Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-164-RC1, 2021

Comment on essd-2021-164

Martijn van den Ende (Referee)

Referee comment on "INSTANCE – the Italian seismic dataset for machine learning" by Alberto Michelini et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-164-RC1, 2021

This manuscript describes a newly created dataset comprising broadband seismic recordings of earthquakes and ambient noise in Italy, called INSTANCE. Public datasets are of utmost importance to advance Machine Learning in seismology, and ensure their reproducibility. I therefore applaud the authors for their efforts to create a dataset with Machine Learning applications in mind. The authors perform various quality checks and provide metadata based on which the user can determine which data to include in their investigations. Moreover, the INSTANCE dataset could be suitable for analyses outside of the domain of Machine Learning, and the authors facilitate these by including detailed metadata of both the earthquake and noise recordings.

The manuscript that describes the dataset is well written and provides an extensive analysis of various summary statistics of the dataset, as well as a useful review of other datasets that are currently available. I have a few minor suggestions for clarifications in the text (see below), but other than that I think that the manuscript is in good shape.

The project webpage and GitHub repository are clearly structured, and include Python notebooks to reproduce the figures presented in the manuscript. The example data represent only a subset of the data analysed in the manuscript, and so the figures look different, but I understand that it could be too computationally expensive to re-run the analyses on the full dataset. In principle the users could apply the same code to the full dataset if they wish to exactly reproduce this study. One suggestion I would like to make here is to consider some form of versioning of the INSTANCE dataset. It is not unimaginable that the dataset be modified or expanded in the future, for instance to re-classify misclassified earthquake or noise traces, or to include future seismic events of great significance. By including a version tag (and changelog), users can specify which version of INSTANCE they used in their analyses, which would improve the reproducibility of future studies.

As for the core data files (I reviewed the sample dataset): the waveforms are stored in

datasets labelled in correspondence with their trace name, which makes it convenient to extract only the waveforms that match a metadata query. I did notice that in many cases (about 1 in 3 traces), the evaluation of EQTransformer has yielded only NaNs. I'm not sure how to interpret this, because I would expect an integer number (zero if no detection was made) instead of NaN. Perhaps the authors could check this and mention how NaN should be interpreted, or replace them with zeroes where applicable. Nonetheless, the fact that EQTransformer fails in 1 out of 3 cases is a bit startling, and underscores the need for dedicated datasets for specific regions to (re)train Machine Learning models. Lastly, I noticed that after bzip decompression the data files take up a lot more disk space, so it could be helpful for users to indicate on the INSTANCE webpage what the data size after decompression is.

Overall, I think that INSTANCE is an important contribution to the seismic community, and I recommend acceptance of the manuscript after some (very) minor revisions.

Kind regards,

Martijn van den Ende

Minor comments on the manuscript (P = page; L = line):

1. P1, L19-20: in addition to the URLs, I would add the names of the ML platforms to make these more recognisable (and in case the URLs are changed in the future).
2. P6, L1-2: I don't quite understand that is meant here. Could the authors clarify this?
3. Figure 3, panel d: I would plot these data on a linear scale to better see the trends. It doesn't seem very logical to me to plot azimuth on a log scale.
4. Figure 1, panel a: the colour scale does not really provide any insights into the depth of the vast majority of earthquakes on this map (practically all the events are coloured deep red). To make this panel more informative, I would tailor the colour scale to the crustal earthquakes and accept clipping for the deep earthquakes. Also, the jet colourmap is hard to interpret in grayscale (when printing) and by readers with colour vision deficiencies. I would recommend re-rendering this figure with a perceptually uniform colourmap like viridis or cividis.
5. P16, L8, Fig. 6: maybe I missed a mention of this earlier in the manuscript, but which velocity model is used to calculate the theoretical arrivals / residuals? Is the model consistent for the entire dataset?
6. P17, L11-12: since the mean is removed from each trace, I would be surprised if the mean would be anything else than zero. The figures that show histograms of the mean are therefore not very informative and could be omitted, in my opinion.
7. P27, L11: could the authors confirm that the 120s time window is different for each station? In the way it is written now, it is not fully clear to me. I think it's unlikely that one would find a 120s window common to all stations that matches all of the criteria

simultaneously, but this is not explicitly mentioned in the text. This would be an important consideration when making any assumptions regarding the spatial coherence of the noise.
8. P32, L4: personally I would also have created a dataset with low SNR waveforms to complete the “spectrum” from noise to clear earthquakes (useful for microseismicity studies), but I understand that this may be beyond the scope of the project. Perhaps this is something that the authors could consider including in the future (see my comment about versioning).