16th World Conference on Earthquake, 16WCEE 2017

Santiago Chile, January 9th to 13th 2017

Paper N° 4941

Registration Code: S-S1464658037

GROUND MODEL ENSEMBLE SELECTION BASED ON INFORMATION THEORY AND GLOBAL INVERSION OF SURFACE WAVE DISPERSION DATA

A. Savvaidis[(1)], M. Ohrnberger[(2)], M. Whatelet[(3)], and C. Cornou[(4)]

[(1)] *Project Manager of TexNet, Bureau of Economic Geology, Jackson School of Geoscience, The University of Texas at Austin, Texas, U.S.A., Alexandros.Savvaidis@beg.utexas.edu*

[(2)] *Research Associate, Institute of Earth and Environmental Sciences, University of Potsdam, Karl-Liebknecht-Str. 24, 14476 Golm, Germany, Matthias.Ohrnberger@geo.uni-potsdam.de*

[(3)] *Researcher, Institut des Sciences de la Terre (ISTerre), Maison des Géosciences BP53X 38041 Grenoble Cédex 9, France, Marc.Wathelet@ujf-grenoble.fr.*

[(4)] *Researcher, Institut des Sciences de la Terre (ISTerre), Maison des Géosciences BP53X 38041 Grenoble Cédex 9, France, Cecile.Cornou@obs.ujf-grenoble.fr*

Abstract

The purpose of this study is to obtain a "reasonable ensemble" of one dimensional shear wave velocity models that explain an observed surface wave dispersion curve with its corresponding measurement errors. The difficulty of finding this set of models lies in the definition of what is to be considered "reasonable" on one hand and that we are unaware of the correct velocity depth function description to be used. Therefore we need to be able to compare different velocity depth model parameterizations among each other in a correct way. This particular problem is approached by testing all perceived velocity depth model parameterizations until complexity (related to number of free parameters) doesn't allow the exploration of the model space (curse of dimensionality) and using a bias correction term (here AICc) for comparing the model fit to the data. Having found the overall best earth model, we can predict the maximal acceptable threshold for the misfit (including bias correction) of the mean curve. Consequently, we present a strategy for deriving a combined trans dimensional model ensemble using data uncertainties resulting in site characterization defined as a distribution of most likely models explaining our data.

Keywords: Surface Wave Data; Uncertainties on Earth Models; Genetic Algorithm; Akaike

1. Introduction

In applied geophysics we may use the microtremor array method to measure Dispersion Curve (DC) of surface waves and then derive the shear wave velocity profile. Given the nonlinear relation between ground model parameters and observation, directed Monte Carlo search techniques are common approaches for obtaining solutions to the inverse problem. Compared to linear methods, the goal is not necessarily to find a 'best fitting' ground model to the data, but a set of models explaining the data within its uncertainty bounds. Parameterization of the model space has a significant influence on the resulting misfit and the capabilities of different inversion algorithms to fit to the data: what should then be the criteria for ranking different model ensemble related to different parameterization? For answering this issue, we use the corrected Akaike criteria (AICc) [1] that incorporates the Degrees of Freedom (DoF) of the parameter space for bias-free comparison of model fits.

A synthetic model was considered in order to study the variability of shear wave velocity profiles obtained from inversion. In the current paper we are investigating the effect of the different model selection after the calculation of a different number for different degrees of freedom for a model decision using the corrected Akaike criteria [1], for the special case of a least-squares estimation with normally distributed errors. Based on this we attempt to define a strategy in order to calculate a model ensemble using data uncertainties and a global search algorithm (Monte-Carlo), for site characterization.

2. Data and Parameterizations Used

The Dispersion Curve (Fig. 1) used is the theoretical curve corresponding to an earth model with a boundary between loose sediments and rock at 180 meters. A Dispersion Curve (DC) with sixty-five data points is used ranging from 1 to 20 Hz.

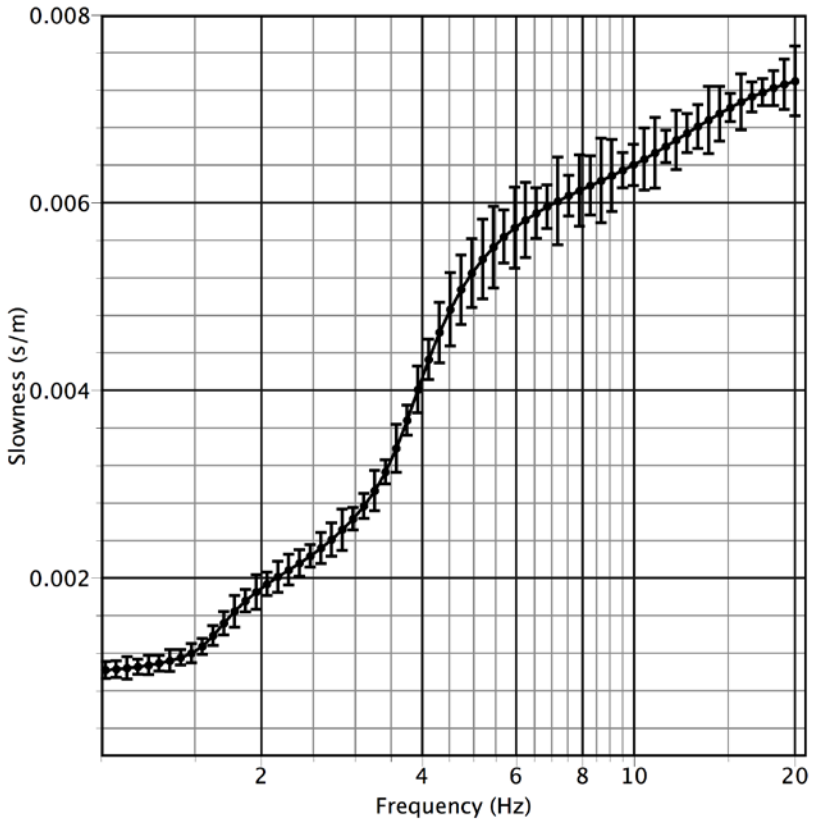

Fig. 1 – Synthetic Dispersion Curve used in this study

For our test four different parameterization groups (Fig. 2) are used as for the shear-wave velocity. Uniform layers were used as the simplest case. However, since in most of the sedimentary environments the upper shallow layers are not as

much consolidated as the deeper ones we considered that a power-law and a linear increase for the shear wave velocity is a good approximation for the first layer. Between the first layer and the Half-Space (HS) an increasing number of uniform layers is added each time for each group of parameterizations until we reach a maximum number of nine layers over HS.

The four different groups of parameterizations correspond to: (a) uniform layers, (b) one power law layer over uniform layers, (c) a linear increase layer over uniform layers, and (d) uniform layers like in (a) but with the depth interval of each layer to be defined by a geometrical progression based on the wavelength range. In the later case the thickness of the first layer is set to a quarter of minimum wavelength and the depth of the last layer is set to half of the maximum wavelength. For groups (a)-(c) the Bottom Depth of each layer was varying from 3-300m, as this results from the 1/3 of the min and max wavelength of the DC.

In all parameterization, the compressional wave velocity follows the same model as for the shear wave velocity and the depth for its layer is linked to the shear wave velocity layer. Only for group (d) the compressional wave velocity follow a linear increase, with five sub layers until the Half-Space. The shear wave velocity is varying for all layers from 50 to 2500 *m/sec*, except for the half space that is between 150 to 3500 *m/sec*. The compressional wave velocity is varying from 200 to 5000 *m/sec*. The density is uniform for all the layers and equals to 2000 kg/m^3.

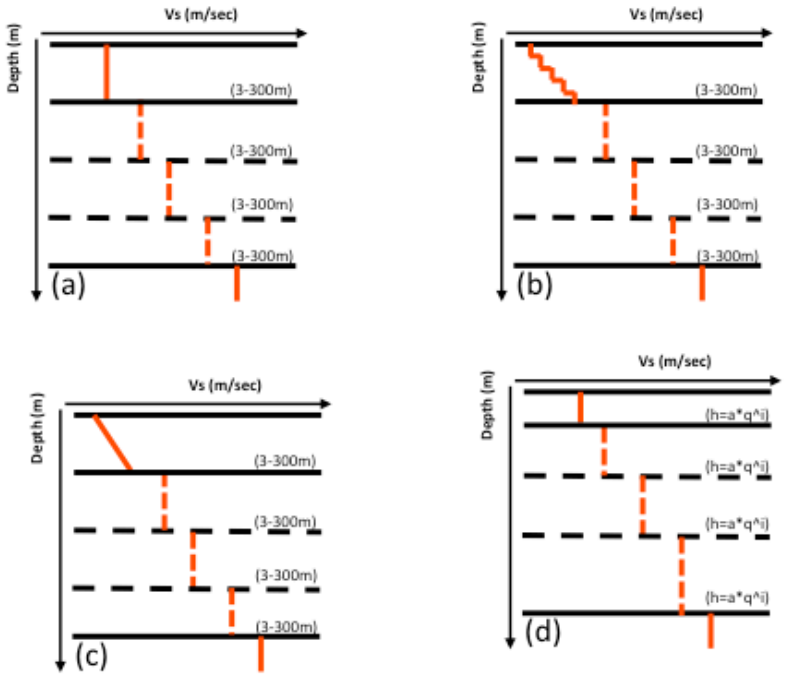

Fig. 2 – Four groups of parameterizations used for the inversion. Uniform layers are used except for the first layer, that is either uniform (a) and (d), either follow a power law with depth (b), or a linear increase (c). The depth is varying for each layer from 3-300m, as defined from minimum and maximum wavelength divided by three. Only for (d) the depth is progressively increasing

3. Method

A direct search algorithm was used for the surface-wave inversion [2,3]. This scheme is based on a neighborhood algorithm (NA) that is a stochastic direct-search method for finding models of acceptable data fit inside a multidimensional parameter space [4]. Pseudo-random samples are generated in the parameter space (V_p, V_s, density and depth) and the dispersion curves are computed for all these models. The comparison of the resulted dispersion curve for each model of the parameter space with the measured dispersion curve provides one

misfit value that indicates how far the generated model is from the true solution. Like most of the direct search methods NA make use of previous samples for guiding the search in order to get improved models. Once the data misfit function is known at all previous samples (forward computations), the neighborhood algorithm provides a simple way of interpolating an irregular distribution of points, making use of Voronoi geometry to find and investigate the most promising parts of the parameter space. In order to check the robustness of the method four different seeds where used since one could claim that your resulted models depend on the random seed used in the inversion.

The misfit value is evaluated using the uncertainty of the real data,

$$misfit = \sqrt{\sum_{i=0}^{n_F} \frac{(x_{di}-x_{ci})^2}{\sigma_i^2 n_F}} \quad (1)$$

where x_{di} is the phase velocity of the data curve at frequency F_i, x_{ci} is the velocity of the calculated curve at frequency i, σ_i is the uncertainty of the frequency samples considered and n_F is the number of frequency samples considered. The misfit value described here will be referred as Normalized Misfit (NM) in the text.

Subsequently we used an information criterion, in order to conclude to the best model estimate for different parameterizations, and also investigate the uncertainty due to the number of models estimated for the specific data set described before. Akaike Information Criteria (AIC) is an estimator for such a problem that also includes the DoF (number of parameters) of a model into the best model criteria. As inferred in [1], in the case of Least Squares estimation supposing normally distributed errors for all our models AIC can be expressed as,

$$AIC = n_F log_e(misfit) + 2K \quad (2)$$

where, K is the number of parameters, i.e. DoF. Since in our case $n/K < 40$ the AICc is used where similar to AIC is expressed as,

$$AIC_c = n_F log_e(misfit) + 2K + \frac{2K(K+1)}{(n_F-K-1)} \quad (3)$$

In Figure 3 we present the AICc values for an increasing number of DoF for 65 data points (n_F) and a given misfit function (blue line),

$$misfit = \frac{\sqrt{DoF}}{10^{1.5*DoF}} \quad (4)$$

The AICc values that correspond to one standard deviation are shown with a dashed black line. The blue dashed line point the DoF with the lowest AICc value and it is also crossing the one standard deviation line. The area prescribed between the dashed black line and the blue line bounds AICc values between one standard deviation and minimum AICc for different DoF. One should select a model ensemble out of the total models calculated through an inversion scheme. Of course the models with the lower AICc values are the lower limit line but one should include all possible models that fit the data to one, two or to even three times the standard deviation. In that case the ensemble will include a numerous amount of models. However, could one suppress the models based on an AICc limit? Even if we select as an upper boundary the one standard deviation limit one should not use the complete ensemble. Of course one could use different AICc upper boundaries, presented with red color lines in Figure 3. But still, one should not accommodate the models having a high DoF removing away the most complex ones (Occam's razor). For this reason, we suggest to use as an upper AICc limit the AICc value that correspond to misfit equal to one for the parameterization with the lowest AICc value among all tested models. This limit is denoted with a continuous red line in Figure 3.

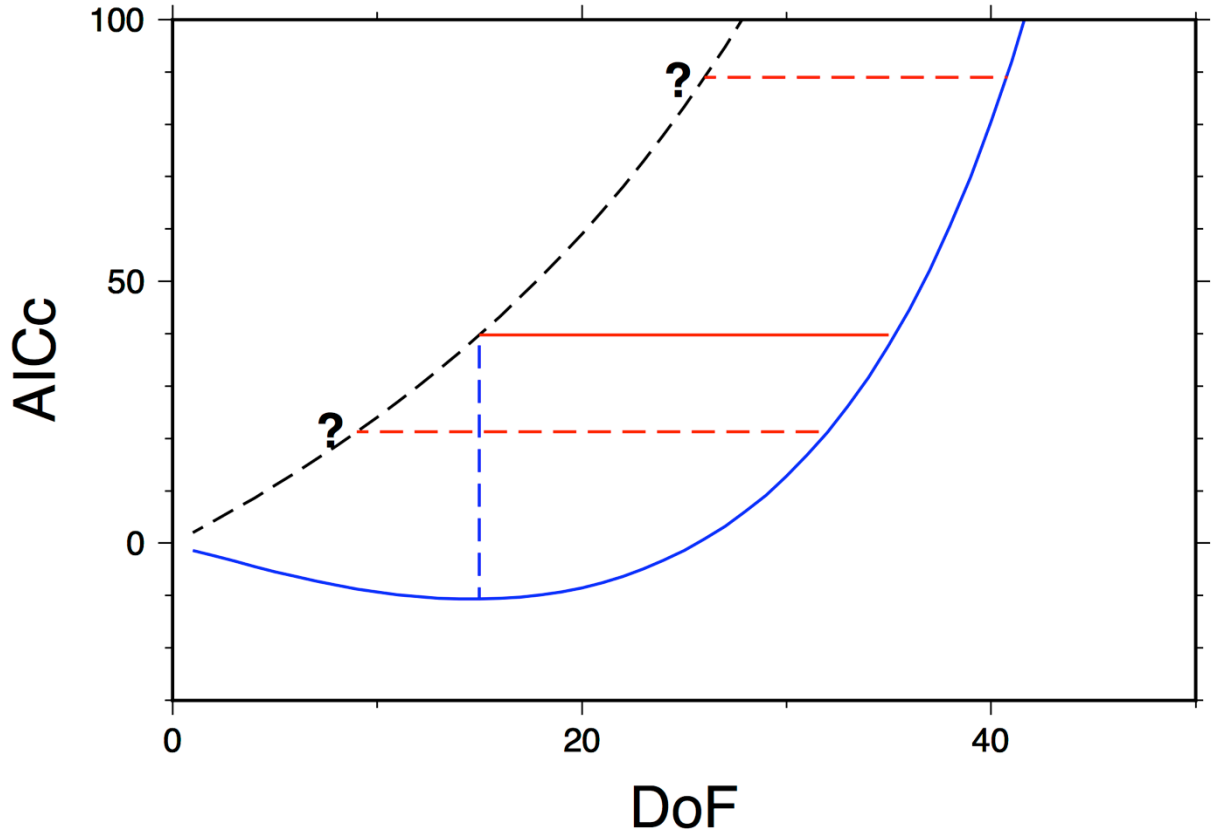

Fig. 3 – Evolution of AICc with Degrees of Freedom (blue curve) for a misfit function along with the AICc for misfit of 1 (black dashed curve). The blue dashed curve points the minimum AICc value and the corresponding AICc value for misfit of 1. The area prescribed between the dashed black line and the blue line show the models of different DoF that has misfit equal to one standard deviation and minimum AICc values. The red lines correspond to different AICc ceilings with the continuous line matching the AICc value of one standard deviation of the lowest AICc value for all DoF.

4. Model Ensemble Calculation

Extensive calculation was implemented for each run (4 runs for each set of parameterization) using the Neighborhood Algorithm with increasing number of calculations for high Degrees of Freedom until the misfit converged to its lowest value. This resulted in at least 4 million model calculations for each parameterization reaching a value of up to 64 million models in some cases to reach convergence. The evolution of Minimum misfit and corrected Akaike with the Degrees of Freedom for different parameterizations are presented in Figures 4a and 4b, respectively. As expected for the Minimum Misfit plot after a considerable number of increasing degrees of freedom the minimum misfit values are very similar with small dependence on the parameterization selected for the inversion. However, for the AICc evolution, as expected there is a lowest AICc value that is different for each group of parameterizations. The lowest AICc value corresponds to a different parameterization (DoF) than the Minimum Misfit value. The AICc is favoring the simplest models since although parameterizations of high DoF manage to have a low misfit they are over-interpreting the data [5].

Among all tested models the one with seven uniform layers over half space with the depth interval of each layer to be defined by a geometrical progression, is the most preferable parameterization that succeed in resolving the lowest AICc (Figure 4b). The corresponding AICc value for misfit equal to one is also denoted

with a dashed magenta line for this parameterization, along with an upper limit as of AICc value that corresponds to that misfit (magenta line).

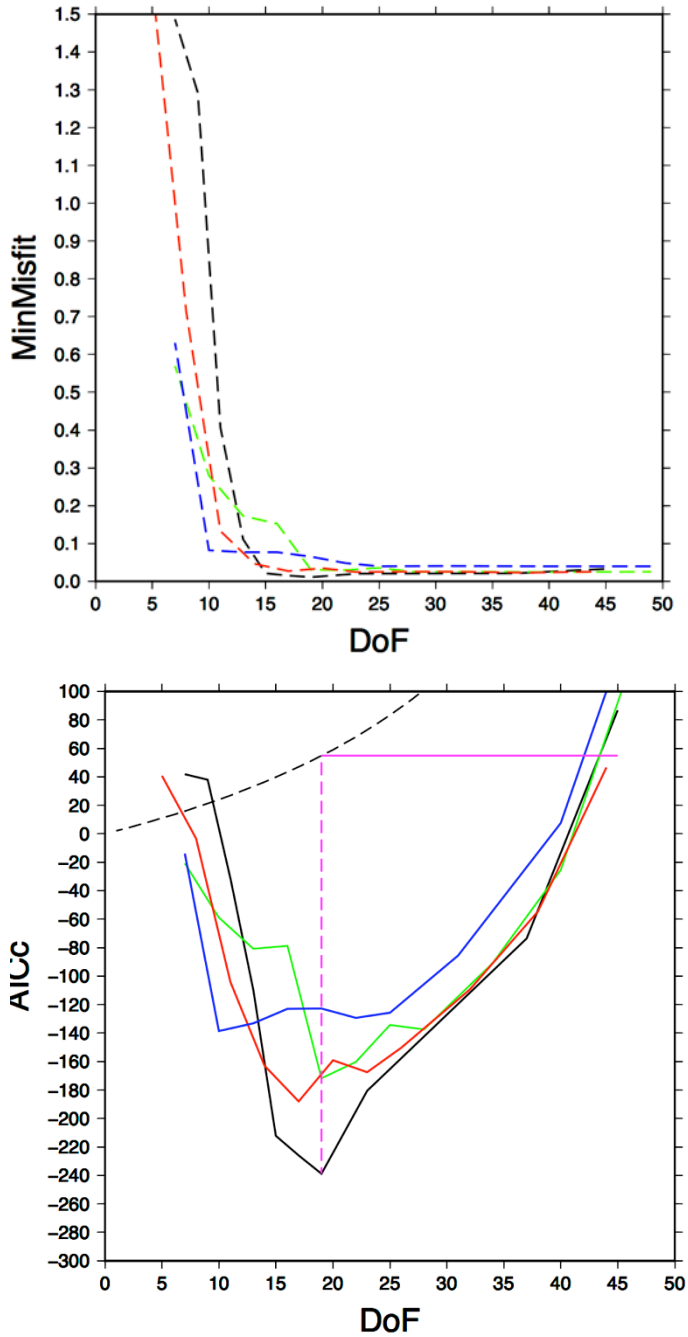

Fig. 4 – Minimum Misfit (a) and AICc (b) Plots with the Degrees of Freedom (DoF) for the different parameterizations used. Results for the parameterizations of uniform (red), power-law (blue), linear-increase (green), and uniform with progressive depth limits (black) are shown.

5. Discussion and Conclusions

It is obvious from the information presented that the AICc could be used in order to select not only the best solutions for different parameterizations but also a model ensemble that sufficiently describes the Dispersion

Curve. As expected an intensive number of iterations can provide lowest misfit values even for parameterizations of higher degrees of freedom.

We suggest to utilize a model ensemble including the models with the lowest AICc for each parameterization up to the AICc value that corresponds to a misfit equal to one (*misfit*=1) for the parameterization with the lowest AICc value among all tested models. However, further investigation is needed on comparing the SH response of these models with the one of the starting model used during this exercise.

7. Aknowledgments

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Consortium of Organizations for Strong-Motion Observation Systems (COSMOS) Facilitation Committee for the Development of the COSMOS International Guidelines for the Application of NonInvasive Geophysical Techniques to Characterize Seismic Site Conditions.

6. References

References must be cited in the text in square brackets [1, 2], numbered according to the order in which they appear in the text, and listed at the end of the manuscript in a section called References, in the following format:

[1] Burnham K, Anderson D (2001): Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, **28**, 111-119.

[2] Wathelet M, Jongmans D, Ohrnberger M (2004): Surface wave inversion using a direct search algorithm and its application to ambient vibration measurements. *Near Surface Geophysics*, **2**, 211-221.

[3] Wathelet M (2008): An improved neighbourhood algorithm: Parameter conditions and dynamic scaling. *Geophysical Research Letters*, **35**, L09301. doi: 10.1029/2008GL033256.

[4] Sambridge M (1999): Geophysical inversion with a neighbourhood algorithm searching a parameter space. *Journal of Geophysical Research*, **103**, 4839–4878.

[5] Scherbaum F, Hinzen KG, Ohrnberger M (2003): Determination of shallow shear wave velocity profiles in the Cologne, Germany area using ambient vibrations. *Geophysical Journal International*, **152** (3), 597-612.