Geometry-independent realistic noise models for synthetic data generation

C. Birnie, M. Ravasi

December 15, 2021

Abstract

Synthetic datasets are vital for the development and benchmarking of new processing and imaging algorithms as well as in the training of machine learning models. It is therefore important that such datasets are generated with realistic noise conditions making them resemble as much as possible their corresponding field datasets. Building on previously developed covariance-based noise modelling, we propose an extension of such an approach that aims to translate a noise model onto a user-defined geometry by means of Gaussian Process Regression. Starting from a synthetic data, we show that noise models can be generated and transformed into a desired geometry whilst keeping the same underlying statistical properties (i.e., covariance and variogram). The modelling procedure is subsequently applied to the ToC2ME passive noise dataset transforming the actual 69-sensor acquisition geometry into a gridded, 56-sensor array. The ability to generate realistic, geometry-independent noise models opens up a host of new opportunities in the area of survey design. We argue that by coupling the noise generation and monitoring algorithms, the placement of sensors could be further optimised based on the expected microseismic signatures as well as the surrounding noise behaviour.

Geometry-independent realistic noise models for synthetic data generation

Introduction

Over the years significant effort has been put into the development of noise modelling procedures for the generation of realistic noise to be incorporated in synthetic datasets. Whilst until recently this effort found primarily applications in the testing of processing and imaging algorithms, with the rise of interest in Machine Learning for seismological applications, synthetic seismic datasets have also become the go-to option for training datasets. As such, it is vital that these datasets are indistinguishable from the field data onto which the trained model will be applied. Despite the common use of additive White, Gaussian Noise (WGN), it is well documented that noise is the summation of various noise sources and that noise, as a whole, is rarely, if ever, stationary, white or Gaussian. Focussing on microseismic monitoring, Birnie et al. (2020) compared how algorithms benchmarked on WGN compare to synthetics generated with recorded noise. It was shown that by adding WGN, the performance of detection procedures was underestimated whilst that of localisation and moment tensor inversion was overestimated. This result highlighted that WGN can fail to reveal pitfalls that arise due to the presence of coherent noise.

A variety of noise modelling techniques have been proposed over the years, each with their own strengths and drawbacks based on the assumptions onto which the methods are built. For example, Pearce and Barley (1977) assume that noise is stationary - an unrealistic assumption. Whereas another commonly used method of distributed surface sources generally fails to capture the complexities of recorded noise (Dean et al., 2015). The Isolated-CoVAriance (ICOVA) method of Birnie et al. (2016) overcame these assumptions to generate noise models that accurately imitate recorded noise, nevertheless these models are restricted to the array geometry in which the sample noise has been recorded.

In this work we extend the ICOVA modelling workflow to allow for changes in the acquisition geometry as part of the noise modelling procedure. Such a goal is achieved by using geospatial interpolation in the form of Gaussian Process Regression (GPR) (Rasmussen, 2003). The generation of geometry-independent noise models opens up the opportunity to incorporate realistic noise effects alongside other parameters related to the expected location and focal mechanisms of events in the survey design process for microseismic monitoring.

Methodology

The proposed workflow for generating geometry-independent noise models is outlined in Figure 1. It comprises of two steps: the first step is responsible for the characterisation of the recorded noise field and the generation of multiple realisations with same statistical properties, constrained to the original acquisition geometry. The second step leverages Gaussian Processes to transform the noise realisations to a user-defined geometry. Below we discuss in greater detail these steps.

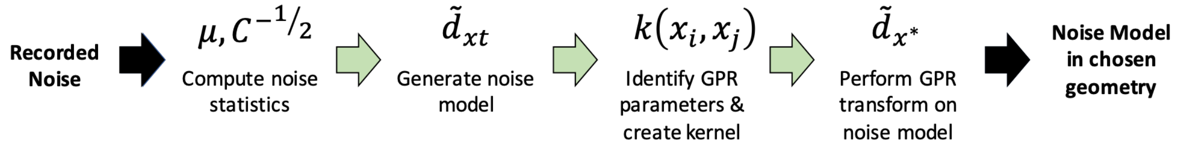


Figure 1 *Workflow for generating geometry-independent noise models.*

Given a noise field recorded by a set of stations $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ for a certain time period T, Birnie et al. (2016)'s ICOVA noise modelling starts by separating it into N noise realisations of length $T_{reals} = T/N$, which are assembled in a matrix $\hat{\mathbf{D}} = [\hat{\mathbf{d}}_{xt,1} - \boldsymbol{\mu}, \hat{\mathbf{d}}_{xt,2} - \boldsymbol{\mu}, ..., \hat{\mathbf{d}}_{xt,N} - \boldsymbol{\mu}]$. Here $\boldsymbol{\mu}$ represents the sample mean computed for the each receiver and the subscript $\cdot_{xt}$ indicates that the each vector $\hat{\mathbf{d}}_{xt}$ contains the vectorized time-space noise patch. The sample covariance, $\mathbf{C}$, is further computed as illustrated in equation 1a. This is followed by a decomposition of the covariance matrix into its upper and lower triangular matrices (equation 1b), which is here accomplished by means of randomised SVD. Finally, multiple noise realisations, $\tilde{\mathbf{d}}_{xt}$, can be generated by multiplying a base random vector, **b**, with the lower triangular matrix and summing back the sample mean (equation 1c).

$$\mathbf{C} = \hat{\mathbf{D}}\hat{\mathbf{D}}^T/N \tag{1a}$$

$$\mathbf{C} = \mathbf{C}^{1/2}(\mathbf{C}^{1/2})^T \tag{1b}$$

$$\tilde{\mathbf{d}}_{xt} = \mathbf{C}^{-1/2}\mathbf{b} + \boldsymbol{\mu} \tag{1c}$$

Up to this point, the modelling procedure is limited in that it is constrained to the geometry onto which the noise was recorded. To overcome this limitation, GPR is subsequently applied to each time slice of every noise model. GPR is a statistical modelling technique that assumes the input data to originate from a multivariate normal distribution whose spatial correlation is estimated from the available data points (i.e., noise values at each original station for a given time step). This is defined by the so-called kernel of the underlying Gaussian Process that in our case is chosen to be a square-exponential function of distance d between two stations at locations $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$k(\mathbf{x}_i, \mathbf{x}_j) = c \cdot \exp \frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2l^2}, \tag{2}$$

where c and l are two scalars. In the context of spatial data, GPR is closely linked to Kriging and the aforementioned scalars are also referred to as the sill and range of the variogram of the sample data. In our implementation, these values are initialized by fitting the sample variogram and later on further optimised as part of the GPR process. Ultimately, each time slice of the previously generated noise, $\tilde{\mathbf{d}}_x$, is transferred to a new set of stations $\mathbf{X}^* = \{\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_m^*\}$ by evaluating the following equation:

$$\tilde{\mathbf{d}}_{x^*} = \mathbf{K}_*^T(\mathbf{K} + \varepsilon\mathbf{I})^{-1}\tilde{\mathbf{d}}_x \tag{2}$$

where **K** is a matrix containing the kernel in equation 2 evaluated for each station pair, whilst $\mathbf{K}_*$ contains the kernel evaluated for each combination of physical and new stations. ε is a stabilization factor and **I** represents the identity matrix.

Results

A synthetic noise dataset is first used to validate the modelling workflow proposed in this paper. Given a dense grid geometry illustrated by grey boxes in Figure 2(a), the noise field is here generated by defining its covariance matrix as $\mathbf{C} = \mathbf{C}_t \otimes \mathbf{C}_x$ where $\mathbf{C}_t$ and $\mathbf{C}_x$ are modelled using exponential and squared-exponential covariance functions, respectively. The modelled noise is then subsampled onto an irregular grid (black circles in Figure 2(a)) and run through the first step of the ICOVA modelling procedure. Finally, GPR is applied to each generated noise realization to transform the original geometry to a new, desired geometry (red triangles in Figure 2(a)). Figure 2(c,d) illustrate the amplitudes of the noise from the original noise model (circular points) and transformed noise model (triangular points) at timesteps $t = 0$ and $t = 10$. In order to understand if the noise statistics are maintained throughout the entire process, both the covariance and variogram of each noise model is computed and illustrated in Figure 3. Figures 3a and b illustrates the initial covariance and variogram values computed from the synthetic generated noise (on the dense grey grid). Figures 3e and g show the sample covariance matrix computed from the generated noise on the original and desired geometries. Similarly Figures 3f and h display their respective sample (red) and fitted (blue) variograms. By comparison of the top and bottom rows, we can observe that minimal errors are introduced in the ICOVA and GPR modelling steps. Whilst there are a few spurious artefacts present in the bottom row, in general we observe a high alignment which emphasises the validity of the geometry-independent noise modelling procedure. After validating our procedure on the synthetic dataset, our workflow is applied to the ToC2ME field dataset from Alberta, Canada. The array consists of 69 geophones in a pseudo-random pattern, as shown by black circle in Figure 4(a). The desired geometry, illustrated by red dots, is a gridded array of 56 sensors spanning a

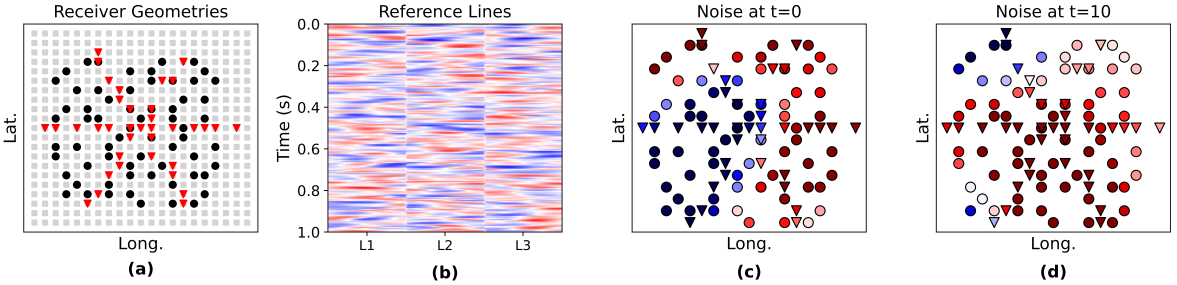


Figure 2 *(a) Survey geometries of a dense grid for reference, a sparser grid for imitating field collection, black circles, and a star-shape array on which to compute the noise levels, red triangles. (b) Reference synthetically generated noise for the full survey area. (c) and (d) illustrate stations where noise was recorded, circular, and stations where GPs have been used to estimate the noise, triangular.*

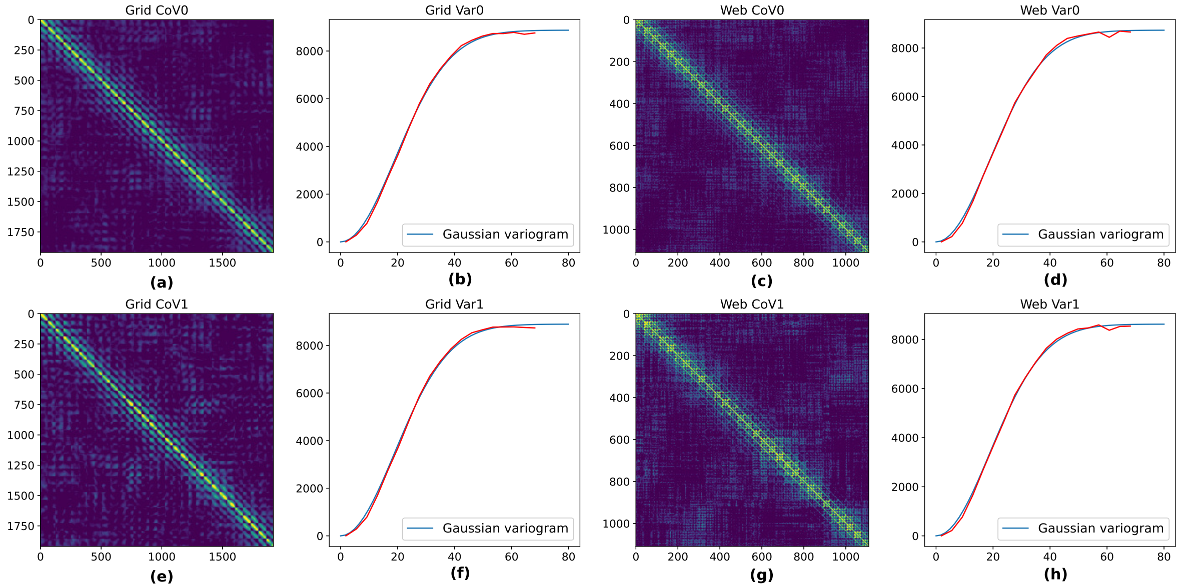
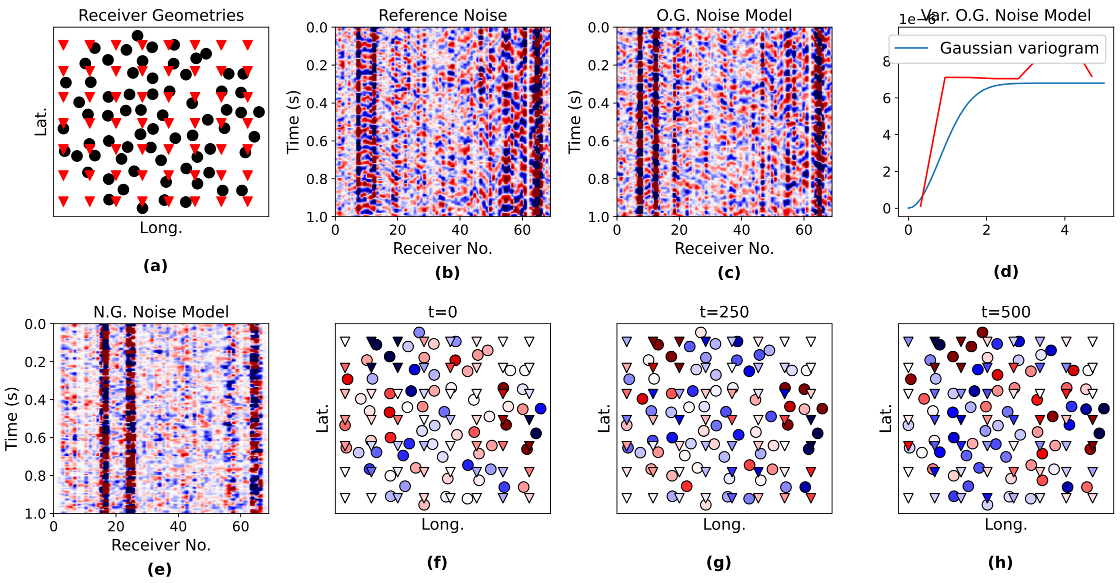


Figure 3 *Covariance matrices for (a,c)reference noise from the grid and star geometries respectively, for (e) ICOVA modelled noise on the grid array and for (g) estimated modelled noise on the web array computed from the ICOVA modelled noise. (b,d,f,h) show variograms average over time for the same noise and geometries portrayed to their left with a Gaussian variogram comparison.*

similar range to the original geometry. One hour of recorded noise is used to generate the ICOVA noise model with Figure 4(b,c) illustrating a one second recording of the recorded noise and modelled noise, respectively. The range and sill for the Gaussian Processes are identified from the sample variogram of the ICOVA noise model as shown in Figure 4(d). A noise realisation from the desired receiver geometry is illustrated in Figure 4(e) with the remaining plots providing a comparison between the noise models at different timestamps. In general we observe a high similarity between the ICOVA noise model and the transformed model, although the results are not perfect. This is likely due to the fact that whilst the noise field as a whole presents some spatial correlation (as shown by the sample variogram), it does still contain some small-scale variations that are difficult to capture via Gaussian Processes.

Discussion and Conclusion

In this work we have shown that it is possible to fully characterize highly complex noise fields by means of their spatio-temporal covariance matrix. One of the key benefits of such an approach, is that by simply retaining the sample noise statistics $\mathbf{C}$ and $\boldsymbol{\mu}$ along with the original geometry, a new noise realization



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4 *(a) original survey geometry, black circle, and desired survey geometry, red triangles. (b) One second of recorded noise and (c) one second of modelled noise for original geometry. (d) Variogram of modelled noise. (e) One second of modelled noise transformed onto the desired geometry with (e,f,g) illustrating the original, circle, and transformed, triangle, noise values at different timesteps.*

can always be generated. Whilst outperforming all other noise modelling procedures with respect to the reality of the noise generated, the ICOVA noise modelling method has a severe drawback of being restricted to the geometry in which the sample noise was recorded. We have further shown here that GPR represents a valid approach to alleviate such a limitation and generate realistic, geometry-independent noise models. This opens up the possibility for in-depth analysis of optimum geometries based not only on the expected microseismic signatures but also site conditions, for example, location of nearby roads, forests or industrial buildings.

Finally, future work will consider how noise models from different monitoring sites can be combined to generate tailor-made noise models based on the expected environment conditions of new site prior to any deployment of sensors. As an example, for a potential carbon storage site this may allow evaluating the possible impact of a processing plant and/or a road system yet to have been recorded at such a site.

Acknowledgements

The authors would like to thank the University of Calgary for releasing the ToC2ME dataset.

References

Birnie, C., Chambers, K., Angus, D. and Stork, A. [2016] Analysis and models of pre-injection surface seismic array noise recorded at the Aquistore carbon storage site. *Geophysical Journal International*.

Birnie, C., Chambers, K., Angus, D. and Stork, A.L. [2020] On the importance of benchmarking algorithms under realistic noise conditions. *Geophysical Journal International*, **221**(1), 504–520.

Dean, T., Dupuis, J.C. and Hassan, R. [2015] The coherency of ambient seismic noise recorded during land surveys and the resulting implications for the effectiveness of geophone arrays. *Geophysics*, **80**(3), P1–P10.

Pearce, R. and Barley, B. [1977] The effect of noise on seismograms. *Geophysical Journal International*, **48**(3), 543–547.

Rasmussen, C.E. [2003] Gaussian processes in machine learning. In: *Summer School on Machine Learning*. Springer, 63–71.