Outliers in Strong-Motion Datasets

Fleur O. Strasser, Julian J. Bommer
Dept. of Civil & Environmental Engineering, Imperial College London, SW7 2BU, UK

Introduction

Ground-motion prediction equations are an essential element of PSHA. The equations generally predict median values of ground-motion parameters as a function of explanatory variables such as magnitude, distance, site classification and style-of-faulting. The aleatory variability associated with the predictions, represented by the distribution of the residuals with respect to the equation, is generally modeled as a Gaussian distribution of the logarithmic residuals, referred to as a lognormal distribution, characterized by a standard deviation σ. An observation or an estimate of the ground-motion parameter can be defined by the number, ε, of logarithmic standard deviations above the logarithmic mean of the equation. When PSHA is performed for very low exceedance frequencies, the non-truncated lognormal distribution can lead to very high estimates of ground motions as a result of large values of ε being considered (Bommer *et al*., 2004). This brief study examines the nature of the distribution of residuals in strong-motion datasets used to derive ground-motion prediction (attenuation) equations and in particular the nature of the highest outliers (i.e. the values with the largest ε values).

Upper tails of the residual distribution

A compilation of predictive equations derived worldwide in the years 1971 to 2003 by Douglas (2003) shows that σ values have not decreased over time, and are virtually unaffected by the use of larger datasets or the incorporation of additional variables. As a result, the nature of the upper tails of the residual distribution will be a key factor in constraining the aleatory uncertainty on ground motion.

The issue of upper tails was specifically addressed in the PEGASOS project (Abrahamson *et al.*, 2002). Although not quite a consensus, there was a general feeling that the available strong-motion data is insufficient to define the nature of the upper tails of the distributions of residuals. Restrepo-Velez & Bommer (2003) applied the upper limit log-normal (ULLN) distribution to the residuals of European strong-motion data calculated using the equation of Tromans & Bommer (2002), from which an upper bound at 6 logarithmic standard deviations above the logarithmic mean was inferred. Since the largest residuals were at about the 3σ level, this extrapolation is neither considered robust nor reliable.

Residual datasets for selected predictive equations

The present study examines the residuals of a number of strong-motion data sets used for the derivation of predictive equations for horizontal spectral ordinates (Ambraseys *et al.*, 1996; Berge-Thierry *et al.*, 2003; Bommer *et al.,* 2003; Chang *et al.*, 2001; Lussou *et al.*, 2001). These equations differ by the extent and provenance of the data, as well as by the number and definitions of the variables used in the regression. However, as illustrated in Figure 1, the largest outliers are consistently at least at the 2.5σ to 3σ level, all equations using a homoscedastic scatter (i.e. σ is independent of the explanatory variables). The overall range of values taken by the residuals varies with the frequency considered, the number of variables included in the regression and the definition of horizontal component used, although no systematic trends could be found.

Examination of the individual residual sets at each frequency shows no correlation with either magnitude or distance, as would be expected for well-conditioned data sets. Neither could a pattern be found with respect to site classification or style-of-faulting, regardless of whether these variables were included in the regression or not. However, all sets displayed a positive correlation between the residuals and the logarithm of the observed ground motion, characterized by a correlation coefficient of about 0.6, indicating that on average, higher residuals should be expected for higher ground motions.

Characteristics of highest outliers

In a second step, subsets comprising the 15 highest residuals of the Bommer *et al.* (2003) and Berge-Thierry *et al.* (2003) datasets were analyzed to check whether the overall lack of correlation between residuals and explanatory variables is also a feature of these extreme outliers. Despite the strong overall correlation with observed ground motion noted above, the highest outliers span a significant range of ground-motion values, but generally fail to include motions of engineering significance, in particular near-source records. Like the complete sets of residuals, these subsets show a lack of correlation with any of the basic explanatory variables used in regression analysis.

It is customary to associate repeated high residuals at a given station with a site-specific response. Similarly, when all records from a single event exhibit high residuals, this is often interpreted as a source characteristic, such as the often cited explanation of “high stress drop” for the 1985 Saguenay earthquake. However, examination of the individual outliers and the associated data shows that in most instances it is not possible to classify them unambiguously as source- or site-related. Figure 2 illustrates this in the case of the 1984 Lazio Abruzzo (Italy) earthquake, providing the highest PGA residual for the Bommer *et al.* (2003) equation: although the residuals for this event are well-correlated with observed ground motion, there does not seem to be any consistent pattern with respect to

distance or site classification, nor any indication that the residuals can be related to a gross source characteristic.

Systematic examination of outlying accelerograms can help to identify the factors contributing to these unexpectedly large motions, such as directivity, site effects, or a particular feature of the path (Strasser *et al*., 2004a; Strasser, 2005). Published studies on many notable recordings often tend to assign the high ground-motion amplitudes to a single cause, with different authors identifying different factors, which may be interpreted as indicating that the most extreme motions are the result of the favorable combinations of several factors. Therefore, there is great potential benefit in identifying these factors, their ranges of possible values, and most importantly, the ranges of feasible combinations of these values, perhaps expressed in terms of joint probability distributions. This would then help to define the required input into numerical simulations (Strasser *et al.*, 2004b).

Conclusions

The residual datasets of several recently predictive equations were investigated, with a particular focus on the nature of the highest outliers.

- As a whole, the residual datasets show no correlation with any of the explanatory variables commonly used in predictive equations.
- The highest outliers are mostly related to motions of little engineering significance, despite the fact that overall the residuals are well correlated with observed amplitude.
- Not all of the outliers are of very low amplitude – but few are at short distances.
- Outliers also show no consistent pattern with respect to explanatory variables.
- It is not possible in general to attribute outliers to either source or site effects; the most extreme cases seem to result from combinations of these effects.
- Record processing issues or inadequate definitions of the regression variables may also be the cause of some outliers

Acknowledgements

The first author would like to express thanks to Tom Hanks and the other members of Extreme Ground Motions at Yucca Mountain committee for the invitation to participate in this workshop. We express our thanks to the individuals who assisted us in obtaining copies of the datasets for the residuals analyzed herein: Catherine Berge-Thierry, T-Y. Chang, Fabrice Cotton, John Douglas, Philippe Lussou and Philippe Roth.

References

Abrahamson, N.A., Birkhauser P., Koller, M., Mayer-Rosa, D. Smit, P., Sprecher, C., Tinic, S. and Graf, R. (2002). PEGASOS – a comprehensive probabilistic seismic hazard assessment for nuclear power plants in Switzerland. *12th European Conference on Earthquake Engineering*, London, Paper No. 633.

Ambraseys, N.N., K.A. Simpson & J.J. Bommer (1996). Prediction of horizontal response spectra in Europe. *Earthquake Engineering & Structural Dynamics* **25**, 371-400.

Berge-Thierry, C., F. Cotton, O. Scotti, D.-A. Griot-Pommera & Y. Fukushima (2003). New empirical response spectral attenuation laws for moderate European earthquakes. *Journal of Earthquake Engineering* **7**(2), 193-222.

Bommer, J.J., J. Douglas & F.O. Strasser (2003). Style-of-faulting in ground-motion prediction equations. *Bulletin of Earthquake Engineering* **1**, 171-203.

Bommer, J.J., Abrahamson, N.A., Strasser, F.O., Pecker, A., Bard, P-Y., Bungum, H., Cotton, F., Fäh, D., Sabetta, F., Scherbaum, F. and Studer, J. (2004). The challenge of defining upper bounds on earthquake ground motions. *Seismological Research Letters* **75**(1), 82-95.

Chang, T.-Y., F. Cotton & J. Angelier (2001). Seismic attenuation and peak ground acceleration in Taiwan. *Bulletin of the Seismological Society of America* **91**(5), 1229-1246.

Douglas, J. (2003). Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth-Science Reviews* **61**, 41-104.

Lussou, P., P.Y. Bard & F. Cotton (2001). Seismic design regulation codes: Contribution of K-Net data to site effect evaluation. *Journal of Earthquake Engineering* **5**(1), 13-33.

Restrepo-Vélez, L.F. and Bommer, J.J. (2003). An exploration of the nature of the scatter in ground-motion prediction equations and the implications for seismic hazard assessment. *Journal of Earthquake Engineering* **7**(special issue no.1), 171-199.

Strasser, F.O. (2005). Interpretation and modeling of extreme ground motions. *PhD Thesis*, Imperial College London.

Strasser, F.O., Bommer, J.J. and Boore, D.M. (2004a). What produces large earthquake motions? (Abstract). *Seismological Research Letters* **75**(2), 289.

Strasser, F.O., Priolo, E., Vuan, A., Bommer, J.J., Klinc, P. and Laurenzano, G. (2004b). Preliminary results of simulations exploring the nature of extreme ground motions using a kinematic deterministic-stochastic finite-fault model: EXWIM (Abstract). *Seismological Research Letters* **75**(2), 283.

Tromans, I.J. and Bommer, J.J. (2002). The attenuation of strong-motion peaks in Europe. *Proceedings of the 12th European Conference on Earthquake Engineering*, London Paper no. 394.

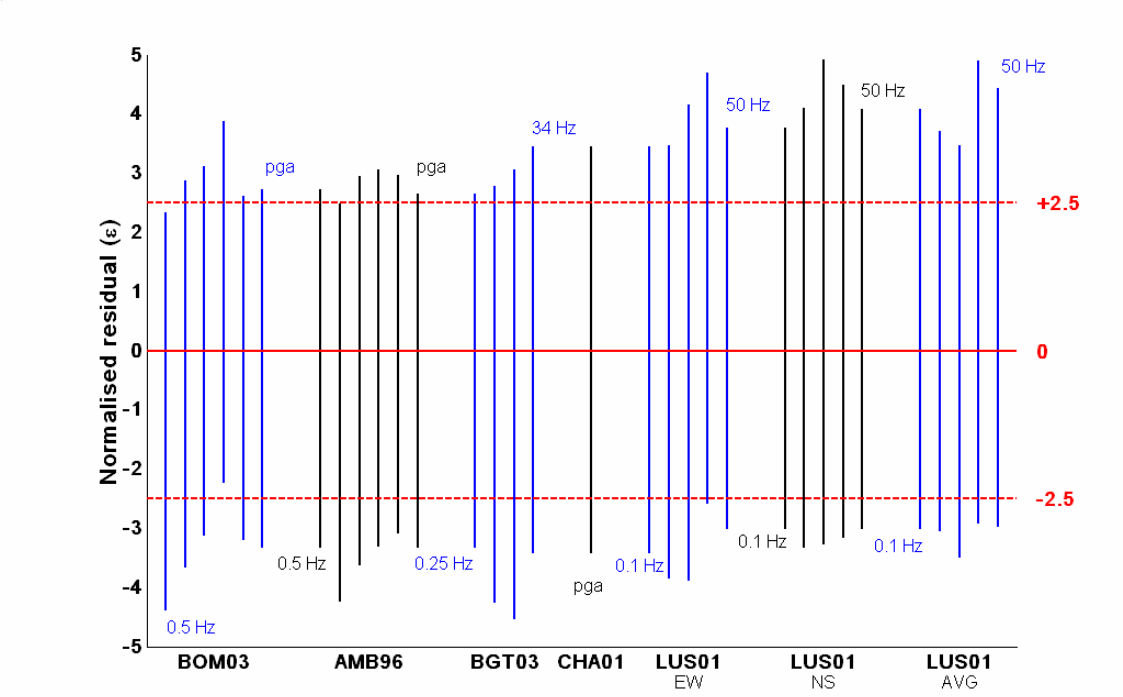

Figure 1. Ranges of values taken by ε in the data sets used, at different frequencies and for different definitions of the horizontal component.

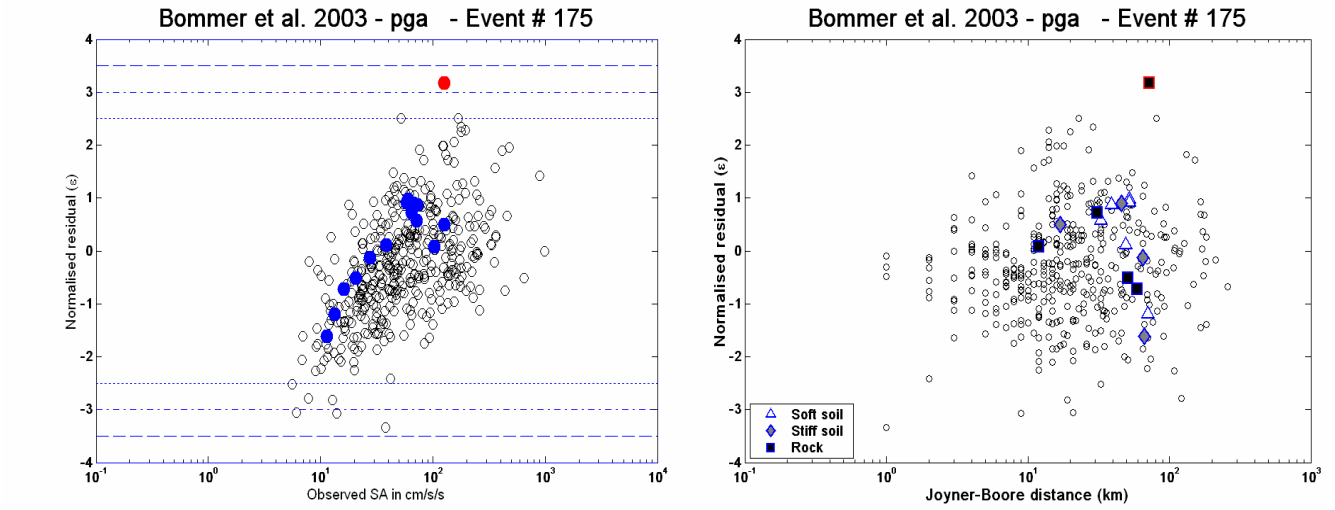

Figure 2. Locations with respect to whole dataset of normalized ε residuals from all stations available for a single event (Lazio Abruzzo, 07/05/1984, M_S=5.8) plotted against log (SA) *(left)*, and distance, including site classification *(right)*.