



UNIVERSIDAD DEL VALLE
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
DOCTORADO EN INGENIERÍA
ÉNFASIS EN CIENCIAS DE LA COMPUTACIÓN

DOCTORAL THESIS
**ARTIFICIAL INTELLIGENCE AND COMPLEX NETWORK-
BASED MODEL FOR SEISMICITY ANALYSIS IN THE
COLOMBIAN REGION.**

DANIEL ANDRÉS LEÓN VARGAS

SUPERVISORS:
VICTOR A. BUCHELI, PH.D.
JUAN ALEJANDRO VALDIVIA, PH.D

2025



UNIVERSIDAD DEL VALLE
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
DOCTORADO EN INGENIERÍA
ÉNFASIS EN CIENCIAS DE LA COMPUTACIÓN

DOCTORAL THESIS
ARTIFICIAL INTELLIGENCE AND COMPLEX NETWORK-BASED
MODEL FOR SEISMICITY ANALYSIS IN THE COLOMBIAN REGION.

DANIEL ANDRÉS LEÓN VARGAS

SUPERVISORS: VICTOR A. BUCHELI PH.D.
JUAN ALEJANDRO VALDIVIA PH.D

EVALUATION COMMITTEE: ANDRÉS M. CASTILLO, PH.D.
BENJAMÍN A. TOLEDO, PH.D.
MAX A. RAMÍREZ, PH.D.

PRESENTED AND DEFENDED TO OBTAIN THE DEGREE OF
DOCTOR OF ENGINEERING

2025

To Tatiana and Sofía, the force and the drive behind all my efforts.

Daniel Andrés León

Declaration of Authorship

I, Daniel Andrés León Vargas, declare that the thesis entitled "ARTIFICIAL INTELLIGENCE AND COMPLEX NETWORK-BASED MODEL FOR SEISMICITY ANALYSIS IN THE COLOMBIAN REGION" and the work presented are entirely of my own authorship. I affirm the following:

- This work was developed while I was a candidate for a research degree at this university.
- If any part of this thesis has been previously submitted for the attainment of a degree or any other qualification at this university or any other institution, it has been explicitly stated.
- Proper attribution is always given to any published work consulted from other individuals.
- The source of any cited work from others is always acknowledged. Except for these citations, this thesis is entirely my own work.
- I have acknowledged all significant sources of assistance.
- If the thesis is based on work conducted in collaboration with others, I have clearly stated what has been contributed by others and what I have contributed myself.

Acknowledgements

During the elaboration of this thesis, I received great support from many people. I want to thank Tatiana Calderón, Sofía León, Claudia Vargas, Gladys García, Henry Calderón, and all my wonderful family who always stood by my side.

To my advisors, Victor Bucheli and Juan Alejandro Valdivia, who guided me through all this road with patience and support.

To professors Andrés Castillo, Max Ramírez and Benjamín Toledo, for their generous comments and suggestions, which improved this thesis.

To Giovanny Hidalgo, Luis Ordóñez, José Llanos and Jefferson Peña, who brought listening ears and advise anytime I asked them.

To Universidad del Valle, and particularly the Escuela de Ingeniería de Sistemas y Computación and all their teachers, for their support and understanding.

To Santiago Giraldo Alegría, with whom I was able to understand the Deep Learning Models applied to graphs.

To Laura Rodríguez, Javier Vargas, Paola Guerrero, Sandra Arango, Fabián Macías, Vivian Ospina and Juan Carlos Ramirez for their generosity and friendship

To everyone at SSI, Corporación OSSO, Ofiled Colombia and Solfix in Spain where I was able to work, learn and grow as a researcher.

To Jorge Mejia, who taught me everything I know about seismology.

To Ramon, Felipe, David, Juan, Martha, Jesús and Ramon Andrés, who brought a soundtrack for this thesis. To all my friends. Thank you all. You all made possible this accomplishment.

Table of Contents

DECLARATION OF AUTHORSHIP	III
ACKNOWLEDGEMENTS	IV
LIST OF FIGURES.....	VII
LIST OF TABLES.....	XI
LIST OF ABBREVIATIONS.....	XIII
ABSTRACT	XIV
INTRODUCTION	1
1.1. MOTIVATION	2
1.2. PROBLEM STATEMENT.....	3
1.3. GOALS AND SCOPE.....	4
1.4. DELIVERIES FROM THIS RESEARCH.....	5
1.5. THESIS OUTLINE	7
STATE OF THE ART	8
2.1. A REVISION OF SEISMICITY MODELS BASED ON COMPLEX SYSTEMS AND EARTHQUAKE NETWORKS.....	8
2.1.1. <i>Complexity in Seismicity and Complex Networks.....</i>	<i>9</i>
2.1.2. <i>Earthquake networks</i>	<i>11</i>
2.1.3. <i>Seismicity information from earthquake networks.....</i>	<i>16</i>
2.1.4. <i>Role of cell size in Cells as Nodes Earthquake networks.....</i>	<i>19</i>
2.1.5. <i>Conclusions and Discussion.....</i>	<i>21</i>
2.2. ARTIFICIAL INTELLIGENCE IN SEISMOLOGY AND COMPLEX NETWORKS.....	23
2.2.1. <i>AI applications in seismology.....</i>	<i>23</i>
2.2.2. <i>Machine learning on Complex networks.....</i>	<i>27</i>
2.3. CHAPTER HIGHLIGHTS	29
EARTHQUAKE COMPLEX NETWORKS IN COLOMBIA	30
3.1. EARTHQUAKE NETWORK ANALYSIS	34
3.1.1. <i>Cell size assessment</i>	<i>34</i>
3.1.2. <i>Experiments design</i>	<i>36</i>
3.2. RESULTS.....	37

3.3.	DISCUSSION AND CONCLUSIONS.....	45
3.4.	CHAPTER HIGHLIGHTS	48
	SEISMICITY MODELS USING EARTHQUAKE COMPLEX NETWORKS AND MACHINE LEARNING	49
4.1.	EXPLORATORY DATA ANALYSIS.....	49
4.1.1.	<i>Catalog contributors.....</i>	<i>50</i>
4.1.2.	<i>Seismicity Analysis: Energy and time.</i>	<i>51</i>
4.1.3.	<i>Geographic sectorization of seismicity.</i>	<i>58</i>
4.1.4.	<i>Input Features and outputs</i>	<i>61</i>
4.1.5.	<i>Models and learning algorithms</i>	<i>63</i>
4.1.6.	<i>Ontology definition.....</i>	<i>66</i>
4.2.	CLASSIFICATION MODELS	68
4.3.	REGRESSION MODELS.....	75
4.3.1	<i>Days for the next earthquake regression</i>	<i>76</i>
4.3.2	<i>Magnitude for the next event regression</i>	<i>78</i>
4.3.3	<i>Logarithm of seismicity for the next events regression</i>	<i>81</i>
4.4.	CHAPTER HIGHLIGHTS.....	87
	CONCLUSIONS AND DISCUSSION	88
	BIBLIOGRAPHY	91
	APPENDICES	99
	<i>APPENDIX 2: REGRESSION METRICS EXTENDED.</i>	<i>99</i>
	<i>APPENDIX 1: MEASURES TAKEN FROM THE EARTHQUAKE NETWORKS (EN) CONSTRUCTED WITH DIFFERENT CELL SIZES (DX).</i>	<i>100</i>
	<i>APPENDIX 2: REGRESSION METRICS EXTENDED.</i>	<i>102</i>
	<i>APPENDIX 3: PUBLICATIONS MADE DURING THE DOCTORAL STUDIES.</i>	<i>104</i>

List of Figures

FIGURE 1: DIAGRAM OF THE EARTHQUAKE NETWORK PRODUCED BY THE SPATIOTEMPORAL SEQUENCE OF EARTHQUAKES.....	13
FIGURE 2: EARTHQUAKE NETWORK CONSTRUCTED FROM A VISIBILITY GRAPH.	15
FIGURE 3: ML METHODS AND THEIR APPLICATIONS..	23
FIGURE 4: THE RATIO OF VARIOUS ML APPROACHES USED FOR EACH SEISMOLOGICAL TASK.	25
FIGURE 5: GENERAL EARTHQUAKE PREDICTION MODEL.	26
FIGURE 6: SEISMICITY IN COLOMBIA - JANUARY 1973 TO OCTOBER 2023.	31
FIGURE 7: GUTENBERG-RICHTER DISTRIBUTION OF THE SEISMICITY IN THE REGION.	32
FIGURE 8: HISTOGRAM ILLUSTRATING THE DISTRIBUTION OF SEISMIC EVENT DEPTHS.	33
FIGURE 9: DIAGRAM OF THE CONSTRUCTION OF AN EARTHQUAKE NETWORK FROM THE SPATIOTEMPORAL SEQUENCE OF EARTHQUAKES; INCLUDING THE DEPTH.	34
FIGURE 10: NUMBER OF NODES VS. CELL SIZE.....	38
FIGURE 11: DEGREE OF NODES VS. CELL SIZE.	38
FIGURE 12: CLUSTERING COEFFICIENT VS. CELL SIZE.....	39
FIGURE 13: AVERAGE PATH LENGTH VS. CELL SIE.	39
FIGURE 14: SMALL WORLDNESS VS. CELL SIZE.	40
FIGURE 15: Γ EXPONENT OF THE POWER LAW FOR THE DEGREE DISTRIBUTION VS. CELL SIZE.....	40
FIGURE 16: ENTROPIES OF ECN (BLUE DOTS) AND SIMULATED ER (RED), BA (YELLOW) AND WS (GREEN) NETWORKS VS. CELL SIZE.....	42
FIGURE 17: DISEQUILIBRIUM OF ECN (BLUE DOTS) AND SIMULATED ER (RED), BA (YELLOW) AND WS (GREEN) NETWORKS VS. CELL SIZE.	42
FIGURE 18: DIVERGENCES BETWEEN ECN AND SIMULATED ER (RED), BA (YELLOW) AND WS (GREEN) NETWORKS VS. CELL SIZE.....	43
FIGURE 19: ENTROPY-COMPLEXITY PLANE FOR ECN (CIRCLES), ER (EXES), WS (SQUARES), AND BA (CROSSES).	45

FIGURE 20: DISTRIBUTION OF GLOBAL SEISMOGRAPHIC NETWORK (GSN) STATIONS.....	50
FIGURE 21: CONTRIBUTORS TO THE SEISMICITY CATALOG OVER TIME	51
FIGURE 22: HISTOGRAM OF THE MAGNITUDE OF THE EVENTS IN THE CATALOG.....	52
FIGURE 23: DISTRIBUTION OF MAGNITUDE OF EVENTS OVER TIME.....	53
FIGURE 24: SEISMIC ENERGY RELEASED BY EACH EARTHQUAKE OVER TIME	53
FIGURE 25: NUMBER OF DAYS BETWEEN CONSECUTIVE EARTHQUAKES VS. TIME.....	54
FIGURE 26: HISTOGRAM OF TIME LAPSES IN DAYS BETWEEN CONSECUTIVE EVENTS	54
FIGURE 27: LOGARITHM OF SEISMICITY USING WINDOWS OF 10 EVENTS VS. TIME.....	55
FIGURE 28. LOGARITHM OF SEISMICITY USING WINDOWS OF 50 EVENTS VS. TIME.....	56
FIGURE 29: LOGARITHM OF SEISMICITY USING WINDOWS OF 100 EVENTS VS. TIME.....	56
FIGURE 30: TIME LAPSE FOR EACH SEISMICITY USING WINDOWS OF 10 EVENTS VS. TIME.	57
FIGURE 31: TIME LAPSE FOR EACH SEISMICITY USING WINDOWS OF 50 EVENTS VS. TIME	57
FIGURE 32: TIME LAPSE FOR EACH SEISMICITY USING WINDOWS OF 100 EVENTS VS. TIME	58
FIGURE 33: DEFINITION OF 4 SEISMICITY SECTORS SELECTED MANUALLY. IN THE MAP, THE COLOR OF THE DOTS INDICATES THE ASSIGNED TYPE ACCORDING TO THE SECTOR.....	59
FIGURE 34: DEFINITION OF 4 SEISMICITY SECTORS SELECTED MANUALLY. THE BAR PLOT SHOWS THE NUMBER OF EVENTS FOR EACH SECTOR.	59
FIGURE 35: DEFINITION OF 5 SEISMICITY SECTORS SELECTED MANUALLY. IN THE MAP, THE COLOR OF THE DOTS INDICATES THE ASSIGNED TYPE ACCORDING TO THE SECTOR.....	60
FIGURE 36: DEFINITION OF 5 SEISMICITY SECTORS SELECTED MANUALLY. THE BAR PLOT SHOWS THE NUMBER OF EVENTS FOR EACH SECTOR.	60
FIGURE 37: ML MODELS CREATION PIPELINE	63

FIGURE 38: PIPELINE FOR THE IMPLEMENTATION OF DGCNN CLASSIFICATION MODELS	64
FIGURE 39: DGCNN ARCHITECTURE.	65
FIGURE 40: CONFUSION MATRIX FOR GCNN MODEL ON 2 CLASS MAGNITUDE BINNING AND 3D NETWORKS.....	73
FIGURE 41: CONFUSION MATRIX FOR GCNN MODEL ON 3 CLASS MAGNITUDE BINNING AND 2D NETWORKS.....	74
FIGURE 42: RANDOM FOREST REGRESSION OF DAYS FOR THE NEXT EVENT USING ONLY NETWORK FEATURES FROM 2D NETWORKS WITH 50 EVENTS WINDOWS.....	76
FIGURE 43: RANDOM FOREST REGRESSION OF DAYS FOR THE NEXT EVENT USING ONLY NETWORK FEATURES FROM 3D NETWORKS WITH 50 EVENTS WINDOWS.....	77
FIGURE 44: RANDOM FOREST REGRESSION OF DAYS FOR THE NEXT EVENT USING NETWORK AND CATALOG FEATURES FROM 2D NETWORKS WITH 50 EVENTS WINDOWS.....	77
FIGURE 45: RANDOM FOREST REGRESSION OF DAYS FOR THE NEXT EVENT USING NETWORK AND CATALOG FEATURES FROM 3D NETWORKS WITH 50 EVENTS WINDOWS.....	78
FIGURE 46: RANDOM FOREST REGRESSION OF MAGNITUDE FOR THE NEXT EVENT USING ONLY NETWORK FEATURES FROM 2D NETWORKS WITH 50 EVENTS WINDOWS.....	79
FIGURE 47: RANDOM FOREST REGRESSION OF MAGNITUDE FOR THE NEXT EVENT USING ONLY NETWORK FEATURES FROM 3D NETWORKS WITH 50 EVENTS WINDOWS.....	79
FIGURE 48: RANDOM FOREST REGRESSION OF MAGNITUDE FOR THE NEXT EVENT USING NETWORK AND CATALOG FEATURES FROM 2D NETWORKS WITH 50 EVENTS WINDOWS.....	80
FIGURE 49: RANDOM FOREST REGRESSION OF MAGNITUDE FOR THE NEXT EVENT USING NETWORK AND CATALOG FEATURES FROM 3D NETWORKS WITH 50 EVENTS WINDOWS.....	80
FIGURE 50: RANDOM FOREST REGRESSION OF SEISMICITY OF THE NEXT 10 EVENTS USING NETWORK FEATURES AND 2D NETWORKS WITH 50 EVENTS WINDOWS.....	81

FIGURE 51: RANDOM FOREST REGRESSION OF SEISMICITY OF THE NEXT 10 EVENTS USING NETWORK FEATURES AND 3D NETWORKS WITH 50 EVENTS WINDOWS.....	82
FIGURE 52: RANDOM FOREST REGRESSION OF SEISMICITY OF THE NEXT 10 EVENTS USING NETWORK AND CATALOG FEATURES AND 2D NETWORKS WITH 50 EVENTS WINDOWS.....	82
FIGURE 53: RANDOM FOREST REGRESSION OF SEISMICITY OF THE NEXT 10 EVENTS USING NETWORK AND CATALOG FEATURES AND 3D NETWORKS WITH 50 EVENTS WINDOWS.....	83
FIGURE 54: RANDOM FOREST REGRESSION OF SEISMICITY OF THE NEXT 20 EVENTS USING NETWORK AND CATALOG FEATURES AND 2D NETWORKS WITH 50 EVENTS WINDOWS.....	84
FIGURE 55: RANDOM FOREST REGRESSION OF SEISMICITY OF THE NEXT 20 EVENTS USING NETWORK FEATURES AND 2D NETWORKS WITH 50 EVENTS WINDOWS.....	85
FIGURE 56: RANDOM FOREST REGRESSION OF SEISMICITY OF THE NEXT 20 EVENTS USING NETWORK FEATURES AND 2D NETWORKS WITH 50 EVENTS WINDOWS.....	85
FIGURE 57: RANDOM FOREST REGRESSION OF SEISMICITY OF THE NEXT 20 EVENTS USING NETWORK AND CATALOG FEATURES AND 3D NETWORKS WITH 50 EVENTS WINDOWS.....	86
FIGURE 61: PIPELINE FOR THE CREATION OF THE SEISMICITY MODEL SUPPORTED BY EARTHQUAKE NETWORKS AND ML.....	88

List of Tables

TABLE 1: DETAILED LIST OF DOCUMENTS THAT CONTRIBUTED FOR EACH OBJECTIVE OF THE THESIS.....	5
TABLE 2: SEVEN NETWORK MEASURES IN COMPLEX NETWORKS.	10
TABLE 3: RULES USED TO DEFINE NODES AND LINKS IN EARTHQUAKE NETWORKS.....	16
TABLE 4: EARTHQUAKE NETWORK PARAMETERS	17
TABLE 5: PREVIOUS STUDIES ABOUT CELL SIZES IN EARTHQUAKE NETWORKS	20
TABLE 6: APPLICATIONS OF ML IN EARTHQUAKE SEISMOLOGY. FROM [59].	24
TABLE 7: CLASSIFICATION OF FORECAST PERFORMANCES FOR CLASSIFICATION AND REGRESSION IN SEISMOLOGY.	27
TABLE 8: PARAMETERS OF THE EXPERIMENTS.....	37
TABLE 9: ML ALGORITHMS AND INPUT FEATURES USED TO CREATE THE MODELS	62
TABLE 10: ML PROBLEMS AND TARGET OUTPUTS.	62
TABLE 11: PARAMETERS USED TO CREATE DGCNN MODELS FOR EN.	65
TABLE 12: METRICS USED TO QUALIFY CLASSIFICATIONS.	68
TABLE 13: CLASSIFICATION METRICS FOR THE 4 CLASS GEOGRAPHIC SECTORIZATION MODELS.	69
TABLE 14: CLASSIFICATION METRICS FOR THE 5 CLASS GEOGRAPHIC SECTORIZATION MODELS.	70
TABLE 15: CLASSIFICATION METRICS FOR THE 2 CLASS MAGNITUDE BINNING MODELS.	71
TABLE 16: CLASSIFICATION METRICS FOR THE 3 CLASS MAGNITUDE BINNING MODELS.	72
TABLE 17: DESCRIPTION OF REGRESSION METRICS.....	75
TABLE 18: METRICS FOR THE RANDOM FOREST REGRESSION OF DAYS FOR THE NEXT EVENT.	76
TABLE 19: METRICS FOR THE RANDOM FOREST REGRESSION OF MAGNITUDE FOR THE NEXT EVENT.	78

TABLE 20: METRICS FOR THE RANDOM FOREST REGRESSION OF SEISMICITY RELEASED BY THE NEXT 10 EVENTS.	81
TABLE 21: METRICS FOR THE RANDOM FOREST REGRESSION OF SEISMICITY RELEASED BY THE NEXT 20 EVENTS.	83

List of abbreviations

AI	ARTIFICIAL INTELLIGENCE
ECN	EARTHQUAKE COMPLEX NETWORKS
ER	ERDÖS-RENYI
BA	BARABÁSI-ALBERT
WS	WATTS-STRÖGATZ
SF	SCALE-FREE
SW	SMALL-WORLD
APL	AVERAGE PATH LENGTH
GCC	GLOBAL CLUSTERING COEFFICIENT
ML	MACHINE LEARNING
SVM	SUPPORT VECTOR MACHINE
RF	RANDOM FOREST
DGCNN	DEEP GRAPH CONVOLUTIONAL NEURAL NETWORKS
SGC	SERVICIO GEOLÓGICO DE COLOMBIA (COLOMBIAN GEOLOGICAL SURVEY)
GR	GÜTENBERG-RICHTER
USGS	UNITED STATES GEOLOGICAL SURVEY
NEIC	NATIONAL EARTHQUAKE INFORMATION CENTRE

Abstract

ARTIFICIAL INTELLIGENCE AND COMPLEX NETWORK-BASED MODEL FOR SEISMICITY ANALYSIS IN THE COLOMBIAN REGION.

BY DANIEL ANDRÉS LEÓN VARGAS. M.Sc

Research about Earthquake Networks has been carried on for many years, and they have been proved as a tool to analyze the seismicity. However, there has been no comprehensive integration of this representation with artificial intelligence methods to model the dynamics of seismicity. This thesis presents a novel approach to find how Earthquake Networks, supported by Machine Learning methods, can be used to look into the dynamics of the seismicity in the Colombian region.

This thesis shows how can Earthquake Networks be constructed to represent the seismicity in the region of study. The information of such networks has been analyzed, contrasting their characteristics with the variations of the seismicity in time, in order to find an appropriate way of creating the Earthquake Networks. Also, data analysis of a seismic catalog is performed to develop an annotation schema and a pipeline for the application of Machine Learning algorithms with Earthquake Networks information. The outputs of the Machine Learning models trained have been analyzed using different metrics, and some models appeared fit to forecast some aspects of the seismicity such as the seismic energy released by the expected events, or the moment of occurrence of the next event.

The findings indicate that earthquake networks, combined with machine learning techniques, can provide valuable insights into seismic behavior, and may serve as a predictive tool for seismic monitoring. These models can potentially be integrated into real-time systems for seismic observatories, offering forecasting and alerts to users of interest.

Keywords: Earthquake Networks, Seismicity, Machine Learning, Graph Neural Networks,

CHAPTER 1

Introduction

The processes that occur inside the Earth have an aura of mystery that has inspired works like Jules Verne's "Journey to the Center of the Earth," or myths about civilizations, enormous caverns, oceans, and lost cities. This imagination stems from the fact that it is very difficult for humans to see the interior of the Earth. This is precisely what has led to this phrase from Stein and Wyss's book: "seismology is the primary tool for the study of the Earth's interior because little of the planet is accessible to direct observation" [1].

Seismology, as a science, plays a crucial role in understanding the structure of the Earth's interior and finding out the mechanisms underlying earthquake occurrence. Through the analysis of the vibrations measured on the Earth's surface (which can be generated by natural events or human-induced activities), seismologists can infer information such as the composition, density, and elasticity of Earth's layers; and provide information about geophysical properties and tectonic processes. On the other hand, the outcome of Artificial Intelligence (AI) has allowed to model very complex phenomena, such as climate, financial markets, or natural-language processing; but with all these advances in processing and modeling, to this day, earthquakes still can't be predicted [2]. There are various reasons which make this challenge unsolved by the day: the stochastic and nonlinear nature of earthquake generation, involving complex interactions among geological faults, stress accumulation, and release mechanisms; and the impossibility to actually see the interior of the Earth, due to the pressure, strength and temperature of the materials that compose this planet.

This thesis intends to use a novel lens to grasp into the seismic phenomena: to use the representation of seismicity as Earthquake Networks and Artificial Intelligence algorithms to observe and model the *seismicity* in a region (A remark for this thesis: seismicity is a word often used in this thesis, and it is important to make it clear that it is a quantification of the seismic activity in a region [3]). The idea is to use these concepts in the seismicity near Colombia, where many tectonic plates interact between them, which results in a good region to investigate new advances in seismology.

1.1. Motivation

This thesis has been the result of a long journey. It started long time ago, working as a research assistant at Corporación OSSO, processing seismological data from oil and gas fields. In my master's thesis I developed a Classification system for seismograms, and the annotation of such signals was made using Earthquake networks; and during that research, with my advisors, we published the first Earthquake Networks analysis for the Colombian seismicity "Modeling of Colombian Seismicity as Small-World Networks" on the Seismological Research Letters journal on 2018 [4]. That was the first seed of the research that continues to this day.

In seismology, there are laws such as the Gutenberg-Richter [5] and Omori laws [6], which delve into the complex nature of the seismic phenomena. And the word "complex" is chosen because they are governed by power laws in their formulations. Power laws play a pivotal role in the understanding of complex systems, offering a lens through which the scale-invariant properties can be grasped and emergent patterns inherent in these systems.

About complex systems, in the introduction of Nino Boccara's book "Modeling Complex Systems" [7], it is described that a complex system is composed of multiple agents where a global dynamic emerges as a result of the actions of these agents rather than being imposed by a central controller, exhibiting a behavior of self-organization. An excellent example of complex systems are social networks, which, viewed as graphs, exhibit a high degree of self-organization, evidenced by power laws governing some of their characteristics. Since these same types of laws are present in seismology, it suggests that seismicity can be treated as a complex phenomenon. 'Earthquake Complex Networks'; abbreviated in this thesis as ECN networks (not to be confused with seismic or seismological networks, which are the set of seismometers and geophones that measure Earth vibrations), have been proposed as a way to observe seismicity using tools from complex systems and particularly from network science. An ECN network is built from a catalog of earthquakes and rules defined by an observer to configure nodes and links. The construction of earthquake networks was proposed by Abe and Suzuki [8], [9] to study seismicity in various regions using concepts from network science, applying concepts from Erdős and Renyi's random networks [10], Watts and Strogatz's (WS) Small-World (SW) networks [11], and Scale-Free (SF) networks by Albert and Barabasi [12]. Erdős-Renyi networks are referred in this thesis as ER networks, Watts-Strogatz networks as WS networks and Barabási-Albert networks as BA networks.

Currently, the best seismicity models suggest the possibility of forecasting earthquake occurrences, under the best circumstances (i.e., a well-instrumented region with an extensive event catalog), within a few years [13]. Any advance in this research field is very important to help organizations and governments to be prepared for events that may have a great impact on the life of the citizens. As this thesis was written, no work has been made using a seismicity representation as ECN networks, which makes it interesting to see if this approach might help in these tasks.

1.2. Problem statement

Characterizing ECN networks at different spatial scales and their dynamics could eventually provide information with added value in a timely manner, allowing a region to prepare for changes in seismicity behavior. This remains one of the most important tasks in geophysics and seismology. The purpose of this thesis is the creation of a seismicity model that allows for the characterization of its dynamics, that is, both the current state and how such system may change in time. For this purpose, a description will be used where the dynamics occur on complex networks of earthquakes, which characterizes the spatiotemporal heterogeneity of the phenomenon. A first step in this direction for the case of Colombia can be seen in the publication by León et al. [4].

The representation of seismicity as ECN networks has been a research topic for several years. Different ways of creating such networks have been proposed, and the analysis of the characteristics from the constructed network shows that they have a very close relationship with some parameters calculated from seismicity. Studies of these networks are reported for specific cases in some countries or in a region within a country (see Table 5), and so far, there has been no study that tries to define a model that integrates the observations made in a region at different spatial scales and their variation over time.

On the other hand, the forecast problem in seismology has been tackled also with the power of Artificial Intelligence (AI) and Machine Learning (ML), looking to improve models of seismicity. To this day, the scarcity of ML applications on ECN networks opens a research question which can be stated as:

¿How Earthquake Complex Networks, supported by Machine Learning algorithms, can be used to forecast seismicity characteristics?

1.3. Goals and scope

Following the problem statement, a general and some specific objectives have been defined:

General objective

To develop a dynamical regional model based on complex networks concepts and supported by ML algorithms, which allows to study the behavior and changes of seismicity in a region.

Specific objectives

- a) To define a scheme of complex network construction from a seismicity catalog, which allows to study the dynamical behavior of seismicity in a region.
- b) To define characterizations of the earthquake networks from different scales and a semi-automatic annotation scheme from the analysis of seismicity, which can be used to apply different AI techniques.
- c) To create a dynamical regional model of seismicity from the representation and annotation scheme.
- d) To integrate AI techniques to describe the behavior of seismicity through different scales of the earthquake networks.
- e) To evaluate the proposed model from the perspective of forecasting the dynamics of the next states of the regional seismicity.

It is important to clarify that this thesis has some limitations. This research has been made using only the seismicity around a specific region, and not all regions have the same behavior in terms of this phenomenon. Also, the data analyzed spans a period of approximately 40 years, which in terms of geologic times is a small period.

1.4. Deliveries from this research.

To accomplish the objectives of this thesis, some papers and products were made. Table 1 presents a detailed list of the publications, chapter sections, and documents where a contribution was made to reach a completion of each objective.

Table 1: Detailed list of documents that contributed for each objective of the thesis.

Objective	Type	Reference	Contribution
To define a scheme of complex network construction from a seismicity catalog, which allows to study the dynamical behavior of seismicity in a region.	Publication	León-Vargas, D., Bucheli-Guerrero, V. A., y Ordoñez-Erazo, H. A., (2023). Modelado de la sismicidad colombiana como una red compleja espacio-secuencial. Revista Científica, 48(3), 67-81 [14].	Description of the way to construct earthquake networks in the region, and some of the metrics which might be of interest to study the seismicity.
	Chapter	Chapter 3: Earthquake complex networks in Colombia	Full detailed definition of the network's construction scheme and assess about the best cell size to create Earthquake Networks for the seismicity in Colombia.
To define characterizations of the earthquake networks from different scales and a semi-automatic annotation scheme from the analysis of seismicity, which can be used to apply different AI techniques.	Publication	D. Leon, J. Valdivia, and V. Bucheli, "A revision of seismicity models based on complex systems and earthquake networks," J. Seismol., vol. 26, no. 1, pp. 137–145, Feb. 2022 [15]	Definition of the state of the art in ECN Networks as a tool to analyze seismicity.
	Publication	D. Leon-Vargas., V. Bucheli-Guerrero., H. Ordoñez-Erazo "TagEQ-CN: Clasificación de Eventos Sísmicos Basado en Redes Complejas", Investigación e Innovación en Ingenierías, vol. 9, n°1, 136-148, 2021. [16]	Example of annotation of earthquakes using the ECN networks information for an earthquake classification system.
	Chapter	Chapter 3: Earthquake complex networks in Colombia	Experiment on networks constructed with different cell sizes to assert which cell size is the best to analyze the network
	Chapter subsection	Section 4.1.4: Input Features and outputs	Definition of windows to construct the networks over time, features of these networks to train the ML models, and output targets for the models. Definition of categories
	Chapter subsection	Section 4.1.5: Ontology definition	Proposal of rules to create an ontology to represent the target outputs.
To create a dynamical regional model of seismicity from the	Chapter section	Section 4.1: Exploratory Data Analysis	Inspection of the target outputs and how they change over time, and with the cell size.

representation and annotation scheme.			
To integrate AI techniques to describe the behavior of seismicity through different scales of the earthquake networks.	Chapter	Chapter 4: Seismicity Models using Earthquake Complex Networks and Machine Learning	Creation of ML models (Random Forest and Support Vector Machine) and DGCNN models to forecast target outputs. Metrics for regressions and classifications in Tables 13, 14,15, 16, 18, 19, 20 and 21.
To evaluate the proposed model from the perspective of forecasting the dynamics of the next states of the regional seismicity.			

It is important to mention that during the development of this research, I contributed to the paper *“Predicción de radiación solar en sistemas fotovoltaicos utilizando técnicas de aprendizaje automático”* [17], on some concepts about the photovoltaic systems and ML models. Also, to afford my doctoral studies, I worked as a consultant on some companies on different projects related with data processing and, machine learning, which due to confidentiality agreements, no publications can’t be made.

1.5. Thesis outline

This thesis presents the results of a research intended to use earthquake networks to model the seismicity supported by ML techniques and has been outlined in 5 chapters:

Chapter 1 serves as an introduction of the main topics that this thesis deals, depicting some of the concepts used on this text, defining what is the research problem, the objectives, and the scope of the thesis.

Chapter 2 presents the state of the art on the three main subjects of this thesis: Earthquake Networks, ML applications on seismology and ML applications on complex networks, with the intention of gathering the main aspects to consider when creating the models.

Chapter 3 tackles the problem of representing the best possible way the seismicity as ECN networks using an earthquake catalog for a region surrounding Colombia.

Chapter 4 presents the development of the model performing first a data analysis of the catalog to explicitly state the desired features of the seismicity to model; an ontology that defines the concepts used in the models, the inputs, and the outputs; and the application of various machine learning techniques on the seismicity representation outlined on chapter 3 and the desired outputs with the results of such applications.

Finally, Chapter 5 compiles the conclusions for this thesis and presents some aspects which can be improved in future research on this topic.

CHAPTER 2

State of the Art

The purpose of this chapter is to present what has been done regarding seismicity models using ECN networks, and how these models have been applied. Additionally, since the goal is to use machine learning tools in conjunction with representations of seismicity through earthquake networks, this chapter also intends to show the extent to which ML has been applied in both seismology and complex networks.

Earthquake Complex networks (ECN) were introduced by Abe and Suzuki in 2004 [8] as a way to represent the seismicity and use the methodological tools that graph theory can provide to study complex systems. And, thanks to the advances in computing sciences, artificial intelligence has been used in seismology for several applications. This chapter provides the state of art of these two subjects. The state of art in ECN networks was published in 2021 in the Journal of Seismology [15], updated to the date; and a revision of the use of artificial intelligence in seismology. Section 2.1 presents that publication, section 2.2 a review on applications of ML algorithms in seismology and section 2.3 a review on the uses of ML algorithms on complex networks.

2.1. A revision of seismicity models based on complex systems and earthquake networks

Through the statistical analysis of seismicity, two of the essential seismology laws have been formulated: the Omori law [6], which describes the frequency of occurrence of aftershocks after a mainshock, and the Gutenberg-Richter (GR) law [18], which describes the scaling relationship between earthquake magnitude and frequency of occurrence. These two laws display power law distributions, which are common in *complex systems*. A complex system may be defined as «*a system formed out of many components whose behavior is emergent, that is, the behavior of the system cannot be simply inferred from the behavior of its components*» [19]. In the sense of this type of interaction, complexity has been a critical concept of several models in seismology, such as the earthquake nucleation process (which is, how an earthquake is generated) as a cellular automaton [20], [21] or seismicity in general as *earthquake networks* [8], [22], [23], [24], [25].

This sub-chapter aims to collect a good part of the efforts made in the analysis of earthquake networks to know what questions remain to be resolved in this research area. Some theoretical foundations are discussed that support seismicity analysis as a complex phenomenon and complex network analysis. Then, the methodologies used to build earthquake networks are shown, the most relevant characteristics used in the analysis of these networks, and what kind of information about seismicity may be inferred from those characteristics. Finally, some conclusions are gathered, and some open topics are named that can serve as motivation for future works.

2.1.1. Complexity in Seismicity and Complex Networks

Power-law distributions in seismology are evidence of the complexity inherent to this phenomenon. Thus, many works have been carried to explore seismicity with this light. In the work of Bak and Tang [21], seismicity is modeled as a phenomenon with critical self-organization using blocks linked by springs, and the power laws of Gutenberg-Richter (GR) and Omori are related to behaviors of critical self-organization that emerge from such system. This work defines an initial framework to study seismicity as a complex phenomenon, which continues with Ito [26] discussing fractality and multifractality in seismicity. In the preface of "Introduction to Nonextensive Statistical Mechanics" by Tsallis [27] it is read that *"One of the main – necessary but by no means sufficient – features of complexity has to do with the fact that both very ordered and very disordered systems are, in the sense of plectics, considered to be simple, not complex"* (*Plectics* has to be understood as the science that studies complexity). In that work are described some of the analytical tools used in complex systems. It is also explained that entropy in complex systems differs from entropy in random type or chaotic systems (such as a fluid or a gas), in the sense that the system's total entropy is not the sum of the entropy of its parts, which requires a *non-extensive statistical mechanics*. This type of tools, applied to seismology, derives in a branch that can be described as non-extensive statistical seismology, collected by Vallianatos et al. [28], where many features that make seismicity seen as a complex phenomenon are described, such as scaling relationships, spatial and temporal correlations, and critical phenomena; as well as the way to apply non-extensive statistical mechanics in seismology.

Boccaletti et al. presented a review about calculations made in complex networks [29]. Numerous parameters and measures related to complex networks and some

real network characteristics are described. Many of the characteristics inherent to complex systems are found in complex networks, such as self-similarity, fractality, and scalability. There are different types of complex networks. The most important ones can be classified as *Small-world (SW)*, *Scale-free (SF)* and *random* (random networks are referenced in the literature as *Erdős-Renyi (ER)* networks). Completely random networks began to be studied by Erdős and Renyi [10], which became a first effort to lay the foundations with which the networks were studied later. SW networks were coined by Watts and Strogatz [11], and are presented as a type of networks that lie between regularly connected networks and networks connected completely randomly. SF networks are a type of networks that demonstrate a distribution that follows a power law in the number of connections that nodes have (known as the degree of a node) and have the property of having a few nodes with many connections and many nodes with few [30]. Real-world networks are normally compared to and studied from the perspective of one of these three types of networks. Table 2 is modified from Chorooglou et al. [31], and some parameters of the networks that are most used for analysis are described.

Table 2. Seven network measures in complex networks. Inspired from Chorooglou et al.[31]

Network parameter	Description
Degree centrality	Average overall nodes of either the number or the strength of connections.
Clustering coefficient	Average over all nodes of the fraction of connections between the nodes within their neighborhood, divided by all possible connections.
Characteristic path length	Average of the shortest path lengths in the network, computed over all pairs of nodes.
Global efficiency	Average inverse shortest path length in the network, computed over all pairs of nodes.
Betweenness centrality	Average over all nodes of the node betweenness centrality, which is the fraction of all shortest paths in the network that contain the node, divided by all possible paths.
Modularity	Shows the strength of a network that is divided into components.
Eccentricity	Average over all nodes of the maximal shortest path length between a node and any other node

Some complementary parameters have been proposed to analyze networks, such as small-worldness S which, for a network i is defined as

$$S_i = \frac{C_i/C_r}{L_i/L_r},$$

where C_i and L_i corresponds to the Global Clustering Coefficient (GCC) and Average Path Length (APL) of the network respectively, and C_r and L_r to the same parameters

of an ER network similar to the original network; which means that it is calculated comparing the clustering and the path lengths in the network with the same parameters of an ER random network [32]; Renyi's entropy [33]; Jenson-Shannon entropy and statistical divergence respect an equilibrium state (which for the purpose of complex networks is chosen as a similar ER random network) [34]. Other types of entropies can be reviewed, as entropy can be estimated from the distribution of any measure from the nodes, such as degree, centrality, or average path length. While the calculation of entropy is valid to check the amount of information in a system, using Shannon's entropy may not be the best option in earthquake networks, since as commented by Vallianatos [28] seismicity has properties which can be better described using non-extensive statistical mechanics, so the use of Tsallis entropy could be more appropriate to describe such systems.

There are many computational tools and frameworks for analyzing complex networks. SNAP [35], written in C++ and is available in Python, has many tools to analyze big networks. Gephi [36] is a platform built in Java and uses OpenGL for visualization. It is available in Windows, Linux, and Mac OS x, intended to visualize large networks and manipulate them. GraphViz [37] is also an open-source graph visualization software, intended mainly to visualize graphs. On Wolfram Mathematica [38], there is a library with many commands and algorithms intended to analyze networks and graphs. In python NetworkX [39] has many algorithms to draw and analyze networks, and StellarGraph provide a ML framework to work with graphs [40].

2.1.2. Earthquake networks

A sequence of earthquakes is a series of events that occur in space and time. In Gao et al. [41], different methods to construct complex networks from time series were presented, such as visibility graphs, recurrence networks, and networks based on transitions. This type of analysis could open a place for cross-fertilization between methods to analyze time series and complex networks theory. Interestingly, as earthquakes follow a temporal sequence (thus a seismicity catalog can be seen as a time series of events), they can be expressed as complex networks in ways that have not yet been contemplated in the literature. The joint use of complex networks and time series is also evident in Pastén et al. [42], which shows a procedure to transform an earthquake network into a time series through a non-random walk, following the sequence of occurrence of earthquakes. The time series constructed from this

connectivity shows that multifractality decreases when seismic activity increases differently in two different regions. So, analyzing complex networks as time series seems to generate exciting information about the activity and deserves to be considered an analysis tool.

Abe and Suzuki [8] were pioneers in the analysis of seismicity using complex networks. The constructed earthquake networks contain various properties that can be directly related to parameters estimated from traditional seismicity analyzes, showing that the analysis of earthquake networks can be a tool that allows analyzing the complexity of seismicity. Different ways of building earthquake networks have been developed from that work, which can be differentiated by how nodes are defined. On the one hand, earthquake networks are constructed by dividing a region into cells and a cell is selected as a node when a seismic event occurs in it (*cells as nodes*). On the other hand, there are earthquake networks in which earthquakes are the nodes (*earthquakes as nodes*).

2.1.2.1. Cells as nodes

Abe and Suzuki describe a way to create an earthquake network from a seismicity catalog, performing a division of the studied region in cubic cells [9]. A cell is chosen as a network node if earthquakes occur within it, and nodes are connected following the earthquake's temporal sequence. Figure 1 shows a step-by-step schema that describes the process of constructing a cell-as-node network.

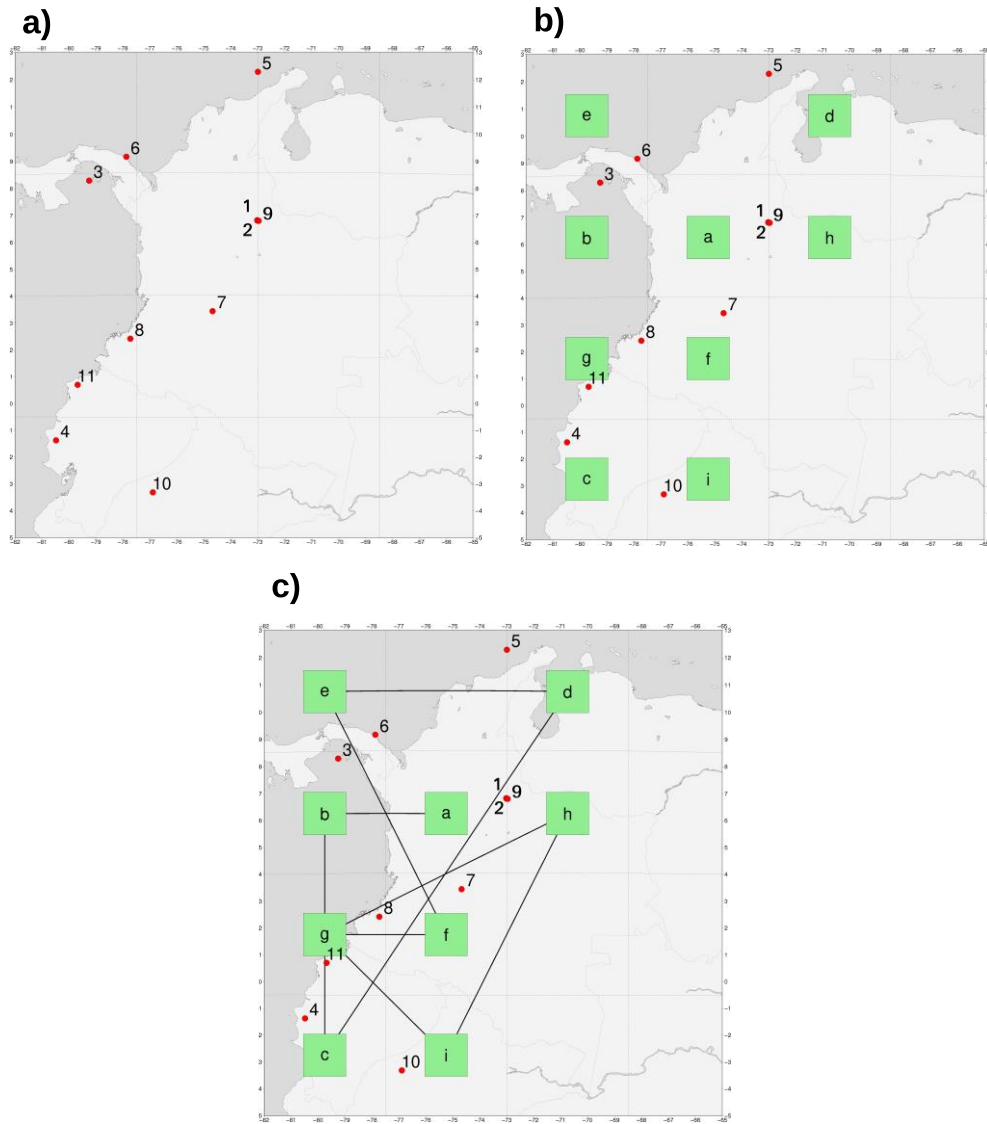


Figure 1: Diagram of the earthquake network produced by the spatiotemporal sequence of earthquakes. (a) The sequence of events (numbered dots) is mapped on a grid. The number near the events follows the temporal order from oldest to newest. (b) A cell from the grid is selected as a node if there are events mapped inside the cell. The letter inside the node (squares) identifies the node. (c) Nodes are connected following the temporal sequence of events, linking two nodes with successive events.

The observation of the earthquake networks reported by Abe and Suzuki [8] for seismicity in California and Japan gave as a result degree distribution which follows a power law. Earthquake networks can be built considering or not the directivity of the

connections; however, since nodes have the same input and output degree (except the nodes that correspond to the first and last earthquakes), the networks can be analyzed as non-directed. This is because any inward link comes from a previous earthquake and follows an outward link to the next earthquake. Also, to analyze the earthquake network's clustering coefficient, *loops* (links of a node with itself) and multiple links between two nodes (due to several times when an earthquake in a specific node follows an earthquake on another specific node) should not be considered, as two nodes are neighbors either there are many or just one link between them. In this case, the earthquake network's small-world nature comes to light, as the clustering in these networks is strikingly different from that of completely random networks [4], [9], [43]. Abe et al. [44] also analyze the cell size influence in the earthquake network structure, and it is shown that clustering C and the exponent γ in the power law present in the connectivity distribution remain constant ($C \sim 0.85$ and $\gamma \sim 1.0$) after exceeding a given size; which is a consequence of bigger cells grouping more events (quantity of nodes decrease) and greater connectivity (the links on specific nodes increase). This invariance is an important finding as to the smallest cell size that generates an earthquake network with these values for C and γ may constitute optimum earthquake networks for characterizing the seismicity.

2.1.2.2. *Earthquakes as nodes*

Earthquake networks can be constructed by linking earthquakes between them. Telesca and Lovallo [25] propose to build earthquake networks using *visibility graphs*; which were introduced by Lacasa et al. [45] as a simple computational tool to study series from the perspective of graphs, connecting different nodes from some measure of amplitude. Telesca and Lovallo introduce this technique based on the magnitude of earthquakes. Events are plotted in a magnitude vs. time plot, drawing lines from each dot perpendicular to the time axis. Then, links are made between any two earthquakes that do not intersect any perpendicular line. Reviewing a sequence of events in Italy, the technique shows that the degree distribution follows a power law. Figure 2 is made similar to an example shown on Telesca and Lovallo and illustrates the method with 18 hypothetical earthquakes.

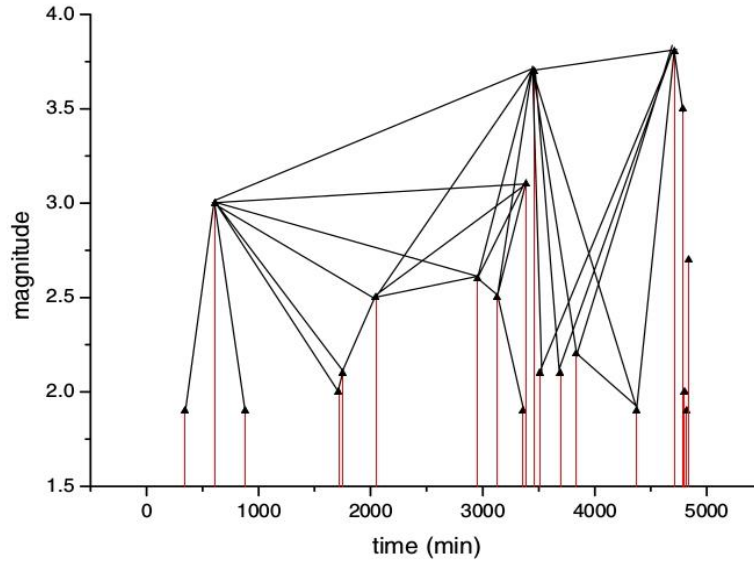


Figure 2: Earthquake network constructed from a visibility graph. Earthquakes are mapped as dots in a magnitude vs. time plane; and links are made so they don't intersect any vertical line from the dots and the time axis. Inspired in [25]

Baiesi and Paczuski [46] describe another way to create earthquake networks using an *event correlation metric* to link earthquakes. Correlation c_{ij} between two events is defined as $c_{ij} = 1/n_{ij}$; where $n_{ij} \equiv K t l^{d_f} 10^{-b m_i} \Delta m$: this correlation metric takes into account the difference of the magnitudes Δm , the distance l and the time t between events, the analysis of the GR curve in the study region (parameter b), the fractal dimension of seismicity d_f , and the calculation of a constant K which is calculated for a region, time interval and underlying seismicity. This correlation metric should not be confused with the statistical correlation; however, it also tries to find how related are two events using some of their features. Networks built with this method are of the scale-free type (few *hubs*, many nodes with few links), consistent with few mainshocks connected with many aftershocks. Baiesi and Paczuski conclude that although the correlation threshold may be a bit arbitrary (although for Southern California, the default value was 10^4), networks maintain in their structure a constant exponential parameter that scales the magnitude of an earthquake with the distance of each aftershock; and that there is evidence that seismicity is dominantly triggered by small earthquakes.

Suteanu [47] proposes a way to create earthquake networks similar to that proposed Baiesi and Paczuski [46], for volcanic events in Hawaii; however, the links have a

weight which is a combination of a spatial, a temporal and a magnitude variables. In the formulas used to estimate the weights, constants are used which, for the purposes of the study, are tested with different values. Suteanu finds that the variation of clustering coefficient C on networks constructed on successive temporal windows can be related to variations on the Gutenberg-Richter β parameter, which shows that earthquake networks' structures may be associated with fundamental properties of seismicity.

Xu et al. [48] propose another type of threshold. They state that each earthquake has a threshold of influence in time and distance that depends on the magnitude; whereby links are generated between events within the domain of that space-time influence. By using this method, the division by node communities was explored, concluding that using the communities can slightly improve the prediction of events (although the accuracy of the predictions is still below 0.5).

2.1.3. Seismicity information from earthquake networks

Table 3 summarizes the methods for creating earthquake networks, indicating how the nodes are defined, the rules governing the links between nodes, and some references where such networks have been used.

Table 3: Rules used to define nodes and links in earthquake networks.

Definition of nodes	Definitions of links
Cells as nodes	Cells are linked following the temporal sequence of earthquakes [4], [8], [9], [43], [49].
	Cells are linked if a visibility condition between nodes is met, based on the magnitudes and the time between events [24].
	Cells are linked if it is met that an earthquake in the following cell is inside a space-temporal influence domain for the previous event in its cell [23].
Events as nodes	Events are linked if a visibility criterion is met based on the magnitude and the time between events (visibility graph) [25].
	Events are linked if they are inside the space-temporal influence domain [48].
	Events are linked using a correlation factor from magnitudes, occurrence times, and distances [46], [47].

Measurements from the earthquake networks have a different meaning, depending on the type of earthquake network constructed. For instance, the degree in a *cells as*

nodes network implies how many times an earthquake can occur in the sector corresponding with that cell after an earthquake has occurred on other sectors. Hence, a high degree for a node implies many earthquakes in a sector. For the *earthquakes as nodes* networks, degree implies how many earthquakes are inside its influence radius (that is, how many earthquakes can be related as aftershocks or foreshocks). Hence, a high degree implies a significant earthquake in terms of the number of foreshocks and aftershocks. The other parameters' meaning merits a more detailed analysis of the implications in the seismicity dynamics.

From the analysis of the earthquake networks' parameters, some conclusions found in the literature can be summarized in Table 4.

Table 4: Earthquake network parameters

Network parameter	Seismicity information
Community division	Earthquake prediction by the network can be improved using a community division [48].
Quantity of links	It is shown that the Gutenberg-Richter data can be estimated using the Abe-Suzuki model and the Omori law using the Telesca-Lovallo model [24]
Connectivity distribution	It follows a power law, which agrees with the Gutenberg-Richter law, showing the presence of giant components (earthquakes that generate aftershocks) and the behavior of <i>preferential attachment</i> [50]
Multifractality	Multifractality is evaluated in a temporal series created from an earthquake network, and it is shown a change by the presence of strong earthquakes [42]
Small-worldness and clustering	Changes in small-worldness and clustering can be of use to indicate a possible seismic risk in the short term [49]

Other conclusions that emerge from other works can be listed below. Chorozoglou et al. [49] made a review of the evolution of the earthquake networks created using the Abe and Suzuki method before and after large earthquakes, analyzing the cases with and without weights in the links, and using 30-day windows which move in periods comprising 300 days before and 150 days after a large earthquake. Network parameters such as clustering coefficient, eccentricity, small-worldness, path length, and efficiency of the networks are reviewed (efficiency in a network can be seen as the inverse of the path length and indicates how effective is the transmission of data by the network); as well the calculation of a statistical factor p in which the earthquake network is compared with random networks similar by the degree distribution (for

the case without weights) and by the connectivity (for the case with weights). The conclusions show that in the case of links with weights, there is a behavior that seems to announce the appearance of a large earthquake in a short time, and an increase in centrality is appreciated before a large event. In this work can be seen that earthquake networks could be applied as short-term prediction tools (a few months) before a large earthquake. However, it is noted the importance to continue investigating to corroborate this statement.

Deyasi et al. [51] made another review of earthquake networks created with the Abe and Suzuki method, for different regions in the world that are located between different tectonic plates. A measurement of the similarity of the networks from the Jensen-Shannon divergence is used, and it was observed that seismic regions near the same tectonic plate could generate similar networks. The paper raises the difficulty of measuring the distance between two earthquake networks built with different amounts of seismic events and generating networks with different numbers of nodes. This difficulty was addressed by calculating the spectrum of adjacency matrices, and although it is not mentioned, it is shown in the article that the structure of the networks has some characteristics which are conditioned by the tectonics of the region and that this conditioning covers the regions that are close to such tectonic structures.

Ohsawa [52] presents a method to detect foreshocks of large events using a Regional Entropy of Seismic Information index. Entropy, as in almost all work done in complex systems, is evaluated from the amount of information in a system. The method is applied in seismic clusters, which are created by dividing the study region into a mesh, and the cumulus sizes in each mesh cell are estimated. It is observed that some gaps appear in the seismicity before the appearance of a large earthquake, which has been seen to correspond to changes in the behavior of this Entropy index. Thus, the index can detect a precursor activity up to a year before a large earthquake. Although in this work a seismic network is not created, the idea of using the changes in the entropy of the system, that is, the changes in the information of the seismic system, shows promising results for an earthquake prediction system, which could well be used in an earthquake network where the calculation of entropy is easier.

About what is the right size to build a cells-as-nodes earthquake network, some works have been found. Lotfi and Darooneh [53] revise a resolution parameter that is inversely proportional to the size of the cells, and it is concluded that when the resolution increases, the characteristics of the networks become very dependent on each other.

2.1.4. Role of cell size in Cells as Nodes Earthquake networks

Cell-as-nodes earthquake networks are sensitive to the cell size. In the work of Lotfi and Darooneh [53], earthquake networks were constructed to analyze seismicity in Iran, in a region that covered 24° of longitude and 20° of latitude. By using different cell sizes to create the network, the authors concluded that as the resolution increases, that is, with smaller cell sizes, the network becomes random and uncorrelated. However, questions arise about how well these networks represent seismicity. Seismicity is a complex phenomenon, which is confirmed by laws such as the Gutenberg-Richter law and the Omori law, and a random representation would be inadequate given the complex nature of this phenomenon.

On the other hand, Abe and Suzuki presented in [54] a study on the cell size in earthquake networks constructed in California, Japan, and Iran. The study showed that when the cell is larger than a certain value, the scale-free behavior remains constant, that is, the clustering coefficient reaches a constant level near 0.8 and the degree distribution follows a power law with an exponent $\gamma \sim 1$. Following this, if many networks are created from a single catalog that spans different periods, larger cell sizes than this value can create similar networks, and smaller cell sizes could be able to capture the dynamics of seismicity over time.

In the work of Pavez-Orrego and Pastén [55], a survey in Norway was also made looking for a cell size to construct earthquake networks that is independent of location errors. It is concluded that it is not easy to determine the best range of cell sizes, although some relatively small sizes are suggested so that the network is not homogeneous, that is, that the γ exponent of the power law of the degree distribution is greater than 1.

Also, Pastén et al. [56] evaluated seismicity in Chile before and after an $M_w=8.3$ earthquake using 3D earthquake networks with cell sizes of 5 and 10 km. In that work, changes in the networks could be seen, particularly in the exponent of the power law of the degree distribution, using 5 km cells, whereas 10 km cells did not detect significant changes in these distributions. Table 5 presents a synthesis of the works related to the cell sizes in earthquake networks.

Table 5: Previous studies about cell sizes in earthquake networks

Reference	Region	Cell size	Cell size criteria	Control parameters	Results
Abe & Suzuki [54]	California $\Delta\text{Lat: } 11.4^\circ$ $\Delta\text{Long: } 11.4^\circ$	5 km, 10 km and 20 km	2D and 3D cells $l_2 = L / (L_{\text{lat}} * L_{\text{lon}})^{0.5}$ $l_3 = L / (L_{\text{lat}} * L_{\text{lon}} * L_{\text{Dep}})^{0.33}$	Clustering < 0.85	$l_2 = 0.04 \sim 0.05$ is the minimum value where clustering ~ 0.85
	Japan $\Delta\text{Lat: } 31.4^\circ$ $\Delta\text{Long: } 35^\circ$	5 km, 10 km and 20 km			
Abe et al. [44]	Irán. $\Delta\text{Lat: } 20^\circ$ $\Delta\text{Long: } 27.6^\circ$	20 km			
	Chile $\Delta\text{Lat: } 6.5^\circ$ $\Delta\text{Long: } 4.5^\circ$	20 km			
Lotfi & Darooneh [53]	Iran. $\Delta\text{Lat: } 20^\circ$ $\Delta\text{Long: } 24^\circ$	$\sim 4 \text{ km} - \sim 20 \text{ km}$	2D Cells Resolution N Latitude divided in $12 * N$ Longitude in $10 * N$	Degree distribution power law $\gamma > 1$ Clustering < 0.85	Cells between 5 and 40km do not saturate Clustering, networks are not random according to the degree distribution.
Pavez-Orrego & Pastén [55]	Norway. $\Delta\text{Lat: } 7^\circ$ $\Delta\text{Long: } 9^\circ$	5 km - 20 km	2D and 3D cells between 5 and 20 km.	Power law degree distribution > 1	In cells between 8 and 16km there is no homogenization in the power law exponent.
Pastén et al. [56]	Chile. $\Delta\text{Lat: } 6.5^\circ$ $\Delta\text{Long: } 5^\circ$	3D cells - 5 km and 10 km	N/A	Detection of changes in the degree distribution power laws after the 2015 Mw=8.3 Illapel earthquake	Networks with 5km cells detect changes. 10km cells don't detect.
Leon et al. [4]	Colombia $\Delta\text{Lat: } 18^\circ$ $\Delta\text{Long: } 17^\circ$	2D cells between 10 and 100 km	N/A	Small-worldness.	10km cells detect zones with greater seismicity, and 100km cells detect more zones.

Previous work on cell size has focused on characteristics such as the clustering coefficient or the γ exponent of the power law that appears in the probability distribution of degree. Some other elements can be added to this discussion, such as the *small-worldness*, *entropy*, *complexity*, *disequilibrium*, and *divergence* metrics.

2.1.5. Conclusions and Discussion

From all these elements, the possibility of building a seismicity model from the perspective of complex networks is raised. The following are questions that remain open:

What is the best cell size to create an earthquake network of the type “cell as a node”?

From the representation with all the links (both with loops and multiple links between nodes) in the cells-as-nodes earthquake networks, how much can be predicted about the seismological activity? And how much from the simplified representation that evaluates the small-world characteristics of earthquake networks?

Entropy, complexity, and divergence, as estimated by Wiedermann et al. [34], have not been evaluated in detail in earthquake networks. How much information do these parameters provide about the seismological activity?

In seismology, the analysis of the dynamics of mechanical stresses (internal forces within the earth) resulting from the seismological activity is of great interest. How can a model that relates earthquake networks and the mechanical stress release due to the occurrence of earthquakes be created?

Creating a model that relates stress release with earthquake networks can be very complicated, and it would be interesting to devise some first steps to outline a model of this type. However, based on what is already known about earthquake networks, another research question can be raised that brings together several of these issues in this way:

How can a dynamical model of seismicity be developed from the representation of an earthquake network?

The research question has several questions implicit that should be answered. The first has to do with what are the best ways to create earthquake networks and what information each one delivers. The problem could be limited by reviewing only the representation of networks of the cells as nodes type, and the problem of cell size should be reviewed. The type of networks constructed has to be reviewed; hence the statistical divergence used by Wiedermann et al. can be used as a similitude measurement for the three main types of networks. Also, it must be answered which are the more appropriate measures from the networks to create a dynamical model of seismicity. Many have been used in previous works, but other measures can be used such as the Renyi's entropy used by Duan [33], and the Tsallis-Shannon entropy

used by Yulmetyev [57]; or based on the distribution of other measures such as centrality, clustering, or path lengths, which can involve different scenarios (in this respect it can be remembered that the entropy seen from Tsallis' vision generated the branch of non-extensive statistical mechanics [27]). The proposed model should somehow describe some characteristics of the dynamics of seismicity, so the measurement of time in this sense becomes important in the formulation of the model. Although it cannot be specified from now on whether specific seismic events can be predicted with the proposed model, it is expected that with this model, some kind of behavior can be foreseen; and this is already a sufficient motivation to say that this research is relevant: every step that is taken in the way of understanding the seismic phenomenon and that can announce a seismicity behavior will be very important for science and in particular for the nearby population to the region studied. It is expected that the proposed model can be used in different regions, and just as with the Gutenberg-Richter and Omori laws, some parameters may change depending on the region or the type of events that are presented, so it is expected that the model can characterize a seismic region and allow comparison with the model applied to the seismicity of other regions. In this regard, the question also arises of which regions should be analyzed to raise the seismicity model.

2.2. Artificial Intelligence in seismology and complex networks

2.2.1. AI applications in seismology

Seismology is a scientific field that encompasses a wide range of processes, from signal processing and phase picking in seismograms to source location, earthquake source modeling, event classification, and event forecasting. The use of ML in seismology has been reviewed in several publications such as the ones by Jiao and Alavi [58], Mousavi and Beroza [59], Kubo et al [60], among others. In Bergen et al. [61] a compilation of ML methods and their applications has been presented as in Figure 3.

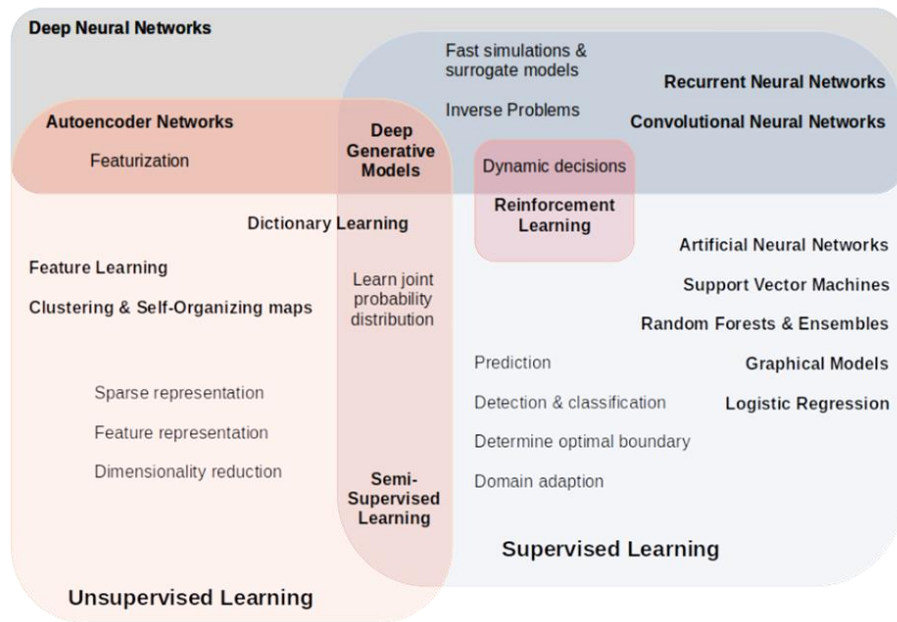


Figure 3: ML methods and their applications. From [61].

In [61] is stated that ML in solid Earth geosciences can be grouped in three main categories: *automation*, *modeling*, and *inverse problems*:

- Automation applications are meant to perform complex tasks that cannot be described by a set of explicit commands. In the case of earthquake seismology, analysis most of the time is performed on large-scale data and the tasks are sometimes repetitive and require the eye of an expert, such as earthquake detection, phase picking, or similar waveform searching.

- Modeling applications try to create a representation that captures the relationships and structure in a dataset. In earthquake seismology this means for instance the relationship between source parameters and peak ground acceleration for ground motion prediction, or the relationship between features and outputs such as magnitude or time of occurrence for next events.
- Inverse problems connect, as stated in [61], “*observational data, computational models, and physics to enable inference about physical systems in the geosciences*”. In earthquake seismology this means how to convert observations to results such as earthquake location or full source modeling which is to model the faults and movement that generated the earthquakes.

The review presented by Kubo et al. [60] focuses on four major categories for applications of ML, as shown on Table 6.

Table 6: Applications of ML in earthquake seismology. From [59].

Earthquake catalog development	Earthquake risk evaluation & seismicity analysis	Ground motion prediction	Application to geodetic data related to crustal deformation
• Catalog pipeline ○ Event detection/classification ○ Arrival time picking ○ Phase association ○ Earthquake location ○ Quality control • Similar waveform searching • Focal Mechanism Analysis • Paleoseismic record analysis	• Rupture forecasting of laboratory earthquakes • Risk evaluation of induced earthquakes • Other applications in seismicity analysis	• Prediction of ground motion intensity and times series from features and time series • Related studies: Approach to imbalanced ground-motion data	• Clustering analysis • Detection of geodetic signals due to seismic/aseismic phenomena • Modeling & prediction of crustal deformation

About the four categories outlined in [60], seismicity analysis to forecast events have been tackled by a Deep Learning based seismicity model.

Mousavi et al. [59] present a very nice compilation of publications about applications about ML applications in seismology. Figure 4 shows the ratio of ML approaches for each seismological task.

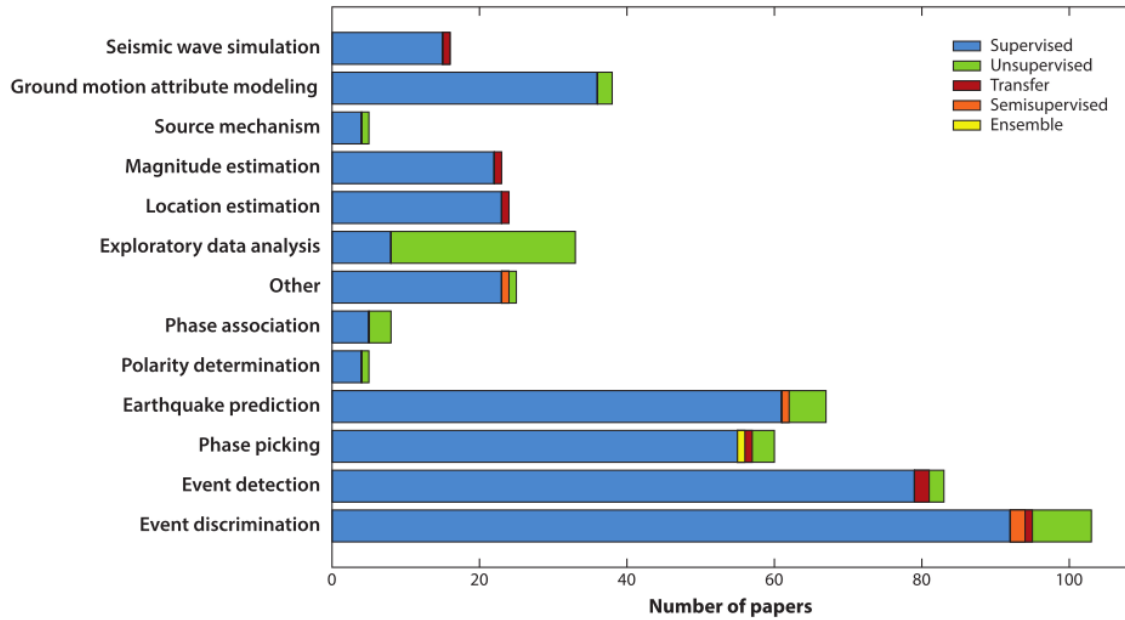


Figure 4: The ratio of various ML approaches used for each seismological task. Figure taken from [59].

As shown on Figure 4, much of the applications have tackled signal processing tasks such as event detection, phase picking and event discrimination; and earthquake prediction has also become a major application of ML in seismology. Although there is skepticism about earthquake prediction [62], there are many attempts to predict earthquakes using ML. Banna et al. [63] have compiled an extensive review about this matter, and state that earthquake prediction can be classified into short-term and long-term processes. According to Banna et al., the prediction of earthquakes follows a model as described in Figure 5.

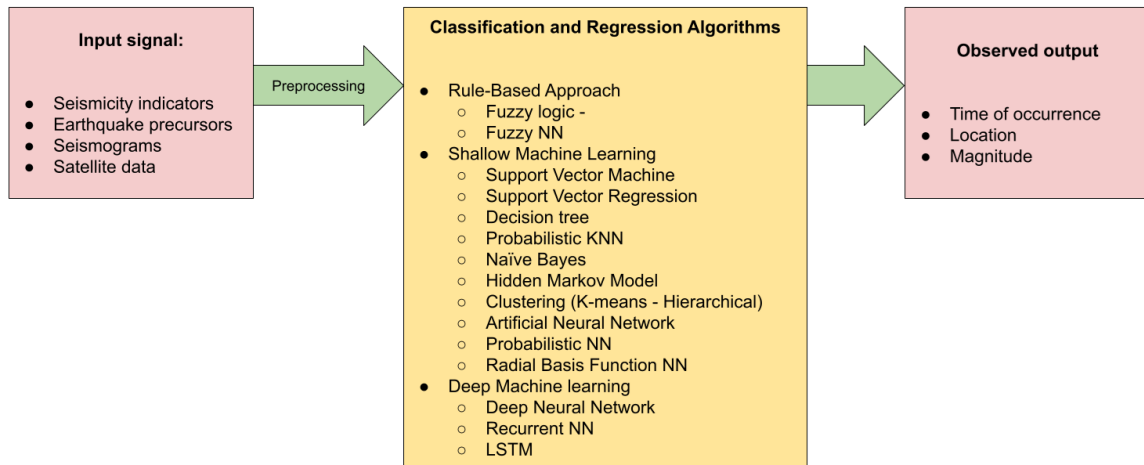


Figure 5: General earthquake prediction model.

The type of input signals mentioned on Figure 5 refer to the type of data available to train an ML algorithm. Numerous studies have highlighted the effectiveness of various ML algorithms, including Random Forest Regressors, Neural Networks, PhaseNet, and Deep Learning models; using historical seismic data, geographical information, and seismic time series to predict earthquake characteristics like magnitude, depth, and location. Also, incorporating ML techniques like Support Vector Machines, Convolutional Neural Networks, and K-Nearest Neighbors alongside regression methods can further enhance these predictive capabilities. This paves the way for early detection of seismic events on a global scale and facilitates real-time monitoring of seismicity in specific regions [64], [65], [66], [67].

About the metrics of the forecast, Banna et al. [63] considered a classification of forecasting performance following accuracy (for classifications) and root mean square error (for regressions) as shown on Table 7.

Table 7: Classification of forecast performances for classification and regression in seismology.

Class	Accuracy (classifications)	Root Mean Square Error (regressions)
High	$75\% < \text{Acc} \leq 100\%$	$0 < \text{RMSE} \leq 25\%$
Medium	$25\% < \text{Acc} \leq 75\%$	$25\% < \text{RMSE} \leq 75\%$
Low	$\text{Acc} \leq 25\%$	$\text{RMSE} > 75\%$

In the state of art presented by Banna et al., there are various cases with high performance in forecasting, using ML techniques such as SVM, Neural Networks, Random Forest, Linear Regression or Deep Learning among other techniques; and using features such as Gutenberg-Richter Law parameters, seismic signals, source parameters, seismicity indicators, but no research has been made using earthquake complex networks.

2.2.2. Machine learning on Complex networks.

In the chapter 2 of the book “Graph Machine Learning” by Stamile et al. [68], on page 56, it is said that graphs can be analyzed at different levels: *node*, *link* and *graph*. On the node level, the analysis goal can be to classify such nodes between different classes. On the link level, the goal might be to predict links between nodes. And on the graph level the problem might be to classify different graphs between some classes. ML can be a great tool for these tasks, as well as some other such as generation of graphs, detection of communities in a network, or prediction of how a network might grow over time. The *learning representations* for graph-structured data define also how a ML algorithm can be used to perform a specific task.

Chami et al. [69] present a taxonomy on Graph Representation Learning, where they state that there are four main categories for such methods: *Shallow Network embedding*, *auto-encoding*, *Graph Regularized Networks (GRN)* and *Graph Neural Networks (GNN)*:

Network embedding can be viewed as a dimensionality reduction technique for graph structured data: graphs are mapped to a continuous domain using vectors that

represent local or global properties of the nodes. On the other hand, both GRN and GNN use the structure and the node features in the learning process, but they differ as GRN are often used on semi-supervised learning tasks, smoothing the differences between networks to test similarities between networks, whereas GNN have a broader application on supervised learning tasks.

In the same work, Chami et al. [69] introduce a *Graph Convolution Framework* (GCF), where they characterize the use of graph convolution layers to provide a generalization of standard convolutions on grids (in this case, the network itself) to graph-structured data. In Zhou et al. [70] a review of GNN methods and applications can be found. About the implementation of such models, Stamile et al. [68] present a book with an introduction of many models and their implementation on python using libraries such as NetworkX and StellarGraph.

2.3. Chapter Highlights

Earthquake complex networks can be used to model the seismicity. There are two ways of creating such networks: the *cells as nodes* way and *earthquake as nodes* way. Earthquake as nodes networks are often used to analyze the aftershock activity, and Cells as nodes networks are often used to analyze the seismicity in a region. Therefore, the *Cells as nodes* earthquake networks are used in this work. One of the problems for these networks is choosing the adequate cell size, so the resulting network can have enough information about the seismicity.

In seismology, ML techniques have been used extensively to process seismograms in order to detect, locate and classify earthquakes. Most applications are of the supervised type. However, the use of these techniques has increased in forecasting certain aspects of the seismicity, such as the expected time of occurrence, magnitude, or location; and accuracies above 75% or RMSE lower than 0.25 are considered as the result of high-performance models. Nevertheless, in these cases it is important to consider that the power laws involved in earthquakes might result in a data set highly skewed.

On complex networks, ML can be used for applications in three levels: node, link, and graph. On graph levels, the problems might be related to classifying the network, or to use many networks to predict a value in a regression way.

Many features of seismicity have been used in these forecast models, such as magnitude of events, Gutenberg-Richter parameters, location tags; but no model involving Complex Earthquake Networks has been implemented. This work, as far as it is known, would be the first seismicity model using earthquake networks and ML.

Chapter 3

Earthquake complex networks in Colombia

The first publication about earthquake complex networks in Colombia was made in 2018 by León, Bucheli and Valdivia [4]. This paper was a first glimpse into the seismicity in Colombia represented as Earthquake Networks, and these networks have been used to label earthquakes in a classification system, identifying nodes of interest [16]. However, this analysis fell short in answering some questions such as what the best cell size is to create networks and what criteria should be used to say that certain networks represent the seismicity better than others. This chapter presents a deeper analysis on the earthquake networks in Colombia, in order to find what is the best cell size to construct an EN; and what features can be used to characterize the seismicity.

Colombia is located in a high seismicity region, where the Nazca, Caribbean and South American tectonic plates converge [71], [72], [73], [74]. Most cities are in areas with intermediate to high seismic hazard levels; and as a developing country, this heightened seismicity poses great challenges in terms of infrastructure resilience and disaster preparedness.

For this thesis, the seismicity analyzed spans from 1973 to 2023 and encompasses a region bounded by latitudes -5° and 13° and longitudes -82° and -66° . Recognizing that seismicity analysis benefits from longer periods, the USGS-NEIC (United States Geological Survey - National Earthquake Information Center - <https://www.usgs.gov/programs/earthquake-hazards/lists-maps-and-statistics>) catalog was employed instead of the Colombian Geological Survey (SGC) catalog, as the SGC catalog contains events from 1993 onward, while the USGS catalog holds events dating back much earlier for this region. The seismicity catalog used in this study comprises events occurring between January 1, 1973, and October 11, 2023, totaling 7759 events. In Chapter 4 there is a more detailed description of this catalog.

Figure 6 illustrates the distribution of epicenters in the region, color-coding earthquakes according to depth and representing magnitude using circle diameter.

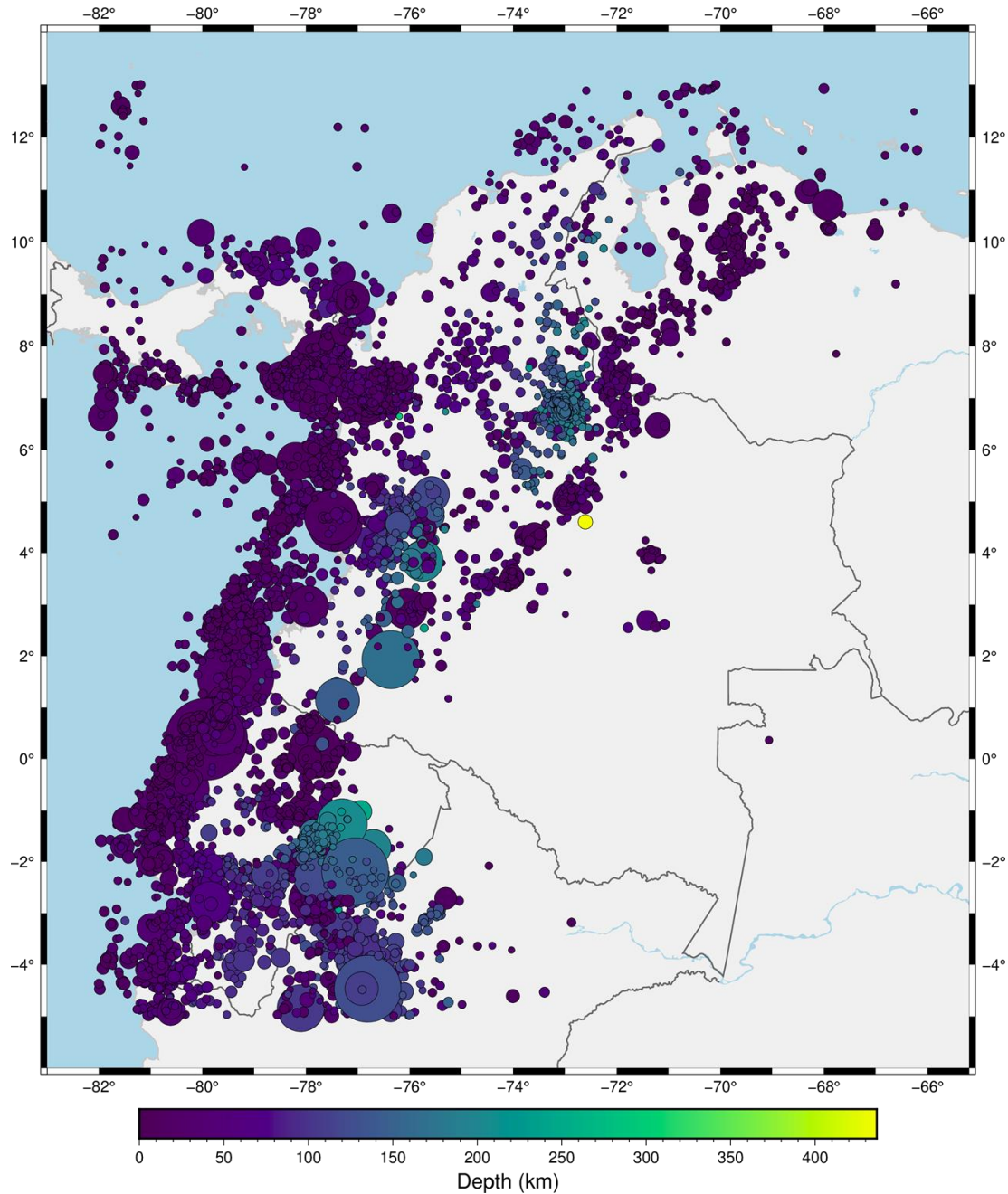


Figure 6: Seismicity in Colombia - January 1973 to October 2023.

Figure 7 shows the Gutenberg-Richter distribution for the region, where the X-axis represents the magnitude, and the Y-axis represents the logarithm of the number of events with magnitude lower than the X value. The dark line represents the logarithm

of the number of events with magnitudes greater than or equal to the magnitude specified on the X-axis. The red line represents the linear portion of this graph, considering only data for magnitudes exceeding the blue line, which in this case is 4.1 Mw. The line is fitted up to a magnitude of 7.5, as events with magnitudes exceeding this value are scarce. Based on this distribution, a completeness magnitude of 4.1 Mw was chosen for the catalog, indicating that all events with magnitudes greater or equal than 4.1 for the selected period are confidently included in the catalog. This magnitude serves as the cutoff for selecting events used to construct earthquake networks, resulting in a final catalog of 6262 events.

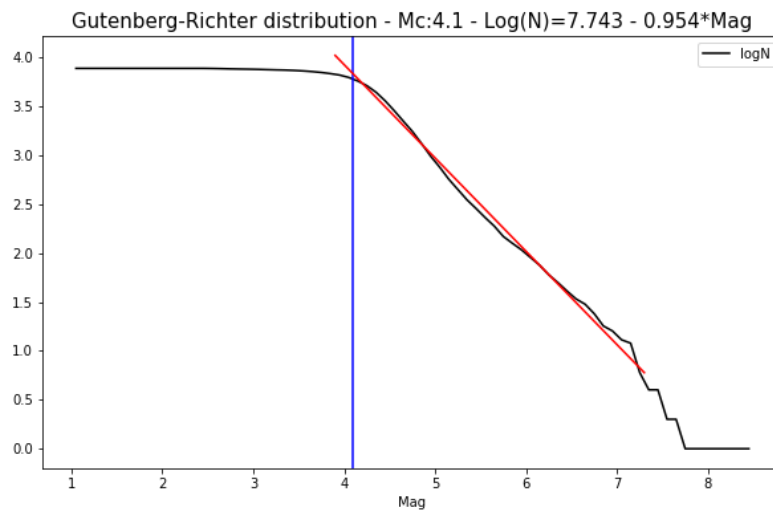


Figure 7: Gutenberg-Richter distribution of the seismicity in the region. A completeness magnitude of 4.1 Mw is assumed

Regarding the depth distribution of seismic events, Figure 8 shows a histogram of the distribution of the depths of events. In the figure, distinct ranges have been identified within the region. The first two ranges span from 10 to 49 km and 50 to 100 km depth which accounts for the activity along the pacific subduction zone: shallower seismicity towards the oceanic platform and deeper seismicity towards the continental platform. A third depth range, from 150 to 200 km depth, can be associated mainly with the Bucaramanga nest, which is located on an intermediate depth.

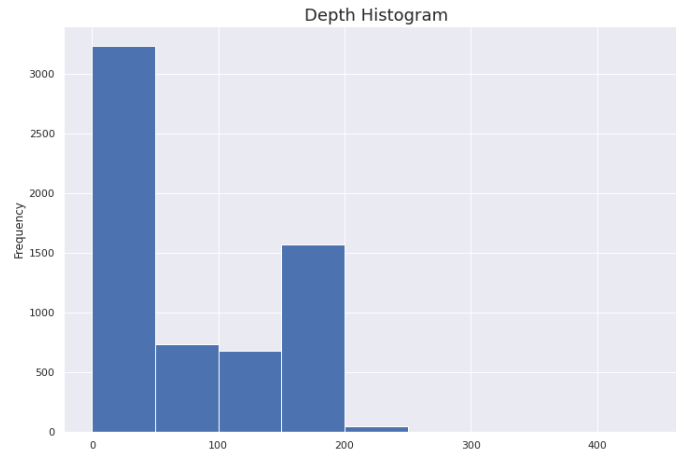


Figure 8: Histogram illustrating the distribution of seismic event depths.

To represent seismicity as an earthquake network, two methods were implemented: the method proposed by Abe and Suzuki [8] was implemented, as described in León et al. [14] (3D networks), and a variation where only latitude and longitude were considered, as implemented by León et al. [4] (2D networks). To create 3D networks, the region is partitioned into cubic cells with an edge length of Δd . Cells containing seismic events are selected as network nodes. A degree-kilometer conversion factor of 111 km/degree is assumed. The mapped region measures approximately 2000 km in length, 1900 km in width, and 400 km in depth. Figure 9 illustrates the earthquake network construction process. The region is divided into cubic cells, each labeled with an index (Figure 9a). Seismic events within the region are assigned to specific cells. Figure 9b depicts hypothetical events, labeled chronologically, and color-coded according to depth. Cells containing events are selected as nodes of the earthquake network, and connections are established between these nodes based on the temporal order of events (Figure 9c). This approach preserves information regarding the event sequence, location, and depth. For network analysis, loops (a node connected to itself) and multiple connections between any pair of nodes are disregarded; consequently, connections are binary. The method to create 2D networks is the same, but depth was not considered.

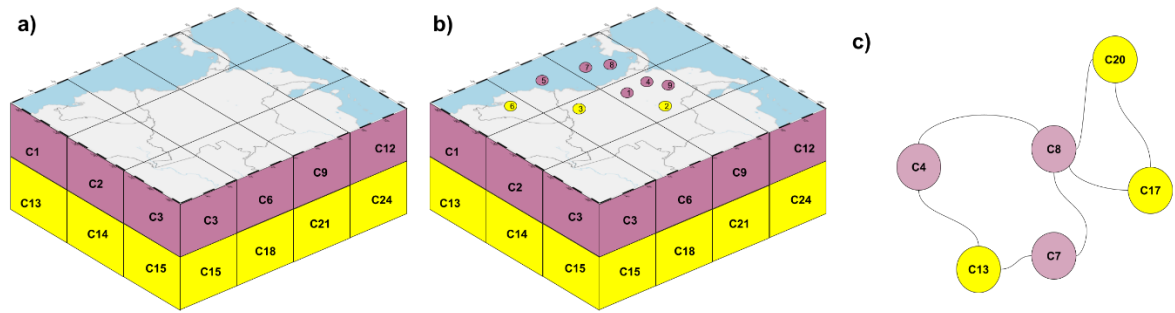


Figure 9: Diagram of the construction of an earthquake network from the spatiotemporal sequence of earthquakes; including the depth. In this example, the first event occurs on cell 8 and the second on cell 20, which is near the first but deeper, which results on a link between cells 8 and 20. The third event occurs on cell 17, so a link is set between cells 20 and 17. The rest of links follow the timely sequence in the same way as the first two links.

3.1. Earthquake Network analysis

The representation of seismicity using ECN should be able to capture its dynamics. Using *cells-as-nodes* networks, there are two parameters that must be assessed in order to create a good representation: the size of cells, and the time windows to capture changes on networks over time. In this section, the cell size will be assessed.

3.1.1. Cell size assessment

An initial analysis of an earthquake network in Colombia for the full catalog was made, in order to characterize network features for the region.

Network analysis typically employs metrics such as average degree, clustering coefficient, and average path length. The average degree represents the mean degree of all nodes in the network, where the degree refers to the number of connections a node possesses. This measure is also known as the network's connectivity. The clustering coefficient quantifies the tendency of a node's neighbors to be connected, indicating the presence of clusters or communities within the network. The path length represents the minimum number of connections required to traverse between two nodes, and the average path length across all paths indicates the network's information transmission efficiency. Additionally, for this study, other metrics were considered such as small-worldness, entropy, complexity, disequilibrium, and divergence (distance between networks).

Small-worldness Sw , as introduced by Humphries & Gurney [32], assesses whether a network exhibits small-world network characteristics. This measure must exceed 1 and is calculated using the formula:

$$Sw = \frac{C \times N \times \log N}{L \times k \times \log k} \quad (1)$$

where C represents the clustering coefficient, N represents the number of nodes, L represents the average path length, and k represents the average degree. As evident from the formula, small-worldness is proportional to the clustering coefficient and inversely proportional to the average path length.

Entropy, as defined in information theory, represents the amount of information associated with a system [75], [76]. According to Shannon's formulation, entropy is calculated using the formula:

$$H(P) = -\sum_i p_i \log_2 p_i \quad (2)$$

where p_i represents the probability of occurrence of event i in the system. In the context of networks, entropy for the degree quantifies the heterogeneity of the degree distribution. If all nodes have similar degrees, the entropy will be very low, approaching zero; conversely, if the number of nodes with small degrees is similar to that of nodes with high degrees, the entropy will be higher. Utilizing this entropy definition, Wiedermann et al. [32] proposed a complexity-entropy plane to distinguish between different networks. Complexity is estimated using the formula:

$$C(P_i) = D \times H(P) \quad (3)$$

where D is a measure of *disequilibrium*, calculated as the distance to an equilibrium state defined as an Erdős-Rényi network of similar size (equal number of nodes and similar connectivity). This distance between probability distributions is determined using the *Jensen-Shannon statistical divergence*, expressed as:

$$Q(P_1, P_2) = \frac{1}{N \times \log 2} \times \left[H\left(\frac{P_1 + P_2}{2}\right) - \frac{[H(P_1) + H(P_2)]}{2} \right] \quad (4)$$

where P_i represents the probability degree distribution of each network. Following this, *disequilibrium* can be defined as $D = Q(P_i, P_{er})$, where P_i corresponds to the degree distribution of i network, and P_{er} to a similar Erdős-Rényi network. It is important to remark that, as *disequilibrium* can be measured as a distance to an equilibrium state defined as a ER network, this type of distance can be measured respect other types of networks such as Barabási-Albert or Watts-Strogatz networks.

Observed Earthquake Networks can be compared with simulated networks based on different models. Most common models are the *Erdős-Rényi* (ER), *Barabási-Albert* (BA) and *Watts-Strogatz* (WS). ER networks are created by defining a number of nodes and assigning the probability that links exist between any pair of nodes i and j , which is defined as $p_{i,j}=k_{mean}/N$. In this type of networks, a normal distribution of degrees is presented.

BA are simulated networks of the scale-free type, with a power-law degree distribution [12]. The concept of *preferential attachment* is used to build the network from a small set of nodes connected as a star, and where each new node is added with a predetermined number of links connecting it to the initial set in such a way that the links are more likely to be towards the nodes with the highest degree. This is done until all the nodes defined as the target size are added to the network. This type of network configures hubs, a few nodes with a high degree, which connect to many nodes with a reduced degree.

WS are simulated networks of the small-world type, in which a high clustering coefficient and a short path length between each pair of nodes are favored [11]. This type of network is built from a ring of N nodes, that is, each node is connected to two neighbors. Then, each node is connected to $k-2$ neighbors, where k is the desired average degree for the network. Finally, from a reconnection probability β , each link is reassigned to some other node in the network.

In this work, all networks (ER, WS and BA) were generated and all metrics were calculated using the NetworkX Python library [39].

3.1.2. Experiments design

For each cell size, a set of 100 random networks are generated for the ER, WA and BA types; so these networks have the same number of nodes and a similar average degree than the earthquake networks observed. The BA set is simulated from the network size and the mean degree. It is worth to note that BA and WS networks generate networks which have rounded integer average degrees, and which will be the closest to the corresponding earthquake network. In the case of ER networks, for each size, an extra ensemble of networks is generated that will be used to measure the disequilibrium. For WS networks, the choice of the reconnection probability β is done as follows: an ensemble of 20 WS networks with probabilities β between 0.1 and 1.0 is generated, increasing by 0.1 each iteration; the small-worldness is calculated and then the probability β that generates the WS networks with small-worldness on

average most similar to the ECN earthquake network is chosen. Table 8 compiles the parameters with which the experiments have been designed, the description of the parameters, and the values they take.

Table 8: Parameters of the experiments

Parameter	Description	Range of Values
Cell size	Length of the cubic cell edge with which an earthquake network is constructed.	10 – 120 km.
Number of WS networks	Number of WS networks created with a particular β reconnection probability, in order to find the correct β for the experiment with the correct WS networks.	20
β	Reconnection probability to simulate WS networks.	0.1 – 1
Number of experiments	Number of simulated ER, WS and BA networks for comparison with a particular ECN network.	100
Network size	Number of nodes in the observed ECN network.	214-4345
Average degree	Average of the node's degree in the observed ECN network.	2.8 – 22.1

3.2. Results

For each observed network, the number of nodes, links, average degree, power-law exponent for the degree probability, clustering, average path length, small-worldness, entropy, disequilibrium, and complexity are calculated. The observed (EN) networks were constructed for cell sizes between 10 and 120 km. Figure 10 presents the number of nodes, Figure 11 the average degree, Figure 12 the clustering coefficient, Figure 13 the average path length, Figure 14 the small worldness and Figure 15 the γ exponent of the power law distribution of probabilities of degree for the networks constructed with different cell sizes. The full table of these results can be found in the appendices of this thesis.

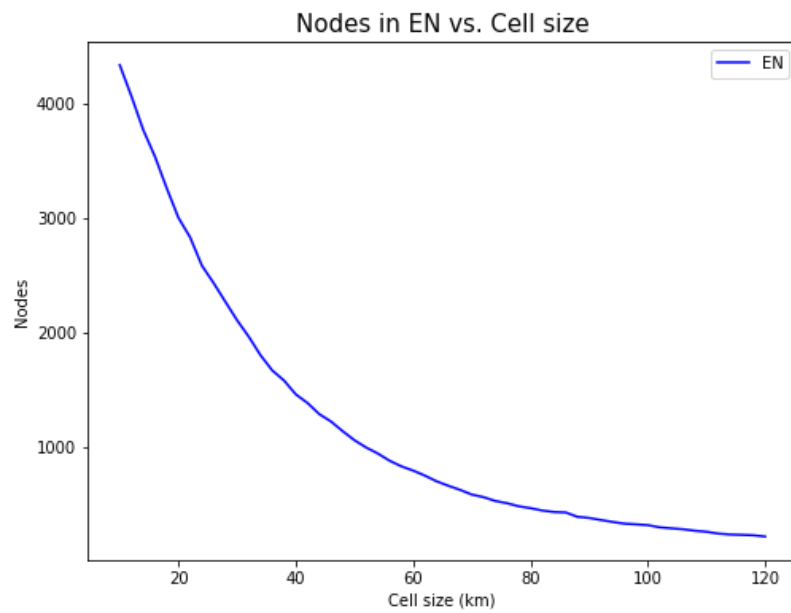


Figure 10: Number of nodes vs. Cell size.

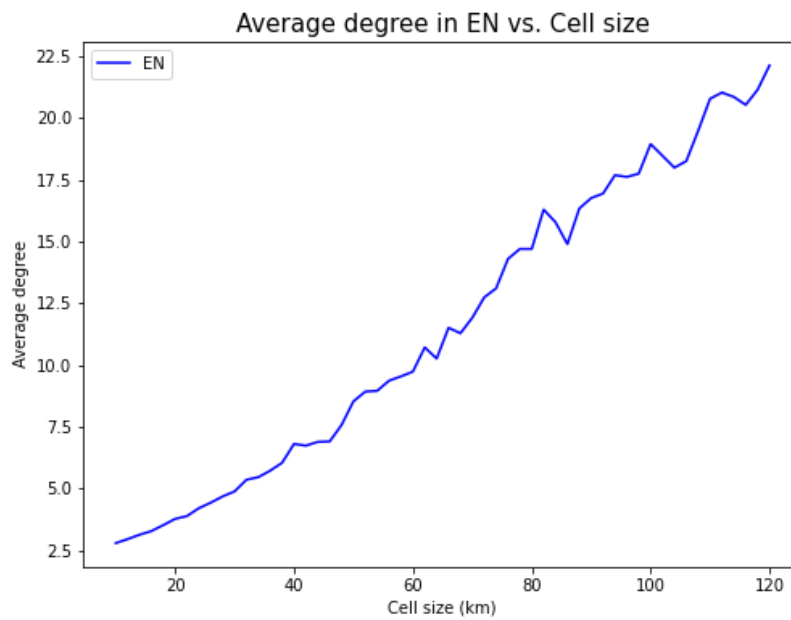


Figure 11: Degree of nodes vs. Cell size.

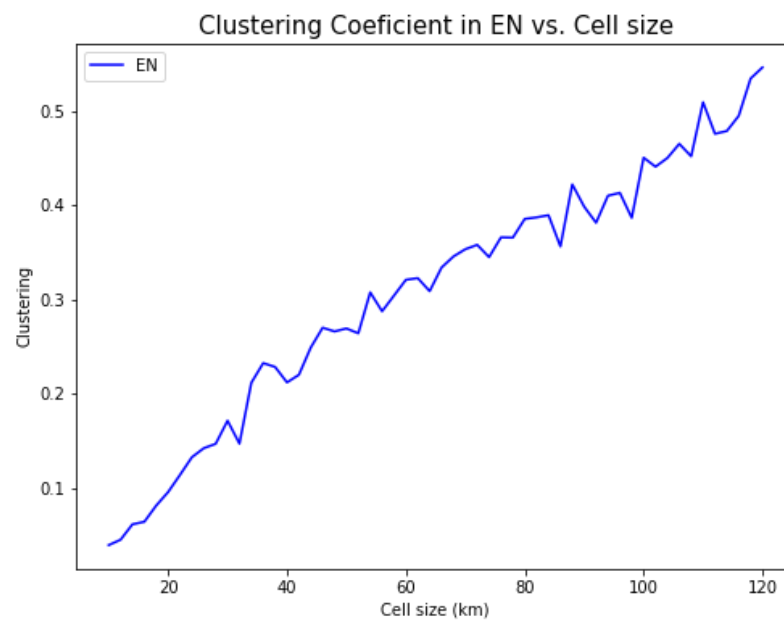


Figure 12: Clustering coefficient vs. Cell size.

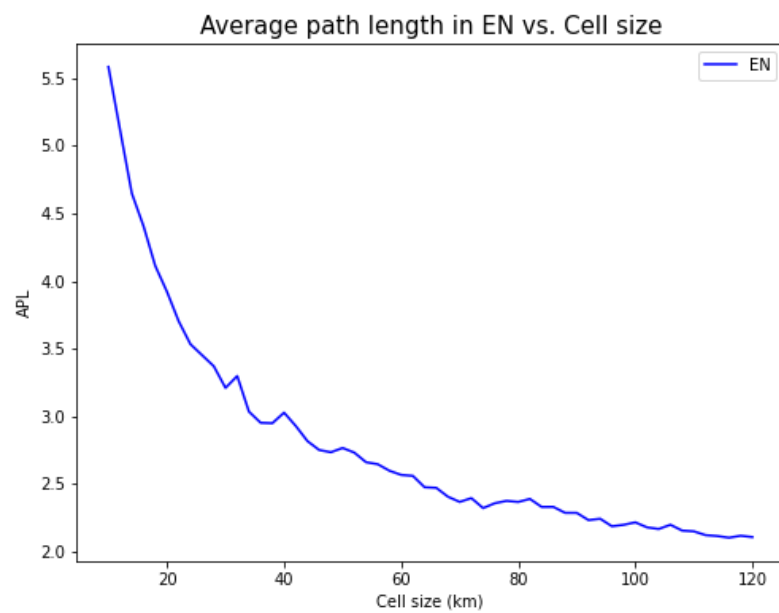


Figure 13: Average path length vs. Cell size.

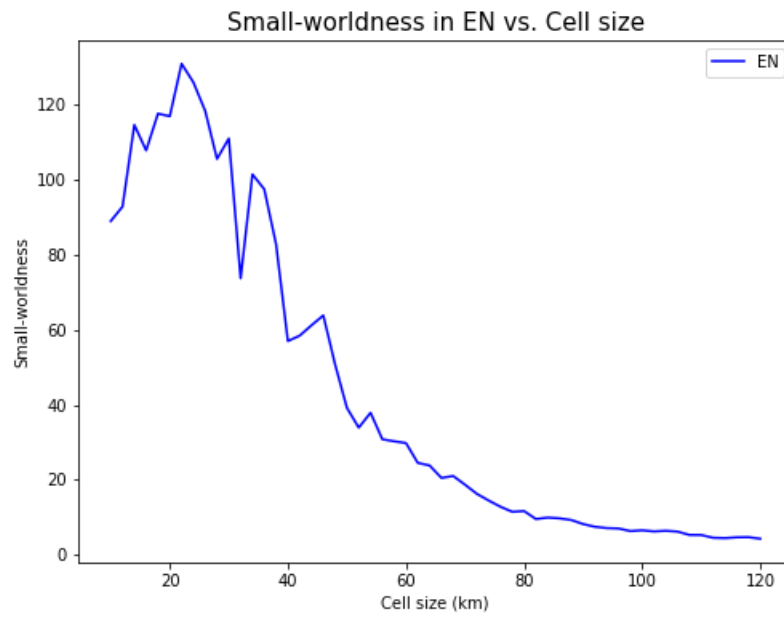


Figure 14: Small worldness vs. Cell size.

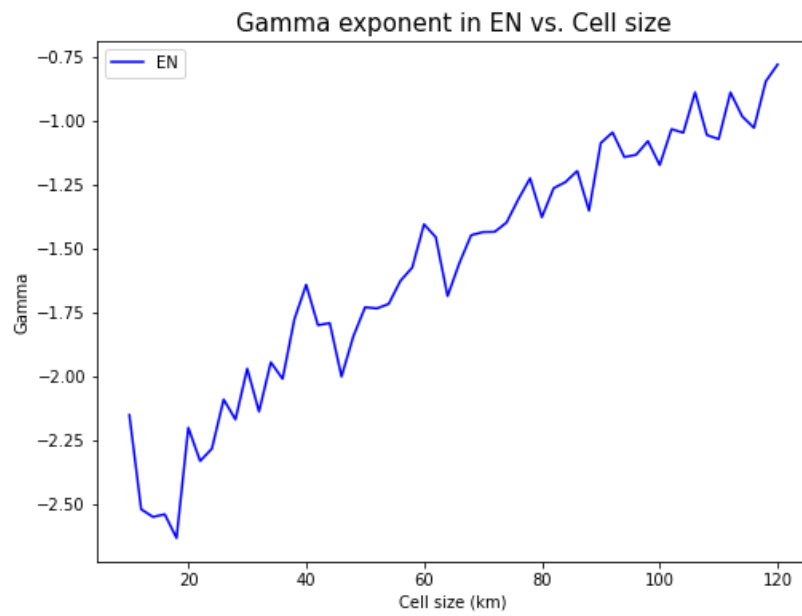


Figure 15: γ exponent of the power law for the degree distribution vs. Cell size.

Figure 10 shows that as the cell size increases, the observed networks will have fewer nodes but a higher degree. This is expected, since large cells collect more events, and will connect with more zones, increasing connectivity, which is evident in Figure 11 as it shows the average degree of the networks with different cell sizes. Figure 12 and Figure 13 show that clustering and Average path length are consistent with the effect of having less nodes but increasing the degree. Figure 14 shows that the small-worldness measure for all cases is greater than 1, so it can be said that all sizes generate networks with small-world characteristics. Networks with larger cell sizes have a greater connectivity, and present nodes that connect to each other with greater probability. In the small-worldness graph, all networks have a value greater than 1, with a maximum in cells of 22 km. After this cell size, the value falls exponentially. These values do not behave monotonically, which accounts for the complexity of the behavior of this metric in relation to the cell size.

Regarding the degree distribution; in [55] the degree distribution is reviewed in log-log plots of the degree of nodes k vs the probability of such values $P(k)$; and the slope γ of the linear distribution that appears corresponds to the exponent that relates the degree with their probability. Figure 15 shows that for the 10 km cell the slope is -2.15, while in the case of the 100 km cell the slope is -1.2. For this parameter, γ values closer to 0 indicate that the system is more stable and predictable, while values with higher absolute value indicate a lower probability of finding extreme states, which in this case would be nodes with a high degree.

About the comparison with synthetic networks, the calculation of the reconnection probabilities β to generate the WS network ensembles has a meaning for networks with cell sizes greater than 18 km. For smaller sizes, the small-worldness of the WS networks was very different from the ECN networks given that they had a clustering close to 0. For cells greater than 18 km, the beta probabilities were mostly 0.1, with a maximum of 0.3 in cell sizes 18, 20, and 32 km. In other words, the simulated WS networks have very little reconnection probability.

Given that the disequilibrium is measured as the statistical divergence with respect to similar ER networks, a divergence can be made also between ECN networks and the other BA and WS simulated networks.

Figure 16 shows the entropies of the ECN and the simulated ER, BA and WS networks when changing the cell size; Figure 17 the disequilibriums of these networks, and Figure 18 the divergence between ECN and ER, BA and WS synthetic networks.

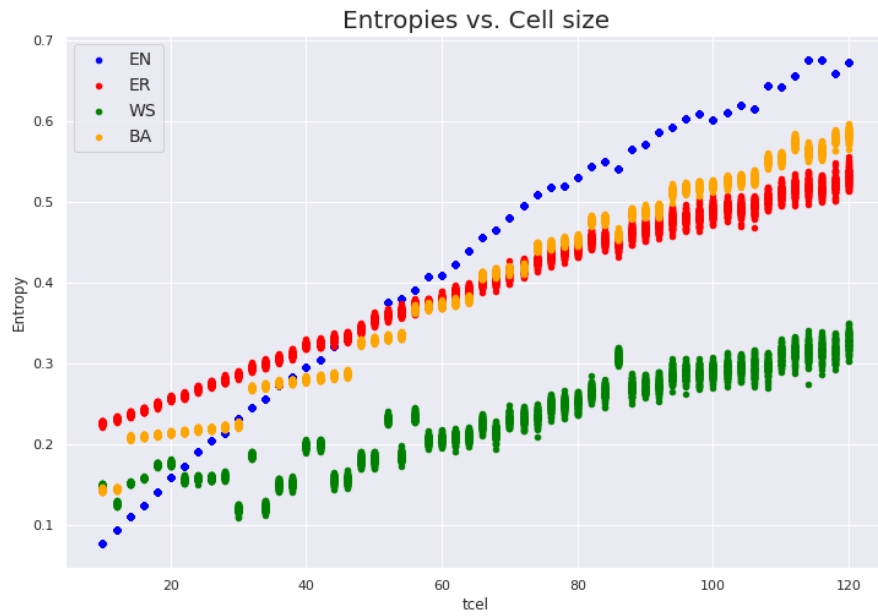


Figure 16: Entropies of ECN (blue dots) and simulated ER (red), BA (yellow) and WS (green) Networks vs. Cell size.

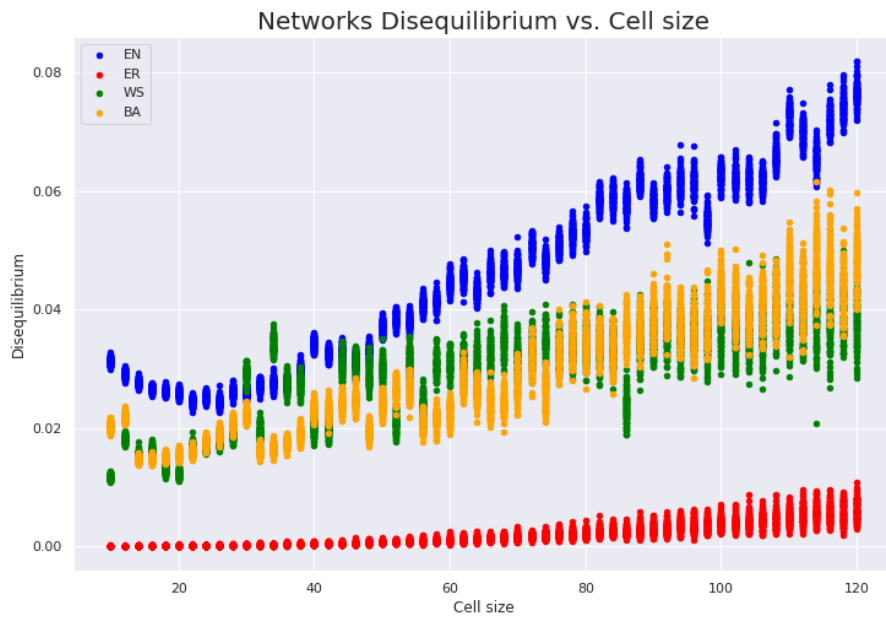


Figure 17: Disequilibrium of ECN (blue dots) and simulated ER (red), BA (yellow) and WS (green) Networks vs. Cell size.

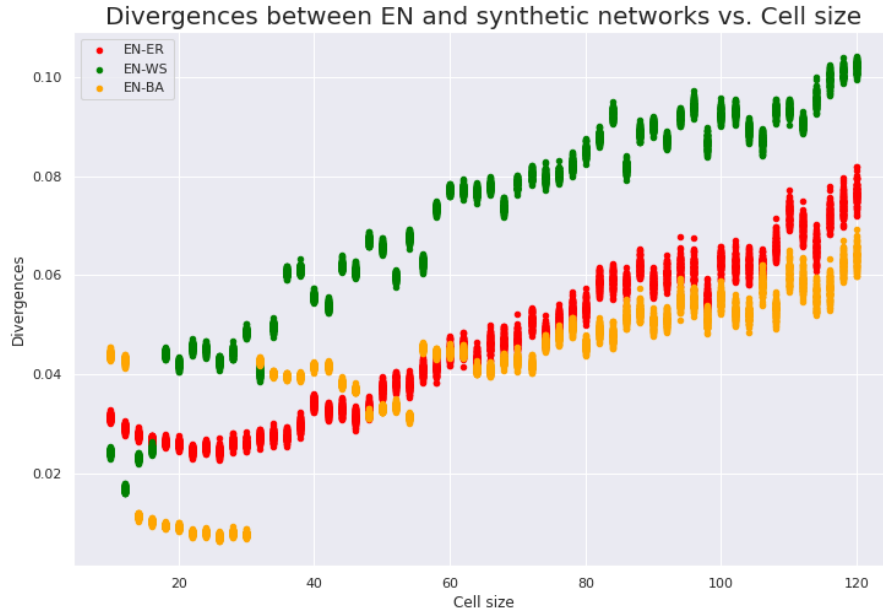


Figure 18: Divergences between ECN and simulated ER (red), BA (yellow) and WS (green) Networks vs. Cell size.

Entropy measures the information of a system in terms of the number of states it can have and the probability of each of them, which in this case will be the probability of node degree. An entropy is maximum if all states (degrees) of the nodes have the same probability, and decreases if there are few possible states, i.e., if there are only a few probable states. Figure 16 shows that WS networks have the lowest entropy due to the low beta probability and most nodes have the same state. ER networks have a higher entropy given that the randomness of connections leads to greater heterogeneity in the states. In BA networks, which are characterized by the power law in the degree distribution, there are more states than WS networks, and in some cases may have the same entropy as ER networks. With ECN networks, it is observed that there are ranges of cell sizes where the entropy is lower than the ER networks (cell sizes less than 50 km) and others where it is higher than for ER networks (cells greater than 50 km).

Regarding the disequilibrium of the networks, Figure 17 shows that as the cell size increases, so does the disequilibrium. This can be understood as if the experiments moved from a non-equiprobable distribution to one in which the degree distribution is equiprobable; that is, as if the networks with small cell sizes had distributions similar

to the power law and moved to distributions more similar to a normal distribution. This also occurs with the disequilibrium in WS and BA networks, although they do not increase in the same way as ECN networks. Up to cells near 60 km, the disequilibriums are similar with respect to WS and BA, indicating that perhaps up to these sizes the networks have more information.

About the divergences between networks, Figure 18 shows that for sizes between 20 and 30 km, the smallest values were found with respect to BA networks. Then, there is a leap, and the smallest divergence is with respect to the ER networks. Respect the WS networks, the divergence is greater, showing that although the networks have similar characteristics to small-world networks, they differ in their structure from the WS networks.

In [32], the entropy/complexity plane is proposed to distinguish different types of networks, where the complexity is calculated according to formula (3). Figure 19 shows the average values of the entropies and complexities in the entropy/complexity plane for each cell size.

Figure 19 shows that ER networks have the lowest complexity. This is because, while they have a higher entropy in general than the other types of networks, the disequilibrium is the lowest of all. For WS, BA, and ECN networks, as the cell size increases, so does the complexity, although in a more accelerated way in the case of ECN networks. In general, the growth is linear for each type of network, although the slope with which the ECN network grows is greater and for cell sizes of 60 km, there are increasingly differentiated complexities from those present in BA and WS networks. Figure 19 also shows that ECN and BA networks change their position in the plane, compared to the other types of networks. The sector occupied by ER networks is conditioned by the low complexity of these networks. ECN networks, with sizes between 20 and 50 km are in a sector of the plane very similar to the one occupied by all WS networks. BA networks follow a little more the region occupied by ECN networks, although the maximum complexity is lower than that of ECN networks.

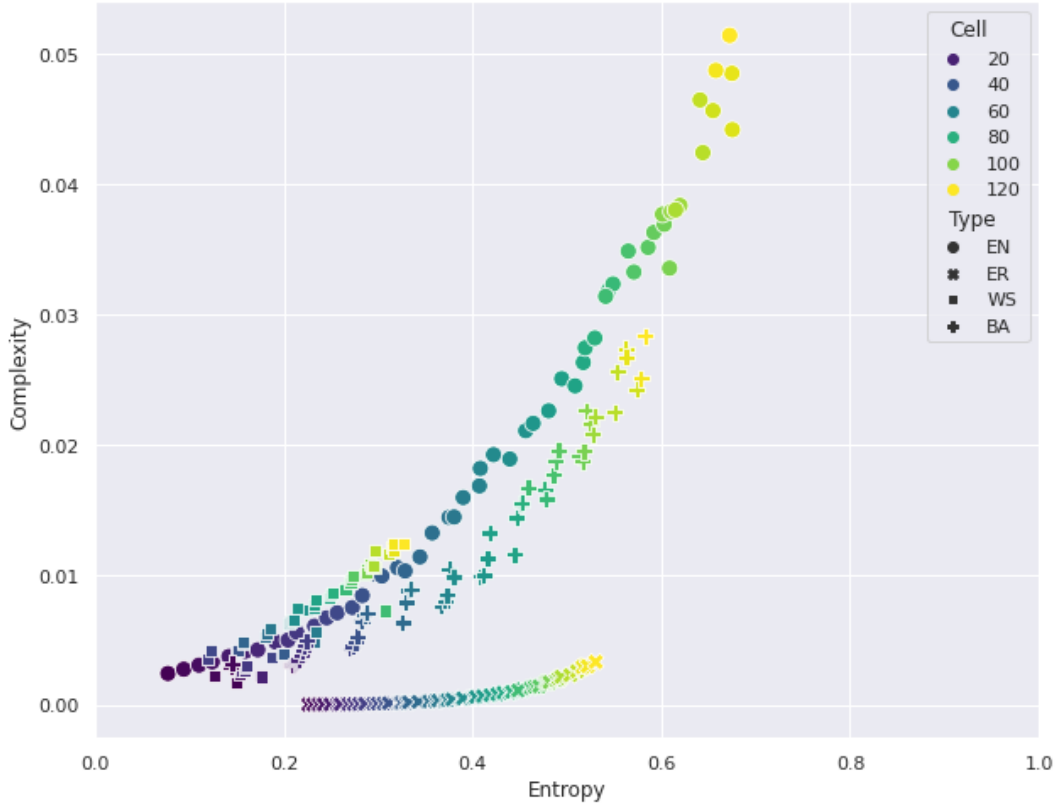


Figure 19: Entropy-complexity plane for ECN (circles), ER (exes), WS (squares), and BA (crosses). The color indicates the cell size with which the networks were generated. The X axis corresponds to entropy and the Y axis to complexity.

3.3. Discussion and conclusions

As can be seen in Table 5, many studies show that the clustering coefficient $C \sim 0.85$ becomes invariant when the cells become very large and exceed a certain threshold that can be estimated from the size of the monitored region and the cell sizes. This was first reported by Abe et al. in [54], and many previous observations of earthquake networks show this pattern. However, for the networks constructed in this work, there was no invariance detected regarding the clustering when increasing the cell size. It is possible that this difference with respect to previous studies could be related with the fact that other studies revised earthquake networks corresponding to a

seismicity where hypocenters were located in a limited depth range, while for the case studied in this paper, there is a diverse range of depths of the events.

Another invariant reported by previous studies, as was shown in Table 5, is the exponent of the power law that appears in the probability distribution of degrees. In this case, this exponent $\gamma \sim 1$ with very large cells [44], [54]; that is, there is a cell size beyond which the γ exponent no longer changes for the constructed earthquake networks. In the case of this study, this size is close to 90 km. On the other hand, the power laws that appear in the degree distribution for cell sizes smaller than 50 km are governed by exponents greater than 1.75. The review of the literature, in particular the case presented by Pastén et al. [56], showed that earthquake networks built with cell sizes of 5 km had a higher power law ($\gamma > 3$) than networks built with cell sizes of 10 km ($\gamma < 3$); and the conclusion of this work was that with smaller cells, a change in the degree distribution of the earthquake networks built before and after the 2015 Illapel earthquake could be evidenced. This conclusion could indicate that higher values of the γ exponent of power laws may be suitable for detecting changes before and after large earthquakes.

In previous works, as seen on Table 5, degree probabilities distribution and clustering were used as proxies for limits in cell sizes where invariants appear. In this work, it is seen that other variables might give other information, such as the entropy of the degree probability distribution of the nodes.

Earthquake networks, for all cell sizes, have characteristics of small-world and scale-free networks. The values of small-worldness, clustering, average path length, and disequilibrium show that these networks do not behave like random networks. Statistical divergences between earthquake networks and other types of networks show that for small values (between 20 and 30 km) the networks are similar to scale-free networks. Cells between 30 and 40 km configure networks with disequilibriums very similar to WS networks. Between 40 and 60 km, earthquake networks have smaller divergences with respect to ER than with other networks, but disequilibriums similar to those of WS networks. Although the disequilibriums are smaller, the networks have greater complexity than ER, and after sizes of 60 km, they continue to have similarities with BA networks. Earthquake networks have a disequilibrium similar to WS networks up to cell sizes close to 46 km, after which the disequilibrium grows above the other networks. BA networks after cell sizes close to 60 km present a disequilibrium very similar to WS networks. The high disequilibriums of ECN networks indicates that they have a more organized topology than the other types of networks.

When observing the map in Figure 6, it is observed that most of the seismicity is distributed in a homogeneous manner throughout the subduction zone in the Pacific and the Bucaramanga seismic nest, although there are some areas with little seismicity towards the Atlantic coast or the east of Colombia. For this reason, it could be thought that an earthquake network representative of seismicity should have an entropy that represents this condition, in which there is a limited range of states for which the probability would be higher. In this sense, the range of cells smaller than 50 km, which corresponds to systems with fewer probable states than ER random networks, would represent the seismic phenomenon in the region in a better way.

The cell size conditions the structure of the constructed earthquake networks, which is evident on the network's metrics. On smaller cell sizes entropy shows that earthquake networks tend to be more heterogeneous, that is, few states with high probability and many with low probability; the γ exponent that the ECN networks are more complex; and disequilibrium that the ECN networks have great differences with respect to ER similar systems. Also, the divergence between different types of networks shows that there are also ranges where the ECN networks are similar to BA networks (smaller than 30 km).

The different topologies found for earthquake networks when changing the cell size may favor different aspects which might be of interest. In [4] it was mentioned that bigger cell sizes might help to automatically identify many nodes or regions with great seismicity, whereas smaller cell sizes might help to highlight the most important seismic zone in the region. In this work, it was seen that regardless of the cell size, the earthquake networks have characteristics of small-world and scale-free networks, but when the cell size changes some features are enhanced or diminished. A proper cell size choice should be made according to the goal of the task to be achieved.

As future work, it would be interesting to review what information the entropy can provide from the distributions of other node measures, such as clustering or centrality; but it would have to be carefully reviewed how to construct such distributions, given that these measures in the nodes are not integers and the entropy would have to be estimated from continuous probability functions. A limitation of the work is that the definition of an optimal or universal cell size will depend on the context and the seismicity of each zone, however this is a contribution to the discussion of the cell size and other network metrics have been reviewed and proposed to aid in the choice of the cell size to build earthquake networks in the form of *cell as node*.

3.4. Chapter Highlights

Most of the highlights were commented on section 3.3, Discussion and conclusions. Regarding this study, a cell size of 50 km seems appropriate to create EN, as this size creates a network with characteristics similar to those of ER and BA networks, as seen on Figure 16, which works as a ground base of the behavior of the region; and it could be modeling changes in the seismicity when finding a behavior which differs from this ground base. What follows now is to split the catalog in smaller sets to create earthquake networks in time and use them in ML instances to model seismic features, which is the goal of chapter 4.

Chapter 4

Seismicity Models using Earthquake Complex Networks and Machine Learning

This thesis deals with the implementation of Machine Learning models to forecast some aspects about the seismicity in the region around Colombia using data from earthquake complex networks. There are some desired characteristics for these models: they must forecast some features of the seismicity, the models should be able to classify the ECN in different classes, or to make a regression of some value. This chapter evaluates first what characteristics of the seismicity should be used, and then the creation of different models to forecast such characteristics.

The desired model would be a *specialized model*, where features of the seismicity are input into a model which outputs some characteristic of the following seismicity, which might be for a single event or for the activity for a period. The flow of data should follow this:

$$\text{Input features} \rightarrow \text{Model} \rightarrow \text{Output}$$

; where the *input features* would be taken from Earthquake Networks and the output can be a label (High-Medium-Low Magnitude Event, id of a geographical sector, etc.), or a continuous variable (Magnitude of next event, time of occurrence, seismicity for next N events, etc.).

This chapter present an *Exploratory Data Analysis* to choose the input the features and the desired outputs, as well the training of different models to forecast the outputs. For each output different models were tested and the best results are shown. Finally, a compilation of highlights and conclusions are presented.

4.1. Exploratory Data Analysis

Exploratory Data Analysis, according to Tukey [77], is a detective work, about finding out what the data has to tell. There are many relationships known in seismology, but

nonetheless it is important to look with detail how is the data in order to make better interpretations from it.

4.1.1. Catalog contributors

In Chapter 3 the catalog of events was introduced, and Figure 6 showed the map of the region with circles representing the location of earthquakes and their magnitude represented by the size of each circle. The USGS-NEIC catalog has many sources, mainly from the WWSSN (World-Wide Standardized Seismograph Network [78]) and the GSN (Global Seismographic Network [79]) global networks. However, there are some other institutions that contribute to this catalog. A full list of contributors to the catalog can be consulted in the USGS-NEIC web site¹. As mentioned before, the seismicity catalog used spans from 1973 to 2023, encompasses a region bounded by latitudes -5° and 13° and longitudes -82° and -66° , the events have a magnitude greater or equal than 4.1, and encompass a set of 6262 events. Figure 20 shows the distribution of the seismological stations from GSN.

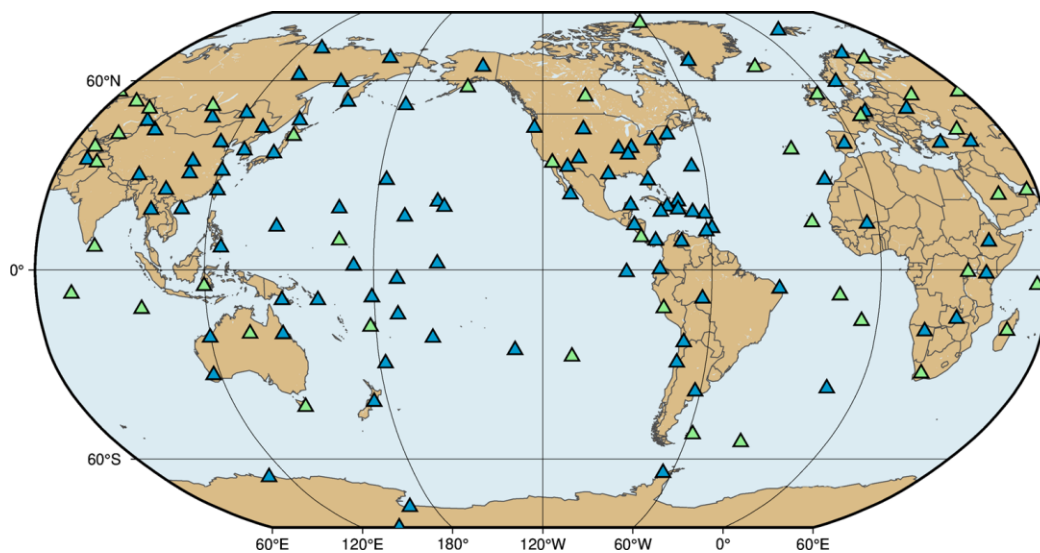


Figure 20: Distribution of Global Seismographic Network (GSN) stations. USGS GSN sites are shown in blue and IRIS/IDA stations are shown in green. Image from <https://www.usgs.gov/media/images/global-seismographic-network-gsn-stations>

¹<https://earthquake.usgs.gov/data/comcat/contributor/>

About the seismicity analyzed in this work, Figure 21 shows that the main contributor over time has been the global network (code *us* - blue bars), with some episodes where the Panama network (code *upa* - red bar), Colombian network (codes *rsnc/us_rsnc* – purple and light blue bars) and Costa Rican network (code *car* – fuchsia bar) barely contributed to the catalog.

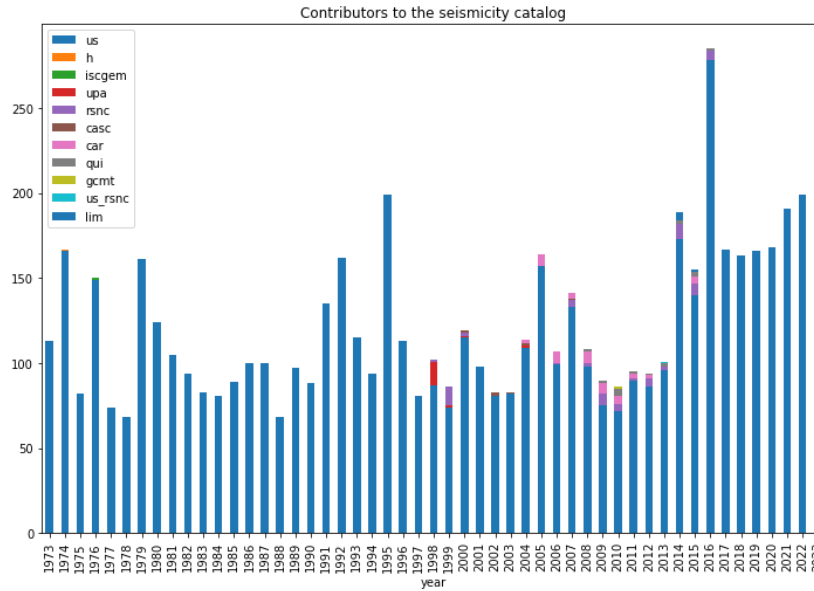


Figure 21: Contributors to the seismicity catalog over time

4.1.2. Seismicity Analysis: Energy and time.

Seismicity models strive to forecast something about magnitude or location of earthquakes. As the famous Charles Richter stated, there's a difference between magnitude and intensity: intensity measures how the ground's movement was in a place; and for a single earthquake, different places can have different intensities depending on the distance to the epicenter, or ground properties. Magnitude on the other hand, is a characteristic of the earthquake as a whole [80], and it should be equally measured independent of the place. According to Gutenberg and Richter, the seismicity quantifies the seismic activity in a region, and it can be calculated using the formula:

$$S = \frac{\sum_i^n E_i}{\delta Lat_0 \delta Long_0 \delta Depth_0 \delta Time_0} \quad (5)$$

, where E_i is the released seismic energy by an earthquake i , and δLat_0 , $\delta Long_0$, $\delta Depth_0$, and $\delta Time_0$ represent an interval of latitude, longitude, depth and time where all n earthquakes occurred. According to Kanamori [81], the magnitude-energy relation $\log E = 1.5M + 11.8$ was developed by Gutenberg and Richter [82], where energy E is expressed in ergs.

Figure 7 depicts the Gutenberg-Richter distribution in the region. As said before in the beginning of chapter 3, the Gutenberg-Richter distribution shows the logarithm of the cumulative number of events with magnitude equal or less than any of the values. The shape of the plot shows that there are more events with low magnitude than events with high magnitude, and it follows a power-law. A histogram of the magnitude of the events is shown on Figure 22.

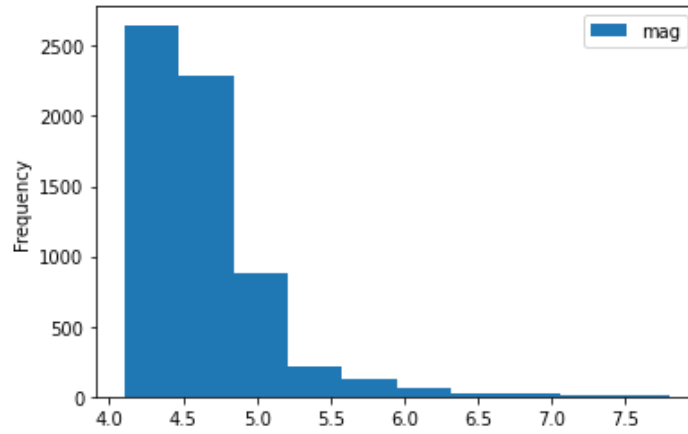


Figure 22: Histogram of the magnitude of the events in the catalog.

As shown in Figure 22, there is a high number of events with low magnitude and a low number with high magnitude. This observation is important, as it will derive in unbalanced data when trying to classify events between high and low magnitude events.

Figure 23 shows the distribution of magnitude of events over time and using the Gutenberg-Richter magnitude-energy relation the seismic energy released (in Jules) by each event on Figure 24.

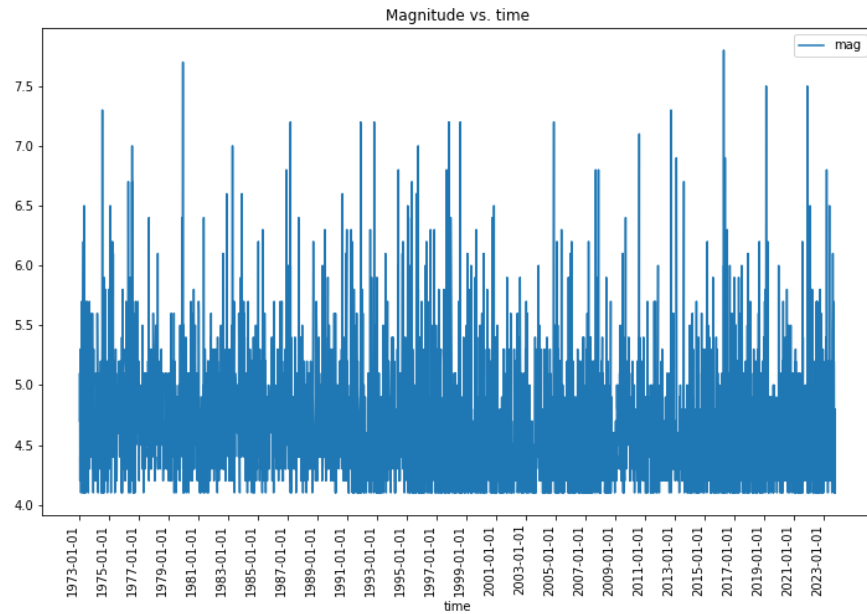


Figure 23: Distribution of magnitude of events over time

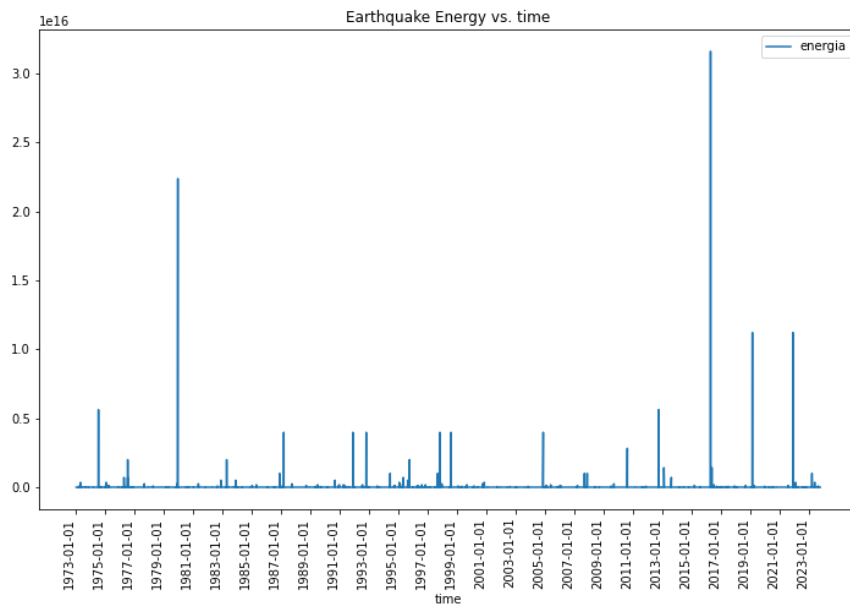


Figure 24: Seismic energy released by each earthquake over time

Figure 23 and Figure 24 show that most of the released seismic energy comes from only a few events. Seismicity, according to the Gutenberg-Richter seismicity formula

(5), can be calculated as the sum of the energy of n events that occurred on a region in a lapse of time. As the time lapse is important to define the seismicity, the time lapses (in days) between earthquakes are shown on Figure 25, and in Figure 26 a histogram of such time lapses.

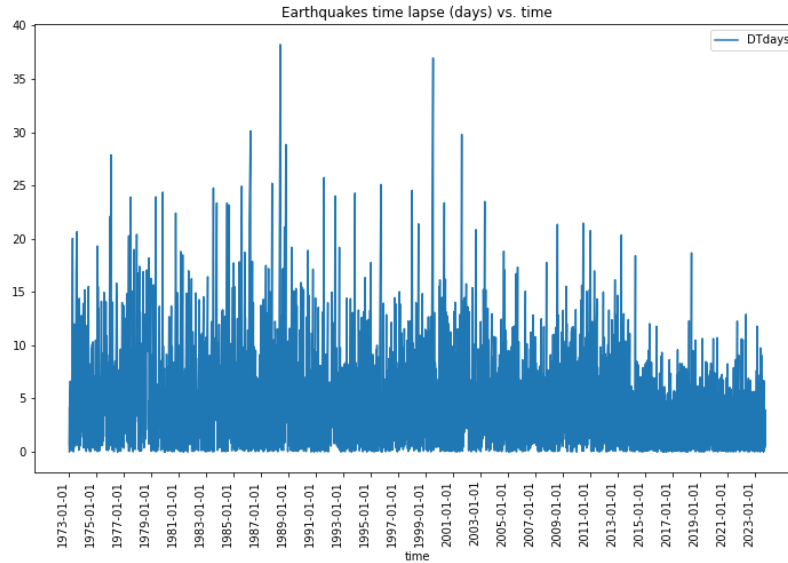


Figure 25: Number of days between consecutive earthquakes vs. time

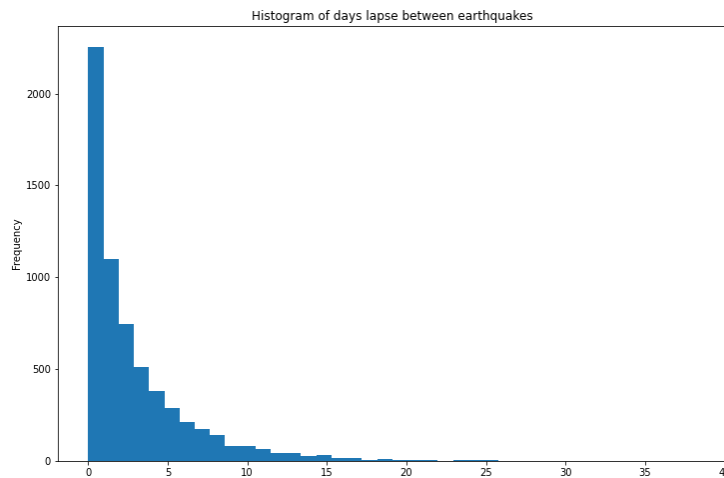


Figure 26: Histogram of time lapses in days between consecutive events

The mean time lapse between consecutive earthquakes is 2.96 days. However, Figure 26 shows that there is a great portion of events that occur on the same day; and Figure 25 shows that the time lapse can be greater than 35 days. If a set of consecutive

events is chosen to calculate the seismicity, not only the energy of those events must be summed but also the time must be accounted dividing the sum. To analyze the seismicity over time the following steps are performed:

- The full list of n events $[x_0, x_1, \dots, x_n]$ are ordered by their time of occurrence, where x_1 is the oldest event and x_n the last event.
- A set window $[x_0, x_1, \dots, x_{w-1}]$ of the first w events is used to calculate the seismicity according to the Gutenberg-Richter seismicity formula (5), summing their energy and dividing by the difference in time between the first and last event of the set.
- The window is moved $w/2$ events, so the next set window is $[x_{(w/2)}, x_{(w/2)+1}, \dots, x_{(w/2)+w-1}]$; and the seismicity is calculated as mentioned in step ii.
- The window is moved for the full list of events.

Figure 27 shows the logarithm of seismicity using windows of 10 events vs time, Figure 28 for windows of 50 events and Figure 29 for windows of 100 events. Figure 30 shows the time lapse in days for the windows of 10 events vs time, Figure 31 for windows of 50 events and Figure 32 for windows of 100 events.

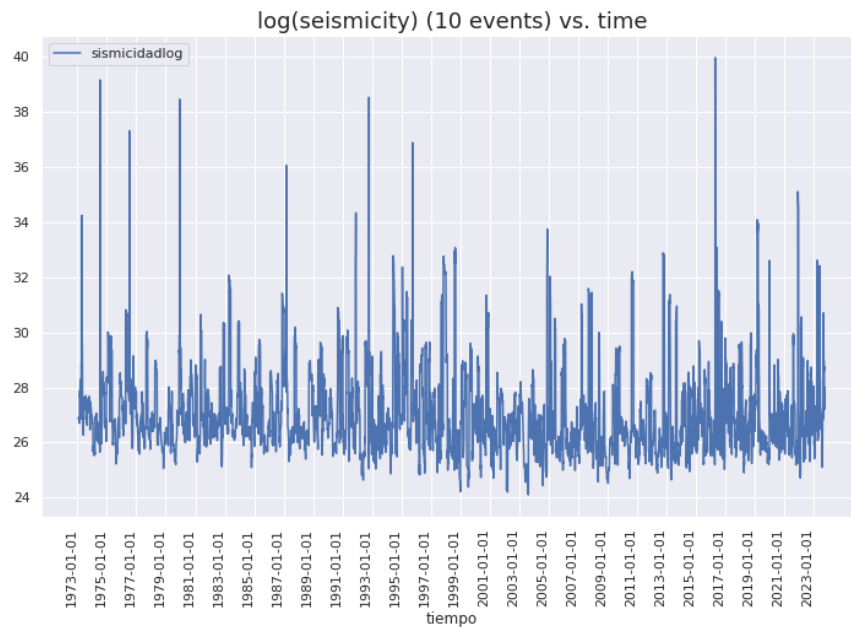


Figure 27: Logarithm of seismicity using windows of 10 events vs. time.

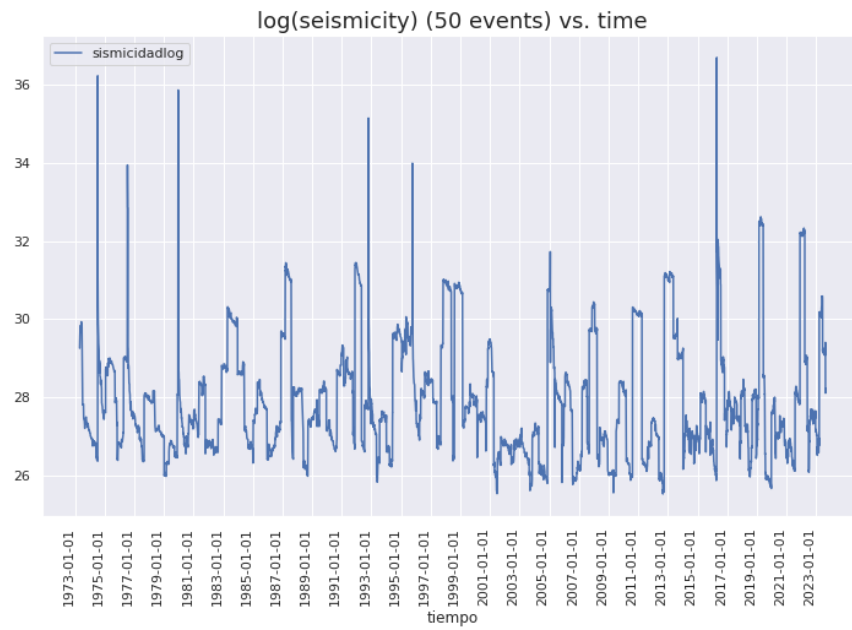


Figure 28. Logarithm of seismicity using windows of 50 events vs. time.

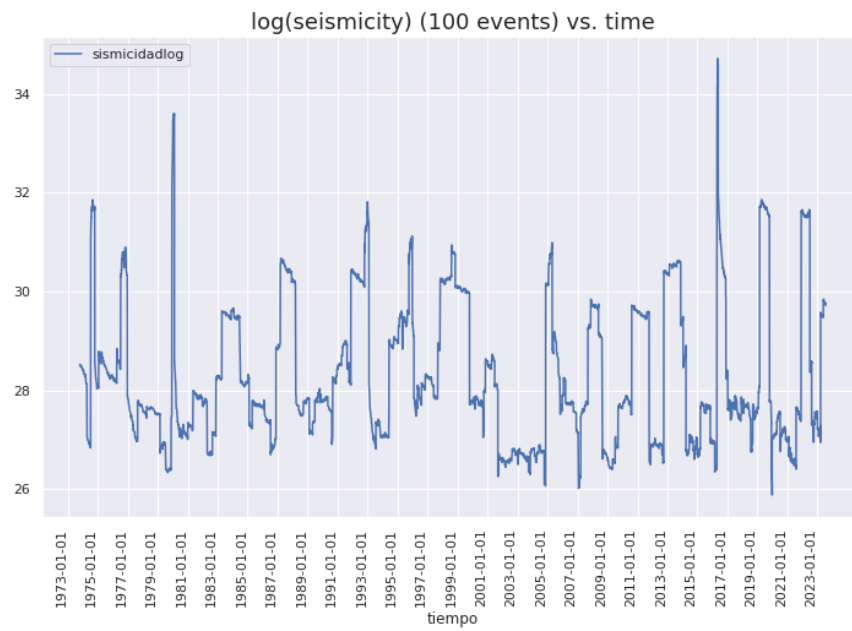


Figure 29: Logarithm of seismicity using windows of 100 events vs. time.

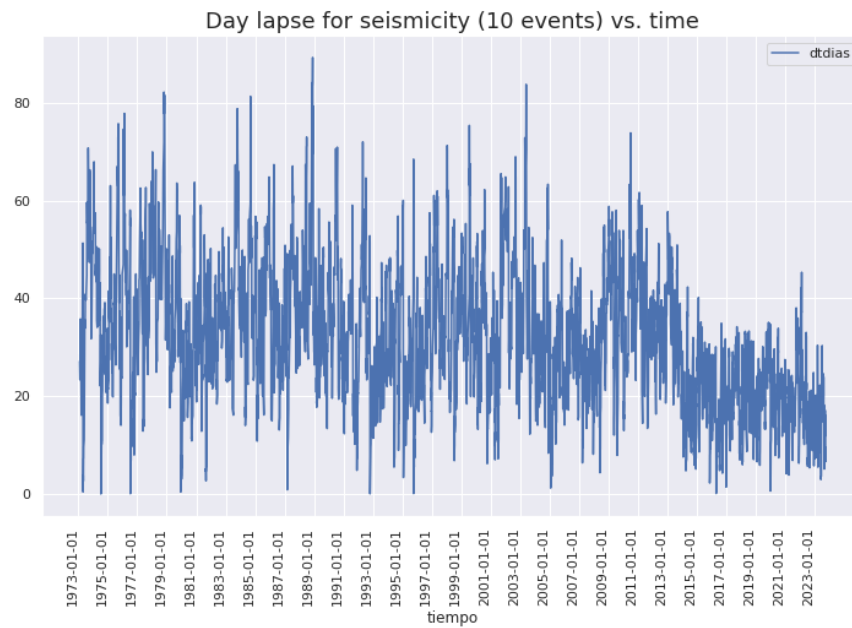


Figure 30: Time lapse for each seismicity using windows of 10 events vs. time.

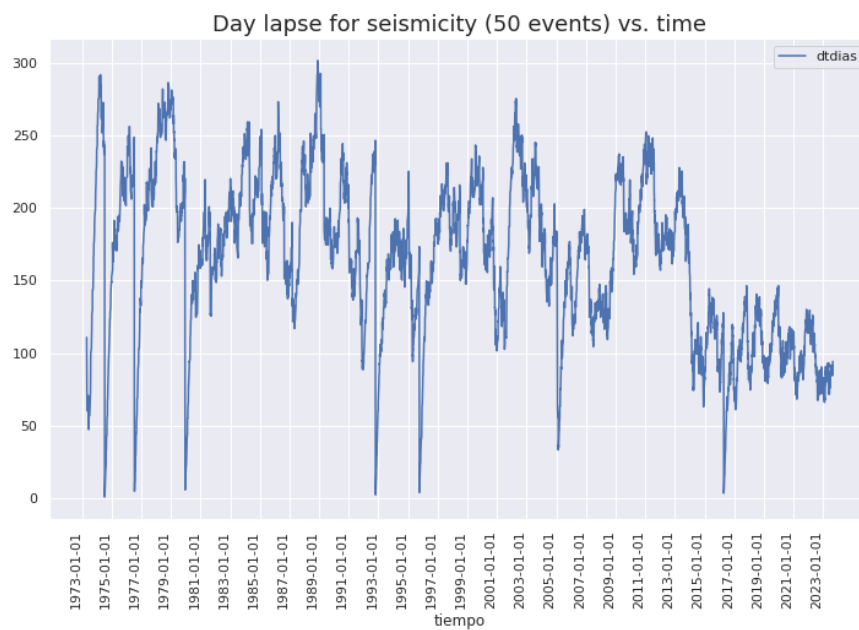


Figure 31: Time lapse for each seismicity using windows of 50 events vs. time

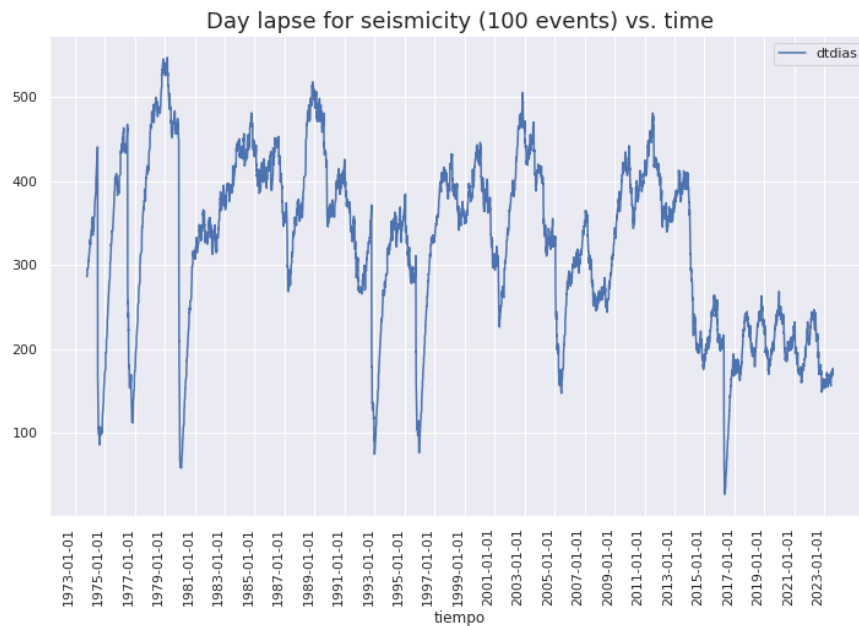


Figure 32: Time lapse for each seismicity using windows of 100 events vs. time

About the day lapse to calculate the seismicity using different windows, Figure 30 shows that a window of $n=10$ events derives in rapid changes for each timestamp and results in a noisy trend. Maximum lapse is near 90 days and many instances have lapses of few days, even less than a day. For windows $n=50$ and $n=100$ events, Figure 31 and Figure 32 looks similar, but maximum lapses times might be very extended for a window with 100 events (almost 550 days).

4.1.3. Geographic sectorization of seismicity.

An ideal forecast system would predict the location, magnitude and time when an event will occur. However, a simpler system which indicates the region where the next event will occur could be also of great use in order to be prepared. Figure 33 and Figure 34 show a proposition of a 4-sector geographic sectorization, and Figure 35 and Figure 36 a proposition of a 5-sector geographic sectorization.

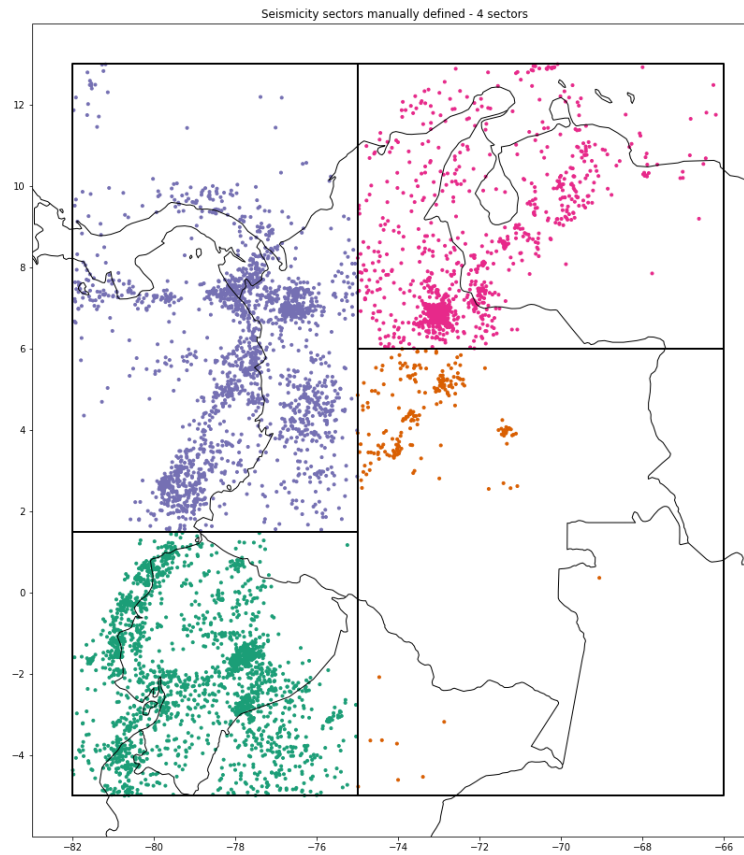


Figure 33: Definition of 4 seismicity sectors selected manually. In the map, the color of the dots indicates the assigned type according to the sector.

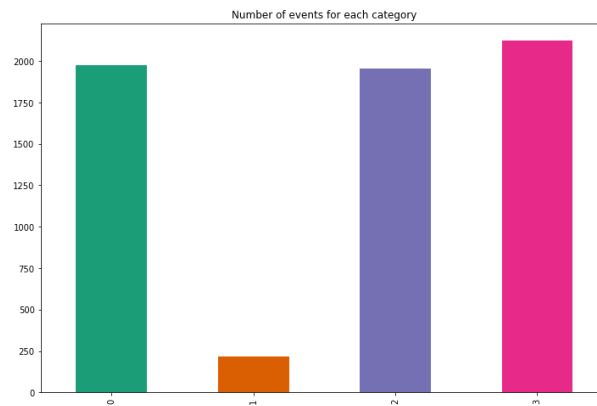


Figure 34: Definition of 4 seismicity sectors selected manually. The bar plot shows the number of events for each sector.

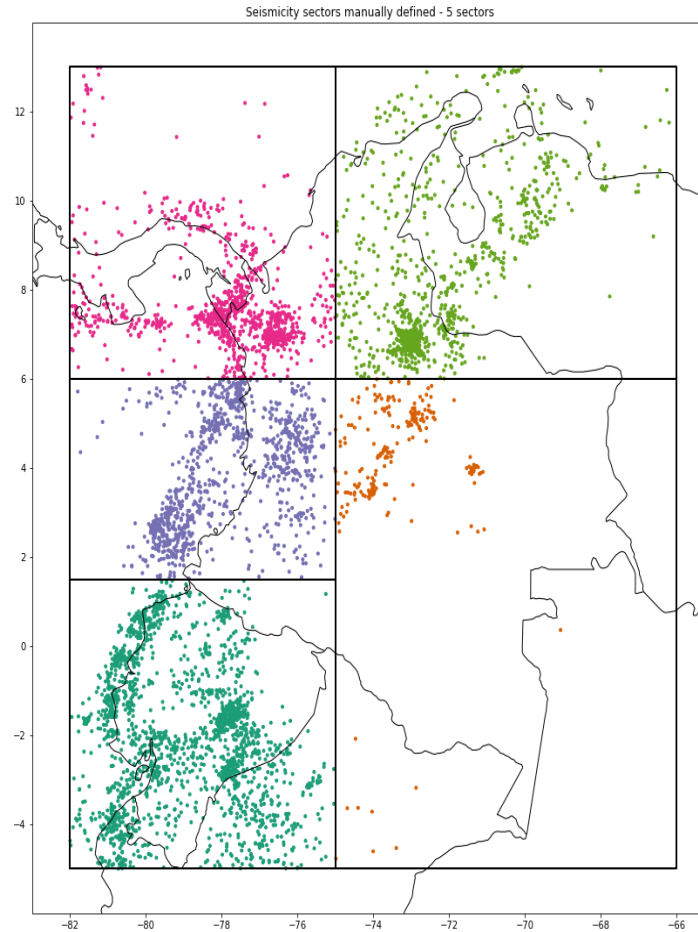


Figure 35: Definition of 5 seismicity sectors selected manually. In the map, the color of the dots indicates the assigned type according to the sector.

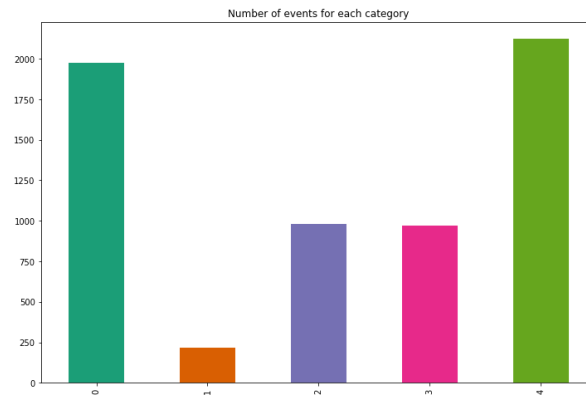


Figure 36: Definition of 5 seismicity sectors selected manually. The bar plot shows the number of events for each sector.

According to Figure 33 and Figure 34, a four-sector classification gives that sector 0 (near Ecuador), sector 2 (Pacific Coast) and sector 3 (Bucaramanga and northeast)

have a similar number of earthquakes. However, the extension of sector 2 is large, which is why a five-sector classification could also be relevant. In this case, Figure 35 and Figure 36 show that sector 0 (near Ecuador) and sector 4 (Bucaramanga and northeast) have a similar number of earthquakes; but in sectors 2 and 3, which corresponds to the seismicity in the Pacific Coast, the number of events have half the number of events of sectors 0 and 4. This might be relevant when training a model with these labels.

4.1.4. Input Features and outputs

According to the analysis of the Earthquake networks presented in Chapter 3, and the catalog in section 4.1, the following remarks about the features of seismicity and earthquake networks can be said:

- The earthquake networks should be created using a cell size that allows a good representation of seismicity. According to Figure 10 to Figure 18 and conclusions from Chapter 3, a size of 50km is chosen.
- Networks can be constructed considering the depth of the events (named as *3d networks*) or discarding the depth (named *2d networks*).
- Figure 27, Figure 28, Figure 29, Figure 30, Figure 31 and Figure 32 show the seismicity and time elapsed for set windows of 10, 50 and 100 events, moving the windows half the size. The 10 event window results in a rapid changing seismicity and a maximum time lapse near 90 days, the 50-events window with a maximum time lapse near 300 days, and the 100-events window with a maximum time lapse between 500 and 600 days.
- Most of the released seismic energy comes from a few events with magnitude $M > 7.0$.
- A model can be constructed using feature vectors from each network constructed to train ML algorithms such as Random Forest or Support Vector Machine; or using the networks themselves to train Graph Machine Learning models such as Graph Neural Networks.

Following this, Table 9 presents the algorithms and features selected to create the models, and Table 10 the output targets for the models.

Table 9: ML algorithms and input features used to create the models

ML Algorithm	Source	Feature
Random Forest Support Vector Machine	Earthquake Networks $\Delta=50\text{km}$ N. events: [10, 50, 100]	Nodes
		Edges
		Mean Degree
		Clustering Coefficient
		Average Path Length
		Entropy
		Small-worldness
		Betweenness Centrality
	Catalog	Energy
		Days
		Seismicity
Graph Neural Networks	Earthquake Networks $\Delta=50\text{km}$ N. events: [10, 50, 100]	

Table 10: ML problems and Target outputs.

M.L. Problem	Target	Description
Regression	Time of next event (days)	Number of days for the next event
	Magnitude	Magnitude of next event
	Seismicity 10 events	Logarithm of seismicity for the next 10 events
	Seismicity 20 events	Logarithm of seismicity for the next 20 events
Classification	Geographic sector (4 sectors)	Id of sector where next event will occur
	Geographic sector (5 sectors)	
	2 Magnitude bins	Level of magnitude. Low: $M < 5$ High: $5 < M$
	3 Magnitude bins	Level of magnitude. Low: $4 < M < 5$ Medium: $5 < M < 6$ High: $6 < M$

4.1.5. Models and learning algorithms

There are two major types of algorithms used in this work: traditional ML algorithms (Support Vector Machine and Random Forest) and Deep Learning models (Deep Graph Convolutional Neural Networks).

To create traditional ML models, feature vectors are created for each network. Table 9 shows the features selected to create the vectors, and Table 10 the targets. Figure 37 depicts the process to train a ML model from a catalog: create multiple catalogs grouping sets of N events; then construct networks for each network and calculate the target variables. Then, joining the feature vectors and the target variables, a model is trained with 80% of the vectors and tested with the remaining 20%.

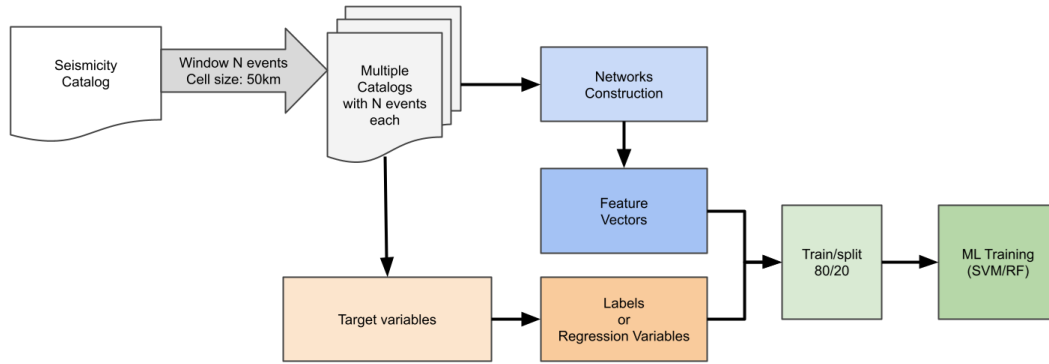


Figure 37: ML models creation pipeline

As shown on Figure 22, Figure 34 and Figure 36, the target values are imbalanced; that is, there are classes or categories with much more members than others. This has led to the use of classification models suited for unbalanced data sets. In this thesis two methods have been used to train a model suited for imbalanced data: oversampling on categories with few members, and association of weights to each class. The programs and the models have been trained using python and **scikit-learn** library. According to the documentation of this library², a Random Forest classifier with configured in the *balanced* mode adjusts the weights for a class C automatically in an inversely proportional way to the frequency F of this class following the formula $W_c = S / (N * S_c)$ where S is the total number of observations, N the number of classes, and S_c is the number of observations for the class C . The RF models were trained using

²<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

100 trees and the balanced class weight. Also, to apply oversampling on the data, the SMOTE method from the **imbalanced-learn** library in python is used³ to oversample and balance the data on the training vectors; the test vectors are kept the same to validate the model.

As for the SVM models, they are not recommended for imbalanced data sets, but the use of a high cost in the implementation could help to penalize bad classifications and train better models. The SVM models were trained using a Radial Basis Function kernel, $\gamma=0.7$ and $C=25$.

About deep learning models, Deep Graph Convolutional Neural Networks (DGCNN) was used in this thesis from the StellarGraph library. According to the documentation⁴, DGCNN was proposed by Zhang et al. [83], using the Graph Convolutional Layers from Kipf and Welling [84]. A change in the pipeline for the DGCNN models is made, as feature vectors for the nodes are needed but graph feature vectors not. The feature vectors for the nodes have information about the geographic location of the node: Latitude, Longitude and Depth. This architecture has three sequential stages: Graph convolution layers extract nodes local substructure features and define a consistent sorting, then a Sort Pooling layer sorts the nodes feature vectors following the previous defined sorting and unifies input sizes, and finally traditional convolutional and dense layers read the sorted graph representations and perform the classifications. Figure 38 shows the pipeline for the implementation of DGCNN models for classifications.

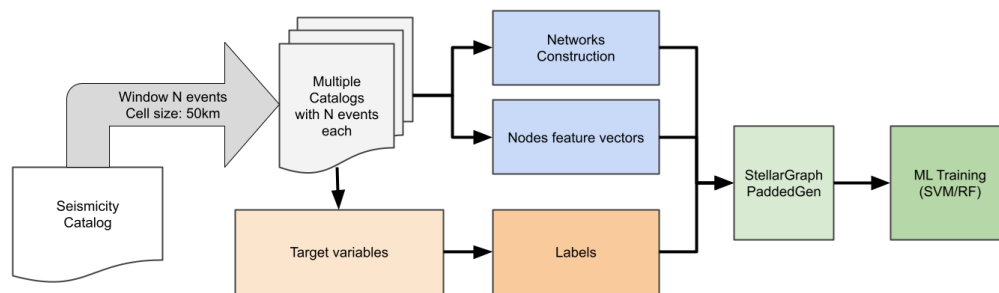


Figure 38: Pipeline for the implementation of DGCNN classification models

³ https://imbalanced-learn.org/stable/over_sampling.html

⁴ <https://stellargraph.readthedocs.io/en/stable/demos/graph-classification/dgcnn-graph-classification.html>

To create the DGCNN models, different hyperparameters must be configured, as well settings to create the EN. The architecture of the DGCNN model follows the schema shown on Figure 39.

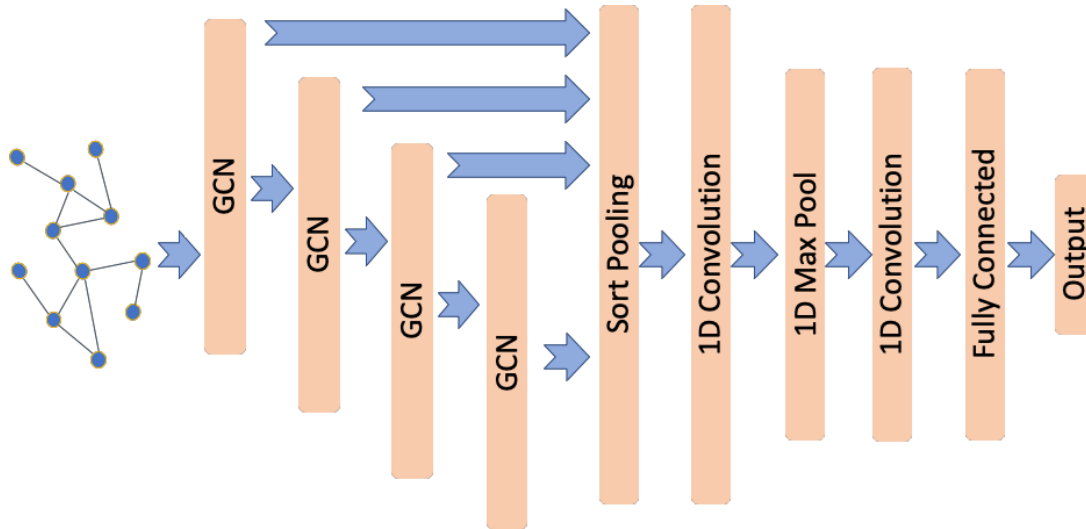


Figure 39: DGCNN architecture. From <https://stellargraph.readthedocs.io/en/stable/demos/graph-classification/dgcnn-graph-classification.html>

Table 11 shows the different parameters, a description of the parameters, and the different values used to compile the DGCNN models.

Table 11: Parameters used to create DGCNN models for EN.

Name	Description	Range of values
N. Events	Number of events used to create the EN.	[10, 50,100]
EN Type	Type of ECN created.	[2D, 3D]
Train Batch size	Number of training examples that will be propagated through the network for the training	50
Test Batch size	Number of test samples that will be tested for each network training	1
N. Rows	Number of rows for the output tensor	30
Layer Dimensions	Size of the convolution layers. A vector with the size for each layer.	[30, 30, 30, 4]
Filters 1	Number of filters used on the first convolutional layer	16
Pool Size	Size of the pooling in the MaxPool layer.	2
Filters 2	Number of filters on the second convolutional layer	[25, 32]
Kernel size	Size of the filters used on the first convolutional layer	[5, 8]
Units	Number of neurons on the fully dense layer before the output.	128

Activation 1	Activation function for the dense layer.	Relu
Activation 2	Activation function for the output layer	Softmax
Learning rate	Setting to control the weights adjusting respect the loss function	[0.0001, 0.0005, 0.0007, 0.001 , 0.002]

Some remarks: The output layer is fully connected also, and the number of outputs corresponds with the number of classes. Also, the type of loss function used is sparse categorical cross-entropy, as the classifications are made for multi-class instances.

4.1.6. Ontology definition

An ontology definition for target outputs, that is, a semantic model, should define the source of the data, boundaries and time periods used to analyze the data, the name and type of targets, conditions for labeling a specific vector, and the type of ML paradigm to implement (classification or regression).

Some axioms for an ontology can be stated as the following:

- A *Dataset* definition, stating the source of the *Catalog*, and the spatial and time boundaries.
- A *Catalog*, which is a compilation of *Earthquakes*.
- An *Earthquake* occurred in a place and in a specific instant, and has a *Hypocenter* and a *Magnitude*.
- A *Source Time* has a date and a time.
- *Spatial Coordinates* have latitude and longitude in degrees, and depth in kilometers.
- *Magnitude* is expressed in degrees of moment magnitude. *Energy* for an earthquake can be calculated from the magnitude using the Gutenberg-Richter magnitude-energy relation from subsection 4.1.2.
- A *Hypocenter* is the description of the time and place of the occurrence of an *Earthquake*, and has a *Source Time* and *Spatial Coordinates*.
- An *Events Window* has an excerpt of N events from the *Catalog*, and have characteristics such as *Days* which is the difference of the *Source Time* for the first and the last event, *Energy* which is the sum of energies for the *Earthquakes*, and *Seismicity* which is calculated for the *Earthquakes* in the window using formula 5.
- A *Cell Size* is expressed in kilometers.

- An *Earthquake Network* is constructed from a *Dataset*, using a specific *Cell Size* and *Events Window*.
- *Network Measurements* are calculated from an *Earthquake Network*, and have the features from table 10: Nodes, Edges, Main Degree, Clustering Coefficient, Average Path Length, Entropy, Small-worldness, and Betweenness Centrality.
- A *Feature Vector* is calculated for an *Earthquake Network*, and has *Network Measurements* and *Events Window* characteristics.
- A *Target Output*, which can be either of a *Regression* type or a *Classification* type. *Target Outputs* are used together with *Feature Vectors* to train ML models.
- *Time of Next Event* is a *Target Output* of the *Regression* type, which counts, for a *Feature Vector*, the time in days for the next *Earthquake*.
- *Magnitude of Next Event* is a *Target Output* of the *Regression* type, which says for a *Feature Vector* what is the magnitude of the next *Earthquake*.
- *Seismicity of the Next Events* is a *Target Output* of the *Regression* type, which counts for a *Feature Vector* what the seismicity is for the next N events. Seismicity is calculated for the N events following formula 5.
- *Geographic Sectorization* is the splitting of the region defined by the spatial boundaries of the *Dataset* in different *Sectors*. Each *Sector* has *boundaries* and a *label* to identify it.
- *Sector of Next Event* is a *Target Output* of the *Classification* type. This target output corresponds to the *label* that identifies the *Sector* from the *Geographic Sectorization* definition where, for a *Feature Vector*, the next *Earthquake* occurs.
- *Magnitude Binning* is the splitting of the *Magnitude* range into N discrete intervals or *Magnitude Bin*, each one identified with a specific *label*.
- *Magnitude Bin of Next Event* is a *Target Output* of the *Classification* type. This target output corresponds to the *label* that identifies the *Magnitude Bin* from the *Magnitude Binning* definition that, for a *Feature Vector*, the next *Earthquake*'s magnitude will fall into.

This ontology could be implemented using knowledge representation languages such as OWL, JSON-LD, or RDF. The actual implementation of an ontology following the previous definitions is left for future work.

4.2. Classification Models

As mentioned before, the catalog consists of 6262 events. Earthquake networks were created as mentioned on section 4.1.2. This resulted in 6251 networks when the set window is 10 events, 6211 networks when the window is 50 events, and 6161 networks when the window is 100 events. According to Table 10, classification models are trained for geographic sectorization and magnitude level forecasting. Network features were used to create vectors, and the target labels corresponded to Geographic Sectorization with 4 and 5 sectors, following the cases shown on Figure 33 and Figure 35 respectively; and Magnitude Binning with 2 and 3 bins, where a binary classification was made for 2 bins defining a low magnitude event with $M < 5.0$ and a high magnitude with $M > 5.0$, and a 3 classes classification with class 0 being $M < 5.0$, class 1 where $5.0 < M < 6.0$ and class 2 is $M > 6.0$.

As for the metrics used to qualify the classification models, the *Precision*, *Recall* and *F1 score* are used. Table 12 shows the metrics, the descriptions of the metrics and the formula to calculate them.

Table 12: Metrics used to qualify classifications.

Metric	Description	Formula	Condition
Precision	Proportion of good predictions among all the predictions made for a specific class: what is actually well predicted among the predictions.	$P = TP / (TP + FP)$, where TP is True Positives and FP is False Positives.	$P > 0.8$ good.
Recall	Also called sensitivity, measures the proportion of good predictions among the actual instances: how good the model can predict a specific class.	$R = TP / (TP + FN)$, where FN is False Negatives.	$R > 0.8$ good.
F1 Score	balances the Precision and Recall metrics: seeks a balance between high precision and high recall.	$F1 = 2 * P * R / (P + R)$, where P is precision and R Recall.	$F1 > 0.8$ good

The metrics can be calculated for each class and as a weighted average for all classes. The weighted average takes a metric for each class and multiplies it by the number of samples of each class, all are summed and then divided by the total number of samples. This weighted average is a better indication for the model than other types of averages as it considers the case of imbalanced data.

Metrics for the 4 class Geographic sectorization are shown in Table 13, for the 5 class Geographic sectorization in Table 14, Table 15 for the 2 class magnitude binning and

Table 16 for the 3 class magnitude binning. In the DGCNN case, the metrics corresponds with the parameters which resulted on the highest accuracy.

Table 13: Classification metrics for the 4 class Geographic sectorization models.

ML Algorithm	Net type	N. Events	Target	Precision	Recall	F1 score
Random Forest (class weights balanced)	2D	10	4 GeoSector	0.36	0.36	0.36
		50	4 GeoSector	0.40	0.41	0.40
		100	4 GeoSector	0.41	0.41	0.40
	3D	10	4 GeoSector	0.38	0.38	0.38
		50	4 GeoSector	0.39	0.40	0.39
		100	4 GeoSector	0.42	0.42	0.42
Random Forest (Oversampling SMOTE)	2D	10	4 GeoSector	0.36	0.36	0.36
		50	4 GeoSector	0.41	0.40	0.40
		100	4 GeoSector	0.40	0.39	0.39
	3D	10	4 GeoSector	0.38	0.37	0.38
		50	4 GeoSector	0.39	0.38	0.39
		100	4 GeoSector	0.40	0.39	0.40
SVM	2D	10	4 GeoSector	0.34	0.34	0.34
		50	4 GeoSector	0.36	0.37	0.36
		100	4 GeoSector	0.37	0.37	0.36
	3D	10	4 GeoSector	0.33	0.33	0.33
		50	4 GeoSector	0.36	0.36	0.36
		100	4 GeoSector	0.37	0.37	0.37
DGCNN	2D	10	4 GeoSector	0.47	0.41	0.36
		50	4 GeoSector	0.44	0.42	0.39
		100	4 GeoSector	0.43	0.42	0.41
	3D	10	4 GeoSector	0.42	0.41	0.38
		50	4 GeoSector	0.48	0.43	0.39
		100	4 GeoSector	0.42	0.41	0.38

Table 14: Classification metrics for the 5 class Geographic sectorization models.

ML Algorithm	Net type	N. Events	Target	Precision	Recall	F1 score
Random Forest (class weights balanced)	2D	10	5 GeoSector	0.33	0.34	0.33
		50	5 GeoSector	0.40	0.40	0.40
		100	5 GeoSector	0.40	0.40	0.39
	3D	10	5 GeoSector	0.32	0.33	0.32
		50	5 GeoSector	0.40	0.41	0.40
		100	5 GeoSector	0.39	0.39	0.39
Random Forest (Oversampling SMOTE)	2D	10	5 GeoSector	0.33	0.35	0.34
		50	5 GeoSector	0.37	0.37	0.37
		100	5 GeoSector	0.38	0.38	0.38
	3D	10	5 GeoSector	0.34	0.34	0.34
		50	5 GeoSector	0.35	0.35	0.5
		100	5 GeoSector	0.38	0.37	0.38
SVM	2D	10	5 GeoSector	0.30	0.31	0.30
		50	5 GeoSector	0.33	0.35	0.32
		100	5 GeoSector	0.35	0.36	0.33
	3D	10	5 GeoSector	0.30	0.30	0.30
		50	5 GeoSector	0.32	0.34	0.32
		100	5 GeoSector	0.34	0.35	0.33
DGCNN	2D	10	5 GeoSector	0.48	0.43	0.40
		50	5 GeoSector	0.55	0.42	0.36
		100	5 GeoSector	0.49	0.40	0.34
	3D	50	5 GeoSector	0.53	0.42	0.36

Table 15: Classification metrics for the 2 class Magnitude binning models.

ML Algorithm	Net type	N. Events	Target	Precision	Recall	F1 score
Random Forest (class weights balanced)	2D	10	2 MagBin	0.75	0.83	0.78
		50	2 MagBin	0.75	0.83	0.78
		100	2 MagBin	0.74	0.82	0.77
	3D	10	2 MagBin	0.76	0.83	0.78
		50	2 MagBin	0.75	0.83	0.78
		100	2 MagBin	0.73	0.82	0.77
Random Forest (Oversampling SMOTE)	2D	10	2 MagBin	0.75	0.78	0.76
		50	2 MagBin	0.77	0.74	0.75
		100	2 MagBin	0.77	0.74	0.75
	3D	10	2 MagBin	0.75	0.78	0.76
		50	2 MagBin	0.77	0.75	0.76
		100	2 MagBin	0.75	0.73	0.74
SVM	2D	10	2 MagBin	0.70	0.75	0.73
		50	2 MagBin	0.77	0.82	0.79
		100	2 MagBin	0.71	0.79	0.74
	3D	10	2 MagBin	0.71	0.77	0.73
		50	2 MagBin	0.76	0.82	0.78
		100	2 MagBin	0.72	0.79	0.75
DGCNN	2D	10	2 MagBin	0.71	0.84	0.77
		50	2 MagBin	0.71	0.84	0.77
		100	2 MagBin	0.71	0.84	0.77
	3D	50	2 MagBin	0.87	0.84	0.77

Table 16: Classification metrics for the 3 class Magnitude binning models.

ML Algorithm	Net type	N. Events	Target	Precision	Recall	F1 score
Random Forest (class weights balanced)	2D	10	3 MagBin	0.73	0.83	0.77
		50	3 MagBin	0.74	0.83	0.78
		100	3 MagBin	0.74	0.83	0.77
	3D	10	3 MagBin	0.73	0.83	0.77
		50	3 MagBin	0.75	0.83	0.78
		100	3 MagBin	0.74	0.83	0.77
Random Forest (Oversampling SMOTE)	2D	10	3 MagBin	0.74	0.78	0.76
		50	3 MagBin	0.76	0.73	0.74
		100	3 MagBin	0.75	0.73	0.74
	3D	10	3 MagBin	0.74	0.77	0.75
		50	3 MagBin	0.76	0.74	0.75
		100	3 MagBin	0.75	0.72	0.74
SVM	2D	10	3 MagBin	0.70	0.76	0.73
		50	3 MagBin	0.75	0.82	0.78
		100	3 MagBin	0.71	0.79	0.75
	3D	10	3 MagBin	0.71	0.77	0.73
		50	3 MagBin	0.75	0.81	0.77
		100	3 MagBin	0.72	0.79	0.75
DGCNN	2D	10	3 MagBin	0.85	0.84	0.77
		50	3 MagBin	0.85	0.85	0.77
		100	3 MagBin	0.71	0.84	0.77
	3D	50	3 MagBin	0.71	0.84	0.77

According to Table 13 and Table 14, performance was bad on all metrics for geographic sectorization. This suggests that the network approach is not fit to forecast the location of earthquakes. Using oversampling didn't improve the model significantly compared with the balanced weight method.

Table 15 and Table 16 show better results for magnitude bins forecasting. Oversampling improved the precision compared with balanced class weights, compromising recall and f1 score. This is because the models improved the

classifications for the classes with few members but not in a way that these categories had good metrics, but the majority class decreased their metrics.

Highest precisions were achieved using GCNN models on 50 events windows: for 2 class magnitude binning using 3D networks, and for 3 class magnitude binning using 2D networks. Figure 40 presents the confusion matrix for GCNN model on 2 class magnitude binning and Figure 41 for 3 class magnitude binning.

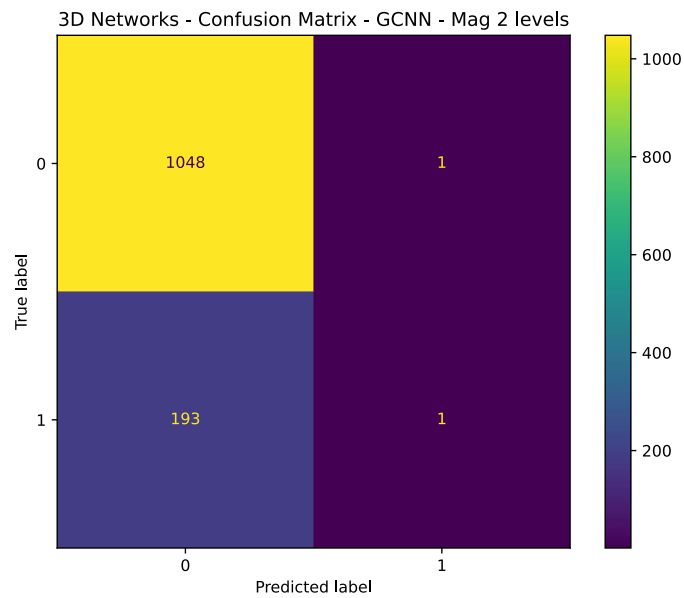


Figure 40: Confusion matrix for GCNN model on 2 class magnitude binning and 3D networks

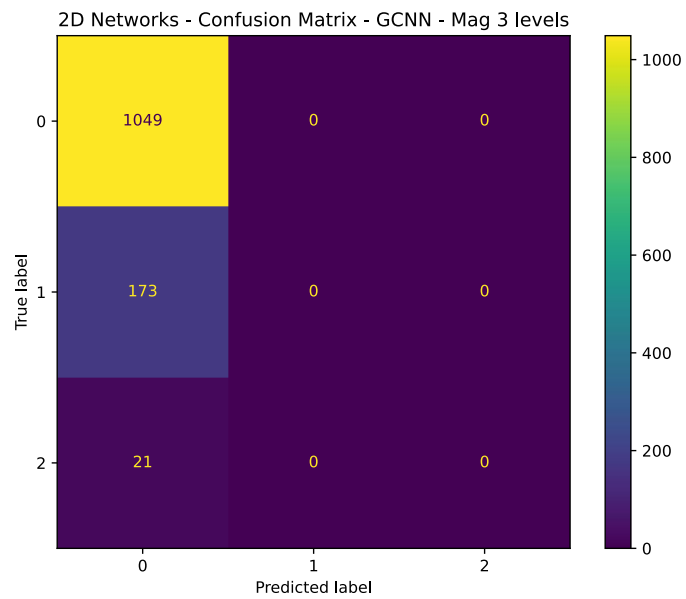


Figure 41: Confusion matrix for GCNN model on 3 class magnitude binning and 2D networks

Figure 40 and Figure 41 show that the models can't predict correctly if an event will have high magnitude. In general, the classifications using any of these models fall short on being of high quality.

4.3. Regression Models

In the case of regressions, two exercises were made: one where the feature vector had both network and catalog characteristics, and the other with only network characteristics (see Table 9). Two algorithms were used: Random Forest Regression and Support Vector Regression (SVR); and were trained using the scikit-learn library from python. However, no results for SVR are shown, as the performance was worst in every possible scenario. Table 10 shows the targets for regression. The metrics employed to measure the quality of the regression models are shown on Table 17.

Table 17: Description of regression metrics. Information taken from the sklearn documentation web page: https://scikit-learn.org/stable/modules/model_evaluation.html

Metric	Description	Formula	Condition
MAE (Mean Absolute Error)	Corresponding to the expected value of the absolute error loss.	$MAE = \frac{1}{n} \sum_i y_i - \hat{y}_i $	<i>Good:</i> $0 < MAE \leq \infty$: <i>Bad</i>
MAPE (Mean Absolute Percentage Error)	The idea of this metric is to be sensitive to relative errors. It is for example not changed by a global scaling of the target variable. As some values could be 0, a very small arbitrary value is selected to avoid indetermination.	$MAPE = \frac{1}{n} \sum_i \frac{ y_i - \hat{y}_i }{\max(\epsilon, y_i)}, \epsilon \approx 0$	<i>Good:</i> $0 < MAPE \leq \infty$: <i>Bad</i>
MSE (Mean Square Error)	Corresponding to the expected value of the squared (quadratic) error.	$MSE = \frac{1}{n} \sum_i (y_i - \hat{y}_i)^2$	<i>Good:</i> $0 < MSE \leq \infty$: <i>Bad</i>
R2 (Coefficient of Determination)	Proportion of variance (of y) that has been explained by the independent variables in the model.	$R^2(y, \hat{y}) = 1 - \frac{\sum_i ((y_i - \hat{y}_i)^2)}{\sum_i ((y_i - \bar{y})^2)}$	<i>Bad:</i> $-\infty < R^2 \leq 1$: <i>Good</i>

Results of regressions are presented for 2D and 3D networks constructed with sets of 50 events, using 1000 trees for the random forest algorithm, creating input vectors using network features, and comparing when catalog features are also used. An extended view of the results of the regressions can be found in the appendices.

4.3.1 Days for the next earthquake regression

The metrics of the regression of the number of days for the next earthquake are shown on Table 18. This table shows that the R2 metric for these experiments is below 0.1; with some being negative, which means that the model is not suited to forecast. The catalog parameters improved a little the models, but even on those cases, the models were far from good.

Table 18: Metrics for the Random Forest Regression of days for the next event.

Net. Type	N. Evs.	N. trees	Features	MAE	MSE	MAPE	R2
2d	50	1000	Net	2.587	12.993	35.487	-0.019
3d	50	1000	Net	2.601	13.126	37.081	-0.030
2d	50	1000	Net/catalog	2.491	12.352	64.603	0.031
3d	50	1000	Net/catalog	2.449	12.041	56.554	0.056

To have a better look on these experiment, Figure 42, Figure 43, Figure 44 and Figure 45 show the comparison between days for the next event real (X axis) vs. predicted (Y axis) using 2d and 3d networks, and using only network features and network plus catalog features. Important to have in mind that a perfect regression would be seen in this plots as a straight line.

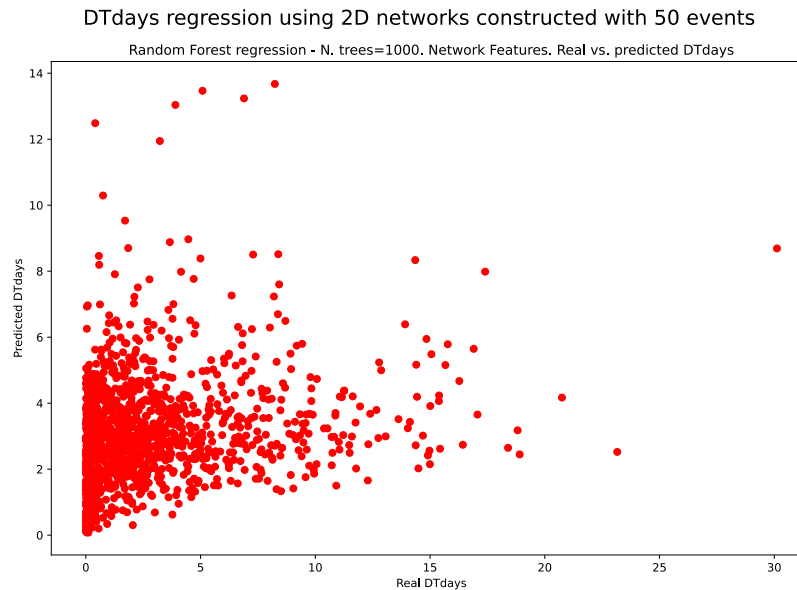


Figure 42: Random Forest Regression of days for the next event using only network features from 2D networks with 50 events windows.

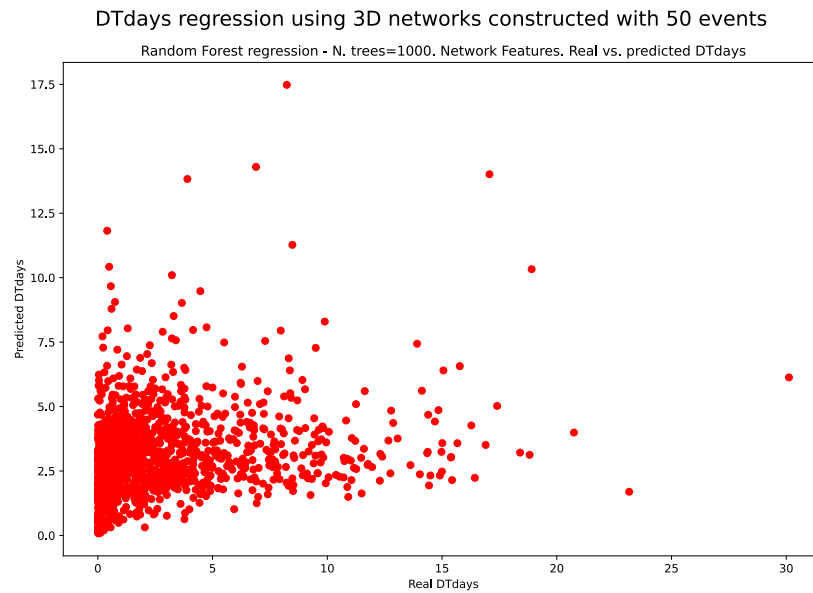


Figure 43: Random Forest Regression of days for the next event using only network features from 3D networks with 50 events windows.

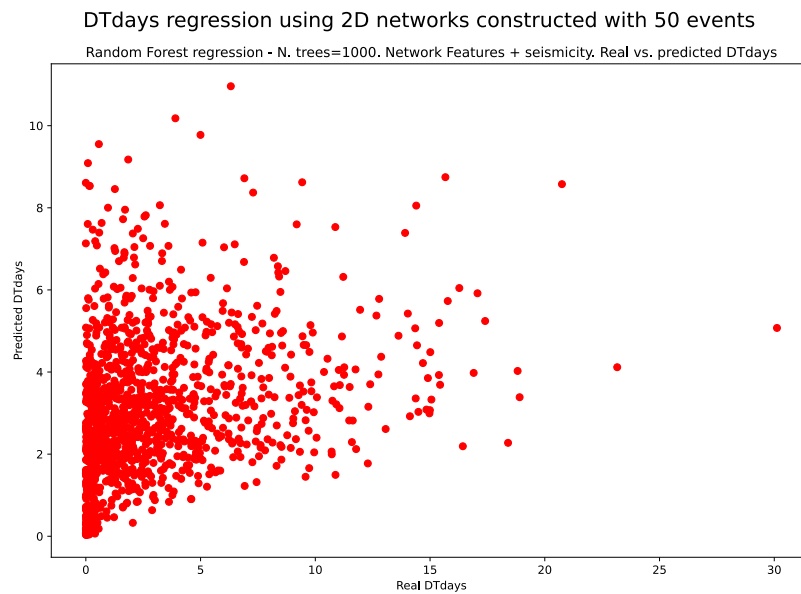


Figure 44: Random Forest Regression of days for the next event using network and catalog features from 2D networks with 50 events windows.

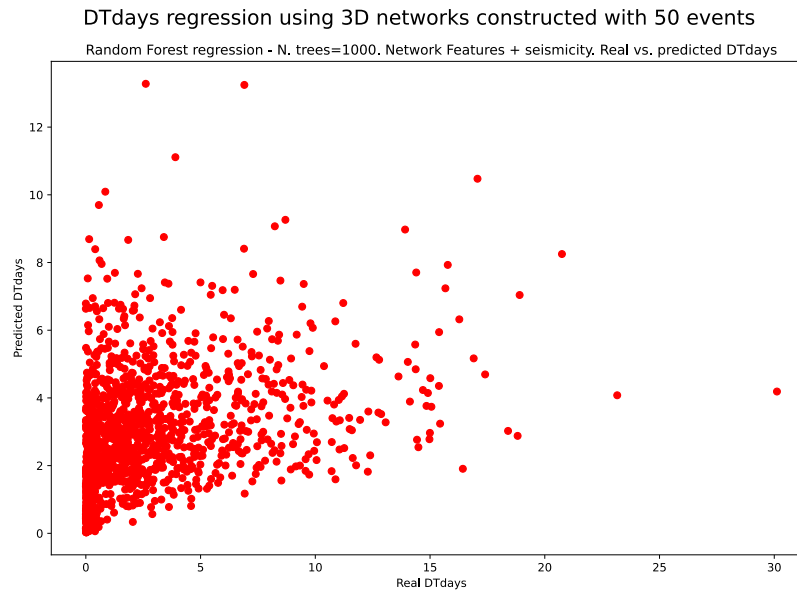


Figure 45: Random Forest Regression of days for the next event using network and catalog features from 3D networks with 50 events windows.

4.3.2 Magnitude for the next event regression

The metrics for the regression of magnitude for next event are shown on Table 19. This table shows that the magnitude regression has a $R^2 > 0$ for all the experiments, which also accounts as a poor model to forecast magnitude.

Table 19: Metrics for the Random Forest Regression of magnitude for the next event.

Net. Type	N. evs.	N. trees	Features	MAE	MSE	MAPE	R2
2d	50	1000	Net	0.318	0.180	0.068	-0.123
3d	50	1000	Net	0.317	0.177	0.068	-0.104
2d	50	1000	Net/catalog	0.306	0.169	0.065	-0.055
3d	50	1000	Net/catalog	0.302	0.163	0.065	-0.021

As with the days for the next earthquake, Figure 46, Figure 47, Figure 48 and Figure 49 show the comparison between real magnitude vs predicted magnitude for the next event using 2d and 3d networks, and using only network features and network plus catalog features. In these figures it is clear that the models are not appropriate for the desired task.

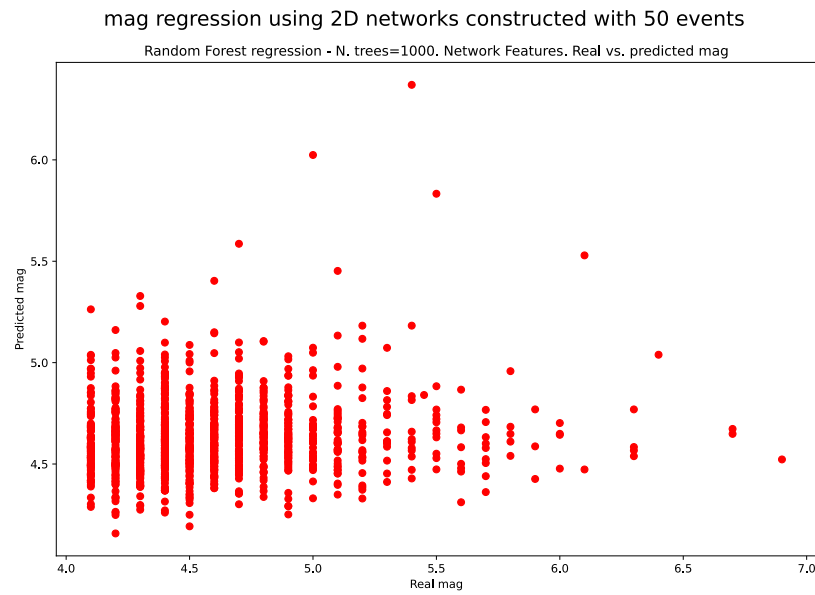


Figure 46: Random Forest Regression of magnitude for the next event using only network features from 2D networks with 50 events windows.

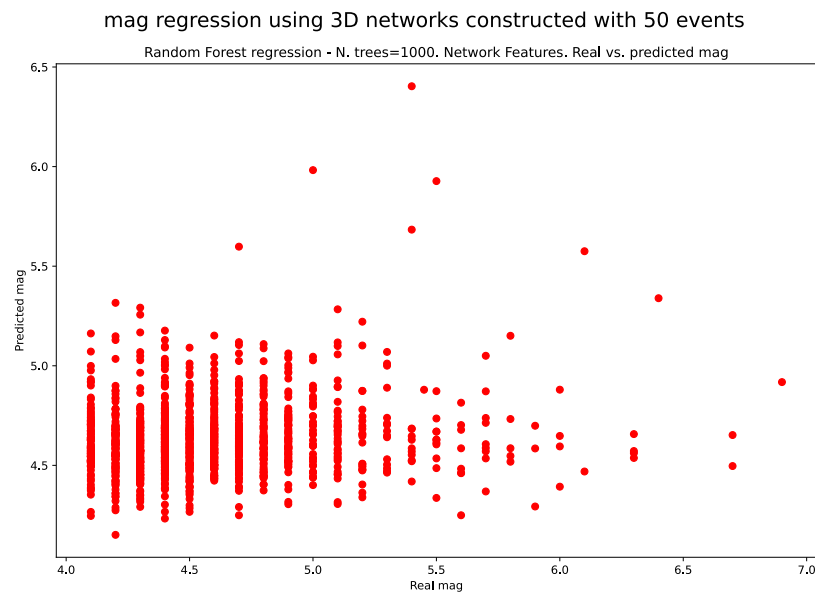


Figure 47: Random Forest Regression of magnitude for the next event using only network features from 3D networks with 50 events windows.

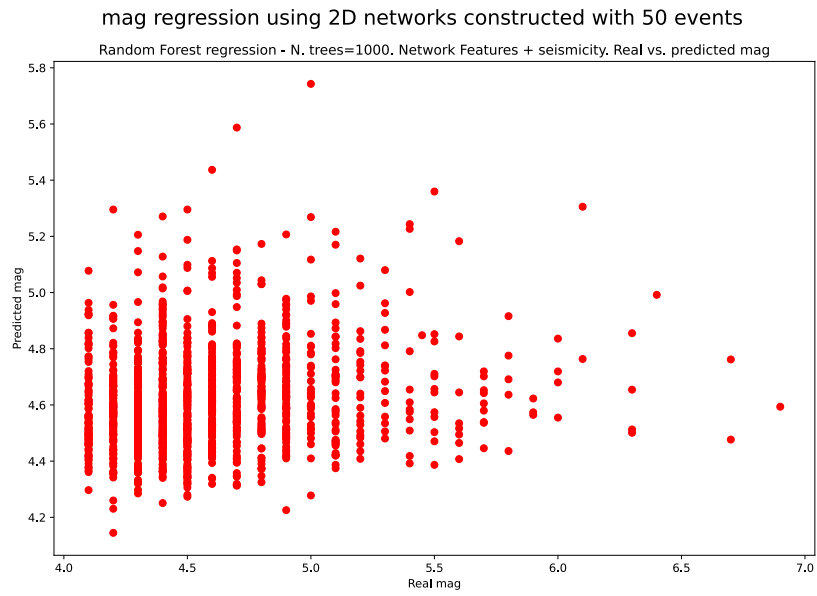


Figure 48: Random Forest Regression of magnitude for the next event using network and catalog features from 2D networks with 50 events windows.

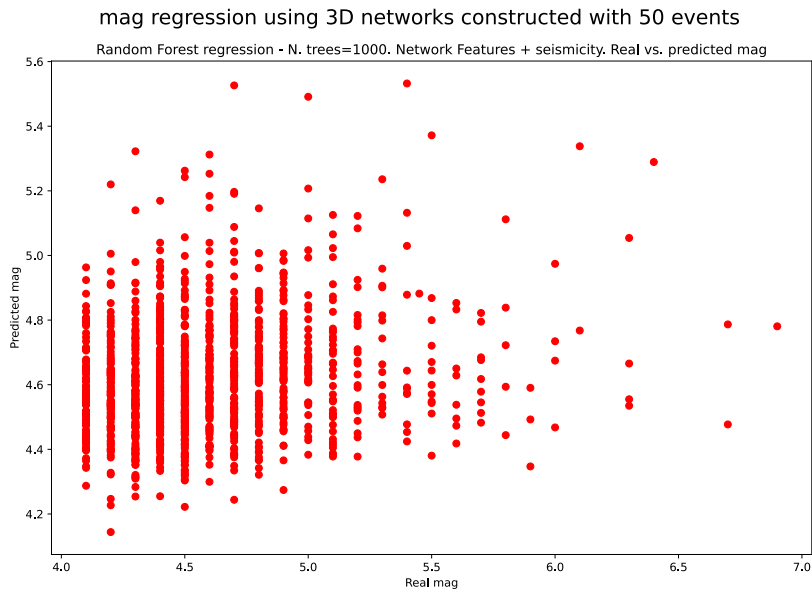


Figure 49: Random Forest Regression of magnitude for the next event using network and catalog features from 3D networks with 50 events windows.

4.3.3 Logarithm of seismicity for the next events regression

Models to forecast the seismicity of the next 10 and 20 events were made. The metrics for the regression of logarithm of the seismicity for the next 10 events are shown on Table 20. This table shows a $R^2 > 0.5$ for the experiments, and a better performance on the network and catalog features case near 0.8.

Table 20: Metrics for the Random Forest Regression of seismicity released by the next 10 events.

Net. Type	N. evs.	N. trees	Features	MAE	MSE	MAPE	R2
2d	50	1000	Net	0.949	1.867	0.057	0.523
3d	50	1000	Net	0.936	1.831	0.056	0.533
2d	50	1000	Net/catalog	0.604	0.858	0.037	0.781
3d	50	1000	Net/catalog	0.607	0.870	0.037	0.778

Seismicity regression for the next 10 events using 2D networks constructed with 50 events

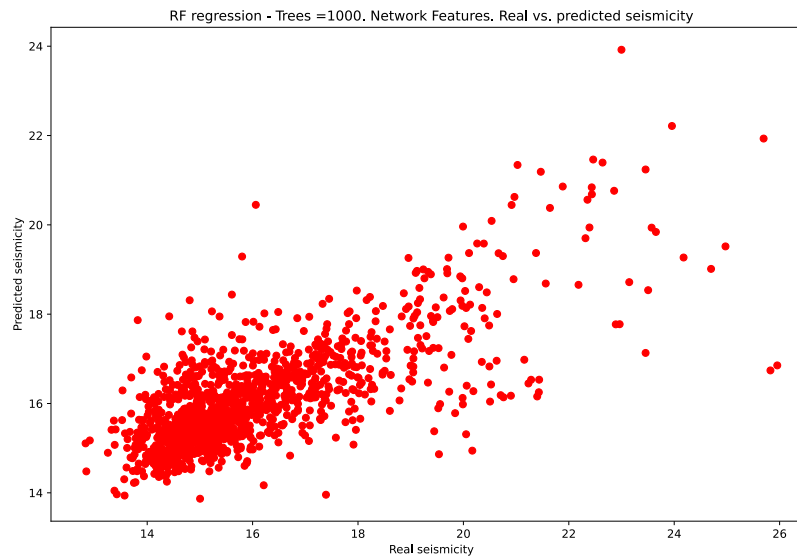


Figure 50: Random Forest Regression of seismicity of the next 10 events using network features and 2D networks with 50 events windows.

Seismicity regression for the next 10 events using 3D networks constructed with 50 events

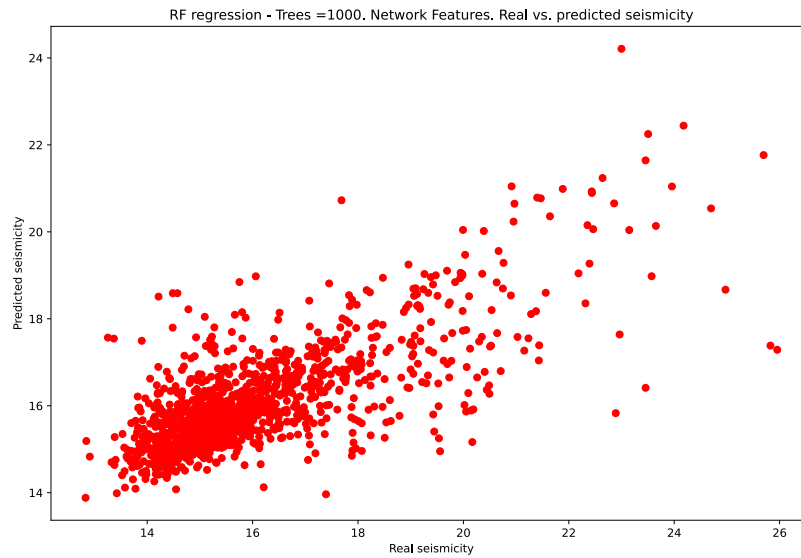


Figure 51: Random Forest Regression of seismicity of the next 10 events using network features and 3D networks with 50 events windows.

Seismicity regression for the next 10 events using 2D networks constructed with 50 events

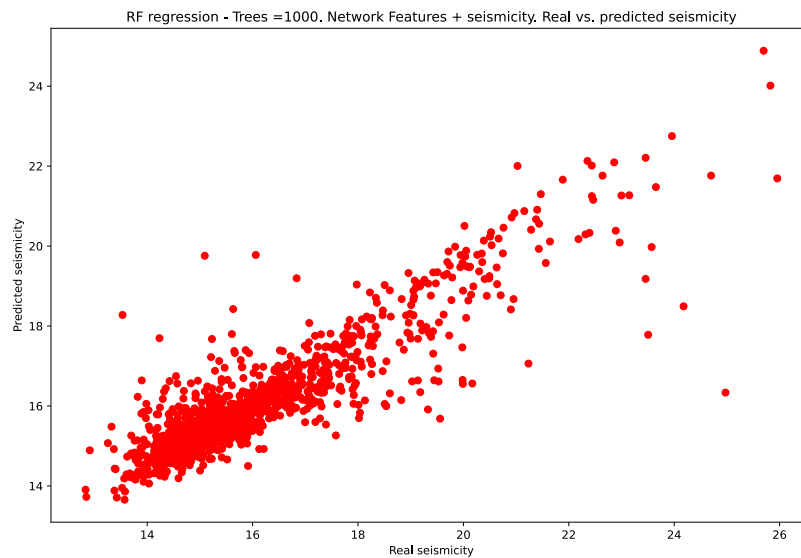


Figure 52: Random Forest Regression of seismicity of the next 10 events using network and catalog features and 2D networks with 50 events windows.

Seismicity regression for the next 10 events using 3D networks constructed with 50 events

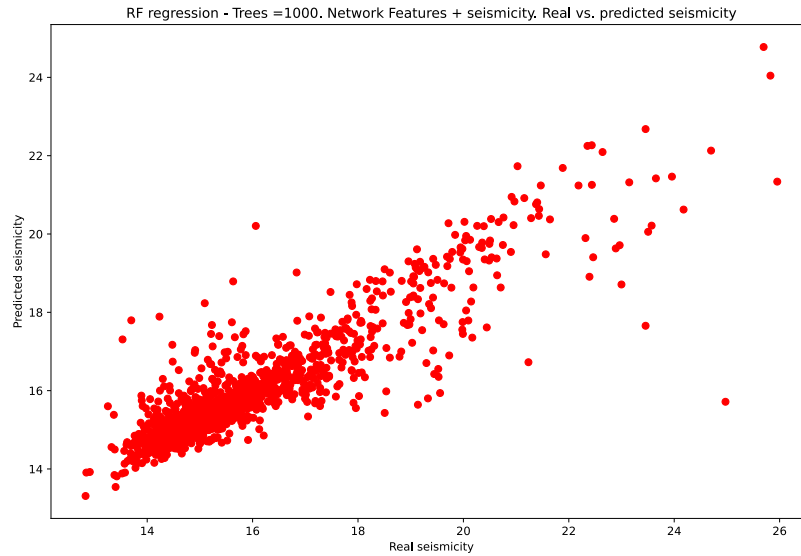


Figure 53: Random Forest Regression of seismicity of the next 10 events using network and catalog features and 3D networks with 50 events windows.

Figure 50, Figure 51, Figure 52 and Figure 53 show the comparison between real vs predicted seismicity for the next 10 events using 2D and 3D networks, and using only network features and network plus catalog features. These figures agree with the observations shown on Table 20, that using both network and catalog features can improve the results.

The metrics for the regression of logarithm of the seismicity for the next 20 events are shown on Table 21. This table shows also a $R^2 > 0.5$ for the experiments, and a better performance on the network and catalog features case.

Table 21: Metrics for the Random Forest Regression of seismicity released by the next 20 events.

Net. Type	N. evs.	N. trees	Full/mod	MAE	MSE	MAPĚ	R2
2d	50	1000	Net	0.884	1.712	0.052	0.546
3d	50	1000	Net	0.891	1.689	0.053	0.552
2d	50	1000	Net/catalog	0.475	0.631	0.028	0.833
3d	50	1000	Net/catalog	0.493	0.720	0.029	0.810

Table 21 show that the confidence levels are better for the next 20 events compared to the next 10 events case when looking at the R2 metric, and that the best values are when using the catalog information.

Figure 55, Figure 56, Figure 54 and Figure 57 show the comparison between real vs predicted seismicity for the next 20 events using 2D and 3D networks, and using only network features and network plus catalog features. These figures also show that using catalog features improve the quality of the regression.

Seismicity regression for the next 20 events using 2D networks constructed with 50 events

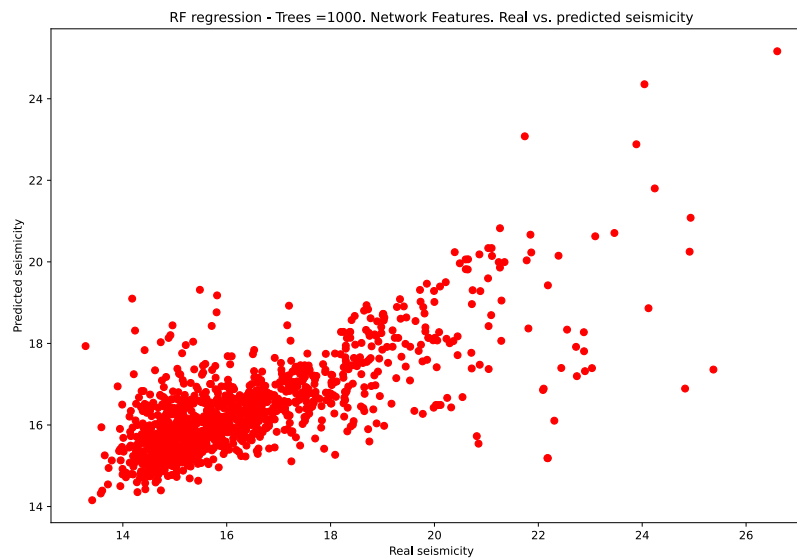


Figure 54: Random Forest Regression of seismicity of the next 20 events using network and catalog features and 2D networks with 50 events windows.

Seismicity regression for the next 20 events using 3D networks constructed with 50 events

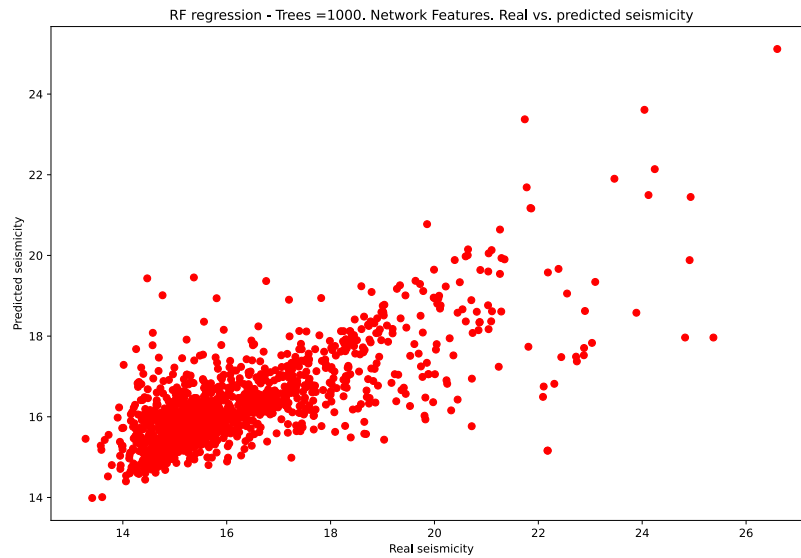


Figure 55: Random Forest Regression of seismicity of the next 20 events using network features and 2D networks with 50 events windows.

Seismicity regression for the next 20 events using 2D networks constructed with 50 events



Figure 56: Random Forest Regression of seismicity of the next 20 events using network features and 2D networks with 50 events windows.

Seismicity regression for the next 20 events using 3D networks constructed with 50 events



Figure 57: Random Forest Regression of seismicity of the next 20 events using network and catalog features and 3D networks with 50 events windows.

4.4. Chapter highlights

In this chapter different ML models have been trained using data from Earthquake Networks, to predict characteristics for next events. In general, classification models don't have good results, as the data is highly unbalanced. ML algorithms were trained taking this into account.

The time lapse between events is, in average, a little more than 4 days. Windows of 50 events seem to be the better to create models, and there are two reasons to say this: The time lapse this number of events is not that long (maximum near a year, and on average a network spans half a year), compared to networks with 100 events (maximum near 500 days). The second reason is that results are similar and sometimes better than using windows of 100 events. On the other side, 10-events window seem to be short to forecast using EN.

Classification models had poor results, no matter which model was used. Forecasting characteristics for a next earthquake is a hard task and the use of these techniques to create feature vectors for a ML algorithm is not enough.

The regression models using RF had low errors for seismicity target outputs, but high for time and magnitude for the next event. The R^2 coefficient was very low when trying to forecast the magnitude and the day for the next event, which fits in the conclusion about forecasting characteristics for a next earthquake. However, when forecasting seismicity for the next set of events, the model were promising, and improved a lot when network features were complemented with seismicity characteristics, resulting in R^2 coefficients near 0.8 which is an indicator of a good model in seismology.

Chapter 5

Conclusions and Discussion

Forecasting of seismic events is a very difficult task, due to the impossibility to actually see what is in the interior of the earth. However, a pipeline of the process of creation of the seismicity model supported by ECN networks and ML can be summarized by Figure 58.

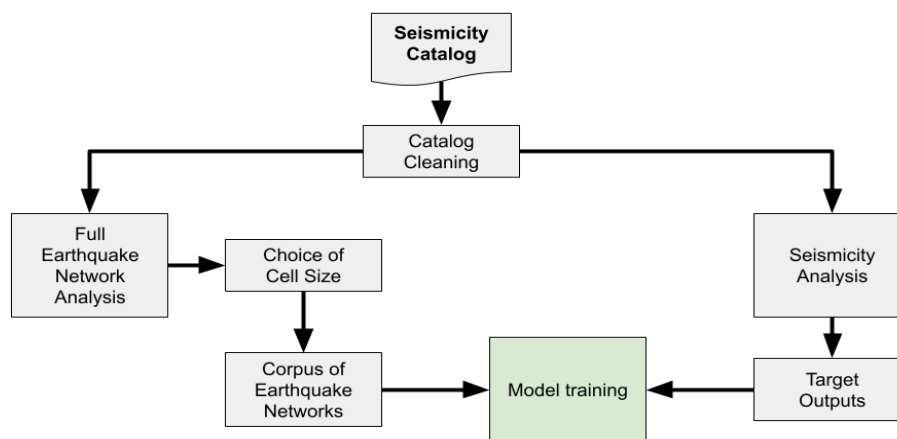


Figure 58: Pipeline for the creation of the seismicity model supported by Earthquake Networks and ML

This thesis, as far as it is known, is the first use of earthquake networks and ML models to forecast characteristics of the seismicity. Following the ontology described in Chapter 4, regression and classification models were trained. In general, the regression models showed good metrics, and the classification models poor metrics.

Earthquake Networks are a seismicity representation model that captures the relationship between events. In the literature, characteristics of these networks such as mean degree, or global clustering, have been analyzed and in some cases, they change before a big event occurs, which suggest their use to forecasting models. A good choice of a cell size for the Earthquake Networks can be achieved by plotting the entropy of the degree probability distribution versus cell size and comparing it with the entropy of similar Erdős-Renyi networks, and selecting when the ECN plot crosses the ER plot.

The analysis of seismicity must be able to evaluate how the targets change over time. In the case of this experiment, seismicity for a set of events, days between events, magnitude and location targets were evaluated. The classification models were not good enough to forecast the sub-region where a subsequent event could occur, neither if the next event can be classified with a label associated with the magnitude. However, the regression experiments were successful in forecasting how many days until the next event, or what would be the seismicity for the next 10 or 20 events. This information is of great importance, as it makes it possible to say if the energy released by the next events will be high or low and help the respective agencies to prepare for such scenarios.

The fact that a simple ML model such as Regression Random Forest can have a good representation of the changes in time and seismicity for the next events suggests that the feature engineering was adequate.

In seismology, there are some parameters related to the complexity of the phenomenon. In the Gutenberg register distribution, the b parameter can be seen as the relationship between magnitude of events and their frequency. This parameter tends to be $b \sim 1$ in all seismic regions (Stein et al, p. 274 [1]). Deviation from this value sometimes occurs, and sometimes they are associated with changes in the seismic regime. The same might happen with parameters from the ECN in a region: there are some cell sizes to construct ECN where the entropy is similar to random ER networks, and when a network has an entropy that varies from a similar ER Network then some variation in the seismicity might be announcing.

Traditional ML models such as RF and SVM, and DGCNN have a similar response to classify the target outputs. As the literature has shown, predicting aspects of the seismicity in a region is a hard task, and training the classification models have shown poor results. However, when creating regression models, the forecasting of the time for the next event, the magnitude of next event, and the seismicity for the next 10 and 20 events have shown good metrics.

Although the DGCNN didn't show good results to classify the target outputs, the parameters used to train these models such as the kernel size of the first convolutional layer might have an impact in the quality of the models. These models were a little better than those dealing with the imbalance, but still were not good.

The possibility to integrate the proposed ontology to real-time operations in a seismic Network to forecast aspects of the seismicity exists. In this scenario, the models can

be updated with new observations using traditional ML models, because these models needed a short time to be trained; unlike the DGCNN, where the training was long, and given that both types of models resulted in similar metrics, a simpler model would be preferred. The next phase of this research might be using convolutional networks with graphs, in this case EN, to make regressions for the variables proposed in this thesis: the time and magnitude for the next event, and the energy release by the next events.

The use of depth in the creation of the earthquake networks, that is, to construct 3D or 2D networks, doesn't have a significant impact on the results, except for the DGCNN model created to classify the 2 class Magnitude Binning with a high precision for high class despite a very low recall. However, a 3D ECN seems to be a better fit for the creation of these models. Also, the use of catalog information; that is, the energy, the time span for each window and the seismicity; helps the models to improve.

This thesis fulfilled in a way all the objectives that were set at the beginning. However, some of them were partially fulfilled, such as the semi-automatic annotation scheme. The ontology defined aimed to complete this objective, but the actual implementation using some language such as JSON-LD or OWL was not actually made, because at the completion of this thesis the resulting model is more conceptual and is not fully implemented as a software with services to ingest data from observatories or so. A future step of this research could be aimed towards a fully working implementation, defining a database schema to store results and configurations, web services to ingest data when a recent earthquake occurs and create new Earthquake Networks and update the model, and maybe generating messages to selected people following the estimated states from the working model.

The best models were for the seismicity for the next set of earthquakes, using both network and catalog features. Although the best metrics was reached forecasting the seismicity for the next 20 events, it is preferred to forecast a shorter time span, as it is closer to trigger an immediate response if necessary. A summary of the preferred parameters to train a ML algorithm to model the seismicity in the region surrounding Colombia might be the following:

- Cell sizes of 50 km.
- Event windows of 50 events.
- 2D Earthquake Network.
- Random Forest with 1000 trees.
- Network and catalog features.

Bibliography

- [1] S. Stein and M. Wyss, *An introduction to seismology, earthquakes, and earth structure*. Malden, MA: Blackwell Pub, 2003.
- [2] “Why can’t we predict earthquakes?,” *Nat. Comput. Sci.*, vol. 3, no. 2, pp. 115–115, Feb. 2023, doi: 10.1038/s43588-023-00418-1.
- [3] B. Gutenberg and C. F. Richter, “Seismicity of the Earth,” in *Geological Society of America Special Papers*, vol. 34, Geological Society of America, 1941, pp. 1–126. doi: 10.1130/SPE34-p1.
- [4] D. A. León, J. A. Valdivia, and V. A. Bucheli, “Modeling of Colombian Seismicity as Small-World Networks,” *Seismol. Res. Lett.*, vol. 89, no. 5, pp. 1807–1816, Sep. 2018, doi: 10.1785/0220180076.
- [5] B. Gutenberg and C. Richter, “Earthquake magnitude, intensity, energy, and acceleration,” *Bull. Seismol. Soc. Am.*, vol. 32, no. 3, pp. 163–191, 1942.
- [6] B. Enescu, J. Mori, M. Miyazawa, and Y. Kano, “Omori-Utsu Law c-Values Associated with Recent Moderate Earthquakes in Japan,” *Bull. Seismol. Soc. Am.*, vol. 99, no. 2A, pp. 884–891, Apr. 2009, doi: 10.1785/0120080211.
- [7] N. Boccara, *Modeling complex systems*, 2. ed., Softcover reprint of the hardcover 2. ed. 2010. in Graduate texts in physics. New York: Springer, 2010.
- [8] S. Abe and N. Suzuki, “Scale-free network of earthquakes,” *Europhys. Lett. EPL*, vol. 65, no. 4, pp. 581–586, Feb. 2004, doi: 10.1209/epl/i2003-10108-1.
- [9] S. Abe and N. Suzuki, “Small-world structure of earthquake network,” *Phys. Stat. Mech. Its Appl.*, vol. 337, no. 1–2, pp. 357–362, Jun. 2004, doi: 10.1016/j.physa.2004.01.059.
- [10] P. Erdos and A. Renyi, “On the evolution of random graphs,” *Publ Math Inst Hung. Acad Sci*, vol. 5, pp. 17–61, 1960.
- [11] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998, doi: 10.1038/30918.
- [12] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, Jan. 2002, doi: 10.1103/RevModPhys.74.47.

- [13] A. Donnellan *et al.*, "Potential for a large earthquake near Los Angeles inferred from the 2014 La Habra earthquake: POTENTIAL FOR A LARGE EARTHQUAKE NEAR LA," *Earth Space Sci.*, vol. 2, no. 9, pp. 378–385, Sep. 2015, doi: 10.1002/2015EA000113.
- [14] D. León-Vargas, V.-A. Bucheli-Guerrero, and H.-A. Ordoñez-Erazo, "Modelado de la sismicidad colombiana como una red compleja espacio-secuencial," *Rev. Científica*, vol. 48, no. 3, pp. 67–81, Sep. 2023, doi: 10.14483/23448350.20963.
- [15] D. Leon, J. Valdivia, and V. Bucheli, "A revision of seismicity models based on complex systems and earthquake networks," *J. Seismol.*, vol. 26, no. 1, pp. 137–145, Feb. 2022, doi: 10.1007/s10950-021-10017-0.
- [16] H. Ordoñez Erazo, D. A. Leon Vargas, and V. A. Bucheli Guerrero, "Tagq TagEQ-CN: Clasificación de Eventos Sísmicos Basado en Redes Complejas," *Investig. E Innov. En Ing.*, vol. 9, no. 1, pp. 136–148, May 2021, doi: 10.17081/invinno.9.1.4856.
- [17] L. E. Ordoñez-Palacios, D. A. León-Vargas, V. A. Bucheli-Guerrero, and H. A. Ordoñez-Eraso, "Predicción de radiación solar en sistemas fotovoltaicos utilizando técnicas de aprendizaje automático," *Rev. Fac. Ing.*, vol. 29, no. 54, p. e11751, Sep. 2020, doi: 10.19053/01211129.v29.n54.2020.11751.
- [18] B. Gutenberg and C. F. Richter, "Frequency of earthquakes in California," *Bull. Seismol. Soc. Am.*, vol. 34, no. 4, pp. 185–188, 1944.
- [19] Y. Bar-Yam, *Dynamics of complex systems*. in Studies in nonlinearity. Reading, Mass: Addison-Wesley, 1997.
- [20] Z. Olami, H. J. S. Feder, and K. Christensen, "Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes," *Phys. Rev. Lett.*, vol. 68, no. 8, pp. 1244–1247, Feb. 1992, doi: 10.1103/PhysRevLett.68.1244.
- [21] P. Bak and C. Tang, "Earthquakes as a self-organized critical phenomenon," *J. Geophys. Res. Solid Earth*, vol. 94, no. B11, pp. 15635–15637, Nov. 1989, doi: 10.1029/JB094iB11p15635.
- [22] M. Baiesi and M. Paczuski, "Scale-free networks of earthquakes and aftershocks," *Phys. Rev. E*, vol. 69, no. 6, p. 066106, 2004.
- [23] X. He, H. Zhao, W. Cai, Z. Liu, and S.-Z. Si, "Earthquake networks based on space–time influence domain," *Phys. Stat. Mech. Its Appl.*, vol. 407, pp. 175–184, Aug. 2014, doi: 10.1016/j.physa.2014.03.093.

- [24] S. Rezaei, H. Moghaddasi, and A. H. Darooneh, "Preferential attachment in evolutionary earthquake networks," *Phys. Stat. Mech. Its Appl.*, vol. 495, pp. 172–179, Apr. 2018, doi: 10.1016/j.physa.2017.12.063.
- [25] L. Telesca and M. Lovallo, "Analysis of seismic sequences by using the method of visibility graph," *EPL Europhys. Lett.*, vol. 97, no. 5, p. 50002, Mar. 2012, doi: 10.1209/0295-5075/97/50002.
- [26] K. Ito, "Towards a new view of earthquake phenomena," *Pure Appl. Geophys. PAGEOPH*, vol. 138, no. 4, pp. 531–548, 1992, doi: 10.1007/BF00876337.
- [27] C. Tsallis, *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*. New York, NY: Springer New York, 2009.
- [28] F. Vallianatos, G. Michas, and G. Papadakis, "Nonextensive Statistical Seismology," in *Complexity of Seismic Time Series*, Elsevier, 2018, pp. 25–59. doi: 10.1016/B978-0-12-813138-1.00002-X.
- [29] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, "Complex networks: Structure and dynamics," *Phys. Rep.*, vol. 424, no. 4–5, pp. 175–308, Feb. 2006, doi: 10.1016/j.physrep.2005.10.009.
- [30] A. Barabási, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999, doi: 10.1126/science.286.5439.509.
- [31] D. Chorozoglou, A. Iliopoulos, C. Kourouklas, O. Mangira, and E. Papadimitriou, "Earthquake Networks as a Tool for Seismicity Investigation: a Review," *Pure Appl. Geophys.*, Jun. 2019, doi: 10.1007/s00024-019-02253-w.
- [32] M. D. Humphries and K. Gurney, "Network 'small-world-ness': a quantitative method for determining canonical network equivalence," *PloS One*, vol. 3, no. 4, p. e0002051, 2008, doi: 10.1371/journal.pone.0002051.
- [33] S. Duan, T. Wen, and W. Jiang, "A new information dimension of complex network based on Rényi entropy," *Phys. Stat. Mech. Its Appl.*, vol. 516, pp. 529–542, Feb. 2019, doi: 10.1016/j.physa.2018.10.045.
- [34] M. Wiedermann, J. F. Donges, J. Kurths, and R. V. Donner, "Mapping and discrimination of networks in the complexity-entropy plane," *Phys. Rev. E*, vol. 96, no. 4, Oct. 2017, doi: 10.1103/PhysRevE.96.042304.

- [35] J. Leskovec and R. Sosič, "SNAP: A General-Purpose Network Analysis and Graph-Mining Library," *ACM Trans. Intell. Syst. Technol. TIST*, vol. 8, no. 1, p. 1, 2016.
- [36] M. Bastian, S. Heymann, M. Jacomy, and others, "Gephi: an open source software for exploring and manipulating networks.," *Icwsn*, vol. 8, no. 2009, pp. 361–362, 2009.
- [37] E. R. Gansner and S. C. North, "An open graph visualization system and its applications to software engineering," *Softw. - Pract. Exp.*, vol. 30, no. 11, pp. 1203–1233, 2000.
- [38] Wolfram Research Inc., *Mathematica, Version 11.0*. 2017. [Online]. Available: <https://www.wolfram.com/mathematica>
- [39] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring Network Structure, Dynamics, and Function using NetworkX," in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, and J. Millman, Eds., Pasadena, CA USA, 2008, pp. 11–15.
- [40] C. Data61, "StellarGraph Machine Learning Library," *GitHub Repository*. GitHub, 2018. [Online]. Available: <https://github.com/stellargraph/stellargraph>
- [41] Z.-K. Gao, M. Small, and J. Kurths, "Complex network analysis of time series," *EPL Europhys. Lett.*, vol. 116, no. 5, p. 50001, Dec. 2016, doi: 10.1209/0295-5075/116/50001.
- [42] D. Pastén, Z. Czechowski, and B. Toledo, "Time series analysis in earthquake complex networks," *Chaos Interdiscip. J. Nonlinear Sci.*, vol. 28, no. 8, p. 083128, Aug. 2018, doi: 10.1063/1.5023923.
- [43] S. Abe, D. Pastén, and N. Suzuki, "Finite data-size scaling of clustering in earthquake networks," *Phys. Stat. Mech. Its Appl.*, vol. 390, no. 7, pp. 1343–1349, Apr. 2011, doi: 10.1016/j.physa.2010.11.033.
- [44] S. Abe, D. Pastén, V. Muñoz, and N. Suzuki, "Universalities of earthquake-network characteristics," *Chin. Sci. Bull.*, vol. 56, no. 34, pp. 3697–3701, Dec. 2011, doi: 10.1007/s11434-011-4767-6.
- [45] L. Lacasa, A. Nuñez, É. Roldán, J. M. R. Parrondo, and B. Luque, "Time series irreversibility: a visibility graph approach," *Eur. Phys. J. B*, vol. 85, no. 6, Jun. 2012, doi: 10.1140/epjb/e2012-20809-8.

- [46] M. Baiesi and M. Paczuski, "Complex networks of earthquakes and aftershocks," *Nonlinear Process. Geophys.*, vol. 12, pp. 1–11, 2005.
- [47] M. Suteanu, "Aspects of Structure in Earthquake Networks," *Pure Appl. Geophys.*, vol. 172, no. 7, pp. 1865–1878, Jul. 2015, doi: 10.1007/s00024-014-0939-z.
- [48] Y. Xu, T. Ren, Y. Liu, and Z. Li, "Earthquake prediction based on community division," *Phys. Stat. Mech. Its Appl.*, vol. 506, pp. 969–974, Sep. 2018, doi: 10.1016/j.physa.2018.05.035.
- [49] D. Chorozoglou, D. Kugiumtzis, and E. Papadimitriou, "Application of complex network theory to the recent foreshock sequences of Methoni (2008) and Kefalonia (2014) in Greece," *Acta Geophys.*, vol. 65, no. 3, pp. 543–553, Jun. 2017, doi: 10.1007/s11600-017-0039-4.
- [50] S. Abe and N. Suzuki, "Complex-network description of seismicity," *Nonlinear Process. Geophys.*, vol. 13, no. 2, pp. 145–150, May 2006, doi: 10.5194/npg-13-145-2006.
- [51] K. Deyasi, A. Chakraborty, and A. Banerjee, "Network similarity and statistical analysis of earthquake seismic data," *Phys. Stat. Mech. Its Appl.*, vol. 481, pp. 224–234, Sep. 2017, doi: 10.1016/j.physa.2017.04.050.
- [52] Y. Ohsawa, "Regional Seismic Information Entropy to Detect Earthquake Activation Precursors," *Entropy*, vol. 20, no. 11, p. 861, Nov. 2018, doi: 10.3390/e20110861.
- [53] N. Lotfi and A. H. Darooneh, "The earthquakes network: the role of cell size," *Eur. Phys. J. B*, vol. 85, no. 1, Jan. 2012, doi: 10.1140/epjb/e2011-20623-x.
- [54] S. Abe and N. Suzuki, "Determination of the scale of coarse graining in earthquake networks," *EPL Europhys. Lett.*, vol. 87, no. 4, p. 48008, Aug. 2009, doi: 10.1209/0295-5075/87/48008.
- [55] C. Pavez-Orrego and D. Pastén, "Defining the Scale to Build Complex Networks with a 40-Year Norwegian Intraplate Seismicity Dataset," *Entropy*, vol. 25, no. 9, p. 1284, Aug. 2023, doi: 10.3390/e25091284.
- [56] D. Pastén, F. Torres, B. Toledo, V. Muñoz, J. Rogan, and J. A. Valdivia, "Time-Based Network Analysis Before and After the Mw 8.3 Illapel Earthquake 2015

Chile,” *Pure Appl. Geophys.*, vol. 173, no. 7, pp. 2267–2275, Jul. 2016, doi: 10.1007/s00024-016-1335-7.

- [57] R. M. Yulmetyev, N. A. Emelyanova, and F. M. Gafarov, “Dynamical Shannon entropy and information Tsallis entropy in complex systems,” *Phys. Stat. Mech. Its Appl.*, vol. 341, pp. 649–676, Oct. 2004, doi: 10.1016/j.physa.2004.03.094.
- [58] P. Jiao and A. H. Alavi, “Artificial intelligence in seismology: Advent, performance and future trends,” *Geosci. Front.*, vol. 11, no. 3, pp. 739–744, May 2020, doi: 10.1016/j.gsf.2019.10.004.
- [59] S. M. Mousavi and G. C. Beroza, “Machine Learning in Earthquake Seismology,” *Annu. Rev. Earth Planet. Sci.*, vol. 51, no. 1, pp. 105–129, May 2023, doi: 10.1146/annurev-earth-071822-100323.
- [60] H. Kubo, M. Naoi, and M. Kano, “Recent advances in earthquake seismology using machine learning,” *Earth Planets Space*, vol. 76, no. 1, p. 36, Feb. 2024, doi: 10.1186/s40623-024-01982-0.
- [61] K. J. Bergen, P. A. Johnson, M. V. De Hoop, and G. C. Beroza, “Machine learning for data-driven discovery in solid Earth geoscience,” *Science*, vol. 363, no. 6433, p. eaau0323, Mar. 2019, doi: 10.1126/science.aau0323.
- [62] R. J. Geller, D. D. Jackson, Y. Y. Kagan, and F. Mulargia, “Earthquakes Cannot Be Predicted,” *Science*, vol. 275, no. 5306, pp. 1616–1616, Mar. 1997, doi: 10.1126/science.275.5306.1616.
- [63] Md. H. A. Banna *et al.*, “Application of Artificial Intelligence in Predicting Earthquakes: State-of-the-Art and Future Challenges,” *IEEE Access*, vol. 8, pp. 192880–192923, 2020, doi: 10.1109/ACCESS.2020.3029859.
- [64] G. S. Manral and A. Chaudhary, “Prediction of Earthquake Using Machine Learning Algorithms,” in *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom: IEEE, May 2023, pp. 1–5. doi: 10.1109/ICIEM59379.2023.10166658.
- [65] G. Gürsoy, A. Varol, and A. Nasab, “Importance of Machine Learning and Deep Learning Algorithms in Earthquake Prediction: A Review,” in *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, Chattanooga, TN, USA: IEEE, May 2023, pp. 1–6. doi: 10.1109/ISDFS58141.2023.10131766.
- [66] N. Varshney, G. Kumar, A. Kumar, S. K. Pandey, T. Singh, and K. U. Singh, “Machine Learning Based Algorithm for Earthquake Monitoring,” in *2023 IEEE 12th International Conference on Communication Systems and Network*

- Technologies (CSNT)*, Bhopal, India: IEEE, Apr. 2023, pp. 264–270. doi: 10.1109/CSNT57126.2023.10134640.
- [67] “EARTHQUAKE PREDICTION USING MACHINE LEARNING,” *Int. Res. J. Mod. Eng. Technol. Sci.*, May 2023, doi: 10.56726/IRJMETS39049.
- [68] C. Stamile, A. Marzullo, and E. Deusebio, *Graph machine learning: take graph data to the next level by applying machine learning techniques and algorithms*. Birmingham: Packt Publishing, 2021.
- [69] I. Chami, S. Abu-El-Haija, B. Perozzi, C. Ré, and K. Murphy, “Machine Learning on Graphs: A Model and Comprehensive Taxonomy,” 2020, *arXiv*. doi: 10.48550/ARXIV.2005.03675.
- [70] J. Zhou *et al.*, “Graph neural networks: A review of methods and applications,” *AI Open*, vol. 1, pp. 57–81, 2020, doi: 10.1016/j.aiopen.2021.01.001.
- [71] Y. Chang, L. M. Warren, and G. A. Prieto, “Precise Locations for Intermediate-Depth Earthquakes in the Cauca Cluster, Colombia,” *Bull. Seismol. Soc. Am.*, vol. 107, no. 6, pp. 2649–2663, Dec. 2017, doi: 10.1785/0120170127.
- [72] C. Frohlich, K. Kadinsky-Cade, and S. D. Davis, “A reexamination of the Bucaramanga, Colombia, earthquake nest,” *Bull. Seismol. Soc. Am.*, vol. 85, no. 6, pp. 1622–1634, 1995.
- [73] W. E. G. Moreno, “Analysis of Colombian Seismicity as a Way to Explain and Understand The Bucaramanga Nest,” *J. Geosci. Eng. Environ. Technol.*, vol. 3, no. 3, p. 134, Aug. 2018, doi: 10.24273/jgeet.2018.3.3.1709.
- [74] J. F. Schneider, W. D. Pennington, and R. P. Meyer, “Microseismicity and focal mechanisms of the intermediate-depth Bucaramanga Nest, Colombia,” *J. Geophys. Res.*, vol. 92, no. B13, p. 13913, 1987, doi: 10.1029/JB092iB13p13913.
- [75] J. B. Rundle, A. Giguere, D. L. Turcotte, J. P. Crutchfield, and A. Donnellan, “Global Seismic Nowcasting with Shannon Information Entropy,” *Earth Space Sci.*, p. 2018EA000464, Dec. 2018, doi: 10.1029/2018EA000464.
- [76] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.
- [77] J. W. Tukey, *Exploratory data analysis*. in Addison-Wesley series in behavioral science. Reading, Mass: Addison-Wesley Pub. Co, 1977.

- [78] J. R. Peterson and C. R. Hutt, "World-wide standardized seismograph network: a data users guide," US Geological Survey, 2014.
- [79] R. Butler *et al.*, "The global seismographic network surpasses its design goal," *Eos Trans. Am. Geophys. Union*, vol. 85, no. 23, pp. 225–229, Jun. 2004, doi: 10.1029/2004EO230001.
- [80] C. F. Richter, "An instrumental earthquake magnitude scale," *Bull. Seismol. Soc. Am.*, vol. 25, no. 1, pp. 1–32, 1935.
- [81] H. Kanamori, "The energy release in great earthquakes," *J. Geophys. Res.*, vol. 82, no. 20, pp. 2981–2987, Jul. 1977, doi: 10.1029/JB082i020p02981.
- [82] B. Gutenberg, "The energy of earthquakes," *Q. J. Geol. Soc. Lond.*, vol. 112, no. 1–4, pp. 1–14, Aug. 1956, doi: 10.1144/GSL.JGS.1956.112.01-04.02.
- [83] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, in AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018.
- [84] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations (ICLR)*, 2017.

Appendices

APPENDIX 1: Measures taken from the earthquake networks (EN) constructed with different cell sizes (dx).

APPENDIX 2: Regression metrics extended.

APPENDIX 3: Publications made during the doctorate studies.

- Leon, D., Valdivia, J. & Bucheli, V. A revision of seismicity models based on complex systems and earthquake networks. *J Seismol* 26, 137–145 (2022).
- D. Leon-Vargas., V. Bucheli-Guerrero., H. Ordoñez-Erazo “TagEQ-CN: Clasificación de Eventos Sísmicos Basado en Redes Complejas”, *Investigación e Innovación en Ingenierías*, vol. 9, n°1, 136-148, 2021. DOI: <https://doi.org/10.17081/invinno.9.1.4856>
- León-Vargas, D., Bucheli-Guerrero, V. A., y Ordoñez-Erazo, H. A., (2023). Modelado de la sismicidad colombiana como una red compleja espacio-secuencial. *Revista Científica*, 48(3), 67-81
- Leon, Daniel & Ordoñez, Hugo & Bucheli, Victor. (2020). TagEQ: Sistema de Clasificación de Eventos Sísmicos Basado en Inteligencia Artificial. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*. e27. 359 - 370.
- Ordoñez-Palacios, L. E., León-Vargas, D. A., Bucheli-Guerrero, V. A., & Ordoñez-Eraso, H. A. (2020). Solar Radiation Prediction on Photovoltaic Systems Using Machine Learning Techniques. *Revista Facultad De Ingeniería*, 29(54), e11751. <https://doi.org/10.19053/01211129.v29.n54.2020.11751>

APPENDIX 1: Measures taken from the earthquake networks (EN) constructed with different cell sizes (dx).

<i>dx</i>	<i>Nodes</i>	<i>Edges</i>	<i>k</i>	<i>γ</i>	<i>C</i>	<i>APL</i>	<i>SW</i>	<i>H</i>	<i>D</i>	<i>Complexity</i>
10	4345	6090	2.8032	2.1511	0.0394	5.5843	88.8515	0.0771	0.0312	0.0024
12	4070	6029	2.9627	2.5204	0.0451	5.1144	92.7115	0.0939	0.0291	0.0027
14	3777	5931	3.1406	2.5501	0.0615	4.6479	114.4515	0.1100	0.0277	0.0030
16	3543	5826	3.2887	2.5399	0.0641	4.4053	107.6615	0.1240	0.0266	0.0033
18	3267	5767	3.5305	2.6329	0.0814	4.1120	117.4450	0.1417	0.0262	0.0037
20	3007	5687	3.7825	2.2007	0.0957	3.9207	116.7563	0.1592	0.0259	0.0041
22	2835	5525	3.8977	2.3315	0.1139	3.7026	130.7151	0.1723	0.0245	0.0042
24	2588	5456	4.2164	2.2828	0.1326	3.5332	125.7863	0.1913	0.0252	0.0048
26	2435	5406	4.4402	2.0904	0.1422	3.4512	118.1652	0.2044	0.0243	0.0050
28	2271	5328	4.6922	2.1687	0.1468	3.3695	105.4048	0.2136	0.0261	0.0056
30	2109	5162	4.8952	1.9697	0.1713	3.2091	110.8437	0.2325	0.0261	0.0061
32	1964	5270	5.3666	2.1376	0.1470	3.2974	73.6363	0.2454	0.0272	0.0067
34	1802	4930	5.4717	1.9448	0.2115	3.0337	101.2912	0.2565	0.0276	0.0071
36	1669	4786	5.7352	2.0096	0.2324	2.9513	97.3503	0.2725	0.0274	0.0075
38	1581	4786	6.0544	1.7763	0.2283	2.9480	82.7237	0.2831	0.0297	0.0084
40	1460	4976	6.8164	1.6412	0.2119	3.0265	56.9358	0.2947	0.0342	0.0101
42	1384	4669	6.7471	1.7995	0.2199	2.9278	58.3797	0.3039	0.0326	0.0099
44	1287	4440	6.8998	1.7909	0.2489	2.8155	61.1305	0.3204	0.0329	0.0105
46	1222	4225	6.9149	2.0006	0.2699	2.7499	63.7716	0.3284	0.0313	0.0103
48	1137	4310	7.5814	1.8437	0.2660	2.7334	50.6918	0.3442	0.0330	0.0114
50	1058	4517	8.5388	1.7291	0.2691	2.7649	39.1629	0.3572	0.0370	0.0132
52	994	4440	8.9336	1.7338	0.2640	2.7307	33.8994	0.3749	0.0384	0.0144
54	941	4218	8.9649	1.7164	0.3074	2.6597	37.8708	0.3804	0.0380	0.0144
56	878	4116	9.3759	1.6258	0.2873	2.6448	30.8074	0.3899	0.0409	0.0159
58	829	3958	9.5489	1.5739	0.3040	2.5965	30.2719	0.4072	0.0413	0.0168
60	792	3857	9.7399	1.4044	0.3207	2.5658	29.8031	0.4082	0.0445	0.0182
62	748	4010	10.7219	1.4550	0.3225	2.5587	24.5299	0.4221	0.0456	0.0192
64	698	3584	10.2693	1.6855	0.3087	2.4745	23.8375	0.4394	0.0430	0.0189
66	659	3794	11.5144	1.5563	0.3337	2.4701	20.5372	0.4563	0.0462	0.0211
68	623	3518	11.2937	1.4470	0.3454	2.4042	21.0378	0.4642	0.0466	0.0216
70	582	3469	11.9210	1.4350	0.3532	2.3658	18.7217	0.4806	0.0470	0.0226
72	559	3562	12.7442	1.4339	0.3578	2.3939	16.2957	0.4944	0.0507	0.0251
74	526	3448	13.1103	1.3977	0.3448	2.3201	14.5152	0.5086	0.0482	0.0245
76	506	3618	14.3004	1.3067	0.3657	2.3560	12.8559	0.5173	0.0509	0.0263
78	479	3522	14.7056	1.2245	0.3654	2.3735	11.5132	0.5193	0.0528	0.0274
80	463	3404	14.7041	1.3767	0.3853	2.3657	11.7092	0.5295	0.0532	0.0282
82	442	3602	16.2986	1.2631	0.3869	2.3880	9.5897	0.5441	0.0586	0.0319
84	429	3386	15.7855	1.2392	0.3892	2.3286	9.9799	0.5488	0.0590	0.0324
86	425	3166	14.8988	1.1956	0.3561	2.3290	9.7712	0.5410	0.0580	0.0314

88	387	3162	16.3411	1.3510	0.4219	2.2851	9.3253	0.5649	0.0617	0.0349
90	377	3160	16.7639	1.0862	0.3983	2.2848	8.2489	0.5709	0.0583	0.0333
92	359	3042	16.9471	1.0446	0.3812	2.2312	7.5243	0.5860	0.0600	0.0352
94	342	3025	17.6901	1.1414	0.4099	2.2414	7.1809	0.5920	0.0613	0.0363
96	327	2881	17.6208	1.1322	0.4130	2.1858	7.0758	0.6031	0.0612	0.0369
98	321	2849	17.7508	1.0789	0.3864	2.1964	6.3837	0.6088	0.0551	0.0336
100	314	2975	18.9490	1.1726	0.4502	2.2147	6.5838	0.6012	0.0627	0.0377
102	295	2725	18.4746	1.0318	0.4407	2.1777	6.3007	0.6104	0.0621	0.0379
104	286	2573	17.9930	1.0462	0.4499	2.1656	6.4630	0.6198	0.0619	0.0384
106	278	2538	18.2590	0.8877	0.4649	2.1969	6.2423	0.6151	0.0619	0.0380
108	265	2582	19.4868	1.0549	0.4517	2.1537	5.3583	0.6440	0.0659	0.0424
110	256	2660	20.7813	1.0709	0.5090	2.1481	5.3348	0.6411	0.0725	0.0465
112	240	2524	21.0333	0.8882	0.4756	2.1202	4.6053	0.6547	0.0698	0.0457
114	232	2419	20.8534	0.9819	0.4785	2.1135	4.5165	0.6753	0.0655	0.0442
116	229	2351	20.5328	1.0269	0.4946	2.1011	4.7206	0.6748	0.0719	0.0485
118	225	2379	21.1467	0.8447	0.5342	2.1156	4.7684	0.6578	0.0741	0.0488
120	214	2368	22.1308	0.7787	0.5460	2.1057	4.3439	0.6724	0.0765	0.0515

APPENDIX 2: Regression metrics extended.

Regression metrics - Next event days

Net. Type	N. Evs.	N. trees	Full/mod	MAE	MAPE	MSE	R2
2d	50	1000	full	0.918	1.788	25.292	0.868
2d	100	1000	full	0.919	1.827	16.151	0.864
3d	100	1000	full	0.920	1.834	16.586	0.863
3d	50	1000	full	0.921	1.817	25.266	0.865
2d	10	1000	full	0.927	1.835	22.958	0.864
2d	50	100	full	0.928	1.859	23.792	0.862
3d	100	100	full	0.929	1.895	13.471	0.859
2d	100	100	full	0.930	1.891	16.666	0.859
3d	100	1000	mod	1.010	2.284	15.610	0.830
2d	50	1000	mod	1.011	2.147	39.809	0.841
3d	50	1000	mod	1.011	2.166	39.456	0.839
2d	100	1000	mod	1.014	2.338	17.327	0.825
3d	100	100	mod	1.018	2.344	15.384	0.825
2d	100	100	mod	1.021	2.399	16.167	0.821
2d	50	100	mod	1.024	2.232	38.264	0.835
3d	50	100	mod	1.025	2.256	38.495	0.833

Regression metrics - Magnitude next event

Net. Type	N. evs.	N. trees	Full/mod	MAE	MAPE	MSE	R2
3d	50	1000	full	0.117	0.026	0.025	0.859
2d	50	1000	full	0.118	0.027	0.025	0.858
3d	50	100	full	0.118	0.028	0.025	0.853
3d	100	1000	full	0.119	0.028	0.025	0.854
2d	100	1000	full	0.119	0.027	0.025	0.855
2d	50	100	full	0.119	0.028	0.025	0.852
3d	100	100	full	0.120	0.029	0.025	0.848
2d	100	100	full	0.120	0.029	0.025	0.848
3d	100	1000	mod	0.126	0.032	0.027	0.828
3d	50	1000	mod	0.127	0.032	0.027	0.830
3d	100	100	mod	0.128	0.034	0.027	0.822
3d	50	100	mod	0.128	0.033	0.027	0.826
2d	50	1000	mod	0.129	0.033	0.027	0.823
2d	100	1000	mod	0.129	0.034	0.027	0.820
2d	50	100	mod	0.130	0.034	0.028	0.817
2d	100	100	mod	0.131	0.035	0.028	0.814

Regression metrics - Logarithm of seismicity for the next 10 events

Net. Type	N. evs.	N. trees	Full/mod	MAE	MAPE	MSE	R2
3d	100	1000	full	0.177	0.077	0.011	0.979
3d	100	100	full	0.180	0.081	0.011	0.978
2d	100	1000	full	0.181	0.082	0.011	0.977
2d	100	100	full	0.185	0.087	0.011	0.976
3d	50	1000	full	0.203	0.097	0.012	0.972
2d	50	1000	full	0.205	0.097	0.012	0.972
3d	50	100	full	0.209	0.105	0.013	0.970
2d	50	100	full	0.212	0.106	0.013	0.969
3d	100	1000	mod	0.277	0.173	0.017	0.952
3d	100	100	mod	0.281	0.181	0.017	0.950
2d	100	1000	mod	0.291	0.188	0.018	0.948
2d	100	100	mod	0.296	0.198	0.018	0.945
2d	50	1000	mod	0.321	0.218	0.019	0.937
3d	50	1000	mod	0.321	0.221	0.019	0.936
2d	50	100	mod	0.324	0.229	0.020	0.934
3d	50	100	mod	0.326	0.232	0.020	0.933

Regression metrics: Logarithm of seismicity for the next 20 events

Net. Type	N. evs.	N. trees	Full/mod	MAE	MAPE	MSE	R2
3d	100	1000	full	0.118	0.042	0.007	0.988
3d	100	100	full	0.122	0.044	0.007	0.987
2d	100	1000	full	0.125	0.048	0.007	0.986
2d	100	100	full	0.129	0.052	0.008	0.985
3d	50	1000	full	0.150	0.064	0.009	0.980
2d	50	1000	full	0.154	0.064	0.009	0.980
3d	50	100	full	0.155	0.070	0.009	0.979
2d	50	100	full	0.158	0.068	0.009	0.979
3d	100	1000	mod	0.222	0.120	0.013	0.966
3d	100	100	mod	0.226	0.125	0.013	0.964
2d	100	1000	mod	0.250	0.142	0.015	0.959
2d	100	100	mod	0.255	0.151	0.015	0.957
3d	50	1000	mod	0.283	0.178	0.017	0.946
3d	50	100	mod	0.289	0.189	0.017	0.942
2d	50	1000	mod	0.295	0.185	0.018	0.943
2d	50	100	mod	0.299	0.194	0.018	0.941

APPENDIX 3: Publications made during the doctoral studies.



A revision of seismicity models based on complex systems and earthquake networks

Daniel Leon · Juan Valdivia · Victor Bucheli

Received: 8 July 2020 / Accepted: 23 May 2021 / Published online: 19 November 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract Statistical laws such as the Omori and Gutenberg-Richter laws have become one of the most important tools for seismic hazard. The introduction of earthquake networks by Abe and Suzuki has opened a new research field in seismology, which gives an option to view the seismicity as a complex phenomena in an approachable way and may help to improve seismic hazard assesment. This paper intends to make a revision of the earthquake network models proposed in the literature, and to look for open questions by the date which may be of interest for future research.

Keywords Earthquake networks · Seismicity

1 Introduction

Through the statistical analysis of seismicity, two of the most important laws of seismology have been formulated: the Omori law (Utsu et al. 1995) which describes the frequency of occurrence of aftershocks after a great earthquake, and the Gutenberg-Richter law (Gutenberg and Richter 1944) which describes

the general frequency of occurrence of events with different magnitudes. These two laws involve some behaviors that follow power laws, which are common in complex systems. As a result, different investigations of the seismic phenomenon have been carried out expressing it as a complex system, such as the formulation of models of the earthquake nucleation (Olami et al. 1992; Bak and Tang 1989) (which is, how an earthquake is generated) or seismicity in general as complex networks (Abe and Suzuki 2004b; Baiesi and Paczuski 2004; He et al. 2014; Rezaei et al. 2018; Telesca and Lovallo 2012). In particular, with the presentation of a sequence of seismic events as a complex network, an attempt has been made to understand the structure of these earthquake networks, and in these formulations, it has been specified that seismicity can be seen as a small-world network (Abe and Suzuki 2004b; Chorozoglou et al. 2018; León et al. 2018) and as a Scale-Free network (Abe and Suzuki 2004a; Baiesi and Paczuski 2004). For each of these cases, it has been reviewed the distribution of the degrees, the links, and the effect that large events can generate on the structure of the networks. In this regard, the dynamic analysis of earthquake networks has allowed see some changes in network parameters before and after strong earthquakes (Abe and Suzuki 2007, 2012; Chorozoglou et al. 2018; Pastén et al. 2016; Pastén et al. 2018).

Complex networks are a topic of research of great interest. Much of the generalities of complex networks have been commented by Boccaletti et al. (2006),

D. Leon (✉) · V. Bucheli
Escuela de Ingeniería de Sistemas y Computación,
Universidad del Valle, Cali, Colombia
e-mail: daniel.leon@correounivalle.edu.co

J. Valdivia
Departamento de Física, Universidad de Chile,
Santiago, Chile

where characteristics used to analyze networks such as clustering, degree distribution, connectivity, centrality, and path lengths are described. Some of the research lines in complex networks that may be of interest for seismology have to do with predictions; hence, it is relevant the prediction of links in complex networks using probabilistic and similarity methods (Lü and Zhou 2011) which may help in prediction tasks in an analysis of seismicity.

The purpose of this document is to collect a good part of the efforts that have been made in the analysis of earthquake networks to know what questions remain to be resolved in this area of research. The paper will discuss some theoretical foundations that support seismicity analysis as a complex phenomenon and complex network analysis. Then, we try to show the methodologies used to build earthquake networks, the most relevant characteristics used in the analysis of these networks and the information that these measured parameters of the networks provide about seismicity. Finally, we gather some conclusions and ask some questions that can serve as motivation for future works.

2 Complexity in seismicity and complex networks

The work of Omori (Utsu et al. 1995) and Gutenberg and Richter (1944) on the behavior of seismicity resulted in two of the most important laws of seismology, which describe the frequency of occurrence of aftershocks after an earthquake, as well as the frequency of occurrence of events with different magnitudes. Since in both laws there is a power law behavior, the complexity of the seismicity phenomenon is evidenced. In the work of Bak and Tang (1989), seismicity is modeled as a phenomenon with critical self-organization using blocks linked by springs, and the power laws of Gutenberg-Richter and Omori are related to behaviors of critical self-organization that emerge from such system. This work defines an initial framework to study seismicity as a complex phenomenon, which continues with the work of Ito (1992), discussing fractality and multifractality in seismicity. In the preface of “Introduction to Nonextensive Statistical Mechanics” by Tsallis (2009) it is said that “*One of the main – necessary but by no means sufficient – features of complexity has to do*

with the fact that both very ordered and very disordered systems are, in the sense of plectics, considered to be simple, not complex” (Plectics has to be understood as the science that studies complexity). In this work are described some of the analytical tools used in complex systems, and it is also explained that the entropy of such systems differs from the entropy of systems of the random type such as fluid or chaotic in the sense that the total entropy of the system is not the sum of the entropy of its parts, which requires a *non-extensive statistical mechanics*. This type of tools, applied to seismology, derives in a branch that can be described as non-extensive statistical seismology, collected by Vallianatos et al. (2018), where many of the phenomena that make seismicity seen as a complex phenomenon are described, such as scaling relationships, spatial and temporal correlations, and critical phenomena; as well as the way to apply non-extensive statistical mechanics in seismology.

Boccaletti et al. presented a state of the art (Boccaletti et al. 2006) about calculations made in complex networks; numerous parameters and measures related to complex networks are described, as well as characteristics of real networks. Many of the characteristics inherent to complex systems are found in complex networks, such as self-similarity, fractality, and scalability. There are different types of complex networks. The most important ones can be classified as *small-world*, *scale-free*, and *random* (random networks are referenced in the literature as *Erdős-Renyi* networks). Completely random networks began to be studied by Erdos and Renyi (1960), which became a first effort to lay the foundations with which the networks were studied later. *small-world* networks were coined by Watts and Strogatz (1998), and are presented as a type of networks that lie between regularly connected networks and networks connected completely randomly. *Scale-free* networks are a type of networks that demonstrate a distribution that follows a power law in the number of connections that nodes have (known as the degree of a node), and have the property of having a few nodes with many connections and many nodes with few (Barabási 1999). Real-world networks are normally compared to and studied from the perspective one of these three type of networks. Table 1 is taken from Chorooglou et al. (2019), and some of the parameters of the networks that are most used for analysis are described.

Table 1 Seven network measures in complex networks. Taken from Chorooglou et al. (2019)

Network parameter	Description
Degree centrality	Average over all nodes of either the number or the strength of connections.
Clustering coefficient	Average over all nodes of the fraction of connections between the nodes within their neighborhood, divided by the number of all possible connections.
Characteristic path length	Average of the shortest path lengths in the network, computed over all pairs of nodes.
Global efficiency	Average inverse shortest path length in the network, computed over all pairs of nodes.
Betweenness centrality	Average over all nodes of the node betweenness centrality, which is the fraction of all shortest paths in the network that contain the node, divided by the number of all possible paths.
Modularity	Shows the strength of a network that is divided into components.
Eccentricity	Average over all nodes of the maximal shortest path length between a node and any other node.

Some complementary parameters have been proposed to analyze networks, such as small-worldness (which, for a network i is defined as $S = \frac{C}{C_r}$, and it is calculated comparing the clustering and the path lengths in the network with the same parameters of an Erdős-Renyi random network) (Humphries and Gurney 2008), or Renyi's entropy (Duan et al. 2019). Wiedermann et al. (2017) show a way to characterize different types of complex networks using the entropy calculated from the distribution of node degrees, and statistical complexity calculated from the adjacency matrices of the network, as a product of entropy and divergence of the network from a state of equilibrium, which for the purposes of complex networks is chosen as a similar random network. By plotting different networks on a plane whose coordinates are these two parameters (entropy and complexity), different types of networks can be located in different places of the entropy-complexity plane; delivering a tool that allows to compare different types of networks. Other types of entropies can be reviewed, as entropy can be estimated from the distribution of any measure from the nodes, such as degree, centrality, or average path length. While the calculation of entropy is valid to check the amount of information in a system, using Shannon's entropy may not be the best option in earthquake networks, since as commented by Vallianatos et al. (2018) seismicity has properties which can be better described using non-extensive statistical mechanics, so the use of Tsallis entropy could be more appropriate to describe such systems.

3 Earthquake networks

A sequence of earthquakes is a series of events that occur in space and time. Gao et al. (2016) present a state of art regarding different methods with which complex networks can be constructed from time series, such as visibility graphs (several types of these graphs are listed), recurrence networks, and networks based on transitions between other methods. It is commented that this type of analysis could open a place for cross-fertilization between methods used to analyze time series and the theory of complex networks. It is interesting about this article regarding the research proposed in earthquake networks that as earthquakes follow a temporal sequence, and so a seismicity catalog can be seen as a time series of events, they can be expressed as complex networks in ways than have not yet been contemplated in the literature. The joint use of complex networks and time series is also evident in the work by Pastén et al. (2018), which shows a procedure to transform a complex network of earthquakes into a time series through a non-random walk, following the sequence of occurrence of earthquakes. The time series constructed from this connectivity shows changes in multifractality due to characteristics in the studied areas, and that in general multifractality decreases when seismic activity increases. This work shows that there are other parameters that can be used to discover not so obvious patterns of seismicity, since in one of the areas studied, the increase in multifractality after large earthquakes was very different than in another area because one of the earthquakes had many more aftershocks than the other. Reviewing complex

Table 2 Rules used to define nodes and links in earthquake networks

Definition of nodes	Definitions of links
Region divided in cells, and a cell is chosen as a node if an earthquake occurs into it.	Cells are linked following the temporal sequence of earthquakes Abe et al. (2004a, b, 2011), Chorozoglou et al. (2017), Leon et al. (2018). Cells are linked if a visibility condition between nodes is met, based on the magnitudes and the time between events (Rezaei et al. 2018). Cells are linked if it is met that an earthquake in the following cell is inside a space-temporal influence domain for the previous event in the first cell (He et al. 2014).
Events as nodes	Events are linked if a visibility criteria is met based on the magnitude and the time between events (visibility graph) (Telesca and Lovallo 2012). Events are linked if they are inside the space-temporal influence domain (Xu et al. 2018). Events are linked using a correlation factor from magnitudes, occurrence times and distances (Baiesi and Paczuski 2005; Suteanu 2015).

networks as time series seems to generate very interesting information about the activity and deserves to be taken into account as an analysis tool.

Abe and Suzuki (2004a) were pioneers in the analysis of seismicity using complex networks. The constructed earthquake networks contain various properties that can be directly related to parameters estimated from traditional seismicity analyzes, showing that the analysis of earthquake networks can be a tool that allows to analyze the complexity of the seismicity. From that work, different ways of building earthquake networks have been developed, which can be differentiated in the way in which nodes are defined. On the one hand, there are earthquake networks that are constructed by dividing a region into cells and a cell is selected as a node as long as a seismic event occurs in it. On the other hand, there are earthquake networks in which the earthquakes are the nodes. Table 2 presents different methods of creating earthquake networks, indicating how the nodes are defined, the rules governing the links between nodes, and it is indicated in which works such networks have been used.

3.1 Cells as nodes

Abe and Suzuki (2004b) describe a way to create an earthquake network from a seismicity catalog, performing a division of the studied region in cubic cells which are chosen as nodes of the network if earthquakes occur within the cells, and are connected following the temporal sequence of the earthquakes.

Figure 1 is taken from León et al. (2018), and it is shown how these networks are constructed.

The observation of the earthquake networks reported by Abe and Suzuki (2004a) for seismicity in California and in Japan gave as a result that these networks have a degree distribution which follows a power law, so it was seen that these are scale-free networks: result that somehow agrees with the Omori and Gutenberg-Richter laws, to the extent that large earthquakes that have many aftershocks function as hubs of these networks. Earthquake networks can be built with or without the directivity of the connections; however, since in the case of directed networks the nodes have the same input and output degree, the networks can be analyzed as non-directed. Also, links of a node with itself (loops) as well as multiple or a single link between two nodes may or may not be taken into account to analyze the small-world nature of the earthquake network (Abe and Suzuki 2004b; León et al. 2018). Abe et al. (2011) analyze earthquake networks from both perspectives; and it can be seen that the clustering in these networks is much greater and that the paths are shorter than in random networks. In this work, an analysis of the cell size variation is made, which shows that clustering and connectivity remain constant ($C \sim 0.85$ and $\gamma \sim 1.0$) after exceeding a given size, which is a consequence of the cells grouping more events (number of nodes decrease), but the nodes have greater connectivity (the links on certain nodes increase). In this work, it can also be seen that there are global seismic behaviors which

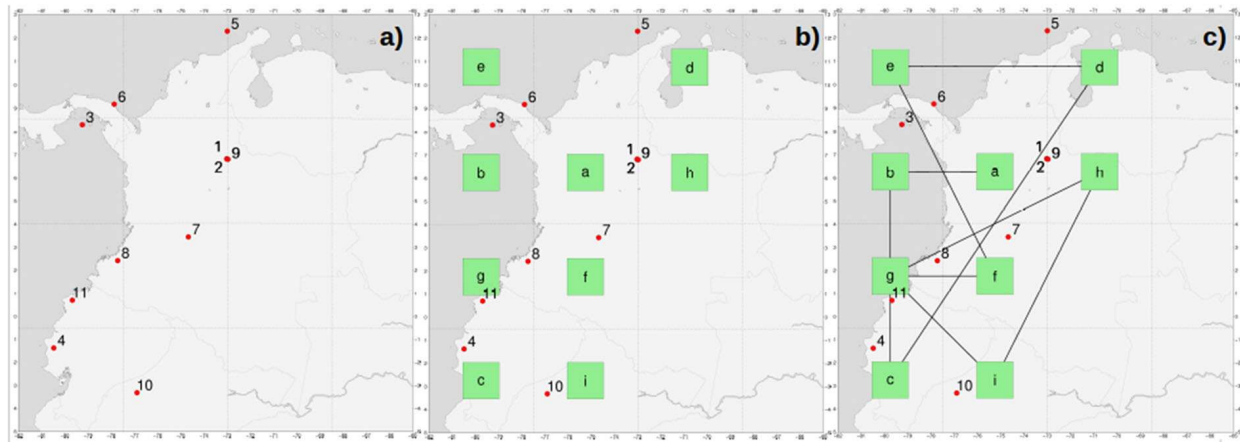


Fig. 1 Diagram of the earthquake network produced by the spatiotemporal sequence of earthquakes. **a** The sequence of events (numbered dots) is mapped on a grid. The number near the events follows the temporal order from oldest to newest. **b** A cell from the grid is selected as a node if there are events mapped

inside the cell. The letter inside the node (squares) identifies the node. **c** Nodes are connected following the temporal sequence of events, linking two nodes with successive events. Taken from León et al. (2018)

can be evaluated from the point of view of complex networks.

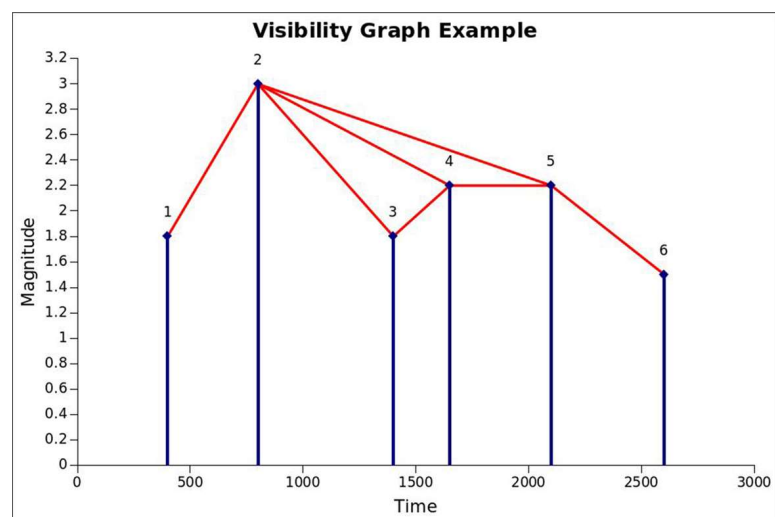
3.2 Earthquakes as nodes

Another way to build the earthquake networks is to link the earthquakes between them based on different rules. There are different criteria that have been used to link earthquakes, ranging from visibility graphs, correlation of events or belonging to the same domain of space-time influence. Telesca and Lovallo (2012) propose to build earthquake networks based on visibility graphs based on the magnitude of the events

and the temporal distance between them. Figure 2, inspired from that same article, shows how these types of graphs are constructed.

Visibility graphs were introduced by Lacasa et al. (2008), as a simple computational tool to study series from the perspective of graphs, connecting different nodes from some measure of amplitude. This type of graph retains many characteristics of the time series they represent, so Telesca and Lovallo in their work introduce this technique to review the series of earthquake magnitudes, showing that the distribution of degrees follows a power law.

Fig. 2 Earthquake network constructed from a visibility graph. Earthquakes are mapped as dots in a magnitude vs. time plane; and links are made so they don't intersect any vertical line from the dots and the time axis. Inspired from Telesca and Lovallo (2012)



Baiesi and Paczuski (2005) describe another way to create a directed earthquake network from a certain earthquake to subsequent earthquakes using an event correlation metric.

Correlation between two events is defined as $c_{ij} = 1/n_{ij}$, where $n_{ij} \equiv K t l^{d_f} 10^{-b m_i} \Delta m$. The correlation metric takes into account the difference of the magnitudes Δm , the distance l , and the time t that occurs between the events, as well as parameters derived from the analysis of the Gutenberg-Richter curve in the study region (parameter b), the fractal dimension of seismicity d_f , and the calculation of a constant K which is calculated for a region, time interval, and underlying seismicity. Paradoxically, this constant depends on many factors, so it would be interesting to make the correlation metric more robust by directly taking the values on which it depends instead of estimating the entire catalog, since it is known the number of earthquakes, the b parameter and the magnitude differences are highly dependent on the occurrence of large events. In the end, the paper shows that the networks built with this method are of the scale-free type, and concludes that the correlation threshold seems a bit arbitrary and generates different networks that, in any case, maintain in their structure a constant exponential parameter that scales the magnitude of an earthquake with the distance of each replica. It also concludes that this approach serves as a sign that some universal properties of networks may exist independent of the type of metric used.

Suteanu (2015) propose a way to create networks of earthquakes similar to that proposed by Baiesi and Paczuski (2005), for volcanic events in Hawaii; however, the links have a weight which is a combination of spatial, temporal, and magnitude variables. In the formulas used to estimate the weights, constants are used which, for the purposes of the study, are tested with different values. These two studies motivate, in some way, the search for parameters and metrics of seismic networks that can characterize seismicity.

Xu et al. (2018) propose another way of building earthquake networks similar to that proposed by Baiesi and Paczuski. In these networks, a correlation threshold is not used as such, but it is stated that each earthquake has a threshold of influence in time and distance that depends on the magnitude, whereby links are generated between events that are within the domain of that space-time influence. By using this method, the division by node communities was

explored, concluding that using the communities can slightly improve the prediction of events (although the accuracy of the predictions is still below 0.5).

3.3 Seismicity information from earthquake networks

Measurements from the earthquake networks have a different meaning, depending on the type of earthquake network constructed. For instance, the degree in a *cells as nodes* network implies how many times an earthquake can occur in the sector which corresponds to that node after an earthquake has occurred on other sectors, so a high degree for a node implies many earthquakes in a sector. For the *earthquakes as nodes* networks, degree implies how many earthquakes are inside its influence radius (that is, how many earthquakes can be related as aftershocks or foreshocks), so a high degree for a node implies an important earthquake in terms of the quantity of foreshocks and aftershocks. The meaning of the other parameters merits a more detailed analysis respect to the implications in the seismicity dynamics.

From the analysis of the parameters found in earthquake networks, some conclusions found in the literature can be summarized on Table 3.

Other conclusions that emerge from other works can be listed below. Chorozoglou et al. (2017) made a review of the evolution of the earthquake networks created using the Abe and Suzuki method before and after large earthquakes, analyzing the cases with and without weights in the links, and using 30-day windows which move in periods comprising 300 days before and 150 days after a large earthquake. The clustering coefficient, eccentricity, small-worldness, path length, and efficiency of the networks are reviewed (efficiency in a network can be seen as the inverse of the path length, and indicates how effective is the transmission of data by the network), and the calculation of a statistical factor p in which the earthquake network is compared with similar random networks in the degree distribution (for the case without weights) and in the connectivity (for the case with weights). The conclusions show that in the case of links with weights, there is a behavior that seems to announce in a short time the appearance of a large earthquake, and an increase in centrality is appreciated before a large event. In this work can be seen that earthquake networks could be applied as short-term prediction tools (a few months) before a large earthquake, although it

Table 3 Earthquake network parameters

Network parameter	Seismicity information
Community division	Earthquake prediction by the network can be improved using a community division (Xu et al. 2018).
Quantity of links	It is shown that the Gutenberg-Richter data can be estimated using the Abe-Suzuki model and the Omori law using the Telesca-Lovallo model (Rezaei et al. 2017)
Connectivity distribution	It follows a power law, which agrees with the Gutenberg-Richter law; showing the presence of giant components (earthquakes that generate aftershocks), and a behavior of <i>preferential attachment</i> (Abe and Suzuki 2006).
Multifractality	Multifractality is evaluated in a temporal series created from an earthquake network, it is shown a chance by the presence of strong earthquakes (Pastén et al. 2018).
Small-worldness and clustering	Changes in small-worldness and clustering can be of use to indicate a possible seismic risk in a short term (Chorozoglou et al. 2017).

is noted that it is important to continue investigating to corroborate this statement.

Deyasi et al. (2017) made another review of earthquake networks created with the Abe and Suzuki method, for different regions in the world that are located between different tectonic plates. A measurement of the similarity of the networks from the Jensen-Shannon divergence is used, and it was observed that seismic regions near the same tectonic plate can generate similar networks. The paper raises the difficulty of measuring the distance between two earthquake networks built with different amounts of seismic events and generating networks with different numbers of nodes. This difficulty was addressed by calculating the spectrum of adjacency matrices, and although it is not mentioned, it is shown in the article that the structure of the networks have some characteristics which are conditioned by the tectonics of the region and that this conditioning covers the regions that are close to such tectonic structures.

Ohsawa (2018) presents a method to detect foreshocks of large events using a Regional Entropy of Seismic Information index. Entropy, as in almost all work done in complex systems, is evaluated from the point of view of the amount of information in a system. The method is applied in seismic clusters, which are created by dividing the study region into a mesh and the cumulus sizes in each mesh cell are estimated. It is observed that some gaps appear in the seismicity before the appearance of a large earthquake, which has been seen to correspond to changes in the behavior of this entropy index; and thus the index can detect

a precursor activity up to a year before a large earthquake. Although in this work a seismic network is not created, the idea of using the changes in the entropy of the system, that is, the changes in the information of the seismic system, shows promising results for an earthquake prediction system, which could well be used in an earthquake network where the calculation of entropy is easier.

About what is the right size to build a cells-as-nodes earthquake network, some works have been found. Lotfi and Darooneh (2012) revise a resolution parameter which is inversely proportional to the size of the cells, and it is concluded that when the resolution increases, the characteristics of the networks become very dependent on each other. Also, Abe and Suzuki (2009) reviewed the cell size, and it is found that increasing the cell size there is a particular size in which the clustering ceases to have considerable variations (about 0.85 for several regions that were studied).

4 Conclusions and discussion

From all these elements, the possibility of building a seismicity model from the perspective of complex networks is raised. The following are questions that remain open:

What is the best cell size to create an earthquake network of the type cell as a node? From the representation with all the links (both with loops and multiple links between nodes) in the seismic networks of the cell type as node, how much can be predicted about the

seismological activity? And how much from the simplified representation that evaluates the small-world characteristics of earthquake networks? Entropy, complexity, and divergence as evaluated by Wiedermann et al. have not been evaluated in detail in earthquake networks. How much information do these parameters provide about seismological activity? In seismology, the analysis of the dynamics of mechanical stresses (internal forces within the earth) as a result of seismological activity is of great interest. How can a model that relates earthquake networks and the mechanical stress release due of the occurrence of earthquakes be created?

For the authors, what has to do with a model that relates stress release with earthquake networks can be very complicated and it would be very interesting to think of some first steps to outline a model of this type. However, based on what is already known about earthquake networks, another research question which can be raised that brings together several of these issues in this way:

How can a dynamical model of seismicity can be developed from the representation of an earthquake network?

The research question has several questions implicit that should be answered. The first has to do with what are the best ways to create earthquake networks and what information each one delivers. The problem could be limited by reviewing only the representation of networks of the cells as nodes type, and the problem of cell size should be reviewed. The type of networks constructed has to be reviewed; hence, the statistical divergence used by Wiedermann et al. can be used as a similitude measurement respect the three main types of networks. Also, it has to be answered which are the more appropriated measures from the networks to create a dynamical model of seismicity. There are many that have been used in previous works, but other measures can be used such as the Renyi's entropy used by Duan et al. (2019), and the Tsallis-Shannon entropy used by Yulmetyev et al. (2004); or based on the distribution of other measures such as centrality, clustering, or path lengths which can involve different scenarios (in this respect it can be remembered that the entropy seen from Tsallis' vision generated the branch of non-extensive statistical mechanics Tsallis 2009). The proposed model should somehow describe some characteristics of the dynamics of seismicity, so the measurement of time

in this sense becomes important in the formulation of the model. Although it cannot be specified from now on whether specific seismic events can be predicted with the proposed model, it is expected that with this model some kind of behavior can be foreseen; and this is already a sufficient motivation to say that this research is relevant: every step that is taken in the way of understanding the seismic phenomenon and that can announce a seismicity behavior will be very important for science and in particular for the nearby population to the region studied. It is expected that the proposed model can be used in different regions, and just as with the Gutenberg-Richter and Omori laws, some parameters may change depending on the region or the type of events that are presented, so it is expected that the model can characterize a seismic region and allow comparison with the model applied to the seismicity of other regions. In this regard, the question also arises of which regions should be analyzed to raise the seismicity model.

References

- Abe S, Suzuki N (2004a) Europhys Lett (EPL) 65(4):581–586. <https://doi.org/10.1209/epl/i2003-10108-1>. <http://stacks.iop.org/0295-5075/65/i=4/a=581?key=crossref.d3566b8167d937b80a185a57db0b8b37>
- Abe S, Suzuki N (2004b) Physica A: Stat Mech Appl 337(1–2):357–362. <https://doi.org/10.1016/j.physa.2004.01.059>, <http://linkinghub.elsevier.com/retrieve/pii/S037843710400144X>
- Abe S, Suzuki N (2006) Phys Rev E 74(2). <https://doi.org/10.1103/PhysRevE.74.026113>
- Abe S, Suzuki N (2007) Eur Phys J B 59(1):93–97. <http://linkinghub.elsevier.com/retrieve/pii/S037843710400144X>
- Abe S, Suzuki N (2009) EPL (Europhys Lett) 87(4):48008. <https://doi.org/10.1209/0295-5075/87/48008>
- Abe S, Pastén D, Suzuki N (2011) Physica A: Stat Mech Appl 390(7):1343–1349. <https://doi.org/10.1016/j.physa.2010.11.033>
- Abe S, Suzuki N (2012) EPL (Europhys Lett) 99(3):39001. <https://doi.org/10.1209/0295-5075/99/39001>, <http://stacks.iop.org/0295-5075/99/i=3/a=39001?key=crossref.8ddf2f1644f77ca1a2261680eb7f0e8c>
- Baiesi M, Paczuski M (2005) Nonlinear Process Geophys 12:1–11
- Baiesi M, Paczuski M (2004) Phys Rev E 69(6):066106
- Bak P, Tang C (1989) J Geophys Res Solid Earth 94(B11):15635–15637. <https://doi.org/10.1029/JB094iB11p15635>
- Barabási A (1999) Science 286(5439):509–512. <https://doi.org/10.1126/science.286.5439.509>
- Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D (2006) Phys Rep 424(4–5):175–308. <https://doi.org/10.1016/j.physrep.2006.01.002>

- physrep.2005.10.009, <https://linkinghub.elsevier.com/retrieve/pii/S037015730500462X>
- Chorozoglou D, Kugiumtzis D, Papadimitriou E (2017) *Acta Geophysica* 65(3):543–553. <https://doi.org/10.1007/s11600-017-0039-4>
- Chorozoglou D, Kugiumtzis D, Papadimitriou E (2018) *Physica A: Stat Mech Appl* 499:28–39. <https://doi.org/10.1016/j.physa.2018.01.033>, <https://linkinghub.elsevier.com/retrieve/pii/S0378437118300530>
- Chorozoglou D, Iliopoulos A, Kourouklas C, Mangira O, Papadimitriou E (2019) *Pure Appl Geophys*. <https://doi.org/10.1007/s00024-019-02253-w>
- Deyasi K, Chakraborty A, Banerjee A (2017) *Physica A: Stat Mech Appl* 481:224–234. <https://doi.org/10.1016/j.physa.2017.04.050>, <https://linkinghub.elsevier.com/retrieve/pii/S0378437117303497>
- Duan S, Wen T, Jiang W (2019) *Physica A: Stat Mech Appl* 516:529–542. <https://doi.org/10.1016/j.physa.2018.10.045>, <https://linkinghub.elsevier.com/retrieve/pii/S0378437118313931>
- Erdos P, Renyi A (1960) *Publ Math Inst Hungary Acad Sci* 5:17–61
- Gao Z-K, Small M, Kurths J (2016) *EPL (Europhys Lett)* 116(5):50001. <https://doi.org/10.1209/0295-5075/116/50001>, <http://stacks.iop.org/0295-5075/116/i=5/a=50001?key=crossref.5603428eebda8ca674ea7bb8a95bf3ba>
- Gutenberg B, Richter CF (1944) *Bull Seismol Soc Am* 34(4):185–188
- He X, Zhao H, Cai W, Liu Z, Si S-Z (2014) *Physica A: Stat Mech Appl* 407:175–184. <https://doi.org/10.1016/j.physa.2014.03.093>, <https://linkinghub.elsevier.com/retrieve/pii/S0378437114003045>
- Humphries MD, Gurney K (2008) *PloS one* 3(4):e0002051. <https://doi.org/10.1371/journal.pone.0002051>, <http://europepmc.org/articles/PMC2323569>
- Ito K (1992) *Pure Appl Geophys PAGEOPH* 138(4):531–548. <https://doi.org/10.1007/BF00876337>
- Lü L, Zhou T (2011) *Physica A: Stat Mech Appl* 390(6):1150–1170. <https://doi.org/10.1016/j.physa.2010.11.027>, <https://linkinghub.elsevier.com/retrieve/pii/S037843711000991X>
- Lacasa L, Luque B, Ballesteros F, Luque J, Nuo JC (2008) *Proc Natl Acad Sci* 105(13):4972–4975. <https://doi.org/10.1073/pnas.0709247105>
- León DA, Valdivia JA, Bucheli VA (2018) *Seismol Res Lett* 89(5):1807–1816. <https://doi.org/10.1785/0220180076>, <https://pubs.geoscienceworld.org/ssa/srl/article/89/5/1807/543355/Modeling-of-Colombian-Seismicity-as-SmallWorld>
- Lotfi N, Darooneh AH (2012) *Eur Phys J B* 85(1). <https://doi.org/10.1140/epjb/e2011-20623-x>
- Ohsawa Y (2018) *Entropy* 20(11):861. <https://doi.org/10.3390/e20110861>, <http://www.mdpi.com/1099-4300/20/11/861>
- Olami Z, Feder HJS, Christensen K (1992) *Phys Rev Lett* 68(8):1244–1247. <https://doi.org/10.1103/PhysRevLett.68.1244>
- Pastén D, Torres F, Toledo B, Muñoz V, Rogan J, Valdivia JA (2016) *Pure Appl Geophys* 173(7):2267–2275. <https://doi.org/10.1007/s00024-016-1335-7>
- Pastén D, Czechowski Z, Toledo B (2018) *Chaos: Interdiscip J Nonlinear Sci* 28(8):083128. <https://doi.org/10.1063/1.5023923>
- Rezaei S, Darooneh AH, Lotfi N, Asaadi N (2017) *Physica A: Stat Mech Appl* 471:80–87. <https://doi.org/10.1016/j.physa.2016.12.003>, <https://linkinghub.elsevier.com/retrieve/pii/S0378437116309748>
- Rezaei S, Moghaddasi H, Darooneh AH (2018) *Physica A: Stat Mech Appl* 495:172–179. <https://doi.org/10.1016/j.physa.2017.12.063>, <https://linkinghub.elsevier.com/retrieve/pii/S0378437117313067>
- Suteanu M (2015) *Pure Appl Geophys* 172(7):1865–1878. <https://doi.org/10.1007/s00024-014-0939-z>
- Telesca L, Lavallo M (2012) Analysis of seismic sequences by using the method of visibility graph. *EPL (Europhys Lett)* 97(5):50002. <https://doi.org/10.1209/0295-5075/97/50002>, <http://stacks.iop.org/0295-5075/97/i=5/a=50002?key=crossref.5d5a6330c049c4a28c8974b45dabd71d>
- Tsallis C (2009) Springer, New York. OCLC: 896249077
- Utsu T, Ogata Y, Matsu'ura RS (1995) *J Phys Earth* 43(1):1–33. <https://doi.org/10.4294/jpe.1952.43.1>, <http://joi.jlc.jst.go.jp/JST.Journalarchive/jpe.1952/43.1?from=CrossRef>
- Vallianatos F, Michas G, Papadakis G (2018) In: *Complexity of Seismic Time Series*. Elsevier, pp 25–59. <https://linkinghub.elsevier.com/retrieve/pii/B978012813138100002X>
- Watts DJ, Strogatz SH (1998) *Nature* 393(6684):440–442. <https://doi.org/10.1038/30918>, <http://www.nature.com/articles/30918>
- Wiedermann M, Donges JF, Kurths J, Donner RV (2017) *Phys Rev E* 96(4). <https://doi.org/10.1103/PhysRevE.96.042304>
- Xu Y, Ren T, Liu Y, Li Z (2018) *Physica A: Statistical Mechanics and its Applications* 506:969–974. <https://doi.org/10.1016/j.physa.2018.05.035>, <https://linkinghub.elsevier.com/retrieve/pii/S0378437118305727>
- Yulmetyev RM, Emelyanova NA, Gafarov FM (2004) *Physica A: Stat Mech Appl* 341:649–676. <https://doi.org/10.1016/j.physa.2004.03.094>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”). Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/352067817>

TagEQ-CN: Clasificación de Eventos Sísmicos Basado en Redes Complejas

Article in *Investigación e Innovación en Ingenierías* · May 2021

DOI: 10.17081/invinno.9.1.4856

CITATIONS

0

READS

80

3 authors, including:



Hugo Ordoñez

Universidad del Cauca

109 PUBLICATIONS 347 CITATIONS

[SEE PROFILE](#)



Victor Bucheli

Universidad del Valle (Colombia)

83 PUBLICATIONS 333 CITATIONS

[SEE PROFILE](#)

TagEQ-CN: Clasificación de Eventos Sísmicos Basado en Redes Complejas

TagEQ-CN: Automatic Classification System for seismic events based on complex networks

Daniel Andres Leon Vargas



Víctor Andrés Bucheli Guerrero



Universidad del Valle, Colombia

Hugo Armando Ordoñez Erazo



Universidad del Cauca, Colombia

OPEN ACCESS

Recibido: 17/12/2020

Aceptado: 25/03/2021

Publicado: 12/05/2021

Correspondencia de autores:

hugoeraso@gmail.com



Copyright 2020
by Investigación e
Innovación en Ingenierías

Resumen

Objetivo: Realizar detección y clasificación de señales sísmicas continúa siendo un tema de investigación, dada la diversidad de tipos de señales, sensores y fuentes de vibración. **Metodología:** Para el proceso de clasificación de eventos utiliza información sobre nodos o lugares que aglomeran eventos sísmicos, así como también comunidades de eventos. **Resultados:** El sistema se testeó con señales provenientes de la estación La Rusia que pertenece a la Red Sismológica Nacional de Colombia, con resultados prometedores. La aplicabilidad de este sistema, permite incluir información nueva para la anotación automática de sismos, así como también reconocer automáticamente eventos de otras fuentes. **Conclusiones:** El sistema desarrollado se basa en el paradigma supervisado, el usuario no escoge directamente cuál es el conjunto de señales que se utilizan para el entrenamiento y prueba del clasificador.

Palabras clave: Redes complejas, aprendizaje de máquina, etiquetado semi-automático de señales, eventos sísmicos, modelos de clasificación.

Abstract

Objective: Performing detection and classification of seismic signals continues to be a topic of research, given the diversity of types of signals, sensors and sources of vibration. **Methodology:** For the event classification process, it uses information about nodes or places that agglomerate seismic events, as well as event communities. **Results:** The system was tested with signals from La Russia station that belongs to the National Seismological Network of Colombia, with promising results. The applicability of this system allows including new information for the automatic annotation of earthquakes, as well as automatically recognizing events from other sources. **Conclusions:** The developed system is based on the supervised paradigm, the user does not directly choose the set of signals used for the training and testing of the classifier.

Keywords: Complex networks, machine learning, semi-automatic labeling of signals, seismic events, classification models.

Introducción

En el área de la sismología constantemente se presentan desarrollos tanto en el procesamiento de señales como en la detección y clasificación de señales sísmicas [1,2,3, 4]. El rango de tipos de señales, sensores y fuentes de vibración es amplio y esto ha llevado a que sea una área activa de investigación [5].

Un sismograma puede describirse como el resultado de la convolución de 3 funciones de transferencia diferentes: La función de la fuente de vibración, el efecto del material en el que la onda se ha propagado, y la respuesta del sensor ante la vibración. Por otro lado, diferentes fenómenos pueden ser fuentes de vibración en un mismo lapso, entre las cuales se tienen los procesos tectónicos, atmosféricos, antrópicos, o de cualquier otra naturaleza [6]. Existen muchas implementaciones de algoritmos para hacer detección de eventos, sin embargo, hay limitaciones en cuanto a la generalidad de estos sistemas, debido a la naturaleza tan heterogénea de las señales registradas. Esto representa un tema de interés en las ciencias de la computación, para tratar de encontrar algoritmos que sean cada vez más generales en cuanto a la clasificación de señales.

Los sismos se pueden entender como ondas P y S, donde sus valores de velocidad de propagación permiten estudiar este tipo de eventos. Sin embargo, la detección es un problema abierto de investigación, debido al gran volumen de datos y señales que genera una red sismológica, así, se hace necesario contar con un sistema automático de detección. Al respecto se han desarrollado diferentes trabajos en el área de procesamiento de señales, así como también, en el área de la detección de eventos. Estas dos áreas son de interés especial para el presente trabajo. Existen actualmente muchas aproximaciones a la solución de la detección automática de sismos usando esquemas de aprendizaje de máquina, que son una rama de la inteligencia artificial. Sin embargo, hay una brecha debido a la heterogeneidad y la cantidad de información que contienen las ondas sísmicas. Así, los diferentes tipos de sensores que hacen el registro de eventos recolectan información de los diferentes tipos de estructuras de la tierra que atraviesan las ondas hasta llegar a tales sensores, los diferentes tipos de eventos sísmicos que se presentan tales como volcanes, entre otros. Entonces, la brecha abierta se puede relacionar con el problema de predecir algún tipo de estos eventos, lo cual puede salvar la vida de muchas personas.

Este artículo propone un sistema prototipo de clasificación de eventos sísmicos a partir de redes complejas, el sistema hace parte de una arquitectura más general que basa la clasificación de eventos sísmicos en inteligencia artificial (TagEQ), una descripción general de dicho fue presentada por [7]. En el sistema prototipo propuesto los eventos sísmicos se representan como una red, lo que permite obtener mayor información para la clasificación automática de eventos, así es posible utilizar información sobre nodos o lugares que aglomeran eventos sísmicos, así como también comunidades de eventos sísmicos, entre otras características topológicas de una red compleja. El artículo presenta TagEQ-CN como un sistema de etiquetado semi-automático de señales sísmicas, que toma sismogramas continuos de una estación y, a partir de un catálogo de eventos sísmicos y una representación de los eventos como red, clasifica automáticamente las señales. Se propone un tipo de anotación a partir de características de los eventos respecto a la topología de la red de sismos previamente construida. En este sistema se integraron algoritmos de aprendizaje de máquina supervisados. De igual manera, se desarrolló un corpus de señales a partir de los datos recopilados de La Rusia (estación que pertenece a la Red Sismológica Nacional de Colombia), de tal manera que se puedan aplicar los algoritmos de aprendizaje de máquina. Para ello, se creó un sistema modular para crear y anotar vectores de características tomadas en formas de onda de eventos sísmicos y características de la red de sismos construida. (módulo ETL), y entrenar un clasificador automático a partir de un conjunto de señales anotadas (módulo de entrenamiento). El etiquetado automático se hace frente a

características propias de un nodo, como el grado o la intermediación (betweenness), o frente a si hace parte de un grupo de nodos como una comunidad.

Métodología

En sismología se han implementado muchos tipos de clasificadores, desde Mapas Auto-Organizados [8, 9], Máquinas de Soporte Vectorial [10], o Redes Neuronales [11], buscando clasificar entre pocos tipos de señales (ruido-sismo, sismos locales-sismos regionales-tele-sismos, voladuras-sismos).

La búsqueda de patrones en la distribución espacial y temporal de eventos sísmicos ha permitido encontrar límites de placas tectónicas, describir sus movimientos sobre el manto terrestre, y evaluar el riesgo sísmico [6]. Existen muchos estudios que tratan de encontrar precursores de eventos grandes usando sismos premonitorios (foreshocks) [12], variaciones magnéticas [13], o emisiones de gas [14]; entre otros. Aunque son muy interesantes, la mayoría de estos estudios proveen resultados de pronóstico no concluyentes, y hasta ahora los estudios sugieren la posibilidad de pronosticar un evento de manera confiable dentro de un rango temporal de unos pocos años en el mejor de los casos [15].

Hay muchos estudios acerca de los patrones espaciales en la distribución espacial de eventos sísmicos. Abe y Suzuki [16] estudian la complejidad espacial de la distribución de eventos sísmicos en el sur de California, llevando a una geometría fractal que se encuentra con frecuencia en sistemas complejos no equilibrados. El uso de la construcción de redes de sismos se ha configurado como una manera novedosa de estudiar la complejidad espacial y temporal de la sismicidad. Por ejemplo, Baiesi y Paczuski [17] crearon redes de sismos de sismos principales y réplicas enlazando eventos que estuvieran altamente correlacionados, encontrando una estructura libre de escala en cada cúmulo identificado. De manera similar, en [18] se crearon redes de sismos para estudiar el efecto del sismo del 2015 de Mw=8.3 en Illapel, Chile, en donde se hallaron cambios en parámetros como la conectividad promedio y la intermediación que varían por la ocurrencia de eventos grandes, evidenciando que eventos grandes afectan la estructura de las redes de sismos creadas. Así mismo, Chorozoglou [19] usaron redes complejas para detectar eventos premonitorios antes de dos sismos en Grecia y encontraron que el método podría ser aplicado, junto con otras medidas, para tareas de riesgo sísmico.

En Abe y Suzuki [20] se dice que “una red de pequeño mundo reside entre la regularidad y la aleatoriedad, análogo al borde del caos en dinámicas no lineales”, lo cual indica que estos tipos de redes tienen un aspecto aleatorio, así como uno estructurado en su interconexión. De acuerdo a León [21], se encontró que una red de sismos para Colombia puede caracterizarse como una red de pequeño mundo, lo cual concuerda con estudios similares en otras regiones del mundo [20].

El uso de características de redes complejas en tareas de machine learning se puede revisar en [22]. En 8, 9 y 10 se muestran casos de estudio en el que se realizan tareas de clasificación automática de datos usando aprendizaje supervisado, semi-supervisado y no supervisado a partir de características estimadas de representaciones de datos como redes complejas. El caso que nosotros proponemos con TagEQ-CN está inmerso en el marco semi-supervisado.

Existen paquetes como SNAP, Gephi y GraphViz que permiten analizar y graficar redes complejas, o librerías de Python como Networkx para estos fines. desarrollado en la universidad de Stanford, o Networkx. Sin embargo, hasta la fecha no se ha encontrado paquetes de software ni frameworks computacionales que construyan redes de sismos.

Este artículo tiene como propósito mostrar una arquitectura modular que pueda usarse para clasificar eventos usando sismogramas de una estación y un catálogo sísmico, aplicando directivas de alto nivel para anotar los eventos, implementando varios algoritmos de extracción de características de redes complejas, detección y clasificación, mostrando cómo se pueden integrar en un sistema automático de clasificación.

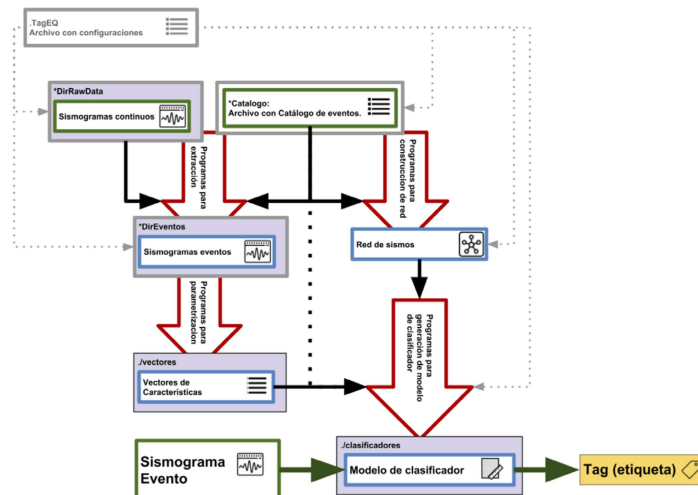
Arquitectura TagEQ-CN

TagEQ-CN recibe un catálogo anotado de eventos sísmicos y los sismogramas de una estación para un tiempo dado. Estos son los datos utilizados para el entrenamiento de un algoritmo de aprendizaje de máquina. A partir del catálogo de sismicidad, se busca en los sismogramas de manera automática las señales que correspondan a cada evento y se parametrizan, creando vectores de características de dichas señales, a partir de la información de una red compleja de sismos. Dicha red se construye de acuerdo con León [21], dividiendo una región en celdas cuadradas, seleccionando como nodos las celdas donde ocurren eventos sísmicos, y enlazando estas celdas siguiendo la secuencia temporal de los eventos. Esta información puede ser utilizada con cualquier algoritmo de aprendizaje que se desee aplicar: máquinas de soporte vectorial (SVM) o bosques de decisión (RF). El proyecto está disponible en <https://gitlab.com/danieleon1/TagEQ>.

Flujo de datos para el sistema TagEQ-CN

Para el funcionamiento del sistema se usa un archivo de configuración que funciona como un mapa para que el sistema sepa dónde están los datos de las señales, los parámetros para los algoritmos de clasificación; los parámetros con los que se asignan las etiquetas, el tipo de clasificación que se desea hacer y las reglas con las que se deben aplicar las etiquetas. En la Figura 1 se presenta el flujo de datos y la arquitectura del sistema.

Figura 1. Flujo de datos y arquitectura del sistema



Fuente: elaboración propia

Las tareas representadas en la Figura 1 son:

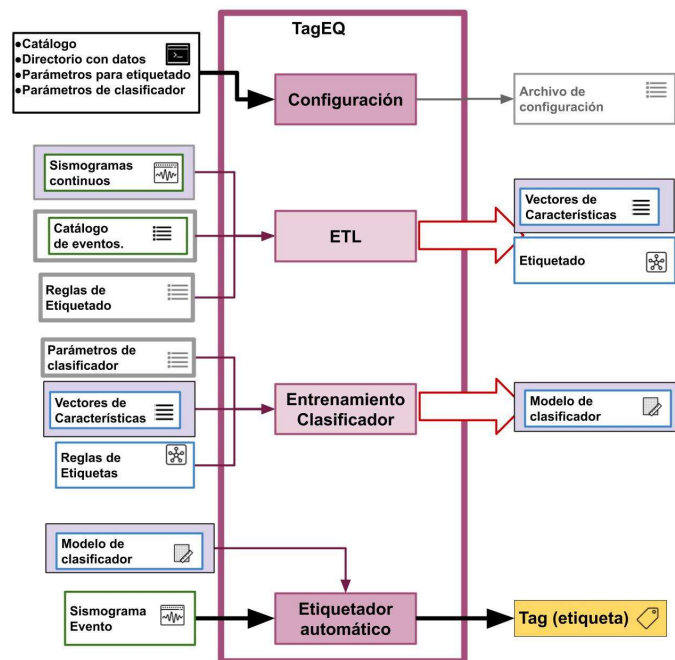
- Ejecución de las tareas de pre-procesamiento de los sismogramas continuos, extracción de las señales de los eventos que hacen parte del catálogo; y definición de las reglas con las que se asignan las etiquetas a los eventos.
- Creación de vectores de características a partir de los sismogramas de los eventos.
- Creación de modelos de clasificadores a partir de los vectores y las reglas definidas para etiquetar los eventos.
- Ya con los modelos construidos, se realiza la lectura del sismograma de un evento para asignarle la etiqueta correspondiente.

Módulos del sistema TagEQ-CN

La Figura 2 muestra los principales módulos del sistema. En ella se presentan las entradas y salidas y la estructura modular del sistema TagEQ-CN. Así se presentan los cuatro módulos principales: ConFiguración, ETL (Extracción, Transformación y Carga), Entrenamiento de clasificador y Etiquetado automático. Cada módulo se ejecuta de manera automática cuando el usuario lo indica, sin embargo, solo los módulos de ConFiguración y Etiquetado automático precisan de interacción con el

usuario para poder entregar su respectiva salida. A continuación, se describe en detalle cada módulo del sistema.

Figura 2. Módulos del sistema TagEQ-CN

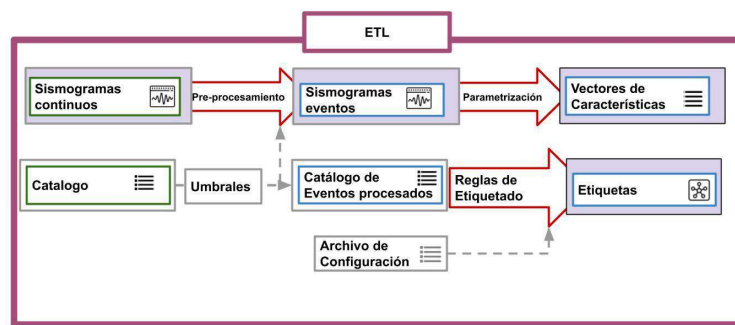


Fuente: Elaboración propia

Módulo de configuración: El módulo de configuración tiene como propósito que el usuario haga explícitas las directivas con las que desea que el sistema realice la clasificación de las señales. Hay dos clases de directivas para el sistema; las directivas generales y las directivas de Machine Learning. En las directivas generales se describe el lugar en que están los sismogramas, cuál es el catálogo con el que se trabaja, dónde se almacenan los datos generados (sismogramas de eventos y catálogos anotados), y qué reglas se usan para anotar los datos. En las directivas de Machine Learning se define el tipo de clasificador y los parámetros con los que se sintonizan tales clasificadores.

Módulo de Extracción, Transformación y Carga (ETL): Realiza todos los procedimientos de pre-procesamiento de datos y parametrización de sismogramas de los eventos. La Figura 3 presenta el esquema general de funcionamiento de este módulo. El módulo ETL toma las entradas (sismogramas continuos y catálogo de sismicidad) y genera los vectores de características que posteriormente se usarán para clasificar eventos. Usando los tiempos de origen de los eventos relacionados en el catálogo de sismicidad, se buscan en los sismogramas continuos las señales que correspondan a los eventos para posteriormente parametrizarlas usando diferentes técnicas, y luego se almacenan estos vectores de características. Para la arquitectura propuesta en el sistema TagEQ, se crearon programas para generar vectores según los siguientes criterios:

Figura 3. Módulo de Extracción, Transformación y Carga (ETL)

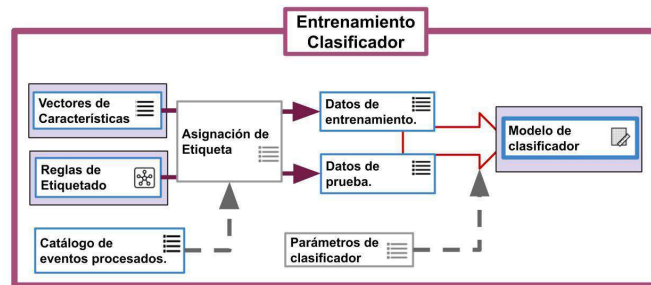


Fuente: elaboración propia

- *Estadísticos:* Se calcula promedio, varianza, asimetría estadística y curtosis en ventanas de 50 segundos para los 3 canales del sismograma.
- *Sonograma:* Calcula el promedio en lapsos de 50 segundos de la energía de 9 bandas de frecuencia, a partir del algoritmo propuesto por [23].
- *Dimensión fractal de la señal:* Calcula la dimensión fractal de la componente vertical para una ventana definida.
- *Variación de la dimensión fractal:* Calcula la varianza de la dimensión fractal calculada en diferentes ventanas.
- *Covarianzas:* Calcula la matriz de covarianzas, rectilinealidad y planaridad de la señal en una definida.

Módulo entrenamiento del clasificador: Para realizar el entrenamiento del clasificador, se toman los vectores de características y se asigna una etiqueta correspondiente. Esta asignación se hace revisando la información consignada en el catálogo de eventos procesados y resolviendo la etiqueta a partir de las reglas que se consignan para la red de sismos, y luego se etiquetan los datos resolviendo la directiva usando el catálogo de sismicidad. El esquema del flujo de información en este módulo se muestra en la Figura 4.

Figura 4. Módulo de entrenamiento del clasificador

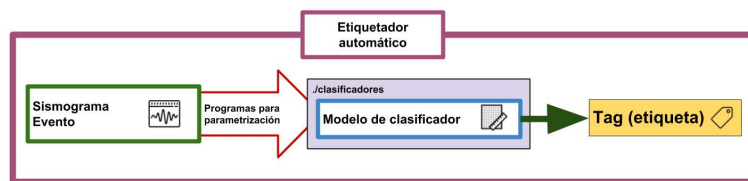


Fuente: elaboración propia

Para crear el modelo de clasificador, se usa una parte del set de datos como entrenamiento y restante como datos de prueba. Para cada caso de algoritmo de clasificación se usan los parámetros introducidos en el archivo de configuración ya sea para generar un solo modelo o para generar muchos modelos y escoger el que mejor desempeño tenga. Para el sistema TagEQ-CN se usaron las librerías LibSVM para las máquinas de soporte vectorial [24]; y para los bosques aleatorios de decisión el programa Ranger [25, 26, 27].

Módulo etiquetador de eventos: el módulo requiere que manualmente se le diga al sistema cuál es el sismograma del evento a clasificar, tras lo cual se parametriza el sismograma, se construye un vector de características, y luego se le ingresa al modelo construido anteriormente. Al ejecutar la predicción del clasificador entrenado se tiene la etiqueta del evento. La Figura 5 muestra el esquema en que funciona el módulo del etiquetador.

Figura 5. Módulo etiquetador de eventos



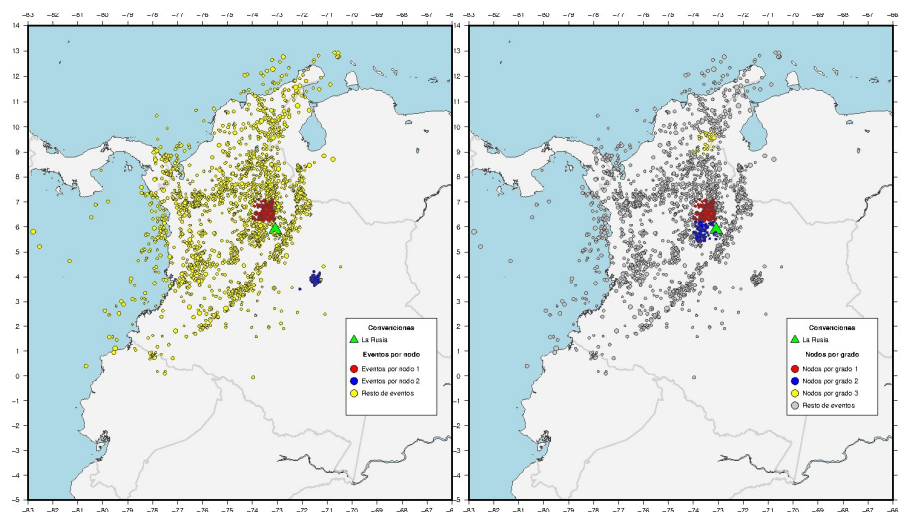
Fuente: elaboración propia

Resultados

El catálogo con el que se trabajó consta de 51520 eventos. De estos eventos, se trabajó solamente los que tenían magnitud $M > 2$, y que tuvieran suficiente energía sísmica en la estación La Rusia; aunque en algunos casos se realizó la prueba con los eventos con magnitud $M > 3$. El conjunto de prueba final consta de 8151 eventos con $M > 2$, y de 502 para $M > 3$, y se probaron algoritmos para realizar clasificación en escenarios según la pertenencia a comunidades de eventos, según el grado del nodo, según la centralidad y según el número de eventos por nodo.

La Figura 6 muestra los eventos etiquetados según la cantidad de eventos por nodo y según el grado del nodo. Para el caso de cantidad de eventos se etiquetaron los eventos que hacían parte de las dos celdas con mayor cantidad y una tercera clase para etiquetar el resto. Para el caso de grado de eventos, se etiquetaron los eventos de los tres nodos con mayor grado, y una cuarta clase para etiquetar el resto de eventos.

Figura 6. Mapas de clasificación de eventos por pertenencia a nodos con mayor cantidad de eventos (izquierda) y por grado (derecha)

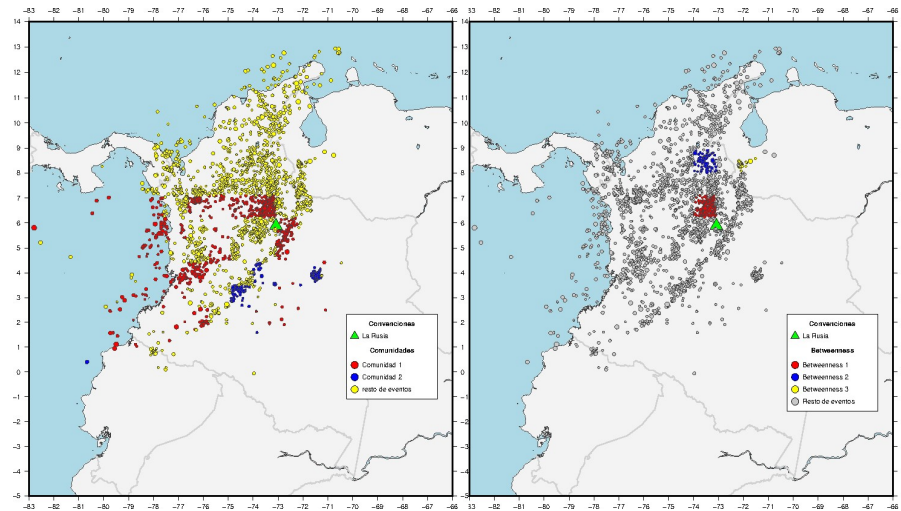


Fuente: elaboración propia

Vale la pena comentar que el etiquetado que se hace muestra en ambos casos la importancia de los eventos que ocurren en el nido de Bucaramanga. Se puede ver también que en el oriente colombiano hay una pequeña región en la que ocurren muchos eventos en un nodo. Y respecto al grado de las celdas, se puede ver que las celdas cerca de Bucaramanga se enlazan mucho con las demás, así como eventos que ocurren más al norte, entre los departamentos de Cesar y Norte de Santander, en la frontera con Venezuela.

La Figura 7 muestra el etiquetado de eventos que pertenecen a las dos comunidades con mayor cantidad de nodos, además una tercera clase para el resto de eventos (izquierda); y etiquetado de eventos que pertenecen a los 3 nodos con mayor índice de centralidad en la red de sismos, además de una cuarta clase para el resto de eventos (derecha).

Figura 7. Mapas de clasificación de eventos por pertenencia a nodos de las mayores comunidades de eventos (izquierda) y a los nodos más centrales de la red de sismos (derecha)

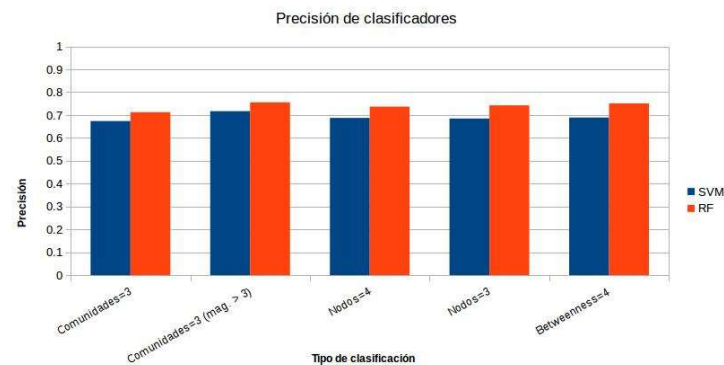


Fuente: elaboración propia

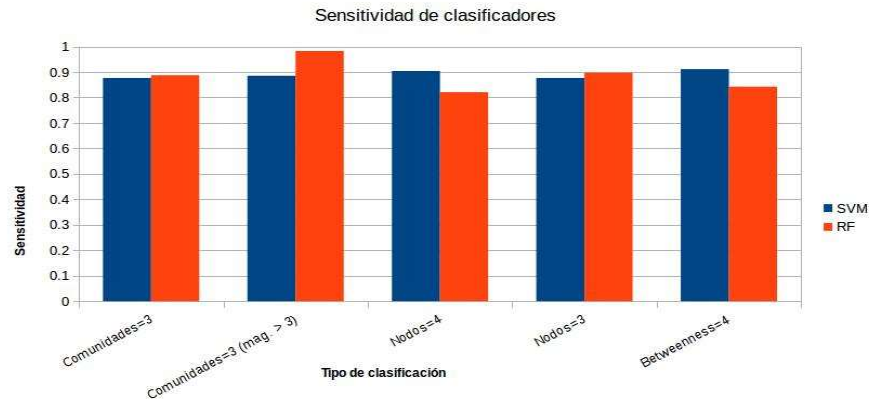
La Figura 7 muestra que el nivel de abstracción necesario para diferenciar eventos de categorías diferentes es elevado. Para el caso de las comunidades (Figura 7 izquierda), se puede ver que los miembros de la misma categoría, por ejemplo, la roja, están localizados en diferentes lugares y muy cerca de miembros de otras categorías; lo cual es un reto para una única estación tratando de diferenciar todos los eventos. Esto mismo sucede para la otra comunidad.

Con los esquemas de anotación mostrados en las Figuras 6 y 7, se entrenaron clasificadores del tipo Máquina de Soporte Vectorial (SVM) y Random Forest (RF); eligiendo aleatoriamente un 80% del conjunto de eventos anotados para entrenamiento y un 20% como conjunto de test. Para las precisiones y sensibilidades mostradas en las Figuras 8 y 9 se usó el mejor clasificador SVM al variar los parámetros, y un clasificador RF con 500 árboles.

Figura 8. Precisión de los clasificadores usando SVM y RF



Fuente: elaboración propia

Figura 9. Sensitividad de los clasificadores usando SVM y RF

Fuente: elaboración propia

De las Figuras 8 y 9 se puede apreciar que el rendimiento es mejor con los RF respecto a las SVM. Esta misma conclusión se encontró en [26], en la que se compararon los rendimientos de SVM, RF y un clasificador Bayesiano. Esto, sumado a que se entrena más rápido y más sencillo, lo ponen con un rendimiento superior a las SVM. De todas maneras, cabe anotar que ambos clasificadores presentan una diferencia pequeña en la precisión y que podría usarse cualquiera de los dos; indicando que la parametrización realizada de las formas de onda tiene información suficiente para crear un buen clasificador a pesar de no tener muchos datos para entrenar (hasta 8151 discriminando con magnitud $M > 2$). De los 51520 eventos, no se revisaron eventos más pequeños porque la energía de estos es muy pequeña y solo podría verse en esta estación si ocurren muy cerca de ella.

Finalmente, los niveles de precisión y sensibilidad superan casi siempre el 70% (excepto para las SVM que se acercan bastante a este nivel). La sensibilidad para este caso multi-clases tiene que ver con cuántos elementos de una categoría pueden ser etiquetados como de otra categoría, por lo que el elevado porcentaje indica que son pocos los elementos que se etiquetan erróneamente.

Conclusiones

El sistema desarrollado se basa en el paradigma supervisado, aunque buena parte del sistema funciona de manera automática en el sentido de que el usuario no escoge directamente cuál es el conjunto de señales que se utilizan para el entrenamiento y prueba del clasificador; ni asigna las etiquetas a los eventos, sino que se asignan a partir de un catálogo de sismos previo, es decir, de una anotación inicial que un analista ha hecho de la sismicidad; y a partir de reglas de alto nivel que defina el usuario. Dado el nivel de abstracción que el clasificador alcanzó, es posible que pueda usarse también con otro tipo de señales, como señales de deslizamientos, explosiones, o movimiento de fluidos. Sin embargo, para poder usar esta técnica se debe crear un catálogo de tiempos de eventos y algún tipo de identificación de la señal.

La clasificación de eventos sísmicos con una sola estación puede ser de gran ayuda cuando no se tiene una red de sensores amplia. Utilizar una única estación para clasificar eventos podría ayudar a un analista a generar una alarma para revisar alguna situación a partir de una única señal adquirida, sin necesidad de

esperar que se tengan las señales de otras estaciones, entregando un reporte preliminar que ponga en “estado de alerta” a un operador de datos sismológicos, con una información preliminar más completa que simplemente decir que una señal superó un nivel de energía.

El sistema desarrollado posee un nivel de modularidad que podría permitir la adición de nuevas características a los vectores desarrollando nuevos programas que extraigan información analizada de un sismograma.

El resultado de las precisiones indica que los parámetros escogidos para representar los eventos en los vectores de características pueden considerarse adecuados; aunque para mejorar el nivel de abstracción con eventos de tantas partes diferentes, con formas de onda diferentes, se podrían necesitar otras características.

Se pudo evidenciar que tanto para los clasificadores Random Forest como para las Support Vector Machine los resultados son buenos. El resultado que se halló también indica que los Random Forest son más adecuados para hacer la clasificación de señales sísmicas.

La clasificación de nuevas señales a partir de las reglas de alto nivel implementadas en el catálogo podría permitir a un analista encontrar rápidamente señales candidatas de algún tipo (eventos profundos/someros, grandes/pequeños, pertenecen/no pertenece) de cierto interés, implementando en conjunto un algoritmo de detección con uno de clasificación.

Como trabajo futuro, dado que este es un prototipo de clasificador, se propone tener un ambiente amigable con el usuario, y definir un marco amplio en el que se puedan definir directivas de alto nivel para asignar etiquetas.

Referencias Bibliográficas

1. J. Akram and D. Eaton, “Adaptive microseismic event detection and automatic time picking.”
2. M. Beyreuther and J. Wassermann, “Continuous earthquake detection and classification using discrete Hidden Markov Models,” *Geophys. J. Int.*, vol. 175, no. 3, pp. 1055–1066, Dec. 2008, doi: <https://doi.org/10.1111/j.1365-246X.2008.03921.x>.
3. P. Bui Quang, P. Gaillard, Y. Cano, and M. Ulzibat, “Detection and classification of seismic events with progressive multi-channel correlation and hidden Markov models,” *Comput. Geosci.*, vol. 83, pp. 110–119, Oct. 2015, DOI: <https://doi.org/10.1016/j.cageo.2015.07.002>.
4. O. Lindenbaum, N. Rabin, Y. Bregman, and A. Averbuch, “Multi-channel fusion for seismic event detection and classification,” Nov. 2016, pp. 1–5, DOI: <https://doi.org/10.1109/ICSEE.2016.7806088>.
5. J. Havskov and L. Ottemöller, *Routine data processing in earthquake seismology: with sample data, exercises and software*. Dordrecht ; New York: Springer, 2010.
6. T. Lay and T. C. Wallace, *Modern global seismology*. San Diego: Academic Press, 1995.
7. D. Leon, H. Ordoñez, and V. Bucheli, “TagEQ: Sistema de Clasificación de Eventos Sísmicos Basado en Inteligencia Artificial,” *RISTI - Rev. Iber. Sist. E Tecnol. Inf.*, vol. e27, pp. 359–370, 2020.
8. B. Sick, M. Guggenmos, and M. Joswig, “Chances and limits of single-station seismic event clustering by unsupervised pattern recognition,” *Geophys. J. Int.*, vol. 201, no. 3, pp. 1801–1813, Apr. 2015, DOI: <https://doi.org/10.1093/gji/ggv126>.

9. A. Köhler, M. Ohrnberger, and F. Scherbaum, "Unsupervised pattern recognition in continuous seismic wavefield records using Self-Organizing Maps: Unsupervised seismic pattern recognition," *Geophys. J. Int.*, vol. 182, no. 3, pp. 1619–1630, Sep. 2010, DOI: <https://doi.org/10.1111/j.1365-246X.2010.04709.x>.
10. A. E. Ruano, G. Madureira, O. Barros, H. R. Khosravani, M. G. Ruano, and P. M. Ferreira, "Seismic detection using support vector machines," *Neurocomputing*, vol. 135, pp. 273–283, Jul. 2014, DOI: <https://doi.org/10.1016/j.neucom.2013.12.020>.
11. S. Gentili and P. Bragato, "A neural-tree-based system for automatic location of earthquakes in Northeastern Italy," *J. Seismol.*, vol. 10, no. 1, pp. 73–89, Jan. 2006, DOI: <https://doi.org/10.1007/s10950-005-9001-z>.
12. J. McGuire, M. S. Boettcher, and T. H. Jordan, "Foreshock sequences and short-term earthquake predictability on East Pacific Rise transform faults," *Nature*, vol. 434, no. 7032, pp. 457–461, Mar. 2005, DOI: <https://doi.org/10.1038/nature03377>.
13. L. N. Doda, V. L. Natyaganov, and I. V. Stepanov, "An empirical scheme of short-term earthquake prediction," *Dokl. Earth Sci.*, vol. 453, no. 2, pp. 1257–1263, Dec. 2013, DOI: <https://doi.org/10.1134/S1028334X1312009X>.
14. P. E. Nikolopoulos D, "Radon-222: A Potential Short-Term Earthquake Precursor," *J. Earth Sci. Clim. Change*, vol. 06, no. 06, 2015, DOI: <https://doi.org/10.4172/2157-7617.1000282>.
15. A. Donnellan *et al.*, "Potential for a large earthquake near Los Angeles inferred from the 2014 La Habra earthquake: POTENTIAL FOR A LARGE EARTHQUAKE NEAR LA," *Earth Space Sci.*, vol. 2, no. 9, pp. 378–385, Sep. 2015, DOI: <https://doi.org/10.1002/2015EA000113>.
16. S. Abe and N. Suzuki, "Law for the distance between successive earthquakes: A NEW LAW FOR DISTANCE BETWEEN EARTHQUAKES," *J. Geophys. Res. Solid Earth*, vol. 108, no. B2, Feb. 2003, DOI: <https://doi.org/10.1029/2002JB002220>.
17. M. Baiesi and M. Paczuski, "Scale-free networks of earthquakes and aftershocks," *Phys. Rev. E*, vol. 69, no. 6, p. 066106, 2004.
18. D. Pastén, F. Torres, B. Toledo, V. Muñoz, J. Rogan, and J. A. Valdivia, "Time-Based Network Analysis Before and After the 8.3 Illapel Earthquake 2015 Chile," *Pure Appl. Geophys.*, vol. 173, no. 7, pp. 2267–2275, Jul. 2016, DOI: <https://doi.org/10.1007/s00024-016-1335-7>.
19. D. Chorozoglou, D. Kugiumtzis, and E. Papadimitriou, "Application of complex network theory to the recent foreshock sequences of Methoni (2008) and Kefalonia (2014) in Greece," *Acta Geophys.*, vol. 65, no. 3, pp. 543–553, Jun. 2017, DOI: <https://doi.org/10.1007/s11600-017-0039-4>.
20. S. Abe and N. Suzuki, "Earthquake networks, Complex," in *Extreme Environmental Events*, R. A. Meyers, Ed. New York, NY: Springer New York, 2011, pp. 312–319.
21. D. A. León, J. A. Valdivia, and V. A. Bucheli, "Modeling of Colombian Seismicity as Smallworld Networks," *Seismol. Res. Lett.*, vol. 89, no. 5, pp. 1807–1816, Sep. 2018, DOI: <https://doi.org/10.1785/0220180076>.
22. T. Christiano Silva and L. Zhao, *Machine Learning in Complex Networks*. Cham: Springer International Publishing, 2016.
23. M. Joswig, "Knowledge-based seismogram processing by mental images," *IEEE Trans. Syst. Man Cybern.*, vol. 24, no. 3, pp. 429–439, Mar. 1994, DOI: <https://doi.org/10.1109/21.278992>.
24. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011, DOI: <https://doi.org/10.1145/1961189.1961199>.

25. M. N. Wright and A. Ziegler, "textbfranger : A Fast Implementation of Random Forests for High Dimensional Data in textitC++ and textitR," *J. Stat. Softw.*, vol. 77, no. 1, 2017, DOI: <https://doi.org/10.18637/jss.v077.i01>.
26. D. Henao - león, A. Camilo Báez - Alarcón, y J. Bethsaid Pedroza - Rojas, "Metodología para determinar la viabilidad de generación de energía eléctrica por medio del recurso eólico", *Revista Investigación e Innovación en Ingenierías*, vol. 6, n°. 2, 2018. DOI: <https://doi.org/10.17081/invinno.6.2.3108>
27. R. Avellaneda, S. Cabrera, P. A. Martínez, y C. G. Donoso Albarracín, "Apoyo tecnológico para la fidelización y captación de nuevos clientes por medio de una aplicación móvil", *Investigación e Innovación en Ingenierías*, vol. 5, n.º 1, pp. 92-101, 2017. DOI: <https://doi.org/10.17081/invinno.5.1.2618>
28. L. Dong, X. Li, and G. Xie, "Nonlinear Methodologies for Identifying Seismic Event and Nuclear Explosion Using Random Forest, Support Vector Machine, and Naive Bayes Classification," *Abstr. Appl. Anal.*, vol. 2014, pp. 1-8, 2014, DOI: <https://doi.org/10.1155/2014/459137>.

Modelado de la sismicidad colombiana como una red compleja espacio-secuencial

Modeling Colombian Seismicity as a Complex Spatial-Sequential Network

Modelagem da sismicidade colombiana como uma complexa rede espaço-sequencial

Daniel León-Vargas¹

Víctor-Andrés Bucheli-Guerrero²

Hugo-Armando Ordoñez-Erazo³

Recibido: noviembre de 2022

Aceptado: julio de 2023

Para citar este artículo: León-Vargas, D., Bucheli-Guerrero, V. A., y Ordoñez-Erazo, H. A., (2023). Modelado de la sismicidad colombiana como una red compleja espacio-secuencial. *Revista Científica*, 48(3), 67-81. <https://doi.org/10.14483/23448350.20963>

Resumen

Este artículo busca describir la sismicidad de Colombia utilizando redes complejas, en las cuales los nodos representan redes cúbicas (con latitud, longitud y profundidad) donde ocurren los eventos sísmicos y los enlaces siguen la secuencia temporal de estos eventos. Si bien se han reportado trabajos similares en la literatura, este trabajo incorpora la profundidad para una mejor comprensión de este fenómeno. Esta investigación considera 5797 eventos del catálogo de terremotos del Servicio Geológico de los Estados Unidos (USGS) correspondientes a Colombia y con una magnitud superior a un determinado umbral, los cuales ocurrieron entre enero 1 de 1975 y enero 18 de 2021. Se describe la estructura de la red mediante una comparación entre redes aleatorias y de pequeño mundo. Así, esta estructura brinda información sobre los mecanismos de liberación de energía y las fuentes que recurrentemente producen eventos sísmicos en Colombia. Los resultados muestran que estas redes presentan características de pequeño mundo, independientemente del tamaño de celda o la granularidad utilizados para construirlas. Este hallazgo es consistente con los resultados reportados para la misma región en redes espaciotemporales bidimensionales. Como una red de pequeño mundo, el esfuerzo y la energía de la región se liberan de acuerdo con una estructura caracterizada por la presencia de hubs y sus relaciones dentro de la red en general. Mediante el aprovechamiento de la información obtenida en este estudio, se pueden entrenar modelos de aprendizaje automático que superen los modelos actuales de base para el pronóstico.

Palabras clave: modelado de la sismicidad; profundidad de los eventos sísmicos; redes complejas; redes de sismicidad.

1. Universidad del Valle (Cali-Valle del Cauca, Colombia). daniel.leon@correounivalle.edu.co.
2. Ph. D. Universidad del Valle (Cali-Valle del Cauca, Colombia). victor.bucheli@correounivalle.edu.co.
3. Ph. D. Universidad del Cauca (Popayán-Cauca, Colombia). hugoordonez@unicauca.edu.co.

Abstract

This article seeks to describe the seismicity of Colombia using complex networks, in which the nodes represent cubic cells (with latitude, longitude, and depth) where seismic events occur, and the links follow the temporal sequence of these events. While similar works have been reported in the literature, this study incorporates depth to gain a more detailed understanding of this phenomenon. This research considers 5797 events from the earthquake catalog of the United States Geological Survey (USGS), corresponding to Colombia and with a magnitude greater than a certain threshold, which occurred between January 1, 1975, and January 18, 2021. The network structure is described by comparing random and small-world networks. Thus, said structure provides information about the energy release mechanisms and the sources that recurrently produce seismic events in Colombia. The results show that these networks exhibit small-world properties, regardless of the cell size or granularity used to construct them. This finding is consistent with the results reported for the same region in two-dimensional spatiotemporal networks. As a small-world network, the effort and energy in the region are released according to a structure represented by the presence of hubs and their relationships within the overall network. By leveraging the information obtained in this study, it is possible to train machine learning models that outperform the current baseline forecasting models.

Keywords: depth of seismic events; complex networks; seismicity modeling; seismicity networks.

Resumo

Este artigo procura descrever a sismicidade da Colômbia usando redes complexas, nas quais os nós representam redes cúbicas (com latitude, longitude e profundidade) onde ocorrem eventos sísmicos e os links seguem a sequência temporal desses eventos. Embora trabalhos semelhantes tenham sido relatados na literatura, este trabalho incorpora profundidade para uma melhor compreensão desse fenômeno. Esta pesquisa considera 5.797 eventos do catálogo de terremotos do Serviço Geológico dos Estados Unidos (USGS) correspondentes à Colômbia e com magnitude superior a um determinado limite, ocorridos entre 1º de janeiro de 1975 e 18 de janeiro de 2021. A estrutura é descrita por meio de uma comparação entre redes random e small world. Assim, esta estrutura fornece informações sobre os mecanismos de liberação de energia e as fontes que produzem eventos sísmicos recorrentemente na Colômbia. Os resultados mostram que essas redes apresentam características de mundo pequeno, independentemente do tamanho da célula ou da granularidade utilizada para construí-las. Esse achado é consistente com os resultados relatados para a mesma região em redes espaço-temporais bidimensionais. Como uma pequena rede mundial, o esforço e a energia da região são liberados de acordo com uma estrutura caracterizada pela presença de hubs e suas relações dentro da rede global. Aproveitando as informações obtidas neste estudo, os modelos de aprendizado de máquina podem ser treinados para superar os modelos de linha de base atuais para previsão.

Palavras-chaves: modelagem da sismicidade; profundidade dos eventos sísmicos; redes complexas; redes de sismicidade.

INTRODUCCIÓN

Buscar patrones en el marco espacial y temporal de los eventos sísmicos ha ayudado en la identificación de los bordes de las placas tectónicas, así como en la caracterización de sus movimientos a través del manto terrestre. Esta información ha permitido la construcción de herramientas para evaluar el riesgo sísmico ([Lay & Wallace, 1995](#)). La predicción de la ubicación y el tiempo de origen de un terremoto es una pregunta abierta de investigación, y diferentes trabajos se han enfocado en identificar precursores de eventos de gran magnitud, tales como eventos sísmicos precursores ([McGuire et al., 2005](#)), variaciones magnéticas ([Doda](#)

[et al., 2013](#)) o emisiones de gas ([Nikolopoulos, 2015](#)), entre otros. Los trabajos relacionados muestran que no hay resultados de pronóstico concluyentes, y se sugiere que la posibilidad de predecir de manera confiable un terremoto se limita a un marco de tiempo de varios años ([Donnellan et al., 2015](#)).

Para la búsqueda de patrones temporales en sismología, la aplicación de leyes empíricas, como es el caso de las leyes de Gutenberg-Richter y Omori, han permitido comprender algunas de las complejidades de los fenómenos sismológicos. La ley de Gutenberg-Richter establece que la frecuencia de los terremotos sigue una distribución de ley de potencia con respecto a la energía liberada. Así, para un terremoto de magnitud M se espera que ocurran cerca de 10 veces más terremotos con una magnitud de $M-1$ ([Gutenberg & Richter, 1944](#)). Por su parte, la ley de Omori establece que la tasa de réplicas después de un terremoto importante disminuye de manera algebraica con respecto al tiempo transcurrido desde la ocurrencia del evento en cuestión ([Utsu et al., 1995](#)).

El trabajo de [Abe y Suzuki \(2003\)](#) muestra que la sismicidad se puede representar como una intrincada red espacial de eventos. Esto, para el caso del sur de California. Los autores afirman que la sismicidad sigue una geometría fractal comúnmente observada en los sistemas complejos que se encuentran en estado de desequilibrio. Por otra parte, se ha observado que la distribución de los tiempos entre terremotos sigue una distribución de ley de potencia, lo cual evidencia la naturaleza intrincada de la actividad sísmica ([Abe & Suzuki, 2002](#)). [León et al. \(2018\)](#) observaron que la red de eventos sísmicos en Colombia sigue la estructura de red de pequeño mundo. Sin embargo, los resultados reportados son para redes espaciotemporales bidimensionales, en las cuales no se incluyó la profundidad. De igual manera, la información de la red de eventos sísmicos en Colombia ha sido utilizada para construir un sistema que permite la anotación automática de sismos, así como reconocer automáticamente eventos de otras fuentes diferentes a las tectónicas ([León et al., 2020](#)).

En este trabajo se construyen redes de sismos en tres dimensiones, considerando la latitud, longitud y profundidad de los eventos, así como su secuencia temporal. Se compara la estructura de las redes aleatorias y las redes de pequeño mundo. Los resultados revelan que las redes son del segundo tipo, independientemente del tamaño de celda o la granularidad utilizada para construirlas. Así, la red de pequeño mundo muestra que el esfuerzo y la energía en la región se liberan de acuerdo con una estructura, lo cual se ve reflejado en la presencia de *hubs* y en las relaciones que tienen estos con el resto de la red en general. En este sentido, este documento presenta cómo la estructura de redes da información acerca de los mecanismos de liberación de energía y las fuentes que recurrentemente producen eventos sísmicos en Colombia.

REDES COMPLEJAS Y SISMICIDAD EN COLOMBIA

Una secuencia de terremotos es una serie de eventos sísmicos que ocurren en el espacio y el tiempo, los cuales pueden ser representados como una red compleja. Según [Gao et al. \(2016\)](#), la literatura reporta diferentes métodos para construir redes complejas a partir de series de tiempo en campos como la dinámica climática, las funciones cerebrales y los sistemas de tráfico, entre otros. El catálogo de eventos sísmicos para una región en un determinado periodo de tiempo sigue una secuencia temporal. Por lo tanto, un catálogo de sismicidad se puede estudiar como una serie de tiempo de eventos y, por tanto, como una red compleja. Algunos trabajos relacionados con el estudio de catálogos de eventos sísmicos como redes complejas son [León et al. \(2018\)](#), [Pastén et al. \(2018\)](#) y [Chorozoglou et al. \(2017\)](#).

Se han desarrollado diferentes formas de construir redes de sismos, *i.e.*, celdas como nodos y eventos como nodos, grafos de visibilidad, entre otros ([León et al., 2022](#)). De acuerdo con [Abe y Suzuki \(2011\)](#), las redes de sismos se construyen dividiendo una región en celdas, y se selecciona una celda como nodo cuando ocurre un evento sísmico en ella, los nodos se relacionan entre sí de acuerdo con la secuencia temporal que ocurre entre los eventos. En [León et al. \(2018\)](#) se crearon redes de sismos de este tipo, pero sin tener en cuenta la profundidad, mapeando los eventos sísmicos de la región de Colombia en una cuadrícula 2D. Para generar dicha cuadrícula, se definió el tamaño de la celda, y posteriormente se conectaron las celdas de acuerdo con la secuencia de los eventos sísmicos. De forma similar, se puede incluir la profundidad del evento sísmico y construir una red 3D, generando cubos que tienen en cuenta el lugar del evento sísmico y su profundidad. Así, se define el tamaño del cubo y se relacionan los cubos entre sí de acuerdo con la secuencia de eventos. Este trabajo modela la sismicidad en Colombia de acuerdo con una red 3D.

En este trabajo se adiciona la dimensión de profundidad para obtener una comprensión más detallada del fenómeno de la sismicidad en Colombia. Esta comprensión es crucial para desarrollar modelos que permitan describir de manera precisa la complejidad de la sismicidad en la región. Para estudiar la complejidad y la estructura de las redes complejas, se describen numerosos parámetros y medidas relacionadas. Muchas de las características inherentes a los sistemas complejos se observan en las redes, tales como la auto-similitud, la fractalidad o la escalabilidad ([Boccaletti et al., 2006](#)). Por otra parte, existen diferentes tipos de redes complejas, *i.e.*, de pequeño mundo, libres de escala y aleatorias (redes Erdős-Rényi). Las redes aleatorias fueron inicialmente estudiadas por [Erdős y Rényi \(1960\)](#) y pueden ser reproducidas por el modelo que lleva su nombre ([Erdős & Rényi, 1959](#)). Así, se pueden describir las redes en las cuales dos nodos se relacionan de acuerdo con una distribución de probabilidad binomial. Esto quiere decir que, en este tipo de redes, la decisión de generar un vínculo entre dos nodos corresponde a lanzar una moneda, y todos los nodos tienen la misma probabilidad de estar conectados.

Las redes de pequeño mundo estudiadas por [Watts y Strogatz \(2011\)](#) tienen una estructura más compleja que una red aleatoria Erdős-Rényi. Por lo general, esta estructura se observa mediante dos métricas del grafo: el coeficiente de agrupamiento global (GCC) y la longitud de camino promedio (APL). Es decir, una red que tiene una estructura de pequeño mundo muestra nodos *hub*, grupos dentro de la estructura de relaciones y una conectividad alta, entre otras características que no se encuentran en las redes Erdős-Rényi. En este trabajo se estudia experimentalmente si la red de sismicidad en Colombia (incluyendo la profundidad, *i.e.*, 3D) sigue un patrón de red Erdős-Rényi o un patrón de pequeño mundo (*small-world*).

Este estudio utiliza una métrica para analizar la tipología de la red conocida como *small-worldness* (*S*), la cual expresa la no aleatoriedad de una red. Esta métrica indica que el GCC de una red de pequeño mundo es mayor que el de una red aleatoria y que su APL es mayor o igual al de una red aleatoria. Para calcular la *small-worldness*, es necesario obtener el GCC y APL de una red Erdős-Rényi con características de tamaño y conectividad similares a la red observada. En otras palabras, *S* se calcula comparando el GCC y APL de la red observada y los de una red aleatoria Erdős-Rényi ([Humphries & Gurney, 2008](#)). Por lo tanto, para caracterizar una red como de pequeño mundo, su *small-worldness* debe ser mayor que 1.

Para analizar la *small-worldness* de las redes sísmicas, se deben considerar varios factores, incluyendo el tamaño de la red, el número de iteraciones utilizadas en el análisis, el umbral de magnitud aplicado y el tipo de catálogo de terremotos utilizado. Además, considerar eventos profundos puede generar patrones de red distintos, especialmente en regiones con diferencias significativas en la actividad sísmica a diferentes profundidades. Otros trabajos similares a los de [Abe y Suzuki \(2009, 2012\)](#) pueden ayudar a comprender la complejidad del fenómeno sísmico e integrarlo con algoritmos de aprendizaje automático.

METODOLOGÍA CONSTRUCCIÓN DEL CATÁLOGO DE DATOS

El catálogo de eventos usado para este estudio se descargó del USGS-NEIC (United States Geological Survey – National Earthquake Information Center) para la región cercana a Colombia. Se escogió este catálogo y no el del Servicio Geológico Colombiano porque es más completo para eventos que ocurrieron antes de 1990. Se configuraron latitudes mínimas y máximas de -5 y 13° respectivamente, así como longitudes mínimas y máximas de -82 y -66° . El período escogido para los eventos del catálogo comprende entre el 1 de enero de 1975 y el 18 de enero de 2021, para un total de 6950 eventos. La magnitud mínima para analizar los datos se calculó a partir de la distribución Gutenberg-Richter del catálogo, como se muestra en la [Figura 1](#), donde el eje X es la magnitud y el eje Y es el logaritmo del número de eventos. La línea oscura es entonces el logaritmo del número de eventos con magnitudes mayores o iguales a la magnitud referida en el eje X, y la línea roja es la parte lineal de esta gráfica, tomando solo los datos con magnitudes mayores a la referida por la línea azul, que en este caso es 4.0. La línea se ajusta a una magnitud de hasta 7.5 porque no hay muchos eventos con magnitudes mayores a este valor. Con esta distribución, se puede elegir una magnitud de 4.0 para la completitud del catálogo, es decir, hay confianza de que todos los eventos con magnitudes mayores a 4.0 para el período seleccionado hacen parte del catálogo. Esta magnitud constituye el límite para elegir los eventos con los que se construirán las redes de sismos, lo cual resulta en un catálogo final de 5797 eventos. Asimismo, la [Figura 1](#) muestra los eventos sísmicos y su ubicación en el mapa para las longitudes y latitudes descritas en las líneas anteriores. El tamaño de los eventos se ilustra según su magnitud, y su color corresponde a la profundidad de los eventos.

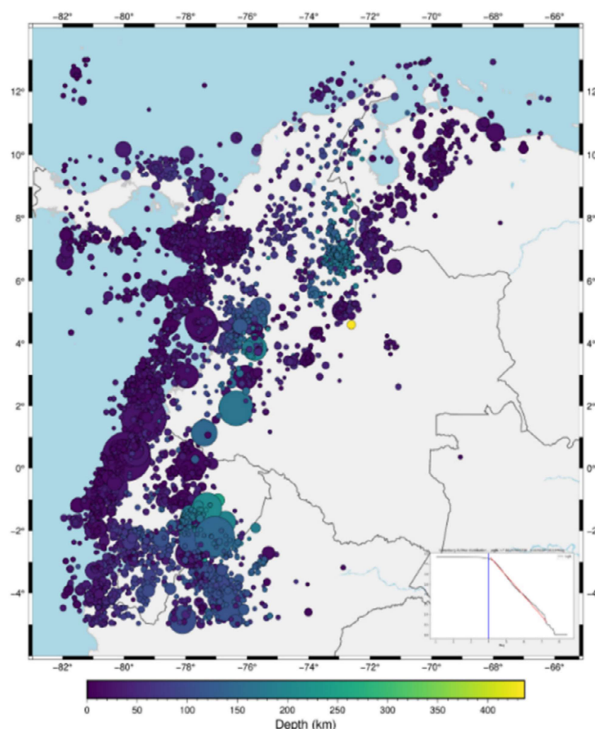


Figura 1. Distribución Gutenberg-Richter del catálogo estudiado

Así, la distribución Gutenberg-Richter de todos los eventos en el catálogo de la región, construido a partir del número total de eventos N con magnitud igual o superior a cada valor M en el eje X , muestra un conjunto completo de magnitudes mayores a 4.0. Como se muestra en la [Figura 1](#), a partir del ajuste de $N * \exp(a - bM)$, se obtienen los parámetros $a = 7.88$ y $b = -0.99$. Así, hay 5797 terremotos con magnitudes mayores a 4.0 entre el 1 de enero de 1975 y el 18 de enero de 2021. Hay otros catálogos de agencias colombianas (Vargas *et al.*, 2018), pero ninguno de ellos se remonta tan atrás en el tiempo como el catálogo del USGS, y, para nuestro análisis, es fundamental contar con el mayor lapso de tiempo posible, en aras de una buena comprensión de la sismicidad de la región.

La ubicación de estos eventos corresponde al límite de cinco placas tectónicas, de acuerdo con [Sornette \(2003\)](#). Esta compleja región de estudio de múltiples placas se ve afectada por la acción combinada de las placas sudamericanas (este), caribeñas (norte), bloques andinos del norte (centro), Cocos (Panamá y Centroamérica) y Nazca (oeste) ([Sornette, 2003](#)). La presencia e influencia de todas estas placas indica una zona con una actividad sísmica alta y compleja que vale la pena analizar con técnicas computacionales.

Teniendo en cuenta lo anterior, la separación temporal entre eventos es relativamente pequeña. De los 5797 eventos seleccionados, la mitad ocurrió con una separación temporal de menos de 41 horas entre eventos consecutivos, y el 80 % con una separación temporal de menos de 118 horas (casi cinco días). Por lo tanto, es altamente probable que dos eventos consecutivos con magnitud mayor que 4.0 ocurran con una separación temporal menor de cinco días.

Se utilizó un método similar a la técnica descrita por [Abe y Suzuki \(2004a\)](#) para crear la red. Los eventos sísmicos de la [Figura 1](#) se mapearon en una cuadrícula 3D para que todas las celdas tuvieran el mismo tamaño, y cada celda se identificó como un nodo de la red si reportaba un evento sísmico. Para este trabajo, se asumió una conversión de grados-km de 111 km/grado, por lo que los eventos sísmicos de nuestros conjuntos de datos se mapearon en una cuadrícula de aproximadamente 2000 km de largo y 1900 km de ancho.

Como se muestra en la [Figura 2](#), una celda se seleccionó como nodo si el epicentro de un evento sísmico estaba dentro de ella. De esta manera, se construyó una celda cúbica con un tamaño dado para alto y largo, y se tuvieron en cuenta las profundidades de los eventos en la celda ([Figura 2a](#)). Esta Figura muestra una escala de profundidad de azul claro (baja profundidad) a negro (alta profundidad). En este trabajo se analizaron redes con diferentes tamaños de celda cúbica, *i.e.*, de 10 a 100 km. Se etiquetó cada celda según su longitud, latitud, tamaño y profundidad si en ella ocurrieron eventos sísmicos ([Figura 2b](#)). Para cada tamaño de celda, se creó una red (una red para 10 kms, otra para 20kms y así sucesivamente), seleccionando los nodos en los que ocurrieron los eventos y estableciendo enlaces entre nodos siguiendo una secuencia temporal, es decir, enlazando eventos sucesivos ([Figura 2c](#)). De esta manera, la red representa una secuencia de eventos, junto con su posición y profundidad. Para analizar dicha red, se ignoraron los bucles (un nodo que se enlaza consigo mismo) y las conexiones múltiples entre cualquier par de nodos.

El proceso que sigue el desarrollo del estudio sigue los pasos descritos en el esquema de la [Figura 3](#), los cuales se construyen a partir del catálogo de sismos. Se construyeron las redes de sismos para diferentes tamaños de celdas, así como matrices de adyacencia, que posteriormente serían la entrada del programa de análisis de redes complejas. Estas matrices se analizaron tomando como referencia el estudio de redes complejas y redes de sismicidad en Colombia de [León *et al.* \(2018\)](#). Las métricas analizadas fueron el número de nodos, el número de enlaces, el grado medio, el APL y el coeficiente de agrupamiento para cada tamaño de celda propuesto. También se obtuvo la representación gráfica de las redes. Los *scripts* se desarrollaron en Wolfram Mathematica. La [Tabla 1](#) presenta los resultados de cada red para cada tamaño de celda y sus respectivas métricas.

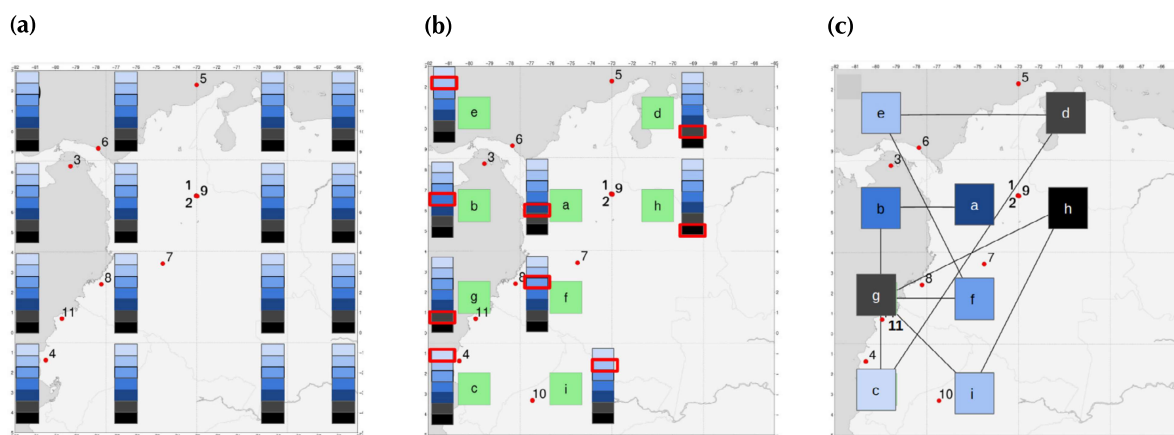


Figura 2. Diagrama de la red de sismos generada con base en su secuencia espaciotemporal e incluyendo la profundidad

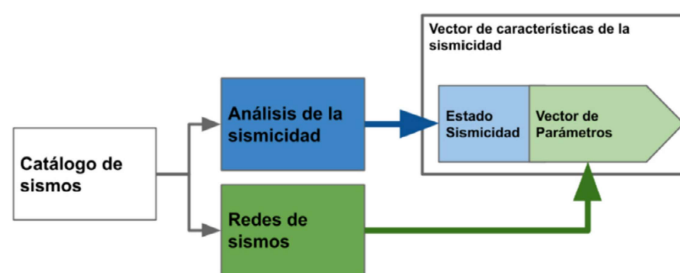


Figura 3. Esquema de construcción de redes de sismicidad

Las herramientas y *software* utilizados fueron Python ver 3.11 ([van Rossum & Python Development Team, 2013](#)) y Wolfram Mathematica v. 13.1 ([Wolfram, 2022](#)). Python es un lenguaje de programación ampliamente usado, a través del cual se procesaron los datos del catálogo de eventos sísmicos y se construyeron las matrices de adyacencia para cada uno de los tamaños de celda. Estos archivos fueron importados a Wolfram Mathematica para el análisis de redes complejas. Las funciones utilizadas fueron *VertexCount*, *EdgeCount*, *DegreeCentrality*, *BetweennessCentrality*, *MeanClusteringCoefficient* y *GraphDistanceMatrix*. Adicionalmente, se aplicaron 100 simulaciones para dos tipos de redes aleatorias (modelo Erdős-Rényi) y redes de pequeño mundo (modelo Watts y Strogatz): *RandomGraph*, *WattsStrogatzGraphDistribution* y *Ba rabasiAlbertGraphDistribution*.

RESULTADOS

Este artículo presenta un análisis de la red de sismicidad en la región de Colombia y Ecuador, la cual utiliza diferentes tamaños de celdas (10 y 100 km). Se observó que las redes obtenidas son de tipo pequeño mundo. La distribución del grado se asemeja a una distribución de Poisson, la cual se utiliza en el análisis de redes de pequeño mundo porque representa una distribución aleatoria de los nodos y conexiones. En

una red de pequeño mundo, los nodos se conectan no solo con sus vecinos directos, sino también con otros nodos más lejanos. Sin embargo, las conexiones con nodos lejanos son eventos menos frecuentes. Dos características de las redes pequeño mundo son estudiadas regularmente en la literatura ([Watts & Strogatz, 2011](#)): el coeficiente de agrupamiento y la longitud media de los caminos más cortos. En este tipo de redes, hay un alto coeficiente de agrupamiento y una baja longitud media de los caminos más cortos, es decir, la red tiene *hubs* y grupos, tiene una estructura de centro-periferia y es eficiente, pues se puede alcanzar cualquier nodo en pocos pasos. Los resultados de estas métricas se presentan en la [Tabla 1](#).

Tabla 1. Características de las redes estudiadas

Tamaño de celda (km)	Número de nodos	Total de enlaces	Media del grado	Tamaño del camino promedio	Media del coeficiente de agrupamiento	Small-worldness (S)	Small-worldness de redes 2D ([11])	S>1
10	3837	5296	2.76	5.703	0.037	73.79	13.79	Sí
20	2739	4973	3.63	3.990	0.089	103.65	9.16	Sí
30	1955	4505	4.61	3.223	0.167	109.11	6.98	Sí
40	1364	4381	6.42	3.040	0.210	56.90	6.40	Sí
50	997	4004	8.03	2.770	0.269	39.89	5.48	Sí
60	746	3439	9.22	2.554	0.316	29.79	4.57	Sí
70	558	3094	11.09	2.388	0.357	19.76	3.95	Sí
80	445	3055	13.73	2.365	0.377	12.02	3.50	Sí
90	364	2828	15.54	2.291	0.389	8.54	3.19	Sí
100	303	2707	17.87	2.218	0.452	6.84	2.80	Sí

De acuerdo con [León et al. \(2018\)](#), la estructura de la red sísmica en la región de Colombia y Ecuador es una red en estado de no equilibrio, donde la información y el estrés pueden propagarse de manera no aleatoria en la red. En este trabajo se confirma que la red no es aleatoria y que las características de la distribución del grado, el coeficiente de agrupamiento y la longitud media de los caminos más cortos lo confirman. En la Tabla se han consignado los resultados del cálculo de la *small-worldness* para redes de sismos 2D, de manera que se puedan comparar con los resultados obtenidos al considerar la profundidad de los eventos.

En [León et al. \(2018\)](#), las redes se construyen en dos dimensiones y la profundidad no se tiene en cuenta. Un aporte de este trabajo es el desarrollo de las redes incluyendo la profundidad. Esto muestra que no solo las zonas siguen una estructura de red no aleatoria, sino también las profundidades en las que ocurren los eventos sísmicos. Así, esta herramienta permite suponer que la estructura de relaciones que subyace a los eventos sísmicos está continuamente liberando estrés por medio de zonas y profundidades que conforman agrupaciones, lo cual permite liberar energía. De igual forma, dicha energía fluye por esta red hacia la periferia de la red en pocos pasos. Consideramos que esta información podría tener implicaciones prácticas para la gestión del riesgo sísmico en Colombia.

La [Figura 4](#) muestra un análisis de la sismicidad desde la perspectiva de la red, con tamaños de celda de 10 km ([Figura 4a](#)) y 100 km ([Figura 4b](#)). Las [Figuras 4](#) y [5](#) muestran estructuras de redes de una complejidad mayor a las redes aleatorias, y se observan agrupaciones de nodos que parecen comunidades

de sismos localizadas en zonas como el nido sísmico de Bucaramanga o algunas en Ecuador y cerca del océano Pacífico, así como también algunas a lo largo de la zona de subducción entre la placa de Nazca y la placa Sudamericana (en la costa del Pacífico colombiano) y cerca de Venezuela. El nido sísmico de Bucaramanga ha sido estudiado ampliamente ([Schneider et al., 1987](#); [Frohlich et al., 1995](#)) y sigue siendo una región de interés ([Poli et al., 2016](#)).

Lo mismo aplica para la actividad sísmica a lo largo de la zona de subducción. Por lo tanto, tener diferentes escalas de observación del fenómeno permite detectar regiones que sísmicamente activas a partir de un catálogo dado sin la necesidad de un experto. Este resultado no es necesariamente evidente al observar los datos en un mapa, como se muestra en la [Figura 5](#). En este sentido, la construcción de las redes sísmicas para diferentes tamaños de celdas permite detectar automáticamente regiones con actividad sísmica importante, o eventos específicos en zonas y profundidades que desencadenan otros eventos.

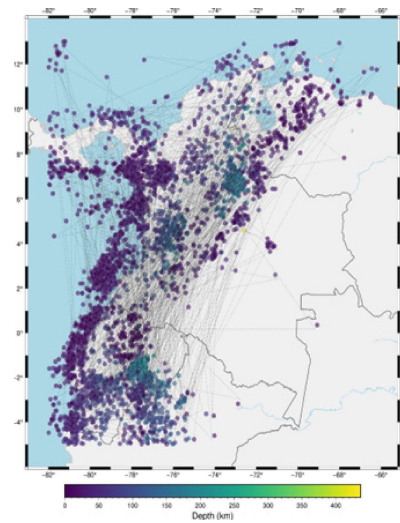


Figura 4. *Nodos de sismos*

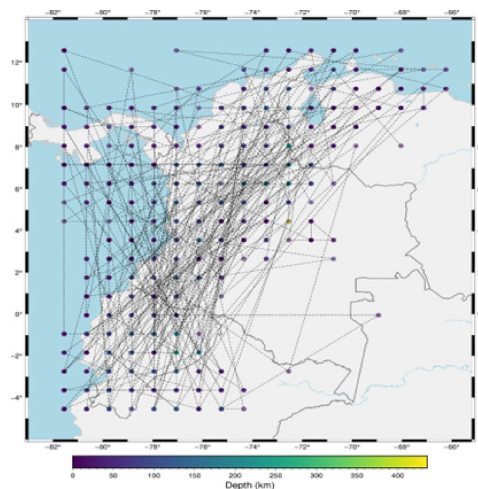


Figura 5. *Celdas de sismos*

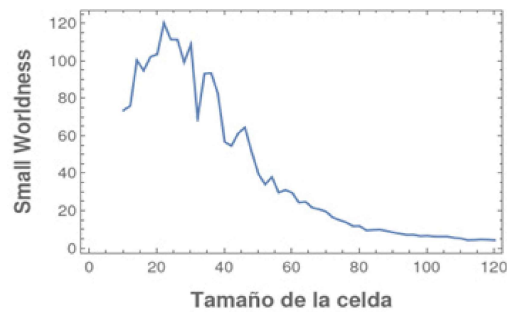


Figura 6. Small-worldness para cada tamaño de celda

La [Figura 6](#) muestra que, para todas las redes observadas entre 10 y 100 km, los valores de *small-worldness* son mayores que 1. Se observa que, para las celdas entre 15 y 30 km, hay un pico en el valor de *small-worldness*. Estos resultados sugieren que las redes observadas se parecen más a una red de pequeño mundo que a una red aleatoria, y permiten representar las secuencias de los eventos sísmicos de esta manera, con una estructura que se aleja del equilibrio.

Para evaluar la robustez de las propiedades de la red encontradas para diferentes tamaños de celda, se realizaron comparaciones con 100 simulaciones para dos tipos de redes aleatorias (modelo Erdős-Rényi) y redes de pequeño mundo (modelo Watts y Strogatz). Para ello, se utilizó el número de vértices de las redes observadas y el número de los enlaces obtenidos para cada tamaño de celda ([Tabla 1](#)). En cada caso, se simulaban redes sísmicas con características similares a las observadas, incluyendo el número de vértices y el grado promedio. Por ejemplo, para redes aleatorias, se mantuvo el mismo tamaño y grado medio, pero, para redes de pequeño mundo, se utilizó el mismo tamaño y una probabilidad de reconexión de 0.25, partiendo de un grafo 2k-regular. Las simulaciones se realizaron utilizando el *software* Mathematica de Wolfram Research Inc. La [Figura 7](#) muestra una gráfica de APL vs. GCC para las redes sísmicas (verde), las redes aleatorias simuladas aleatorias (azul) y las de pequeño mundo (naranja), indicando la media de la distribución de valores para las 100 simulaciones en cada punto de la curva.

Para los datos que se muestran en la [Figura 7](#), tanto para las redes sísmicas como para las simuladas, el APL disminuye de la misma manera. Para valores bajos de APL, que corresponden a tamaños de celda más grandes (granularidad gruesa en el muestreo para crear la red), el comportamiento de la red sísmica parece ser mejor si se reproduce mediante una estructura de pequeño mundo (naranja). Para demostrar este punto, la [Tabla 1](#) presenta el coeficiente de pequeño mundo, calculado utilizando la Ecuación (1) en la red sísmica. Este resultado muestra que las redes con granularidad fina (tamaños de celda más pequeños, más nodos) tienen un coeficiente de pequeño mundo más alto, lo cual se debe principalmente a un GCC pequeño para las redes aleatorias en comparación con las redes sísmicas. Aunque el parámetro de pequeño mundo es más alto con redes más finas, es importante notar que, para todos los casos, $S > 1$, lo que significa que, para las redes de celdas cúbicas de diferentes tamaños (10 y 100 km), la red resultante es de pequeño mundo en cada caso.

En cuanto a las técnicas de análisis, se observó que la configuración de los experimentos de simulación con las reglas propuestas genera redes que comparten características con lo observado en la literatura y en los datos de redes 3D aquí presentados. Se observa un comportamiento de pequeño mundo pequeño en el gráfico de APL vs. GCC, y el grado promedio y las medidas de *small-worldness* también son evidencias de este tipo de red.

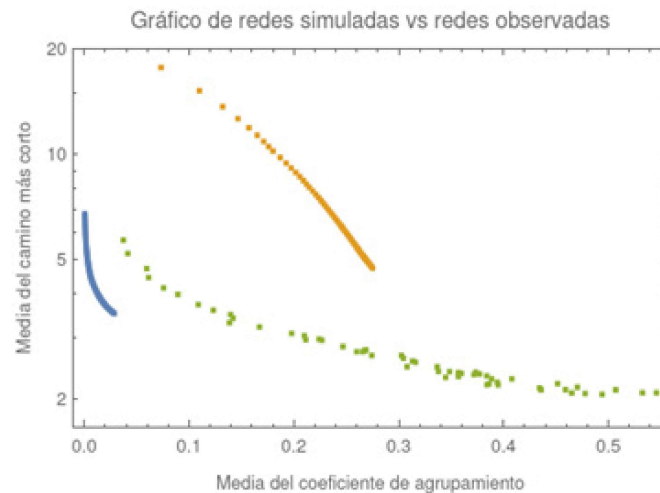


Figura 7. Mapa de pequeño mundo para redes aleatorias, redes observadas y redes de pequeño mundo

CONCLUSIONES

Se encontró que la red de sismos de Colombia, incluyendo la profundidad, puede ser caracterizada como una red de pequeño mundo. Estudios similares en otras regiones del mundo han llegado a una conclusión similar (*e.g.*, Abe & Suzuki, 2011). Según León et al. (2018), la red de sismos en dos dimensiones también es de este tipo. En este sentido, uno de los aportes de este trabajo es la inclusión de la profundidad de los eventos, lo cual proporciona una imagen más completa del fenómeno sísmico en la región. Nuestros resultados son claves debido a que en la región colombiana existen varias placas tectónicas interactuando entre sí a diferentes profundidades y en diferentes zonas del mapa. Esta complejidad requiere incluir la profundidad como una característica necesaria para comprender el fenómeno de la sismicidad en la zona.

De acuerdo con la Figura 4, el parámetro *small-worldness*, que mide qué tanto de la red es de pequeño mundo, es mucho mayor en las redes que consideran la profundidad con respecto a los resultados publicados en León et al. (2018) (Tabla 1). En estas redes de sismos hay una distribución similar a la de Poisson, lo que no se pudo observar en las redes de sismos 2D, donde la *small-worldness* decrece monótonamente al incrementar el tamaño de las celdas. Este comportamiento de la *small-worldness* es interesante porque sugiere rangos de tamaños de celda que mejoran la detección de *hubs* o zonas de interés para estudiar el fenómeno de la sismicidad. Valores muy altos de *small-worldness* indican que, dado el elevado número de nodos, la red mantiene una estructura con las características propias de las redes de pequeño mundo (alto coeficiente de Clustering y longitudes de caminos cortas) en rangos que se alejan de las redes aleatorias tipo Erdős-Rényi. Esto quiere decir que las redes observadas se alejan del equilibrio.

Conocer el tipo particular de red, ya sea aleatoria o de pequeño mundo, es de especial importancia en términos de sus características físicas (o estadísticas), pues podría existir una correlación entre la liberación de estrés en todo el sistema y la estructura de la red sísmica. De esta forma, utilizar el enfoque de redes sísmicas a profundidad en una región donde interactúan varias placas tectónicas puede proporcionar información intrínseca y útil que de otra manera estaría oculta, especialmente debido a que la región sísmica de análisis es única y poco estudiada. Por lo tanto, la caracterización de las redes sísmicas podría ayudar a descubrir aspectos relacionados con las características físicas o estadísticas del fenómeno.

Por ejemplo, la red sísmica de Colombia, a diferentes escalas de celda (10 a 100 km), muestra la importancia del ‘nido’ de Bucaramanga en la dinámica de la sismicidad de la región, independientemente del tamaño de celda utilizado para construir las redes. En este sentido, un nido sísmico se caracteriza por tener una alta tasa de actividad que está aislada de la actividad cercana, y el nido de Bucaramanga representa la mayor concentración de sismicidad de profundidad intermedia en el mundo, lo cual demuestra la importancia sismológica de este lugar. El hecho de que la actividad en esta área no genere un impacto mediático significativo se debe principalmente a que los eventos de este nido ocurren a una profundidad media de 160 km, por lo que rara vez se sienten en otras partes de Colombia, incluso en la misma Bucaramanga. Nuestros resultados resaltan esta región debido al número de eventos que ocurren en ella y sus profundidades. Esto sugiere que este lugar no está exactamente aislado del resto de la actividad en la región, sino que puede tener una influencia importante en la dinámica de los eventos sísmicos. Algo similar sucede con la sismicidad asociada al proceso de subducción en el Pacífico colombiano. La subducción se da cuando una placa tectónica se desliza por debajo de otra, lo cual hace que se generen eventos sísmicos, muchas veces de gran magnitud, a lo largo de toda la superficie de contacto entre las dos placas ([Lallemant et al., 2005](#); [Stern, 2002](#)). Esta interacción entre placas, que genera eventos a diferentes profundidades, cobra también importancia y relevancia en estos resultados, ya que muestra que la agrupación de eventos y/o celdas para una región se da también a profundidad.

Los resultados de los estudios realizados por [Abe, Pastén y Suzuki \(2011\)](#) concluyeron que las redes sísmicas en otras regiones son de pequeño mundo o de escala libre, pero su caracterización varía según el tamaño de las celdas. Las réplicas a menudo regresan al lugar del sismo principal, lo que sugiere que estos eventos pueden convertirse en nodos centrales de una red. En este sentido, el trabajo de [Abe y Suzuki \(2007, 2012\)](#) mostró una forma confiable de distinguir entre sismos principales y réplicas. En nuestro caso, las redes sísmicas tienen un comportamiento similar, y se observa claramente que el tamaño de las celdas cambia el grado, el coeficiente de agrupamiento y la distribución de la longitud del camino en la red resultante. Sin embargo, de acuerdo con los valores de *small-worldness*, cada red fue consistentemente de pequeño mundo. De forma similar a los resultados en [León et al. \(2018\)](#), las redes obtenidas en este trabajo incluyen la profundidad y mantienen el comportamiento observado en trabajos previos.

Al simular redes aleatorias y de mundo pequeño de diferentes tamaños de celda, se evidencia que las conclusiones encontradas por [Abe, Pastén y Suzuki \(2011\)](#) son similares a los valores obtenidos en este trabajo.

El uso de una disposición que optimiza la densidad de los nodos permite resaltar aquellos pertenecientes a zonas de alta sismicidad. Este efecto es particularmente interesante en el caso de un tamaño de celda de 100 km, lo cual sugiere la existencia de varias zonas sísmicas activas significativas. En el otro extremo, en la red sísmica construida con un tamaño de celda de 10 km, vemos un único grupo significativo, pero con el mayor número de eventos. Esta diferencia en ambas redes podría ser relevante para proponer una metodología que permita encontrar la zona con mayor sismicidad utilizando celdas pequeñas, para luego identificar zonas complementarias mediante tamaños de celda más grandes. Este resultado puede ser de interés para la creación de un sistema autónomo que defina etiquetas en los eventos importantes que deban ser estudiados y clasificados más detalladamente en un catálogo. Por lo tanto, la representación de la sismicidad mediante redes podría ayudarnos a entender la forma en que ocurren los eventos. El cambio de estructura de red permite entender la autoorganización del sistema de sismos para liberar estrés eficientemente, llevando la mayor cantidad de energía a los *hubs* y repartiéndola por la red hacia la periferia. Esta representación permite pensar los eventos sísmicos como precursores de otros y, al incluir la profundidad en el análisis, reconocer cuáles eventos pueden disipar más energía que otros. Así, la

profundidad se convierte en un elemento clave para comprender las zonas de mayor actividad e identificar aquellas que activan a otras en eventos posteriores.

La metodología de las redes se utiliza para diferentes tamaños de celda y ventana. Estos parámetros pueden modificar el tipo de red y las características que se observan en tamaños muy pequeños. Así, se pueden obtener redes no conectadas y, en el caso contrario, redes totalmente conectadas. Este estudio muestra que, previo al trabajo con redes de sismos, es necesario estudiar los parámetros para construirlas, pues estos pueden producir diferentes tipologías de red (e.g., redes no conectadas, totalmente conectadas o aleatorias), algunas sin ninguna complejidad. Esta investigación muestra un método para explorar todas las redes obtenidas para distintos tamaños de celda y utilizar la métrica de *small-worldness*, en aras de identificar las redes que ofrecen mayor información. Como trabajo futuro, se propone explorar medidas de entropía, ganancia de información y divergencias para estudiar las redes y la información que ofrece sobre el fenómeno de la sismicidad.

Como trabajo futuro, se propone integrar estos resultados en un sistema de pronóstico de eventos sísmicos para la zona colombiana. Para generar modelos de predicción, sería necesario seguir el tiempo entre eventos como parte de las reglas entre nodos. De manera similar, estudiar la coocurrencia entre nodos centrales y seguir los nodos más conectados también podría proporcionar información útil para diseñar un modelo de pronóstico a partir de redes de sismos.

CONTRIBUCIÓN DE AUTORÍA

Daniel León-Vargas: Conceptualización Metodología, Recursos, Redacción-Borrador original, Supervisión.
Víctor-Andrés Bucheli-Guerrero: Conceptualización, Metodología, Recursos, Redacción-Borrador original, Supervisión.
Hugo-Armando Ordoñez-Erazo: Supervisión, Redacción-Revisión y Edición.

REFERENCIAS

- Abe, S., Pastén, D., Suzuki, N. (2011). Finite data-size scaling of clustering in earthquake networks. *Physica A: Statistical Mechanics and its Applications*, 390(7), 1343-1349. <https://doi.org/10.1016/j.physa.2010.11.033>
- Abe, S., Suzuki, N. (2002). Zipf-Mandelbrot law for time intervals of earthquakes, *arXiv*, 2022, e0208344. <https://doi.org/10.48550/arXiv.cond-mat/0208344>
- Abe, S., Suzuki, N. (2003). Law for the distance between successive earthquakes. *Journal of Geophysical Research: Solid Earth*, 108(B2), e2113. <https://doi.org/10.1029/2002JB002220>
- Abe, S., Suzuki, N. (2007). Dynamical evolution of clustering in complex network of earthquakes. *The European Physical Journal B*, 59(1), 93-97. <https://doi.org/10.1140/epjb/e2007-00259-3>
- Abe, S., Suzuki, N. (2009). Main shocks and evolution of complex earthquake networks. *Brazilian Journal of Physics*, 39(2a), 428-430. <https://doi.org/10.1590/S0103-97332009000400014>
- Abe, S., Suzuki, N. (2011). Earthquake networks, complex. En R. A. Meyers (Ed.), *Extreme Environmental Events* (pp. 312-319). Springer.
- Abe, S., Suzuki, N. (2012). Dynamical evolution of the community structure of complex earthquake network. *EPL (Europhysics Letters)*, 99(3), e39001. <https://doi.org/10.1209/0295-5075/99/39001>
- Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5), 175-308. <https://doi.org/10.1016/j.physrep.2005.10.009>

- Chorozoglou, D., Kugiumtzis, D., & Papadimitriou, E. (2017). Application of complex network theory to the recent foreshock sequences of Methoni (2008) and Kefalonia (2014) in Greece. *Acta Geophysica*, 65(3), 543-553. <https://doi.org/10.1007/s11600-017-0039-4>
- Doda, L. N., Natyaganov, V. L., & Stepanov, I. V. (2013). An empirical scheme of short-term earthquake prediction. *Doklady Earth Sciences*, 453(2), 1257-1263. <https://doi.org/10.1134/S1028334X1312009X>
- Donnellan, A., Grant Ludwig, L., Parker, J. W., Rundle, J. B., Wang, J., Pierce, M., Blewitt, G., Hensley, S. (2015). Potential for a large earthquake near Los Angeles inferred from the 2014 La Habra earthquake. *Earth and Space Science*, 2(9), 378-385. <https://doi.org/10.1002/2015EA000113>
- Erdős, P., Rényi, A. (1959). *On random graphs I*. Publicationes Mathematicae Debrecen. <https://snap.stanford.edu/class/cs224w-readings/erdos59random.pdf>
- Erdos, P., Renyi, A. (1960). *On the evolution of random graphs*. Institute of Mathematics, Hungarian Academy of Sciences. <https://snap.stanford.edu/class/cs224w-readings/erdos60random.pdf>
- Frohlich, C., Kadinsky-Cade, K., Davis, S. D. (1995). A reexamination of the Bucaramanga, Colombia, earthquake nest. *Bulletin of the Seismological Society of America*, 85(6), 1622-1634.
- Gao, Z.-K., Small, M., Kurths, J. (2016). Complex network analysis of time series. *EPL (Europhysics Letters)*, 116(5), e50001. <https://doi.org/10.1209/0295-5075/116/50001>
- Gutenberg, B., Richter, C. F. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, 34(4), 185-188.
- Humphries, M. D., Gurney, K. (2008). Network “small-world-ness”: A quantitative method for determining canonical network equivalence. *PloS One*, 3(4), e0002051. <https://doi.org/10.1371/journal.pone.0002051>
- Lallemant, S., Heuret, A., Boutelier, D. (2005). On the relationships between slab dip, back-arc stress, upper plate absolute motion, and crustal nature in subduction zones. *Geochemistry, Geophysics, Geosystems*, 6(9), e09006. <https://doi.org/10.1029/2005GC000917>
- Lay, T., Wallace, T. C. (1995). *Modern global seismology*. Academic Press.
- León, D. A., Valdivia, J. A., Bucheli, V. A. (2018). Modeling of Colombian seismicity as small-world networks. *Seismological Research Letters*, 89(5), 1807-1816. <https://doi.org/10.1785/0220180076>
- León, D., Ordoñez, H., Bucheli, V. (2020). TagEQ: sistema de clasificación de eventos sísmicos basado en inteligencia artificial. *Revista Ibérica de Sistemas e Tecnologías de Informação*, E37, 359-370.
- Leon, D., Valdivia, J., Bucheli, V. (2022). A revision of seismicity models based on complex systems and earthquake networks. *Journal of Seismology*, 26(1), 137-145. <https://doi.org/10.1007/s10950-021-10017-0>
- McGuire, J. J., Boettcher, M. S., Jordan, T. H. (2005). Foreshock sequences and short-term earthquake predictability on East Pacific Rise transform faults. *Nature*, 434(7032), 457-461. <https://doi.org/10.1038/nature03377>
- Nikolopoulos D., P. E. (2015). Radon-222: A potential short-term earthquake precursor. *Journal of Earth Science & Climatic Change*, 6, e282. <https://doi.org/10.4172/2157-7617.1000282>
- Pastén, D., Czechowski, Z., Toledo, B. (2018). Time series analysis in earthquake complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(8), e083128. <https://doi.org/10.1063/1.5023923>
- Poli, P., Prieto, G. A., Yu, C. Q., Florez, M., Agurto-Detzel, H., Mikesell, T. D., Chen, G., Dionicio, V., Pedraza, P. (2016). Complex rupture of the M 6.3 2015 March 10 Bucaramanga earthquake: Evidence of strong weakening process. *Geophysical Journal International*, 205(2), 988-994. <https://doi.org/10.1093/gji/ggw065>
- Schneider, J. F., Pennington, W. D., Meyer, R. P. (1987). Microseismicity and focal mechanisms of the intermediate-depth Bucaramanga Nest, Colombia. *Journal of Geophysical Research*, 92(B13), e13913. <https://doi.org/10.1029/JB092iB13p13913>
- Sornette, D. (2003). Fractal plate tectonics. *Geophysical Research Letters*, 30(3), e1105. <https://doi.org/10.1029/2002GL015043>

- Stern, R. J. (2002). Subduction zones. *Reviews of Geophysics*, 40(4), e1012. <https://doi.org/10.1029/2001RG000108>
- van Rossum, G., Python Development Team (2013). *The Python language reference*. Python Software Foundation.
- Wolfram Research, Inc. (2022). *Mathematica, version 13.1*.
- Utsu, T., Ogata, Y., S, R., Matsu'ura. (1995). The centenary of the Omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, 43(1), 1-33. <https://doi.org/10.4294/jpe1952.43.1>
- Watts, D. J., Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440-442. <https://doi.org/10.1038/30918>



TagEQ: Sistema de Clasificación de Eventos Sísmicos Basado en Inteligencia Artificial

Daniel León¹, Hugo Ordóñez², Víctor Bucheli¹

daniel.leon@correounivalle.edu.co, hugoordonez@unicauca.edu.co, victor.bucheli@correounivalle.edu.co

¹ Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería, Universidad del Valle, Calle 13 N.º 100-00, 760031, Cali-Valle del Cauca, Colombia.

² Facultad de Ingeniería Electrónica y Telecomunicaciones Universidad del Cauca, departamento de Sistemas, Sector Tulcán Calle 5 N.º 4-70, Popayán, Cauca, Colombia.

Pages: 359–370

Resumen: La detección y clasificación de señales sísmicas es una pregunta abierta, dada la diversidad de tipos de señales, sensores y fuentes de vibración. TagEQ es un sistema de etiquetado semi-automático de señales sísmicas que toma sismogramas continuos de una estación y, a partir de un catálogo de eventos sísmicos y una representación de los sismogramas como vectores de características, clasifica automáticamente las señales. La anotación se construye a partir de características leídas del catálogo, como magnitud o profundidad de los eventos, entre otras. En este sistema se integraron modelos de clasificación que frente a nuevas entradas del sistema clasifica automáticamente los eventos sísmicos. El sistema se testeó con señales provenientes de la estación La Rusia que pertenece a la Red Sismológica Nacional de Colombia, con resultados prometedores. La aplicabilidad de este sistema, dada su modularidad, podría ampliarse a clasificar otros tipos de señales que no necesariamente son sismos tectónicos.

Palabras-clave: Aprendizaje de máquina, etiquetado semi-automático de señales, eventos sísmicos, modelos de clasificación.

TagEQ: Automatic Classification System for seismic events

Abstract: TagEQ is an automatic classification system for seismic signals. The prototype labels semi-automatically the seismic signals. In addition, the system classify seismic events by random forest models and super vector machine models. TagEQ was tested at Colombia and assigns labels semi-automatically. The prototype creates feature vectors from the seismograms, using sonograms, covariance matrix and statistical measures of the signals. The classification models were trained and the results are promising. Seismograms were registered from La Rusia station (this station belongs to the nacional seismicity network of Colombia). We discuss the results and suggest that the TagEQ can be applied to other signals, for instance, vibrations tests in buildings structures, bridges, volcanoes or Oilfields.

Keywords: Machine learning, semi-automatic labeling of signals, seismic events, classification models.

1. Introducción

Un sismo es la vibración que resulta de una súbita liberación de energía acumulada en un material que ya no puede deformarse más ante la presión a la que es sometido. Esta súbita liberación genera dos tipos de ondas principales que se conocen como ondas de cuerpo; y se identifican en sismología como ondas P y S. Debido a la importancia de conocer a fondo este tipo de fenómenos para tratar de predecir su ocurrencia, o de saber qué zonas pueden correr más peligro ante un evento, el procesamiento de la señales sísmicas ha sido un tema de investigación ampliamente tratado desde muchas ramas de la ciencia, como la física y las matemáticas, así como también para las ciencias de la computación.

A partir de las ondas P y S de los sismos y sus valores de velocidad de propagación se busca estudiar y predecir este tipo de eventos. Sin embargo, la detección es un problema abierto de investigación. Debido al gran volumen de datos y señales que genera una red sismológica de varias estaciones, se hace necesario tener un esquema automático de detección. Se han llevado a cabo diferentes trabajos en el área de procesamiento de señales, así como también, en el área de la detección de eventos. Estas dos áreas son de interés especial para este trabajo. Existen actualmente muchas aproximaciones a la solución de la detección automática de sismos usando esquemas de aprendizaje de máquina, que son una rama de la inteligencia artificial. Sin embargo, hay una brecha continuamente abierta debido a la inmensa heterogeneidad y gran cantidad de la información que presentan las ondas sísmicas. Así, los diferentes tipos de sensores que hacen el registro de eventos recolectan información de los diferentes tipos de estructuras de la tierra que atraviesan las ondas hasta llegar a tales sensores, los diferentes tipos de eventos sísmicos que se presentan tales como volcanes, entre otros. Entonces, la brecha abierta se puede relacionar con el problema de predecir algún tipo de estos eventos, lo cual puede salvar la vida de muchas personas y la estabilidad de una población.

Este trabajo propone un sistema prototipo de clasificación de eventos sísmicos basado en inteligencia artificial, que utiliza algoritmos de aprendizaje de máquina supervisados. Se desarrolló un corpus de señales a partir de los datos recopilados de La Rusia (estación que pertenece a la Red Sismológica Nacional de Colombia), de tal manera que se puedan aplicar los algoritmos de detección en términos del rango de señales que pueden detectar. Para ello, se creó un módulo en el que se hace una parametrización de las ondas sísmicas y un módulo de etiquetado automático de señales con base en un catálogo de eventos. A partir de los datos de una estación sismológica, se desarrollaron los modelos para clasificar las señales. La clasificación se hace frente a la magnitud del sismo o de características que se puedan inferir a partir del catálogo.

TagEQ se construyó como un sistema que ejecuta una serie de módulos (programas) y líneas de comando para clasificar señales sísmicas. Para realizar esta clasificación se deben procesar las señales sismológicas, crear vectores de características, implementar un esquema de asignación de etiquetas, y finalmente usar esta información para configurar un algoritmo supervisado de machine learning que aprenda a asignar la etiqueta. Para poder hacer sus tareas, se deben tener un catálogo y los sismogramas continuos de una estación para un lapso con el que se pueda entrenar el sistema de aprendizaje de máquina. A partir del catálogo de sismicidad, se busca en los sismogramas los trozos

de señales que correspondan a cada evento y se parametrizan, creando vectores de características de estas señales, que luego se usarán como entrada para el algoritmo supervisado. En este trabajo se utilizaron las máquinas de soporte vectorial y los bosque aleatorios (Random Forest).

El artículo está organizado de la siguiente manera: en la Sección 2 se presentan trabajos relacionados, en la Sección 3 se presenta la arquitectura del sistema TagEQ, seguido de la Sección 4 en la que se presentan los resultados y finalmente la Sección 5 la discusión.

2. Trabajos relacionados

En sismología se han implementado muchos tipos de clasificadores, desde Mapas Auto-Organizados (Köhler, Ohrnberger, & Scherbaum, 2009, 2010; Sick, Guggenmos, & Joswig, 2015), Máquinas de Soporte Vectorial (Ruano et al., 2014), o Redes Neuronales (Gentili & Bragato, 2006), buscando clasificar entre pocos tipos de señales (ruido-sismo, sismos locales-sismos regionales-tesisismos, voladuras-sismos). Por otro lado, en (Groos & Ritter, 2009) se hace una clasificación de tipos de ruido en el área urbana de Bucarest a partir de un análisis de la distribución Gaussiana de las muestras en 8 diferentes bandas de frecuencia para las señales adquiridas; logrando clasificar entre 6 tipos diferentes de ruidos. Si bien las señales que se tomaron para este análisis son de 4 horas, este trabajo muestra que el análisis estadístico de la distribución de las muestras puede entregar parámetros que sirvan a un sistema de clasificación.

Para hacer detecciones y clasificaciones, las señales sísmicas son procesadas y parametrizadas de diferentes maneras. En (Sabbione & Velis, 2013) se muestra un ejemplo de parametrización que permite detectar microsismos inducidos por actividades relacionadas con el petróleo calculando la envolvente de las señales y aplicando la técnica STA/LTA para detectar señales que puedan indicar sismos, las cuales se corroboran a partir de cálculo de tiempos de llegada en otros sensores, y en (Akram & Eaton, 2012) se hace a partir de umbrales dinámicos. En (Pikoulis & Psarakis, 2015) se usan algunas características estadísticas de las señales en ventanas móviles (promedios y medianas) para segmentar registros entre señal y ruido, y así finalmente hacer detecciones automáticas más estables que las realizadas con algoritmos STA/LTA, en términos de disminución de falsas alarmas. En (Maurer & Deichmann, 1995) se ha implementado un sistema de detección de cúmulos de sismicidad a partir de correlaciones de señales sísmicas, asumiendo que la forma de la onda de eventos muy cercanos debe ser muy parecida al viajar un camino altamente similar, por lo cual serán afectadas por las mismas estructuras del subsuelo. En (Karamzadeh, Voss, & Javan, 2012) se usan espectrogramas calculados en bandas de 2Hz calculados entre 2 y 16 Hz para detectar eventos en una red de Groenlandia, añadiendo elementos que ayuden a discriminar eventos además de los algoritmos STA/LTA. Así mismo, en (Köhler et al., 2010) y (Köhler et al., 2009) se resumen otros tipos de parametrizaciones, ya referidas en el marco referencial de este documento.

Este trabajo tiene como propósito mostrar una arquitectura modular que pueda usarse para clasificar eventos usando sismogramas de una estación y un catálogo sísmico, aplicando directivas de alto nivel para anotar los eventos, implementando varios algoritmos de extracción de características, detección y clasificación, y mostrando cómo se pueden integrar en un sistema de clasificación.

3. Arquitectura TagEQ

Los datos de entrada para TagEQ son un catálogo anotado y los sismogramas de una estación para un lapso fide tiempo. Estos son utilizados como datos de entrenamiento por el algoritmo de aprendizaje de máquina. A partir del catálogo de sismicidad, se busca en los sismogramas de manera automática las señales que correspondan a cada evento y se parametrizan, creando vectores de características de dichas señales, que luego se usarán como la entrada para el algoritmo de aprendizaje que se desee utilizar: máquinas de soporte vectorial (SVM) o bosques de decisión (RF). El proyecto está disponible en <https://gitlab.com/danieleon1/TagEQ>.

3.1. Flujo de datos para el sistema TagEQ

Para el funcionamiento del sistema se usa un archivo de configuración que funciona como un mapa para que el sistema sepa dónde están los datos de las señales, los parámetros para los algoritmos de clasificación; los parámetros con los que se asignan las etiquetas, el tipo de clasificación que se desea hacer y las reglas con las que se deben aplicar las etiquetas. En la Figura 1 se presenta el flujo de datos y la arquitectura del sistema.

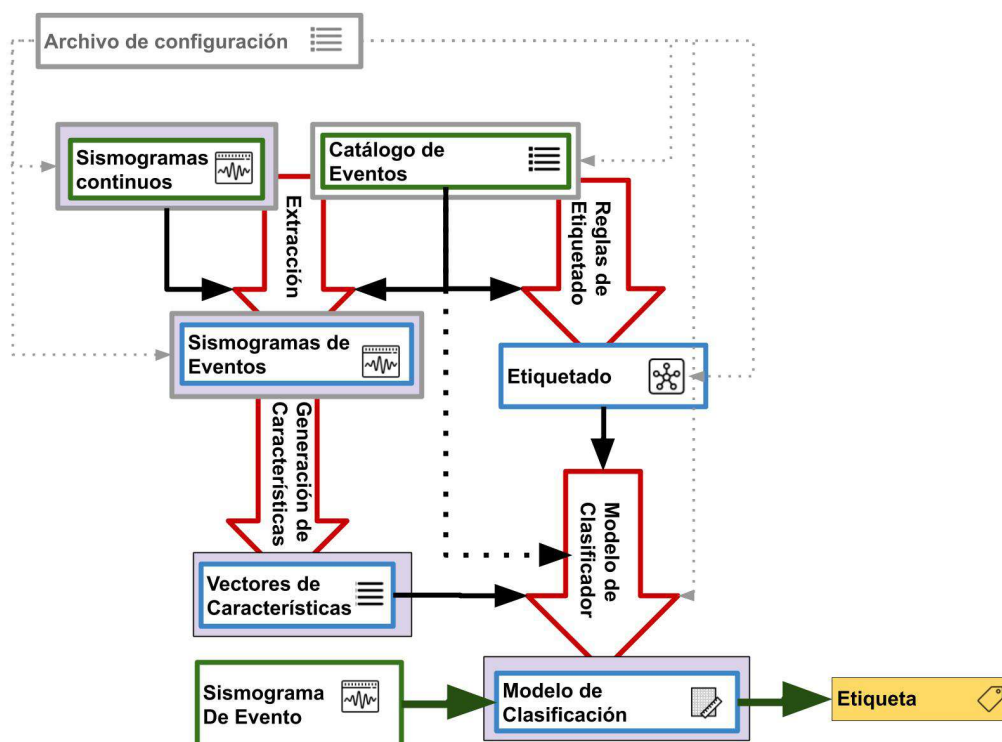


Figura 1 – Flujo de datos y arquitectura del sistema

Las tareas representadas en la Figura 1 son:

- Ejecución de las tareas de pre-procesamiento de los sismogramas continuos, extracción de las señales de los eventos que hacen parte del catálogo; y definición de las reglas con las que se asignan las etiquetas a los eventos.
- Creación de vectores de características a partir de los sismogramas de los eventos.
- Creación de modelos de clasificadores a partir de los vectores y las reglas definidas para etiquetar los eventos.
- Ya con los modelos construidos, lectura del sismograma de un evento para asignarle la etiqueta correspondiente.

3.2. Módulos del sistema TagEQ

La Figura 2 muestra los principales módulos del sistema. En ella se presentan las entradas y salidas y estructura modular del sistema TagEQ. Así se presentan los cuatro módulos principales: Configuración, ETL (Extracción, Transformación y Carga), Entrenamiento de clasificador y Etiquetado automático. Cada módulo se ejecuta de manera automática cuando el usuario lo indica, sin embargo solo los módulos de Configuración y Etiquetado automático precisan de interacción con el usuario para poder entregar su respectiva salida. A continuación se describe en detalle cada módulo del sistema.

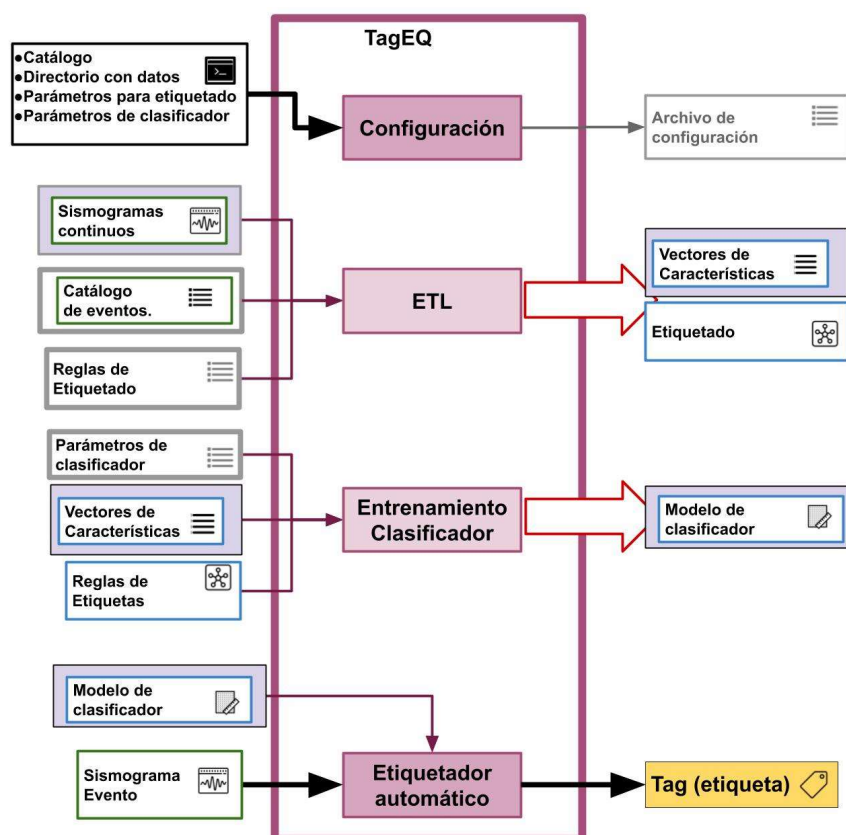


Figura 2 – Módulos del sistema TagEQ

Módulo de configuración: El módulo de configuración tiene como propósito que el usuario haga explícitas las directivas con las que desea que el sistema realice la clasificación de las señales. Hay dos clases de directivas para el sistema; las directivas generales y las directivas de Machine Learning. En las directivas generales se describe el lugar en que están los sismogramas, cuál es el catálogo con el que se trabaja, dónde se almacenan los datos generados (sismogramas de eventos y catálogos anotados), y qué reglas se usan para anotar los datos. En las directivas de Machine Learning se define el tipo de clasificador y los parámetros con los que se sintonizan tales clasificadores.

Módulo de Extracción, Transformación y Carga (ETL): Realiza todos los procedimientos de pre-procesamiento de datos y parametrización de sismogramas de los eventos. La Figura 3 presenta el esquema general de funcionamiento de este módulo. El módulo ETL toma las entradas (sismogramas continuos y catálogo de sismicidad) y genera los vectores de características que posteriormente se usarán para clasificar eventos. Usando los tiempos de origen de los eventos relacionados en el catálogo de sismicidad, se buscan en los sismogramas continuos las señales que correspondan a los eventos para posteriormente parametrizarlas usando diferentes técnicas, y luego se almacenan estos vectores de características. Para la arquitectura propuesta en el sistema TagEQ, se crearon programas para generar vectores según los siguientes criterios:

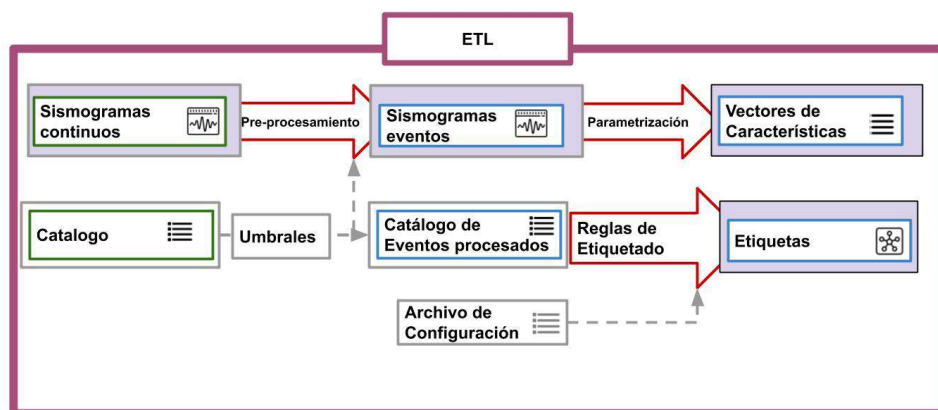


Figura 3 – Módulo de Extracción, Transformación y Carga (ETL)

- *Estadísticos:* Se calcula promedio, varianza, asimetría estadística y curtosis en ventanas de 50 segundos para los 3 canales del sismograma.
- *Sonograma:* Calcula el promedio en lapsos de 50 segundos de la energía de 9 bandas de frecuencia, a partir del algoritmo propuesto por (Joswig, 1994).
- *Dimension fractal de la señal:* Calcula la dimensión fractal de la componente vertical para una ventana definida.
- *Variación de la dimensión fractal:* Calcula la varianza de la dimensión fractal calculada en diferentes ventanas.
- *Covarianzas:* Calcula la matriz de covarianzas, rectilinealidad y planaridad de la señal en una definida.

Módulo entrenamiento del clasificador: Para realizar el entrenamiento del clasificador, se toman los vectores de características y se asigna una etiqueta correspondiente. Esta asignación se hace revisando la información consignada en el catálogo de eventos procesados y resolviendo la etiqueta a partir de las reglas de alto nivel definidas en el módulo de configuración. La idea es que estas reglas de alto nivel se especifiquen como una directiva general (eventos fuertes o pequeños, eventos profundos o someros, eventos que pertenecen o no a la zona X), y luego se etiquetan los datos resolviendo la directiva usando el catálogo de sismicidad. El esquema del flujo de información en este módulo se muestra en la Figura 4.

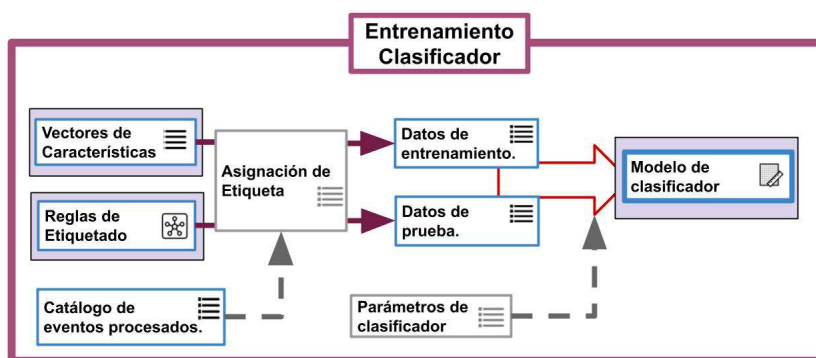


Figura 4 – Módulo de entrenamiento del clasificador

Para crear el modelo de clasificador, se usa una parte del set de datos como entrenamiento y restante como datos de prueba. Para cada caso de algoritmo de clasificación se usan los parámetros introducidos en el archivo de configuración ya sea para generar un solo modelo o para generar muchos modelos y escoger el que mejor desempeño tenga. Para el sistema TagEQ se usaron las librerías LibSVM para las máquinas de soporte vectorial (Chang & Lin, 2011); y para los bosques aleatorios de decisión el programa Ranger (Wright & Ziegler, 2017).

Módulo etiquetador de eventos: el módulo requiere que manualmente se le diga al sistema cuál es el sismograma del evento a clasificar, tras lo cual se parametriza el sismograma, se construye un vector de características, y luego se le ingresa al modelo construido anteriormente. Al ejecutar la predicción del clasificador entrenado se tiene la etiqueta del evento. La Figura 5 muestra el esquema en que funciona el módulo del etiquetador.

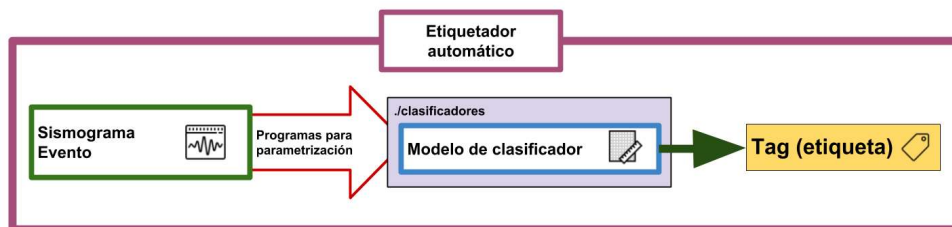


Figura 5 – Módulo etiquetador de eventos

4. Resultados

El catálogo con el que se trabajó consta de 51520 eventos. De estos eventos, se trabajó solamente los que tenían magnitud $M > 2$, y que tuvieran suficiente energía sísmica en la estación La Rusia; aunque en algunos casos se realizó la prueba con los eventos con magnitud $M > 3$. El conjunto de prueba final consta de 8151 eventos con $M > 2$, y de 502 para $M > 3$, y se probaron los algoritmos para realizar clasificación binaria.

Los casos de clasificación binaria se configuraron para eventos que superan un umbral de magnitud. Dos casos se probaron, con el umbral $U=3$ y $U=4$. La Figura 6 muestra el conjunto de eventos con magnitud mayor a cada umbral, y las Figuras 7 y 8 presentan la precisión y la sensibilidad respectivamente de los clasificadores configurados usando tanto SVM como RF.

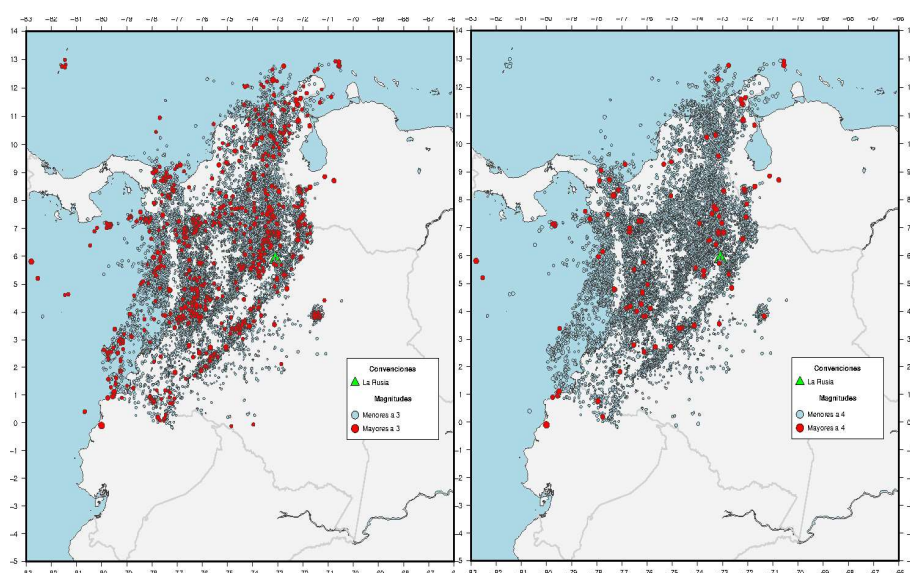


Figura 6 – Mapas de eventos con magnitud mayor a 3 (izquierda) y mayor a 4 (derecha)

La Figura 7 muestra que tanto para las SVM como para los Random Forest los clasificadores presentan una precisión por encima del 80%, y que en algunos casos las SVM se comportan un poco mejor que las Random Forest. Sin embargo, la Figura 8 muestra que los Random Forest tienen una sensibilidad mucho mayor que las SVM. La sensibilidad es una medida de que tan bien el clasificador encuentra los casos positivos en un sistema binario, siendo en este caso la clase positiva los eventos con magnitud por encima del umbral definido. A pesar de que la sensibilidad no parece ser muy buena (cerca del 60% usando los Random Forest), hay que anotar que los conjuntos son altamente desbalanceados, con muchísimos más casos de eventos por debajo de los umbrales propuestos.

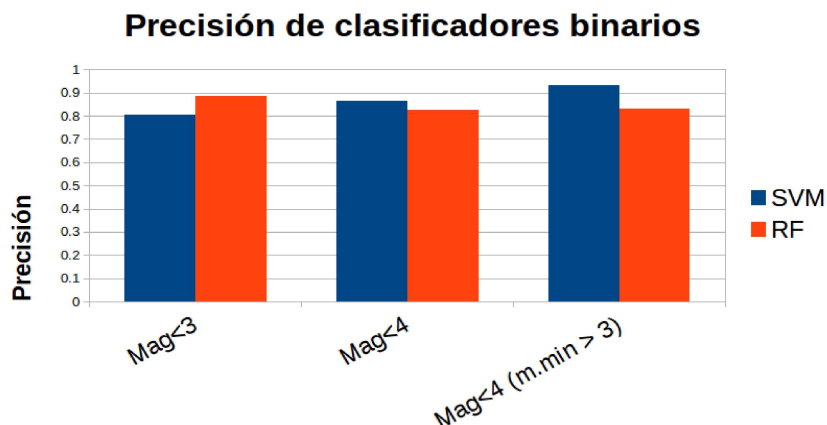


Figura 7 – Precisión de los clasificadores usando SVM y RF

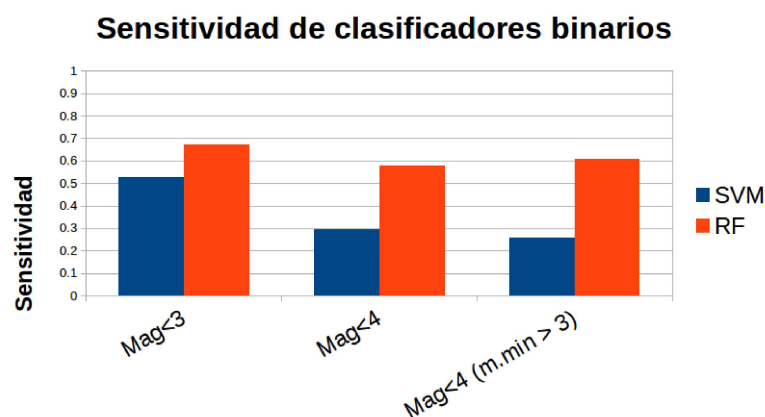


Figura 8 – Sensitividad de los clasificadores usando SVM y RF

5. Discusión

El sistema desarrollado se basa en el paradigma supervisado, aunque buena parte del sistema funciona de manera automática en el sentido de que el usuario no escoge directamente cuál es el conjunto de señales que se utilizan para el entrenamiento y prueba del clasificador; ni asigna las etiquetas a los eventos, sino que se asignan a partir de un catálogo de sismos previo, es decir, de una anotación inicial que un analista ha hecho de la sismicidad; y a partir de reglas de alto nivel que defina el usuario. Dado el nivel de abstracción que el clasificador alcanzó, es posible que pueda usarse también con otro tipo de señales, como señales de deslizamientos, explosiones, o movimiento de

fluidos. Sin embargo para poder usar esta técnica se debe crear un catálogo de tiempos de eventos y algún tipo de identificación de la señal.

La clasificación de eventos sísmicos con una sola estación puede ser de gran ayuda cuando no se tiene una red de sensores amplia. Utilizar una única estación para clasificar eventos podría ayudar a un analista a generar una alarma para revisar alguna situación a partir de una única señal adquirida, sin necesidad de esperar que se tengan las señales de otras estaciones, entregando un reporte preliminar que ponga en “estado de alerta” a un operador de datos sismológicos, con una información preliminar más completa que simplemente decir que una señal superó un nivel de energía.

El sistema desarrollado posee un nivel de modularidad que podría permitir la adición de nuevas características a los vectores desarrollando nuevos programas que extraigan información analizada de un sismograma.

La clasificación binaria obtiene precisiones para ambos tipos de clasificadores que superan el 80%, aunque la sensibilidad muestra un desempeño mucho mejor en el caso de los Random Forest. Existen trabajos con niveles mejores que lo encontrado en este trabajo, lo cual puede ser consecuencia de usar una única estación (estación La Rusia) para clasificar eventos en un rango de distancias considerables y de tener muchos eventos cerca del umbral de decisión, por lo que si se desea detectar sin falta eventos por encima de una magnitud específica debería usarse un umbral de magnitud por lo menos un grado por debajo. El resultado de las precisiones indica que los parámetros escogidos para representar los eventos en los vectores de características pueden considerarse adecuados; aunque para mejorar el nivel de abstracción con eventos de tantas partes diferentes, con formas de onda diferentes, se podrían necesitar otras características.

Se pudo evidenciar que tanto para los clasificadores Random Forest como para las Support Vector Machine los resultados son buenos. El resultado que se halló también indica que los Random Forest son más adecuados para hacer la clasificación de señales sísmicas.

La clasificación de nuevas señales a partir de las reglas de alto nivel implementadas en el catálogo podría permitir a un analista encontrar rápidamente señales candidatas de algún tipo (eventos profundos/someros; grandes/pequeños, pertenecen/no pertenece de cierto interés, implementando en conjunto un algoritmo de detección con uno de clasificación.

Como trabajo futuro, dado que este es un prototipo de clasificador, se propone tener un ambiente amigable con el usuario, y definir un marco amplio en el que se puedan definir directivas de alto nivel para asignar etiquetas.

Referencias

Akram, J., & Eaton, D. (2012, May). “Adaptive microseismic event detection and automatic time picking”. Presented at the CSPG/CSEG/CWLS GeoConvention 2012 Vision, Canada.

- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: “A library for support vector machines”. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1–27. <https://doi.org/10.1145/1961189.1961199>
- Gentili, S., & Bragato, P. (2006). “A neural-tree-based system for automatic location of earthquakes in Northeastern Italy. *Journal of Seismology*”, 10(1), 73–89. <https://doi.org/10.1007/s10950-005-9001-z>
- Groos, J. C., & Ritter, J. R. R. (2009). “Time domain classification and quantification of seismic noise in an urban environment”. *Geophysical Journal International*, 179(2), 1213–1231. <https://doi.org/10.1111/j.1365-246X.2009.04343.x>
- Joswig, M. (1994). “Knowledge-based seismogram processing by mental images”. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(3), 429–439. <https://doi.org/10.1109/21.278992>
- Karamzadeh, N., Voss, P. H., & Javan, G. D. (2012). “Testing of an automatic earthquake detection method on data from Station Nord”, Greenland. *Geological Survey of Denmark and Greenland Bulletin*, 26, 77–80. (Testing-Automatic-Earthquake-detecmethod-greenland-karamzadeh-et-al-2002.pdf).
- Köhler, A., Ohrnberger, M., & Scherbaum, F. (2009). “Unsupervised feature selection and general pattern discovery using Self-Organizing Maps for gaining insights into the nature of seismic wavefields”. *Computers & Geosciences*, 35(9), 1757–1767. <https://doi.org/10.1016/j.cageo.2009.02.004>
- Köhler, A., Ohrnberger, M., & Scherbaum, F. (2010). “Unsupervised pattern recognition in continuous seismic wavefield records using Self-Organizing Maps: Unsupervised seismic pattern recognition”. *Geophysical Journal International*, 182(3), 1619–1630. <https://doi.org/10.1111/j.1365-246X.2010.04709.x>
- Maurer, H., & Deichmann, N. (1995). “Microearthquake cluster detection based on waveform similarities, with an application to the western Swiss Alps”. *Geophysical Journal International*, 123(2), 588–600. <https://doi.org/10.1111/j.1365-246X.1995.tb06873.x>
- Pikoulis, E.-V., & Psarakis, E. Z. (2015). “Automatic Seismic Signal Detection via Record Segmentation”. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7), 3870–2884. <https://doi.org/10.1109/TGRS.2014.2386255>
- Ruano, A. E., Madureira, G., Barros, O., Khosravani, H. R., Ruano, M. G., & Ferreira, P. M. (2014). “Seismic detection using support vector machines”. *Neurocomputing*, 135, 273–283. <https://doi.org/10.1016/j.neucom.2013.12.020>
- Sabbione, J. I., & Velis, D. R. (2013). “A robust method for microseismic event detection based on automatic phase pickers”. *Journal of Applied Geophysics*, 99, 42–50. <https://doi.org/10.1016/j.jappgeo.2013.07.011>

- Sick, B., Guggenmos, M., & Joswig, M. (2015). “Chances and limits of single-station seismic event clustering by unsupervised pattern recognition”. *Geophysical Journal International*, 201(3), 1801–1813. <https://doi.org/10.1093/gji/ggv126>
- Wright, M. N., & Ziegler, A. (2017). ranger: “A Fast Implementation of Random Forests for High Dimensional Data in C++ and R”. *Journal of Statistical Software*, 77(1). <https://doi.org/10.18637/jss.v077.i01>

© 2020. This work is published under
<https://creativecommons.org/licenses/by-nc-nd/4.0/>(the
“License”). Notwithstanding the ProQuest Terms and
Conditions, you may use this content in accordance with the
terms of the License.

Solar Radiation Prediction on Photovoltaic Systems Using Machine Learning Techniques

Luis-Eduardo Ordoñez-Palacios; Daniel-Andrés León-Vargas; Víctor-
Andrés Bucheli-Guerrero; Hugo-Armando Ordoñez-Eraso

Citación: L.-E. Ordoñez-Palacios, D.-A. León-Vargas, V.-A. Bucheli-
Guerrero, H.-A. Ordoñez-Eraso, “Solar Radiation Prediction on
Photovoltaic Systems Using Machine Learning Techniques,” *Revista
Facultad de Ingeniería*, vol. 29 (54), e11751, 2020.

<https://doi.org/10.19053/01211129.v29.n54.2020.11751>

Recibido: Julio 12, 2020; **Aceptado:** Septiembre 17, 2020;

Publicado: Septiembre 18, 2020

Derechos de reproducción: Este es un artículo en acceso abierto
distribuido bajo la licencia [CC BY](#)



Conflicto de intereses: Los autores declaran no tener conflicto de
intereses.

Solar Radiation Prediction on Photovoltaic Systems Using Machine Learning Techniques

Luis-Eduardo Ordoñez-Palacios¹

Daniel-Andrés León-Vargas²

Víctor-Andrés Bucheli-Guerrero³

Hugo-Armando Ordoñez-Eraso⁴

Abstract

Estimation of solar radiation is essential to help decision-makers in the planning of isolated solar energy farms or connected to electricity distribution networks to take advantage of renewable energy sources, reduce the impact produced by climate change, and increase coverage rates in electricity service. The number of existing measurement stations is insufficient to cover the entire geography of a region, and many of them are not capturing solar radiation data. Therefore, it is important to use mathematical, statistical, and artificial intelligence models, which allow predicting solar radiation from meteorological data available. In this work, datasets taken from measurement stations located in the cities of Cali and Villavicencio were used, in addition to a dataset generated by the World Weather Online API for the town of Mocoa, to carry out solar radiation estimations using different machine learning techniques for regression and classification to evaluate their performance. Although in most related works researchers used deep learning to predict solar radiation, this work showed that, while artificial neural networks are the most widely used

¹ Universidad del Valle (Santiago de Cali-Valle del Cauca, Colombia). luis.ordonez.palacios@correounivalle.edu.co. ORCID: [0000-0001-5154-9472](https://orcid.org/0000-0001-5154-9472)

² M. Sc. Universidad del Valle (Santiago de Cali-Valle del Cauca, Colombia). daniel.leon@correounivalle.edu.co. ORCID: [0000-0002-8283-2243](https://orcid.org/0000-0002-8283-2243)

³ Ph. D. Universidad del Valle (Santiago de Cali-Valle del Cauca, Colombia). victor.bucheli@correounivalle.edu.co. ORCID: [0000-0002-0885-8699](https://orcid.org/0000-0002-0885-8699)

⁴ Ph. D. Universidad del Cauca (Popayán-Cauca, Colombia). hugoordonez@unicauca.edu.co. ORCID: [0000-0002-3465-5617](https://orcid.org/0000-0002-3465-5617)

technique, other machine learning algorithms such as Random Forest, Vector Support Machines and AdaBoost, also provide estimates with sufficient precision to be used in this field of study.

Keywords: deep learning; machine learning; photovoltaic systems; prediction model; solar radiation; supervised learning.

Predicción de radiación solar en sistemas fotovoltaicos utilizando técnicas de aprendizaje automático

Resumen

La estimación de la radiación solar es fundamental para quienes participan en la planificación de granjas de energía solar, ya sean aisladas o conectadas a las redes de distribución eléctrica. Esto para el aprovechamiento de las fuentes de energía renovables, reducir el impacto producido por el cambio climático, e incrementar los índices de cobertura en el servicio eléctrico. De igual manera, el número de estaciones de medición existentes es insuficiente para cubrir toda la geografía de una región, y muchas de ellas no están capturando datos de radiación solar. Por consiguiente, es importante hacer uso de modelos matemáticos, estadísticos y de inteligencia artificial que permitan predecir la radiación solar a partir de datos meteorológicos disponibles. En este trabajo se utilizaron conjuntos de datos tomados de estaciones de medición ubicadas en las ciudades de Cali y Villavicencio, además de un conjunto de datos generado por la API World Weather Online para la ciudad de Mocoa. La razón fue realizar estimaciones de radiación solar utilizando distintas técnicas de aprendizaje automático para regresión y clasificación; el principal objetivo fue evaluar su desempeño. Aunque en la mayoría de los trabajos relacionados los investigadores utilizaron el aprendizaje profundo para la predicción de la radiación solar, este estudio demostró que, si bien las redes neuronales artificiales son la técnica más utilizada, otros algoritmos de aprendizaje automático como Random Forest, Máquinas de Soporte Vectorial y AdaBoost también proporcionan estimaciones con suficiente precisión para ser utilizados en este campo de estudio.

Palabras clave: aprendizaje automático; aprendizaje profundo; aprendizaje supervisado; modelo de predicción; radiación solar; sistemas fotovoltaicos.

Predição de radiação solar em sistemas fotovoltaicos utilizando técnicas de aprendizagem automática

Resumo

A estimativa da radiação solar é fundamental para quem participa na planificação de granjas de energia solar, já sejam isoladas ou conectadas às redes de distribuição elétrica. Isto para o aproveitamento das fontes de energia renováveis, reduzir o impacto produzido pela mudança climática, e incrementar os índices de cobertura no serviço elétrico. De igual maneira, o número de estações de medição existentes é insuficiente para cobrir toda a geografia de uma região, e muitas delas não estão capturando dados de radiação solar. Por conseguinte, é importante fazer uso de modelos matemáticos, estatísticos e de inteligência artificial que permitam prever a radiação solar a partir de dados meteorológicos disponíveis. Neste trabalho utilizaram-se conjuntos de dados tomados de estações de medição localizadas nas cidades de Cali e Villavicencio, além de um conjunto de dados gerado pela API World Weather Online para a cidade de Mocoa. A razão foi realizar estimativas de radiação solar utilizando distintas técnicas de aprendizagem automática para regressão e classificação; o principal objetivo foi avaliar seu desempenho. Ainda que na maioria dos trabalhos relacionados os pesquisadores utilizaram a aprendizagem profunda para a predição da radiação solar, este estudo demonstrou que, se bem as redes neurais artificiais são a técnica mais utilizada, outros algoritmos de aprendizagem automática como Random Forest, Máquinas de Suporte Vectorial e AdaBoost também proporcionam estimativas com suficiente precisão para serem utilizados neste campo de estudo.

Palavras chave: aprendizagem automática; aprendizagem profunda; aprendizagem supervisada; modelo de predição; radiação solar; sistemas fotovoltaicos.

I. INTRODUCCIÓN

Con el agotamiento escalonado en las reservas de combustibles fósiles [1], el calentamiento global y los cambios climáticos extremos [2], la mayoría de los países está tomando conciencia acerca de la importancia de utilizar tecnologías innovadoras para la generación de energía limpia, renovable y amigable con el ambiente. Entre estas tecnologías, los sistemas fotovoltaicos se constituyen como una alternativa que se ha fortalecido recientemente, en parte, gracias a que los costos se han reducido gradualmente en los últimos años y el desarrollo tecnológico ha mejorado la eficiencia de sus componentes [3]. Por otro lado, aun cuando la energía solar se caracteriza por ser inagotable y limpia, su explotación y transformación a energía eléctrica depende de factores climáticos, geográficos y técnicos.

La instalación de sistemas fotovoltaicos requiere de una evaluación conceptual previa con el objetivo de conocer la viabilidad del sitio tentativo respecto a las necesidades que se quieren satisfacer. Esta evaluación previa es clave para apoyar a las personas que toman las decisiones en la planificación e implementación de granjas de energía solar (que bien pueden ser aisladas o conectadas a las redes de distribución eléctrica), así como en la operación y la administración de la energía producida. Todo esto se suma al creciente interés mundial por el desarrollo de sistemas de generación de energía limpios y amigables con el ambiente.

Es esencial conocer el comportamiento de la radiación solar global de un lugar específico y, al mismo tiempo, es trascendental en el cálculo del número de módulos solares y los demás componentes que se requieren para producir cierta cantidad de energía eléctrica. Sin embargo, los datos de irradiancia no están disponibles en la mayoría de las estaciones de medición existentes, lo que conlleva a precisar de tiempo adicional para realizar las mediciones y el procesamiento de los datos, ocasionando retrasos y riesgos en la ejecución de este tipo de proyectos.

Actualmente, la Inteligencia Artificial [4] está ampliando las fronteras de la predicción en campos como las finanzas, la justicia [5], la salud, el clima, la agricultura, las ventas, el transporte, entre otros; esto gracias al aprendizaje automático, el procesamiento de imágenes, el procesamiento del lenguaje natural, la robótica, etc.

En el sector de la generación de energía eléctrica, las técnicas del aprendizaje automático han cobrado gran importancia en la estimación de atributos meteorológicos a partir de conjuntos de datos estructurados, semiestructurados y no estructurados. Aunque muchas de las estaciones no están tomando lecturas de brillo solar, existen algunas que sí la miden, junto con otras variables climáticas, con el fin de lograr una mejor comprensión del ambiente. En este trabajo se verifica si estas variables son suficientes para estimar la energía luminosa procedente del sol. El presente estudio logró confirmar que la literatura relacionada con la estimación de la radiación solar en sistemas fotovoltaicos utilizando técnicas de Inteligencia Artificial se produjo en los últimos diez años, en parte, gracias al gran impulso que han tenido las Ciencias de la Computación [6]. También se pudo determinar que la mayoría de los investigadores modernos utilizan el aprendizaje profundo para realizar estimaciones de radiación solar, como se ha presentado en trabajos preliminares [7-12]. A pesar de ello, otros documentos han usado técnicas como árboles de decisión, regresión lineal, no lineal, multivariante y logística [13]; de igual manera, se han utilizado algoritmos como Random Forest [14], clustering [15], máquinas de soporte vectorial [16-17] y modelos híbridos [18].

Sharadga et al. [19] emplean modelos de predicción basados en métodos estadísticos y de inteligencia artificial para la estimación de salida de energía eléctrica a partir de la radiación solar en sistemas fotovoltaicos. Con respecto a la inteligencia artificial, se consideraron los siguientes modelos de Redes Neuronales: Bi-LSTM, agrupamiento Fuzzy c-Means, LSTM, LRNN, MLP y feedforward NNs. De acuerdo con los resultados, el modelo Bi-LSTM tuvo el mejor desempeño, con un coeficiente de correlación del 98%, mientras que RMSE tuvo uno de 0.791, aunque el modelo perceptrón multicapa (MLP) fue el algoritmo más rápido. El estudio demostró que las redes neuronales son más precisas que los modelos estadísticos y requieren menos tiempo de cálculo.

Las redes neuronales y el aprendizaje automático son bastante utilizados en aplicaciones de radiación solar, evaluando el desempeño del perceptrón multicapa (MLP) y los árboles de decisión potenciados mediante la combinación con regresión lineal para la estimación de la energía solar en el sur de Argelia. El conjunto de

datos utilizado incluye las variables radiación solar global diaria, radiación solar global extraterrestre, temperatura del aire y duración de la luz solar; los resultados demuestran que el modelo MLP tuvo un mejor rendimiento según los indicadores coeficiente de determinación R^2 de 97.7% y el RMSE de 0.033 [20].

Otros estudios realizan una mejora en la técnica de regresión Random Forest para predecir la radiación solar en entornos complejos, considerando los factores que influyen en la energía solar, y realizando la selección de parámetros óptimos a partir del análisis de error OOB. Se demostró la validez del modelo gracias a la reducción del error cuadrático medio hasta 19.6085 y la mejora en la precisión de la estimación con un coeficiente de determinación R^2 de 0.9665. Se utilizaron datos de monitoreo de la radiación solar del laboratorio del Instituto de Ingeniería de Shenyang, para una muestra entre 2004 y 2014, que incluye horas de sol, presión atmosférica promedio, velocidad promedio del viento, temperatura promedio del aire, ángulo de elevación solar y humedad promedio [14].

Koo et al. [21] abordan los desafíos relacionados con los patrones espaciales complejos en el promedio mensual de radiación solar diaria en China, así como la carencia de estaciones de observación de esta variable debido a los altos costos de los equipos. El estudio desarrolló un enfoque de estimación mediante técnicas de aprendizaje automático como el algoritmo de agrupamiento K-means y un modelo avanzado de razonamiento basado en casos CBR. Se obtuvo una precisión de estimación promedio de 93.23%, a partir de información recopilada de 97 ciudades de China, en un periodo de 10 años entre 2006 y 2015.

De acuerdo con estudios realizados por Hussain y AlAlili [22], los datos solares no están disponibles en todos los puntos de interés a causa de la insuficiencia de estaciones de medición y sensores solares, por ello, la irradiancia debe estimarse utilizando modelos. Los investigadores desarrollaron una técnica híbrida para mejorar el rendimiento de una red neuronal (ANN), analizando cuatro arquitecturas distintas y utilizando el algoritmo de transformación de wavelet discreta. En el proceso de aprendizaje, los modelos emplearon temperatura, humedad relativa, velocidad del viento y duración del sol para estimar la irradiancia horizontal global sobre Abu Dhabi, Emiratos Árabes Unidos. Los resultados mostraron una mejora

importante en el rendimiento de las ANN con un R^2 de 6.84% para el perceptrón multicapa, y un RMSE de 2.78% para redes neuronales de regresión generalizada.

II. METODOLOGÍA

En esta sección se definen las preguntas de interés que condujeron a la construcción del modelo para la estimación de la Radiación Solar utilizando técnicas de Aprendizaje Automático. También se describen las fuentes de información y la manera en cómo se procesaron los datos.

A. Preguntas de interés

Este trabajo se produce a partir de la necesidad existente en el aprovechamiento de fuentes de energía limpia, como el caso de la energía solar; por lo tanto, es trascendental conocer el comportamiento de los niveles de luminosidad solar para la implementación de sistemas fotovoltaicos en lugares específicos de la geografía colombiana. Sin embargo, las mediciones de brillo solar aún son inexistentes en muchas regiones; esto implica la construcción de modelos de Inteligencia Artificial basados en Aprendizaje Automático para predecir la radiación solar a partir de variables climáticas, pero, ¿esas variables son suficientes para realizar una buena estimación de radiación solar? ¿Cuáles son las técnicas con mejor desempeño en la predicción? Y, ¿cuál es el grado de confianza en dichas estimaciones? Estas interrogantes serán despejadas durante el desarrollo del documento.

B. Fuentes de información

En este estudio se emplearon conjuntos de datos públicos de las estaciones de medición de calidad del aire (Escuela República de Argentina, La Flora, Pance, Compartir) del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA) de la Alcaldía de Cali. También se utilizó un conjunto de datos de la estación Catumare, ubicada en la comuna ocho de la ciudad de Villavicencio, como parte del centro de control de Cormacarena, y un conjunto de datos generado por la API World Weather Online para la ciudad de Mocoa. Todos los conjuntos de datos públicos incluyen observaciones históricas de variables como velocidad y dirección

del viento, temperatura, humedad, lluvia y radiación solar, tomadas por hora durante un periodo de tiempo. En contraste, el conjunto de datos de Mocoa no contiene radiación solar. La Tabla 1 presenta los metadatos de cada conjunto de datos utilizado.

Tabla 1. Descripción de los conjuntos de datos.

Id	Estación	Periodo	Variables	Registros	Uso en el trabajo
1	Escuela República de Argentina	28/01/2018 01/01/2019	12	78241	Entrenamiento de los algoritmos
2	La Flora	06/05/2010 01/01/2019	12	75869	Evaluación del modelo
3	Pance	27/02/2013 01/01/2019	8	51216	
4	Compartir	30/01/2014 01/01/2019	9	43118	
5	Catumare	25/06/2018 31/12/2018	8	4455	
6	Mocoa	01/01/2010 31/12/2019	25	87648	

Se definió la radiación solar como variable objetivo; el mes, la hora, la velocidad y dirección del viento, la temperatura, la humedad y la lluvia se definieron como variables independientes a partir de las cuales se producirán las estimaciones de irradiancia. Considerando que en todos los conjuntos de datos utilizados se conocen los valores de radiación solar, excepto en el conjunto de datos de Mocoa, es preciso subrayar que se seleccionaron técnicas de aprendizaje supervisado como uno de los tipos de Aprendizaje Automático para regresión y clasificación.

C. Algoritmos de aprendizaje automático

En la regresión se emplearon árboles de regresión, regresión lineal múltiple, regresión de soporte vectorial, redes neuronales, Random Forest y Gradient Boosting; para la clasificación se utilizaron árboles de decisión, máquinas de soporte vectorial y AdaBoost. En cuanto a las redes neuronales, se utilizó el perceptrón multicapa para regresión y se probaron diferentes configuraciones, determinando como óptimos los siguientes hiperparámetros: 4 capas ocultas de 100, 75, 50 y 25 neuronas; función de activación ReLU; algoritmo de optimización Adam; término de regularización L2 de 0.00001; tasa de aprendizaje adaptativa, y un número máximo

de 10000 iteraciones. Se utilizó regresión lineal simple para evaluar la correlación entre las variables independientes y la variable objetivo, determinando que la temperatura tiene más reciprocidad con respecto a la radiación solar.

D. Filtrado de datos

El conjunto de datos de la estación Escuela República de Argentina fue utilizado para el entrenamiento del modelo debido a que se encuentra ubicada en el centro de la ciudad de Cali; por ende, se convierte en punto de referencia con respecto a las demás estaciones, las cuales presentan variaciones en las mediciones debido a su ubicación (distancia que las separa). Además, esta estación cuenta con la mayor cantidad de registros. Según datos obtenidos de Google Maps, la estación Escuela República de Argentina está a 3.74 km en línea recta de la estación La Flora, en el norte de la ciudad; a 18.45 km de la estación Pance, en el sur de la ciudad en la zona rural, y a 6.98 km de la estación Compartir, ubicada en el oriente de la ciudad. La Figura 1 muestra la ubicación de las estaciones con respecto a la estación de referencia. El conjunto de datos de la estación Catumare y el de la ciudad de Mocoa son importantes para este estudio debido a que cuentan con características climáticas diferentes, lo que permite evaluar la precisión del modelo en la estimación de la luminosidad proyectada por los rayos solares en otras regiones del país.

Es importante mencionar que los conjuntos de datos presentan discontinuidad (datos faltantes o espacios en blanco) debido a las anomalías que se dan en las estaciones de monitoreo tales como: fallas en los equipos, falta de energía eléctrica en la zona, hurto de equipos o cableado, mantenimiento o cambio de equipos, entre otros. También está la inclusión o exclusión de algunos contaminantes o variables meteorológicas según criterio de los expertos y características de la zona a monitorear.

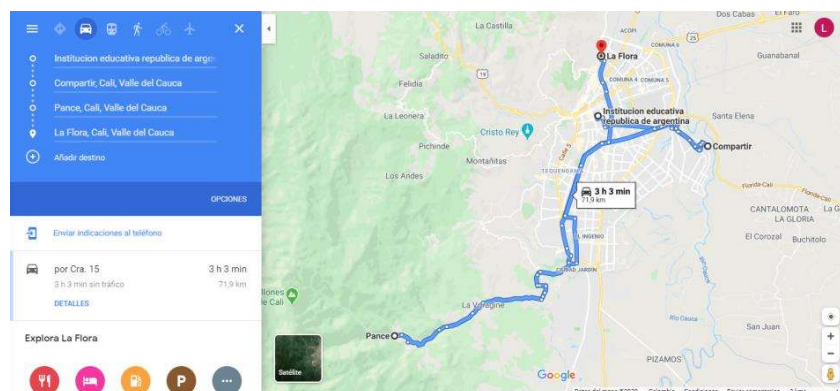


Fig 1. Ubicación de las estaciones de medición. Tomada de Google Maps.

El preprocesamiento de los datos se inicia con la exclusión de las variables de calidad del aire, debido a que no todas las estaciones de monitoreo del clima incluyen estas características, y considerando que los niveles de contaminación del aire en entornos distintos a las ciudades principales son muy bajos o incluso nulos. Además, este estudio pretende estimar la energía luminosa proveniente del sol a partir de datos meteorológicos. Posteriormente, se divide la variable FechaHora en las columnas mes y hora para analizar el comportamiento de la luz solar a lo largo de un día o de un año; por otro lado, se descarta el año y el día de la medición realizada, ya que no se establecen como esenciales para el presente estudio.

Con respecto a la intermitencia en los datos a causa de las razones antes mencionadas, los valores faltantes fueron completados con el promedio entre el valor anterior y el siguiente en cada columna, después se eliminaron aquellos registros que aún contenían valores nulos. Por último, en el caso de la regresión, los datos originales se normalizaron con el método Min-Max, transformando los valores en un rango entre cero y uno. La Tabla 2 muestra la cantidad de registros restantes luego de la eliminación de las observaciones incompletas.

En el modelo de clasificación se transformó la variable radiación solar en las clases 0 y 1; 0 para las observaciones inferiores a 300 Watts/m² y 1 para las demás. El valor definido se tomó como punto de referencia para la transformación de la luz solar en energía eléctrica, considerando la cantidad de paneles y el tamaño de la superficie necesaria para la instalación de sistemas fotovoltaicos. El conjunto de

datos quedó dividido en 26.6% en 1 y 73.4% en 0, dado que, cuando el sol está oculto, la irradiancia es cero, y en las primeras y las últimas horas del día, la radiación solar es baja. Debido a que es recomendable trabajar con datos balanceados, en la clasificación se utilizó la técnica de submuestreo Random Under-sampling para ajustar la distribución de las clases. Así se buscó evitar el *sobreentrenamiento* del algoritmo que conduce al sobreajuste de los datos, quedando finalmente con un conjunto de datos de 9,045 observaciones en cada clase, para un total de 18,090 registros.

Tabla 2. Registros luego de la eliminación.

Id	Estación	Registros antes	Registros después	Pérdida de datos (%)
1	Escuela República de Argentina	78241	33951	56.61
2	La Flora	75869	37411	50.69
3	Pance	51216	304	99.41
4	Compartir	43118	31896	26.03
5	Catumare	4455	4455	0.00
6	Mocoa	87648	87648	0.00

III. RESULTADOS

Posterior al preprocesamiento de los datos, se realizó un re-muestreo en el conjunto de datos de la Escuela República de Argentina, con un 70% para entrenamiento y el 30% restante para evaluar el rendimiento de los algoritmos empleados en el estudio. Luego, se aplicaron distintas técnicas de aprendizaje supervisado para la estimación de la radiación solar. Inmediatamente después del entrenamiento y valoración de los algoritmos, se aplicó la técnica con mayor rendimiento a los datos de prueba de las demás estaciones, para evaluar el desempeño del modelo mediante la comparación entre las predicciones realizadas y las observaciones reales registradas en cada conjunto de datos.

A. Arquitectura

Este apartado presenta el tratamiento aplicado a todos los conjuntos de datos empleados por los diferentes algoritmos de aprendizaje automático, con el propósito de realizar las estimaciones de energía solar mediante regresión y clasificación. La

Figura 2 expone el flujo de la información desde las fuentes hasta la generación de las predicciones en cada una de las etapas del modelo.

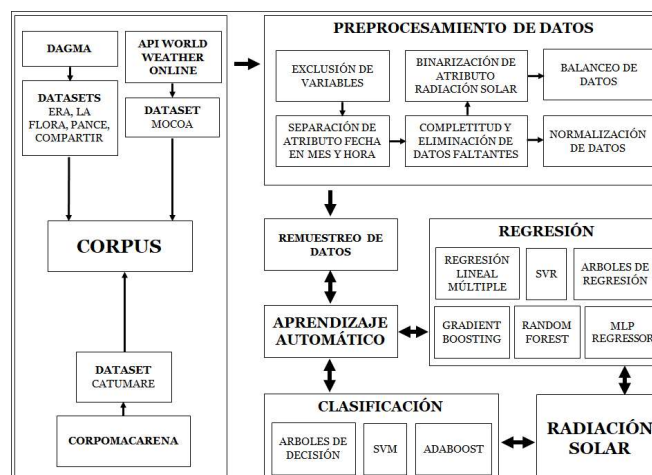


Fig. 2. Flujo de información del modelo.

B. Regresión

En la Tabla 3 se observa que con el algoritmo MLP Regressor se obtuvieron los mejores resultados, según el coeficiente de determinación R^2 en los datos de desarrollo y en la raíz del error cuadrático medio RMSE.

Tabla 3. Resultados obtenidos mediante regresión.

Técnica	R^2 -score entrenamiento	R^2 -score desarrollo	RMSE
Regresión lineal múltiple	0.56	0.55	163.61
Máquinas de Soporte Vectorial para Regresión (SVR)	0.77	0.77	117.73
Árboles de regresión	0.90	0.89	81.66
Gradient Boosting	0.98	0.89	82.26
Random Forest	0.98	0.89	79.99
MLP Regressor	0.90	0.90	77.37

C. Clasificación

La Tabla 4 presenta los resultados obtenidos con los algoritmos de clasificación, resaltando que el método de ensamble AdaBoost tuvo mejor desempeño según las métricas de la matriz de confusión (Accuracy, Precision, Recall y F1-score), tanto para los datos de entrenamiento como para los datos de desarrollo. En contraste,

las máquinas de soporte vectorial superaron al algoritmo AdaBoost en Accuracy en los datos de entrenamiento y en Precisión en los datos de desarrollo.

Tabla 4. Resultados obtenidos mediante clasificación.

Técnica	Entrenamiento	Desarrollo			
	Accuracy	Accuracy	Precisión	Recall	F1-score
Árboles de decisión	0.93	0.93	0.89	0.98	0.93
Máquinas de Soporte Vectorial para Clasificación (SVM)	0.98	0.90	0.91	0.88	0.90
AdaBoost	0.94	0.94	0.90	0.99	0.94

D. Evaluación del modelo

En las Tablas 5 y 6 se pueden apreciar los resultados obtenidos en las predicciones mediante las técnicas con mejor desempeño: Random Forest, MLP Regressor y AdaBoost.

Tabla 5. Resultados obtenidos con Random Forest y MLP Regressor.

Conjunto de datos	Cantidad de datos	Random Forest		MLP Regressor	
		R ² -score	RMSE	R ² -score	RMSE
La Flora	37411	0.88	79.19	0.88	59.84
Pance	304	0.84	51.59	0.82	55.24
Compartir	31896	0.37	114.22	0.44	82.95
Catumare	4455	0.25	154.70	0.74	90.40

Tabla 6. Resultados obtenidos con la técnica de clasificación AdaBoost.

Conjunto de datos	Datos balanceados	Accuracy	Precisión	Recall	F1-score
La Flora	18550	0.94	0.90	0.99	0.94
Pance	66	0.99	0.97	1.00	0.99
Compartir	8902	0.87	0.79	1.00	0.88
Catumare	1532	0.88	0.84	0.95	0.89
Mocoa	63244	0.81	1.00	0.62	0.77

Las Figuras 3 y 4 muestran los diagramas de dispersión obtenidos por la regresión, a partir de la radiación solar real, y estimada en los conjuntos de datos de las estaciones: La Flora, Pance, Compartir y Catumare. Para el conjunto de datos de Mocoa no fue posible evaluar el modelo de regresión, ni tampoco construir la gráfica, dado que no incluye valores de radiación solar.

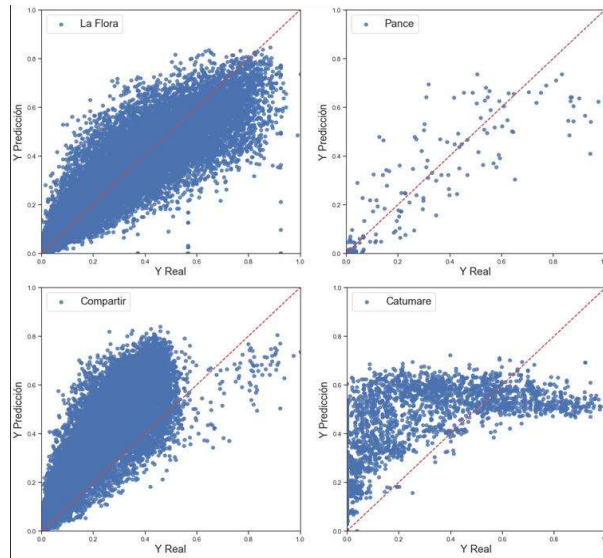


Fig. 3. Radiación solar real y estimada con Random Forest.

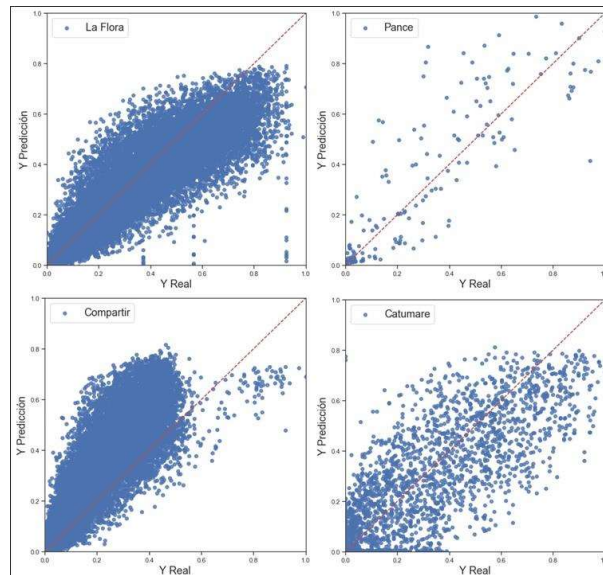


Fig. 4. Radiación solar real y estimada con MLP Regressor.

E. Estimación de radiación solar en la ciudad de Mocoa

La Figura 5 muestra el comportamiento de la radiación solar estimada para la ciudad de Mocoa, la Figura 6 presenta el promedio mensual de radiación solar entre las 6 am y las 4 pm, en el periodo comprendido entre los años 2010 y 2019. Finalmente,

la Figura 7 permite la visualización de las predicciones referentes a la cantidad de horas mensuales con valores de radiación solar mayores o iguales a 300 W/m².

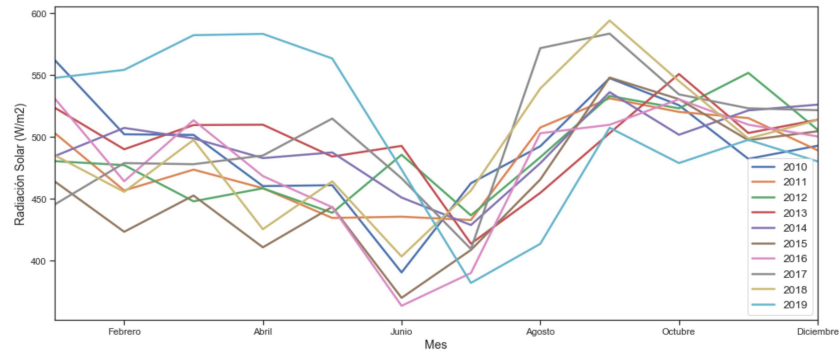


Fig. 5. Radiación solar promedio mensual por año.

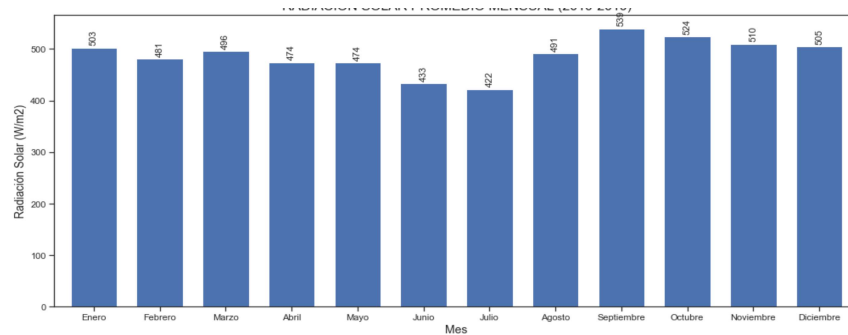


Fig. 6. Radiación solar promedio mensual entre 2010 y 2019.

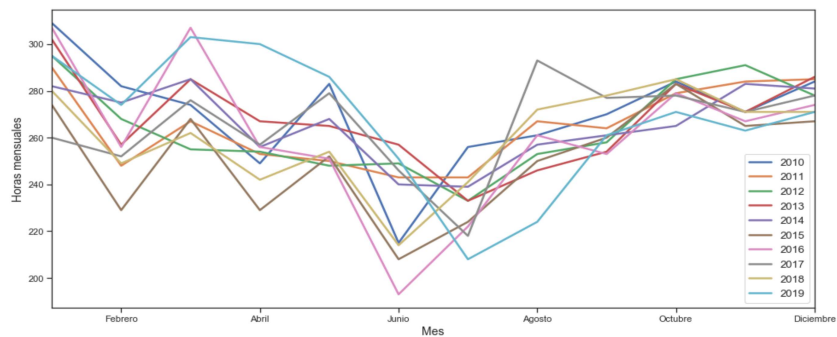


Fig. 7. Horas mensuales con radiación solar mayor o igual a 300 W/m².

IV. DISCUSIÓN

El objetivo de los algoritmos de aprendizaje automático consiste en permitir que las computadoras aprendan. En la actualidad, esta tendencia se ha incrementado en gran medida y es posible observarla en distintos campos como el de la generación de energía; por ende, este estudio ha empleado varias técnicas de aprendizaje supervisado, las cuales han mostrado buen desempeño. Por otro lado, es importante destacar que muchos de los artículos revisados usan redes neuronales artificiales para predecir la energía luminosa que se desprende del sol; por consiguiente, se recomienda continuar este trabajo con la construcción de un modelo robusto, basado en alguno de los distintos marcos de trabajo de aprendizaje profundo, para estimar la irradiancia en cualquier lugar del planeta.

Un modelo de aprendizaje profundo implica la adquisición de muchos más conjuntos de datos de diferentes localizaciones geográficas para incrementar, aún más, la confianza en las estimaciones históricas y el pronóstico del comportamiento futuro en semanas, meses e incluso años. Por otra parte, en este estudio no se tuvieron en cuenta algunos factores que determinan los diferentes tipos de climas regionales como la latitud, la altitud, el relieve, etc., por ello, se considera de suma importancia una investigación más profunda sobre estos aspectos.

Es importante destacar que a todos los conjuntos de datos se les dio el mismo tratamiento en el preprocesamiento de los datos, a pesar de ello, se esperaba un mejor desempeño en la evaluación del modelo para el caso del conjunto de datos de la estación Compartir, dado que su ubicación se encuentra a menos de 7 km de la estación Escuela República de Argentina y ambas se encuentran dentro de la misma ciudad. En el caso de la estación Catumare, también se observa bastante variabilidad en las estimaciones, aunque era algo que se esperaba, debido a que se encuentra situada en otra región a menor altura sobre el nivel del mar.

V. CONCLUSIONES

Comprender el comportamiento de la Radiación Solar en una región determinada es indispensable para la construcción de parques solares correctamente

dimensionados, de modo que conduzcan al máximo aprovechamiento de la energía solar. En este trabajo se logró obtener un buen desempeño en la estimación de la luminosidad del sol a partir las características meteorológicas disponibles en los conjuntos de datos empleados. En el caso del modelo de regresión, la temperatura es la variable que tiene mayor correlación con la radiación solar, mientras que, en el modelo de clasificación, el atributo más importante es la hora del día.

Aunque en la actualidad hay una gran demanda en el uso del aprendizaje profundo para realizar predicciones en diferentes áreas, existe una amplia variedad de técnicas de aprendizaje automático con alto desempeño; por consiguiente, los métodos de ensamble combinan grupos de modelos para incrementar el desempeño en las predicciones. En este trabajo, los mejores resultados para el caso de la regresión se obtuvieron con los algoritmos MLP Regressor, con un coeficiente de determinación R^2 de 0,90 y un RMSE de 77.37, y Random Forest, con un coeficiente de determinación R^2 de 0,89 y un RMSE de 79.99. Por otra parte, en la clasificación, la técnica AdaBoost tuvo mejor desempeño, aun así, las máquinas de soporte vectorial mostraron un rendimiento superior en algunas de las métricas de la matriz de confusión.

Esta investigación propone una evaluación conceptual en la instalación de sistemas fotovoltaicos, a partir de la comprensión de los valores de radiación solar estimados por modelos de inteligencia artificial, como apoyo a las personas encargadas de tomar decisiones. Para el caso de Mocoa, podemos decir que se produce, en promedio, una radiación solar anual de 487.67 W/m² y de 6 a 10 horas diarias con radiación solar mayor a 300 W/m², contemplando un margen de error de entre el 10% y el 60% en las estimaciones, basados en los modelos que utilizan conjuntos de datos donde sí existe información de energía solar para confrontarlos.

CONTRIBUCIÓN DE LOS AUTORES

Luis-Eduardo Ordoñez-Palacios: Conceptualización, Análisis Formal, Investigación, Metodología, Escritura - borrador original, Escritura - revisión y edición.

Daniel-Andrés León-Vargas: Análisis Formal, Investigación, Escritura - revisión y edición.

Víctor-Andrés Bucheli-Guerrero: Análisis Formal, Supervisión, Validación, Escritura - revisión y edición.

Hugo-Armando Ordoñez-Eraso: Análisis Formal, Supervisión, Validación, Escritura - revisión y edición.

REFERENCIAS

- [1] A. Doval Adán, "Los grandes problemas geopolíticos del desarrollo mundial: hacia una planificación global del planeta," in *XVIII Congreso de Geógrafos Españoles* 2003. <https://minerva.usc.es/xmlui/handle/10347/20719>
- [2] Z. Li, "Global Warming: Causes and Effects," in *Southern California Conferences for Undergraduate Research*, 2019. https://www.sccur.org/sccur/fall_2019_conference/poster_session_4/131
- [3] M. R. Gámez, A. V. Pérez, A. M. V. Quiroz, and W. M. S. Arauz, "Mejora de la calidad de la energía con sistemas fotovoltaicos en las zonas rurales," *Revista Científica*, vol. 3 (33), pp. 265-274, 2018. <https://doi.org/10.14483/23448350.13104>
- [4] F. G. Cozman, "O futuro da (pesquisa em) inteligência artificial: algumas direções," *Revista USP*, n. 124, pp. 11-20, 2020. <https://doi.org/10.11606/issn.2316-9036.v0i124p11-20>
- [5] H. Ordóñez, C. Cobos, and V. Bucheli, "Modelo de machine learning para la predicción de las tendencias de hurto en Colombia", *Revista Ibérica de Sistemas y Tecnologías de la Información*, vol. 29, pp. 494-506, 2020.
- [6] M. Haenlein, and A. Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence," *California Management Review*, vol. 61 (4), pp. 5-14, 2019. <https://doi.org/10.1177/0008125619864925>
- [7] N. Premalatha, and A. Arasu, "Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms", *Journal of Applied Research and Technology*, vol. 14 (3), 2016. <https://doi.org/10.1016/j.jart.2016.05.001>
- [8] A. Qazi, H. Fayaz, A. Wadi, R. G. Raj, N. A. Rahim, and W. A. Khan, "The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review," *Journal of Cleaner Production*, vol. 104, pp. 1-12, 2015. <https://doi.org/10.1016/j.jclepro.2015.04.041>
- [9] A. K. Yadav, and S. S. Chandel, "Solar radiation prediction using Artificial Neural Network techniques: A review," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 772-781, 2014. <https://doi.org/10.1016/j.rser.2013.08.055>
- [10] M. Ozgoren, M. Bilgili, and B. Sahin, "Estimation of global solar radiation using ANN over Turkey," *Expert Systems with Applications*, vol. 39 (5), pp. 5043-5051, 2012. <https://doi.org/10.1016/j.eswa.2011.11.036>
- [11] A. Koca, H. F. Oztup, Y. Varol, and G. O. Koca, "Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey," *Expert Systems with Applications*, vol. 38 (7), pp. 8756-8762, 2011. <https://doi.org/10.1016/j.eswa.2011.01.085>
- [12] A. Rahimikhoob, "Estimating global solar radiation using artificial neural network and air temperature data

- in a semi-arid environment," *Renewable Energy*, vol. 35 (9), pp. 2131-2135, 2010.
<https://doi.org/10.1016/j.renene.2010.01.029>
- [13] D. Li, W. Chen, S. Li, and S. Lou, "Estimation of Hourly Global Solar Radiation Using Multivariate Adaptive Regression Spline (MARS) – A Case Study of Hong Kong," *Energy*, vol. 186, e115857, 2019.
<https://doi.org/10.1016/j.energy.2019.115857>
- [14] J. Liu, M. Y. Cao, D. Bai, and R. Zhang, "Solar radiation prediction based on random forest of feature-extraction», *IOP Conference Series: Materials Science Engineering*, vol. 658, e012006, 2019.
<https://doi.org/10.1088/1757-899X/658/1/012006>
- [15] S. Li, H. Ma, and W. Li, "Typical solar radiation year construction using k-means clustering and discrete-time Markov chain," *Applied Energy*, vol. 205, pp. 720-731, 2017.
<https://doi.org/10.1016/j.apenergy.2017.08.067>
- [16] R. Meenal, and A. I. Selvakumar, "Assessment of SVM, Empirical and ANN based solar radiation prediction models with most influencing input parameters," *Renewable Energy*, vol. 121, pp. 324-343, 2017.
<https://doi.org/10.1016/j.renene.2017.12.005>
- [17] M. Lazzaroni, S. Ferrari, V. Piuri, A. Salman, L. Cristaldi, and M. Faifer, "Models for solar radiation prediction based on different measurement sites," *Measurement*, vol. 63, pp. 346-363, 2015.
<https://doi.org/10.1016/j.measurement.2014.11.037>
- [18] E. S. Mostafavi, S. M. Mousavi, and P. Jiao, "Next Generation Prediction Model for Daily Solar Radiation on Horizontal Surface Using a Hybrid Neural Network and Simulated Annealing Method," *Energy conversion and management*, 2017. <http://agris.fao.org/agris-search/search.do?recordID=US201800045628>
- [19] H. Sharadga, S. Hajimirza, and R. S. Balog, "Time series forecasting of solar power generation for large-scale photovoltaic plants," *Renewable Energy*, vol. 150, pp. 797-807, 2020.
<https://doi.org/10.1016/j.renene.2019.12.131>
- [20] A. Rabehi, M. Guermoui, and D. Lalmi, "Hybrid models for global solar radiation prediction: a case study», *International Journal of Ambient Energy*, vol. 41 (1), pp. 31-40, 2020.
<https://doi.org/10.1080/01430750.2018.1443498>
- [21] C. Koo, W. Li, S. H. Cha, and S. Zhang, "A novel estimation approach for the solar radiation potential with its complex spatial pattern via machine-learning techniques," *Renewable Energy*, vol. 133, pp. 575-592, 2019. <https://doi.org/10.1016/j.renene.2018.10.066>
- [22] S. Hussain, and A. Alaili, "A hybrid solar radiation modeling approach using wavelet multiresolution analysis and artificial neural networks," *Applied Energy*, vol. 208, pp. 540-550, 2017.
<https://doi.org/10.1016/j.apenergy.2017.09.100>