Proximity to Past Earthquakes as a Least Astonishing Hypothesis for Forecasting Locations of Future Earthquakes

Alan L. Kafka and John E. Ebel
Weston Observatory
Department of Geology and Geophysics
Boston College
381 Concord Rd.
Weston, MA 02493

Short Note submitted to:
Bulletin of the Seismological Society of America

July 6, 2009

Under Review

Abstract

The Cellular Seismology (CS) method of Kafka (2002, 2007) is presented as a "least astonishing" null hypothesis that serves as a useful standard of comparison for other, more complex, spatial forecast methods (i.e., methods that forecast the locations, but not the times, of earthquakes). Spatial forecast methods based on analyses of earthquakes in California such as that of Ebel *et al.* (2007) and the Pattern Informatics method of Rundle *et al.* (2002, 2007) provide opportunities for comparing methods that incorporate information about rates of seismicity with a method (i.e., CS) that only assumes that future earthquakes will occur near epicenters of past earthquakes. The Ebel *et. al.* (2007) 5-year forecast model (E07) maps the spatial distribution of rates of seismicity, and the Pattern Informatics model (PI) not only considers rates of seismicity, but also incorporates changes in rates of seismicity as a measure of the potential for future earthquakes to occur at some location. Our comparison of success rates of the E07 method and the PI method with CS for earthquakes in California has yet to reveal any compelling evidence that inclusion of rates or changes in rates in the forecast model improves the performance of the forecast.

Introduction

If an earthquake forecast is to be considered successful, it should perform better than a reasonable "least astonishing" null hypothesis. But what should that null hypothesis be? Here, we propose that, for the case of spatial forecasts (i.e., forecasting the locations, but not the times of earthquakes), the Cellular Seismology (CS) method of Kafka (2002, 2007) is an appropriate least astonishing null hypothesis model.

A random spatial distribution of epicenters is an obvious first-order choice as a null hypothesis, and any forecast that is claimed to be successful should certainly be able to perform better than a random distribution. But, performing better than a random spatial distribution is not particularly instructive because (at least in plate boundary regions) it is clear that the spatial distribution of earthquakes is not random. We envision CS, which considers the region surrounding any past earthquake to be a potential source point of future earthquakes, as the next logical step beyond the random hypothesis for the choice of a least astonishing hypothesis for spatial earthquake forecasts. CS does assume that the spatial distribution of earthquakes is not random, but (as we would want of a least astonishing hypothesis) does not depend on any other preconceived notions (beyond the notion of proximity to past earthquakes) regarding the processes that are causing the earthquakes in a given region.

To illustrate how success rates of other forecast models can be tested against a CS null hypothesis, we evaluate two examples of spatial forecast models that incorporate more information than CS. Specifically, two forecast models based on analyses of earthquakes in California, the model of Ebel *et al.* (2007) and the Pattern Informatics model of Rundle *et al.* (2002, 2007), provide opportunities for comparing methods that incorporate information about rates of seismicity with a method (i.e., CS) that only assumes that future earthquakes will occur near epicenters of past earthquakes. The Ebel *et al.* (2007) 5-year forecast method (E07) maps the spatial distribution of *rates* of seismicity, and the Pattern Informatics method (PI) not only considers rates of seismicity, but also incorporates *changes in rates* of seismicity as a measure of the potential for future earthquakes to occur at some location.

The Cellular Seismology Method

Details of the CS method are described by Kafka (2007); here we summarize the essence of the method. Figure 1 shows a hypothetical region in which earthquakes are occurring, and for which the CS method (which is analogous to the configuration of a cellular phone system) is being applied. As illustrated for this hypothetical case, we construct circles of a given distance radius around each epicenter in an earthquake catalog (the “before” catalog), and investigate the percentage of later-occurring earthquakes (the “after” catalog) that were located within that distance of at least one previous earthquake. The distance radius is varied so that the area covered by the interiors of the circles fills a given percentage of the map of the study area. If the epicenter of an after earthquake lies within the given distance of at least one past earthquake, we call that a “hit”, and the observed percentage of hits is investigated for a given combination of before and after catalogs.

Although CS is a very simple method of characterizing the relationship between past seismicity and locations of future earthquakes, we tried more complex approaches (including Gaussian smoothing, following the method of Frankel, 1995) and found the results to be quite similar to what we obtained using this simpler CS method (e.g., Kafka and Levin, 2000). Thus, we have adopted the CS method as a simple and straightforward (time-independent) way of measuring the extent to which past seismicity delineates zones where future (i.e., later occurring) earthquakes occur.

Comparison of Performance of the E07 Method with that of CS

The E07 method was developed as part of the Regional Earthquake Likelihood Models (RELM) study of California (e.g. Field, 2007; Schorlemmer, *et al.*, 2007). The method involves dividing the region into an array of cells 0.3° on a side (Figure 2), and determining the rate of earthquake activity for each cell (based on a before catalog). In implementing this method for the RELM region, Ebel *et al.* (2007) chose to remove foreshocks and aftershocks from the before catalog (using the time-space windows of Gardner and Knopoff, 1974) to avoid the situation in which dependent events in cells near regions of large earthquakes dominate the map of the spatial distribution of the rates.

To compare the success rate of the E07 method with that of CS, we applied the CS and E07 methods to the same region and the same before catalog as was analyzed by E07 (Figure 2). The before catalog for this case consists of $M \geq 4.0$ earthquakes downloaded from the Advanced National Seismic System (ANSS) catalog that fall within the RELM study area and that occurred from the beginning of 1932 to the end of 2004 (see Data and Resources Section). The after catalog for this analysis consists of ANSS earthquakes with $M \geq 4.5$ that occurred from the beginning of 2005 to the end of 2008, and includes 48 events.

When comparing two seismicity-based forecasts, it is important to consider not only the methods themselves, but also the before catalogs that are used to test the forecasts. Rundle *et al.* (2007) noted that it is important to be careful that the before catalog chosen for testing the success rate of a forecast method is not chosen retrospectively so that it is clear that the catalog/forecast method combination is strictly prospective. We follow that principle strictly for the PI/CS comparison described in the next section (a case in which the two methods were developed for regions with significant spatial overlap) by using exactly the same before catalog as that used by Kafka (2002, “K02”) in his development of the CS method (see Data and Resources Section). In the case of this E07/CS comparison, however, we decided that limiting the analysis to the before catalog used by K02 was too restrictive. The K02 before catalog covers a significantly smaller part of California than the E07 study region, and we wanted to take this opportunity to test the applicability of the CS method on a larger scale than if we were to limit the analysis to the region covered by K02. Thus, for the E07/CS comparison, the CS before catalog used here is the same as that used by E07.

For this E07/CS analysis, we varied the size of the CS circles such that the interiors of the circles cover a given percentage of the map area within the RELM study area polygon, and we varied the E07 cutoff for the rates in the cells such that cells with a given cutoff rate or higher cover a given percentage of map area. Figure 3 shows the percentage of hits as a function of map area covered by the CS before catalog circles, as well as the percentage of hits for a given E07 rate cutoff as a function of map area covered by the E07 cells corresponding to that rate cutoff. For this analysis, the success rates are similar for the two methods (Figure 3), and the statistical analysis presented below shows that the differences in success rates is not significant (at the 95% level of significance).

Comparison of Performance of the PI Method with that of CS

To investigate the extent to which incorporating *changes in rates* of seismicity in the forecast model affects the success rate of the model relative to that of CS, we compare the PI method to CS. The following comparison of CS versus PI is based on the CS method as it was published by K02, and it uses exactly the same before catalog as K02. Since the before catalog used by K02 is the same as that used by Kafka and Levin (2000), it was chosen before the occurrence of any of the post-2000 earthquakes that we use for the after catalog in this analysis. Rundle *et al.* (2007) expressed concern that in our previous comparison of PI with CS (Kafka and Ebel, 2007), we used a different before catalog than that of the original CS method, which might have given us the opportunity to modify our forecast model to produce optimal results. By using the identical before catalog as that of K02, and an after catalog consisting of events that occurred after that before catalog was chosen, we insure that this is a strictly prospective test and that no retrospective optimization is occurring here.

In this comparison, we also limit the analysis to earthquakes (both before and after 2000) that lie within the map area common to that covered by both K02 and Rundle *et al.* (2002, "R02"), i.e., the study area shown in Figure 4. The after catalog for this analysis consists of earthquakes in the National Earthquake Information Center (NEIC) catalog with M≥5.0 that occurred from the beginning of 2000 to the end of 2008, and includes 21 events (see Data and Resources Section).

The PI "hotspots" of R02 (Figure 4) cover 14% of the land in the study area, and for that percentage of area there are either 17/21 (81%) or 18/21 (86%) hits, depending on whether or not the 2003 San Simeon earthquake (which is on the boundary of one of the cells) is considered a hit or a miss. Using a CS circle radius that covers that same 14% area, there are 19/21 (90%) hits for the CS method.

Figure 5 shows the percentage of land area versus percentage of hits for the CS forecasts applied to this same after catalog, along with the percentage of hits results for the PI method. For 14% forecast map area, the success rate is similar for these two methods (Figure 5), and the statistical analysis presented below shows that the difference in success rates is not significant (at the 95% level of significance).

Statistical Significance of Differences in Performance of Forecast Models

To test for the statistical significance of observed differences in performance of a given forecast model (FM1) versus some other forecast model (FM2), we model the forecast and testing process as a binomial experiment with success defined as a hit and failure defined as a miss. We treat the FM1 success rate (π_0) as the null hypothesis, and the alternative hypothesis is that FM2 has a higher success rate than FM1. If π is the observed percentage of hits, then the *p-value* for the significance test is the binomial probability of observing π or greater percentage of hits, given that π_0 is the "true" rate of success (i.e., given the null hypothesis).

Using this approach, we calculated *p-values* for all cases shown in Figure 3 where the E07 method yielded a higher percentage of hits than the CS method, and we found that in none of those cases was the difference statistically significant at the 95% level. The *p-values* for this E07/ CS comparison ranged from 0.14 to 0.55, with a mean of 0.34 and a standard deviation of 0.15.

Given the higher success rate for the CS method than the PI method, with the same percentage of map area covered (Figure 5), one might argue that the CS method is an improvement over the PI method. However, using the above formulation, we find that the difference is not statistically significant at the 95% level. *The p-value* for this PI/CS comparison is 0.21 if we assume 17/21 hits for PI, and is 0.41 if we assume 18/21 hits for PI.

Discussion and Conclusions

None of the analyses discussed above reveal any statistically significant evidence that the inclusion of rates and/or changes in rates in the forecast model improves the performance of the forecasts beyond that of the least astonishing CS hypothesis.

Although certainly not an exhaustive analysis of the questions addressed here, the results of this study do suggest the possibility that there might not be any precursory signal in the form of rates of seismicity and/or changes in rates of seismicity at a given location. Does this mean that the search for those signals should be abandoned? We believe not. In this study we have only compared two examples of time-dependent forecast models to the simpler CS model, and both of those comparisons were in the same region. It will require many more such hypothesis tests in many more regions before it would be reasonable to conclude that proximity to past earthquakes is the only predictor of future earthquake locations.

It is intuitively reasonable to expect that patterns in rates of seismicity and/or changes in rates of seismicity at a given location are a reflection of the physical processes that eventually lead up to the occurrence of earthquakes. On the other hand, it is possible that, while patterns in rates of seismicity do reflect the physical processes leading up to an earthquake, any given pattern might be just as likely to indicate an impending large earthquake as to be an indicator of a lower probability of an impending earthquake. If that is indeed the case, then perhaps Cellular

Seismology, simple as it is, may actually be a measure of “all we can know” about the future occurrence of earthquakes. Hard as it may be for seismologists (and the public?) to accept such a conclusion, the results of this study suggest that such a possibility needs to be considered.

In the meantime, we present Cellular Seismology as an appropriate standard of comparison for spatial forecast models. We propose that spatial forecast methods should be tested against the Cellular Seismology method to ascertain whether or not they perform better than this least astonishing hypothesis model.

Data and Resources

Earthquake catalog data collected specifically for the analyses presented in this study were from the Advanced National Seismic System (ANSS) catalog (www.ncedc.org/anss), and the National Earthquake Information Center (NEIC) catalog (neic.usgs.gov). The before catalogs of the K02 and R02 studies were both collected from the Southern California Earthquake Center catalog (www.scec.org).

Acknowledgments

The research and writing for this study was undertaken during a sabbatical leave granted to the first author by Boston College for the 2008-2009 academic year. We are grateful to the BC College of Arts and Sciences, as well to our colleagues in the BC Department of Geology and Geophysics, for their support in providing the first author with the extended research and writing time necessary for us to complete this project.

References

Ebel, J. E., D. W. Chambers, A. L. Kafka, and J.A. Baglivo (2007). Non-poissonian earthquake clustering and the Hidden Markov Model as bases for earthquake forecasting in California, *Seisml. Res. Lett. 78*(1), 57-65.

Field, E. H. (2007). Overview of the Working Group for the Development of Regional Earthquake Likelihood Models, *Seism. Res. Lett. 78*(1), 7-16.

Frankel, A. (1995). Mapping seismic hazard in the central and eastern United States, *Seismol. Res. Lett. 66*(4), 8-21.

Gardner, J. K., and L. Knopoff (1974). Is the sequence of earthquakes in southern California with aftershocks removed Poissonian?, *Bull. Seism. Soc. Am.*, 64, 1363-1367.

Kafka, A. L (2002). Statistical analysis of the hypothesis that seismicity delineates areas where future large earthquakes are likely to occur in the central and eastern United States, *Seism. Res. Lett. 73*(6), 990-1001.

Kafka, A. L. (2007). Does seismicity delineate zones where future large earthquakes are likely to occur in intraplate environments?, in *Continental Intraplate Earthquakes: Science, Hazard, and Policy Issues*, S. Stein and S. Mazzotti (Editors), Geol. Soc. of Am. Special Paper 425, Boulder, Colo., 35–48, doi: 10.1130/2007.2425(03).

Kafka, A. L., and J. E. Ebel (2007). Exaggerated claims about earthquake predictions, *EOS Trans. AGU*, *88*(1), 1, 6.

Kafka, A. L. and S. Z. Levin (2000). Does the spatial distribution of smaller earthquakes delineate areas where larger earthquakes are likely to occur?, *Bull. Seism. Soc. Am.*, 90(3), 724-738.

Rundle, J. B., K. F. Tiampo, W. Klein, and J. S. S. Martins (2002). Self-organization in leaky threshold systems: The influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting, *Proc. Natl. Acad. Sci. U.S.A.* 99, 2514-2521, Suppl. 1.

Rundle, J. B., K. F. Tiampo, and W. Klein (2007). Pattern Informatics and Cellular Seismology: a comparison of methods, *EOS Trans. AGU*, 88(24), 254.

Schorlemmer, D., M. C. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades (2007). Earthquake likelihood model testing, *Seism. Res. Lett.*, 78(1), 17-27.

Figure Captions

Figure 1. Illustration of the Cellular Seismology method showing a map of a hypothetical region in which earthquakes are occurring. The “before” catalog epicenters are shown by the x’s, the CS zones are shown by the shaded zones surrounding those x’s, and the “after” catalog epicenters are shown by the filled circles. For this hypothetical case, there are 8 “after” earthquakes, and 6 of them occur within the chosen distance radius of at least one “before” earthquake, i.e., there are 6/8 (75%) “hits” for this case.

Figure 2. (a) Application of the CS method to earthquakes within the RELM study area (Schorlemmer, *et al.*, 2007; polygon shown here surrounding California). Before catalog is the ANSS catalog of earthquakes with M≥4.0 that occurred between 1932 and 2004 and fall within the RELM study area. CS radius is 6.6 km, which corresponds to 18% of the map area (shaded zones) within the RELM study area. After catalog (open circles) consists of ANSS earthquakes with M≥4.5 that occurred between 2005 and 2008. (b) Application of the E07 method using the same before and after catalogs as that of the CS analysis on the left. Rate threshold for the case shown here (shaded cells) also corresponds to 18% of the area within the RELM polygon. Inset shows percentage of polygon map area as a function of rate threshold (number of observed earthquakes) for the cells. Arrow indicates that 5 earthquakes per cell corresponds the 18% map area example shown here.

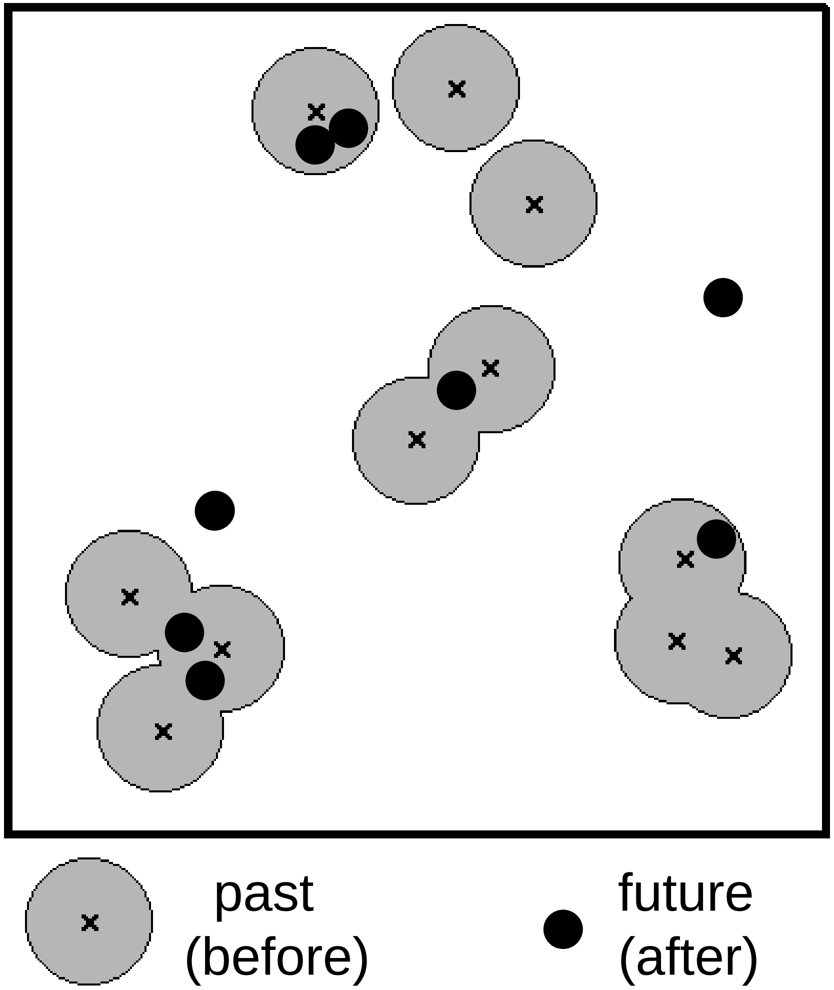
Figure 3. Comparison of performance of E07 versus CS methods of forecasting locations of earthquakes in the study area shown in the maps in Figure 2. Graph shows percentage of hits as a

function of percentage of map area for the two methods.

Figure 4. (a) Application of the CS method to earthquakes within the map area common to that covered by both Kafka (2002) and Rundle *et al.* (2002). Before catalog is the same as that used by Kafka (2002), i.e., 1984-1987, M≥3.0. After catalog (open circles) consists of earthquakes in the NEIC catalog with M≥5.0 that occurred between 2000 and 2008 and fall within the study area shown here. This implementation of CS yields 19/21 hits for 14% map area (shaded zones). (b) PI forecast map of Rundle *et al.* (2002), along with same after catalog as shown on the left (shaded cells are PI hotspots which cover 14% of the map area). Before catalog consists of earthquakes with M≥3.0 from 1932 through 1999. For this after catalog the PI method yields either 17/21 hits (filled circle in Figure 5) or 18/21 hits (open circle in Figure 5), depending on whether or not the 2003 San Simeon earthquake (which is on the boundary of one of the PI cells) is considered a hit or a miss.

Figure 5. Comparison of performance of PI versus CS methods of forecasting locations of earthquakes in the study area shown in the maps in Figure 4. Graph shows percentage of hits as a function of percentage of map area covered by the forecast.

“Cellular Seismology” Method

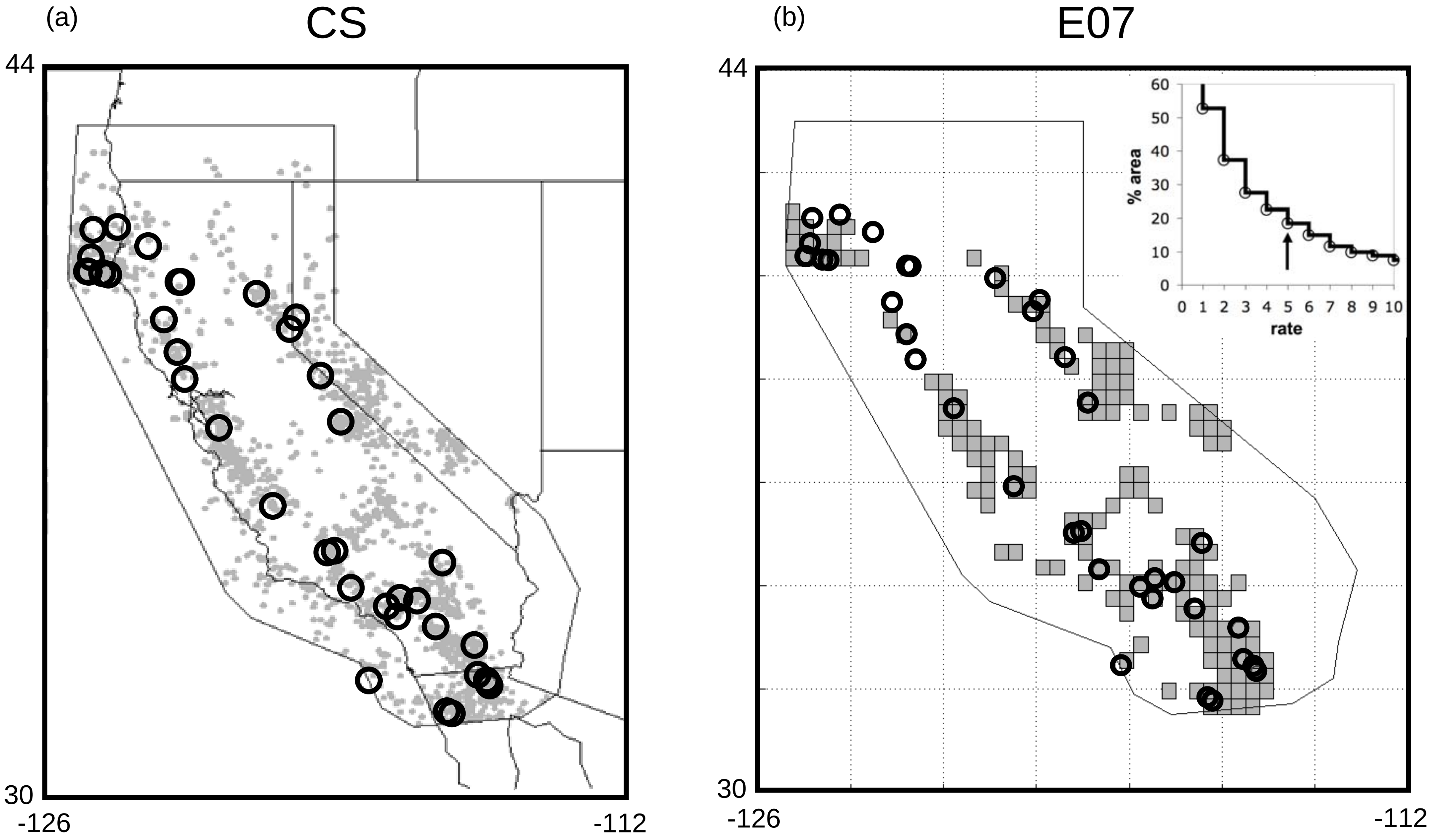


Choose a radius such that interiors of circles fill a given percentage of map area.

Determine percentage of “hits.”

For this hypothetical example, percentage of hits = 6/8 = 75%.

Figure 1



Before: 1932-2004, M≥4.0
After: 2005-2008, M≥4.5

Figure 2

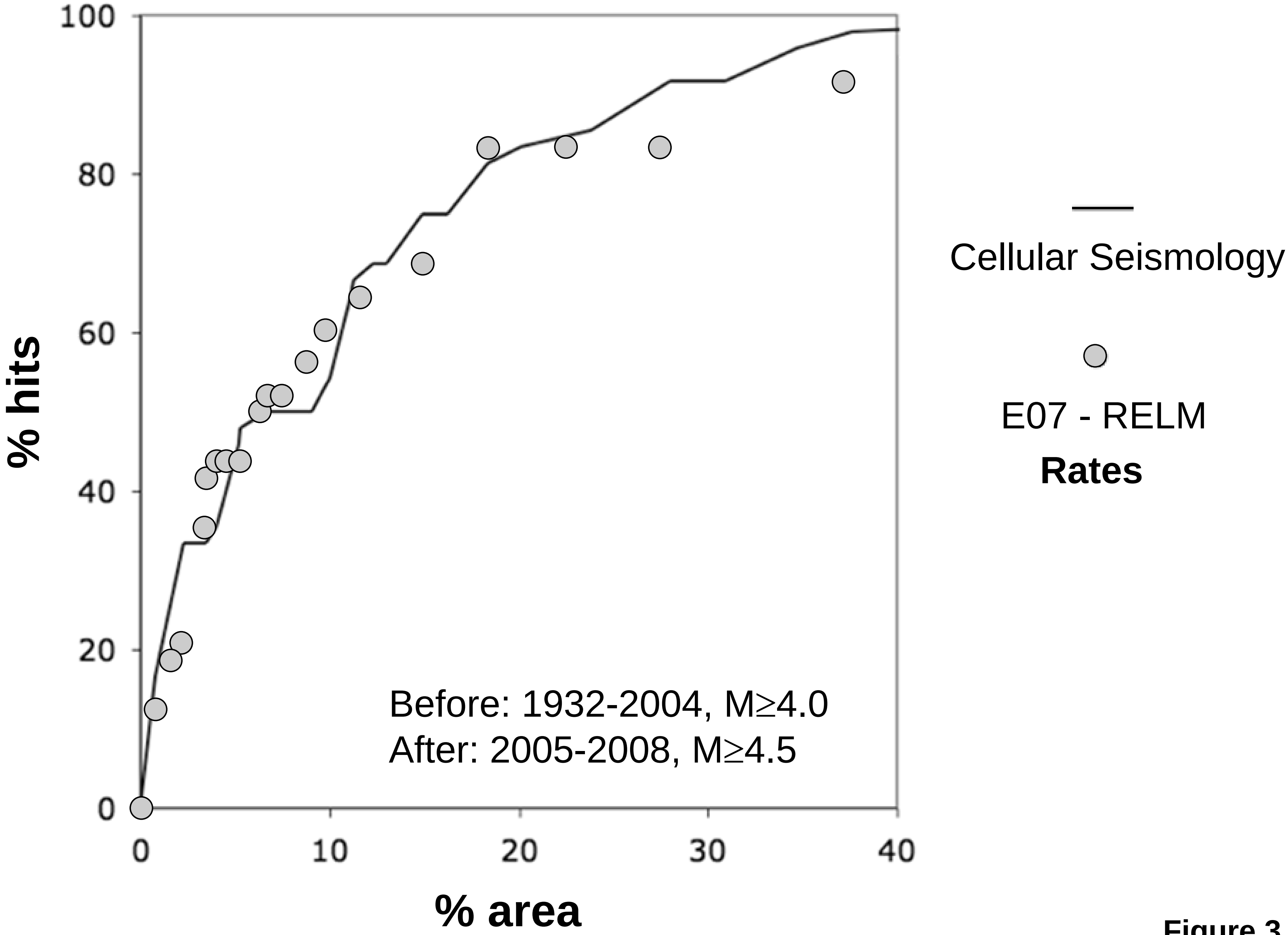
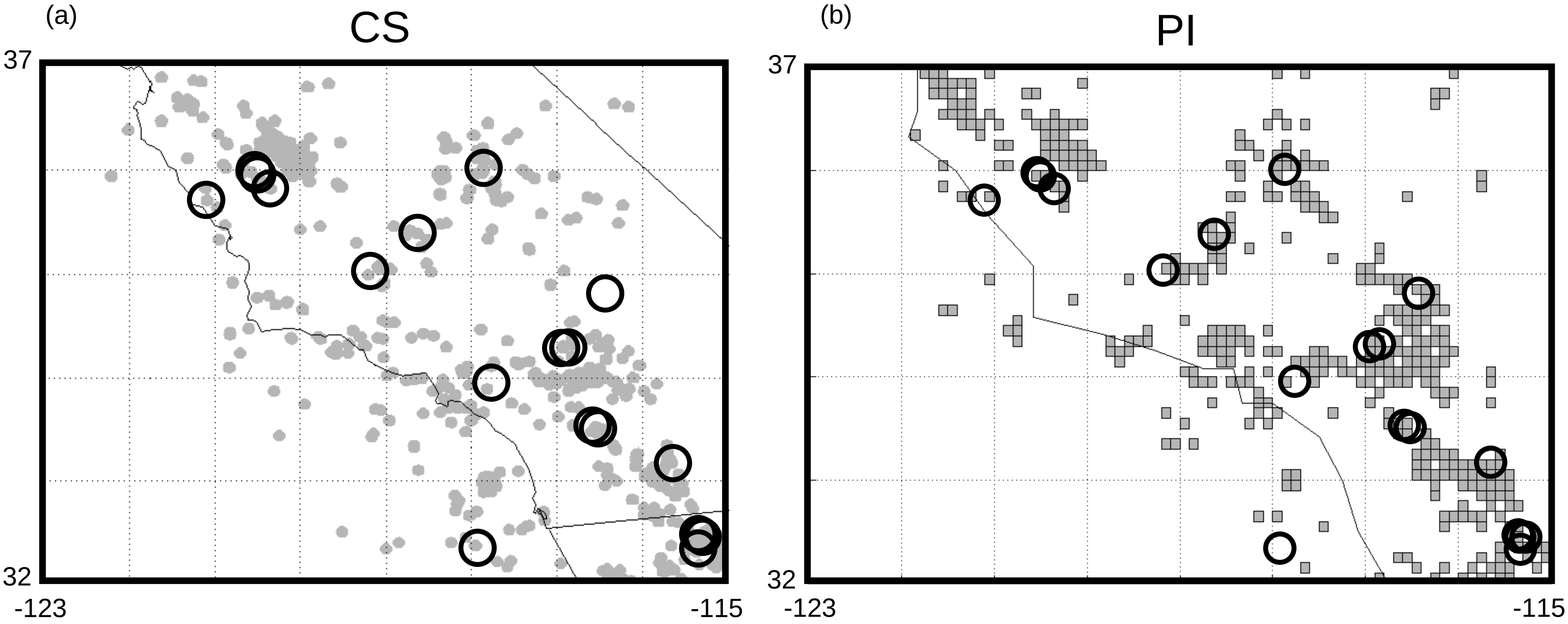

100
80
60
40
20
0
% hits
0
10
20
30
40
% area
Cellular Seismology
E07 - RELM
Rates
Before: 1932-2004, M≥4.0
After: 2005-2008, M≥4.5

Figure 3



Before: 1984-1987, $M \geq 3.0$

Before: 1932-1999, $M \geq 3.0$

After: 2000-2008, $M \geq 5.0$

Figure 4

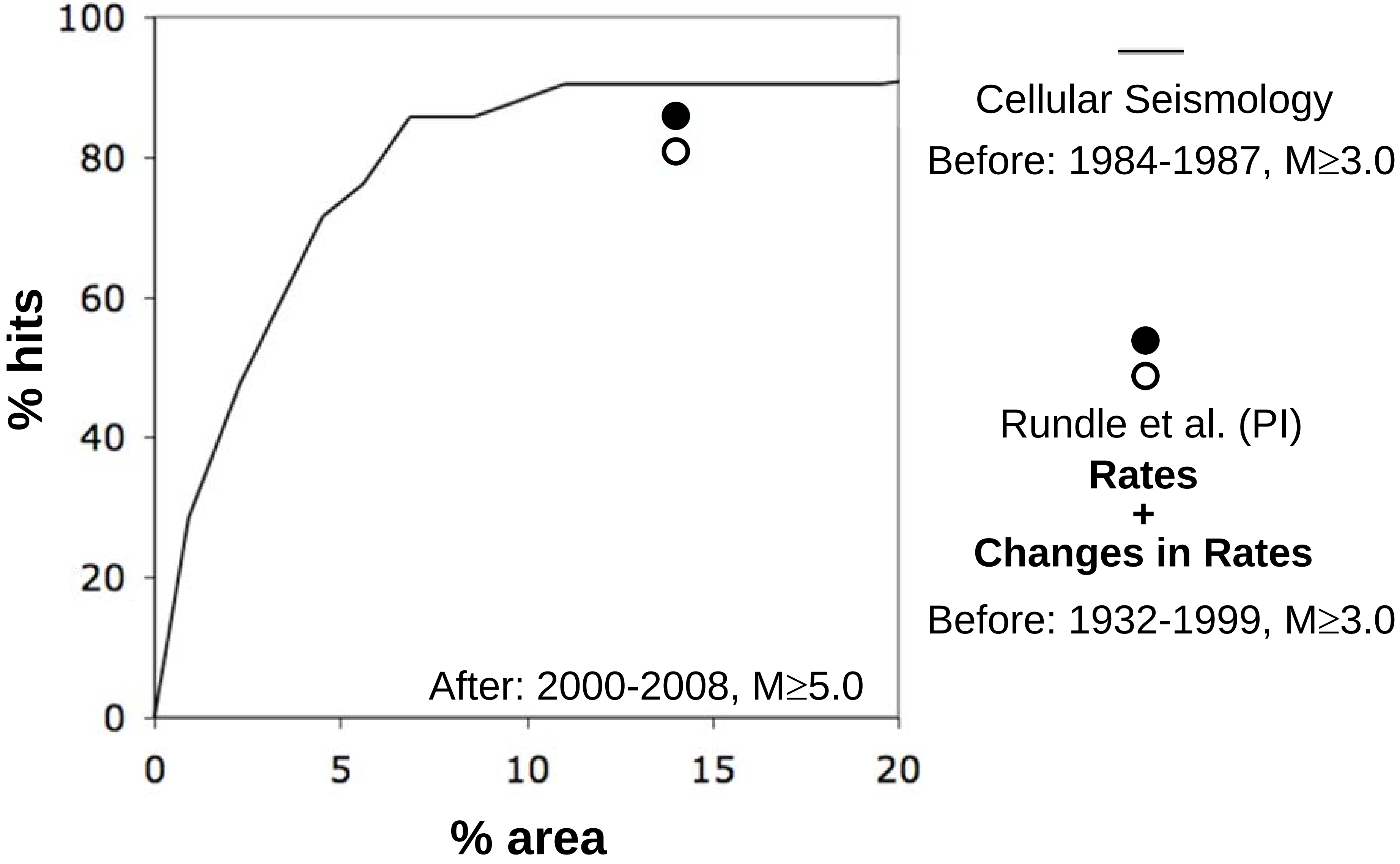

100
80
60
40
20
0
% hits
0
5
10
15
20
% area
After: 2000-2008, M≥5.0
Cellular Seismology
Before: 1984-1987, M≥3.0
●
○
Rundle et al. (PI)
Rates
+
Changes in Rates
Before: 1932-1999, M≥3.0

Figure 5