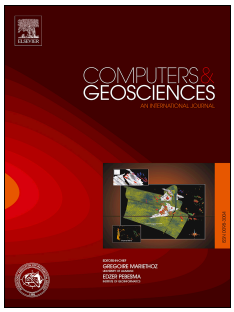
Accepted Manuscript



Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure

G. Asencio–Cortés, A. Morales–Esteban, X. Shang, F. Martínez–Álvarez

PII: S0098-3004(17)30608-8

DOI: 10.1016/j.cageo.2017.10.011

Reference: CAGEO 4039

To appear in: *Computers and Geosciences*

Received Date: 30 May 2017

Revised Date: 4 September 2017

Accepted Date: 22 October 2017

Please cite this article as: Asencio–Cortés, G., Morales–Esteban, A., Shang, X., Martínez–Álvarez, F., Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure, *Computers and Geosciences* (2017), doi: 10.1016/j.cageo.2017.10.011.

Earthquake Prediction in California Using Regression Algorithms and Cloud-based Big Data Infrastructure

G. Asencio–Cortés[a], A. Morales–Esteban[b], X. Shang[c], F. Martínez–Álvarez[a]

[a] *Division of Computer Science, Pablo de Olavide University of Seville, Spain – {guaasecor,fmaralv}@upo.es*
[b] *Department of Building Structures and Geotechnical Engineering, University of Seville, Spain – ame@us.es*
[c] *School of Resources and Safety Engineering, Central South University, China – shangxueyi@csu.edu.cn*

Abstract

Earthquake magnitude prediction is a challenging problem that has been widely studied during the last decades. Statistical, geophysical and machine learning approaches can be found in literature, with no particularly satisfactory results. In recent years, powerful computational techniques to analyze big data have emerged, making possible the analysis of massive datasets. These new methods make use of physical resources like cloud based architectures. California is known for being one of the regions with highest seismic activity in the world and many data are available. In this work, the use of several regression algorithms combined with ensemble learning is explored in the context of big data (1 GB catalog is used), in order to predict earthquakes magnitude within the next seven days. Apache Spark framework, H_2O library in R language and Amazon cloud infrastructure were been used, reporting very promising results.

Key words: Earthquake prediction, big data analytics, cluster computing, regression, ensemble learning

1. Introduction

Modern societies are threatened by natural risks and demand a proper preparation to reduce their impact. During the last years many initiatives have merged from modern societies in order to minimize the economical and human impact of natural disasters.

Natural risk is a concept embedded in the collective consciousness of modern societies. Against expected, an objective and universal definition of risk is yet to be established [1]. Nevertheless, it can be defined as a measure of the combined likelihood of occurrence of a threatening event and its potential consequences.

Natural disasters occur when a probable hazard turns into a real event. Then, potential consequences become real human and economic losses. Among natural disasters such as earthquakes, tsunamis, volcanic eruptions, hurricanes, tornadoes, floods and others, earthquakes stand out due to their devastating effects [2]. Earthquakes arrive suddenly and can destroy a whole city or region within seconds causing lost of lives or injures, property damage, social and economic breaks or environmental damage [3]. Moreover, many populated areas stand on seismic zones. Besides, earthquakes can produce correlated effects such as tsunamis [4], landslides [5] and liquefaction [6].

Seismic risk is a combination of seismic hazard and seismic vulnerability [7]. On the one hand, seismic hazard represents a potentially damaging seismic event that can cause damage. On the other hand, the potential consequences are the existing vulnerabilities that show the susceptibility to the damaging effect of the hazard.

Big data analytics has emerged as a very powerful technique. It is typically used to examine huge datasets in order to extract useful information and discover patterns [8]. When such big datasets must be analyzed, computational resources increase and traditional machine learning algorithms require new parallelized implementations that must be launched in clusters [9].

For all the aforementioned, there is a worldwide trend to enhance our understanding of earthquakes in order to increase our ability to manage them [10]. In this paper, earthquake prediction in one of the most seismic and populated areas of the world -California- is explored. So far, standard machine learning