impunidad', 'droga', 'borrachera']
4	0.4676	ser, hacer, gente, favor, legal, debe, unido, nadie,	['droga', 'alterar', 'gene', 'químicas', 'diferencia', 'religión', 'rastafari', 'simple', 'planta', 'planta', 'respetar']
5	0.4898	legalización, millones, vida, dice, hierba, droga, quiere, consume, consumiendo, bueno dicen, adicción, amigo, pue,	['invertido', 'miles', 'millones', 'pauta', 'promover', 'legalización', 'medio', 'twitter', 'llenos', 'testimonio', 'bueno', 'trabado']
6	0.5503	sabe, droga, stop, cartel, farcsanto, menores	['outfit', 'natural', 'mugre', 'natural', 'toda', 'natural']
7	0.5765	santo, droga, muerte, corrupción, poder, grande, problema, hablar	['liberal', 'partido', 'corrupto', 'lopez_michelsen', 'gaviria', 'samper', 'santo', 'insistir', 'promueven', 'corrupción', 'promotores']
8	0.487	alcohol, radio, sexo, barrio, nacional, sirve, video, vicio droga, muere, vece, social,	['vicio', 'sexo', 'gym', 'video', 'ver', 'dulce',]
9	0.516	dosis, siempre, guerra, paquete, regular	['droga', 'violadore', 'homosexuales', 'ladrone', 'secuestradore', 'cómplice', 'narcoasesino', 'farcsa']

4.6 Conclusion

Location-allocation bi-objective models can play a crucial role in optimizing the allocation of resources related to drug consumption. By considering objectives such as

minimizing drug risk and maximizing accessibility to rehabilitation centers or prevention programs, these models can assist in strategically placing treatment facilities, counseling centers, or harm reduction services in areas with high drug consumption rates. This approach can help improve the efficiency of resource allocation, reduce drug-related harm, and enhance the effectiveness of interventions.

Twitter topic modeling can offer insights into drug consumption patterns, public sentiment, and emerging trends in drug-related discussions on social media. By analyzing tweets related to drug use, researchers and public health professionals can identify key topics, sentiment trends, and influential factors associated with drug consumption behaviors. This information can contribute to targeted prevention campaigns, public awareness initiatives, and policy interventions aimed at reducing drug abuse and promoting healthier behaviors. The combination of location-allocation models and Twitter topic modeling can provide a comprehensive understanding of drug consumption dynamics. Integrating the physical aspect of resource allocation with the social media discourse surrounding drug use allows for a more holistic approach to addressing drug-related challenges. By leveraging these techniques together, decision-makers can identify geographical areas that require specific interventions, tailor prevention strategies to target high-risk populations, and adapt their efforts based on real-time feedback from social media discussions.

Ultimately, the integration of location-allocation bi-objective models and Twitter topic modeling can support evidence-based decision-making, enhance intervention strategies, and contribute to the overall goal of reducing drug consumption and its associated harms. By utilizing these approaches in a coordinated manner, stakeholders can work towards building healthier communities and improving public health outcomes related to drug use. Future research should focus on evaluating the effectiveness of different treatment locations, identifying barriers to access, and developing strategies to optimize the coordination and integration of care across various settings. By expanding and diversifying treatment location options, healthcare systems can improve the availability and effectiveness of drug addiction treatment and support individuals in their journey towards recovery.

References

- [1] M. Isorna, L. Fernández-Ríos, and A. Souto, “Treatment of drug addiction and psychopathology: A field study,” *Eur. J. Psychol. Appl. to Leg. Context*, vol. 2, no. 1, pp. 3–18, 2010.
- [2] E. E. DeVito, P. D. Worhunsky, K. M. Carroll, B. J. Rounsaville, H. Kober, and M. N. Potenza, “A preliminary study of the neural effects of behavioral therapy for substance use disorders,” *Drug Alcohol Depend.*, vol. 122, no. 3, pp. 228–235, May 2012, doi: 10.1016/J.DRUGALCDEP.2011.10.002.
- [3] D. de Andrade, R. A. Elphinston, C. Quinn, J. Allan, and L. Hides, “The effectiveness of residential treatment services for individuals with substance use disorders: A systematic review,” *Drug Alcohol Depend.*, vol. 201, pp. 227–235, Aug. 2019, doi: 10.1016/J.DRUGALCDEP.2019.03.031.
- [4] D. Best, V. Manning, S. Allsop, and D. I. Lubman, “Does the effectiveness of mutual aid depend on compatibility with treatment philosophies offered at residential rehabilitation services?,” *Addict. Behav.*, vol. 103, p. 106221, Apr. 2020, doi: 10.1016/J.ADDBEH.2019.106221.
- [5] K. Beardsley, E. D. Wish, D. B. Fitzelle, K. O’Grady, and A. M. Arria, “Distance traveled to outpatient drug treatment and client retention,” *J. Subst. Abuse Treat.*, vol. 25, no. 4, pp. 279–285, Dec. 2003, doi: 10.1016/S0740-5472(03)00188-0.
- [6] H. K. Smith, P. R. Harper, and C. N. Potts, “Bicriteria efficiency/equity hierarchical location models for public service application,” *J. Oper. Res. Soc.*, vol. 64, no. 4, pp. 500–512, 2013, doi: 10.1057/JORS.2012.68.
- [7] K. Hwang, T. B. Asif, and T. Lee, “Choice-driven location-allocation model for healthcare facility location problem,” *Flex. Serv. Manuf. J.*, vol. 34, no. 4, pp. 1040–1065, Dec. 2022, doi: 10.1007/S10696-021-09441-8/FIGURES/7.
- [8] Y. Zhang, O. Berman, and V. Verter, “Incorporating congestion in preventive healthcare facility network design,” *Eur. J. Oper. Res.*, vol. 198, no. 3, pp. 922–935, Nov. 2009, doi: 10.1016/J.EJOR.2008.10.037.
- [9] Y. Zhang, O. Berman, P. Marcotte, and V. Verter, “A bilevel model for preventive healthcare facility network design with congestion,”

- <http://dx.doi.org/10.1080/0740817X.2010.491500>, vol. 42, no. 12, pp. 865–880, Dec. 2010, doi: 10.1080/0740817X.2010.491500.
- [10] P. Mitropoulos, I. Mitropoulos, I. Giannikos, and A. Sissouras, “A biobjective model for the locational planning of hospitals and health centers,” *Health Care Manag. Sci.*, vol. 9, no. 2, pp. 171–179, May 2006, doi: 10.1007/S10729-006-7664-9/METRICS.
- [11] S. Song, Z. Hou, L. Zhang, Q. Meng, and I. Kawachi, “Can equitable distribution of health resources reduce under-five mortality rate? A cross-sectional study with multilevel analysis of rural counties in China,” *Lancet*, vol. 392, p. S58, Oct. 2018, doi: 10.1016/S0140-6736(18)32687-4.
- [12] T. de M. Sathler, J. F. Almeida, S. V. Conceição, L. R. Pinto, and F. C. de Campos, “Integration of Facility Location and Equipment Allocation in Health Care Management,” *Brazilian J. Oper. Prod. Manag.*, vol. 16, no. 3, pp. 513–527, Aug. 2019, doi: 10.14488/BJOPM.2019.V16.N3.A13.
- [13] A. K. Vatsa and S. Jayaswal, “Capacitated multi-period maximal covering location problem with server uncertainty,” *Eur. J. Oper. Res.*, vol. 289, no. 3, pp. 1107–1126, Mar. 2021, doi: 10.1016/J.EJOR.2020.07.061.
- [14] A. Chouksey, A. K. Agrawal, and A. N. Tanksale, “A hierarchical capacitated facility location-allocation model for planning maternal healthcare facilities in India,” *Comput. Ind. Eng.*, vol. 167, p. 107991, May 2022, doi: 10.1016/J.CIE.2022.107991.
- [15] E. J. Delgado, X. Cabezas, C. Martin-Barreiro, V. Leiva, and F. Rojas, “An Equity-Based Optimization Model to Solve the Location Problem for Healthcare Centers Applied to Hospital Beds and COVID-19 Vaccination,” *Math. 2022, Vol. 10, Page 1825*, vol. 10, no. 11, p. 1825, May 2022, doi: 10.3390/MATH10111825.
- [16] R. Rezaee, F. Rahimi, and A. Goli, “Urban Growth and urban need to fair distribution of healthcare service: a case study on Shiraz Metropolitan area,” *BMC Res. Notes*, vol. 14, no. 1, pp. 1–6, Dec. 2021, doi: 10.1186/S13104-021-05490-2/FIGURES/2.
- [17] R. Marín-Navarrete *et al.*, “Motivational enhancement treatment in outpatient addiction centers: A multisite randomized trial,” *Int. J. Clin. Heal. Psychol.*, vol.

- 17, no. 1, pp. 9–19, Jan. 2017, doi: 10.1016/J.IJCHP.2016.05.001.
- [18] R. Marín-Navarrete *et al.*, “Characteristics of a treatment-seeking population in outpatient addiction treatment centers in Mexico,” *Subst. Use Misuse*, vol. 49, no. 13, pp. 1784–1794, Nov. 2014, doi: 10.3109/10826084.2014.931972.
- [19] C. L. Barksdale, M. Azur, and P. J. Leaf, “Differences in mental health service sector utilization among African American and caucasian youth entering systems of care programs,” *J. Behav. Heal. Serv. Res.*, vol. 37, no. 3, pp. 363–373, Jul. 2010, doi: 10.1007/S11414-009-9166-2.
- [20] J. E. Sturges and J. T. Garlick, “Reality Tour: Adult Attendees’ Perceptions about a Community Based Drug Prevention Program,” *Juv. Fam. Court J.*, vol. 69, no. 4, pp. 59–72, Dec. 2018, doi: 10.1111/JFCJ.12121.
- [21] C. Lin *et al.*, “Telemedicine along the cascade of care for substance use disorders during the COVID-19 pandemic in the United States,” *Drug Alcohol Depend.*, vol. 242, p. 109711, Jan. 2023, doi: 10.1016/J.DRUGALCDEP.2022.109711.
- [22] N. Vidyarthi and O. Kuzgunkaya, “The impact of directed choice on the design of preventive healthcare facility network under congestion,” *Health Care Manag. Sci.*, vol. 18, no. 4, pp. 459–474, Dec. 2015, doi: 10.1007/S10729-014-9274-2.
- [23] R. Selten, “The Equity Principle in Economic Behavior,” *Decis. Theory Soc. Ethics*, pp. 289–301, 1978, doi: 10.1007/978-94-009-9838-4_16.
- [24] G. Lau *et al.*, “Prevalence of Alcohol and Other Drug Use in Patients Presenting to Hospital for Violence-Related Injuries: A Systematic Review,” *Trauma. Violence Abuse*, 2023, doi: 10.1177/15248380221150951.
- [25] K. Palomino, D. Garcia, and C. Berdugo, “A MILP facility location model with distance value adjustments for demand fulfillment using Google Maps,” *J. Eng. Res.*, vol. 10, no. 2A, pp. 270–291, Jun. 2022, doi: 10.36909/JER.10473.
- [26] UNDP, “The 2020 Global Multidimensional Poverty Index (MPI),” Jul. 2020. Accessed: Feb. 24, 2023. [Online]. Available: <https://www.undp.org/botswana/publications/2020-global-multidimensional-poverty-index-mpi>
- [27] G. Mavrotas, “Effective implementation of the ϵ -constraint method in Multi-Objective Mathematical Programming problems,” *Appl. Math. Comput.*, vol. 213,

- no. 2, pp. 455–465, Jul. 2009, doi: 10.1016/J.AMC.2009.03.037.
- [28] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics – Challenges in topic discovery, data collection, and data preparation,” *Int. J. Inf. Manage.*, vol. 39, pp. 156–168, Apr. 2018, doi: 10.1016/J.IJINFOMGT.2017.12.002.
- [29] T. Nasrallah, O. El-Gayar, and Y. Wang, “Social media text mining framework for drug abuse: Development and validation study with an opioid crisis case analysis,” *J. Med. Internet Res.*, vol. 22, no. 8, p. e18350, Aug. 2020, doi: 10.2196/18350.
- [30] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent Dirichlet Allocation Michael I. Jordan,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [31] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent Dirichlet Allocation,” 2003. doi: 10.5555/944919.944937.
- [32] DANE, “Departamento Administrativo Nacional de Estadística. Censo Nacional de Población y Vivienda 2018. Proyecciones de Población 2018-2020, total municipal por área Junio 30.” Bogotá D.C, Colombia, 2018.
- [33] DANE, “Producto Interno Bruto por departamento,” 2021.
- [34] DANE, “Encuesta Nacional de Consumo de Sustancias Psicoactivas en Población General 2019,” 2020. https://microdatos.dane.gov.co/index.php/catalog/680/get_microdata (accessed Jan. 14, 2022).
- [35] Google LLC, “Distance Matrix API: Developer Guide,” *Google Maps Platform*, 2017.
- [36] D. Rico, J. Gonzalo, D. Wiesner, and L. Goyeneche, “Informe del gasto del gobierno de Colombia en lucha antidrogas,” Bogotá D.C, 2016. [Online]. Available: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/3609/Repor_Mayo_2018_Fedesarrollo_y_FIP.pdf?sequence=4&isAllowed=y
- [37] EITI, “Distribución y Ejecución de los Recursos del SGR,” Bogotá , 2019. Accessed: Apr. 28, 2023. [Online]. Available: <https://www.eiticolombia.gov.co/es/informes-eiti/informe-2018/distribucion-y-seguimiento-de-ingresos/distribucion-y-ejecucion-de-los-recursos-del-sgr/>

- [38] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, “More than a Feeling: Accuracy and Application of Sentiment Analysis,” *Int. J. Res. Mark.*, vol. 40, no. 1, pp. 75–87, 2023, doi: <https://doi.org/10.1016/j.ijresmar.2022.05.005>.
- [39] S. Materazzi, G. Peluso, L. Ripani, and R. Risoluti, “High-throughput prediction of AKB48 in emerging illicit products by NIR spectroscopy and chemometrics,” *Microchem. J.*, vol. 134, pp. 277–283, Sep. 2017, doi: [10.1016/j.microc.2017.06.014](https://doi.org/10.1016/j.microc.2017.06.014).
- [40] C. J. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 8, no. 1, pp. 216–225, May 2014, doi: [10.1609/ICWSM.V8I1.14550](https://doi.org/10.1609/ICWSM.V8I1.14550).
- [41] J. Cañete, G. Chaperon, R. Fuentes, and J. Pérez, “Spanish pre-trained BERT model and evaluation data,” PML4DC at ICLR 2020, 2020. Accessed: Apr. 29, 2023. [Online]. Available: https://pml4dc.github.io/iclr2020/program/pml4dc_10.html
- [42] A. G. Fandiño *et al.*, “MarIA: Spanish Language Models,” *Proces. del Leng. Nat.*, vol. 68, 2022, doi: [10.26342/2022-68-3](https://doi.org/10.26342/2022-68-3).
- [43] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, “Optimizing Semantic Coherence in Topic Models,” *Proc. 2011 Conf. Empir. Methods Nat. Lang. Process.*, pp. 262–272, 2011.
- [44] F. Jelinek, “Interpolated estimation of Markov source parameters from sparse data,” 1980.

Chapter 5

5. Data Analytics and Mental Health: Would Ethics Be the Only Safeguard Against the Risks of Identifying "Potential Patients"?

5.1. Abstract

Despite all the prospects for data growth, sharing, and processing, and all the benefits that data can bring, this revolution is not exempt from risks. Even if, at some point, computers may be able to provide diagnoses with greater accuracy than medical professionals, would ethics be the only safeguard against the possible risks? In this article, we implement a pragmatic approach to answer this question by focusing on possible outcomes concerning "potential patients". We define a "potential patient" as an individual who has not yet shown obvious or early signs of disease. Through the outcomes and inferences derived from data analysis, such a patient could potentially require early treatment, more accurate diagnoses, and medications that are better adapted to the patient's conditions.

5.2. Introduction

Mental illness is ranked as the fifth-largest contributor to the worldwide disease burden. An estimation of the economic cost of mental illness was \$2.5 trillion in 2010 and has been estimated to double by 2030 [1]. One out of every four people in the developed world has a mental illness [2]. A key objective of the World Health Organization's Comprehensive Mental Health Action Plan 2013-2030 is to improve mental health information systems, which involves expanding resource capacity for health monitoring [1]. Although some mental health conditions are both preventable and highly responsive to treatment, mental health-related conditions can often be characterized due to delays in seeking help, weak medication adherence, and stigmatized provision of care [2]. Predicting or assessing the risk of treatment response at the individual patient level

continues to be a challenging goal for some conditions, especially mental disorders. For example, selecting an antipsychotic drug is usually a trial-and-error process. Clinical randomized trials and meta-analyses have become cornerstones of evidence-based practice. They have helped find effective treatments for particular disorders by using traditional statistical approaches [3].

Data analytics is a term that denotes large and complex measurement volumes, as well as the speed at which data is generated. An additional important feature of data analytics is the wide range of levels at which data is collected and processed, from the molecular level to medicinal, sociodemographic, managerial, and even social network information [2]. In the medical field, data analytics opens up a whole new field in which there is a shift from group-level evidence, as advocated by evidence-based medicine, to personalized assessments, in particular when it comes to mental health treatment. Furthermore, advances in data analytics help specialists make better diagnoses, develop treatments and prescribe medications adapted to the patient's conditions. Despite all the prospects for data growth, sharing and processing, and all the benefits that data analytics can bring, this revolution is subject to risks [4]. In this regard, it is staked that opportunities come with challenges and risks, such as unethical collection and use of health data; algorithmic bias; and risks to patient safety.

Predictive data analytics models pave the way not only for preventing unfavorable outcomes through targeted strategies for prompt interventions but also for efforts to prevent conversions to disorders. In the midst of these new advances, practitioners find themselves to be the bridge between patient and data outcomes, regularly dealing with patients' demands and technological advancements to deliver the proper standard of care. From the juxtaposition of these developments arises ethical issues that technology alone cannot address. Prediction, on its own, might not be a problem, but the application could be. Even if, at some point, computers may be able to provide diagnoses with greater accuracy than medical professionals, would ethics be the only safeguard against ensuing liability risks? In this article, we implement a pragmatic approach to answer this question by focusing on possible outcomes concerning "potential patients". We define a "potential patient" as an individual who has not yet shown obvious or early signs of disease. Through the outcomes and inferences derived from data analysis, such a patient could potentially

require early treatment, more accurate diagnoses, and medications that are better adapted to the patient's conditions.

5.3. Data Analytics in Mental Healthcare

It is difficult to imagine a human activity these days that does not produce data, considering how close we are connected to electronic devices and, consequently, how interconnected we are with one another. Our actions generate data that can allow others to identify our behavioral patterns and expose our personal information [5]. In this regard, some considerations emerge from the collection and use of this large data flow. Regulation is still developing, several questions are open for debate, and we are possibly looking at some conflicting risks. One, that the data could be mishandled and generate adverse consequences for patients and society; or two, that the awareness of this risk may drive over-regulation that could slow down the benefits of data analytics.

5.3.1. Some Emerging Trends

The promise of data analytics in medicine is quite clear: it can foresee epidemics, diagnose diseases, provide medical insights, improve quality of life, and prevent preventable mortality. Similarly, current trends in applying data analytics in mental health include medication selection, suicide prediction, and symptom/outcome monitoring.

5.3.2. Medication Selection

Although there have been substantial advances in psychiatric research in psychopharmacology, finding the most appropriate medication for a patient today can be a matter of trial and error. Even though clinical judgment is essential in such situations, data analytics provides new information to either the patient or the psychiatrist that can augment drug treatment selection. Just to give some examples, different variables or predictors that range from genetic or molecular characteristics to demographic and social factors may be related to the improved results of one treatment compared to another. Patient-reported data from patients with depression from level 1 of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) was utilized to provide a data-driven method to identify helpful predictor variables within a large number of candidate predictors and to combine them for individual predictions [6]. Also, a recent study used

patient data to develop a predictive model to provide an accurate estimate of patient remission from the antidepressant citalopram, a serotonin reuptake inhibitor. The study aimed to determine whether patients will benefit from a specific treatment based on their clinical history and symptoms [7].

5.3.3. Suicide Prevention

Some other recent proposals include the development of an automatic detection system based on data analytics to improve drug safety monitoring in a hospital and help pharmaceutical surveillance professionals make specific hypotheses on harmful drug events resulting from drug-drug interactions (DDIs) [8]; and a clinical tool capable of predicting responses to risperidone in the first episode of previously untreated schizophrenia patients with an accuracy of 82.5% by analyzing multivariate functional magnetic resonance imaging [9]. Some interesting applications of data analytics offer the prospect of improving suicide prevention by expanding monitoring beyond the clinic. Both logistic regression analysis and survival analysis have been applied to construct the most empirically effective suicide prediction tools to date. In some of these tools, the suicide risk predictors involved consisted entirely of sociodemographic and clinical information collected from medical records [10]; but others have also included some patient-reported scales, such as self-report scales of hopelessness, depression, general psychopathological severity, and attitudes toward suicide [11]. Medical records also represent a useful source of data for machine learning in the field of suicide prevention. Some studies attempt to predict future suicide attempts among patients in the healthcare system [12]–[14].

5.3.4. Outcome Monitoring and Treatment

Furthermore, data analytics can also help guide symptom/outcome monitoring and treatment. Despite the availability of a vast arsenal of antidepressant medication, not all patients with major depressive disorder respond adequately to treatment. For example, a large multicenter study identified demographic and medical factors associated with treatment-resistant depression (TRD) in a representative sample of patients with severe depressive episodes who experienced failure to achieve either treatment response or remission after a minimum of two consecutive antidepressant treatments [15]. Other

studies have associated prolonged duration and severity of existing major depressive episodes (MDEs) and the presence of adverse effects during treatment with TRD [4], [16]–[18].

Finally, it is worth mentioning that current neuroscience imaging techniques significantly impact the precision of diagnosis and prognosis and help us progress toward cures for various mental illnesses. The potential benefits strongly depend on the practical aspects by which large-scale clinical and imaging data can be integrated and examined. A notable study focuses on patient identification using data analytics where a reference identification indicator, referred to as the Alzheimer's Disease Identification Number (ADIN), is calculated using fuzzy processing. This number allows the classification of patients for a particular progression of Alzheimer's disease, estimation of the short-term progression of the disease, and the context for other patients, including adequate treatment, expected life expectancy, Etc. [19].

5.3.5. Ethical Data Usage

There are certain ethical and social aspects of data analytics that apply to medicine, one of which is the importance of privacy and anonymity. Although terms of service generally mention privacy and guarantee the anonymity of personal data, these terms can sometimes be vague in their description. Therefore, there are risks in the use of this anonymized data. Tracing the information back to a particular individual (re-identification process) through reverse engineering techniques may be possible. This aspect is of critical importance in the medical field, as data related to health may contain sensitive information about patients, such as their sexual preference, previous history of abortions, suicide attempts, etc.

Keeping data private can also affect healthcare professionals' work practices and legal responsibilities. In addition, patients can be vulnerable because of their understanding of their diagnosis or because of concerns about treatment; disclosing their information to unauthorized parties can further complicate how they experience their conditions or how they respond to treatment. Data collection must also be transparent and describe what information is being collected, and clearly indicate the potential uses of that data. In other words, when a patient shares their data, it is essential that they are aware of the ethical

principles of the institution in charge of the collection, namely what is planned to be done with the information and what is forbidden. All the techniques that encompass what is known as data analytics are a fundamental part of the transformation of multiple communication facets, which can be seen in the growing links between technology, communication, and health. Another relevant ethical aspect regarding data analytics is using suitable data sets. Using incomplete, biased, or out-of-context analysis of medical data could produce inaccurate results, and these conclusions could lead to detrimental actions or decisions. Last but not least is the question of ownership, it is a concern that an organization may own patients' personal information, as well as their preferences and behavior, and use it to influence future decisions and preferences. However, the boundaries are also unclear here, as it is necessary to establish which data may be public, and which should remain private. Medical data are not necessarily private property as they are co-constructed, meaning that a proprietary account must not necessarily confer unique rights to the patient.

5.4. Distinctive Ethical Questions

5.4.1. How Devastating Could a Risk Assessment be for a Medical Patient?

One should be concerned about how devastating a risk assessment could be for a medical patient. For example, some mental health disorders can be predicted by a simple genetic test. A positive result can tell a patient that, in the next few years, they will experience progressive and severe brain function loss while still healthy. It is possible that the stressful burden of knowing may result in accelerating the disorder or even lead to another one, such as a subsequent depressive episode or drug abuse, in the coming years prior to the start of the condition being predicted. Even though data analytics can have considerable potential benefits and significant social implications for preventing negative health outcomes, it is arguable that there is no such thing as 100% accurate risk assessment and that the patient may have the freedom to choose how the results will affect their life and decisions. However, from other legal or social perspectives, hiring or enlisting for military service a person with a high probability of developing a mental health disorder may cause health care and pension expenses. Not to mention, if the symptoms of a disorder first show during their daily activities, there is a risk that they could jeopardize both their safety and the safety of others. This question, then, is of

paramount importance in the ethics of data analytics and mental health - how can medical patients deal with such risk assessment about their prospective future and carefully consider the harms and gains of their use?

5.4.2. Can Mental Health Risk Assessments Affect Insurance Costs?

As previously mentioned, predictive mental health care can improve outcomes and help prevent serious disability or injury. However, it can also cause significant problems, impacting other aspects of health beyond patients' health. Even if predicting future mental health problems may not be a concern, the actual implementation of treatment after that may be. If told that they are on a course toward mental illness, a patient would likely use health services more and therefore be charged more for a health plan. How attractive is this to private health companies? To what extent can mental health risk assessments impact insurance costs? Knowing that a patient is highly likely to develop a mental illness can potentially lead to early detection of symptoms of the disorder and allow for prompt intervention. It is possible that the patient may start attending an outpatient mental health center before the onset of the disorder or before the most noticeable symptoms occur. Also, the individual may need familial and professional support during this prodromal stage. Once again, there is a possibility that they will not develop the illness because the risk assessment is not accurate, and the patient may go through the whole treatment experience unnecessarily, incurring expenses accordingly.

5.4.3. How Should Social Media be Used to Promote Mental Health?

Psychiatrists have long been examining human behavior, trying to describe and treat pathological alterations in human behavior. In this context, the data analytics advances of recent years may help psychiatry develop more accurate and objective measurement tools. By examining individuals' data, from behavior to social activity, signs of depression, anxiety, bipolar disorder, and other syndromes may be detected. Even if the data collected for analysis has been made public, is the individual aware that their information may be used to evaluate their health? How should we obtain consent? It is also worth mentioning that a patient's results may label them as targets for eugenic policies or even social prejudice. From an ethical standpoint, there is concern that the use of data blurs the boundaries about what should be classified, regulated, and protected as medical information. Processing this type of data could reveal more about a patient's mental state

to outsiders than what is in their confidential medical records. Moreover, that is difficult to hide since practically any collected data can become informed about a patient's state.

5.4.4. How Much Informed Consent Do Doctors Require for Data Analytics Work?

Informed consent for medical practices and health research is the process where the patient expresses their will based on clear, precise, and adequate information about some salient points, proposed procedure, expected benefits and risks. Informed consent in mental health research is often intended to indicate what kind of data will be collected and for what purpose. This presents a challenge because one of the aims of data analysis is to bring out new insights or patterns from the data, and it is possible that some of these may not be within the original statement of consent. Therefore, including the unpredictable in informed consent is a real challenge. In addition, it is also worth considering that the data may be of more benefit if it is shared, integrated, and reanalyzed among diverse groups of patients. This increases the complexity of giving consent but provides even more helpful medical information. In other words, informed consent was designed to address issues relevant to evidence-based medicine, with predetermined questions and a limited number of expected outcomes. We can consider two critical questions. Firstly, what is the best way to adapt this consent to the new reality offered by data analytics? And secondly, how would such adaptations impact patients?

5.5. Ethics, the Only Safeguard?

As the use of data analytics in healthcare becomes more widespread, it is important for healthcare professionals and patients alike to be aware of these issues and to work together to ensure that data is used in a responsible and ethical manner. While there are still many questions to be answered and challenges to be overcome, it is clear that data analytics will play a significant role in shaping the future of healthcare. How devastating could a risk assessment be for a medical patient? To what extent can mental health risk assessments impact insurance costs? How should social media be used to promote mental health? What degree of informed consent do doctors need when working with data analytics? Even though most of the techniques and models in current data analytics studies are still in concept testing stages so far, it is likely that patients and healthcare professionals will need to deal with these questions in the foreseeable future.

5.5.1. Balancing Precision and Ethics in Data Analytics

The technology still depends on massive amounts of data input. Mental disorders should be defined precisely for the machines to interpret accurately, and if the incoming data is imprecise or incorrect, the output might be useless. At some point, computers may be able to provide diagnoses with greater accuracy than medical professionals, as some algorithms are already achieving greater accuracy with machine learning to diagnose certain conditions. In this little-known area, there is no clear policy or guidelines on how to proceed. It is possible that medical guidelines should address the problem of "potential" patients who do not show symptoms at the moment of diagnosis. This power to define "who is ill" and "who is well" needs a robust system of ethics. Then, should ethics be the only safeguard? And if so, which system of ethics? When considering ethical issues about data, the critical aspect is to remember that the techniques and treatment of the results have been designed and programmed by human beings. Therefore, it is necessary to note that it is not just about ethical principles but also a redefinition of responsibility and developing a solid institutional framework.

Regarding ethical principles, we should first remember that some of our cherished values today may/might evolve as we face new challenges. Ethical policies regarding data analytics should not be limited to rigid, absolute principles. Instead, it should be flexible and able to adapt to new challenges and unanticipated outcomes. In relation to medical ethics, today, two main schools of thought exist: deontological and utilitarian. According to the deontological approach, the outcomes may not just justify the methods that are used to achieve them. By contrast, utilitarians believe that the outcomes justify the means. Those outcomes seek the maximum expected benefits for the maximum number of people. Unlike utilitarians, deontologists believe that an ethical justification of each action hinges upon the action itself. In other words, killing a few people to develop a cure for a disease for which millions suffer is ethically justified, according to utilitarians but not deontologists.

At the risk of oversimplification, deontology tends to focus on the patient, while utilitarianism tends to focus on society. However, we contend that both of these ethical schools raise valid points and that a balanced midway position is needed by the healthcare system to achieve optimal mental health practices. Deontological advocates are often

driven to the utilitarian approach by public medical professionals, managers inside the health sector, politicians, and third-party payment systems. Taking a utilitarian perspective, the resources of the healthcare system are finite and must be properly managed to achieve the best service for society. As long as the optimal good is achieved for the greatest number with the resources available, any possible iatrogenic harm or loss of confidentiality between the doctor-patient relationships that data analytics policies might cause should, in our view, be considered acceptable. In today's scenario, the utilitarian perspective is seen to counterbalance the deontological perspective and, therefore, most of the ethical and moral dilemmas related to data analytics. Hopefully, balancing these two approaches might bring greater justice and fairness to the practice of mental health.

5.5.2. Who Is Responsible for Considering Treatment Decisions and Ethics?

Focusing on the responsibility aspect, it is a fact that data analytics will have an impact on important decisions in the mental healthcare industry. Of course, the risk may be mitigated in a medical context: potential patients, as long as they are clearly informed, will have some degree of risk or responsibility if they rely on the results expected from the data analysis. A patient with a mental disorder may be open to trying medication because it may bring hope that, even if he or she does not directly benefit from it, their trial might help other patients in the future. However, consider a case where a patient is expected to attempt suicide. Should he or she remain in the hospital for a longer period of time or return home and have their family monitor and regularly follow-up with appointments? Who is responsible? In addition, if this person lives alone, should they also receive in-home monitoring? On the one hand, this prediction may provide a better allocation of resources or treatment to those expected to commit suicide, but it may also result in the neglect of those who are not expected to be at risk of committing suicide. Exactly what responsibility should the patient, the healthcare professional, the hospital, or the data analysts have in these situations?

In any medical responsibility issue, there must be an acceptable compromise between risk and benefit. Not all risks are totally unacceptable, and while it is also the function of the ethics committee to ensure that risks are reduced as much as possible, one could argue that it is the responsibility of all potential participants to decide whether or not they want

to accept such risks. It also stands to reason that the greater the probability that analysis will have an impact, the more likely it is that there will be a risk associated with it. This is more difficult when critical decisions are made based on data processing results. If patients are capable of consenting willingly, and if the scenario where benefits have been justified and risk has been reduced to a minimum, ethics committees must learn to accept this; otherwise, they will fail to demonstrate respect for the patients themselves.

5.5.3. Ethical Imperatives in Healthcare Institutions

Responsibility must be supported by autonomy, beneficence, and fairness. Firstly, autonomy and the sense of respect for potential patients raise questions concerning consent, such as: Is the patient capable of giving consent? Have they fully understood what they are consenting to? Can they withdraw their consent at any time? Etc. Secondly, when considering beneficence, defining the term as the act of doing good, the potential benefit and risks must be clear, and questions such as what the interests of the persons or parties involved are, who might be harmed and how, or who might benefit, need to be answered. When looking at justice, it is necessary to discuss how risks and benefits are shared with patients, as well as the eventual impact of data analytics results on a broader spectrum. When an ethics committee focuses too much on the risks, it leads to a completely new type of risk: excessive regulation of technology applications which will limit their own space for innovation —this is not only detrimental to the medical sector but also to society.

Finally, ethics has to go hand in hand with a solid institutional framework. Detailed regulations are often unwieldy and hard to fully implement. However, general regulations may fail to provide sufficient guidelines about how to respond to a specific medical problem. In this way, ethics cannot be avoided. If people expect healthcare institutions or companies in the sector to do the right thing, it is first necessary to define what the "right" thing is. A problem is there is often more than one "right" thing. What is "right" according to insurance companies often differs from what is "right" according to optimal health care specialists. Many different competing ethical imperatives exist. For example: do the least harm, make the most financial profit, do whatever is the prevailing norm, Etc. When people think of medical institutions, they think of institutions with a particular function

in the community. Such institutions can promote positive social values, and healthcare professionals can promote of such values within their institutions.

How can medical institutions promote value in various data analytics scenarios? Part of the answer is that value must be promoted by respecting patients, and therefore the institutions that work are institutions that allow people to trust each other, people must trust that society will treat them as rights holders. The core question is "value for whom?" Although it sounds cynical, from a purely financial perspective, "maximum value" for health care providers might extend an illness as long as possible to obtain the most financial remuneration. This hard reality should not be overlooked. In short, economic interests clash. Another part is to be found in society itself. Medical institutions must trust people to use their possibilities as rights holders according to national standards. These institutions do not try to take on the job of promoting the good moral but rather provide a solid framework in which the good is respected because that is the framework in which healthcare professionals and patients coexist and cooperate.

5.5.4. Discussion and Conclusion

It is evident that the ideas presented in this article are interrelated and have significant implications for the topic at hand. Firstly, data analytics not only help us to understand mental illness better but also improve the delivery of psychiatric healthcare for potential patients. Recently, researchers have been publishing several studies that show the potential of these methods in improving mental health, preventing suicide, assigning the correct medication, enhancing the efficiency of healthcare, and also monitoring patients. Considering the emergence of data analytics being applied to healthcare, it is no surprise that the role and potential of the field are still being explored. However, caution is also warranted despite the emerging achievements and ongoing prospects of data analytics for mental health. Retrospective patient record reviews have become particularly popular because of their capacity to simultaneously simplify and standardize medicine in order to obtain more accurate results. Thus, a tricky balance must be achieved for effective and responsible mental health practices, and ethics seem to be the only safeguard.

Secondly, when the question of "Is ethics the only safeguard?" was addressed, it was necessary to point out that it is not only a question of ethical principles but also of

redefining responsibilities and establishing a solid institutional framework. current ethical regulation privileges consent above all other ethical aspects and have been depicted as a way to counter the possible risk of autocratic practices. We do not advocate that consent must be bypassed but that regulatory processes should allow for a proper discussion of consent in an ethical context. That is, we should consider consent as fluid and evolving. The current need for regulations on managing and using data can lead to unforeseeable harm to individuals and society as a whole. If society is swayed by concerns regarding this technology, overreaction and preventive barriers can occur, resulting in a restrictive and over-regulated policy that can impede not only the treatment of potential patients but also the advancements and benefits that could improve healthcare. The ability to access data at a large scale is not enough. What we gain from the large volume of data and coverage could be offset by misinformed interpretation and analysis or collection for purposes other than patients' health. Health records, as may happen with any bureaucratic records, could be shaped by administrative convenience more than by the search for medical evidence.

Lastly, involving patients with mental health conditions often causes concerns for ethics committees. A typical reason for this is adopting a protectionist position, as these patients are generally considered a vulnerable population that is more likely to suffer through participation. Although some degree of risk is unavoidable, the requirement that patients be adequately informed of the exact magnitude and level of risk prior to giving consent makes it far more likely that every effort will be made to reduce risk to a minimum and maximize benefit. In this regard, consent is of critical relevance and cannot be overstated; it is at the core of respect for medical patients. People with health conditions might experience consent-based vulnerability because they cannot voice their autonomy or experiences based on equity. A mental health diagnosis makes them part of a group that lacks appropriate opportunities or freedoms, which also makes them prone to be coerced. As much as an ethics committee may disagree about the outcomes or inferences of a particular medical study, it would be difficult to dispute the importance of the principle of respect for patients. The mere awareness of the risk of harm may reduce the occurrence of harm, besides increasing the ability of practitioners to cope positively with it.

The future of data analytics is fraught with challenges and opportunities that will require creative and innovative solutions to overcome. One of the significant ethical challenges

is ensuring that the algorithms used to analyze patient data are unbiased and do not discriminate against certain groups of patients. This can result in certain groups of people being underrepresented or overrepresented in identifying potential patients, leading to inequities in healthcare. Another challenge is the potential for making decisions that are not aligned with human values. AI algorithms are trained on data, which means that they are designed to optimize for specific outcomes. This can lead to situations where AI algorithms make technically correct but morally wrong decisions. Despite the challenges, identifying potential patients also presents significant opportunities to improve the accuracy and consistency of diagnoses. Human healthcare providers are fallible and subject to biases, which can lead to misdiagnoses and inappropriate treatments. However, AI algorithms can analyze large amounts of data and identify patterns humans may miss, leading to more accurate diagnoses and effective treatments.

References

- [1] M. Conway and D. O'Connor, "Social media, big data, and mental health: Current advances and ethical implications," *Current Opinion in Psychology*, vol. 9. ELSEVIER SCIENCE BV, PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS, pp. 77–82, Jun. 2016. doi: 10.1016/j.copsyc.2016.01.004.
- [2] I. Passos, P. Ballester, J. Pinto, B. Mwangi, and F. Kapczinski, "Big Data and Machine Learning Meet the Health Sciences: Big Data Analytics in Mental Health," in *Personalized Psychiatry: Big Data Analytics in Mental Health*, 2019, pp. 1–13. doi: 10.1007/978-3-030-03553-2_1.
- [3] A. Pavlova and P. Berkers, "Mental health discourse and social media: Which mechanisms of cultural power drive discourse on Twitter," *Soc. Sci. Med.*, vol. 263, p. 113250, 2020, doi: <https://doi.org/10.1016/j.socscimed.2020.113250>.
- [4] M. Balestri *et al.*, "Socio-demographic and clinical predictors of treatment resistant depression: A prospective European multicenter study.," *J. Affect. Disord.*, vol. 189, pp. 224–232, Jan. 2016, doi: 10.1016/j.jad.2015.09.033.
- [5] A. L. Beam and I. S. Kohane, "Big Data and Machine Learning in Health Care," *JAMA*, vol. 319, no. 13, pp. 1317–1318, Apr. 2018, doi: 10.1001/jama.2017.18391.
- [6] A. M. Chekroud *et al.*, "Cross-trial prediction of treatment outcome in depression: a machine learning approach.," *The lancet. Psychiatry*, vol. 3, no. 3, pp. 243–250, Mar. 2016, doi: 10.1016/S2215-0366(15)00471-X.

- [7] A. M. Chekroud, R. Gueorguieva, H. M. Krumholz, M. H. Trivedi, J. H. Krystal, and G. McCarthy, “Reevaluating the Efficacy and Predictability of Antidepressant Treatments: A Symptom Clustering Approach,” *JAMA psychiatry*, vol. 74, no. 4, pp. 370–378, Apr. 2017, doi: 10.1001/jamapsychiatry.2017.0025.
- [8] G. Bouzillé *et al.*, “An Automated Detection System of Drug-Drug Interactions from Electronic Patient Records Using Big Data Analytics,” *Stud. Health Technol. Inform.*, vol. 264, pp. 45–49, Aug. 2019, doi: 10.3233/SHTI190180.
- [9] B. Cao *et al.*, “Treatment response prediction and individualized identification of first-episode drug-naïve schizophrenia using brain functional connectivity,” *Mol. Psychiatry*, vol. 25, no. 4, pp. 906–913, 2020, doi: 10.1038/s41380-018-0106-5.
- [10] M. J. Spittal, J. Pirkis, M. Miller, G. Carter, and D. M. Studdert, “The Repeated Episodes of Self-Harm (RESH) score: A tool for predicting risk of future episodes of self-harm by hospital patients,” *J. Affect. Disord.*, vol. 161, pp. 36–42, 2014, doi: <https://doi.org/10.1016/j.jad.2014.02.032>.
- [11] K. Bilén, S. Ponzer, C. Ottosson, M. Castrén, and H. Pettersson, “Deliberate self-harm patients in the emergency department: who will repeat and who will not? Validation and development of clinical decision rules,” *Emerg. Med. J.*, vol. 30, no. 8, pp. 650 LP – 656, Aug. 2013, doi: 10.1136/emered-2012-201235.
- [12] S. B. Choi, W. Lee, J.-H. Yoon, J.-U. Won, and D. W. Kim, “Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea,” *J. Affect. Disord.*, vol. 231, pp. 8–14, 2018, doi: <https://doi.org/10.1016/j.jad.2018.01.019>.
- [13] G. E. Simon *et al.*, “Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records,” *Am. J. Psychiatry*, vol. 175, no. 10, pp. 951–960, Oct. 2018, doi: 10.1176/appi.ajp.2018.17101167.
- [14] S.-E. Cho, Z. W. Geem, and K.-S. Na, “Prediction of suicide among 372,813 individuals under medical check-up,” *J. Psychiatr. Res.*, vol. 131, pp. 9–14, Dec. 2020, doi: 10.1016/j.jpsychires.2020.08.035.
- [15] D. Souery *et al.*, “Clinical factors associated with treatment resistance in major depressive disorder: results from a European multicenter study,” *J. Clin. Psychiatry*, vol. 68, no. 7, pp. 1062–1070, Jul. 2007, doi: 10.4088/jcp.v68n0713.
- [16] A. Kautzky *et al.*, “A New Prediction Model for Evaluating Treatment-Resistant Depression,” *J. Clin. Psychiatry*, vol. 78, no. 2, pp. 215–222, Feb. 2017, doi: 10.4088/JCP.15m10381.

- [17] R. Musil *et al.*, “Subtypes of depression and their overlap in a naturalistic inpatient sample of major depressive disorder.,” *Int. J. Methods Psychiatr. Res.*, vol. 27, no. 1, Mar. 2018, doi: 10.1002/mpr.1569.
- [18] A. Tomlinson *et al.*, “Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol,” *Evid. Based Ment. Heal.*, vol. 23, no. 2, pp. 52 LP – 56, May 2020, doi: 10.1136/ebmental-2019-300118.
- [19] K. Munir, A. de Ramón-Fernández, S. Iqbal, and N. Javaid, “Neuroscience patient identification using big data and fuzzy logic—An Alzheimer’s disease case study,” *Expert Syst. Appl.*, vol. 136, pp. 410–425, 2019, doi: <https://doi.org/10.1016/j.eswa.2019.06.049>.

Chapter 6

6. Conclusions and Further Directions

This chapter summarizes the most important conclusions derived from the main questions posed in the research framework. Additionally, it discusses both theoretical and practical implications resulting from the study, presents the encountered limitations and offers recommendations for future research in this field.

6.1. Conclusions

A variety of conclusions and inferences arose from this research. Firstly, regarding the bibliometric analysis of statistical techniques and machine learning in psychoactive substance use, this study revealed significant growth in annual articles over the past two decades. The United States, China, Brazil, and India were identified as the most significant contributors to publications in this area. The study also identified the most prolific authors and journals and noted an increase in the use of advanced technological methods, particularly Bayesian techniques. While spatial analysis papers were limited, machine learning and multivariate or univariate statistical analysis were found to remain critical components of understanding psychoactive substances. Overall, the study provides solid evidence of the increasing acceptance and validation of psychoactive substance use research in broader geographical regions, research fields, and periods. The findings of this bibliometric analysis can be useful for researchers in identifying potential collaborators, journals, and future research directions.

In addition, the bibliometric analysis examined publications related to Bayesian analysis, multivariate or univariate statistical analysis, spatial analysis, and machine learning techniques in the context of psychoactive substance use, risks and patterns, and prevalence, mainly in adolescents. The results showed that the majority of publications related to Bayesian analysis were associated with epidemiological models. On the other hand, publications related to multivariate or univariate statistical analysis were mainly

associated with predictive models and techniques involving drug use and abuse, addiction, and prevalence in adolescents. Spatial analysis research was mostly conducted on psychoactive substance use, risks and patterns, and the presence of crime, violence, and child abuse in urban or suburban environments. Machine learning techniques were commonly used in these studies, including social media monitoring, natural language processing, random forest, and deep learning. The results suggest that there is a growing global concern surrounding psychoactive substance use (PSU), and while there is a lot of research on this topic, there is a noticeable gap in the literature regarding the application of spatial analysis. Implementing a GEOTELO framework that employs geostatistical modeling and incorporates language and location models could address this deficiency and provide more sophisticated insights into PSU patterns and associated risks, particularly in urban and suburban areas. This could also help mitigate the methodological heterogeneity in existing research, generating more consistent and reliable findings, and ultimately enhancing the generalizability of the results. Developing a GEOTELO-based intervention framework could revolutionize the field and provide new, data-driven approaches to tackle this complex issue, making it a valuable tool for researchers and policymakers.

Concerning the design of an ensemble model that integrates an autoencoder with clustering and spatial models to find sociodemographic and spatial patterns of georeferenced data, We developed and tested a Deep Neural Network-based Clustering-oriented Embedding algorithm to identify patterns of psychoactive substance (PAS) use and abuse in Colombia. Our model automatically extracts features from input data, such as sex, age, socioeconomic status, and housing type, to determine whether an individual has consumed PAS. After the training process, we generated a latent feature space (LFS) and analyzed the results. Our findings show clearly marked clusters where the prevalence of individuals who use or do not use PAS is notable. We also found that region, sex, housing type, socioeconomic strata, age, and whether individuals contribute to household finances have a statistically significant impact on the clustering structure. Our proposed model, CAE-DEC, performed better than the CAE-Spectral model based on several metrics, including the Silhouette statistic, Calinski-Harabasz index, and Davies-Bouldin index. We identified three distinct clusters, where individuals who are more likely to consume PAS are grouped in cluster 2, while cluster 1 consisted of individuals who did not consume PAS.

Females, those in socioeconomic strata 1, and those 40 years old or older without economic contribution to the household predominantly characterize cluster 1. On the other hand, cluster 2 is characterized by a higher proportion of males aged between 20 and 40 in socioeconomic strata 1 and 2 who do not contribute to the household finances. Finally, cluster 0 is characterized by a small proportion of males, a higher proportion of individuals in strata 3, 4, 5, and 6, and individuals more likely to contribute to the household economy. Our study also identified that legal drugs, such as alcohol, have a high prevalence in all regions of Colombia, with a slight tendency to more consumption in coastal areas. Coastal areas have a higher demand for alcoholic beverages due to tourism, fishing, and maritime culture. The Northern region has lower consumption of illegal drugs but a higher proportion of non-prescription tranquilizers, opioids, ketamine, GHB, and heroin. The Central region has a diverse consumption pattern, with the largest cities in Colombia, and a higher proportion of tobacco use. The Eastern region has a higher consumption of energy drinks and a diverse consumption pattern, with a prevalence of heroin, basuco, non-prescription tranquilizers, methamphetamines, opioids, and ketamine.

Besides, the research findings suggest that the Southern region of Colombia has a higher likelihood of consuming illegal drugs, such as basuco, heroin, and Yagé. This is due to the region's favorable environmental characteristics for drug consumption and production, as it is the second largest illegal drug-producing region in Colombia. Additionally, the region's high percentage of rurality (52%) and low level of development, as measured by GDP, make it more susceptible to drug use. In contrast, the Western region, also known as the Pacific region, is characterized by geographical isolation, poverty, and ongoing conflict, contributing to the growth of drug production and trafficking in the area. This region has a similar percentage of urban (53%) and rural (47%) populations compared to the Southern region and ranks second among the regions with the lowest levels of development. Consumption in this region mainly includes methylene chloride, GHB, heroin, opioids, and methamphetamines. Poverty is one of the primary factors driving drug production in the Pacific region, leading many people to turn to drug cultivation and trafficking for survival. The region's rugged terrain and limited infrastructure have also made it challenging for the Colombian government to establish a strong presence, allowing drug traffickers to operate with relative impunity.

With respect to the design of an optimization model to identify intervention points, integrating location-allocation models and Twitter topic modeling enhances resource allocation and provides insights into drug consumption patterns and public sentiment. This comprehensive approach supports evidence-based decision-making, improves intervention strategies, and aims to reduce drug consumption and associated harms.

Last but not least, it is relevant to mention that the application of data analytics in the field of mental health has showcased significant potential to revolutionize healthcare by providing better understanding, prevention, and treatment options for mental illnesses. The recent studies discussed in this article have demonstrated the benefits of data analytics in various aspects of mental health. Nevertheless, it is important to strike a delicate balance between leveraging the power of data analytics and upholding ethical standards to ensure responsible mental health practices. Ethics play a crucial role in safeguarding against potential misuse and abuse of sensitive patient data. While the current ethical regulations place significant emphasis on consent, it is important to consider consent as a fluid and evolving concept that must be discussed in a broader ethical context. Overemphasis on consent may lead to restrictive and over-regulated policies that could impede advancements and hinder the potential benefits of data analytics in mental healthcare. To address these concerns, it is essential to redefine responsibilities and establish a solid institutional framework that goes beyond ethical principles.

This framework must ensure that data analytics practices are transparent, accountable, and carried out with the utmost consideration for patient privacy and well-being. Moreover, the importance of accurate interpretation and analysis of the data collected should not be overlooked, as misinformed interpretation can result in unintended harm to patients and society. Thus, the integration of data analytics in mental health has the potential to bring about significant improvements in patient care and healthcare efficiency. However, it is of utmost importance to maintain a careful balance between harnessing the power of data analytics and adhering to ethical standards. This can be achieved by fostering a comprehensive and evolving ethical discourse, redefining responsibilities, and establishing a strong institutional framework to ensure that the benefits of this technology are maximized while minimizing any potential harm. The results presented highlight the importance of respecting individuals' autonomy and ensuring that they are fully informed before participating in medical studies, especially

for vulnerable populations such as those with mental health conditions. Ethics committees play a crucial role in protecting people's rights and minimizing risks, and consent is at the core of this principle. The use of data analytics and algorithms in healthcare presents both challenges and opportunities. Ensuring that these algorithms are unbiased and aligned with human values is a significant ethical challenge that must be addressed. However, the use of advanced techniques can also improve the accuracy and consistency of diagnoses, leading to more effective treatments. Overall, the future of healthcare will require creative and innovative solutions to balance the benefits of technology with the need to protect individuals' rights and ensure equitable access to healthcare.

6.2. Limitations and Future Research

There are several limitations to the current research that future studies should address. Firstly, the bibliometric analysis was restricted to the Web of Science database, potentially overlooking relevant research in other databases or published in non-indexed sources. Future research should consider employing a more comprehensive search strategy, including databases such as Scopus, PubMed, and Google Scholar, to ensure a more exhaustive representation of the research landscape. Another limitation of this study is the absence of systematic review and meta-analysis, which may offer different and more detailed perspectives on the research. Future studies should incorporate these methods to provide more comprehensive guidance on the topic, including the effectiveness of different interventions and the impact of various factors on psychoactive substance use. Furthermore, the heterogeneity of the studies included in the bibliometric analysis makes it challenging to draw broad conclusions about the field. Future research should consider organizing the studies into more homogeneous groups based on factors such as methodology, population, or research question, allowing for more targeted and informative comparisons. Additionally, the current study did not thoroughly examine the design and approach of the publications reviewed, limiting the understanding of their outcomes and potential for future research. To address this gap, future studies should undertake a detailed examination of the methodologies, statistical analyses, and results of each publication, enabling the identification of best practices, common pitfalls, and areas of improvement.

While the implementation of a GEOTELO framework in psychoactive substance use (PSU) research presents a promising opportunity to address gaps and limitations in the literature, several concerns should be considered: Data quality and privacy are crucial in ensuring the success of a GEOTELO framework. However, the collection and utilization of this data may raise ethical concerns regarding privacy and informed consent. Future research should examine ways to anonymize and protect sensitive information while maintaining the integrity of the data. Spatial analysis may inadvertently introduce biases in the findings, as certain regions or populations might be over- or under-represented in the data. It is crucial for future research to ensure that the data used is representative of the target populations and addresses potential biases that may skew the results. Cross-cultural and contextual factors can also influence PSU patterns and behaviors, and the application of a GEOTELO framework may not fully account for these unique factors. Researchers should consider incorporating qualitative methods to complement the quantitative data generated to better understand the nuances of PSU across different settings. Technical limitations, such as ensuring the scalability, reliability, and adaptability of the system, may pose significant challenges.

In addition, while the study provides valuable insights into the patterns of psychoactive substance (PAS) use and abuse in Colombia using the Deep Neural Network-based Clustering-oriented Embedding algorithm, there are several limitations that should be considered. First, the model relies on self-reported input data, which can introduce biases and inaccuracies. Participants may underreport or overreport their substance use due to social desirability bias or recall errors. Additionally, the input data only includes factors such as sex, age, socioeconomic status, and housing type, which may not capture the full range of factors influencing PAS consumption. Second, the cross-sectional nature of the study limits its ability to infer causal relationships between the identified factors and PAS consumption. Longitudinal research designs are needed to establish whether these factors contribute to PAS use or if they are merely correlated. Third, the study focuses on the Colombian context, which might limit the generalizability of the findings to other countries or regions with different cultural, social, and economic contexts. It is essential to replicate the study in different settings to validate the model and improve its external validity. Future research could address these limitations by incorporating additional variables such as mental health status, social support, and peer influences, which may contribute to PAS consumption. Expanding the sample to include diverse populations in

other countries or regions would enhance generalizability and validate the model across different contexts. Developing interventions targeting the identified risk factors and assessing their effectiveness in reducing PAS consumption in various population groups is another crucial area for future research. Exploring the use of alternative data sources, such as medical records or biological markers, to validate self-reported substance use and obtain more accurate estimates of PAS consumption can also help address the limitations of self-reported data.

With respect to important insights into the regional patterns of illegal drug consumption in Colombia, the study relies on data collected through self-reported surveys, which may be subject to biases because the true prevalence of drug use might be underestimated due to the participants' reluctance to disclose their consumption habits. Future research could employ more objective methods, such as toxicological screenings, to validate self-reported data. Also, it should consider a broader range of contextual factors to provide a more comprehensive understanding of the dynamics driving drug consumption in different regions. The proposed CAE-DEC model, while innovative, may not capture all relevant features influencing drug use patterns. Further refinement of the model could include the incorporation of additional data sources, such as crime rates or drug-related hospitalizations, to enhance the model's predictive capabilities. Moreover, the current model does not account for potential interactions between different features, which could provide valuable insights into the complex interplay of factors influencing drug use. Future research should explore more advanced modeling techniques to better capture the complexity of the underlying factors. Furthermore, the study does not explore the potential impact of government policies and interventions on drug consumption patterns. An important avenue for future research would be to assess the effectiveness of different prevention and treatment strategies in reducing drug use and its associated harms. This could involve conducting controlled experiments or natural experiments to evaluate the impact of specific policies or programs. Future research should focus on evaluating treatment effectiveness, addressing barriers to access, and optimizing care coordination across settings.

Lastly, another limitation of the current research is the reliance on retrospective people record reviews. While these reviews provide valuable insights, they are limited by the quality and completeness of the records themselves. Future research should consider incorporating prospective data collection methods to improve the quality and

generalizability of the findings. Moreover, as the field of data analytics continues to evolve, it is crucial to continually assess the effectiveness and ethical implications of these methods in mental health care. Future research should also explore the role of interdisciplinary collaboration in addressing the ethical challenges associated with data analytics in mental health care. Collaboration between data scientists, clinicians, ethicists, and advocacy groups could lead to the development of more robust and ethically sound data analysis practices. This collaboration can also help ensure that the regulatory processes remain responsive to the evolving ethical landscape in data-driven healthcare. Another area of future research should focus on developing methods to ensure that the benefits of data analytics in mental health care are equitably distributed across all populations. This includes addressing potential biases in the algorithms and data collection processes to ensure that marginalized and underrepresented groups are not disproportionately affected by any negative consequences. Additionally, researchers should investigate the potential for data analytics to improve access to mental health care, particularly in underserved communities. As technology continues to advance, it is crucial to explore the potential for incorporating new data sources, such as wearable devices and smartphone apps, in mental health care. These sources can provide real-time, continuous data that can enhance the accuracy and efficiency of diagnoses and treatment plans. However, the use of these data sources raises ethical concerns related to privacy, consent, and data security that must be carefully considered and addressed.

Appendix A

The source code is available on GitHub at the following path:



`<https://github.com/Mental-Health-Framework/Stage-1.git>`

Appendix B

The source code is available on GitHub at the following path:



`<https://github.com/Mental-Health-Framework/Stage-2.git>`

Appendix C

The source code is available on GitHub at the following path:



`<https://github.com/Mental-Health-Framework/Stage-3.git>`
