

Article

Not peer-reviewed version

Earthquake Footprints for Predicting Events

[Kieran Greer](#) *

Posted Date: 11 July 2023

doi: [10.20944/preprints202307.0584.v2](https://doi.org/10.20944/preprints202307.0584.v2)

Keywords: earthquake, footprint, predict events, cluster, frequency grid



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

Earthquake Footprints for Predicting Events

Kieran Greer

Distributed Computing Systems, Belfast, UK; jelisavcic@mdpi.com

Abstract: This paper considers the problem of predicting earthquakes. It uses a small amount of information to create a descriptive key that can be used as a footprint to describe an event. A frequency grid clusters events that occurred at the same time and then the algorithm measures the history of these events over preceding days, in particular the gaps when the events did not occur. The clusters with the time gaps can be used to create the footprint description and results suggest that seismic events can in fact be traced using this, and subsequently recognised again, if the same conditions occur. Greek and USA datasets have been looked at and the prediction accuracy can be 70% or better. The author therefore suggests that this is an interesting method that deserves attention.

Keywords: earthquake; footprint; predict events; cluster; frequency grid

1. Introduction

This paper considers the problem of predicting earthquakes. This difficult problem is a relatively new science that has been approached from different directions, but primarily by studying seismic and electromagnetic activity these days. What exactly causes earthquakes is still not clear, including the underlying mechanisms, and modelling this to allow a system to predict future events has to date yielded only modest results. After an earthquake has happened for example, do the same conditions still exist, so that a future event could be predicted. This is in fact a recognised problem where the answer is that some earthquakes repeat and some do not, so only some future earthquakes would be predictable. Most research has attempted to understand how the earthquake works and tried to build a model of that. Until recently, the available data has been a bit patchy, but now it is possible to map the whole planet, with regard to its electromagnetic footprint. This should make it easier to build machine learning models in the future. This paper suggests a machine-learning approach that uses a relatively small amount of data. It proposes that earthquakes do in fact have a footprint that can be traced and recognised. Rather than trying to understand the underlying mechanisms that cause the earthquake, the clustering method is evidence-based and builds up a picture based only on the evidence that is provided. This has some advantages over knowledge or model-based methods, including neural networks, because it can more easily learn stochastic data where random events may occur and which may be more appropriate for environmental data. If the clustering method can produce slightly different cluster sets to other algorithms, then it may be possible to realise different conclusions and in fact create a prediction algorithm. The author is unlikely to continue the research and so the software is available to download and use for free [6].

The rest of the paper is organised as follows: section 2 gives some related work. Section 3 describes the methodology of the new prediction algorithm. Section 4 gives the results of running tests using the new method. Finally, section 5 gives some conclusions on the work.

2. Related Work

An earlier machine learning system called VAN [15,16] was shown at the time to produce better results for the Greece earthquakes and was a temporal clustering method. The VAN method tries to recognise changes in the rock's electromagnetic emissions, with the underlying theory that rocks under stress emit different types of signal. It has since been updated [15] with the concept of natural time, which is a time series analysis technique that puts weight on a process based on the ordering of events. However, the prediction results of the method were questioned and it has both supporters

and critics. As stated in [10]: 'Why is temporal clustering such an important issue? Primarily because some variation in natural phenomena, such as electric field variation, which might follow earthquakes, would typically precede late events in a cluster. The electrical variations might thus appear to have some predictive capability, but this would actually come purely from the clustering of earthquakes.' This indicates that earthquake clusters are important. The paper [14] introduces a new model which considers that the fundamental aspects of the strain accumulation and release processes are critical to causing earthquakes. They also state that a problem with current models is that they assume that large earthquakes release all accumulated strain, despite evidence for partial strain release in earthquake histories showing clusters and gaps. The following sections will show an agreement with both these papers. While their design may be model-based however, this paper uses an evidence-based design.

In [13] they suggest that an acceptable test for earthquake accuracy might be the ability to predict an earthquake in a region of 50km from the epicentre and up to 21 days before the event. However, only a 5% success rate with these criteria is deemed a good result. A more recent summary about machine learning methods can be found in [2], for example. It notes that much of the progress has been in developing the data catalogs to train the AI models on, such as their own STEAD dataset. In fact, there have been recent claims of success using AI models [1,9,11,12], where some accuracy quotes are over 90%. The paper [1] measures the upper atmosphere's Ionospheric total electron content, which originates in the rock, while the paper [9] considers water vapour in volcanic regions. Not only different methods, but also different aspects of the earthquake are now predicted and in practice there is still no generally accepted reliable method.

The design of this paper makes use of a new clustering algorithm called a Frequency Grid [8]. This was also used in [7] to predict energy usage in households, where Dr. Yixin Bi is also an expert in modelling electromagnetic data [3]. The frequency grid is an event-based clustering method. It reads a dataset where each row lists events that occurred at the same time. These associations produce sets of count values that represent which events more often occurred together. The grid is entropy-based however, rather than local associations only, where the aggregation into a single table can produce a holistic view of the associations. Because the clustering process is event-based, it does not have to produce a consistent underlying model or theory. A neural network, for example, may need to map from input to output using a consistent or continuous function, but the frequency grid does not have to do this. As a result, it can maybe model stochastic data more easily and so environmental data, which may include random elements, can be modelled more easily. Another idea taken from the energy paper [7] is to discretise the seismic data into bands, thus allowing the grid a finer level of granularity. This is described further in the following sections.

3. Methodology

The proposed method is to try to recognise the events that led up to a major earthquake and represent them in some unique way. If clusters of earthquakes that occur together can be realised, then it should be easier to produce a unique description. If these clusters produced one major event, then they were not involved in a different major event, for example. The method uses the frequency grid to cluster the seismic events, which are represented by their location, magnitude, and time that they occurred. From these clusters, the larger seismic events can be found and it can be determined what occurred with them. The frequency grid is category-based and so the events need to be translated from numerical values to text-based values. Each event can therefore be represented by a key that stores its longitude – latitude location, and then a magnitude representation. The magnitude part is discretised into bands as follows: 1 band represents 0.5 of a magnitude size, so an earthquake of size 5 would have a band value of 10. Therefore, a seismic event would be represented by something like 30:20:8, where the longitude of the event would be 30, the latitude would be 20 and the actual magnitude would be 4. Then for each time unit, which is currently set to days, all events that occurred during that time unit are added to the dataset as a row of data. When the frequency grid reads the dataset, it will try to associate the events in the same row.

3.1. Discretise the Frequency Grid

To make it slightly more accurate, it has been decided to discretise the earthquake magnitude, rather than use a real value. Then, each row or column in the frequency grid is represented by a location plus a discretised band. In this way for example, location A can be associated with location B when the seismic magnitude is small, but with location C when the magnitude value is large. Consider the following example: locations A (1,1), B (1,2) and C (2,1) all have earthquake events that occur on the same days, as shown in Table 1. On day 1, only 2 events occur, both with a magnitude of 1 and so the frequency grid would cluster A and B together with keys something like A:1 and B:1. On day 2 there are 4 events, but with different magnitudes. For the smaller magnitude, A and B still occur together, but for the larger magnitude A occurs with C. The frequency grid would therefore produce a second cluster for the larger magnitude that would be A:2 and C:2. Putting these counts into the frequency grid leads to Table 2, where tracing through this manually can even show the clusters.

Table 1. Example of earthquake events.

Date	Token	Longitude	Latitude	Magnitude
1	A	1	1	1
1	B	1	2	1
2	C	2	1	2
2	A	1	1	2
2	B	1	2	1
2	A	1	1	1

Table 2. Frequency Grid for the Earthquake Events. Clusters are A:1, B:1 and A:2, C:2.

	A:1	A:2	B:1	B:2	C:1	C:2
A:1	x	0	2	0	0	0
A:2	0	x	0	0	0	1
B:1	2	0	x	0	0	0
B:2	0	0	0	x	0	0
C:1	0	0	0	0	x	0
C:2	0	1	0	0	0	x

3.2. Creating the Event Footprint

The frequency grid clustering is only the first stage of the full prediction algorithm. After the dataset is created and the frequency grid clusters generated, the significant events can be found, based on their magnitude key part. Then, from the original dataset, the date of the significant event can be retrieved. The algorithm then wants to determine if the days leading up to the significant event indicated that it would occur. The algorithm can look at data rows in the dataset for x days before the event. After selecting that subset, rows in it are then removed if they do not contain any of the events in the significant cluster. This then leaves the algorithm with the cluster values and some data rows for some days before the event. It is these data rows that are then able to produce the footprint that can describe the significant event. In fact, the current version makes use of the dates only. Not all days would store related events and so there may be gaps in the time series. A first part to the key is therefore to look at these gaps and average over them. A second part then counts how many of the cluster events occurred during each day and averages that. An event footprint can then be created using these average values and the number of days that they were created from. Because the significant event itself occurs only once, it is really the more minor events clustered with it that are

being traced, to create the footprint. A surprising result when looking at the key value as a footprint, was that it was mostly unique for the significant event, but the accuracy did drop when the significant threshold was reduced.

3.3. Prediction Algorithm

The test of accuracy would then be for the program to be able to predict the significant events by summarising the days before it and creating the footprint value from them. If a recognised footprint was realised, then that would indicate that the related significant event was likely to happen. The prediction algorithm was therefore quite similar to the one that created the footprints, with a few changes. It started at day 1 of the dataset and would cumulate the data rows in turn until a maximum number for a time window w was used. Then for each of the significant event clusters, a new subset would be created that contained only the data rows relevant to that cluster. From that subset the footprint key could be created. If the key was within a certain error margin of the known footprint for the cluster, then the significant event would be flagged for the last day in the time window. This would be repeated, a row at a time, through the whole dataset, where earlier days would be removed when the window moved past them. At the end of this process, the program would print out what days it considered each significant event to potentially occur on.

4. Testing and Results

A program has been written in the Java programming language to test the theory. The original Greece [4] and USA [5] datasets have been re-formatted slightly to be read by the program. The band size was set to 0.5, which meant that an increase of this amount in the seismic magnitude would place the event in a different category. The threshold was set to 12, meaning that only seismic events of size 6 or above would be considered as significant. The time window was set to 200, meaning that the program would look at 200 days before the significant event day, but clearly these values can be changed for a new set of tests. The Greece data was used from the start of 2005 only, so the earlier rows were removed first. It was decided that a time unit of days would be best and so the USGS dataset, which has a smaller timeframe, was aggregated to be in days rather than hours. The only columns that are required are the date, longitude, latitude, and the event magnitude. The example in the following sections is for the Greece dataset.

4.1. Data Rows and Clusters

The program firstly reads the formatted dataset and creates lists of events for each day in the record. It then feeds these lists into the frequency grid that generates clusters of events that most often occur together. The program is then given a lower threshold for what a significant event might be. A threshold value of 12 would look for major events of size 6 and above, for example, where a data row could be represented by something like the following.

Date (in days)	Events on that day
01/01/2005 00:00:00	38:24:3, 25:20:4, 40:42:5

The data row describes 3 events that occurred on the 1 of January 2005. The program can process in hours, but the preferred time unit is days and so the time part is set to 0. The first event '38:24:3' indicates a latitude of 38, a longitude of 24 and a discrete band size of 3. The frequency grid would therefore have associated these events together and clusters with events that pass the threshold are retrieved and stored as significant or target events. For example, the following was a significant cluster for the Greece dataset, because one event had a discrete value of 13.

Significant Cluster	38:22:6, 40:19:7, 6:30:8, 40:23:8, 39:25:8, 38:22:13
---------------------	--

4.2. Footprint Key

The footprint is created from all events in a cluster, even if they are smaller in size. This is translated over into a list of gaps between days when events occurred and also the number of related events that occurred on each day. To determine if a related event occurred, the longitude, latitude and magnitude were all considered. The footprint key can then be created from this list and may look something like the following.

Footprint (Day Count / Av Day Gap / Av Event Count)	50 / 2.5 / 3
---	--------------

This represents the days leading up to the event as follows: there were 50 days in the time window before the event that contained any of the events in the cluster. In these 50 days, the average gap between these days was 2.5 and there was an average of 3 relevant events each day. It turns out that this description is quite unique and can describe the significant event of the cluster while excluding descriptions for the other significant clusters.

4.3. Analysis and Predictions

The program can then run an analysis phase, where it selects a time window related to a significant event and analyses the data rows that occurred in that time window. It notes the events that are part of the significant event cluster and notes when they occurred. From this it is able to generate the footprint key and store it with the event description, as has just been described. This can then be used as the marker that the prediction should try to find. The analysis data is written to a file and then read by a simulation program that makes predictions on when the significant events are likely to happen, as described in section 3.3. The footprint does not have to be exact. For this set of tests the error margin was set to 5% difference in any of the 3 parts to the footprint key. So the footprint would be considered to be the same, only if all 3 parts were within an error margin of 5% of the related actual footprint part. One thing to note however is that a significant event does not necessarily have a history that can be traced. Some events appear to happen without warning, while other events have a long history trace that can be used to calculate the footprint.

4.3.1. Greece Dataset

For the Greek dataset, most of the significant events did in fact have a footprint. Figure 1 shows the significant events with a threshold of value 12 or above (magnitude value of 6 or above) and only 1 event occurred without an earlier trace. It is therefore not possible to make a prediction for that event. For the other events however, the predictions are listed in Figure 2. Most of the significant events have been predicted and within reasonable time of the real event. Some of the time spans are very short, but if the data was tested first, then it would be known to expect a short time span there. One event is a lot more messy however. It occurred on 16 July 2008, but its footprint has been repeated several times over the dataset. It happens to be the case that the other flagged events were clustered mostly in the same region and so that may have produced a similar footprint. Even if that event is not included, the author would suggest that the accuracy is 6 good predictions from 9 plus one half-prediction, which still gives 70% accuracy.

With Footprint		Without Footprint	
Key	Date	Key	Date
2/46.5/3 @ 36:23:12	9 January 2006	0/0.0/3 @ 38:27:13	31 October 2020
18/10.7/6 @ 37:23:12	7 January 2008		
83/2.4/2 @ 38:22:13	9 June 2008		
37/5.4/3 @ 36:28:12	16 July 2008		
12/2.7/12 @ 36:27:12	2 April 2011		

13/14.8/29 @ 40:25:12	25 May 2014		
28/6.6/18 @ 39:21:12	18 November 2015		
1/142.0/17 @ 40:22:12	4 March 2021		

Figure 1. Significant events for the Greece Dataset – Footprint at Event.

Footprint	Event	Date	Predictions
2/46.5/3	36:23:12	9 January 2006	13 October 2005 to 9 January 2006
18/10.7/6	37:23:12	7 January 2008	6 January 2008 to 8 January 2008
83/2.4/2	38:22:13	9 June 2008	15 February 2008 to 9 June 2008
37/5.4/3	36:28:12	16 July 2008	13 June 2006 to 22 June 2006 3 April 2007 to 6 April 2007 9 May 2007 to 13 May 2007 5 July 2008 to 19 July 2008 3 June 2010 to 13 June 2010 14 June 2010 to 20 June 2010 29 September 2010 to 30 September 2010 3 October 2010 15 August 2020 to 29 August 2020 21 September 2020 to 24 Sept 2020 14 December 2020 to 15 December 2020
12/2.7/12	36:27:12	2 April 2011	29 March 2011 to 4 April 2011
13/14.8/29	40:25:12	25 May 2014	20 May 2014 to 25 May 2014
28/6.6/18	39:21:12	18 November 2015	17 March 2012 to 22 March 2012 16 November 2015 to 18 November 2015
1/142.0/17	40:22:12	4 March 2021	1 March 2021 to 4 March 2021

Figure 2. Predictions for the Greece Dataset.

4.3.2. USA Dataset

For the USA dataset however, over 50% of the significant events do not have a footprint. Figure 3 shows the significant events with a threshold of value 12 or above and only 8 have a traceable footprint, while 9 do not. Of the 8 that have a footprint, predictions could be made for 7 of them, as listed in Figure 4. These predictions are quite close however and so in real terms that is still a 40% accuracy.

With Footprint		Without Footprint	
Key	Date	Key	Date
2/1.0/95 @ -59:-25:12	7 March 2022	0/0.0/0 @ 12:-87:13	22 April 2022
3/1.0/108 @ -20:-178:12	8 March 2022	0/0.0/0 @ 24:123:12	10 May 2022
90/1.0/92 @ -58:149:12	5 June 2022	0/0.0/0 @ -16:-174:12	20 May 2022
92/1.0/138 @ -9:-71:13	9 June 2022	0/0.0/0 @ 24:121:12	21 June 2022
156/1.0/124 @ 44:148:12	8 August 2022	0/0.0/0 @ 56:166:12	21 Sept 2022
1/192.0/139 @ 23:121:13	Twice	0/0.0/0 @ 4:96:12	23 Sept 2022
1/50.0/123 @ -5:101:13	19 November	0/0.0/0 @ 18:-103:13	23 Sept 2022
53/3.6/110 @ -15:-173:13	5 December 2022	0/0.0/155 @ 23:121:13	Twice
		0/0.0/0 @ -26:178:13	10 Nov 2022 x3

Figure 3. Significant events for the USA Dataset – Footprint at Event.

Footprint	Event	Date	Predictions
2/1.0/95	-59:-25:12	7 March 2022	6 March 2022 to 7 March 2022
3/1.0/108	-20:-178:12	8 March 2022	7 March 2022 to 8 March 2022
90/1.0/92	-58:149:12	5 June 2022	31 May 2022 to 9 June 2022
92/1.0/138	-9:-71:13	9 June 2022	4 June 2022 to 9 June 2022
156/1.0/124	44:148:12	8 August 2022	31 July 2022 to 15 August 2022
1/50.0/123	-5:101:13	19 November	2 November 2022 to 12 Dec 2022
53/3.6/110	-15:-173:13	5 December 2022	21 November 2022 to 12 Dec 2022

Figure 4. Predictions for the USA Dataset.

The test was also run on both datasets with a threshold of 10 or more (magnitude of 5 or more), and 11 or more, where Table 3 gives a summary of all the results. For the lower threshold, the clusters were not predicted as accurately and some of them could be very incorrect. The story, for example, is a bit different for the USA data with a 10 threshold, but most of the misses are because no footprint was available. Of the 18 predictions that were made, 15 of them would be acceptable. The table therefore gives an accuracy score for significant events, only when predictions were made and then also for all significant events.

Table 3. Summary of the test results.

Dataset	Magnitude	% with Footprint	% Prediction with Result	% Prediction All
Greece	6+	88	81	70
Greece	5.5+	95	63	60
Greece	5+	94	53	50
USA	6+	47	88	40
USA	5.5+	41	100	36
USA	5+	33	80	20

5. Conclusions

This paper has suggested a method of producing footprints with which to recognise seismic events. The process requires training an algorithm on previous events and so if the conditions change, the process may not work very well. It is also evidence-based, which means that it can map more easily to the data, but that it has very little knowledge or understanding of that data. It might therefore not be possible to transfer results from one region to another region. One idea when measuring gaps between days was that there were more gaps leading up to an event, so this could be a type of general indicator, but further tests did not show this as being definitive. The gaps may indicate when forces are acting against each other, for example, when if one set of forces gives way to another, it could lead to catastrophic failure.

The footprint is generated from only a small amount of information and sometimes it cannot be created. It would be possible to add other information, but the author does not know what that might be. With the clusters however, even the current key is mostly unique, for the Greece and USA datasets. Because the significant event itself occurs only once, it is really the more minor events clustered with it that are being traced, to create the footprint. The current program is only a first attempt, but it has been able to show a proof of concept and the accuracy of the predictions are probably at an acceptable level. Therefore, the author would suggest that this is an interesting method that deserves further investigation.

References

1. Asaly, S., Gottlieb, L.-A., Inbar, N. and Reuveni, Y. (2022). Using Support Vector Machine (SVM) with GPS Ionospheric TEC Estimations to Potentially Predict Earthquake Events. *Remote Sens.*, Vol. 14, p. 2822. <https://doi.org/10.3390/rs14122822>.
2. Beroza, G.C., Segou, M. and Mostafa Mousavi, S. (2021). Machine learning and earthquake forecasting - next steps. *Nat Commun*, Vol. 12, No. 4761. <https://doi.org/10.1038/s41467-021-24952-6>.
3. Christodoulou, V., Bi, Y. and Wilkie G. (2019). A tool for Swarm satellite data analysis and anomaly detection. *PLoS ONE* Vol. 14, No. 4: e0212098. <https://doi.org/10.1371/journal.pone.0212098>.
4. Dataset, Greece. (2023). Kaggle, <https://www.kaggle.com/datasets/nickdoulos/greeces-earthquakes>.
5. Dataset, USA. (2023). Kaggle, <https://www.kaggle.com/datasets/the-devastator/uncovering-geophysical-insights-analyzing-usgs-e>.
6. Greer, K. (2023). Earthquake Footprint Software, <https://github.com/discompsys/Earthquake-Footprint>.
7. Greer, K. and Bi, Y. (2022). Event-Based Clustering with Energy Data, *WSEAS Transactions on Design, Construction, Maintenance*, Vol. 2, Art. #26, pp. 197-207. DOI: 10.37394/232022.2022.2.26.
8. Greer, K. (2019). New Ideas for Brain Modelling 3, *Cognitive Systems Research*, Vol. 55, pp. 1-13, Elsevier. DOI: <https://doi.org/10.1016/j.cogsys.2018.12.016>.
9. Hiraishi, H. (2022). Earthquake Prediction Software on Global Scale. *Journal of Geoscience and Environment Protection*, Vol. 10, pp. 34 - 45. <https://doi.org/10.4236/cep.2022.103003>.
10. Kagan, Y.Y. and Jackson, D.D. (1996). Statistical tests of VAN earthquake predictions: Comments and reflections, *Geophysical Research Letters*, Vol. 23, No. 11, pp. 1433 - 1436.
11. Kavianpour, P., Kavianpour, M., Jahani, E. and Ramezani, A. (2023). A cnn-bilstm model with attention mechanism for earthquake prediction. *The Journal of Supercomputing*, pp. 1 - 33.
12. Laurenti, L., Tinti, E., Galasso, F., Franco, L. and Marone, C. (2022). Deep learning for laboratory earthquake prediction and autoregressive forecasting of fault zone stress, *Earth and Planetary Science Letters*, Vol. 598, p.117825.
13. Luen, B. and Stark, P.B. (2008). Testing earthquake predictions. In *Probability and statistics: Essays in honor of David A. Freedman* (Vol. 2, pp. 302-316). Institute of Mathematical Statistics.
14. Neely, J.S., Salditch, L., Spencer, B.D. and Stein, S. (2023). A More Realistic Earthquake Probability Model Using Long-Term Fault Memory. *Bulletin of the Seismological Society of America*, Vol. 113, No. 2, pp. 843 - 855.
15. Varotsos, P., Sarlis, N. and Skordas, E. (2002), Long-range correlations in the electric signals that precede rupture, *Physical Review E*, Vol. 66, No. 1: 011902, Bibcode:2002PhRvE..66a1902V, doi:10.1103/physreve.66.011902, PMID 12241379.
16. Varotsos, P., Alexopoulos, K. and Nomicos, K. (1981). Seven-hour precursors to earthquakes determined from telluric currents, *Praktika of the Academy of Athens*, Vol. 56, pp. 417-433.