Originally published as:

Mak, S., Clements, R., Schorlemmer, D. (2014): The Statistical Power of Testing Probabilistic Seismic-Hazard Assessments. - *Seismological Research Letters*, 85, 4, p. 781-783

DOI: <http://doi.org/10.1785/0220140012>

The statistical power of testing probabilistic seismic hazard assessments

Sum Mak, Robert Alan Clements and Danijel Schorlemmer

GFZ German Research Centre for Geosciences, Helmholtzstraße 6, 14467 Potsdam, Germany.

March 29, 2014

1 Introduction

Probabilistic Seismic hazard assessment (PSHA) consists of components that are largely empirical with some physical insights (e.g., empirical ground-motion prediction equations and the Gutenberg-Richter magnitude distribution), components that are primarily based on simplified physics with some support from empirical data (e.g., the characteristic earthquake model, Schwartz and Coppersmith, 1984), and geological information that requires a lot of resources to collect but is yet difficult to assure completeness (e.g., location and dimension of faults). Empirical models do not always have clear physical meanings, and physical models are not always empirically verifiable in all situations. Together with the geological information that is almost always incomplete (and the degree of completeness unknown), assessing the “true” hazard level is a formidable task. The current practice of conducting PSHA aims at providing defensible justifications, based on the best available technology and information, for allocating resources to earthquake-resistant design of engineering structures and hazard mitigation in general.

Testing predictive models using observations has gained momentum in earthquake forecasting (e.g., Schorlemmer et al., 2007; Schorlemmer and Gerstenberger, 2007) and became a standard through the Collaboratory for the Study of Earthquake Predictability (Jordan, 2006). In the realm of PSHA, the need of testing has long been recognized (Ward, 1995). After the 2011 Tohoku earthquake, furious debates of whether existing SHAs have failed their purpose (Stein et al., 2011; Hanks et al., 2012; Stirling, 2012; Stein et al., 2012; Frankel, 2013a,b; Stein et al., 2013) reiterate the importance of testing SHAs with observations. In fact, PSHA tests proliferate in recent years (Ordaz and Reyes, 1999; Stirling and Petersen, 2006; Albarello and DAmico, 2008; Beauval et al., 2008; Miyazawa and Mori, 2009; Fujiwara et al., 2009; Stirling and Gersten-

berger, 2010; Mezcua et al., 2013). These tests, with some exceptions (Ward, 1995; Miyazawa and Mori, 2009; Fujiwara et al., 2009), apply point comparisons between modeled and observed seismic hazards. This is the most intuitive and straightforward way of testing and, if successful, the result will be directly relevant to the end-users of the PSHA.

2 The usefulness of a PSHA test

Beauval et al. (2008) pointed out that it requires a hopelessly long time to collect the necessary observations to test a point modeled hazard, implying that such tests are practically useless. Their argument is based on the parameter of coefficient of variation (COV, defined as the ratio of the standard deviation to the mean). By simulating the Poisson process, they showed that it takes about 12000 years for the observed number of ground-motion exceedances to achieve a $\text{COV} \leq 0.2$ when the return period of the ground-motion is set to 475 years (their Figures 6–8). Analytically, the COV of a Poisson process is written as $1/\sqrt{N}$, where N is the expected number of exceedances. If a $\text{COV} \leq n$ is desired, then $N \geq 1/n^2$ is required. The case presented by Beauval et al. (2008) used a COV of 0.2 as an example, thus $n = 0.2$, resulting in $N = 25$. Because the expected number of exceedances (N) is one for a time window equal to the return period of a ground-motion level, it takes 25 return periods to achieve $N = 25$. For a ground-motion of a 475-year return period, this translates into approximately 12000 years.

Although a small COV is generally desired, it remains unclear how low it needs to be for a PSHA test to be useful, and so the interpretation of a testing result based on the COV is not general. In addition, the COV is not a formal parameter to measure the usefulness of a statistical test. Here, we present the calculation of the power of testing a point seismic hazard. The power of a statistical test is a formal and general description of its usefulness. It quantitatively illustrates the conditions for such a test to be useful.

3 The power of testing ground-motion exceedance as a Poisson process

A popular assumption made for time-independent PSHAs is that earthquake occurrences are a Poisson process. Using this assumption, the modeled hazard level (in terms of rate of exceedance) can be tested against the observed number of ground-motion exceedances in a way that if the observed number is larger or smaller than some pre-defined thresholds (i.e., a two-tail test), the model is interpreted as rejected by the observation. This approach was used by a number of previous studies (Ordaz and Reyes, 1999; Stirling and Petersen, 2006; Albarello and DAmico, 2008; Stirling and Gerstenberger, 2010; Mezcua et al., 2013).

Even if the hazard level is correctly modeled, there is still a probability that the observed number of ground-motion exceedances will exceed the thresholds. Typically, we want to keep this probability (α , the probability for committing a Type I error) small. In many fields of natural and social sciences, α values of 0.1, 0.05 or 0.01 are often prescribed. On the other hand, even if the actual hazard level is different from the modeled one, there is also a probability that the observed number of exceedances does not exceed the thresholds, resulting in failure in revealing such difference (i.e., committing a Type II error). Such a case provides a false comfort that the PSHA is confirmed by the observations. The power of a statistical test is formally described as the probability for not committing a Type II error. That is, when the actual hazard level is different from the modeled one, a useful test should have a high probability to reveal such difference. The power depends on the prescribed α value, as well as how different the actual and modeled hazard levels are.

Let the modeled expected number of exceedances be N_m . The corresponding probability mass function is plotted as crosses ($\times$) in Figure 1. Using $\alpha \leq 0.1$ (*less than or equal* instead of *equal* because of the discrete distribution), the two thresholds of critical regions are determined (dashed lines). If the model is correct, it is unlikely (with a probability of α) that the observed number will fall into the critical region. Let the true hazard level be different from the modeled one such that the true expected number of exceedances is N_t . The corresponding probability mass function is plotted as filled circles ($\bullet$). Then, the probability for the observed number to fall into the critical regions equals to the area of the shaded regions. Such a probability is the power of the test.

4 Conditions for an informative PSHA test

Figure 2 shows the power of the test under various conditions. The ratio of the true to the modeled annual rate of exceedance (r) and the time window length (N_T , counted as multiples of the return period) jointly determine the true and the modeled expected numbers of exceedances, which are used to calculate the power of the test. The commonly used hazard level of 10% in 50 years translates into a return period of 475 years. In a few regions of the world (e.g., Miyazawa and Mori, 2009; Mezcua et al., 2013), the strong-motion record of the past 475 years is presumable complete in the form of macroseismic intensity. If such a dataset is used to test against the modeled hazard at the hazard level of 10% in 50 years, the test will likely reveal a significant difference if the true hazard level is drastically underestimated (e.g., $r \geq 7$ if a power of $\geq 90\%$ is desired).

Because the time window of available observations is usually short, PSHAs have also been tested at hazard levels of shorter return periods (Ordaz and Reyes, 1999; Stirling and Petersen, 2006; Albarello and

Damico, 2008; Beauval et al., 2008; Stirling and Gerstenberger, 2010; Mezcua et al., 2013) so that the time window of available observations equals to a larger multiple of the selected return period. The results of such tests, even when all assumptions behind the tests (e.g., conversion from macroseismic intensities to peak ground accelerations) are correct, will not be directly usable to the end-user of the PSHA because engineering structures are not designed for such short return periods. Nevertheless, those tests may provide a measure of the correctness of the parameters and assumptions used by a PSHA, and therefore useful for hazard modelers in improving their models. When the observed number of exceedances does not fall into the critical region of the modeled hazard level (c.f., shaded regions in Figure 2), the modeled hazard level can either be interpreted as “not rejected by” or “consistent to” the observations. The former is technically correct but does not imply the test is useful, and so does not provide all information the reader needs to know. The latter implicitly assumes the test is useful (i.e., of high power), which is not always the case. Considering the power of the test, we suggest that the interpretation of a test result of non-rejection should be cautiously phrased as like “*the observed hazard level is unlikely to be more than r times the modeled one*”, where r depends on the time window of the available observations, and the term “unlikely” refers to the power of the test, which should be reasonably close to one.

For example, if the time window of available observations is five times the concerned return period of a ground-motion level, r has to be larger than three for the power of the test to be at least 90% (Figure 2). In this case, the suggested inference implies that if the true hazard level is more than three times the modeled one, the observed number of exceedances will likely (with $\geq 90\%$ chance) fall into the critical region given the power of the test. If it has not fallen into the critical region, such an extent of discrepancy between the modeled and true hazard level is deemed unlikely. Similarly, to test whether the true hazard level is at most twice the modeled one, a test with a power of $\geq 90\%$ requires a time window of available observations to be at least 15 times the return period.

If the time window of available observations equals to a small multiple of return period (c.f., $r < 3$ in Figure 2), a test can still achieve sufficient power when the true hazard is severely underestimated. Interestingly, when the true hazard is overestimated, even if the discrepancy between the true and modeled hazard is huge, using short time windows generally leads low test powers. It is therefore logistically more difficult to achieve a powerful test when the hazard is overestimated. A test of low power is more likely to commit a Type II error and yield the false comfort that the modeled hazard is “consistent” with the observed one. Such false comfort, however, could be a minor issue because overdesigned buildings will not harm one’s life, although they may be detrimental to one’s financial health.

5 Acknowledgement

This work is developed within the Test and Evaluation project supported by Global Earthquake Model (www.globalquakemodel.org).

References

- Albareello, D. and DAmico, V. (2008). Testing probabilistic seismic hazard estimates by comparison with observations: an example in Italy. *Geophysical Journal International*, 175:1088–1094.
- Beauval, C., P.-Y. Bard, S. H., and Guguen, P. (2008). Can strong-motion observations be used to constrain probabilistic seismic-hazard estimates? *Bulletin of the Seismological Society of America*, 98(2):509–520.
- Frankel, A. (2013a). Comment on “Why earthquake hazard maps often fail and what to do about it” by S. Stein, R. Geller, and M. Liu. *Tectonophysics*, 592:200–206.
- Frankel, A. (2013b). Corrigendum to comment on “Why earthquake hazard maps often fail and what to do about it” by S. Stein, R. Geller, and M. Liu [TECTO 592(2013) 200-206]. *Tectonophysics*, 608:1453–1454.
- Fujiwara, H., Morikawa, N., Ishikawa, Y., Okumura, T., Miyakoshi, J., Nojima, N., and Fukushima, Y. (2009). Statistical comparison of national probabilistic seismic hazard maps and frequency of recorded JMA seismic intensities from the K-NET strong-motion observation network in Japan during 1997-2006. *Seismological Research Letters*, 80(3):458–464.
- Hanks, T. C., Beroza, G. C., and Toda, S. (2012). Have recent earthquakes exposed flaws in or misunderstandings of probabilistic seismic hazard analysis? *Seismological Research Letters*, 83:759–764.
- Jordan, T. (2006). Earthquake predictability, brick by brick. *Seismological Research Letters*, 77(1):3–6.
- Mezcua, J., Rueda, J., and Blanco, R. M. G. (2013). Observed and calculated intensities as a test of a probabilistic seismic-hazard analysis of Spain. *Seismological Research Letters*, 84(5):772–780.
- Miyazawa, M. and Mori, J. (2009). Test of seismic hazard map from 500 years of recorded intensity data in Japan. *Bulletin of the Seismological Society of America*, 99(6):3140–3149.
- Ordaz, M. and Reyes, C. (1999). Earthquake hazard in Mexico City: observations versus computations. *Bulletin of the Seismological Society of America*, 89(5):1379–1383.
- Schorlemmer, D. and Gerstenberger, M. (2007). RELM testing center. *Seismological Research Letters*, 78(1):30–36.

143 Schorlemmer, D., Gerstenberger, M., Wiemer, S., Jackson, D. D., and Rhoades, D. A. (2007). Earthquake
144 likelihood model testing. *Seismological Research Letters*, 78(1):17–29.

145 Schwartz, D. P. and Coppersmith, K. J. (1984). Fault behavior and characteristic earthquakes: Examples
146 from the Wasatch and San Andreas fault zones. *Journal of Geophysical Research*, 89(B7):5681–5698.

147 Stein, S., Geller, R. J., and Liu, M. (2011). Bad assumptions or bad luck: why earthquake hazard maps
148 need objective testing. *Seismological Research Letters*, 82:623–626.

149 Stein, S., Geller, R. J., and Liu, M. (2012). Why earthquake hazard maps often fail and what to do about
150 it. *Tectonophysics*, 562-563:1–25.

151 Stein, S., Geller, R. J., and Liu, M. (2013). Reply to comment by Arthur Frankel on “Why earthquake
152 hazard maps often fail and what to do about it”. *Tectonophysics*, 592:207–209.

153 Stirling, M. and Gerstenberger, M. (2010). Ground motionbased testing of seismic hazard models in New
154 Zealand. *Bulletin of the Seismological Society of America*, 100(4):1407–1414.

155 Stirling, M. and Petersen, M. (2006). Comparison of the historical record of earthquake hazard with seismic-
156 hazard models for New Zealand and the continental United States. *Bulletin of the Seismological Society
157 of America*, 96(6):1978–1994.

158 Stirling, M. W. (2012). Earthquake hazard maps and objective testing: The hazard mappers point of view.
159 *Seismological Research Letters*, 83(2):231–232.

160 Ward, S. N. (1995). Area-based tests of long-term seismic hazard predictions. *Bulletin of the Seismological
161 Society of America*, 85(5):1285–1298.

Figure 1: Example of computing the power of a test. The modeled (N_m) and true (N_t) expected number of exceedances of a Poisson process are 10 and 13, respectively. If the observed number is > 15 or ≤ 5 (dashed lines), the modeled hazard level is interpreted as rejected by the observed one for $\alpha = 0.1$. Under the true hazard level, the probability for achieving such an interpretation equals to the total area of the two shaded regions. This probability is defined as the power of the test. This example is equivalent to taking $r = 4/3$ and $N_T = 10$ in Figure 2, which leads to a power of 27%.

Figure 2: The power of various scenarios for SHA test for $\alpha = 0.1$. The power is given as a function of the ratio of the true to the modeled rate of exceedance, r , and the time window length of the available observations in terms of multiples of return periods, N_T . Power values $\geq 90\%$ are indicated in white.