Epicentral Region Estimation using Convolutional Neural Networks

Leonel Cruz[1], Rubén Tous[1], Beatriz Otero[1], Leonardo Alvarado[2,3], Sergi Mus[1], and Otilio Rojas[2,4]

[1] Universitat Politécnica de Catalunya, 08034, Barcelona, Spain
{lcruz, rtous, botero, smus}@ac.upc.edu
[2] Universidad Central de Venezuela, 1070, Caracas, Venezuela
[3] Fundación Venezolana de Investigaciones Sismológicas, Caracas, 1070, Venezuela
joleonar@gmail.com
[4] Barcelona Supercomputing Center, Barcelona, Spain
rojasotilio@gmail.com

Abstract. Recent works have assessed the capability of deep neural networks of estimating the epicentral source region of a seismic event from a single-station three-channel signal. In all the cases, the geographical partitioning is performed by automatic tessellation algorithms such as the Voronoi decomposition. This paper evaluates the hypothesis that the source region estimation accuracy is significantly increased if the geographical partitioning is performed considering the regional geological characteristics such as the tectonic plate boundaries. Also, it raises the transformation of the training data to increase the accuracy of the predictive model based on a Projected Coordinate Reference (PCR) System. A deep convolutional neural network (CNN) is applied over the data recorded by the broadband stations of the Venezuelan Foundation of Seismological Research (FUNVISIS) in the region of 9.5 to 11.5ºN and 67.0 to 69.0ºW between April 2018 and April 2019. In order to estimate the epicentral source region of a detected event, several geographical tessellations provided by seismologists from the area are employed. These tessellations, with different number of partitions, consider the fault systems of the study region (San Sebastián, La Victoria and Morón fault systems). The results are compared to the ones obtained with automatic partitioning performed by the k-means algorithm.

Keywords: Earthquake location estimation · Supervised learning · Convolutional neural networks · Deep learning.

1 Introduction and Related Works

Reliable earthquake detection and location algorithms are needed to properly catalog and analyze steadily growing seismic records. Artificial neural networks (ANN) have been employed in [12], [5], [18], [15], [17] and [25] to study earthquake detection and location of seismics events. In [6], ANNs are employed to pick P- and S-wave arrivals using the amplitude of three-component seismic traces, resulting successful on more of the 90% of the testing data. Same Authors in [7], later applied back propagation ANN to identify wave type, either P or S by using signal polarization, and achieved a performance near to 80% in the case of P waves, compared to 60% for S waves. Parallel efforts in [24] and [23] also developed some ANN for P and S detection, and estimate the onset times using a variety of features as input data, including STA/LTA time series [1], [2], windowed spectrograms, and autoregressive coefficients. Results were highly satisfactory for both phases, and effectiveness for S detection was about 86%. Later, a ANN with a problem adaptive structure was used by [8] for P and S picking, and results compared to arrival times chosen by a trained analyst, present deviations close to 0.1 sec for both phases. Recent advances of this sort for earthquake detection are ConvNetQuake [20] and networks in [12]. Using cataloged events as reference, ConvNetQuake and best candidates in [12] show a 100% accuracy detection. In addition, ConvNetQuake also detects an important amount of uncataloged earthquakes in a month of continuous data, where 94% of these events were later confirmed by autocorrelation. Such performances, along with the lower computational costs of ANN compared to established detection schemes (see again [20]), lend themselves a tremendous potential for processing large seismic datasets and real time detection.

This paper presents the outcomes of an approach, called UPC-UCV-GEO, to apply CNN over single-station 3-channel waveforms for earthquake location estimation in north-central Venezuela. The hypothesis that the source region estimation accuracy is significantly increased if the geographical partitioning is performed considering the regional geological characteristics is evaluated. The network applied is UPC-UCV, whose results for P-wave detection and basic source region estimation (with k-means based tessellation) are described in [22]. We apply our technique to seismic data collected by broadband stations at north-central Venezuela, during the time period of 2018 to 2019. The seismicity in the region results from the right-lateral strike-slip faulting experienced along the interface between the Caribbean and South American plates, as the former moves to the east with respect to the latter. An important amount of this movement seems to be accommodated along the Boconó - El Pilar fault system, that extends across Venezuela from west to east. Several geographical tessellations provided by seismologists from the area are employed. These tessellations, with different number of partitions, consider the fault systems of the study region (San Sebastián, La Victoria and Morón fault systems). The results are compared to the ones obtained with automatic partitioning performed by the k-means algorithm. A review of the seismic history and tectonic of related regions can be found in [19], [3] and [4], and references therein.

2 Methodology

2.1 Data

CARABOBO dataset is made of data provided by FUNVISIS, the official agency for monitoring and reporting seismic activity in Venezuela. The FUNVISIS network comprises 35 stations that continuously record signals on three channels at 100 Hz; this is spread over the geographical area containing high seismic activity. The dataset was provisioned by data collected among April 2018 and April 2019 through 5 seismological stations (BAUV, BENV, MAPV, TACV, and TURV) in northcentral Venezuela, specifically on the region between the coordinates 9.5^o to 11.5^o N and 67.0^o to 69.0^o W and this consists of seismic data miniSEED format and includes a catalog with metadata regarding these events(hypocenter, P-wave arrival times, magnitude, etc.) in Nordic format.

The earthquakes have a magnitude ranging from 1.7 to 5.2 Mw distributed over the region shown (see Figure 1), whose epicenters are located on the north-central states of Carabobo, Aragua and Miranda. These zones contain a set of interconnected faults such as: San Sebastián and La Victoria that make up an important seismic area, belong to a continental scale, converging to a larger fault system called Boconó. As well as El Pilar fault system that lies the Caribbean and South American plates.

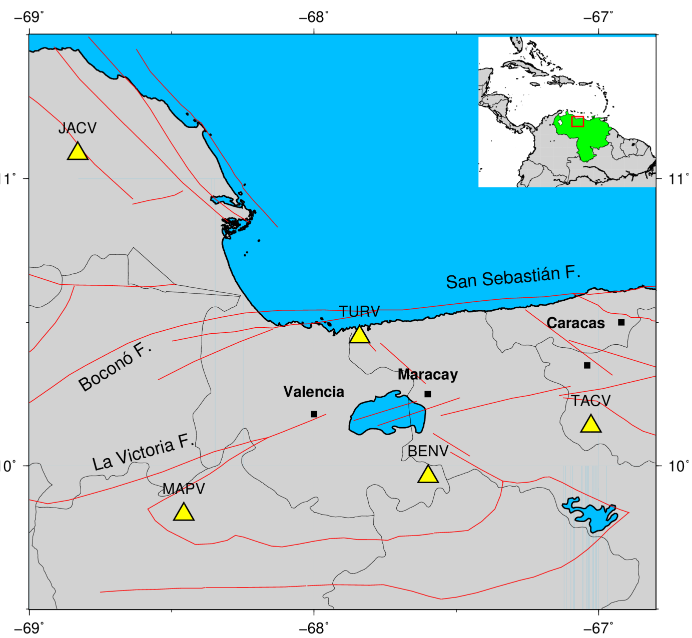

Fig. 1. Geographical distribution of the Boconó, La Victoria and San Sebastián faults, and locations of FUNVISIS stations.

2.2 Earthquake Source Region Estimation

Source region estimation is a relaxed version of the earthquake location problem that consists on, first, partitioning a study area into k geographic subdivisions and, second, attempting to determine to which one the earthquake epicenter belongs. Several works have demonstrated the possibility to estimate the source region of an earthquake from a single-station 3-channel waveform [21], [22], [20]. In this work, source region estimation is approached as a multiclass classification problem. First, the wave is divided into three-channel temporal windows of a fixed-size, and these windows will be classified into k+1 classes: one class for windows not containing a P-wave (negatives) and k classes for windows containing a P-wave (positives) but classified into the k geographic regions that we want to discriminate.

2.3 Partitioning the Study Area into Geographic subdivisions

In order to perform source region estimation, it is first necessary to partition the study area into multiple geographic subdivisions and to define the membership criteria. Previous works, such as [21] and [22], attempted to automatically infer a relevant geographic partitioning directly from the dataset, with the help of a clustering algorithm. Contrary to those methods, in this work the geographic partitioning is provided by seismologists from the study area.

Based on the geographical subdivision that the FUNVISIS expert gave us (see Figures 2 and 3), four different partitions were obtained (k = 3, 4, 5, 10). All these delimited by irregular polygons covering the main seismic faults of Venezuela, according to the following study [14].

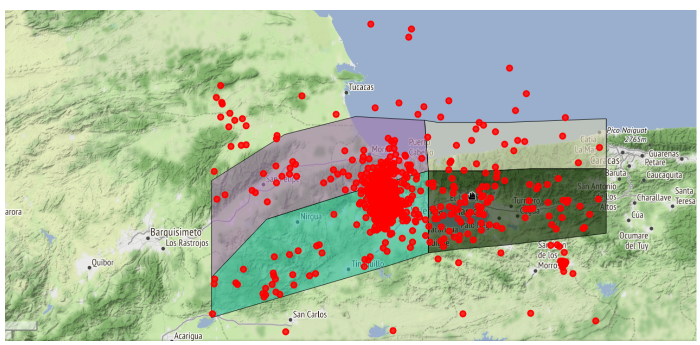

Fig. 2. Clustering the locations of all the earthquakes with fault-based geographic partitioning on 4 zones.

Seismic hazard studies carried out north and central Venezuela by FUNVISIS analyze the seismic activity in shallow seismogenic regions ([10], [11]). These regions were defined as polygons around each fault segment, which were chosen according to their tectonics and degree of seismicity (see Figure 3). At first, 10 regions were delimited, which were later regrouped in new bigger regions with similar seismic-tectonic characteristics, leaving the whole study area split into the 4 regions shown in Figure 2.

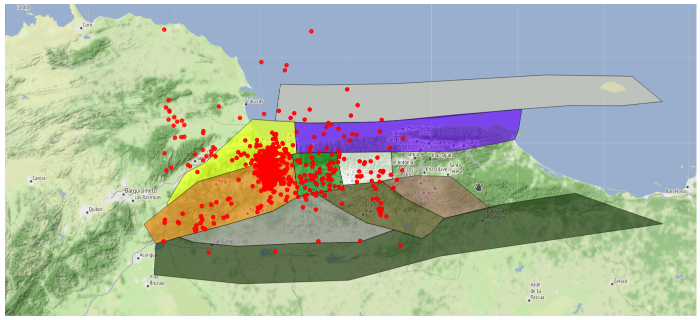

Fig. 3. Clustering the locations of all the earthquakes with fault-based geographic partitioning on 10 zones.

2.4 Data Preprocessing

For the training of the model, a filtering is carried out first, purging the most dispersed events over the study region in order to reduce overfitting.

Input waveforms are first normalized and divided into single-station streams. Each 3-channel single-station stream is split into multiple 3-channel temporal windows of a fixed size. With the events information obtained from the metadata files, we divide the windows into $k + 1$ classes: one class for windows not containing a P-wave (negatives) and k classes for windows containing a P-wave (positives) but classified into the K geographic regions that we want to discriminate. With a 50 sec./window our preprocessing stage yields $12,685$ positives and $84,911$ negatives. The classification of windows among the different K regions is done using clustering the locations of all the earthquakes with a fault-based geographic partitioning provided by a seismologist.

In Figures 2 and 3, we can observe the several segmentations granted by FUNVISIS and the distribution of the events over the study region that will be used for the comparison of the k-means clustering method.

2.5 Spatial Data Preprocessing

In the Carabobo dataset, each of its records contains coordinates that indicate the epicenter of seismic events. To deal with this information is necessary to have a reference frame capable of making sense of these data to view and manipulate them and thus feed the CNN model. This reference frame is known as Coordinate Reference System(CRS) [13], classified in Geographical Coordinate System (GCS) and Projected Coordinate System (PCS). Figure 4 shows the representation of a specific place on earth,on the left, we have its three-dimensional visualization on the globe with a GCS, and on the right, the two-dimensional projection of this same point projected in PCS.

The first is a reference framework that defines the locations of features on a model of the earth. It's shaped like a globe-spherical. Its units are angular,

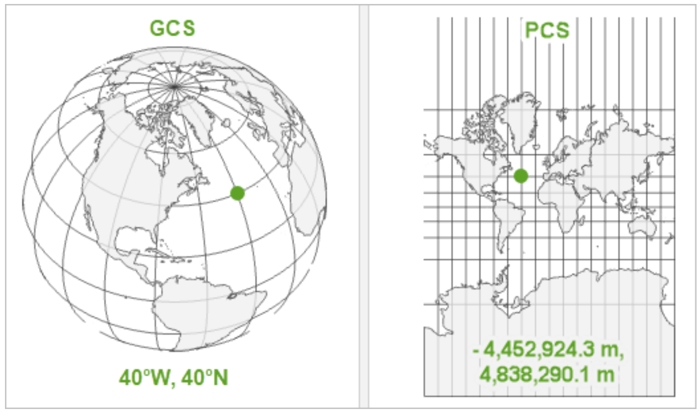

Fig. 4. Types of coordination reference system.

commonly degrees. It's the best for the location and visualization of elements, but distance measurements have a distortion when using latitude and longitude due to the earth's shape. The second is flat. It contains a GCS, but it converts that GCS into a flat surface using a projection algorithm [13] and it is excellent for performing calculations involving distance measurements over geographic areas.

For our dataset, we have selected the GCS called WGS84 to work with our initial data and its corresponding PCR named Pseudo-Mercator, which is by de facto the standard for web services such as Google Maps, Bing Maps, OpenStreet, among others. The coordinates provided generate perimeters with undefined shapes, which in Spatial Data are known as irregular polygons. In order to handle this information, it is necessary to create a JSON (See Listing 1) structure that could store an indeterminate number of points that formed these polygons.

The JSON file contains a key called clusters, whose value is composed of a list of elements that describe the n-regions of which the coordinates define each region's perimeter. An essential section of the file is the points key, because here the coordinates that delimit the boundary and shape of the area are stored; unlike other approaches here, there is no fixed number of elements when forming the list.

For instance, the list 1 shows an irregular polygon is made up of 3 points while the second zone is made up of 5. After defining the type of CRS and the JSON structure that stores the shape of the polygons, we proceeded to use Computational Geometry techniques to assign each of the events to a specific cluster and label it for use model training process. To this, we face the Point in polygon (PIP) problem in which it is decided whether or not a point is in an irregular polygon, described below:

Given a point R and a polygon P represented by n points: $P_0, P_1, ..., P_{n-1}, P_n = P_0$, determine whether R is inside or outside the polygon P. When a line is drawn from R to other point S that is wagered to extend outside the polygon. If this

```
{
 "cluster_type": "RCPoligons",
 "clusters":[

  {
  "id": 1,
  "label": "Zone_01",
  "points": [
     {"x": 10.41413, "y": -67.90393},
     {"x": 10.42613, "y": -66.10873},
     {"x": 10.39413, "y": -66.90323}]
  },

  {
  "id": 2,
  "label": "Zone_02",
  "points": [
     {"x": 10.71413, "y": -66.20393},
     {"x": 10.62613, "y": -66.40873},
     {"x": 10.79413, "y": -67.10323},
     {"x": 10.22763, "y": -66.30822},
     {"x": 10.11113, "y": -67.09223}]
  },

   ...
            ]
}
```

Listing 1: JSON Structure

line $\overline{RS}$ crosses the edges $e_i = \overline{P_i P_{i+1}}$ of the polygon an odd number of times, the points is inside P, otherwise it is outside.

To carry out the PIP queries, we used Shapely's binary predicates [9] that implement these algorithms to assess the topological relationship between geographic objects. It is based on the widely deployed GEOS (Geometry Engine Open Source), allowing work with three main Point, Line String, and Polygons objects. These algorithms are used to determine whether a seismic event falls within the perimeter of a given zone.

2.6 Network Architecture

Our model, called UPC-UCV, is a CNN that takes a multiple-channel single-station seismogram window as input and outputs $k+1$ probabilities estimating if the window contains a P-wave originated at one of the k given locations (one of the outputs is the probability that the window does not contain any P-wave). Figure 5 illustrates the overall network architecture.

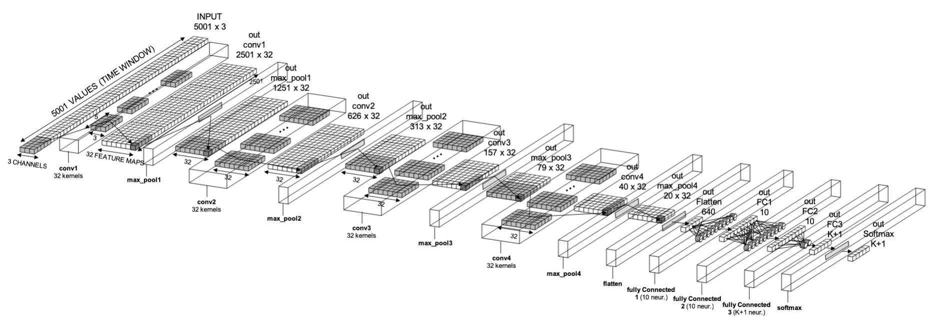

Fig. 5. Network architecture with a 3-channel input, 4 convolutional+max_pooling layers, 3 fully connected layers and a softmax layer

The network has four 32-kernels convolutional layers, each one of them with an associated max pooling layer. The convolutions are applied in a 1D fashion (only through the temporal axis), but the kernels are 2D to process the multiple input channels. Given a 2D kernel k of size $s \times c$ (width s over c channels) of a given layer l, $W^{k,l} \in \mathbb{R}^{2\mathrm{s}+1\times \mathrm{c}}$:

$$W^{(k,l)} = \begin{pmatrix} W^{(k,l)}_{-s,1} & & \dots & & W^{(k,l)}_{-s,c} \\ \vdots & & \vdots & & \vdots \\ W^{(k,l)}_{0,1} & \dots & W^{(k,l)}_{0,0} & \dots & W^{(k,l)}_{0,c} \\ \vdots & & \vdots & & \vdots \\ W^{(k,l)}_{s,1} & & \dots & & W^{(k,l)}_{s,c} \end{pmatrix} \tag{1}$$

The discrete convolution of the input 2D tensor Y with filter $W^{k,l}$ at position t (the kernel only moves in 1D) is given by:

$$(Y * W^{k,l})_t = \sum_{u=-s}^{s} \sum_{v=1}^{c} W^{k,l}_{u,v} * Y_{t+u,v} \tag{2}$$

The output t of a convolutional layer l and kernel k is computed as:

$$Y^{(l)}_{k,t} = \sigma(b^{(l)}_k + (Y^{(l-1)} * W^{k,l})_{t'}) \tag{3}$$

where t' is index of the previous layer where the convolution will be applied. t' depends on t but also on the stride (2 in our case). $\sigma(\dot{}) = max(0, \dot{})$ is the nonlinear Rectified Linear Unit) (ReLU) activation function and $b^{(l)}_k$ is the bias term for kernel k in layer l. We employ 32 kernels with shape 5×3 in the first layer and 5×32 in the subsequent convnet layers. After each convolutional layer we apply a max-pooling layer with a pooling window of size 5 and stride 3.

After the last convolutional layer we flatten the resulting 2D tensor into a 1D tensor. This feature vector is then processed by three fully connected layers with ReLU activation functions (10 neurons the first ones, $k + 1$ neurons the last one). Finally, a softmax function is applied to the class scores to obtain a properly normalized probability distribution.

3 Experiments and Results

This work's main objective is to assess the effectiveness of source region estimation using a geographic partitioning provided by an expert and determine the impact on the prediction's improvement using a PCR transformation, compared with the approach automatically generated with k-means as in [22]. The UPC-UVC network's best configuration was taken as a basis, considering the network's parametrization and geometry to carry out this new approach.

The experiments were performed on equipment provided by the Computer Architecture Department of the UPC. The device had Intel(R) Core(TM) i7-3770 CPU running at 3.40GHz and 8GB in RAM. With this configuration, the training process was carried out between 3 to 4 hours per model.

In the first instance, a set of experiments were performed to determine if Spatial-Data techniques within the pre-processing of the training data increased the accuracy of the model prediction results. The data in the Table 1 describe the experiment; in the first column, we partition the study area into k regions; in the second column, we have the experiment's result without using a pre-processing, taking by default a Geographical Coordinate System(GCS) as in the works [22], [16]. In the last column, we have the effect after transforming the original data and converting them into a Projected Coordinate System(PCS). As shown in the table, the post-transformation result obtains an accuracy higher than 90% in each of the cases.

Table 2 summarizes the results of UPC-UCV-GEO obtained using spatial data pre-procesing. The results of ConvNetQuake and UPC-UVC are provided for comparison. For a small number of geographic subdivisions (3-5), the obtained results don't enable to confirm the target hypothesis. The partitioning into 4 regions recommended by the expert (UPC-UCV-GEO with K=4) provided an accuracy of 95.43%, just slightly above than the results for a k-means based partitioning (UPC-UCV) with K=5 (93.36%) and slightly below than the results for a k-means based partitioning (UPC-UCV) with K=3 (95.68%). However, the hypothesis seems to be confirmed for a more fine-grained partitioning (K=10), as UPC-UCV-GEO obtains an accuracy of 91.43% while the accuracy of UPC-UCV degradates to 66.10%.

Table 1. Forecasting impact of source region estimation based on Coordinate Reference System

K Zones	**Accuracy GCS**	**Accuracy PCS**
4	86.52%	95.43%
10	87.72%	91.78%
16	88.93%	91.43%

Table 2. Source region estimation results

K Zones	**Model**	**Accuracy**
3	ConvNetQuake	84.58%
	UPC-UCV	95.68%
4	UPC-UCV-GEO	95.43%
5	ConvNetQuake	82.08%
	UPC-UCV	93.36%
10	UPC-UCV-GEO	91.78%
	UPC-UCV	66.10%

4 Conclusions

In this paper, we have evaluated the hypothesis that the accuracy of methods (such as [22] and [20]) for the automated estimation of the epicentral source region of a seismic event is increased if the geographical partitioning is performed considering the regional geophysical characteristics. The UPC-UCV-GEO deep convolutional neural network is applied over the CARABOBO dataset, consisting of three-channel seismic waveforms recorded in north-central Venezuela from April 2018 to April 2019. Instead of partitioning the data with K-means, we have applied several geographical tessellations provided by seismologists from the study area.

While the obtained results for a small number of geographic subdivisions are not better than the ones obtained with k-means clustering, the good results obtained with a large number of subdivisions (91.78% with K=10) outperform the k-means approach (66.10%). It should be noted that to obtain these results, the use of spatial-based techniques significantly improved the final model. This confirms the target hypothesis that the source region estimation accuracy is significantly increased if the geographical partitioning is performed considering the regional geophysical characteristics such as the tectonic plate boundaries.

5 Data and resources

In order to enable the reproducibility the results, the data and the source code used in this work are publicly available on https://github.com/rtous/deepquake.

Acknowledgements

This work is partially supported by the Spanish Ministry of Economy and Competitivity under contract TIN2015-65316-P, by the Spanish Ministry of Science and Innovation under contract PID2019-107255GB-C22, and by the SGR programmes (2014-SGR-1051 and 2017-SGR-962) of the Catalan Government and has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 777778 (MATHROCKS). We thank FUNVISIS for providing the seismic data subject of our current studies.

References

1. Allen, R.: Automatic earthquake recognition and timing from single traces. Bulletin of the Seismological Society of America **68**(5), 1521–1532 (1978)
2. Allen, R.: Automatic phase pickers: Their present use and future prospects. Bulletin of the Seismological Society of America **72**(6B), S225–S242 (1982)
3. Audemard, F.: Ruptura de los grandes sismos históricos venezolanos de los siglos xix y xx revelados por la sismicidad instrumental contemporánea (2002)
4. Audemard, F.: Revised seismic history of the el pilar fault, northeastern venezuela, from the cariaco 1997 earthquake and recent preliminary paleoseismic results. Journal of Seismology **11**(3), 311–326 (2007)
5. Chen, Y.: Automatic microseismic event picking via unsupervised machine learning. Geophysical Journal International **212**(1), 88–102 (2018)
6. Dai, H., MacBeth, C.: Automatic picking of seismic arrivals in local earthquake data using an artificial neural network. Geophysical Journal International **128**(3), 758–774 (1995)
7. Dai, H., MacBeth, C.: The application of back-propagation neural network to automatic picking seismic arrivals from single-component recordings. Journal of Geophysical Research: Solid Earth **102**(B7), 15105–15113 (1997)
8. Gentili, S., Michelini, A.: Automatic picking of p and s phases using a neural tree. Journal of Seismology **10**, 39–63 (2006)

9. Gillies, S., Bierbaum, A., Lautaportti, K., Tonnhofer, O.: Shapely: manipulation and analysis of geometric objects. Available online: github. com/Toblerity/Shapely (accessed on 15 June 2019) (2007)
10. Hernández, J.: Revisión de la sismicidad y modelo sismogénico para actualización de las evaluaciones de amenaza sísmica en la región norcentral de venezuela. In: IX Congreso Venezolano de Sismología e Ingeniería Sísmica. Caracas (May 2009)
11. Hernández, J., Schmitz, M.: Modelo sismogénico de venezuela para evaluaciones de la amenaza sísmica. In: XI Congreso Venezolano de Sismología e Ingeniería Sísmica. Caracas (July 2017)
12. Ibrahim, M.A., Park, J., Athens, N.: Earthquake warning system: Detecting earthquake precursor signals using deep neural networks (2018)
13. Janssen, V.: Understanding coordinate reference systems, datums and transformations. International Journal of Geoinformatics **5** (01 2009)
14. Michael, S., Julio, H., Cecilio, M., Jean, D., Victor, R., Vallée, M., Tagliaferro, M., Delavaud, E., Singer, A., Amaris, E., Danna, M., González, M., Leal, V., el., C., Franck, A.: Principales resultados y recomendaciones del proyecto de microzonificación sísmica de caracas. Revista Facultad de Ingeniería **26**, 113–127 (01 2011)
15. Mus, S., Gutiérrez, N., Tous, R., Otero, B., Cruz, L., Llácer, D., Alvarado, L., Rojas, O.: Long short-term memory networks for earthquake detection in venezuelan regions. In: Nicosia, G., Pardalos, P., Umeton, R., Giuffrida, G., Sciacca, V. (eds.) Machine Learning, Optimization, and Data Science. pp. 751–754. Springer International Publishing, Cham (2019)
16. Novianti, P., Setyorini, D., Rafflesia, U.: K-means cluster analysis in earthquake epicenter clustering. International Journal of Advances in Intelligent Informatics **3**(2), 81–89 (2017). https://doi.org/10.26555/ijain.v3i2.100, http://ijain.org/index.php/IJAIN/article/view/100
17. Rojas, O., Otero, B., Alvarado, L., Mus, S., Tous, R.: Artificial neural networks as emerging tools for earthquake detection. Computación y Sistemas **23**(2), 335–350 (08 2019). https://doi.org/10.13053/CyS-23-2-3197
18. Ross, Z.E., Meier, M.A., Hauksson, E., Heaton, T.H.: Generalized seismic phase detection with deep learning. Bulletin of the Seismological Society of America (2018)
19. Suárez, G., Nábelek, J.: The 1967 caracas earthquake: Fault geometry, direction of rupture propagation and seismotectonic implications. Journal of Geophysical Research: Solid Earth **95**(B11), 17459–17474 (1990)
20. Thibaut, P., Michaél, G., Marine, D.: Convolutional neural network for earthquake detection and location. Science Advances **4**(2) (2018)
21. Tiira, T.: Detecting teleseismic events using artificial neural networks. Computers & Geosciences **25**, 929–938 (09 1999). https://doi.org/10.1016/S0098-3004(99)00056-4
22. Tous, R., Alvarado, L., Otero, B., Cruz, L., Rojas, O.: Deep neural networks for earthquake detection and source region estimation in north-central venezuela. Bulletin of the Seismological Society of America **110**(5), 2519–2529 (06 2020). https://doi.org/10.1785/0120190172
23. Wang, J., liang Teng, T.: Identification and picking of s phase using an artificial neural network. Bulletin of the Seismological Society of America **87**(5), 1140–1149 (1997)
24. Wang, J., Teng, T.L.: Artificial neural network-based seismic detector. Bulletin of the Seismological Society of America **85**(1), 308–319 (1995)

25. Zhu, W., Beroza, G.C.: Phasenet: A deep-neural-network-based seismic arrival time picking method. http://arxiv.org/abs/1803.03211v1 (2018)