Available online at www.sciencedirect.com
ScienceDirect
Journal of Applied Geophysics 61 (2007) 39–55
JOURNAL OF APPLIED GEOPHYSICS
www.elsevier.com/locate/jappgeo

Rayleigh wave dispersion curve inversion via genetic algorithms and Marginal Posterior Probability Density estimation

Giancarlo Dal Moro *, Michele Pipan, Paolo Gabrielli

Department of Geological Environmental and Marine Sciences, University of Trieste, Via Weiss 1, 34127 Trieste, Italy

Received 30 September 2005; accepted 17 April 2006

Abstract

Surface wave dispersion curve inversion is a challenging problem for linear inversion procedures due to its highly non-linear nature and to the large numbers of local minima and maxima of the objective function (multi-modality). In order to improve the reliability of the inversion results, we implemented and tested a two-step inversion scheme based on *Genetic Algorithms* (GAs). The proposed scheme performs several preliminary "parallel" runs (first step) and a final global run using the previously-determined fittest models as starting population.

In this work we focus on the inversion of shear-wave velocity and layer thickness while fixing compressional-wave velocity and density according to user-defined Poisson's ratios and velocity–density relationship respectively. The procedure can nonetheless perform the inversion under different degrees of regularization, depending on the a priori information and the desired degree of freedom of the system.

Thanks to the large number of considered models, in addition to the fittest model, a mean model and its accuracy are evaluated by means of a statistical approach based on the estimation of the *Marginal Posterior Probability Density* (MPPD).

We tested the proposed GA-based inversion scheme on three synthetic models reproducing a complex structure with low-to-moderate velocity cover (also including a low-velocity channel) lying over hard bedrock. For all the considered cases the bedrock velocity and depth were properly identified, and velocity inversion was reconstructed with minor uncertainties.

The performed tests also investigate the influence of the first higher mode, the reduction of the frequency range of the considered dispersion curve as well as the use of different number of strata. While a limited frequency range of the dispersion curve (maximum frequency reduced from 80 to 40 Hz) does not seem to significantly limit the accuracy of the retrieved model, the adoption of the correct number of strata and the addition of the first higher mode help better focus the final solution.

In conclusion, the proposed approach represents an improvement of a purely GA-based optimization scheme and the MPPD-based mean model typically offers a more significant and precise solution than the fittest one.

Results of the inversion performed on a field data set were validated by borehole stratigraphy.

Keywords: Geophysical data inversion; Genetic algorithms (GAs); Surface waves; Rayleigh waves; Dispersion curve inversion; Posterior probability density (PPD)

1. Introduction

Surface wave analysis is an efficient tool to obtain the vertical shear-wave profile (Park et al., 1999). The large

* Corresponding author. Fax: +39 040 5582290.
E-mail address: dalmoro@units.it (G. Dal Moro).

0926-9851/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jappgeo.2006.04.002

amplitude and low attenuation allow an accurate reconstruction of the subsurface structure via inversion of the observed dispersion curve and smoothed seismic sections can be reconstructed by considering successive shots (e.g. Park, 2002).

The inversion procedure follows the determination of the dispersion curve (e.g. Park et al., 1999; Dal Moro et al., 2003a, in press) and allows reconstructing the vertical shear-wave velocity distribution from the observed dispersion curve. As surface wave inversion is a typical example of non-linear multi-parameter problem, the classical solution consists in a linearization of the set of considered equations.

The linear approach can be considered as an acceptable and computationally-effective solution only when some robust a priori information is available and a good starting model can thus be established that is close to the real solution. In fact, the main problem of the traditional approach is that the final solution intrinsically depends on the starting model and that poor or missing a priori information makes the final solution particularly weak.

When the considered problem is multi-modal[1] (i.e. the objective function exhibits several local minima and maxima), an approach based on the *Jacobian* matrix can fail because the starting model can be close to some local minimum that will attract it.

A simple example can highlight this aspect. We calculated the synthetic fundamental mode dispersion curve for a model consisting of six layers and then calculated the objective function [*root-mean-square*—see Eq. (1)] by varying only thickness and shear-wave velocity of the second and third layers[2] and by fixing the values of the other parameters to their actual values. In other words, as surface wave velocities are function of shear- and compressional-wave velocities, density and thickness of the layers, we let free only four out of twenty-three possible variables (the bottom layer is semi-infinite).

Fig. 1a shows a 3D plot of objective functions (calculated for the synthetic curve) for 4000 random models. Velocities and thickness are fixed to their correct values but for the second and third layers. For such layers velocities vary in the range between 150 and 400 m/s and thickness between 1 and 3 m. Actual values are 350 m/s and 2 m for the second layer and 200 m/s and 2 m for the third one (see model #1 in Fig. 2). The abscissas report the ratio between shear-wave velocity in the second and third layers (V_S ratio) and the ratio between the thickness of the second and third layers (THK ratio).

The complex result from this extremely simple and constrained example gives evidence of several local minima. The possible use of derivative-based linear methods would cause an evident *starting-model-dependent* solution.

To further highlight the problem, Fig. 1b shows a 2D plot for 4000 objective functions versus the V_S ratio, by keeping constant the layer thicknesses to the proper values (THK ratio equal to 1). In this case velocities vary in the range between 380 and 320 m/s in one layer and between 230 and 170 m/s in the other, confining the velocity ratio between 2.2 and 1.4. The distribution of points characterized by low-value objective functions is large even in the surrounding of the global-maximum area and gives evidence of the extreme non-linearity and multi-modality of the problem. Small variations of the V_S ratio produce "misleading" values of the objective function also in the surrounding of the correct solution. In such conditions, linear methods often fail to converge on the correct solution and, as far as heuristic methods are concerned, the scope of the search-space exploration becomes a crucial issue.

Several algorithms tackle this problem with the main goal of sampling a wide search space to detect the global minimum (or maximum) of a given function (e.g. Smith et al., 1992; Sambridge and Mosegaard, 2002).

Heuristic optimization schemes can be divided into enumerative, random search (uniform distribution of the search space sampling, such as Monte Carlo methods) and "importance sampling" (the search space is non-uniformly sampled because some function drives the search, such as Evolutionary or Genetic Algorithms).

Genetic algorithms (GAs) have been used to invert seismic velocities (Louis et al., 1999), seismic waveform (Stoffa and Sen, 1991) and shallow elastic parameters (Rodriguez-Zuniga et al., 1997).

In GAs, a series of "genetic operations" (namely *selection*, *crossover* and *mutation*) acts along various successive steps (*generations*) with the aim of working out a solution able to minimize (or maximize) a certain fitness function that measures how good a certain model is with respect to a desired characteristic. The final solution is a model that shows the best fitness value. This kind of procedure, or at least its basic form, does not provide any evaluation of accuracy or uncertainty of the proposed final solution. This is a crucial issue that

[1] It must be underlined that the term *multi-modality* used in the optimization literature has nothing to do with the concept of *mode* as used in surface wave analysis.

[2] The considered model is later on presented in some detail—see Fig. 2, model #1.

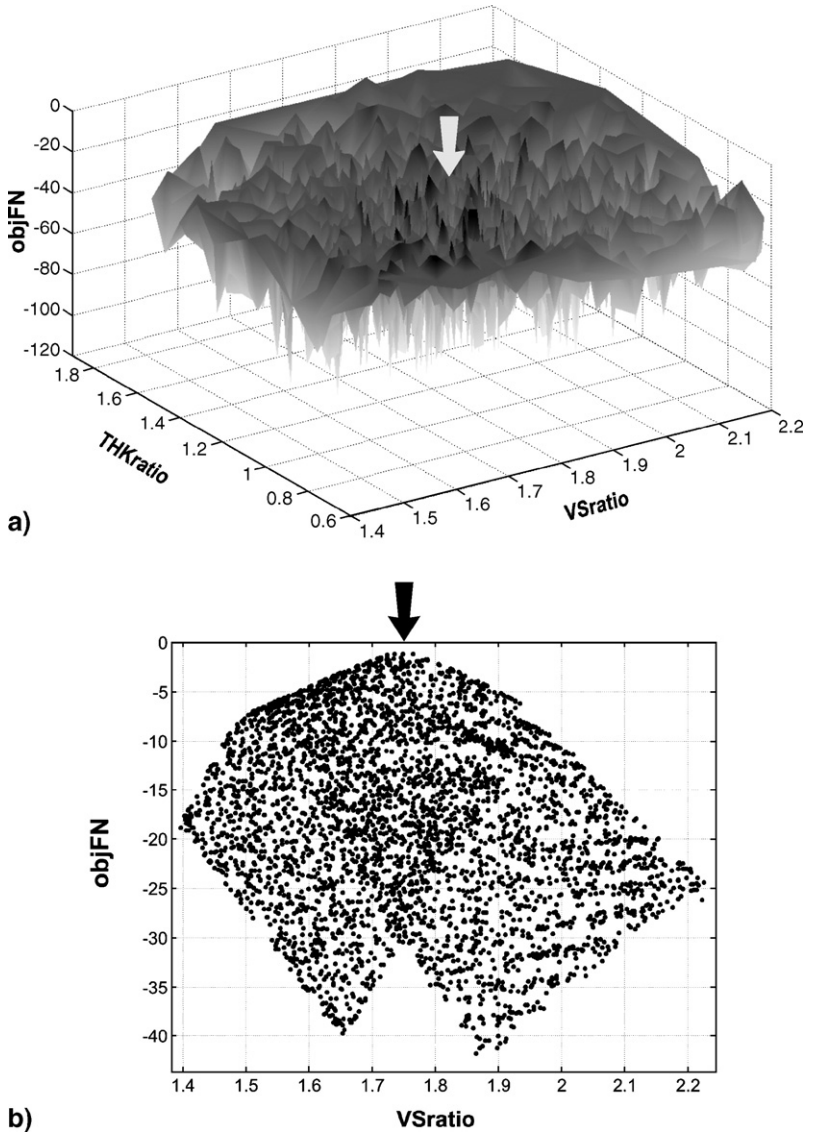

Fig. 1. Multi-modality of the surface wave dispersion curve inversion problem. (a) In abscissa: the ratios between the shear-wave velocities (V_S) of two overlying strata and their thickness (THK), in ordinate: the objective functions for 4000 models. (b) A 2D plot of 4000 objective functions versus the V_S ratio, by keeping constant the THK ratio at the proper value.

we tackle by estimating the *Posterior Probability Density* (PPD).

In our scheme, a Marginal PPD analysis integrates the final model obtained from the GA procedure, with a mean model and the standard deviation of each considered variable.

This approach couples the final solution with an estimate of its statistical relevance which is not provided by the fittest model identified by standard heuristic methods.

We designed three synthetic models based on a similar subsurface structure to check the identification capacity of the proposed scheme as a function of different-scale features: a low-velocity (V_S<400 m/s) surface cover (also including a low-velocity channel—in refraction seismics often referred to as "hidden layer") is then followed by a higher velocity layer (V_S=800) overlying a hard bedrock (V_S=2000 m/s) (see Fig. 2).

Some tests were performed also to evaluate critical aspects of inversion: frequency range of the dispersion curve, number of layers and use of higher modes.

In order to evaluate the results for a real case we eventually considered a data set acquired on a waste

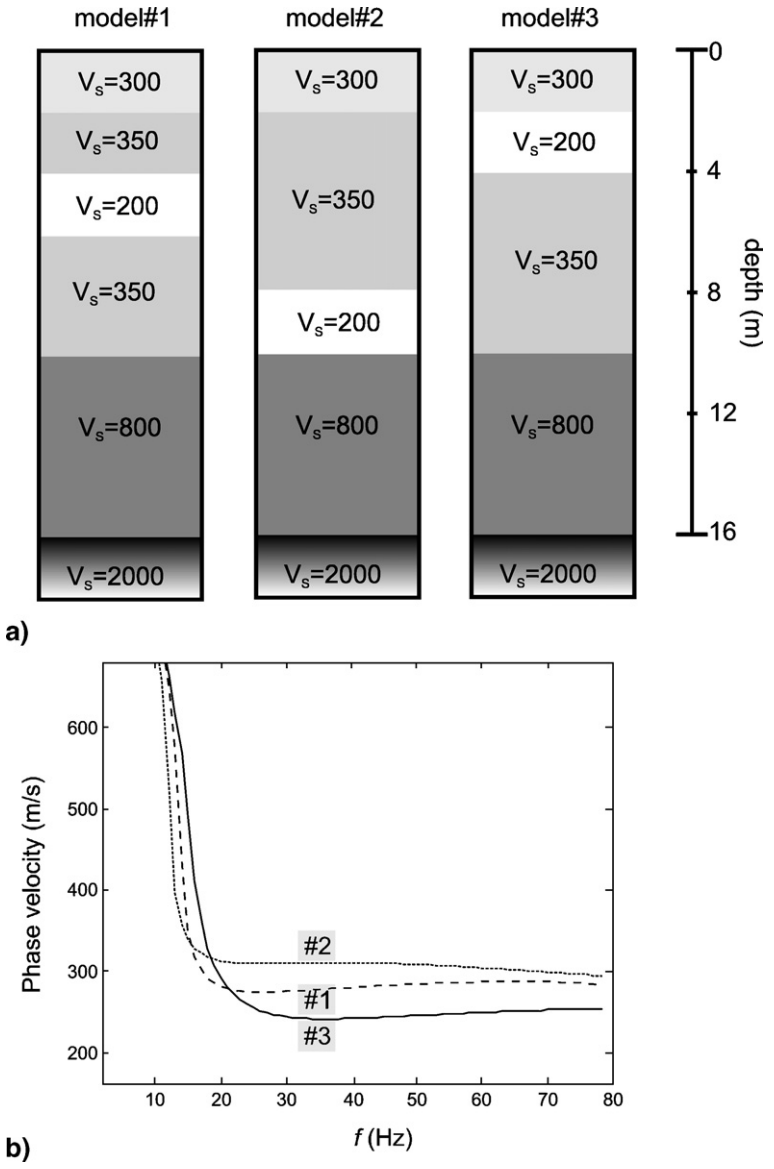

Fig. 2. (a) The three models utilized for the tests (see also Table 1) and (b) their respective dispersion curves.

disposal site in NE Italy (Monfalcone) for which a rich data set is available, also including some boreholes (Dal Moro et al., 2003b).

2. Fundamentals on genetic algorithms

Genetic Algorithms (GAs) have been originally introduced by John Holland and his group at the University of Michigan in the 1970s (Holland, 1975). The fundamental aspect characterizing a genetically-based evolutional scheme is the *Darwinist* paradigm that the fittest survive and reproduce, the others disappear.

Among the several appealing features that characterize GAs some are particularly relevant to solve optimization problems. The main advantage of this class of optimizers is that they tend to elude the attraction of local minima and their *random-but-driven* search schemes try to reach an optimal solution by considering all of the regions of a user-defined search space.

Differently from common linear methods, they do not require an initial model to start the optimization. The user designs a search space within which the possible solutions are searched and evaluated. Moreover, as other heuristic methods, GAs can be applied to problems

where the function to be optimized exhibits discontinuities that would prevent from the use of derivative-based approaches.

An initial *population* composed by an arbitrarily-fixed number of *individuals* (candidate solutions or, to use the genetic lingo, *chromosomes*) is randomly generated and their *fitness* determined according to the discrepancy with respect to a desired characteristic.

This *fitness* value (determined by means of an *objective function*) is then considered in the successive *selection* and *crossover* operations: the fittest individuals (i.e. the ones with the highest fitness values) are chosen to generate offspring whose characteristics are partly taken from one parent and partly from the other. *Elitism* is a strategy often implemented to pass the best individual(s) of each generation unchanged to the next generation in order to avoid possible loss of good individuals. *Mutation* operators allow good genes (that have never appeared before) to be selected and should also ensure that a potentially good component is not lost during reproduction and crossover operations (Goldberg, 1989; Man et al., 2001).

The process can stop after a fixed number of generations or when the *fitness* of an individual reaches a certain previously-fixed value.

3. GAs for surface wave analysis

We implemented a series of *Matlab* tools on the basis of the GAOT (Genetic Algorithms for Optimization Toolbox) optimization routines designed by Houck et al. (1995).

The codes (SWIGA—Surface Wave Inversion via Genetic Algorithms) can perform the inversion of dispersion curves following five different procedures. The user can select the optimum one according to various possible a priori considerations and to the desired degree of freedom of the system. The forward modelling is calculated according to Lai and Rix (1998) as solution of the eigenvalue problem of Rayleigh waves in elastic vertically-heterogeneous media.

A more natural real-valued (floating-point format) formulation of the problem was adopted rather than a binary-encoded one. In fact GAs often prove to be more efficient in the real-valued formulation than in the binary one (Houck et al., 1995; Reeves and Rowe, 2003) and such formulation allows a more straightforward coding. Elitism was adopted in the generation offspring.

In the less constrained case all the four parameters affecting Rayleigh wave propagation (V_S, V_P, density and thickness) are set free.

As the most important parameters affecting Rayleigh wave propagation are shear-wave velocity and layer thickness (e.g. Xia et al., 1999) and with the aim of decreasing the computational load by reducing the number of variables we also adopted two further strategies. For the first one we set the V_P values according to user-defined Poisson's values while for the second one are fixed the densities.

We implemented two additional solutions to cope with further specific situations. A fourth case performs only the inversion of layer thickness by fixing the vertical shear-wave velocity distribution. This can be useful to invert several dispersion curves when geological data suggest that lateral variations are mainly due to geometrical variations (i.e. layer thickness) along the profile rather than to modifications of the elastic properties of the materials.

The last SWIGA procedure handles user-defined relations between shear-wave velocity and density while compressional-wave velocities are fixed on the basis of user-defined Poisson's ratios—consequently, strictly speaking, only V_S and thickness represent variables.

In this paper we focus on the results obtained using this last procedure by deferring a detailed comparison of the results obtained by fixing different constraints to a next communication. Such choice is justified by the fact that layer thickness and shear-wave velocity are by far the most important parameters that influence Rayleigh wave propagation. The number of layers and the range of each variable for each layer must be defined by the user (thus defining the search space).

The key element for any kind of optimization tool is the model *evaluation*, which is performed by means of an *objective function* (objFN) that allows a quantitative estimation of the model (also called *individual* or *candidate solution*).

In the present case we considered the *root-mean-square* value of the difference between the observed and calculated phase velocities (i.e. the dispersion curve):

$$\mathrm{objFN} = -\sqrt{\frac{\sum_{i=1}^{n}\left(v_{\mathrm{obs}_i}-v_{\mathrm{cal}_i}\right)^2}{n}} \qquad (1)$$

where n represents the number of observed frequency–velocity couples, v_{obs_i} the observed phase velocity at the ith frequency and v_{cal_i} the calculated velocity for the considered model (individual of the current population). This kind of formulation is also referred to as the $\ell 2$-norm.

GAOT's search procedure seeks maxima of the objective function and this justifies the negative sign on the right side of Eq. (1).

The inversion scheme we designed takes advantage of the partial results of several preliminary and independent "parallel" runs by selecting the best individuals (models) of all of them as starting population for the final run (Fig. 3 shows the scheme of such algorithm).

The main objectives of such architecture are:

- to obtain a highly-accurate sampling of the search space;
- to avoid problems related to the severe multi-modality of the considered inversion problem;
- to obtain a large number of models to perform a statistical assessment of the solution uncertainty via *Marginal Posterior Probability Density* estimation (see next paragraph).

The number of generations should be kept small for the preliminary parallel runs and increased for the final one.

The selection of the models for the starting population of the final run is determined on the basis of the maximum value of the objective function for each singular preliminary run. By recalling that the objective-function values are negative and the search procedure seeks a maximum, we define the initial population for the final run by selecting all the individuals with an absolute objective-function value higher than n times the best one (where n is a user-defined real number greater than 1).

The peculiarity of such search procedure is justified by the fact that models located in various and distant regions of the search space can, due to the severe multi-modality of the problem, have similar and relatively good fitness values. The preliminary parallel runs aim at identifying promising regions of the search space to be eventually considered in the final-run optimization. In the final inversion step, their genes are genetically processed through *selection* and *crossover* to determine new models that assume the best characteristics of each of them.

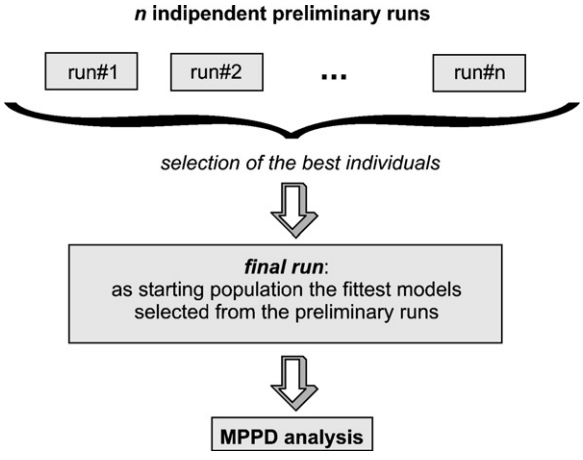

Fig. 3. Code architecture.

We tested different selection, crossover and mutation functions as well as parameter values but, as observed also by Sen and Stoffa (1992), we noticed that non-extreme variations of such functions and parameters scarcely influence the performances of the algorithms and we eventually adopted the values suggested by Houck et al. (1995).

The selection function exploits three probabilistic schemes for parents' selection: roulette wheel, ranking method and tournament selection. Roulette wheel selection method assigns to an individual a selection probability proportional to its fitness while ranking method and tournament selection use the evaluation function to map individuals in an ordered set.

For the crossover operations we considered the following three methods: simple point crossover, arithmetic (complementary linear combinations of the parents) and heuristic (linear extrapolation of the two parents). Simple and arithmetic crossover require as input only the two parents and the limits of each variable, while the heuristic function also needs the number of trials to be performed. As for the latter function, we set a value of 3 so that if after three attempts the offspring fitness value is worse than that of the parents, the offspring is then set equal to the parents.

We adopted the following mutation functions (see Houck et al., 1995 for details): boundary, which changes randomly one of the parameters of the model either to its upper or lower bound; uniform and non-uniform, that change one of the parameters on the basis of a uniform or non-uniform probability distribution respectively; multi-non-uniform, which changes all the model parameters according to a non-uniform probability distribution.

4. Mean model and standard deviations via MPPD

The determination of the statistical meaning and accuracy of the solution of an inversion process in a multi-modal problem is a critical and often under-estimated issue. The approach we explored is based on the determination of the *Posterior Probability Density* (PPD) solution (e.g. Frazer and Basu, 1990; Stoffa and Sen, 1991; Sen and Stoffa, 1992; Gerstoft and Mecklenbrauker, 1998).

The result of an inversion procedure by means of heuristic methodologies is usually simply the fittest

model. No statistical accuracy is then provided and, due to the intrinsic indeterminacy of most of the inversion problems, several different models that exhibit similar fitness values may be equally likely.

We can estimate inversion accuracy by determining mean value and standard deviations of each considered variable through MPPD when a sufficiently large model population is available.

As a general statistical definition, the a posteriori mean model is determined according to the

$$\langle m\rangle = \int \mathrm{d}m m\sigma(m) \tag{2}$$

where $\sigma(m)$ represents the joint posterior probability density defined as

$$\sigma(m) = \frac{\mathrm{e}^{E(m)}}{\sum \mathrm{e}^{E(m)}} \tag{3}$$

and $E(m)$ is the fitness value for each considered model.

The standard deviations of the final model parameters identified via GA procedures are then calculated by considering the square roots of the diagonal terms of the covariance matrix $\mathbf{C}_\mathrm{M}$ determined by

$$\mathbf{C}_\mathrm{M} = \int \mathrm{d}m(m-\langle m\rangle)(m-\langle m\rangle)^T\sigma(m) \tag{4}$$

The MPPD for the ith parameter is then determined according to (e.g. Gerstoft and Mecklenbrauker, 1998):

$$\sigma^i(m^i) = \int \sigma(m)\mathrm{d}m^1 \ldots \mathrm{d}m^{i-1}\mathrm{d}m^{i+1}\ldots \mathrm{d}m^M \tag{5}$$

where $\sigma(m)$ represents the joint posterior probability density previously evaluated and M the number of model parameters.

The characteristics of the most appropriate population to use to perform these calculations will be discussed later on in this paper.

MPPD computation is a simple and powerful tool that provides valuable insight into the solution uncertainty. Nonetheless, in order to be statistically meaningful, it requires a large number of data (i.e. models).

5. Inversion results

5.1. Synthetic data

To study the performance of the proposed inversion scheme, we used a synthetic model simulating (from top to bottom) soft low-velocity sediments and more compact sediments (characterized by higher velocities) lying over hard-rock basement. In the upper part we also introduced a velocity inversion.

We considered only the fundamental mode, because higher modes are rarely identifiable when such a soft unconsolidated cover is present.

In the performed inversions, shear-wave velocity and thickness are the independent variables because they are the most important parameters that determine Rayleigh wave propagation (Xia et al., 1999). Such choice improves computation time and efficiency because the number of variables is reduced: densities and compressional-wave velocities are fixed on the basis of user-defined V_S–ρ relation and Poisson's ratios.

To test the performance of the method in more realistic conditions (in which the user does not know the real relationships), in the inversion of the synthetic curves we intentionally introduced some errors in the density and compressional-wave velocities. Poisson's ratios were fixed to values slightly different from those actually used in the calculation of the synthetic dispersion curves.

We adopted a simple rule and fixed a Poisson's value of 0.35 when V_S is lower than 1500 m/s and 0.25 for higher velocities (the bottom bedrock half-space), thus introducing an average error of approximately 7% (compare values in Table 1).

A V_S–ρ relation ($\rho = 0.77 * \log_{10}(V_\mathrm{S}) + 0.15$) was used, which introduces a further error up to 12%. This was derived from the V_P–ρ relation proposed by Gardner et al. (1974).

A crucial issue in dispersion curve inversion is the number of layers. If no additional data are available, subsurface layering is basically unknown and its reconstruction is the main goal of any non-invasive geophysical investigation. The number of layers can be in principle considered an additional variable in our optimization problem but this would determine a

Table 1
Parameters of the three models utilized for the tests (V_S is the shear-wave velocity (m/s), ρ is the density (g/cm^3), THK is the thickness (m))

Layer	Model#1			Model#2			Model#3		
	V_S	ρ	THK	V_S	ρ	THK	V_S	ρ	THK
1	300	1.8	2	300	1.8	2	300	1.8	2
2	350	1.9	2	350	1.9	6	200	1.7	2
3	200	1.7	2	200	1.7	2	350	1.9	6
4	350	1.9	4	800	2.1	6	800	2.1	6
5	800	2.1	6	2000	2.4	∞	2000	2.4	∞
6	2000	2.4	∞						

V_P values are defined according to Poisson's ratio values equal to 0.4 when $V_\mathrm{S} \leq 400$, 0.3 for $V_\mathrm{S} = 800$ and 0.25 for $V_\mathrm{S} = 2000$.

dramatic increase in the computation time. An arbitrary and relatively large number of strata (i.e. an *over-parameterization* of the problem) is then probably a reasonable solution and a good compromise between computation effort and effectiveness of the inversion procedure and results. Therefore we basically adopted a 10-layer structure with thickness and velocity boundaries (search space) reported in Table 2, even if the synthetic dispersion curves to invert were calculated with an Earth model consisting of 5 or 6 layers (see Fig. 2 and Table 1). If not otherwise specified we considered dispersion curves with frequencies up to 80 Hz.

Nine preliminary parallel runs, with a population of 7000 each, were considered to generate the models that were successively passed to the final inversion step. The number of generations for the preliminary runs was set to 10 while for the final one to 250.

It can be noticed that the number of models necessary to cover the entire search space in an enumerative search scheme would be larger than $6*10^{15}$, in case velocity and thickness increments of 10 m/s and 10 cm are considered. Therefore almost eleven orders of magnitude separate the enumerative case from the total number of models actually considered for the inversions here presented (less than 70000). Such a rough assessment offers an idea about the performances of a GA-based optimization scheme.

The initial population for the final inversion was determined by considering all the individuals with an objective-function value greater than five times the best one for each preliminary parallel run. For instance, if the best individual of a preliminary parallel run has an objective value equal to −10 all the models characterized by an objective-function value greater than −50 are passed to the initial population of the final run. Fig. 4 shows an example of this kind of model selection. All the models with an objective function greater than −30 (highlighted in the light-grey box) are selected for the final-run initial population because the best model for this preliminary run obtained an objective function value of about −6. The procedure is repeated for all the preliminary runs.

Table 2
Search space for the inversion in the 10-layer case (V_S in m/s; THK in m)

Layer	Adopted boundaries	
	Max–Min V_S	Max–Min THK
1	400–150	3–0.5
2	400–150	3–0.5
3	400–150	3–0.5
4	400–150	3–0.5
5	400–150	3–0.5
6	400–150	3–0.5
7	900–200	5–1
8	1000–300	5–1
9	3000–1000	5–1
10	3000–1500	Half-space

The basic idea for this procedure is that heuristic methods provide surely good but not necessarily optimal solutions in case the problem is particularly complex. In the present case for example (in which we basically used a 10-layer model and only thickness and shear-wave velocity are considered as variables), it can be noticed that the number of resulting variables is equal to 19. As previously stated, preliminary runs aim at identifying promising regions in the search space. The models containing such genes are passed to the final run for a further and computationally-intensive genetic selection.

Fig. 5 highlights how the different preliminary runs identify different promising regions for the fourth-layer shear-wave velocity. It is worth underling that the classes with higher frequencies are characterized by similar objective functions (approximately ranging from −11 up to −7) in spite of the fact that they span a large area of the user-defined search space for that variable. This is evidence that the solutions provided by heuristic methods still contain an intrinsic indeterminacy that can be properly handled only by means of a statistical analysis.

Moreover, the entire search space is covered so that any velocity value can still be evaluated during the final run even if belonging to models that were labelled as not particularly fit during the preliminary runs.

The models selected from these preliminary runs are then used as starting population for the final run in which the number of generations is now much higher (10 for the preliminary runs and 250 for the final one—see Fig. 6).

Fig. 7a shows the inversion results for model#1. We can notice that the general velocity trend is properly identified and allows identifying the two deepest velocity discontinuities. Moreover, the MPPD-derived mean model appears more significant than the fittest one, being typically closer to the real model (see also inversions presented later on).

To evaluate further aspects involved in the dispersion curve inversion, we performed three further tests over the same model. In the first one, we included the first higher mode and tested the results to evaluate whether higher modes better constrain the inversion process. Inversion was performed with the same parameters previously adopted (see Table 2). Fig. 7b shows the shear-wave vertical profile thus obtained.

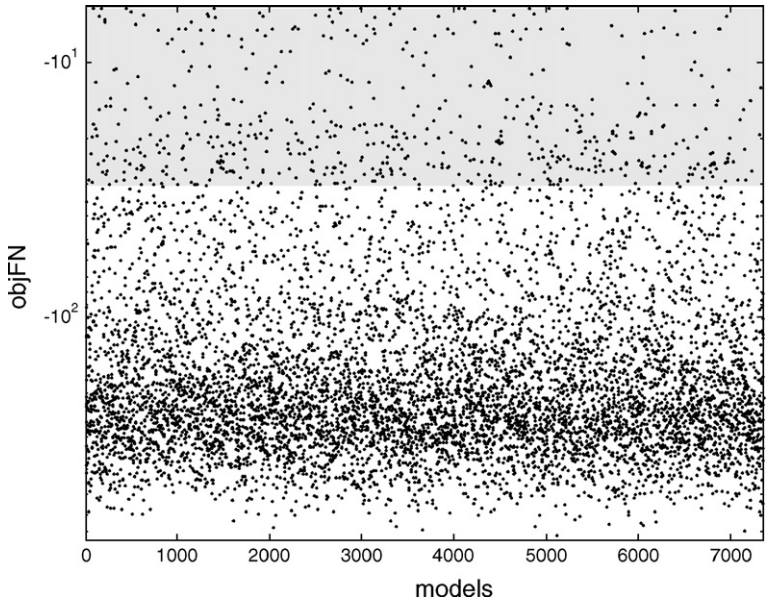

Fig. 4. Model #1, first preliminary run: objective functions for 7357 models, partly (7000) present in the initial population, partly (357) generated in the successive 10 generations. Models in the light-grey box are selected for the final-run initial population. Y-axis in logarithmic scale.

The comparison with the results obtained by considering the fundamental mode only (see Fig. 7a) gives evidence of an improved focusing of the inversion with smaller standard deviations that indicate a better-constrained solution.

We evaluated the possible improvement of the inversion results in a further test that was performed by considering the same number of layers (six) actually used to generate the synthetic dispersion curve (model#1 in Fig. 2 and Table 1). Table 3 shows the

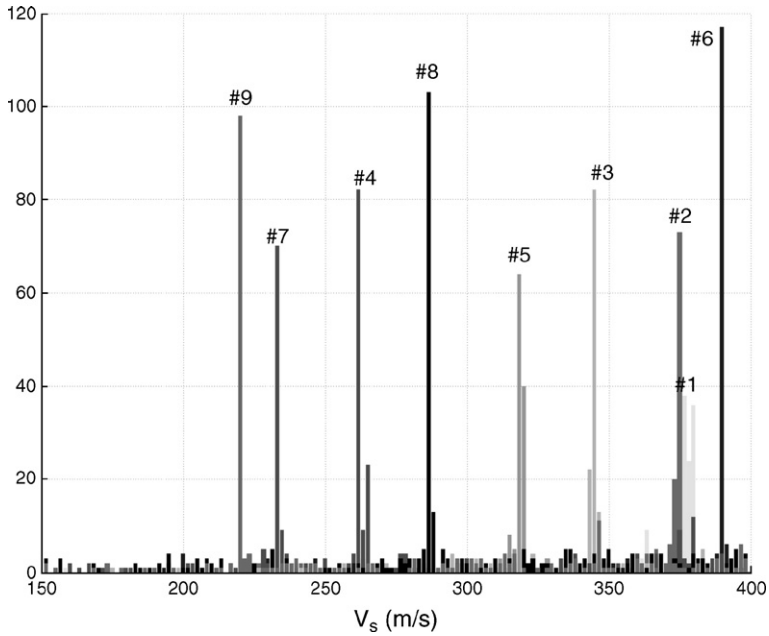

Fig. 5. Histogram of the values for the seventh variable (fourth-layer shear-wave velocity). Bin dimension approximately equal to 2 m/s, average population size for the nine indicated populations equal to 229 models.

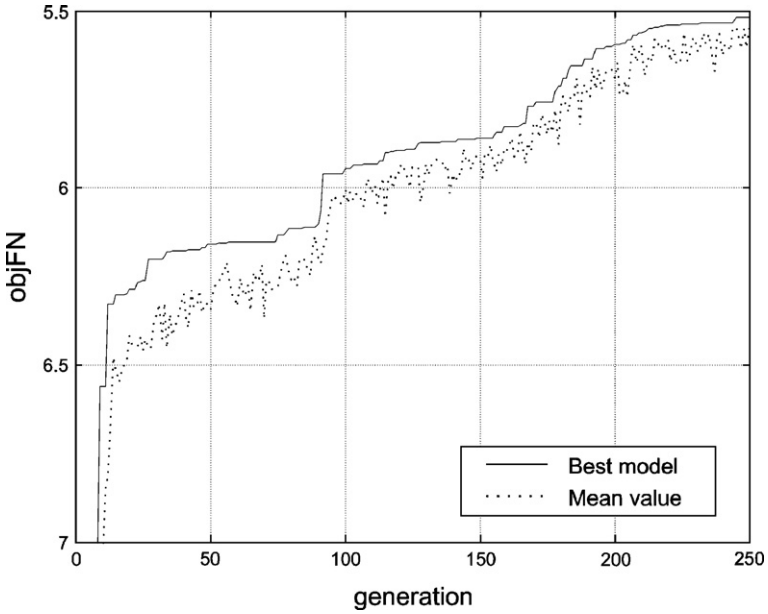

Fig. 6. Fitness values over the passing generations.

boundary conditions for the inversion. Fig. 7c shows the results to compare with those obtained from the 10-layer model reported in Fig. 7a.

As easily predictable, the final results in the 6-layer case indicate a much better focused solution than in the 10-layer, one because of the lower degrees of freedom of the system.

We performed another test by considering a limited frequency range. In fact, in real-world seismics the frequency spectrum resulting from the application of a low-frequency source (as for instance a sledgehammer) over unconsolidated sediments often suffers from lack of high frequencies.

Fig. 7d shows the shear-wave vertical profile obtained when only frequencies lower than 40 Hz are used to a 10-layer structure. As it can be noticed, mean model and standard deviations indicate that the solution does not significantly differ from the one obtained by using a larger spectrum (80 Hz; Fig. 7a) and, in both cases, the mean model appears somehow more meaningful than the purely GA-based solution (fittest model). However values in the uppermost portion of the sequence appear somehow less precise that in the previous case (compare with Fig. 7a).

We remark that if we consider a V_S equal to 300 m/s (uppermost layer) and a 40 Hz component, the corresponding wavelength is 7.5 m. According to the $\lambda/2$ rule of thumb (e.g. Xia et al., 2004) this component is able to sense depths of approximately 4 m and consequently it would be rather inadequate to discriminate the first layer from the second (see model parameters in Table 1).

Two further tests were performed in order to evaluate the identification power of the method. The models are variations of the model#1, with the low-velocity channel placed at different depths (see Fig. 2a, models #2 and #3).

As for model#1 depth and velocity of the bedrock are identified with good precision (see Fig. 8a and b) and the low-velocity channel is also identified with reasonable accuracy.

It is worth stressing that the resulting mean model is not a genetically-created model. Such a model is not the result of any genetic event but the product of the MPPD analysis that comparatively evaluated each singular variable with respect to the rest of them on the basis of the fitness of each model.

A key issue in MPPD calculation is the population of models to use for such operation. Three options are actually possible based on the designed inversion scheme: (a) to consider only the models randomly generated for the initial populations of the preliminary runs, (b) to consider all the models generated and evaluated in the preliminary runs (the initial populations plus the models generated during the ten generations adopted for the present case) or (c) to collect the entire set of models generated in the whole process, thus including also the models generated during the 250 generations of the final run.

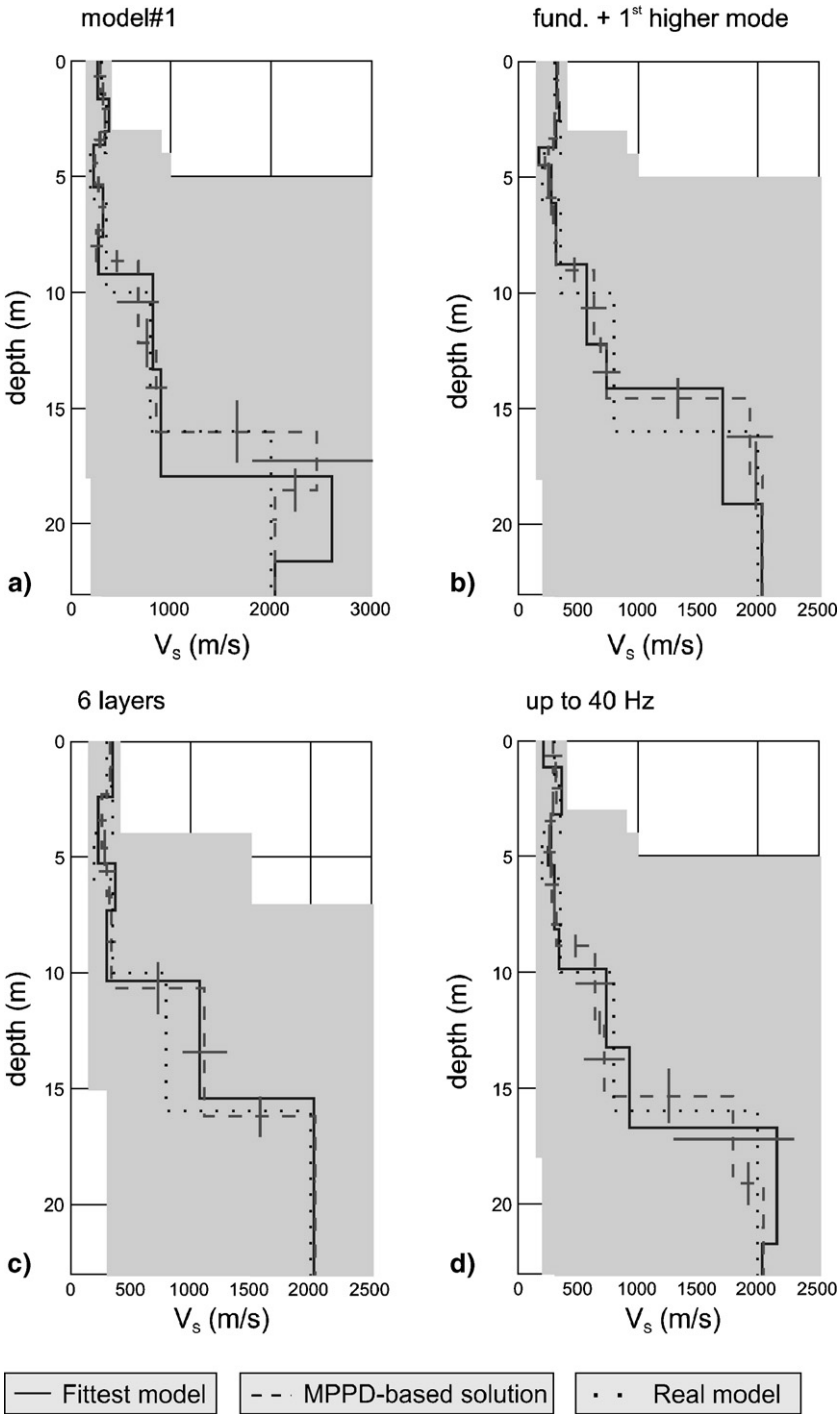

Fig. 7. Final solutions (fittest model and MPPD-defined mean model) for model#1 when (a) only fundamental mode is considered, (b) fundamental and first higher mode are considered, (c) inversion is performed by considering a 6-layer model (instead than a 10-layer one) and (d) dispersion curve is cut to 40 Hz (instead of 80). The background grey area represents the search space.

The decision about the population to be considered must be clearly taken by considering the computational and statistical consequences (see also Stoffa and Sen, 1991). It is apparent that the models of the preliminary initial populations (case a) can provide a somehow poor estimate due to the fact that most of such randomly-

Table 3
Search space for the inversion of model #1 in the 6-layer case (V_S in m/s; THK in m)

Layer	Adopted boundaries	
	Max–Min V_S	Max–Min THK
1	400–150	3–0.5
2	400–150	3–0.5
3	400–150	3–1
4	400–150	6–2
5	1500–300	8–3
6	3000–1500	Half-space

generated models are clearly "weak" (i.e. often extremely far from any good model and then characterized by very low fitness values). As a consequence, the estimated mean model is quite unstable and the standard deviations are large (see Fig. 9a). On the other extreme side (case c), a MPPD calculation performed by considering the entire set of generated models easily leads to a final mean model very close to the one genetically determined (best fitness) and to very small standard deviations (Fig. 9c). This is easily understood when we consider that, during the final run, the offspring converge towards a certain solution. Such model will strongly bias the MPPD analysis due to its high fitness values.

We then decided to select a population constituted of the models generated for the initial populations of the preliminary runs plus their 10-generation offspring models (case b). This is a kind of compromise between the two extremes that avoids the indeterminacy of the first case and the strongly-biased and exceedingly-constrained solution of the second (Fig. 9b). In this respect it is also worth noticing the even distribution of models in Fig. 5.

To further clarify this aspect Fig. 10 shows a close up of the MPPD values just for the shear-wave velocity of the seventh layer for all of these three possible choices (example is taken from the model#1 inversion). If we consider only the purely-random models of the preliminary-run initial populations the MPPD values are very dispersed and the standard deviation accordingly large (Fig. 10a). In the opposite case (the complete set of models evaluated in the entire GA-based procedure), a concentration of solutions occurs around a specific value (very close to the final genetic fittest solution) thus determining an extremely small standard deviation because of an extremely-biased population (Fig. 10c).

We can further analyse the problem by considering the histograms of the objective-function distribution for the three possible cases (Fig. 11). The purely-random

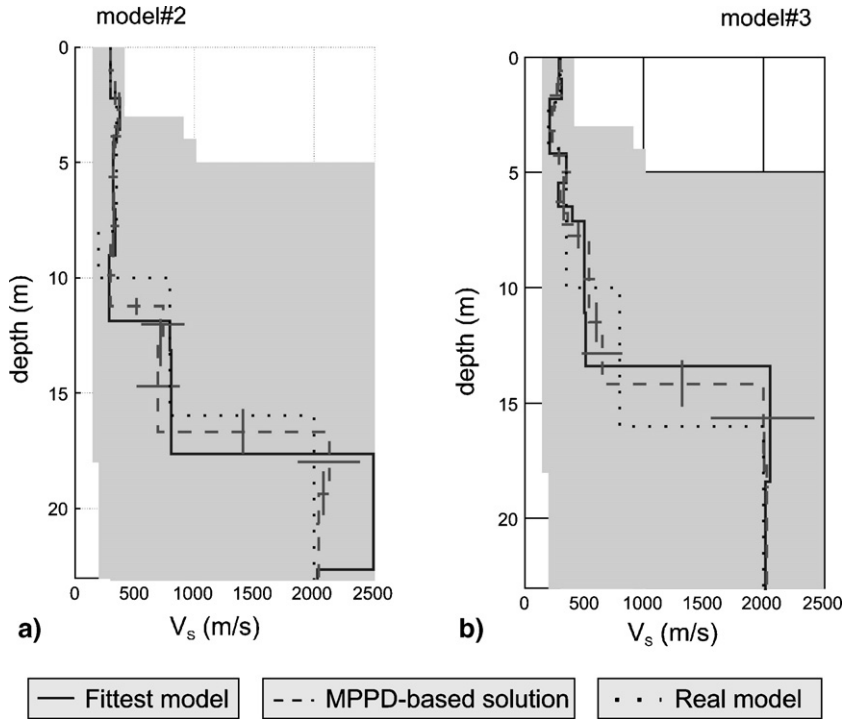

Fig. 8. Final solutions (fittest model and MPPD-defined mean model) for models #2 and #3. The background grey area represents the search space.

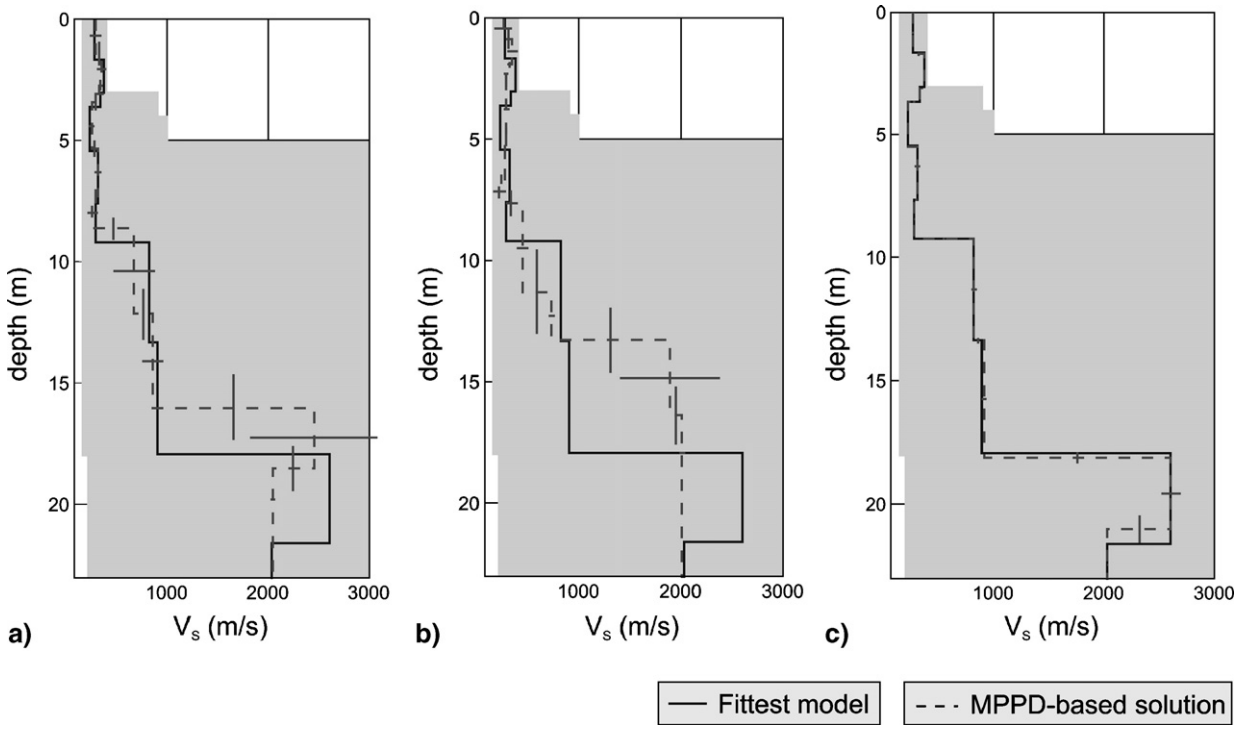

Fig. 9. Model#1 inversion. Mean model and standard deviations calculated according to MPPD evaluation performed by considering: (a) the models generated for the initial populations of the preliminary runs; (b) all the models evaluated during the preliminary runs; (c) all the models of the entire GA-based procedure (see text).

case (Fig. 11a) shows a distribution very close to the Gaussian while the high narrow peak determined by the complete set of evaluated models (Fig. 11c) puts in evidence a strongly-biased population.

The small peak (Fig. 11b) determined when considering all the models generated during the preliminary runs (then also considering the 10-generation offspring) must be regarded as evidence of a meaningful set of models for the MPPD evaluation, whose favourable characteristics are highlighted by the data shown in Figs. 9b and 10b.

The population was checked and redundant models deleted to avoid the biasing effect of duplicate models in the MPPD calculation.

5.2. Field data

The designed algorithm was used to invert a data set acquired in a waste disposal site in NE Italy. This is essentially characterized by an 18-m-thick unconsolidated-sediment sequence lying over a limestone basement. A number of geophysical surveys (surface Ground-Penetrating Radar—GPR, borehole Vertical Radar Profiling—VRP, resistivity, magnetometry) were conducted (Dal Moro et al., 2003b) and results were compared with borehole data and measurements on samples with the goal of determining the identification power of each methodology and classifying waste typology and extension to plan future remediation acts.

Because of the poor geotechnical characteristics of the uppermost layers and in particular the extremely heterogeneous and unconsolidated superficial waste level (about 2 m thick), seismic records present a limited spectral content and, as far it concerns the P-wave component, a low signal-to-noise ratio strongly dominated by ground roll components.

An example of velocity spectrum computed according to the *phase shift* method (Park et al., 1998; Dal Moro et al., 2003a) is reported in Fig. 12a. As the small differences among the dispersion curves acquired along the acquisition profile are more evident at the high frequencies, they are mostly due to lateral heterogeneities of the uppermost layer evidenced also by the GPR and resistivity surveys (Dal Moro et al., 2003b).

Similarly to the strategy adopted for the inversion of the synthetic data, we considered as variables only V_S and thickness while fixing ρ and V_P according to the V_S–ρ relation mentioned in the previous section and the Poisson's ratios reported in Table 4. These latter were

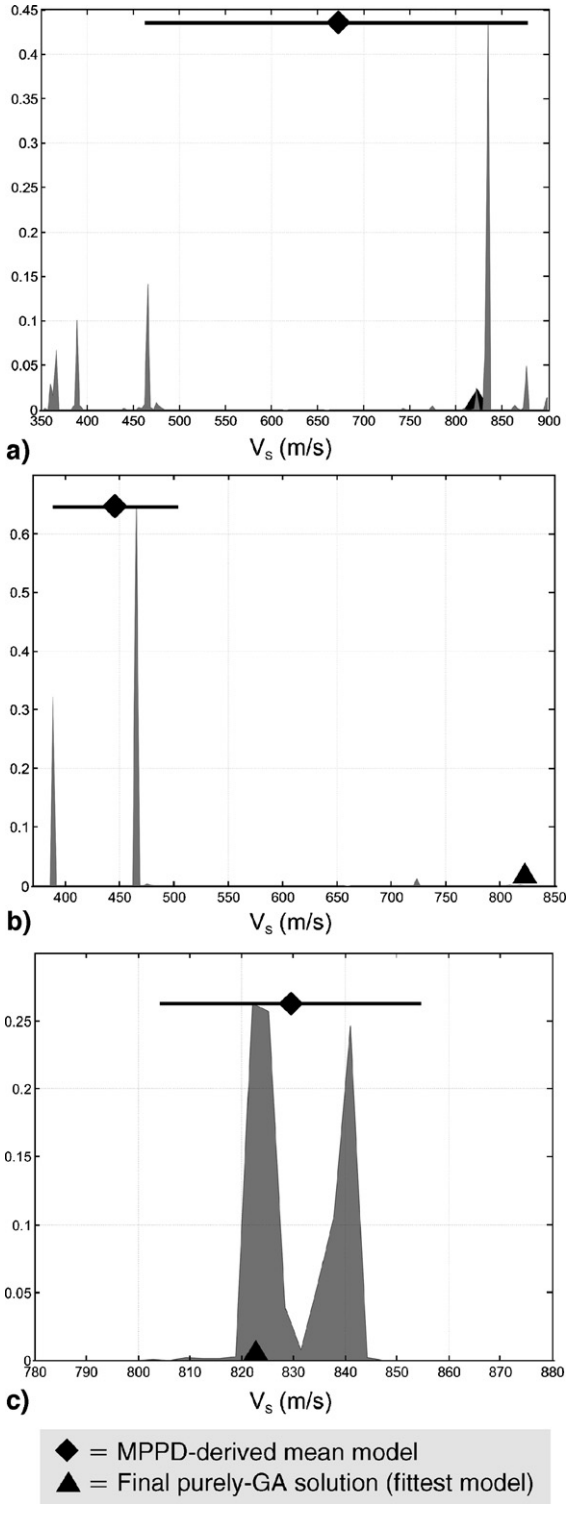

Fig. 10. Model#1 inversion. MPPD distribution: shear-wave velocity values for the seventh layer (same cases as for Fig. 9) (diamond: MPPD-derived mean model; triangle: final purely GA solution—fittest model). Different scales are used to highlight the data.

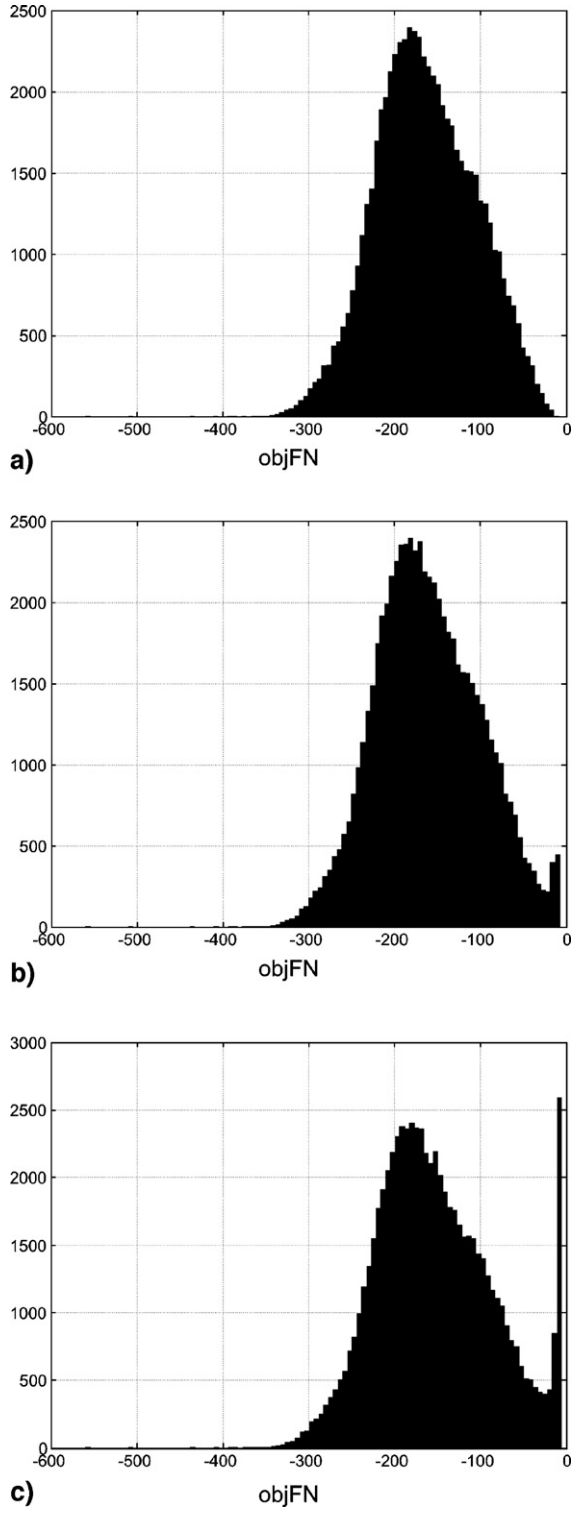

Fig. 11. Model#1 inversion: histograms of the objective functions (same cases as for Figs. 9 and 10).

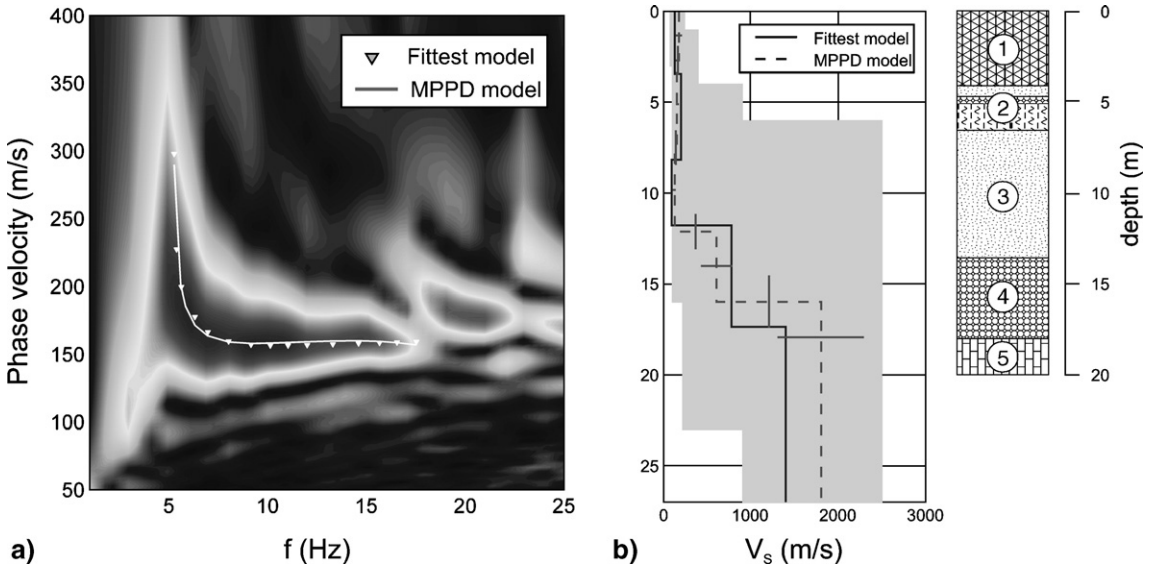

Fig. 12. Field dataset: (a) observed velocity spectrum and retrieved dispersion curves; (b) vertical shear-wave velocity profiles from dispersion curve inversion; the background grey area represents the search space (see Table 4). Also shown is the borehole stratigraphy: (1) waste, (2) mixed sand and clay, (3) sandy silt, (4) gravel, (5) fractured limestone.

set according to values commonly adopted for the relevant stratigraphic units (0.45–0.4 for variously silty alluvia, 0.25 for gravel and 0.2 for the limestone bedrock) (e.g. Ivanov et al., 2000; Adme, 2004).

The results of the performed inversion (Fig. 12b) seem to be characterized by a certain degree of uncertainty in particular for the deepest part of the vertical profile (see standard deviations associated with the retrieved shear-wave velocities) very likely due to the limited frequency range of the dispersion curves (approximately 4–19 Hz). In spite of this, the best and the MPPD-defined models still appear furnishing a meaningful solution that appears in fairly-good agreement with the borehole stratigraphy also shown in Fig. 12. In fact it can be noticed that the depths of the two deepest velocity discontinuities result coherent with the borehole data and can then be interpreted as the effect of the silt–gravel and gravel–limestone contacts (see also Dal Moro et al., 2003b).

Table 4
Field dataset: search space and Poisson's values adopted for the dispersion curve inversion

Layer	Max–Min V_S	Max–Min THK	Poisson's ratio
1	250–80	4–1	0.45
2	400–100	6–1	0.45
3	400–100	6–2	0.40
4	900–220	7–2	0.25
5	2500–900	Half-space	0.20

6. Conclusions

Genetic algorithms, as members of the class of global-search optimization schemes, represent an appealing inversion tool because generally little sensitive to local minima/maxima and therefore particularly suitable for non-linear multi-modal problems.

Similarly to the other heuristic methods, GAs adopt a user-defined search space within which solutions are sought and evaluated. This means that, unlike linear methods, they do not require any starting model that would risk to be attracted by some local minimum (or maximum) thus eventually furnishing an erroneous solution.

The wide search space boundaries here adopted simulate conditions in which no a priori knowledge is available: the number of layers is large and velocity and thickness boundaries are kept remarkably broad. The general trend in which velocity increases with depth can be considered as representative of a very general velocity distribution based on the shape of the considered dispersion curve, on the V_S–V_R relationship (e.g. Viktorov, 1967) and on the $\lambda/2$ rule of thumb (e.g. Xia et al., 2004).

The proposed approach (composed of several preliminary "parallel" runs and a final one) can be considered as an improvement of the GA-based inversion scheme with the following extensions: the identification of a mean model and the computation of standard deviations for each considered variable.

Relevant aspects to consider for the performance analysis are the identification of the bedrock and the contact between the uppermost low-velocity cover (V_S<400 m/s) and a deeper higher velocity layer (V_S=800 m/s).

We noticed that in all the examined cases the bedrock depth and velocity are identified with good precision. This represents a first important achievement because it demonstrates an identification ability that cannot usually be equalled by common linear inversion procedures.

The contact between the uppermost low-velocity part (V_S<400 m/s) and the higher velocity one (V_S=800 m/s) was also properly imaged in all the cases. The identification of the velocity inversion (located at different depths in the different models) seems more challenging. The shallow low-velocity channel was identified by all the solutions obtained from the dispersion curve inversion but its depth and thickness are not always precise probably due to an intrinsic indeterminacy of the method (Klimeš, 2004). The reliability of the retrieved model must be then regarded as a key point and efforts should be made in order to design survey and inversion techniques to eventually exploit different wavefield components (reflection, refraction, P and S waves) in a cooperative inversion scheme (Dal Moro and Pipan, 2006-this issue).

We also investigated the effect of two velocity spectrum features to analyse their effects on the accuracy of the final solution: the inclusion of the first higher mode and a frequency band limitation.

When the first higher mode is considered together with the fundamental one the results give evidence of a reduction of standard deviations and a correlated better-focused solution. As for the bandwidth, for the considered case we did not notice a significant improvement in the result by doubling the frequency range from 40 to 80 Hz.

An apparent improvement in the final model occurs when the number of layers specified for the inversion process matches the real one (i.e. the one actually used in the synthetic curve computation). In the present case (we considered the correct 6-layer model instead of the 10-layer one), this is most likely due to the fact that the number of variables is considerably reduced (from nineteen to eleven) and the optimization procedure can better handle their evaluation.

In several cases we observed that the mean model derived from the MPPD analysis is better than the purely genetic one (the fittest model). This shows that the extreme complexity of the problem (its multi-modality and the number of variables) requires a statistical approach to provide meaningful and robust solutions and uncertainty evaluation, being that even global-search methods (like GAs) do not necessarily ensure optimal solutions.

A detailed comparative analysis of role and effects of different inversion strategies, initial and final-run population size, preliminary-run number and genetic operators is in progress and will be the subject of a future communication.

Results of the inversion performed on a field dataset acquired in Monfalcone (NE Italy) were presented and validated by borehole stratigraphy.

Acknowledgements

This research was supported by CNR (Italy), National Group for defence against chemical, industrial and ecologic hazards, Grant No. 00.00623.PF37 and by the European contract EVK4-CT-2001-00046, HYGE-IA. Part of the present research was performed during a stay of one of the authors (Giancarlo Dal Moro) at the Department of Geophysics, Charles University, Prague (Czech Republic), in the framework of the European contract MAGMA (European contract EVG3-CT-2002-80006). Authors want to express their gratitude to Ludìk Klimeš (from this latter Department) for his valuable suggestions regarding the way to approach the accuracy estimation and to Peter Gerstoft for his kindness and willingness during a brief e-mail exchange.

References

Adme, Z.G., 2004. Analysis of NATM tunnel responses due to earthquake loading in various soils, REUJAT (Research Experiences for undergraduates in Japan in Advanced Technology) 2004, open file (http://wusceel.cive.wustl.edu/reujat/2004/adme.pdf).

Dal Moro, G., Pipan, M., 2006. Joint inversion of surface wave dispersion curves and reflection travel times via multi-objective evolutionary algorithms. Journal of Applied Geophysics 61, 56–81 doi:10.1016/j.jappgeo.2006.04.001 (this issue).

Dal Moro, G., Pipan, M., Forte, E., Finetti, I., 2003. Determination of Rayleigh wave dispersion curves for near surface applications in unconsolidated sediments. In: Proceedings SEG (Society of Exploration Geophysicists) 2003, 73rd Annual Meeting, Dallas, Texas, October 26–31, 2003, pp. 1247–1250.

Dal Moro, G., Pipan, M., Forte, E., Sugan, M., Finetti, I., 2003. Integrated non-invasive characterization of waste disposal site. In: Proceedings Symposium EEGS 2003 (Environmental and Engineer Geophysical Society), Prague (Czech Republic), August 31–September 4, 2003, O-085.

Dal Moro, G., Forte, E., Pipan, M., Sugan, M., in press. Velocity spectra and seismic signal identification for surface wave analysis. Near-Surface Geophysics.

Frazer, L.N., Basu, A., 1990. Freeze bath inversion. In: Proceedings SEG (Society of Exploration Geophysicists) 1990, 60th Annual Meeting, San Francisco, CA, September 23–27, 1990, pp. 1123–1125.

Gardner, G.H.F., Gardner, L.W., Gregory, A.R., 1974. Formation velocity and density—the diagnostic basic for stratigraphic trap. Geophysics 39, 770–780.

Gerstoft, P., Mecklenbrauker, C.F., 1998. Ocean acoustic inversion with estimation of a posteriori probability distributions. Journal of the Acoustical Society of America 104, 808–819.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Company Inc.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor.

Houck, C.R., Joines, J.A., Kay, M.G., 1995. A genetic algorithm for function optimization: a Matlab implementation, open file (http://www.ie.ncsu.edu/mirage/GAToolBox/gaot/).

Ivanov, J., Park, C.B., Miller, R.D., Xia, J., 2000. Mapping Poisson's ratio of unconsolidated materials from a joint analysis of surface-wave and refraction events. In: Proceedings of the Symposium on the Application of Geophysics to Engineering and Environmental Problems (SAGEEP 2000), Arlington, Va., February 20–24, pp. 11–19.

Klimeš, L., personal communication. Tests on linear inversions of surface wave dispersion curves.

Lai, C.G., Rix, G.J., 1998. Simultaneous inversion of Rayleigh phase velocity and attenuation for near-surface site characterization. Georgia Institute of Technology, School of Civil and Environmental Engineering, Report No.GIT-CEE/GEO-98-2, July 1998, 258 pp.

Louis, S.J., Chen, Q., Pullammanappallil, S., 1999. Seismic velocity inversion with genetic algorithms. CEC99, 1999 Congress on Evolutionary Computation, Mayflower Hotel, Washington D.C., July 6–9, 1998, pp. 855–861.

Man, K.F., Tang, K., Kwong, S., 2001. Genetic Algorithms. Sringer.

Park, C.B., 2002. Multichannel analysis of surface waves (MASW). MASW Workshop Notes, open file (http://www.terrajp.co.jp/MASW_Workshop_Tokyo.pdf).

Park, C.B., Xia, J., Miller, R.D., 1998. Imaging dispersion curves of surface waves on multichannel record. In: Proceedings SEG (Society of Exploration Geophysicists) 2003, 68th Annual Meeting, New Orleans, Louisiana, September 13–18, 1998, pp. 1377–1380.

Park, C.B., Miller, R.D., Xia, J., 1999. Multichannel analysis of surface waves. Geophysics 64, 800–808.

Reeves, C.R., Rowe, J.E., 2003. Genetic Algorithms—Principles and Perspectives. Kluwer, Norwell, Kluwer Academic Publisher.

Rodriguez-Zuniga, J.L., Ortiz-Aleman, C., Padilla, G., Gaulon, R., 1997. Application of genetic algorithms to constrain shallow elastic parameters using "in situ" ground inclination measurements. Soil Dynamics and Earthquake Engineering 16, 223–234.

Sambridge, M., Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems. Reviews of Geophysics 40, 3.1–3.29.

Sen, M.K., Stoffa, P.L., 1992. Rapid sampling of model space using genetic algorithms: examples from seismic waveform inversion. Geophysical Journal International 108, 281–292.

Smith, M.L., Scales, J.A., Fischer, T.L., 1992. Global search and genetic algorithms. Geophysics: The Leading Edge of Exploration 11, 22–26.

Stoffa, P.L., Sen, M.K., 1991. Nonlinear multiparameter optimisation using genetic algorithms: inversion of plane wave seismograms. Geophysics 56, 1794–1810.

Viktorov, I.A., 1967. Rayleigh and Lamb waves: Physical Theory and Applications. Plenum Press, New York.

Xia, J., Miller, R.D., Park, C.B., 1999. Estimation of near-surface shear-wave velocity by inversion of Rayleigh waves. Geophysics 64, 691–700.

Xia, J., Miller, R.D., Park, C.B., Ivanov, J., Tian, G., Chen, C., 2004. Utilization of high-frequency Rayleigh waves in near-surface geophysics. The Leading Edge 23 (8), 753–759.