Earthquake intensity measures prediction on Southern Italy: A non parametric approach

Introduction

Ground motion prediction equations (GMPEs) relate ground motion intensity measures to variables describing earthquake source, path, and site effects.

The designing of good models, able to produce precise estimates of such measures, is a key point in different fields related to earthquakes (seismic hazard assessment, to make an example). Thus, a broad literature on the topic was developed (and is in developing), and find one of his peek tools in non linear mixed effect models. During last semester, we and other two colleagues (credits to Mirko Giovio and Daniele Venturini) tried to train a model based on the work of Lanzano, Luzi, Pacor et al [1], with respect to the prediction of a derived measure of the Fourier Amplitude Spectrum.

In this work, we applied non parametric tools to enhance previous considerations, evaluate a non parametric regression approach and finally compare the performances in prediction with the classical model, in order to see if it could be a valid alternative.

Dataset

The dataset we worked with was provided by INGV. In principle, it contains recordings of events happened in Southern Italy (see Figure 1), with respect to the station which did the recording, and all the covariates necessary to train the model.

Here's a list of the most important ones (for our analysis):

- Event ID: to identify the event. Note that multiple measurements per event are present (taken from different stations).
- Station ID: to identify the station.

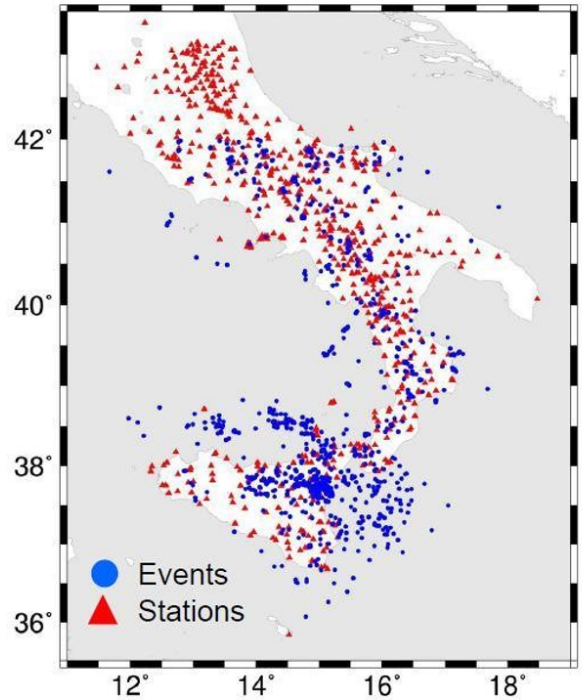

Figure 1: Events and Stations

- FAS: Fourier Amplitude Spectrum components (x,y,z) at given frequency [1].
- ML-MStat: magnitude at location (ML), derived from magnitude recorded at station (MStat).
- Dipo-Depi:hypocentral and epicentral distance
- VS30: Shear wave velocity in the upper 30 metres; basically a parameter accounting for the soil characteristics [2].
- Depth: depth at which an event takes place.

[1] As we will point out later, the model calibration is frequency dependent: a dataset per each frequency of the FAS must be provided. In this work we focus on the lowest one: 0.50HZ

[2] Very few empirically measured values are available; a coarse approximation is provided in most of the cases

- EC8: factor providing a hand-written classification of each event, given by experts. It consists of seven classes and numerous unlabeled observations are present.

Classical approach

To completely understand, it is necessary to show briefly what we have done following the classical approach. Our quantity of interest was (and still is) the logarithm of the geometric mean of the two horizontal components of the Fourier Amplitude Spectrum (at frequency 0.50HZ). The regression variables taken into account in the classical approach were a term for magnitude (ML), a term for distance (Dipo) and a term for soil characteristics (VS30), plus an interaction term between magnitude and distance [3]. Note that the model is highly nonlinear and dependent on frequency-scaling parameters(Mh, Mref, h):

$$log_{10}Y = a + F_{\mathrm{M}} + F_{\mathrm{D}} + F_{\mathrm{S}} + \epsilon \quad (1)$$

$$F_{\mathrm{M}} = \begin{cases} b1(M - Mh) & if \quad M <= Mh \\ b2(M - Mh) & if \quad M > Mh \end{cases} \quad (2)$$

$$F_{\mathrm{D}} = c1[(M - Mref) + c2]log_{10}D + c3D \quad (3)$$

$$F_{\mathrm{S}} = klog_{10}(VS30)/800 \quad (4)$$

The procedure followed to train the model was composed by two step: first we perform a nonlinear least square minimization, focusing our attention on the frequency-dependent hyperparameters [4]; then, keeping those parameters fixed, we perform a mixed effect regression. We did so mainly because the training of the model is difficult and highly unstable. In the first step (which was the most time consuming one) we were able to spot good starting values for the hyperparameters, looking at plots such as the ones in Figure 2 and Figure 3.

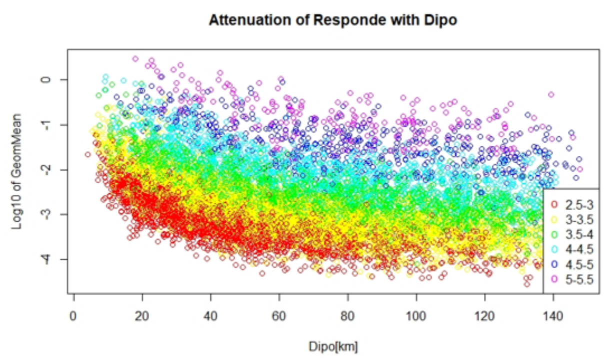

Figure 2: Scaling of LogY with respect to distance

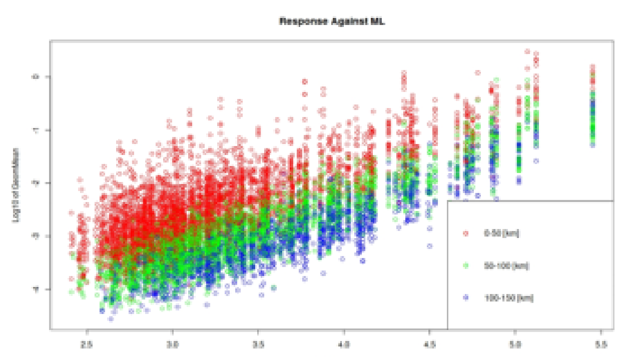

Figure 3: Scaling of LogY with respect to magnitude

Thus, we were able to regularize the tuning of the hyperparameters (which we recall scale with frequency), and consequently to regularize also the coefficients values of the final mixed effect model.

[3]we actually consider to enhance the model with the introduction of other regressors in our previous work, but here we consider the one proposed (and validated) in literature, as we need it for benchmarking

[4]Levenberg - Marquardt algorithm

Outliers detection

Dataset cleaning and outliers detection was a very important topic in our previous work. Actually, one of the model assumptions was response normality, which was not completely achieved, simply because the (supposed) true distribution is fat-tailed. That was a major concern: we could not rely on straightforward and automatic analysis, as most of them would have discarded points in which we were actually really interested. We were able to spot outliers, wrong measurements and highly influential points in a raw manner, achieving some sort of normality. Needless to say, that was heavily time consuming.

This time, we followed an approach based on Tukey depths to see if it was satisfactory in spotting problematic points (which we already knew). Basically, we look at different kind of bagplots, identifying in those ones relating important covariates and the response the most promising ones to spot outliers. In Figure 4 and 5 are reported, as an example, the bagplots relative to hypocentral distance (vs response) and depth (vs response).

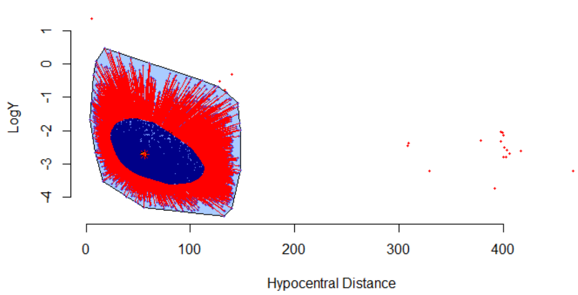

Figure 4: Clear cluster of high distanced points

To conclude, we we have been surprised by the effectiveness of this approach: basically all the problematic points we knew about we're spotted, and actually some other ones too, keeping the number of discarded points

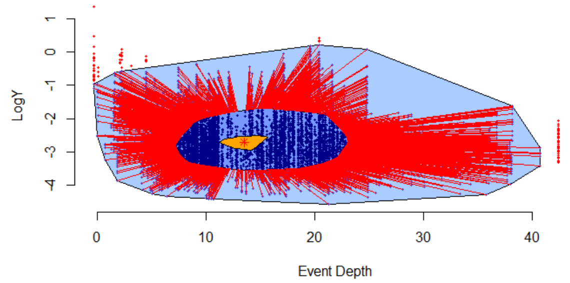

Figure 5: Both too deep and too shallow events can be misleading

to an acceptable level; moreover, the time required to clean data had been much less.

Testing

In this part of the report are collected and presented all the tests we performed to validate our beliefs. Adopting a non parametric approach for testing was fundamental, as most of the times we weren't in the position to rely on parametric ones.

1. High distanced data: we performed a permutational test to assess equality in distribution between the cluster of points with high distance, spotted during outlier detection, and the rest of data. L^2 norm of the multivariate mean was considered as test statistic. As expected, the hypothesis of the two samples coming from the same distribution was rejected at any significance level.

2. ML-MStat relation: we wanted to validate the hypothesis that MStat could have be seen as ML plus a random error (and thus backing up the idea of using ML as regressor and a random effect dependent on station).To do so, we perform a regression with ML as regressor: we got a coefficient estimate near one and a (permutational) significance at maximum level, validating our beliefs.

3. EC8 significance: we perform a permutational ANOVA (we had non normality between groups) to assess the effect EC8 on LogY. As we expected (see

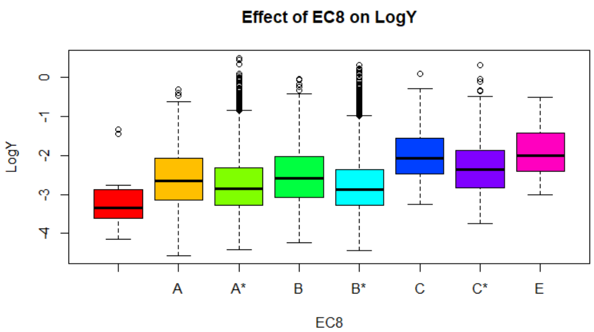

Figure 6: Differences can be seen between different classes

also Figure 6) the factor effect was actually relevant; however, going deeper:

- class A against class B: effect was not significant at any reasonable level (pvalue of 0.31).
- class A* against class B*: effect was not significant, even if the resulting p value was near 0.1 (0.109).
- class A against class E: effect was significant at 95% but not at 99% (p value of 0.013).

In conclusion, surely EC8 factor need a redesigning to be included in a model for FAS prediction. [5]

In addition, covariate selection was performed via permutational assessment of the significance level of the regressors, due to the fact that we could not rely on normality of residuals (see 'Non parametric approach' paragraph).

[5]We have to recall that being a classification done 'by hand' by some expert, it makes sense to not include it (it has a high cost to be collected)

Non parametric approach

As we were studying a complex phenomenon, which we knew might present problems at boundaries, we immediately chose splines as main tool for the non parametric regression.

As an introductory work, we investigated the relationships between the most important covariates and the response, evaluating the behaviour of different kind of splines (classical, natural and smoothing). We also considered different set of hyperparameters (knots and degree) and smoothing levels. As an example, in Figure 7 is reported the comparison between natural and smoothing splines modeling the Hypocentral distance-response relationship.

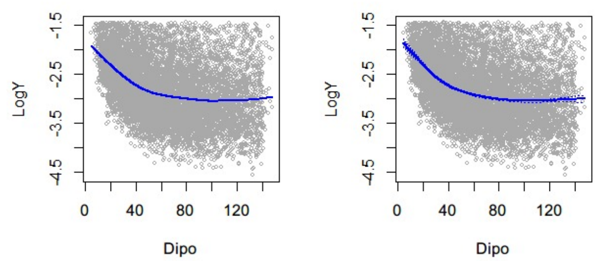

Figure 7: Smoothing and natural splines seems to perform similarly in this case

In the following step, we built a first, complete Generalized Additive model, including as regressors a magnitude term (ML), a distance term (Dipo), an interaction term between the two, a depth term (EvDpt) and a term accounting for soil characteristics(VS30). Of course, we also introduce a station-random effect. [6]

For all the smoothing terms we considered cubic regression splines, while the interaction term was modeled through a thin-plate spline.

Terms relative to magnitude, distance and

[6]Such model mimics the classical one

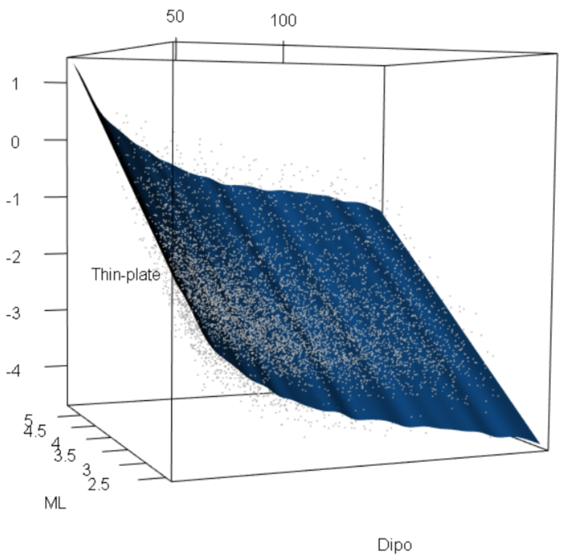

Figure 8: Thin-plate spline modeling the magnitude-distance interaction term

depth constituted the backbone of our model; interaction and VS30 terms significance was investigated through permutational tests (due to the fact that we could not rely on gaussianity of residuals):

- Interaction Term: significant at 95%
- VS30 term: not significant at 95%. It was significant at 90% level, but we finally decided to discard it. [7]

Note that, also in the non parametric approach, the random effect term dependent on stations kept a primary role in the regression.

Thus, the chosen version of the model accounted for three cubic regression splines (based on magnitude, distance and depth), a thin-plate one modeling the interaction between magnitude and distance and the random effect station term.

A note on VS30

Shear wave velocity in the upper 30 metres

[7] see 'A note on VS30' section for more details

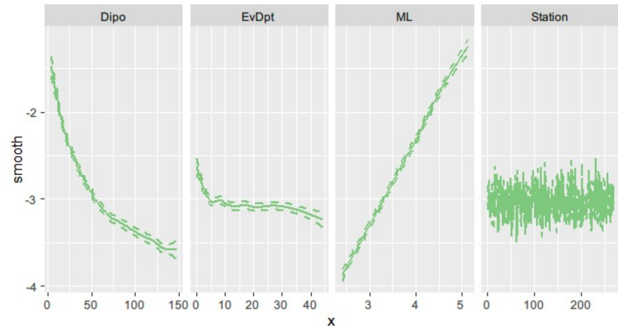

Figure 9: GAM components

is considered in literature about ground motion prediction equations as a proxy which accounts for soil condition; said so, it is straightforward to understand the theoretical reason why is included in models like the one proposed by Lanzano [1]. However, just a very small fraction of points was provided of real and precise measurement of this quantity; the great majority was furnished of an estimate, computed using the topographical slope (see Wald and Allen [2]). That estimator was developed looking at crustal condition in United States (mostly), and thus produces very coarse estimates in our context. Moreover, the whole approach seems a bit restrictive if applied to territory such as Southern Italy, which shows a great variability of soil conditions. In addition, in about 15% of data still had no value available, and thus when we followed the classical approach we had to discard them during fitting.

Keeping mind of all these previous considerations (and also the fact that the permutational significance test relative to the VS30 term produces a pvalue of 6%), we had the following idea: as the values of VS30 are measured (or computed) in each station location, discarding them we would have produced an increase of the station-dependent random effect, but no notable drawback in prediction. The turning point of such reasoning is that, not introducing explicitly VS30 in the regression, we were able to exploit the data discarded by the classical approach, making a more precise training of the GAM. As we will see in the next section, this strategy was the winning one: we obtained an enhancement in prevision performances, even evaluating the two models only on data for which VS30 was available.

Comparison between the models

The crucial step of our work was to make a comparison between the classical approach model (taken as benchmark) and the non parametric one (the GAM we selected in the section 'Non parametric approach'), in terms both on goodness of fit and, above all, prediction performances.

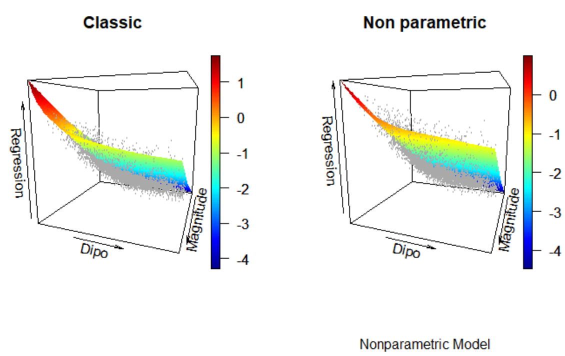

Figure 10: Regression Surfaces with respect to magnitude-distance grid

Firstly, we noted that both models showed homoschedastic, but not normal, residuals. At first sight, the residuals seems to follow a pretty similar behaviour. To assess that, we initially compute some goodness-of-fit indexes, taking into account different perspectives: Akaike Information Criterion (classical and second order) and Bayesian Information Criterion. As you can see from Table 1, they were not completely resolutive, as first two indicated the GAM approach, while the BIC the classical one. [8]

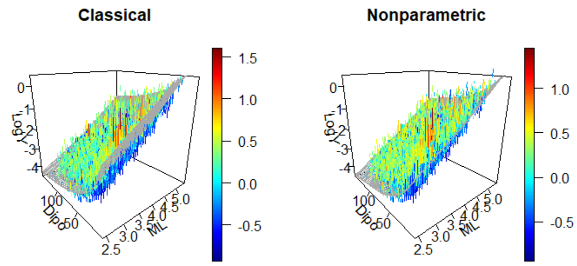

Figure 11: Residuals visualization on magnitude-distance grid

GoF	Classical	GAM
AIC	2075.16	1207.52
AICc	2075.16	1229.48
BIC	2140.04	3400.65

Table 1: Summary of GoF criteria

We finally chose to evaluate the performances of our models in prediction computing the root mean square error (RMSE) on a test set, composed by 30 observations not used in training. We found out that the prevision performances of the two models were comparable, with the non parametric one producing a slightly better score. In Figure 12 we reported in a graphical manner the behaviour of the two models on the test set, also providing reverse percentile intervals built upon the non parametric model. In addition, the plot also shows the errors made by the two models per each observation, through the bars in the bottom.

Figure 13 report the errors distribution and the RMSE scores, testifying the idea that the new approach can be preferred (or at least be considered comparable).

Of course such a small testing set could not provide us enough certainty, thus we refined our RMSE estimates performing a ten-fold

[8] In addition, due to the highly flexible nature of the considered models, goodness of fit criteria might not be too reliable

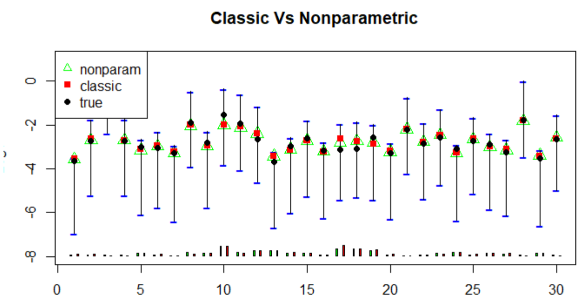

Figure 12: Classical and Non parametric performances in prediction

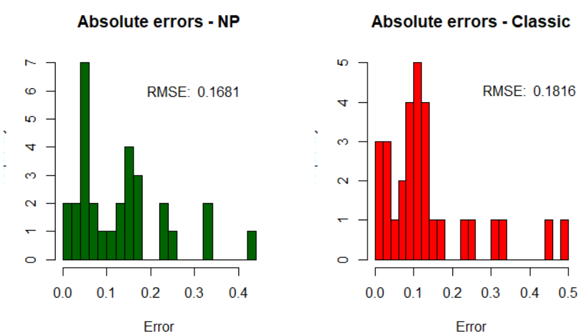

Figure 13: Errors distribution

cross validation: [9]

	Classical	GAM
RMSE	0.2629	0.2565

Table 2: 10fold CV RMSE estimates

Conclusions

The whole non parametric approach was based upon an assumption we made about the true generative mechanism of data: instead of performing the two-step procedure to identify the frequency-dependent parameters, and then perform a linear regression with mixed effects, we thought that a non parametric regression approach (performed one-shot on the dataset of reference frequency 0.50HZ) would have been able to grasp all the interesting patterns. Basically, instead of fixing a specific, hand-designed shape for the model, which needed additional parameters to account for the chosen frequency, we allowed more flexibility to do that implicitly. That is a key point: performances are comparable (even better) but the time needed to train the model significantly drop. [10]

To sum up, we can conclude that the non parametric approach in predicting the intensity measure of interest was surprisingly efficient, and could represent a worthy alternative to be explored in future studies.

References

[1] Giovanni Lanzano et al. “A Revised Ground-Motion Prediction Model for Shallow Crustal Earthquakes in Italy”. In: *Bulletin of the Seismological Society of America* XX.XX (2019).

[2] David Wald and Trevor Allen. “Topographic Slope as a Proxy for Seismic Site Conditions and Amplification”. In: *Bulletin of the Seismological Society of America* 97.5 (2007), pp. 1379–1395.

[9] Evaluation of the two models performances was done considering only the observations which had an available measurement of VS30

[10] A further analysis could be the scaling of performances of nonparametric models with frequency: we actually verified that the trend showed here is confirmed also at slightly higher frequencies [0.53HZ, 0.56HZ], but it could scale until values bigger than 20