REVIEW ARTICLES

Reply to the comment on "Assessing CN earthquake predictions in Italy" by G. Molchan, A. Peresan, G.F. Panza, L. Romashkova, V. Kossobokov

Matteo Taroni[1], Warner Marzocchi[1], Pamela Roselli[1]

[1] *Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy*

Article history
Received April 3, 2017; accepted January 16, 2018.
Subject classification:
Earthquake prediction; Model validation; Italy.

Molchan et al. [2018] raised concerns on the reliability of the main Taroni et al.'s [2016] conclusion that reads "Considering the data available so far, the Molchan Test does not show that CN prediction performance is significantly better than predictions based on the stationary Poisson model." In particular, Molchan et al. [2018] discuss two main issues: 1) the Taroni et al.'s [2016] results are based on too few data to achieve robust conclusions, and 2) the parimutuel gambling score (PGS) produce unfair results in comparing predictive models.

We thank Molchan et al. [2018] to give us this opportunity to clarify further some aspects of our paper, but we anticipate that we do not see any compelling reason to modify our original conclusion.

1. Conclusions drawn from too few data

Molchan et al. [2018] dissert on the influence of few data on the outcome of a statistical test, claiming that "*A priori the standard statistical methods may not be effective in any of the CN sub-regions*". In particular, Molchan et al. [2018] argue that the rejection of the null hypothesis at a specific significance level of the test is unstable when the test is made with a few data. We agree with the general (and trivial) fact that the fewer the data, the smaller the power of the test. At the same time, we do not agree that with this number of target earthquakes in each sub-region, unavoidably, we cannot reach any meaningful conclusion. For example, if the CN alarm rate would have been smaller, say 0.10, 4/5 hits in the Northern region would have been an excellent and robust proof of a significant superiority of CN predictive capability with respect to the Poisson reference model (the null hypothesis of the test); in fact for 4/5 hits the *P*-value is 4.6 10^{-4}, and if we remove one hit, i.e., 3/5 hits, the *P*-value is still very low (8.6 10^{-3}). So, it is not only matter of how many target earthquakes we have used for the test, but also of the alarm rate imposed by the model that does not depend on the number of target earthquakes. If the alarm rate is high (as for CN), of course we will need many more earthquakes to detect a possible difference with the reference Poisson model. But, in this case, we can still draw some conclusions about the CN predictive capability, e.g., (at best) it requires the occurrence of many earthquakes to show some statistically significant gain with respect to the reference Poisson model.

In Taroni et al. [2016] we have tested each sub-region independently to avoid the problem of sub-region overlapping. Conversely, Molchan et al. [2017] adopt a strategy to test the CN model aggregating the three sub-regions. This aggregation introduces some further assumptions, but it allows the authors to test the model using simultaneously all target earthquakes. Worthy of note, the results of Molchan et al. [2017] do not contradict the ones showed in Taroni et al. [2016]; in fact Molchan et al. [2017] find a *P*-value of about 0.08 for the pooled test. The same result (*P*-value of 0.08) can be obtained aggregating the three *P*-values shown in Taroni et al. [2016] through a classical Fisher method [Fisher 1925], once the *P*-values have been corrected for continuity. This is certainly interesting because it confirms the reliability of the analysis and results shown in Taroni et al. [2016]. In their Table 3, Molchan et al. [2017] show other apparently "more significant" results (even if they never achieve a *P*-value less than 0.01) only in two cases that we deem as inappropriate:

i) when the target earthquake that occurred both in Central and Southern region (9/9/1998) is considered only as a success for the Central region, and not also as a missed target earthquake for the Southern region despite the latter was not in alarm (note that in Taroni et al., 2016 this case has been considered both as a success for Central region and as a missed target earthquake for Southern region);

ii) when the Norcia earthquake (10/30/2017) is included as a target earthquake. The latter is unacceptable because it contradicts the rules imposed by the authors, which aim at predicting only independent mainshocks. In fact, the Norcia earthquake is clearly related to the target Amatrice earthquake (8/24/2017) that occurred nearby about two months before [Marzocchi et al. 2017]; hence, if the Norcia earthquake is considered as a target earthquake, the Poisson assumption used for the reference model is violated, and the results of the test become unacceptable.

As a final consideration on this section, we underline that in Taroni et al. [2016] the null hypothesis H_0 consists of the equality of predictive performance of CN with a reference model that is the Poisson model (CN aims to predict only mainshocks and do not consider aftershocks). Taroni et al. [2016] concluded that with the available data we do not reject this H_0. At the risk of being trivial we emphasize that this does not mean that CN and Poisson prediction performance are really equal, but only that the available data do not show any significant discrepancy from the null hypothesis. Note that we never use the terms "verified" or "proved" as claimed by Molchan et al. [2018]. At this point, it is worth explaining why we decided to carry out this analysis with this limited amount of data. Since 2011, Peresan et al. [2011] claimed that CN provides *"successful and stable results"* in Italy when compared to the Poisson model, CN predictions are *"so far the only formally validated tools for anticipating the occurrence of strong Italian earthquakes"* [Peresan et al. 2012], and the same authors repeatedly advocate through Italian mass-media that CN predictions should be used by Civil Protection for risk reduction purposes (they provide these predictions to the Civil Protection of Friuli Venezia Giulia since many years). We think that any prediction model to be used for practical purposes must prove its reliability and superior skill with respect to the present state of knowledge through prospective tests, i.e., using independent data [Jordan et al. 2011]; the Taroni et al.'s [2016] results show that so far CN prediction capability does not appear superior to random guess, even aggregating the results of the sub-regions (see before).

2. On the Parimutuel Gambling Score (PGS)

Before addressing this comment, we emphasize that PGS is not an alternative method to the approach discussed before. PGS is a scoring procedure that tends to give higher score to the best performing model according to a specific metric. In other words, PGS is not used for hypothesis testing, and it gives a different type of information with respect to the previous test.

In Taroni et al. [2016] we apply correctly PGS, comparing the CN model with a random forecast (RF). Molchan et al. [2018] is right when they claim that, under some specific conditions, PGS rewards more any reasonable random process with respect to the CN prediction scheme. According to equation 12 in Molchan et al. [2017], these conditions are met when the alarm rate is much higher that the target earthquake rate (using as time unit the length of the prediction time interval), as in the present case of CN model. However, it is worth noting that when this condition is met, it says a lot on the prediction capability of the model, because it implies that the false alarm rate is large.

At the same time, we acknowledge the fact that if we decide that a false alarm is less important from a practical point of view than a missed target earthquake, PGS should be applied weighing differently these kinds of errors. We do not do that in Taroni et al. [2016], where both successes and errors are weighed symmetrically (i.e. hit and the correct negative give the same gain, likewise missed target earthquakes and false alarm give the same loss). The choice to weigh any kind of error in the same way is neither right nor wrong. It is just a possible decision for scoring, and others are possible. In Taroni et al. [2016] we avoid to give different weights to a failure in terms of missing target earthquakes or false alarms, because this requires non-scientific information (e.g. the cost of a false alarm versus the cost of a missed target earthquake) that has to be decided by the decision makers.

Molchan et al. [2018] also claim that using RF instead of random guessing (RG) is not the same thing. As mentioned at the beginning of this section, we did not apply PGS to add a further significance test to the CN model, but to score CN against RF from a uniform Poisson model, which is a standard model used in such a kind of comparison [e.g. Rhoades et al. 2011].

3. Additional considerations

Finally, we take this opportunity to add a few further considerations. In Taroni et al. [2016] we have used the classical Neyman-Pearson approach of statistical testing with a significance level of 0.05. Although this value is still widely used in science, we argue that this choice is very generous for the CN model; in fact, the recent general tendency is to use much smaller significance levels in testing any null hypothesis [Singh Chawla 2017]. On the other hand, if we decide to interpret the results of Taroni et al. [2016] avoiding the somehow subjective choice of a significance level, we may interpret the *P*-value as a graded measure of the strength of evidence against the null hypothesis or the reference model [Amrhein et al. 2017]. This raises the problem of the scientific quality of the reference model that has been used to evaluate the CN model; in fact, a weak reference model may easily lead to get small *P*-values [e.g. Marzocchi et al. 2003]. In essence, the reference model used for testing is a spatially homogeneous Poisson process inside each sub-region (and in all aggregated sub-regions). We think that this is a very crude reference model, and it does not represent what seismologists already know. In fact, we know that there is a substantial spatial variability of independent mainshock occurrences inside these regions (see, e.g., Werner et al. [2010], and references therein; and the spatial variability of the seismicity rate model used for the new seismic hazard model for Italy; Meletti et al. [2017]). We also know that, even after applying some declustering procedures, in time intervals less than one year the Poisson assumption for the mainshocks is not always applicable. More in general, since forecasting models for Italian seismicity already exist [e.g. Marzocchi et al. 2014], they could be more proper reference models to be used in testing CN.

Moreover, in Taroni et al. [2016] we follow the same testing rules used by the CN authors. However, we think that these rules should be modified in the future to make more meaningful tests. For instance, if the ultimate goal is to predict earthquakes inside the Italian territory, target earthquakes that occur in Italy outside the macro-zones used by CN must be considered as failures, because the definition of the spatial macro-zones is part of the CN model that we aim to test. The two most remarkable examples of earthquakes in the Italian territory, but not considered for CN testing are the M 6 earthquake occurred just offshore Palermo on Sept 9, 2002, and the M5.9 occurred in Molise region on October 31, 2002 (a new region covering this part of Italy has been included only after this earthquake). Of course, the inclusion of these failures would strengthen the Taroni et al.'s [2016] conclusions.

4. Conclusions

We thank again Molchan et al. [2018] for their comments and for pointing out some interesting features of the PGS test. We would like also to emphasize that this scientific discussion is possible because the CN authors did a good job in allowing independent researchers to evaluate their predictions; this is a positive and rare attitude in this field. In this reply we show that we did not make any mistake in Taroni et al.'s [2016] paper, and we do not feel to have oversold our conclusions. For this reason, we conclude this reply re-stating again that "Considering the data available so far, the Molchan Test does not show that CN prediction performance is significantly better than predictions based on the stationary Poisson model."

References

Amrhein, V., F. Korner-Nievergelt, and T. Roth (2017). The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ, 5, e3544.

Fisher, R.A. (1925). Statistical Methods for Research Workers. Oliver and Boyd (Edinburgh).

Jordan, T.H., Y.-T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka, and J. Zschau (2011). Operational earthquake forecasting: state of knowledge and guidelines for utilization. Ann. Geophys.,54, 315-391.

Marzocchi, W., L. Sandri, and E. Boschi (2003). On the validation of earthquake-forecasting models: the case of pattern recognition algorithms. Bull. Seismol. Soc. Am., 93, 1994-2004.

Marzocchi, W., A.M. Lombardi, and E. Casarotti (2014). The establishment of an operational earthquake forecasting system in Italy. Seismol. Res. Lett., 85(5), 961-969.

Marzocchi, W., M. Taroni, and G. Falcone (2017). Earthquake forecasting during the complex Amatrice-Norcia seismic sequence. Sci. Adv., 3, e1701239.

Meletti, C., Marzocchi, W. and MPS16 Working Group (2017). The 2016 Italian seismic hazard model, 16th World Conference on Earthquake Engineering, Santiago de Chile, 9-13 January.

Molchan, G., A. Peresan, G.F. Panza, L. Romashkova, and V. Kossobokov (2018). Comment on "Assessing CN earthquake predictions in Italy" by M. Taroni, W. Marzocchi, P. Roselli. Ann. Geophys., this volume.

Molchan, G., L. Romashkova, and A. Peresan (2017).

On some methods for assessing earthquake predictions. Geophys. J. Int., 210, 1474-1480.

Peresan, A., E. Zuccolo, F. Vaccari, A. Gorshov, and G.F. Panza (2011). Neo-Deterministic Seismic Hazard and Pattern Recognition Techniques: Time-Dependent Scenarios for North-Eastern Italy, Pure Appl. Geophys. 168, 583-607.

Peresan, A., V.G. Kossobokov, and G.F. Panza (2012). Operational earthquake forecast/prediction, Rend. Fis. Acc. Lincei.

Rhoades, D., D. Schorlemmer, M. Gerstenberger, A. Christophersen, J. Zechar and M. Imoto (2011). Efficient testing of earthquake forecasting models. Acta Geophys., 59(4), 728-747.

Singh Chawla, D. (2017). Big names in statistics want to shake up much-maligned P value. Nature, 548, 16-17

Taroni, M., W. Marzocchi, and P. Roselli (2016). Assessing CN earthquake predictions in Italy. Ann. Geophys., 59(6), S0648.

Werner, M. J., J. D. Zechar, W. Marzocchi, and S. Wiemer (2010). Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts. Ann. Geophys., 53, 11-30.

*Corresponding author: Matteo Taroni,
Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy;
email: matteo.taroni@ingv.it