Zoning of Ecuador According to Maximum Magnitudes of Earthquakes and their Frequency of Occurrence using Statistical Models Estimated by Maximum Likelihood

Sandra GARCIA-BUSTOS[1,*] (iD) , Steven NAVARRETE[2] (iD) , Ángel CHANCAY[1] (iD) , Marcos MENDOZA[1] (iD) ,

Marcela PINCAY [3] (iD) , Miguel TERAN [3] (iD)

[1] Faculty of Natural Sciences and Mathematics, Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Km 30.5 Vía Perimetral, Guayaquil, Ecuador.
[2] Faculty of Engineering in Earth Sciences, Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Km 30.5 Vía Perimetral, Guayaquil, Ecuador.
[3] Universidad Estatal del Sur de Manabí, Campus Los Ángeles, Jipijapa Manabí, Ecuador.

Highlights
• This work presents models to estimate the maximum magnitudes of earthquakes and their frequency.
• A zoning considering maximum magnitudes through clustering methods was carried out.
• The zoning obtained has an important relationship with the geological nature of the active faults.

Article Info	Abstract
	This work applied the extreme value theory and Poisson regression considering maximum likelihood estimates to characterize the 5 seismicity zones of Ecuador obtained through cluster analysis. The Gumbel model was optimal in describing maximum magnitudes in these areas. The Poisson regression model was used to estimate the frequency of earthquakes as a function of magnitude (MW) for each zone. Zones 1 (portions of Esmeraldas and Manabí and their respective maritime zones) and 5 (north central of Ecuador) were the most representative in terms of recurrence and seismic magnitudes.

1. INTRODUCTION

The presence of important seismogenic structures, such as the subduction zone between the plates of Nazca and South America and the interaction between the South American continental plate and the NorAndino block define Ecuador as an area of important seismic activity. Several studies on the kinematics of plates and seismic cycles in Ecuador have indicated the presence of mechanical couplings in the subduction zone and in the NorAndino block [1-4]. These types of couplings lead to the accumulation of shearing stresses that, after a certain threshold, cause sliding, which results in earthquakes [1].

Strong earthquakes—such as the one in Esmeraldas in 1906 of 8.8 M_W [5] and the most recent event in Pedernales in 2016, of 7.8 Mw [6] and the associated serious social and economic consequences justify the need to present models of seismic magnitude, recurrence frequency, and zonal categorization.

The mathematical models applied to seismology have established relationships on seismic hazards in each region of the country in order to represent seismic hazard maps in terms of peak ground acceleration (PGA) and spectrum acceleration (SA). The Ecuadorian construction code [7] published its first seismic hazard maps in 2001, classifying the region into 4 zones on the basis of PGA, for a return period of 475 years. In

*Corresponding author, e-mail: slgarcia@espol.edu.ec

contrast, current construction regulations in Ecuador, the NEC-15, established 6 seismic hazard zones on the basis of PGA. These zones were determined using the following methods: evaluation and re-evaluation of Bakun and Wentworth magnitudes [8], analysis of homogeneity and completeness, modelling of seismic events and of earthquake occurrence, among others. [9]. The results of the seismic hazard investigations by [10] are similar to those presented by the Ecuadorian Standard in terms of rock acceleration; however, they differ in the southern part of the country [10].

On the other hand, the seismicity of a place can be characterized using probabilistic models such as the study carried out by [11]. They developed seismic hazard maps using probabilistic models of PGA, differing from [9] and [10] in their zoning.

Statistics provides some estimation and analysis techniques to characterize variables of a population. The earthquakes that cause greater problems to a population are those that have greater intensities and magnitudes. For this reason, this study will analyze those largest magnitudes that occurred in Ecuador. The categorization of earthquakes in the region and the data collected by the Military Geographical Institute (IGM) since 1906 allow the construction of statistical models of Ecuador's earthquake magnitudes.

One of the most commonly used techniques for model estimations is the maximum likelihood method, because asymptotically normal, efficient, and consistent estimates are obtained [12]. The objective of this study is to construct a non-relational statistical model estimated by means of the maximum likelihood method, utilizing the theory of extreme values for estimating the probability of occurrence of maximum-magnitude earthquakes in continental Ecuador through two approaches: block maxima (BM) and peaks over threshold (POT). K-means and Ward clustering methods were also applied to construct an Ecuador zoning according to seismic characteristics (maximum magnitudes, recurrence of high magnitudes, numbers of earthquakes).

To characterize each of these seismicity zones, extreme value theory models were estimated to understand the behaviour of the maximum seismic magnitudes. Poisson regression models obtained by maximum likelihood methodology were also used to estimate earthquake counts as a function of M_W (scale magnitude) intervals for each of these zones. R software was used for the implementation of these methodologies [13].

In this investigation, the characteristics of the soil where the earthquake occurred were not considered in the analysis, nor was any attempt made to quantify the damage that would be caused by the earthquake or the rupture length of the earthquakes. However, the results of the statistical analysis, with concepts and geological information of the region, were addressed to present subsequent conclusions and evaluations.

2. METHODOLOGY AND MATERIALS

2.1. Seismic Catalogue

Seismic catalogues are a fundamental component of a seismic hazard analysis, since they are required to model the magnitudes, frequencies, and probabilistic distributions within each analysis zone [14]. The seismic catalogues used in this research were instrumental records and integration with historical records. The data considered in this study were provided by the Geophysical Institute of Ecuador [15] and were homogenized on the M_W scale. The database contained a total of 12,531 records of earthquakes between 1906 and 2017 occurring in the seismogenic structures of the subduction zone between the Nazca plate and the South American plate, and the interplate with the NorAndino block. This information was verified with the earthquake's information available on the Incorporated Research Institutions for Seismology of Washington, DC website [16].

2.2. Tectonic Region

In general, tectonic structures on Ecuador are complex, it is formed in 4 main structures [17]:

Continental platform: Actives continental subduction margins located among Tiputini shield and Paleo-Cretaceus subduction system on real cordillera, composed of accreted Precambrian-Paleozoic belt. Marine Continental Crust: Formed by Triassic, Jurassic, Cretaceous continental belts and accreted oceanic slabs. Late Accreted Ocean Crust: Located in much of the western cordillera, extending to the current location of subduction trench. Ocean Crust: Located in the Galapagos islands

Within these components are the following seismogenic sources:

The main seismogenic source in the region is the subduction zone located in the Pacific Ocean at a distance between 60 and 150 km from the coastline of Continental Ecuador. This zone has 756 kilometers considering the subduction geodynamics from the north coast of Peru to the southern Andean part of territory [18].

Parra suggests differentiating in two subduction zones, interphase and in-slab zones. The difference is the location and the restriction of movement due to the Nazca plate and the angle of immersion of the oceanic plate roof is considerably less in contrast to in-slab zone, producing an accumulation of tension [17]

The seismic recurrence for this seismogenic structure is every 20 years, of which two have been very strong. The estimation of the maximum magnitude considering the 280 kilometers of structural length failure could reach earthquakes in the order of 8 to 8.1 Magnitude [18].

Another seismogenic source is greater dextral system that separates the North-Andean Block from the South American continental plate. This zone is a collision strip made up of crustal geological faults with different trends and directions [19]. Generating cortical earthquakes of high to moderate intensity [18].

2.3. Extreme Values Theory

The extreme values theory is used in statistical analysis of rare events and has been applied in several fields of research, including neurology, hydrology, economics, and actuarial and earth sciences. This theory has also been used in the analysis of the greatest earthquakes [20, 21]. More recent works have used it for the estimation of the upper limit of Japan's seismic hazard curve [22]. [23] used this theory to estimate the distributions of the maximum magnitudes of earthquakes for the coast of Ecuador, based on statistics from the 1906–2016 period. Zones were also categorized according to the seismic hazard, but from a purely statistical point of view.

In the next section, the two approaches of extreme values theory that were used in this work are described: Block Maxima and Peaks Over Threshold.

The Block Maxima approach is used to estimate the probability distribution of the maximum value of a random variable, in this case, of earthquakes. To this end, the maximum earthquakes per period, or block, must be determined, which for this study corresponded to one year.

In this approach, the Fisher–Tippett–Gnedenko theorem must be considered, which proposes the following:

$$P\left(\frac{M_n - b_n}{a_n} \le z\right) \to G(z), \; when \; n \to \infty \tag{1}$$

where $M_n = \max(X_1, X_2, \dots, X_n)$, $X_1, X_2, \dots, X_n$ are independent and identically distributed variables, and the sequences of constants a_n and $b_n > 0$. This theorem indicates that if the maximum of a random variable is appropriately normalized, that is, a trend parameter is subtracted and divided by a dispersion parameter, its distribution function will adjust to $G(z)$ when n tends to infinity, and can be one of the following distributions [24-26]:

Cumulative Gumbel distribution [27, 28]:

$$G(z) = e^{-e^{\frac{-(z-\mu)}{\sigma}}} \qquad\qquad z \in (-\infty, +\infty) \qquad\qquad (2)$$

Cumulative Fréchet distribution [24]:

$$G(z) = e^{-\left(\frac{z-\mu}{\sigma}\right)^{-\xi}}, \; if \; z > \mu \qquad\qquad (3)$$

Weibull probability distribution [29, 24]:

$$G(z) = \begin{cases} 0, & if \; z < \mu \\ 1 - e^{-\left(\frac{z-\mu}{\sigma}\right)^{\xi}}, & if \; z \geq \mu \end{cases} \qquad\qquad (4)$$

where: μ is a position/location parameter; σ is a dispersion parameter and ξ is a shape parameter.

[30] and [31] proposed a single expression (5) to generalize the three distributions of Gumbel, Fréchet, and Weibull. This expression is called the generalized extreme-value (GEV) distribution and depends on the shape parameter (ξ), so that if ξ> 0, the Fréchet distribution is used; in contrast, if ξ = 0, the Gumbel distribution is chosen, and if ξ <0, the Weibull distribution is applied

$$G(z) = \begin{cases} e^{-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}}, & if \; \xi \neq 0 \\ e^{-e^{\frac{-(z-\mu)}{\sigma}}}, & if \; \xi = 0 \end{cases} \qquad\qquad (5)$$

The domain of the function is: $\left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right) > 0$, if $\xi \neq 0$. If $\xi = 0$, then x $\in (-\infty, +\infty)$.

An alternative approach used in application of the extreme value theory is the POT model [32], where it is necessary to set a threshold u and any value that exceeds this threshold is considered an extreme value. The objective is to estimate the probability distribution of these extreme values.

The key requirement is the selection of the threshold; it is necessary to choose the highest possible value yet ensure that there is still a large number of observations that exceed that threshold. These conditions are explained in the Pickands-Balkema-De Hann Theorem. This theorem indicates that the limit distribution of probabilities for excesses over the threshold ($F_u(x)$) approximates the generalized Pareto (GP) distribution [33-35].

The GP distribution contains 3 fundamental parameters: central tendency parameter (μ), a scale parameter (σ), and a shape parameter (ξ), and is defined by:

$$F(x) = \begin{cases} 1 - \left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-\frac{1}{\xi}}, & if \; \xi \neq 0 \\ 1 - e^{-\left(\frac{x-\mu}{\sigma}\right)}, & if \; \xi = 0 \end{cases} \qquad\qquad (6)$$

for $x \geq \mu$, when $\xi \geq 0$, and $\mu \leq x \leq \mu - \frac{\sigma}{\xi}$ when $\xi < 0$.

If ξ = 0, the asymptotic distribution of threshold excesses is exponential.

2.4. Poisson Regression

Generalized linear models (GLM) use the maximum likelihood method to estimate the coefficients of a model that relates a response variable (which is part of the exponential family of distributions), with one or several explanatory variables. If the response variable is of the counting type, the Poisson distribution is

commonly assumed, which is also part of the exponential family of distributions. Then, the Poisson regression that is a GLM, is the methodology that would be used to estimate the average μ_i counts of the Poisson variable. The following model is commonly used for Poisson regression:

$$log(\mu_i) = \alpha + \beta x_i + \varepsilon_i \tag{7}$$

where x_i is the independent variable, α and β are the coefficients of the model, and ε_i are the model errors. More information about Poisson regression and GLM can be found in [36].

The model (7) can also be estimated by least squares, except in the case when μ_i is equal to 0 and it is simply a classical linear regression model, although the maximum likelihood methodology used in GLM can be applied even when μ_i is equal to 0.

2.5. Maximum Likelihood Estimation Method

Likelihood in statistical language means the probability that a sample fits a distribution model with specific parameters. If the probability associated with this adjustment is greater, then the behavior of the sample resembles the given function.

This method requires assuming a probability distribution so that given a sample, the values of the maximum likelihood estimators for the probability distribution parameters assumed are those that maximize the probability of occurrence of the sample. It is commonly used for parameter estimation of a model [36, 37].

When models are constructed with this method, the likelihoods of these models are usually compared to find the one that fits the data best, through hypothesis tests that use a quotient of those likelihoods as a statistic. The preferred models are those that significantly improve the likelihood or that are simpler (with fewer explanatory variables) but with a similar likelihood to another more complex model that is being compared in the hypothesis test. Thus:

H_0 = Model 1 is adequate
H_1 = Model 2 is adequate

where model 1 is a shorter model with fewer explanatory variables, while model 2 is longer. The decision on this contrast is based on the p-value. If the p-value is greater than 0.05, H_0 is not usually rejected, otherwise H_0 would be rejected.

Another criterion based on likelihood considers the values of the Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC) statistics [38, 39]. These are defined as follows:

$$\mathbf{AIC} = 2k - 2\ln(L) \tag{8}$$

$$\mathbf{BIC} = -2\ln(L) + kln(n) \tag{9}$$

where k is the number of parameters of the analyzed model, L is the maximum value of the likelihood function and n the size of the sample. Models with lower values of AIC and BIC will be preferred since they indicate greater likelihood.

2.6. Cluster Analysis

Cluster analysis is a mathematical method that allows classification of a complex amount of information into small groups or clusters, where the members in each cluster share characteristics similar to each other

and are different from those of other groups. To create these groups, it is necessary to use a measure of distance or similarity, and one of the most common is Euclidean distance [40], which was used in this work. For the calculation of these distances, it is common to work with standardized variables, that is, with a similar scale and zero mean, in order to eliminate the scale effect. The Euclidean distance between two vectors that represent observations of a sample where p variables are measured: $x = (x_1, x_2, \dots, x_p)'$ and $y = (y_1, y_2, \dots, y_p)'$ is defined as:

$$d(x, y) = \sqrt{(x - y)'\,(x - y)} = \sqrt{\Sigma_{j=1}^{p}(x_j - y_j)^2}. \tag{10}$$

Clustering algorithms can be classified into two large groups: hierarchical and non-hierarchical. Hierarchical methods consist of grouping clusters to form a new cluster, or separating an existing cluster to give rise to two new clusters in such a way that, if this process of cluster agglomeration or division is carried out successively, some distance is minimized or some similarity measure is maximized [40]. One of the most commonly used hierarchical methods is Ward's method, because it is more discriminative in determining group levels [41]. The method, first proposed in 1963 [42], creates groups such that the variance within each is minimal.

In non-hierarchical methods, a K number of groups must be set in advance and the groups are formed as a function of the proximity between their elements. The K-means is the most commonly used non-hierarchical method, due to its simplicity and efficiency [43]. In this method, each group is represented by a centroid, which is not more than the group mean. In this manner, each element is assigned to only one of those K groups, whose centroid is the closest. Further details of these methods can be found in [40]. In this work, the Ward and K-means clustering techniques were used.

2.7. Levels and Return Periods

The return levels are considered as the quantiles of order (1-p) of the extreme value distribution and are represented as Zp, where p is the probability that an earthquake of magnitude Zp will be exceeded once a year. In contrast, the return period, 1/p, is the number of units of time that will elapse on average for an earthquake greater than Zp to occur [44].

3. THE RESEARCH FINDINGS AND DISCUSSION

This section presents a brief description of the general characteristics of the earthquakes recorded in Ecuador during the period 1906–2017, in addition to the main results of the statistical models detailed in section 2.

3.1. Generalities of the Seismic Catalogue

The seismic catalogue used contains 12,531 records of earthquakes. Most of the Ecuadorian zone experiences significant seismic activity, with Sucumbíos, Orellana, Loja, El Oro, and Zamora Chinchipe registering less seismic activity than other provinces (Figure 1).

The depths of the earthquakes, along with their geographic coordinates, represent important variables to describe the geospatial location of geological faults. These depths vary throughout the region; however, in coastal areas, they are shallow [0 to 50 km] and 80% of recorded earthquakes are superficial in nature. The information presented in Figure 1 can also be shown as a function of low, moderate, and severe with respect to their seismic magnitudes (Table 1).
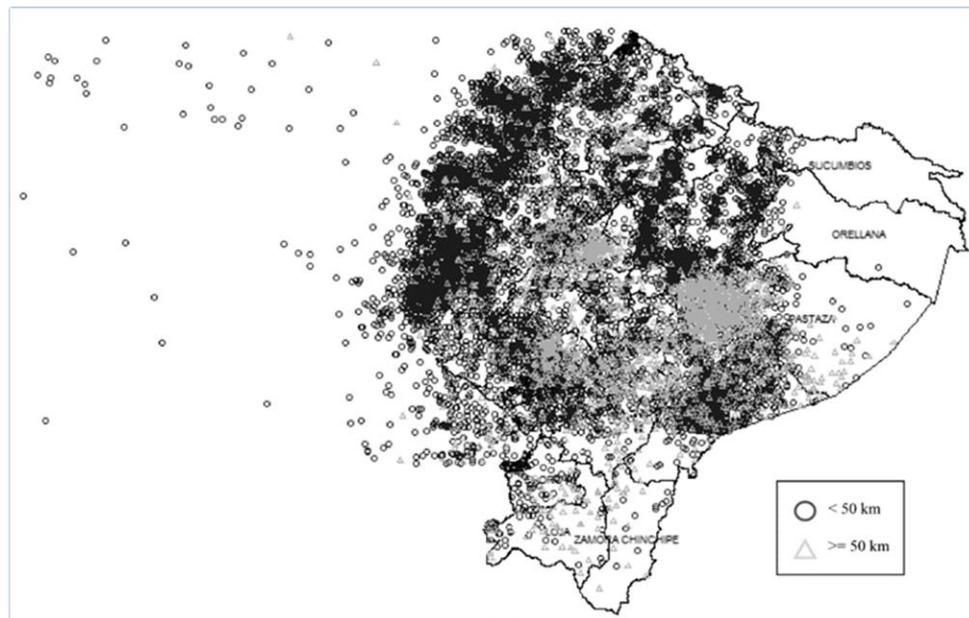
Figure 1. *A map of the locations and depths of earthquakes in Ecuador (1906–2017)*

Table 1. *Classification of Seismic Magnitudes*

Class	Magnitude (M_w)
Low	< 4.5
Moderate	4.5–5.5
Severe	> 5.5

This classification allows the separation of earthquakes associated with volcanic activities from those caused by geological faults. For example, as seen in Figure 2, earthquakes in volcanic regions are of low magnitude.
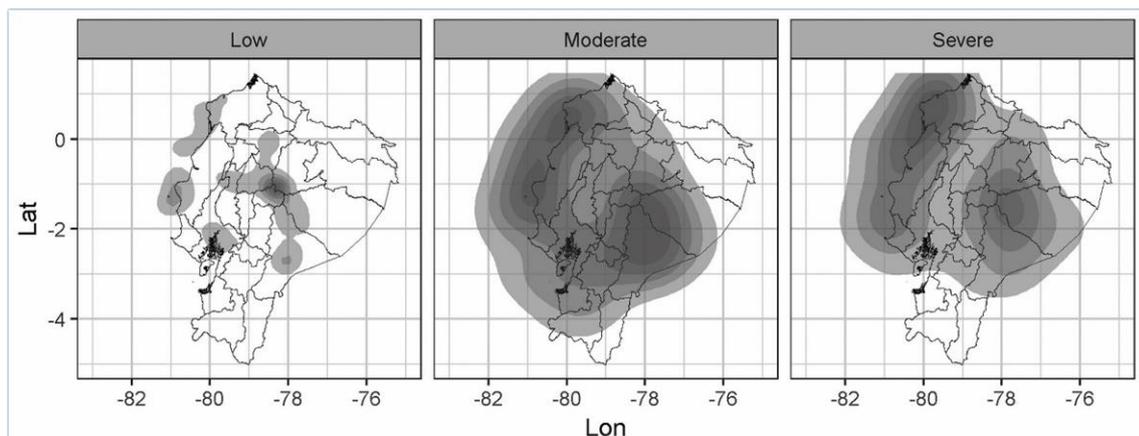
Figure 2. *Level Curves demonstrating the frequency of earthquakes in Ecuador by magnitude (1906–2017)*

Greater concentrations of moderate and severe earthquakes were recorded in the southwest of the province of Manabí, in its border area with Esmeraldas (Pedernales and Muisne cantons), and at the intersection of Morona Santiago and Pastaza. Consequently, the most active provinces where severe earthquakes occur are Manabí and Esmeraldas, on the coast, and Morona Santiago and Pastaza, in the east (Figure 2).

3.2. Distribution Models of Extreme Values Applied to Ecuador (Without Zoning)

The analyzed extreme value theory models are presented using two approaches: BM and POT.
The *BM method* estimates the probability distribution of the maximum magnitude of earthquakes recorded each year in Ecuador. The estimation of the GEV distribution parameters was performed through the maximum likelihood method using R's extremes library [13].

To analyze the fit of the models, the likelihood-ratio test and the AIC and BIC criteria were calculated. The probability distribution model that would adjust to the maximum annual earthquakes recorded in Ecuador was the Weibull, since the shape parameter was negative (Table 2). However, because the value of the shape parameter was close to zero, it may not be significant. Consequently, the Gumbel distribution was also considered as the possible distribution model, since it was characterized by a shape parameter of 0 (Table 3).

Table 2. GEV model applied to the maximum magnitudes in Ecuador

	Parameters		
	Location	**Scale**	**Shape**
Estimate	5. 74	0.72	-0.15
Standard error	0.08	0.06	0.06
AIC	266.42	**BIC**	274.58

Table 3. Gumbel model applied to the maximum magnitudes in Ecuador

	Parameters		
	Location		**Scale**
Estimate	5.68		0.69
Standard error	0.07		0.05
AIC	267.95	**BIC**	273.39

The Weibull model had a slightly lower AIC value than the Gumbel model, while the latter had a slightly lower BIC value. Thus, these two models were compared by means of the likelihood ratio test to verify whether the shape parameter ξ was equal to 0.

From Table 4, the p-value of the likelihood ratio test demonstrated that there was insufficient statistical evidence to reject the hypothesis that the shape parameter was equal to zero. However, for this particular case, the GEV (Weibull) model was preferred, since it exhibited a better graphic fit (Figures 3 and 4, "Model Quantiles" and "Max Empirical Quantiles") and the return levels obtained with it were closer to the historical reality of Ecuador. The graphs of empirical versus theoretical density modelled and the graph of return levels of the two models are also presented in Figures 3 and 4.

Table 4. Likelihood ratio test of Gumbel versus GEV

Likelihood Ratio Test	
Likelihood Ratio	3.53
p-value	0.06

Return periods were determined for 2, 5, 10, 50, and 100 years at a confidence level of 95%. For a return period of 50 years, for example, an earthquake greater than or equal to 7.8 M_W would be expected to occur in Ecuador (Table 5).
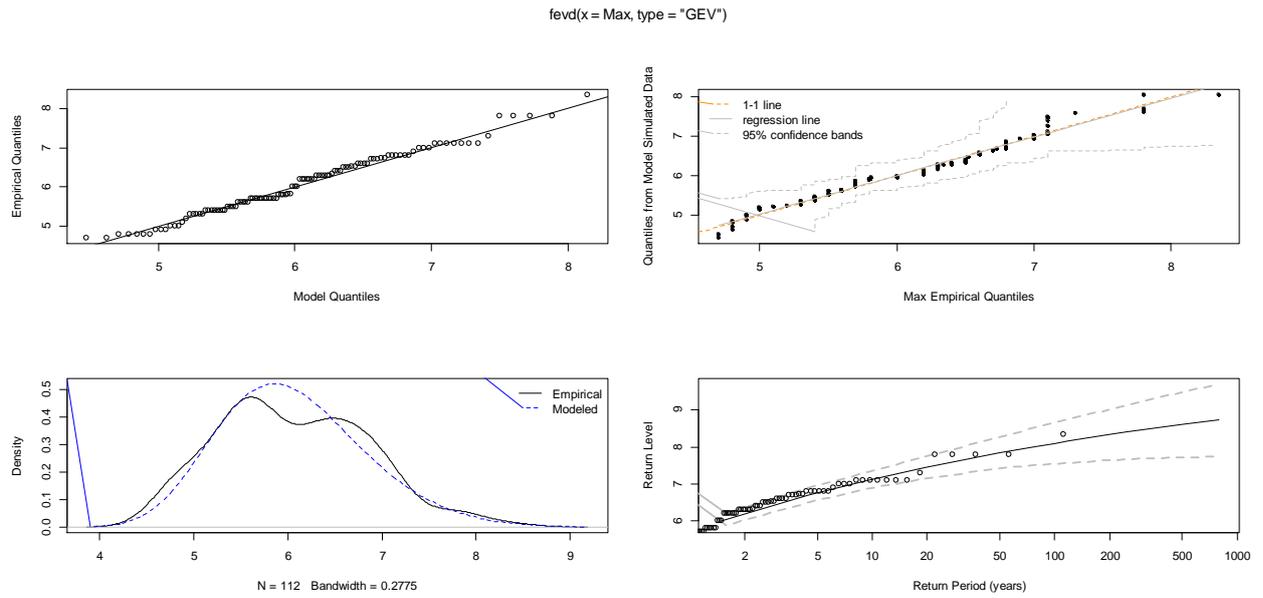
fevd(x = Max, type = "GEV")

Figure 3. *Graphs of the GEV distribution model*
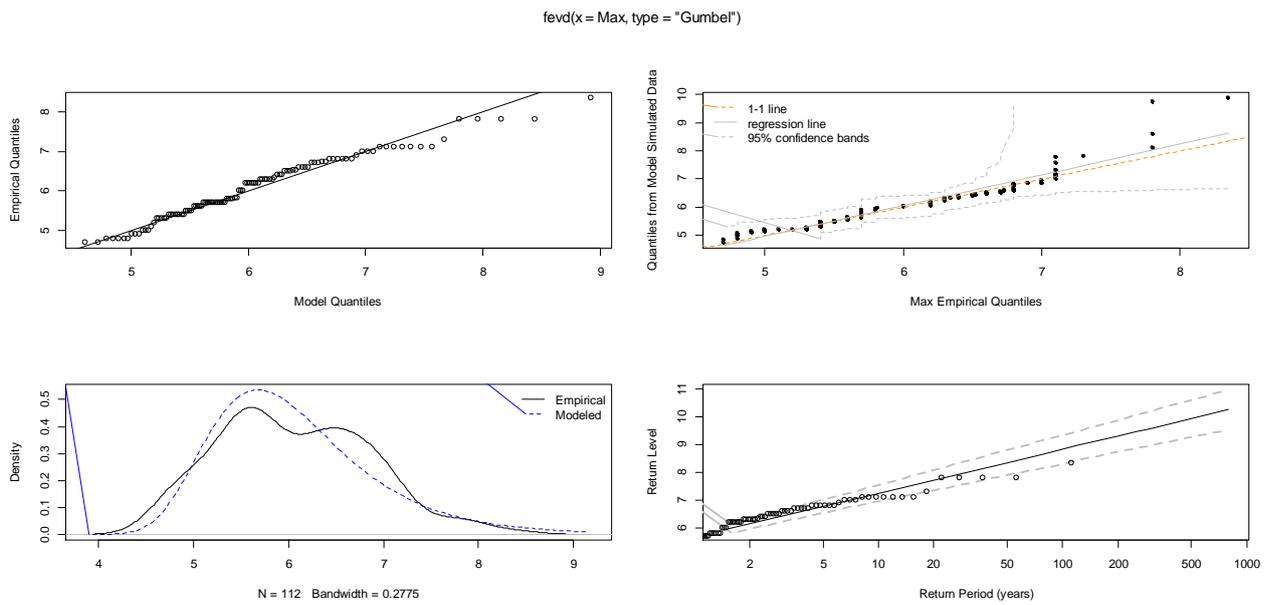
fevd(x = Max, type = "Gumbel")

**Figure 4.** *Graphs of the Gumbel distribution model*

Table 5. *Periods and return levels of the GEV model*

Return period	Lower Limit of Interval 95%	Estimated Magnitude	Upper Limit of Interval 95%
2 years	5.83	5.99	6.15
5 years	6.51	6.70	6.89
10 years	6.871	7.10	7.33
50 years	7.41	7.84	8.27
100 years	7.55	8.10	8.65

The *POT approach* uses the GP distribution as the limit distribution for earthquake magnitudes that exceed a threshold. It is critically important to obtain the threshold; in this work, the graph of the threshold relationship and the number of exceedances obtained through the R library [13] were analyzed (Figure 5). The optimum is to have a high threshold with many observations (exceedances) and therefore, the potential threshold values considered were those between 4.90 to 5.27 M_W. Models were constructed with each of these thresholds and the one that exhibited the lowest AIC and BIC was selected. The chosen threshold was 5.2 M_W.
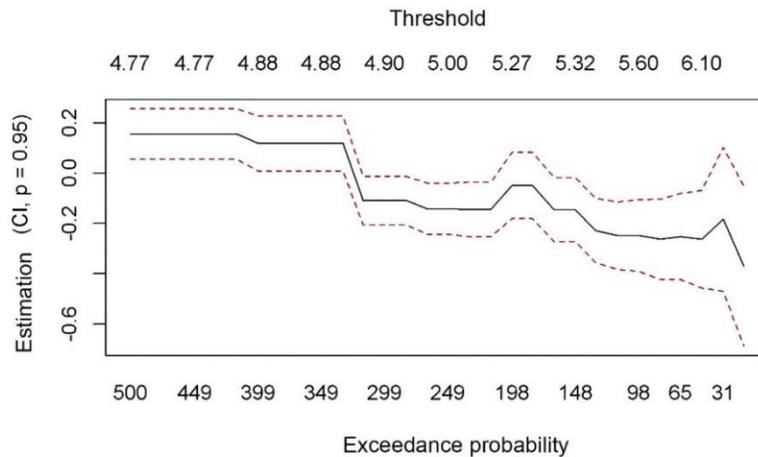

Figure 5. *Variation of the shape parameter of the POT model applied to the magnitude of earthquakes in Ecuador*

Table 6 presents the parameters estimated using maximum likelihood for GP model with a threshold value of 5.2 MW. Figure 6 shows the observed distribution versus the GP theory, together with the quantile-quantile graphs of the respective model.

Table 6. *GP model applied to the magnitudes of the earthquakes in Ecuador*

	Parameters			
	Scale		**Shape**	
Estimate	0.79		-0.16	
Standard error	0.075		0.064	
AIC	251.02		**BIC**	257.64
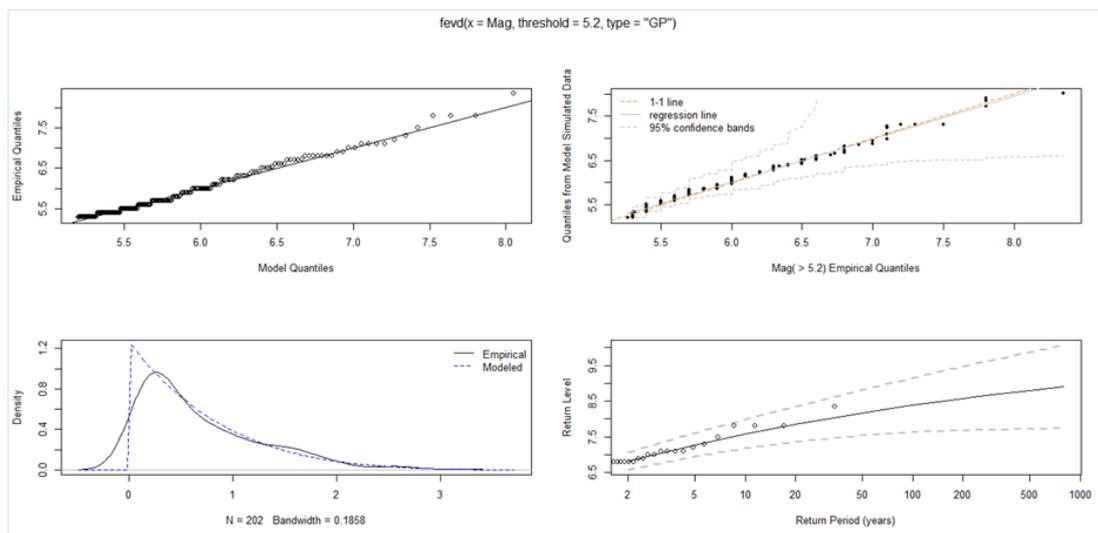
Figure 6. *Graph of the generalized Pareto (GP) distribution model*

Return periods were determined for 2, 5, 10, 50, and 100 years at a confidence level of 95%. For a return period of 50 years, for example, an earthquake exceeding 8.18 M_W would be expected to occur (Table 7).

Table 7. *Return periods of the GP model*

Return Period	Lower Limit of Interval 95%	Estimated Magnitude	Upper Limit of Interval 95%
2 years	6 .58	6.82	7.07
5 years	6.96	7.28	7.60
10 years	7.19	7.58	7.98
50 years	7.54	8.18	8.82
100 years	7.62	8.39	9.15

3.3. Zoning of Ecuador through Clustering Methods

Characteristic zones were determined as a function of the analyzed variables: the Universal Transverse Mercator (UTM) geographic coordinates of the epicentre, depth, and magnitude of the earthquake. Two clustering methodologies were considered: K-Means and Ward. The Silhouette and Calinski criteria were applied to choose the optimal number K of groups, because these indicators are measures of similarity of a group's elements.

The hierarchy method was not useful for zoning, given that the clusters overlapped and were dispersed throughout Ecuador

To determine the optimal number of groups in the K-Means methodology, indicators that quantify the homogeneity within the group and heterogeneity among the groups can be used. The Silhouette index, which is a measure of how similar an object is to its own group compared to other groups, was applied in this study. A high value indicates that the object fits well with its own group and does not correspond to neighbouring groups [45]. Figure 7 graphically presents the relationship between number of clusters and the Silhouette index; according to this criterion, 7 groups must be chosen, since the Silhouette index was maximized.
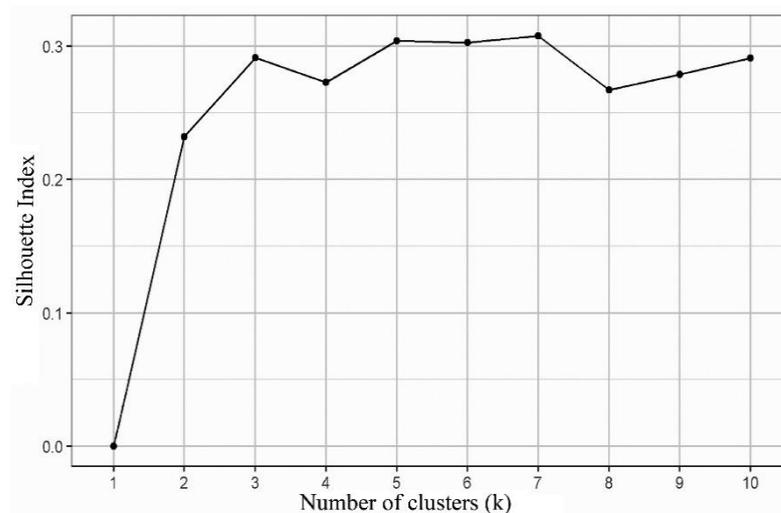
Figure 7. *Silhouette index (K-Means)*

The Calinski–Harabasz index is a ratio that has a measure of the separation of groups as the numerator and a measure of the way in which points are grouped within clusters as the denominator [46]. When clusters are properly formed, groups would be expected to be well separated, so that the value of the numerator is

large, while the points within a group should be as close as possible to each other, that is, a smaller denominator. Therefore, the optimal number of groups according to this criterion is the one that maximizes the Calinski–Harabasz index. Figure 8 suggests that the optimal number of clusters is 5.
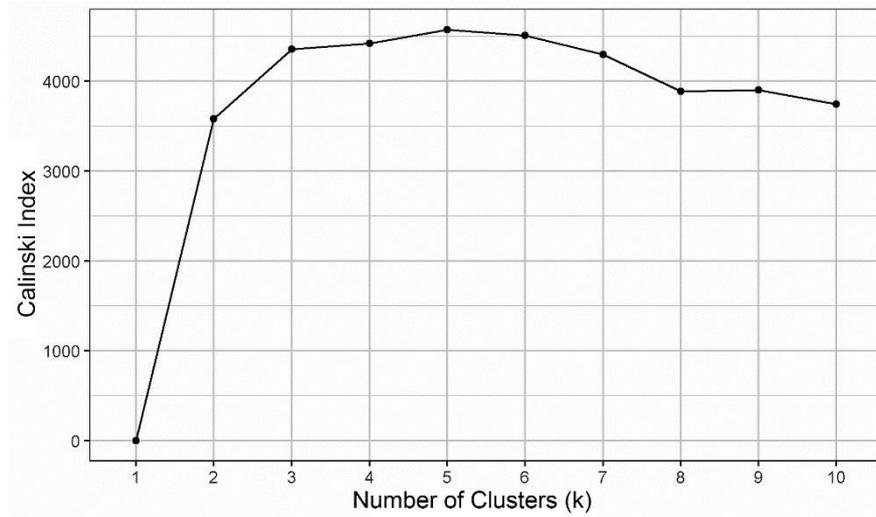
Figure 8. A plot of the Calinski–Harabasz index as a function of K

Initially, the criterion of the Silhouette index was considered and a K=7 groups was chosen. However, this number of groups did not allow good discrimination of the seismic zones because they overlapped. In addition, two of the seven groups had few records of earthquakes, which did not allow the BM and POT models to be built reliably. Therefore, the value of K = 5 zones was considered (Figure 9).

Table 8 summarizes the average characteristics for each zone. Zones 1 and 4 have the highest magnitude earthquakes. Zone 1 experiences more shallow earthquakes because it is located on the coast of Ecuador, while zone 4 is in the mountainous region so that earthquakes have a greater depth on average.
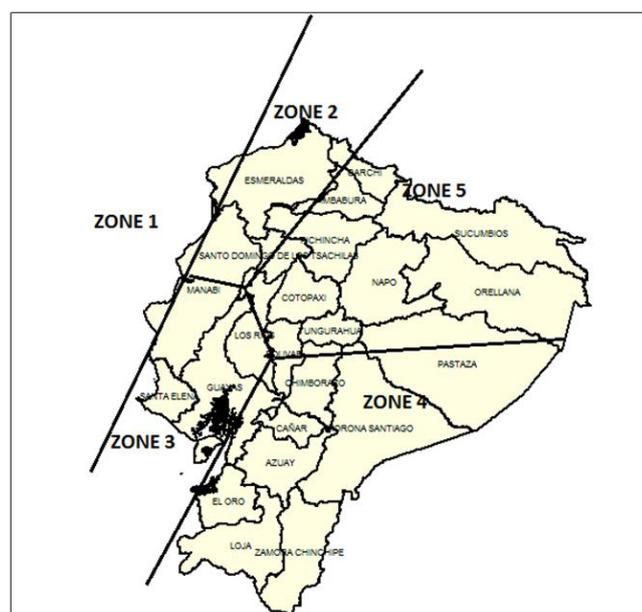
Figure 9. Ecuador divided into 5 seismicity zones

Table 8. *Average characteristics by zones*

Zone	Mag. (M_W)	Depth km	UTM-X	UTM-Y
1	3.71	14.05	528370.4	9955005.7
2	3.39	25.02	690794.6	10047886.2
3	3.54	30.06	794232.5	9790216.2
4	3.61	53.10	790853,9	9727706.0
5	3.3626	36.99	795122.0	9909860.0

Zone 1 falls in the northwest of Ecuador, covering portions of Esmeraldas and Manabí and their respective maritime zones. Zone 2 also falls in the northwest, covering much of the province of Esmeraldas and parts of Manabí, Pichincha, Santo Domingo de los Tsáchilas, Carchi, and Imbabura. Zone 3 is in the southwest, and includes parts of the provinces of Manabí, Los Ríos, Guayas and Santa Elena, the latter 2 with their maritime zones. Zone 4 is located in the south of Ecuador and includes the provinces of Morona Santiago, Zamora Chinchipe, El Oro, Loja, Azuay, Cañar, Chimborazo and parts of Guayas, Los Ríos, Pastaza and Bolívar. Finally, Zone 5 is located in north central Ecuador, covering the provinces of Napo, Cotopaxi, Tungurahua, Sucumbíos, Orellana and parts of Pichincha, Santo Domingo de los Tsáchilas, Chimborazo, Bolívar, Los Ríos, Imbabura, Carchi, and Pastaza.

3.4. Statistical Models Applied to the Zoning with the K-Means Clustering Method

Once the zoning of the Ecuadorian territory was established, the extreme value theory models were applied to each of the zones. The model that best accounted for the maximum magnitude earthquakes of each of the zones was the one obtained with the BM method, specifically the Gumbel distribution model. This was concluded through an analysis similar to the one without zoning.

Tables 9A and 9B show the 95% confidence intervals of the return levels and the return periods. Magnitudes greater than or equal to 7.1 M_W would be expected in zone 1 with a return period of 50 years; for zone 2, there will be an earthquake greater than or equal to 6.16 M_W every 50 years. This was similar to zone 3, where a magnitude higher than 6.43 M_W is expected for the same return period. In zones 4 and 5, an earthquake greater than or equal to 6.9 M_W is expected to occur every 50 years on average. In zones 1, 2, and 3, more superficial earthquakes are expected, while deeper ones are expected in zones 4 and 5.

Table 9A. *Return periods of seismic magnitudes (M_W) for each zone*

Return period	Zone 1			Zone 2			Zone 3		
	Lower Limit of 95% Interval	*Estimated Magnitude*	*Upper Limit of 5% Interval*	*Lower Limit of 95% Interval*	*Estimated Magnitude*	*Upper Limit of Interval 95%*	*Lower Limit of Interval 95%*	*Estimated Magnitude*	*Upper Limit of 95% Interval*
2 years	4.93	5.10	5.24	4.35	4.50	4.65	4.76	4.88	5
5 years	5.48	5.72	5.95	4.80	5.04	5.27	5.19	5.38	5.56
10 years	5.83	6.13	6.44	5.09	5.39	5.69	5.47	5.70	5.94
50 years	6.58	7.1	7.52	5.71	6.16	6.62	6.07	6.43	6.79
100 years	6.89	7.44	7.98	5.97	6.49	7.01	6.32	6.73	7.14

Table 9B. *Return periods of seismic magnitudes (M$_W$) for each zone*

Return period	Zone 4			Zone 5		
	Lower Limit of Interval 95%	*Estimated Magnitude*	*Upper Limit of Interval 95%*	*Lower Limit of Interval 95%*	*Estimated Magnitude*	*Upper Limit of Interval 95%*
2 years	5.11	5.24	5.38	5.01	5.14	5.28
5 years	5.58	5.79	5.99	5.51	5.72	5.93
10 years	5.88	6.14	6.40	5.83	6.10	6.37
50 years	6.54	6.93	7.33	6.52	6.94	7.35
100 years	6.81	7.27	7.72	6.82	7.29	7.77

Regression models were constructed to estimate the number of earthquakes that occurred during the period 1906–2017, for different magnitude bands, for each of the zones. The estimated relation for each of the models was: $log(\mu_i) = \hat{\alpha} + \hat{\beta}IM_W$ where μ_i is the average number of earthquakes by zone and by magnitude interval of earthquakes in M$_W$, and IM_W are magnitude intervals of the earthquakes. Poisson regression models were used through a maximum likelihood estimation and the classical linear regression model that utilized a least squares estimation. Coefficients for each zone are presented in Table 10.

Tables 11A and 11B present the observed and estimated values of the number, or frequency, of earthquakes for each band of earthquake magnitudes on the M$_W$ scale. The mean absolute error (MAE) is also presented. For zones 1 and 3, the values of MAE were similar for both models; for zones 2, 4, and 5, the estimates obtained by Poisson regression were better, which is reflected in lower MAE values. The NA (not available) values mean that classical regression models could not estimate frequencies of zero. In general, the Poisson regression can estimate frequencies better than a classical regression. It is noteworthy that the classical regression is commonly used in the Gutenberg–Richter relationship [47].

Table 10. *Estimated parameters*

Zone	Poisson Regression		Classical Regression	
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
1	13.51	-1.44	13.71	-1.49
2	15.94	-2.31	13.51	-1.80
3	14.34	-1.78	14.33	-1.78
4	13.82	-1.53	14.31	-1.64
5	17.48	-2.32	15.14	-1.82

Table 11A. *Estimated parameters derived from Poisson and classical regression*

M$_W$	Frequencies of earthquakes								
	Zone 1			Zone 2			Zone 3		
	Real	*Poisson regression*	*Classical regression*	*Real*	*Poisson Regression*	*Classical regression*	*Real*	*Poisson Regression*	*Classical regression*
Mw ≤ 4	2289	2344.66	2334.143	812	804.66	555.68	1368	1374.83	1366.76
4< Mw < 5	660	557.07	526.368	68	79.6	92.2	244	232.21	231.25
5 < Mw < 6	93	132.35	118.7	9	7.88	15.3	35	39.22	39.13
6 < Mw < 7	24	31.45	26.768	4	0.78	2.54	7	6.62	6.62
Mw > 7	7	7.47	6.036	0	0.08	NA	0	1.12	NA
MAE		**41.17**	**41.64**		**4.67**	**72.07**		**4.87**	**4.62**

Table 11B. *Estimated parameters derived from Poisson and classical regression*

| M_W	Zone 4			Zone 5		
	Real	*Poisson Regression*	*Classical Regression*	*Real*	*Poisson Regression*	*Classical Regression*
Mw ≤ 4	2162	2213.74	2287.47	3715	3682.96	2558.56
4< Mw < 5	571	479.83	441.64	312	363.32	413.23
5 < Mw < 6	76	104	85.27	44	35.84	66.74
6 < Mw < 7	12	22.54	16.46	13	3.54	10.78
Mw ≥ 7	4	4.89	3.18	2	0.35	1.74
MAE		**36.47**	**53.9**		**20.53**	**256.6**

Zones 1 and 5 were the most representative in terms of recurrence and seismic magnitude, with 24% and 32% of the total earthquakes recorded, respectively, and the highest magnitudes associated with the return periods shown (Tables 8A, 8B, 10A, and 10B).

3.5. Discussion

From a geological point of view, each zone obtained in this work generally shows characteristics of their location. Studies on the characterization of faults capable of generating macro seismic events provided by [18] and shown in Figure 7 indicate that inverse and shear faults are associated with Zones 1, 4, and 5; in contrast, the normal type of faults dominate in Zone 2.

In general, the shear type of fault (lateral movements) is the one present in the Andean region (Andes mountain range). For regions with low altitudes, inverse faults dominate, except for the southern part of the Ecuadorian coast in the province of Guayas, where normal faults predominate.

The maritime region of Zone 1 experiences earthquakes predominantly originating in the subduction zone and, being close to this seismogenic source, explains the occurrence of high-magnitude earthquakes. Additionally, the zone shows high activity given that the faults are active, as evident from the fact that the most recent events occurred there in the current decade.
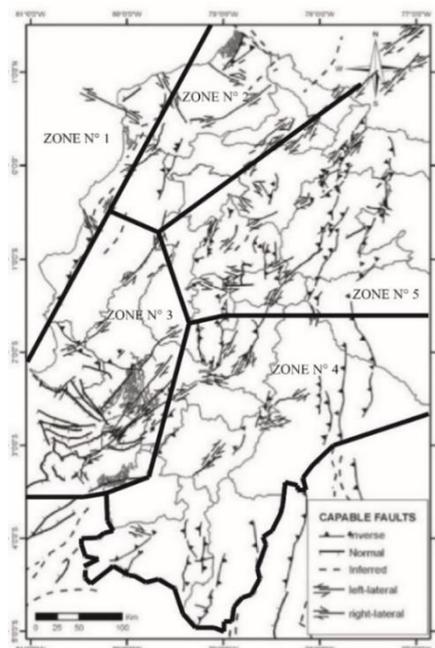
Figure 7. *Comparison of zones obtained in the investigation versus kinematics, type of capable faults mapped in Ecuador [45]*

Zone 2 mostly experiences a uniform distribution of earthquakes throughout its area. The low occurrence of inverse and sinistral faults (left lateral movement) in the area explains the high return periods for strong earthquakes in this area. Furthermore, the faults are relatively short.

Zone 3 is located in a system of normal faults that constitute the Mega-fault Guayaquil-Caracas. Other active faults associated with dextral (right lateral movement) and inverse movements are also present around the Santa Elena Province, but their lengths are not long. Longer lengths of capable faults are normally associated with higher magnitudes [48]. Return periods for strong earthquakes are high due to the type of displacement expected.

Zone 4 mostly contains faults of reverse character, except for the normal and sinistral faults present in the pattern of the Mega-fault Guayaquil-Caracas that extends by the Andes mountain range. The return periods for high magnitude earthquakes are very low, similar to Zone 1. The presence of a number of active faults throughout the area justifies the high seismic activity in the catalogue used.

In Zone 5, there are mostly sinistral, dextral, and inverse earthquakes, with a slightly low return period associated with high-magnitude earthquakes. This zone experiences the highest seismic activity, given the large number of active faults present; however, when compared to Zones 1 and 4, the lengths of its faults are less long.

Table 12 summarizes the results of a comparison of expected magnitudes derived from the probabilistic model (return period 475 y) with maximum magnitudes estimated in the subduction models as proposed by [10] and [18]. The authors' most representative zoning was compared with that of the K-Means clustering method for the lower and upper limits of the 95% confidence interval.

Table 12. Comparison of maximum magnitudes in subduction models

Zone	*Authors*	*(Chunga et al., 2016)*	*(Parra et al., 2016)*
1	$[7.60 < M_W < 9.01]$	$7.7 < M_W < 8.8$	$5.3 < M_W < 8.8$
2	$[6.55 < M_W < 7.90]$	$7.7 < M_W < 8.8$	$5.3 < M_W < 8.8$
3	$[6.88 < M_W < 7.95]$	$7.1 < M_W < 7.8$	$5.3 < M_W < 7.9$
4	$[7.42 < M_W < 8.61]$	$7.1 < M_W < 7.8$	$7.5 < M_W < 7.7$
5	$[7.47 < M_W < 8.70]$	$7.7 < M_W < 7.8$	$7.0 < M_W < 7.2$

The two subduction models presented by [18] followed the methodology used by [49], and described maximum magnitudes as a function of fault length through modelling of displacement planes between the Nazca and Continental plates. They predicted maximum magnitudes of 7.7–8.8 M_W for the northern region and 7.1–7.8 M_W for the southern region.

In their subduction model, [10] defined seven zones based on the adjusted model of [47], taking into account geological criteria and the maximum probable magnitude in the area.

The magnitudes obtained in this investigation were similar to those of [18] and [10]. For return periods of 475 years, these values would allow the determination of maximum possible magnitudes, since a mathematical approach does not have upper limits in terms of M_W.

4. RESULTS

In this paper, a statistical analysis of the seismic magnitudes at the Ecuadorian level in the continental region was presented, using the maximum likelihood estimation of the BM, POT, and Poisson regression models. To identify zones within Ecuador according to the seismic hazard, the Ward and K-Means clustering methodologies were used. The Ward's clustering method was not useful in the creation of zones because it produced zones that overlapped and were undifferentiated. In contrast, the K-Means method

allowed the definition of 5 zones with different characteristics in terms of magnitude, depth, and position. For each of these zones, the best extreme values theory model was that of Gumbel, with zones 1, 4 and 5 having the highest expected magnitudes for different return periods.

The Poisson regression provided better results than the classical regression for earthquake frequencies for each band of magnitudes, and so it can be concluded that the Gutenberg-Ritcher relationship could also be estimated with a Poisson regression. The results of the extreme value theory models and the Poisson regression demonstrate that zones 1 and 4 would be highly seismic.

It can also be concluded that the zoning obtained has an important relationship with the geological nature of the active faults and the seismicity rate present in Ecuador. Areas with a greater number of faults result in high earthquake frequency, and zones with faults of inverse character will experience higher magnitude earthquakes with low return periods.

It is concluded that provinces such as Manabí and Esmeraldas are prone to high-magnitude earthquakes for much lower return periods than other areas, with a possibility of reaching a magnitude equal to or greater than 7.0 M_W in 50 years.

It is recommended for subsequent studies to adjust the statistical models proposed in this research with existing geological studies and criteria, in such a way that more precise return periods can be obtained in accordance with the geological history of the region.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Nocquet, J., Mothes, P., and Alvarado, A., "Geodesia, geodinámica y ciclo sísmico en Ecuador", Geología y Geofísica Marina y Terrestre Del Ecuador (Comisión Nacional del Derecho del Mar, Ecuador), 83-94, (2009).

[2] Witt, C., and Bourgois, J., "Relaciones entre la evolución de la cuenca del Golfo de Guayaquil-Tumbes y el escape del Bloque Nor-Andino", Geología y Geofísica Marina y Terrestre Del Ecuador (Comisión Nacional del Derecho del Mar, Ecuador), 95-106, (2009).

[3] Segovia, M., and Alvarado, A., "Breve análisis de la sismicidad y del campo de esfuerzos en el Ecuador", Geología y Geofísica Marina y Terrestre Del Ecuador (Comisión Nacional del Derecho del Mar, Ecuador), 131-149, (2009).

[4] Vaca, S., Régnier, M., Bethoux, N., Álvarez, V. and Pontoise, B, "Sismicidad de la región de Manta: Enjambre sísmico de Manta-2005", Geología y Geofísica Marina y Terrestre Del Ecuador (Comisión Nacional del Derecho del Mar, Ecuador), 151-166, (2009).

[5] IGEPN, "Terremoto de Esmeraldas de 1906 - Uno de los sismos más grandes de la historia reciente", Quito, Ecuador, recovered from https://www.igepn.edu.ec/servicios/noticias/575-terremoto-de-esmeraldas-de-1906-uno-de-los-sismos-m%C3%A1s-grandes-la-historia[Access date: 21 November 2019], (2012).

[6] IGEPN, "Informe sísmico para el año 2016", [online] Quito: Instituto Geofísico – EPN, recovered from https://www.igepn.edu.ec/informes-sismicos/sismicos-anuales/20518-informe-sismico-para-el-ecuador-ano-2016/file [Access date: 2 April 2019], (2016).

[7]   CEC, "Requisitos generales de diseño: peligro sísmico, espectros de diseño y requisitos mínimos de cálculos para diseño sismo resistente", Código Ecuatoriano de la Construcción. Registro Oficial No. 382, (2001).

[8]   Bakun, W. H., and Wentworth, C. M., "Erratum to Estimating earthquake locations and magnitude from seismic intensity data", Bulletin of the Seismological Society of America., 89: 557, (1999).

[9]   NEC-15, "Norma Ecuatoriana de la Construcción", Registro Oficial No. 413 del 10 de enero de 2015, recovered from http://www.normaconstruccion.ec/ [Access date: 1 December 2019], (2015).

[10] Parra, H., Benito, M., and Gaspar-Escribano, J, "Seismic hazard assessment in continental Ecuador", Bullentin of Earthquake Engineering, 14(8): 2129-2159, (2016).

[11] Beauval, C., Mariniére, J., Yepez, H., Audin, L., Nocquet, J., Alvarado, A., Baize, S., Aguilar, J., Singaucho, J. and Jomard, H., "A New Seismic Hazard Model for Ecuador", Bulletin of The Seismological Society of America, 108(3A): 1443-1464, (2018).

[12] Anderson, E., "Asymptotic Properties of Conditional Maximum Likelihood Estimators", Journal of the Royal Statistical Society, 32: 283-301, (1970).

[13] R Core Team., "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/[Access date 11 December 2019], (2008).

[14] Beauval, C., Yepes, H., Palacios, P., Segovia, M., Alvarado, A., Font, Y., Aguilar, J., Troncoso y L., and Vaca S., "An Earthquake Catalog for Seismic Hazard Assessment in Ecuador", Bulletin of the Seismological Society of America, 103(1): 773-786, doi: 10.1785/0120120270, (2013).

[15] IGEPN, "Instituto Geofísico Escuela Politécnica Nacional", Quito, Ecuador, http://www.igepn.edu.ec/solicitud-de-datos [Access date 1 May 2018], (2018).

[16] IRIS, "Incorporated Research Institutions for Seismology. Interactive Earthquake Browser", http://ds.iris.edu/ieb/ [Access date 11 May 2018], (2017).

[17] Parra, H., "Desarrollos metodológicos y aplicaciones hacia el cálculo de la peligrosidad sísmica en el Ecuador Continental y estudio de riesgo sísmico en la ciudad de Quito" Tesis Doctoral, Departamento de Ingeniería Topográfica y Cartografía, Escuela Técnica Superior de Ingenieros en Topografía, Geodesia y Cartografía, Universidad Politécnica de Madrid, recovered from http://oa.upm.es/39353/1/HUMBERTO_PARRA_CARDENAS_V-2.pdf[Access date: 31 October 2020], (3): 44-54, (2016).

[18] Chunga, K., Gorshkov, A., and Michetti, A., Geología de Terremotos y Tsunami (1st ed), Sección Nacional del Instituto Panamericano de Geografía e Historia, IPGH, Guayaquil, (2016).

[19] Chunga, K., Gorshkov, A., Michetti, A., Panza, G., Soloviev., and Martillo, C., "Aplicación del método de zonación morfo-estructural para identificar nudos sismogénicos en la región costera y cadenas montañosas de los andes septentrionales del Ecuador", Acta Oceanográfica del Pacífico, (16):119-140, (2011).

[20] Epstein, B., and Lomnitz, C, "A model for occurrence of large earthquakes", Nature, 211: 954–956. doi:10. 1038/211954b0, (1966).

[21] Kijko, A., and Sellevoll, MA. "Triple exponential distribution, a modified model for the occurrence of large earthquakes", Bulletin of the Seismological Society of America, 71(6): 2097–2101, (1981).

[22] Pavlenko, V. A., "Estimation of the upper bound of seismic hazard curve by using the generalized extreme value distribution", Natural Hazards, 89(1): 19-33. doi:10.1007/s11069-017-2950-z, (2017).

[23] García-Bustos, S., Landín, J., Moreno, R., Chong, A.S.E., Mulas, M., Mite, M., and Cárdenas, N., "Statistical analysis of the largest possible earthquake magnitudes on the Ecuadorian coast for selected return periods", Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards. 14(1): 56-68, doi: 10.1080/17499518.2018.1542500, (2018)

[24] Fréchet, M., "Sur la loi de probabilité de l'écart máximum", Annales de la Societe Polonaise de Mathematique, Cracow, 6: 93-116, (1927).

[25] Fisher, R., and Tippet, L, "Limiting forms of the frequency distribution of the largest and smallest member of a sample", Mathematical Proceedings of the Cambridge Philosophical Society, 24(2): 180-190, doi:10.1017/s0305004100015681, (1928).

[26] Gnedenko, B. V., "Sur la distribution limite du terme maximum d'une série aléatoire", Annals of Mathematics, 44(3): 423-453, (1943).

[27] Gumbel, E. J., "Leŝ moments des distributions finales de la premiere et de la derniere valeur", Camptes Rendus de l'Académie des Sciences, 198(l): 41-143, (1934).

[28] Gumbel, E. J., "Les valeurs extremes des distributions statistiques", Annales de l'institut Henri Poincaré, 5(2): 115-158, (1935).

[29] Weibull, W., "A statistical distribution function of wide applicability", Journal of Applied Mechanics ASME, 18 (3): 293-297, (1951).

[30] Von Mises, R., "La distribution de la plus grande de n valeurs", American Mathematical Society, Volumen II Providence:R.I., 271-294, (1936).

[31] Jenkinson, A. F., "The frequency distribution of the annual maximum (or minimum) values of meteorological elements", Quarterly Journal of the Royal Meteorological Society, 1(81): 158-171, (1955).

[32] Ferreira, J. A., and Soares, C. G., "An application of the peaks over threshold method to predict extremes of significant wave height", Journal of Offshore Mechanics and Arctic Engineering, 120(3): 165-176, (1998).

[33] Pickands, J., "Statistical inference using extreme order statistics", The Annals of Statistics, 3(1): 19-131, (1975).

[34] Balkema, A., and Haan, L., "Limit distributions for order statistics, I", Theory of Probability and its Applications, 23(1): 77-92, (1978).

[35] Balkema, A., and Haan, L., "Limit distributions for order statistics, I", Theory of Probability and its Applications, 23(1): 341-358, (1978)

[36] Agresti, A., Categorical Data Analysis (2nd edn), John Wiley & Sons, Inc. (eds), Gainesville, Florida, 557-589, (2002).

[37] De Haan, L., and Ferreira, A., Extreme value theory: An introduction, Springer Science and Business Media, LLC, New York, (2006).

[38] Akaike, H., "On entropy maximization principle". In: Krishnaiah, P.R. (Editor), Applications of Statistics, North-Holland, Amsterdam, 27–41, (1977).

[39] Schwarz, E., "Estimating the dimension of a model", Annals of Statistics, 6(2): 461–464, doi:10.1214/aos/1176344136, (1978).

[40] Rencher, A. C., Methods of Multivariate Analysis, Wiley Interscience, USA, (2002).

[41] Kuiper, F. K., and Fisher, L., "A Monte Carlo comparison of six clustering procedures", Biometrics, 31(3): 777-783, (1975).

[42] Ward, J. H., "Hierarchical grouping to optimize an objective function", Journal of the American Statistical Association, 58(301): 236-244, (1963).

[43] Hartigan, J., The K-Means Algorithm, John Wiley & Sons, New York, (1974).

[44] Alam, M.A., Emura, K., Farnham, C., and Yuan, J., "Best-Fit Probability Distributions and Return Periods for Maximum Monthly Rainfall in Bangladesh", Climate, 6(1):1-16, (2018).

[45] Starczewski, A., and Krzyżak, A., "Performance Evaluation of the Silhouette Index", In Lecture Notes in Computer Science, vol 9120. Springer, Cham, (2015).

[46] Calinski, T., and Harabasz, J., "A Dendrite Method for Cluster Analysis", Communications in Statistics, 3(1): 1-27, (1974).

[47] Gutenberg, B. and Richter, C., "Frequency of earthquakes in California", Bulletin of the Seismological Society of America, 34(4): 185–188, (1944).

[48] Wells D. L., and Coppersmith K. J., "New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement", Bulletin of the Seismological Society of America, 84(4): 974-1002, (1994).

[49] Gutscher, M.-A., Malavieille, J., Lallemand, S., and Collot, J.-Y., "Tectonic segmentation of the North Andean margin: impact of the Carnegie Ridge collision", Earth and Planetary Science Letters, 168(3-4): 255-270, (1999).