[COR: Number]

Cover Letter

formikoj: A flexible library for data management and processing in geophysics - Application for seismic refraction data

Matthias Steiner,Adrián Flores Orozco

Dear Editors-in-Chief,

please find the enclosed manuscript "formikoj: A flexible library for data management and processing in geophysics - Application for seismic refraction data" which we are submitting for exclusive consideration for publication in Computers & Geosciences. We confirm that the submission follows all the requirements and includes all the items of the submission checklist.

The manuscript presents the open-source python library formikoj for managing and processing geophysical data collected in environmental and engineering investigations. The library was specifically implemented for multi-platform usage to allow the efficient collaboration and exchange of data between different partners in research projects and academia. In this regard, we believe that this library aids in providing reproducible data and results as well as establishing and maintaining good research practices. Hence, we believe that this manuscript is of relevance to the readership of Computers & Geosciences.

We provide the source codes in a public repository with details listed in the section "Code availability".

We look forward to your decision.

Yours sincerely,

Matthias Steiner and Adrián Flores Orozco
Research Unit Geophysics, Department of Geodesy and Geoinformation, TU Wien; matthias.steiner@geo.tuwien.ac.at

Highlights

formikoj: A flexible library for data management and processing in geophysics - Application for seismic refraction data

Matthias Steiner,Adrián Flores Orozco

- flexible open-source and cross-platform library for managing and processing of geophysical data
- possibility to be deployed for different geophysical methods and/or instruments
- application for the modeling and processing of seismic refraction datasets
- applicable for seismic refraction data collected in 2D and 3D survey geometries
- easily scalable for custom requirements

formikoj: A flexible library for data management and processing in geophysics - Application for seismic refraction data

Matthias Steiner[a,*], Adrián Flores Orozco[a]

[a]Research Unit Geophysics, Department of Geodesy and Geoinformation, TU Wien

ARTICLE INFO

Keywords:
geophysical data processing
seismic refraction
first break picking
seismic waveform modeling
cross-platform application
geophysical python library
flexible open-source librarys
wave based methods

ABSTRACT

We introduce here the open-source library formikoj, which provides a flexible framework for managing and processing geophysical data collected in environmental and engineering investigations. To account for the substantial changes regarding the market shares of operating systems within the last two decades, the library is specifically implemented and tested for cross-plattform usage. We illustrate the applicability of the formikoj library to forward-model seismic refraction waveform data with the `SeismicWaveformModeler` based on a custom subsurface model and survey geometry. Moreover, we present two case studies from seismic refraction tomography field measurements to exemplify the wide range of possibilities provided by the formikoj library for the processing of real data. In particular, we explore different visualization of the seismic traveltime readings to enhance their consistency prior to the inversion. The low-level access provided by our library aims at giving the users the possibility to implement particular modeling, visualization and processing tools specifically designed for their objectives or other geophysical methods.

CRediT authorship contribution statement

Matthias Steiner: Conceptualization and implementation of the library, creating the figures, preparation of the manuscript. **Adrián Flores Orozco:** Conceptualization of the library, preparation of the manuscript.

1. Introduction

The acquisition of spatially quasi-continuous data in a non-invasive manner renders geophysical methods suitable for engineering and environmental investigations. Several studies have demonstrated the successful application of geophysical methods, for instance, to evaluate groundwater remediation techniques (e.g., Flores Orozco et al., 2015), as well as for the investigation of landfills (e.g., Nguyen et al., 2018; Steiner et al., 2022), soil structure (e.g., Romero-Ruiz et al., 2018), or the critical zone (e.g., Parsekian et al., 2015). However, the processing of geophysical data often relies on commercial software solutions and the associated licensing costs might render their use prohibitively expensive, which might be the case for academic projects or institutions. The most popular packages are Res2DInv[1] for electrical methods, Halliburton Landmark SeisSpace ProMAX[2] or ParkSeis [3] for seismic methods, ReflexW[4] for ground-penetrating radar and seismic methods, or the Zond[5] processing tools for various geophysical methods including seismic, gravimetric and electromagnetic data. A common limitation of the aforementioned software solutions

*Corresponding author
ORCID(s): 0000-0002-3595-3616 (M. Steiner); 0000-0003-0905-3718 (A. Flores Orozco)

[1]https://www.geometrics.com/software/res2dinv/, last accessed on July 15, 2022
[2]https://www.landmark.solutions/SeisSpace-ProMAX, last accessed on July 15, 2022
[3]https://www.parkseismic.com/parkseis/, last accessed on July 15, 2022
[4]https://www.sandmeier-geo.de/reflexw.html, last accessed on July 15, 2022
[5]http://zond-geo.com/english/zond-software/, last accessed on July 15, 2022

refers to their specific platform requirements mainly related to the type and version of the operating system; moreover, the possibility to adapt the code are limited if possible at all. Considering the substantial changes regarding the market shares of operating systems within the last two decades, platform-specific software packages are becoming particularly obstructive for academic research and teaching. The increasing popularity of the Python programming language led to the development of various cross-platform open-source software packages for processing, modeling and inverting geophysical data. Available packages can focus on specific geophysical methods, for instance, ResIPy (Blanchy et al., 2020) for electrical data, GPRPy (Plattner, 2020) for ground-penetrating radar data, or ObsPy (Beyreuther et al., 2010) and Pyrocko (Heimann et al., 2017) for seismological data. Other packages provide frameworks for the inversion and permit the inclusion of forward models for different geophysical methods, e.g., SimPEG (Cockett et al., 2015), Fatiando a Terra (Uieda et al., 2013) or pyGIMLi (Rücker et al., 2017).

The seismic refraction tomography (SRT) is a standard technique in environmental and engineering applications. Often applied together with other geophysical methods, the SRT is routinely used, e.g., in alpine and arctic permafrost studies (e.g., Hilbich, 2010; Ramachandran et al., 2011; Krautblatter and Draebing, 2014; Steiner et al., 2021), for the investigation of landslides (e.g., Samyn et al., 2012; Uhlemann et al., 2016; Whiteley et al., 2020), or for hydrogeological characterizations (Ronczka et al., 2018; Bücker et al., 2021). The market for seismic processing software has long been dominated by software packages designed for the processing of large exploration datasets. Accordingly, these seismic processing solutions may not be suited for small-scale projects in environmental and engineering studies, or for teaching purposes. ReflexW overcomes such limitations by providing processing tools specifically designed for near-surface investigations at substantially lower costs. In terms of licensing costs, the developers of Seismic Unix (Stockwell, 1999) went a step further by making their unix-based seismic processing framework available entirely free of charge; whereas Guedes et al. (2022) recently presented RefraPy, a python processing tool for seismic refraction data. Implemented in python, RefraPy is potentially suitable for cross-platform usage, yet Guedes et al. (2022) developed and tested RefraPy solely for Windows operating systems. Moreover, RefraPy does not offer the possibility to generate synthetic seismic waveform data, as required for survey design, as well as teaching and interpretation purposes.

The formikoj library presented here is an open-source framework for creating synthetic datasets, as well as for managing and processing numerical and field data independently from the operating system and without licensing costs; thus, overcoming limitations associated to existing solutions. The design of the library follows the multi-method concept of pyGIMLi and SimPEG, which allows for the implementation of custom designed tools for different geophysical methods. The usage of transparent file formats, e.g., the unified data format (udf[6]), and data management concepts (SQLite database) within the formikoj framework facilitates a simple data exchange between partners in research projects and academia, which is required to guarantee the repeatability of results and good research practices.

[6] http://resistivity.net/bert/data_format.html, last accessed on July 15, 2022

Considering the diverse applications of the SR method we demonstrate the applicability of the proposed library based on tools implemented within the formikoj framework for the modeling and processing seismic waveform data. In particular, we present here a series of illustrative use cases based on the formikoj library referring to (i) the modeling of synthetic seismic refraction (SR) waveform data, (ii) the processing of a 2D SR field dataset collected with a roll-along survey geometry, and (iii) the processing of a 3D SR field data set.

2. Design and structure of the `formikoj` library

As illustrated in Figure 1, the formikoj library comprises a modeling and a processing module, which both rely on a common utilities module. The `DataModeler` and the `MethodManager` class provide the basis to add modeling or processing functionalities for specific geophysical methods. In particular, we present here the `SeismicWaveformModeler` and `SeismicRefractionManager` classes implemented within the formikoj framework, which are used to create and process seismic waveform data, respectively. Similar to RefraPy, these classes are built upon the functionalities of existing packages such as ObsPy for the processing of seismological data (Beyreuther et al., 2010) and pyGIMLi for the modeling and inversion of different geophysical data (Rücker et al., 2017). Other important third party dependencies refer to NumPy (Harris et al., 2020) and Pandas (McKinney, 2010) for general data handling, as well as matplotlib (Hunter, 2007) and PyVista (Sullivan and Kaszynski, 2019) for data visualization. In the current version, we implemented and tested formikoj primarily on Linux machines, yet successful tests and applications can be reported for all major operating systems, i.e., Linux, MacOS and Windows.

2.1. Generation of seismic waveform data for synthetic subsurface models: The `SeismicWaveformModeler`

The SR methods exploits the ground motion recorded by sensors installed in the surface (e.g., geophones) to characterize the propagation of seismic waves generated at well known locations (i.e., shot stations). The visualization of the ground motion as function of time yields a so-called seismogram for each geophone position. Accordingly, the `SeismicWaveformModeler` class provides a flexible way to generate synthetic seismic waveform data either in a python script, interactively in a jupyter notebook or an ipython shell. To create an instance of the class the path to the working directory is provided as parameter to the constructor:

```
from useis import SeismicWaveformModeler
```

```
swm = SeismicWaveformModeler('.')
```

```
INFO    : Created instance of SeismicWaveformModeller
```

The working directory needs to contain a subdirectory *in*, whereas the output directory *out* will be created automatically:

```
working_directory
```

Table 1
Description of the information to be provided in the columns of a measurement scheme in csv format.

Column	**Content**	**Data type**	**Description**
1	x coordinate	float	Station x coordinate, e.g., given in (m)
2	y coordinate	float	Station y coordinate, e.g., given in (m)
3	z coordinate	float	Station z coordinate, e.g., given in (m)
4	Geophone	bool	1 in case of a receiver station, 0 otherwise
5	Shot	bool	1 in case of a shot station, 0 otherwise

```
├── in
└── out
```

The required input files need to be provided via the subdirectory *in* of the working directory. The key input file is the measurement scheme, which contains information regarding the distribution of the shot and geophone stations. If provided in the unified data format the measurement scheme is imported directly with pyGIMLi as a `DataContainer`. In case the measurement scheme is provided as a csv file the `SeismicWaveformModeler` reads the information and writes it to a pyGIMLi `DataContainer`. In the csv format the measurement scheme contains a single line for each station in the survey layout, where a station refers to a location that either hosts a geophone or a shot, or both (see Table 1). The values provided in each line need to be separated by a unique character, preferably a comma (',') as suggested by the file extension csv (comma separated values). Alternatively, tabulator or semicolon separated columns are also possible; yet, for more details regarding separators in csv files we refer to the documentation of the pandas method `read_csv`[7]. In the csv format, the measurement scheme must not contain a header in order to be readable by the `SeismicWaveformManager`.

For the modeling of the seismic waveforms, we need to provide the properties of the base wavelet, the synthetic subsurface model and the resulting waveform datasets, as presented in Table 2. The values for those parameters are provided through a configuration file following the yaml format, e.g.,

[7] https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html, last accessed on July 15, 2022

```yaml
wavelet:
    length: 1.024
    frequency: 100
    sampling_rate: 2000
    pretrigger: 0.02

dataset:
    number: 2
    names: [dataset1, dataset2]
    noise: 0
    noise_level: 1.e-4
    missing_shots: 0
    broken_geophones: 0
    wrong_polarity: 0

model:
    velmap: [[1, 750.], [2, 3000.]]
    layers: [[1, 3.], [2, 10.]]
    quality: 32
    area: 10
    smooth: [1, 10]
    sec_nodes: 3

traveltimes:
    noise_relative: 0.
    noise_absolute: 0.
```

A configuration file located in the input directory can be loaded with the `load` method as follows:

```
swm.load('config')
```

```
WARNING : Multiple configuration files found
INFO    : Provide specific configuration file name
```

```
swm.load('config config_clean.yml')
```

```
INFO    : Configuration loaded
```

By calling the `load` method with the parameter `config`, the `SeismicWaveformModeler` searches the input directory for a configuration file. In case a single yaml file is found, the `SeismicWaveformModeler` automatically loads this file. However, if multiple configuration files are present, e.g., in a project that tests different configurations, the `SeismicWaveformModeler` issues a warning and asks the user to provide the name of the configuration file to be loaded.

In the exemplary yaml file shown above, the first block of the configuration file contains information regarding the wavelet, which controls the modeling of the seismic waveforms as described in Table 2. In the second block, we can set specific names for the datasets to be created and the number of datasets is automatically determined. Alternatively, it is possible to set the number of datasets to be created and the dataset names are automatically generated with the prefix *dataset_*. As described in Table 2, there are various parameters, which control the random error (`noise`) and systematic errors in the modeled seismic waveform data (i.e., missing shots, broken geophones, or traces with wrong polarity). The number and position of the shot and geophone stations affected by the systematic errors are randomly chosen, yet we set the maximum as 5% of the total number of stations in order to avoid a high number of invalid trace data.

The third block contains information regarding the synthetic subsurface model, where for each layer in the model

the corresponding velocity (`velmap`) and layer thickness (`layers`) need to be provided, where all layers are considered to be parallel to the surface topography (geometrical information regarding the stations in the measurement scheme). The remaining parameters, namely `quality`, `area`, `smooth` and `sec_nodes`, define the properties of the mesh to be used for the forward modeling (we refer to the respective pyGIMLi resources[8] for further information). Alternatively, the user can provide a more complex mesh in the binary mesh format (i.e., a bms file) In the same way as for the configuration file, the `SeismicWaveformModeler` first checks whether a single bms file is located in the input directory. Only in case multiple meshes are found the user has to specify which mesh file needs to be loaded:

```
swm.load('mesh')
```

```
WARNING : Multiple mesh files found
INFO    : Provide specific mesh file name
```

```
swm.load('mesh mesh.bms')
```

```
INFO    : Mesh loaded
```

The parameters in the final block do not concern the synthetic waveforms but the forward modeling of the corresponding seismic traveltimes (see Table 2). Based on the exemplary configuration file shown above, the `SeismicWaveformModel` would provide undisturbed traveltimes as both the relative and the absolute noise (`noise_relative` and `noise_absolute`) are set to zero. If we provide values for those parameters we can add random noise to modeled traveltimes in order to mimic picking errors observed in manually picked traveltimes.

Once the `SeismicWaveformModeler` instance is parameterized, we call the `create` method with the parameter `waveforms` to create the synthetic seismic waveform data:

```
swm.create('waveforms')
```

```
INFO    : Measurement scheme loaded
INFO    : Velocity model created
INFO    : Wavelet created
[++++++++++++++ 100% ++++++++++++++] 2048 of 2048 complete
      ...
[++++++++++++++ 100% ++++++++++++++] 2048 of 2048 complete
INFO    : Dataset 'clean' created
```

As can be seen from the `INFO` messages above, the `SeismicWaveformModeler` loads the measurement scheme located in the input directory and creates the velocity model used for the waveform modeling based on the mesh and the model parameters provided in the configuration file. Based on the wavelet parameters provided in the configuration file, the `SeismicWaveformModeler` creates a Ricker wavelet with the given length and frequency by using the pyGIMLi function `ricker` provided in the *seismics* module. Subsequently, mesh, velocity model and Ricker wavelet are used to solve the pressure wave equation for each shot station defined in the measurement scheme with the pyGIMLi function `solvePressureWave`, which is also provided in the *seismics* module. The duration of the waveform modeling depends on the number of shots as well as on the length and resolution of the wavelet. The resultant waveforms at

[8] https://www.pygimli.org/pygimliapi/_generated/pygimli.meshtools.html#pygimli.meshtools.createMesh, last accessed July 15, 2022

Table 2
Description of the parameters, which can be defined in a configuration file used for the modeling of the synthetic seismic data.

Parameter	Unit/ Data type	Description
`wavelet`		
length	s	Length of the base wavelet, which also defines the length of the synthetic seismic waveform data
frequency	Hz	Frequency of the base wavelet
sampling_rate	Hz	Defines temporal resolution of the seismic waveforms
pretrigger	s	Add buffer to the seismic waveforms before the onset of the actual data
`dataset`		
number	int	Number of datasets to be created
names	list (string)	Names of the datasets
noise	bool	1 in case noise should be added to the synthetic waveforms, 0 otherwise
noise_level	-	Level of the seismic background noise
missing_shots	bool	1 in case the datasets should be affected by missing shot files, 0 otherwise
broken_geophones	bool	1 in case the datasets should comprise broken geophones (i.e., no data in the corresponding seismograms), 0 otherwise
wrong_polarity	bool	1 in case the datasets should contain traces with inverse polarity, 0 otherwise
`model`		
velmap	list (float/int)	For each layer the first value defines the marker and the second one the seismic velocity within the layer
layers	list (float/int)	For each layer the first value defines the marker and the second one the thickness
`traveltimes`		
noise_relative	%/100	Relative noise to be added to the forward modeled seismic traveltimes
noise_absolute	s	Absolute noise to be added to the forward modeled seismic traveltimes

the corresponding geophone stations are extracted and stored in an ObsPy `Stream` data structure.

For each dataset listed in the configuration file, a subdirectory with the name of the respective dataset will be created in the output directory (*out*):

```
working_directory
├── in
└── out
    └── dataset1
        └── data
            ├── protocol.txt
            ├── station_coords.csv
            └── Shot_1001.syn
```

```
    ├── ...
    └── Shot_10nn.syn
├── dataset1_tt.pck
└── info.txt
```

In the subdirectory *data*, the synthetic seismic waveforms are stored in a separate shot file for each shot position in the miniseed format (Ahern et al., 2012; Ringler and Evans, 2015) with the file extension *syn* to identify the forward modeled shot files, e.g., Shot_1001.syn. The *data* directory also contains a synthetic measurement protocol (protocol.txt) and the station coordinates provided as a csv file (station_coords.csv). The header of the measurement protocol contains the survey parameters required for the processing of the seismic waveforms, namely sampling rate, recording length, number of geophones and geophone spacing. Moreover, the protocol associates each shot file of the dataset to a specific location within the survey geometry, i.e., with respect to the geophone positions:

```
############################
  Line: SYN_dataset_1
  Sampling rate: 2000 Hz
  Recording length: 1.024 s
  Number of geophones: 48
  Geophone spacing: 6 m
############################

    File number | Station
        1001    |   G001
          :     |     :
        1048    |   G048
```

The auxiliary file info.txt provided in the dataset directory summarizes the parameters from the configuration file and information regarding the simulated systematic errors in the synthetic seismic waveform data:

```
Number of geophones: 48
Number of shots: 48
Recording length (s): 1.024
Sampling frequency (Hz): 2000
Wavelet type: Ricker
Frequency of the wavelet (Hz): 100

Missing shot(s): 16
Broken geophone(s): 6, 14
Wrong polarity geophone(s): 35
```

If the corresponding synthetic traveltimes are computed (`swm.create('traveltimes')`), the resultant values are stored in the dataset directory in an udf file, e.g., dataset1_tt.pck. The file extension *pck* is an abbreviation of the word 'pick' and refers to the first break traveltimes saved in the file.

2.2. Processing of seismic refraction datasets: the `SeismicRefractionManager`

The SR method is based on the measurement of the traveltimes of seismic waves determined form the the first onset of the seismic energy recorded by geophones. In the SRT, the inversion of traveltimes gathered from tens to hundreds of seismograms permits the computation of variations in the seismic velocities in an imaging framework. Measuring the traveltimes in seismograms is referred to as first break picking, which is commonly done manually (semi-automatically) in an iterative process. The signal-to-noise ratio in the recorded seismograms substantially influences the quality of

the traveltimes picked in the seismograms, and thus a proper enhancement of the perceptibility of the first onsets is crucial.

Accordingly, the `SeismicRefractionManager` class provides functionalities that permit the processing of seismic waveform data for a refraction seismic analysis. These functionalities involve the reading of seismic waveform data, combining the data with information about the survey geometry, basic processing of the waveforms (e.g., filtering of high frequency noise, reversal of traces with wrong polarity etc.) as well as the picking of first break traveltimes. Since the first break picking is a highly interactive process, the `SeismicRefractionManager` is designed primarily for usage from within an ipython shell. The `SeismicRefractionManager` also allows to implement python scripts for an automatic picking of first break traveltimes, i.e., without the use of the graphical interface.

2.2.1. Compiling the survey information and creating a project

An instance of the `SeismicRefractionManager` can be created by providing a path to a working directory as the parameter to the constructor. Based on the content of the working directory, the `SeismicRefractionManager` automatically decides whether (i) to start in the data preview mode, (ii) create a new project, or (iii) load an existing project from disk.

The data preview mode is primarily initiated if the provided working directory contains seismic shot files:

```
from useis import SeismicRefractionManager
```

```
srm = SeismicRefractionManager('./01_data/raw')
```

```
INFO    : Starting in data preview mode
Progress <==============================> 100.0% completed
INFO    : Read 48 files
```

To create or load a project, the working directory needs to contain specific subdirectories:

```
working_directory
├── 01_data
│   └── raw
├── 02_geom
└── 03_proc
```

In this directory structure, the seismic shot files are stored in `01_data/raw` and the geometry file (`geometry.csv`) is provided in `02_geom`, thus making them locatable for the `SeismicRefractionManager`.

The geometry file is a csv file that provides an abstract representation of the survey layout. The fundamental element for the description of the survey layout is the station, which refers either to a geophone position, a shot position or a position with co-located shot and geophone. For each station the geometric and semantic information described in Table 3 are provided column-wise, i.e., each line in the geometry file corresponds to a single station with a unique position within the survey layout. Similar to the measurement scheme provided to the `SeismicWaveformManager`, the values in the geometry file are preferably separated by a comma and file must not contain a header in order to be readable by the `SeismicRefractionManager`.

Table 3
Description of the information to be provided in the columns of the geometry file.

Col	Content	Data type	Description
1	x coordinate	float	Station x coordinate, e.g., given in (m)
2	y coordinate	float	Station y coordinate, e.g., given in (m)
3	z coordinate	float	Station z coordinate, e.g., given in (m)
4	Geophone	bool	1 if a geophone was deployed at station, 0 otherwise
5	Shot	int	The numerical part of the file name, e.g., 1001, if a shot was conducted at this station, -1 otherwise
6	First geophone	int	First active geophone (-1 if no shot was conducted at the station)
7	Number of geophones	int	Number of active geophones (-1 if no shot was conducted at the station)

In case the shot files as well as the geometry file are provided and a basic sanity check of the geometry file was successful, the `SeismicRefractionManager` creates a new project:

```
from useis import SeismicRefractionManager
```

```
srm = SeismicRefractionManager('.')
```

```
INFO    : Read geometry information from file
INFO    : Extracted shot geometry
INFO    : Extracted receiver geometry
INFO    : Applied geometry
INFO    : Standard pickset 'picks' created
INFO    : Pickset 'picks' loaded
INFO    : 'picks' set as active pickset
Progress <==============================> 100.0% completed
INFO    : Read 48 files
```

Without further interaction with the user, the `SeismicRefractionManager` creates an SQLite database (prj.db in the working directory) to store the geometry information, whereby the stations are consecutively numbered (see Figure 2). To allow for an efficient data selection through the user the `SeismicRefractionManager` links the station numbers to shot index numbers (SIN) and receiver index numbers (RIN) assigned to shot and receiver stations, respectively, as illustrated in Figure 2. Based on this information, the `SeismicRefractionManager` can apply the geometry, which generates database tables required for the processing of the seismic waveform data. In particular, the first break traveltimes for each SIN-RIN pair are stored in a dedicated database table *fbpicks* together with the name of the corresponding pickset, i.e., the a common label for an entire set of first break traveltimes. By default, each project contains the default pickset 'picks', which is loaded and activated on startup. Once the database is initialized, the waveform data are read from disk and the project is ready for processing.

If a SQLite database is already present in the provided working directory, i.e., the project already exists, the `SeismicRefractionManager` will automatically load the project information, the seismic waveforms as well as the default pickset 'picks':

```
from useis import SeismicRefractionManager
```

```
srm = SeismicRefractionManager('.')
```

```
INFO    : Project information loaded
Progress <==============================> 100.0% completed
INFO    : Read 48 files
INFO    : Pickset 'picks' loaded
INFO    : 'picks' set as active pickset
```

For a project with a successfully applied geometry, the `select` method of the `SeismicRefractionManager` allows to gather the seismic waveforms based on a common shot (`sin`), a common receiver (`rin`) or the common absolute offset (`aoffset`):

```
srm.select('sin 1')
```

```
INFO    : 48 traces selected
```

```
srm.select('rin 10')
```

```
INFO    : 48 traces selected
```

```
srm.select('aoffset 6')
```

```
INFO    : 90 traces selected
```

2.2.2. Visualization and enhancement of the seismic waveform data

The `plot` called without passing any parameter allows for the visualization of the currently selected traces:

```
srm.plot()
```

Once opened, the seismogram plot visualizes the seismic waveforms in a combination of wiggle trace and variable area mode, i.e., the trace data are shown as curves and the area of the curves ('wiggles') is colored according to the sign of the amplitude, i.e., red for negative and blue for positive values, respectively (not shown for brevity). Pressing the up or down arrow key toggles the visualization mode between variable area and variable density. The variable density mode illustrates the amplitudes in the seismograms based on the same red-blue color scheme, yet the strength of the amplitude is reflected by the color saturation, i.e., high amplitudes refer to a stronger shade than low amplitudes (see Figure 3).

The active processing mode as well as the data scaling mode and the traveltime at the current cursor position are displayed in the status bar of the seismogram plot window, as shown in Figure 4. The initial processing mode of the seismogram plot is 'Fb pick', i.e., first break picking is possible, yet the user can switch between the different modes by pressing specific keys on the keyboard. The 'm' key activates the trace mute mode ('Trc mute'), which allows to set the amplitude of a trace to zero by clicking with the left mouse button; clicking again on the same trace restores the amplitude information. The trace reverse mode ('Trc rev') is activated by pressing the 'r' key and enables the user to toggle the polarity of a trace by clicking on it with the left mouse button. The default data scaling mode is 'Zoom', which allows the scaling of the y-axis by turning the mouse wheel. By pressing the 'a' key the amplitude scaling mode

('Amp scal') is activated. Turning the mouse wheel increases or decreases the amplitudes of the traces currently shown in the seismogram plot, and thus might help to enhance the perceptibility of the first onsets. By pressing the key of the currently active mode again, the `SeismicRefractionManager` returns to the default processing or data scaling mode, respectively. As illustrated in Figure 4, the different modes can be activated in any arbitrary order, i.e., it is possible to alternately toggle processing and data scaling modes.

The user can visualize the frequency spectrum of the currently selected trace data by calling the `plot` method with the parameter 'spectrum':

```
srm.plot('spectrum')
```

Figure 5 shows the resulting frequency spectrum, which can be used to discriminate the dominating signal frequencies from those associated to the background noise, and thus can be used to define the adequate filter frequencies. To improve the signal-to-noise ratio it is possible to apply filters on the selected traces through the `filter` method of the `SeismicRefractionManager`, which utilizes the frequency filters implemented in the ObsPy package (lowpass, highpass, bandpass and bandstop; Beyreuther et al., 2010), e.g.:

```
srm.filter('lp 120')
INFO    : Applied 120.0 Hz lowpass filter
srm.filter('bp 10 120')
INFO    : Applied bandpass filter (10.0 to 120.0 Hz)
srm.filter('hold on')
INFO    : Set filter hold on
```

By default, filters are solely applied to the currently selected traces, yet setting the filter on hold allows the filtering of all subsequently selected traces with the same filter settings. Highpass and bandstop filters are applied to the seismic waveforms in a similar way, i.e., `srm.filter('hp 10'))` or `srm.filter('bs 45 55')`, respectively. In case the seismogram plot is opened, the effect of the applied filter on the seismic waveforms is interactively visualized.

2.2.3. Analysis of the seismic waveforms and first break traveltime picking

Based on the visualization of the enhanced seismic waveforms (e.g., after filtering) where the first onsets are perceptible, the data can be analyzed and processed to obtain information about the subsurface conditions. By activating the velocity estimation mode ('Vel est') with the 'v' key, the user can estimate velocities for different wave phases (e.g., originating from a refractor) by pressing the left mouse button and moving the cursor within the seismogram plot. Once the left mouse button is released, a line between the start and end point is drawn and labeled with the corresponding velocity. Such velocity estimation lines can be deleted by clicking with the right mouse button.

If the processing mode 'Fb pick' is activated, picking of first break traveltimes is done individually by clicking with the left mouse button on the respective trace. Clicking again on the same trace will set the first break pick to the

new location as there can only be one traveltime for each SIN-RIN pair. An existing traveltime pick can be deleted by clicking on the corresponding trace with the right mouse button. Alternatively, first break picks can be set for multiple traces by pressing the left mouse button and moving the cursor within the seismogram plot. A line connecting the start point with the current cursor location is drawn and updated upon movement. Once the left mouse button is released, first break picks are defined at the intersections between the line and the seismograms. In the same way, multiple picks can be deleted at the same time if a line is drawn with the right mouse button pressed. The first break traveltimes determined in the seismogram plot are automatically written to the project database when the window is closed. Alternatively, the user can navigate through the datasets by pressing the 'left' or 'right' arrow keys. In this way, the traveltimes for the currently plotted traces are saved in the database and the traces for the next or previous SIN or RIN are automatically selected, respectively.

Sometimes it can be helpful to visualize all first break traveltimes in a single diagram. The traveltime diagram for the currently active pickset can be created through the `plot` method of the `SeismicRefractionManager`:

```
srm.plot('traveltimes')
```

The traveltime diagram provides a schematic illustration of the survey geometry and plots the first break traveltimes according to their common SIN. Figure 6 presents an exemplary traveltime diagram, which is a common way to examine the quality of the first break picking. Such illustration of the traveltimes can be used to identify outliers or erroneous measurements, which are commonly associated to traveltimes with substantial deviations from those observed at adjacent stations such as the first break pick for the SIN-RIN pair (23, 11). Outliers can be removed by clicking on the corresponding symbol ('x') in the traveltime diagram, which is instantly synchronized with the project database. If the seismogram plot and the traveltime diagram are used side-by-side, changes made to the first break picks in one window will interactively trigger an update of the other one and vice versa.

The `SeismicRefractionManager` handles first break picks in so-called picksets, which can be organized and manipulated through the `picksets` method of the `SeismicRefractionManager`:

```
srm.picksets()
```

```
pickset         loaded          active
-----------------------------------------
picks           Y               Y
```

```
srm.picksets('import picking_part1.pck pck_p1')
```

```
INFO    : Created new pickset 'pck_p1'
INFO    : Pickset 'pck_p1' loaded
INFO    : 'pck_p1' set as active pickset
INFO    : Imported 'picking_part1.pck' to pickset 'pck_p1'
```

```
srm.picksets('copy pck_p1 pck_all')
```

```
INFO    : Created new pickset 'pck_all'
INFO    : Pickset 'pck_all' loaded
INFO    : 'pck_all' set as active pickset
INFO    : Copied pickset 'pck_p1' to 'pck_all'
```

```
srm.picksets('unload pck_p1 picks')
```

```
INFO    : Pickset 'pck_p1' removed from workflow
INFO    : Pickset 'picks' removed from workflow
```

```
srm.picksets('delete pck_p1')
```

```
INFO    : Pickset 'pck_p1' deleted
```

```
srm.picksets()
```

```
pickset         loaded          active
-----------------------------------------
picks           N               N
pck_all         Y               Y
```

Calling the `pickset` method without parameters shows the status of all picksets in the project. From the above use case, we see that by default, the pickset 'picks' is loaded (available in the workflow) and activated, i.e., modifications of first breaks are stored in this pickset. First break picks provided by another source (as an udf file) can be stored in *03_proc/picks* and then imported into the current project. For the further processing, it is often sufficient to keep only one pickset in the workflow, i.e., not required picksets can be unloaded. A pickset currently not loaded can be deleted permanently, i.e., the corresponding traveltimes are removed from the database. The `picksets` method can also be used to load picksets from the database:

```
srm.picksets('load picks')
```

```
INFO    : Pickset 'picks' loaded
```

```
srm.picksets()
```

```
pickset         loaded          active
-----------------------------------------
picks           Y               N
pck_all         Y               Y
```

```
srm.picksets('use picks')
```

```
INFO    : Pickset 'picks' loaded
INFO    : 'picks' set as active pickset
```

```
srm.picksets()
```

```
pickset         loaded          active
-----------------------------------------
picks           Y               Y
pck_all         Y               N
```

When a pickset is loaded from the database, it does not become the active pickset automatically; whereas by using the parameter `use` the corresponding pickset is loaded and also becomes the active pickset. The first break traveltimes of a pickset can be exported to an udf file that is stored in *03_proc/picks* subdirectory with the current timestamp as suffix:

```
srm.picksets('export pck_all')
```

```
INFO    : pickset 'pck_all' saved to pck_all_20220428T145648.pck
```

Such exported pick file can then be shared (e.g., with project partners) or used to run an inversion of the corresponding first break traveltimes to solve for a model of the seismic velocities associated to the subsurface materials.

3. Exemplary use cases

3.1. Modeling and processing of synthetic seismic waveform data

To demonstrate the applicability of the `SeismicWaveformModeler` class, we generate two synthetic seismic waveform datasets assuming two horizontal layers in a half-space without topography and an interface between these layers with a constant depth. Table 4 summarizes the parameterization used for the forward modeling of one dataset without noise (Dataset_1) and another dataset with added random noise and systematic errors (Dataset_2). For the visualization of both synthetic datasets, we provide the corresponding geometry files to the `SeismicRefractionManager` in order to have full control regarding data selection and processing capabilities. Alternatively, we could omit the geometry file and start the `SeismicRefractionManager` in the data preview mode.

The synthetic seismic waveforms for SIN 24 of Dataset_1 presented in Figure 7 reveal negative first onsets in all traces. From these seismograms, we can also identify the crossover points, i.e., inflexion points between the first arrivals stemming from the first layer (RIN 17 to 30) and the first onsets associated to the second layer (RIN 1 to 16 and RIN 31 to 48). In Figure 8, we present the synthetic seismic waveforms from Dataset_2 for the same shot (SIN

Table 4
Measurement scheme and parameters provided in the yaml files used to create synthetic seismic waveform datasets with (Dataset_1) and without added noise (Dataset_2).

Measurement scheme		
Number of stations	48	
Station spacing	2 m	
Number of geophones	48	
Number of shots	48	
Model	*Layer 1*	*Layer 2*
Thickness	3 m	10 m
Velocity	750 m/s	3000 m/s
Dataset	*Dataset_1*	*Dataset_2*
Noise	False	True
Noise level	0	1e-5
Missing shots	False	True
Broken geophones	False	True
Wrong polarity	False	True
Wavelet		
Length	1.024 s	
Frequency	100 Hz	
Sampling rate	2000 Hz	

24). In contrast to Dataset_1, the signal-to-noise ratio is a function of the offset between shot and geophone position, i.e., traces farther away from the shot contain a higher level of random noise. From Figure 8, we can also identify systematic errors in the dataset, where RIN 6 and 14 refer to broken geophones, and RIN 35 is an example for readings with a wrong polarity.

Synthetic datasets as presented in Figures 7 and 8 might be useful for the evaluation of processing strategies or investigating the effect of complex survey geometries on the seismic data. The design of the formikoj library and the concept of the `SeismicRefractionManager` allow the implementation of supplementary functionalities for the visualization and processing of seismic refraction waveform data. Such custom extensions are expected to be implemented either as an internal method or as a function in the utilities module, which should then be called through the `compute` method based on a new custom keyword. We believe that the possibility for customization can be exploited in class-room activities but might also be relevant for the design of seismic refraction surveys and processing strategies. Accordingly, we demonstrate the flexibility provided within the formikoj library through implementing simplified versions of an automatic first break picking algorithm as well as the Plus-Minus method (Hagedoorn, 1959).

In the first example, we consider an automatic first break picking algorithm, which determines the traveltimes based on the energy ratio method (e.g., Earle and Shearer, 1994). To use this custom implementation we added the keyword `autopick`, which is passed as the first parameter to the `compute` method:

```
srm.compute('autopick all autopicks')
```

```
INFO    : Created new pickset 'autopicks'
INFO    : Pickset 'autopicks' loaded
INFO    : 'autopicks' set as active pickset
Progress <============================> 100.0% completed
```

```
srm.picksets('export autopicks')
```

```
INFO    : pickset 'autopicks' saved to autopicks_20220429T073017.pck
```

The second parameter defines whether automatic first break picking should be applied to the entire data (`all`) or just to the currently selected traces (`cur`). The third parameter in the example above refers to the name of the pickset that will hold the automatically determined first break traveltimes. Once the auto-picking process is finished, we can export the pickset to a pck file. The synthetic traveltimes computed with the `SeismicWaveformModeler` and the auto-picked traveltimes can then be compared, for instance, in form of a scatter plot as shown in Figure 9. Prior to the comparison we remove outliers from the automatically determined traveltimes, e.g., based on a simple histogram analysis or a more sophisticated approach. In our example, synthetic and auto-picked traveltimes show a strong correlation ($R = 0.99$), yet we observe a systematic deviation from the perfect correlation (1:1 line). This deviation likely indicates that the implementation of the automatic picking algorithm is sensitive to a different part of the waveform, e.g., the first positive wiggle instead of the first negative one indicated by the synthetic traveltimes. At this point, we could verify such interpretation by plotting the waveform data together with both picksets and try to modify the implementation of the auto-picking algorithm accordingly; however, these steps are not shown or discussed here further.

The second example refers to the implementation of the Plus-Minus method (Hagedoorn, 1959). This algorithm is applied to estimate the depth to the interface between layers as well as the velocities within the layers. In particular, we follow Dufour and Foltinek (1996) who propose an automatized analysis of the first break traveltimes. The key step in this approach is the automatic determination of the crossover points based on the traveltime difference analysis (Lawton, 1989) and the branchpoint analysis (Wang and Cheadle, 1995). For adjacent shot gathers we compute the traveltime differences from which we find the location of a specific crossover point as the maximum of the second derivative of the corresponding traveltime differences (Dufour and Foltinek, 1996). In the same way as the auto-picking method we call the Plus-Minus method through the `compute` method of the `SeismicRefractionManager`:

```
srm.compute('plusminus')
```

```
Progress <============================> 100.0% completed
```

Without any further input required from the user, the workflow suggested by (Dufour and Foltinek, 1996) solves for the layer velocities and the interface depth. A comparison of the result obtained for the traveltimes associated to Dataset_1 and the true subsurface conditions is presented in Figure 10. As expected for undisturbed traveltimes, the implemented approach accurately resolves the known properties of the synthetic model. Further tests could investigate the performance of the implemented algorithm for increasing noise levels, complex survey geometries or challenging

Table 5
Extract from the roll-along survey geometry file showing how the information regarding the first geophone assigns the traces in the shot files to the correct stations.

x (m)	y (m)	z (m)	Geo	Shot	1st Geo	# Geo
0.0	0.0	163.5	1	-1	-1	-1
2.0	0.0	163.5	0	1001	**1**	48
⋮	⋮	⋮	⋮	⋮	⋮	⋮
94.0	0	163.5	0	1024	**1**	48
96.0	0	163.5	1	-1	-1	-1
98.0	0.0	163.5	0	1037	**25**	48
⋮	⋮	⋮	⋮	⋮	⋮	⋮
194.0	0.0	163.5	0	1049	**25**	48
196.0	0.0	163.5	1	-1	-1	-1
198.0	0.0	163.5	0	1073	**49**	48
200.0	0.0	163.5	1	-1	-1	-1
202.0	0.0	163.5	0	1050	**25**	48
⋮	⋮	⋮	⋮	⋮	⋮	⋮
286.0	0.0	163.5	0	1084	**49**	48
288.0	0.0	163.5	1	-1	-1	-1
290.0	0.0	163.5	0	1097	**73**	48
⋮	⋮	⋮	⋮	⋮	⋮	⋮
386.0	0.0	163.5	0	1109	**73**	48
388.0	0.0	163.5	1	-1	-1	-1
390.0	0.0	163.5	0	2025	**97**	48
⋮	⋮	⋮	⋮	⋮	⋮	⋮
570.0	0.0	163.5	0	2048	**97**	48
572.0	0.0	163.5	1	-1	-1	-1

surface topography; yet, such detailed analysis is not within the scope of this manuscript.

3.2. First break travel time picking for a 2D roll-along field data set: the Danube island example

The seismic data used in this example were collected at the Danube island (Vienna) in June 2021, using 48 geophones deployed with 2 m spacing between them and shot locations located between the geophone positions. As illustrated in Figure 11, the survey geometry refers to a roll-along geometry, i.e., the geophone spread was moved along a profile with 50% overlap yielding a total of five segments.

In the field, each segment was measured separately, yet for the processing all measurements are combined in a single profile, i.e., the stations of all segments are consecutively numbered. We can implement such measurement layout in the geometry file of the `SeismicRefractionManager` as illustrated in Table 5. The key parameter is '1st Geo' to describe a roll-along survey as it indicates the first active geophone along the profile for each shot file. For the first segment in a roll-along survey geometry, the first active geophone is always geophone 1. From the number of geophones used in each segment, and an overlap of 50%, we can easily find the first geophone for segments two to five to be 25, 49, 73 and 97, respectively. By providing these values in the geometry file as shown in Table 5, the `SeismicRefractionManager` can build the corresponding geometry in the project database.

Based on the shot files and the geometry file stored in the required directory structure we can can create the project.

Once the geometry is applied, i.e., the project is ready for processing, we can obtain a first illustration of the subsurface conditions by computing a common offset stack (COS):

```
srm.compute('cos')
```

```
Progress <==============================> 100.0% completed
INFO    : Computed the common offset stack
```

For the common offset stack, all traces with the same absolute offset from a shot point are stacked, i.e., compute the mean of the summed trace data. In this way, we can reduce the influence of the incoherent noise and thus improve the signal-to-noise ratio. The COS for the Danube island dataset presented in Figure 12 shows first onset for absolute offsets up to approximately 150 m, which suggest a two-layered subsurface model. By using the velocity estimation functionality we can approximate the seismic velocity in the corresponding layers.

After getting a first impression of the subsurface conditions, we can start the first break picking progress by selecting a set of traces, filter the trace data if necessary and show the seismogram plot:

```
srm.select('sin 99')
```

```
INFO    : 48 traces selected
```

```
srm.filter('lp 120')
```

```
INFO    : Applied 120.0 Hz lowpass filter
```

```
srm.plot()
```

For this example, we select the trace data for SIN 99, which refers to a shot position located in segment four. We can easily verify this in the seismogram plot presented in Figure 13, where the lowest RIN (along the x-axis) is 73, corresponding to the first active geophone defined for this segment. Based on the applied geometry, we can determine first break traveltimes in the seismogram plot, which are assigned to the corresponding SIN-RIN pairs. In the exemplary trace data shown in Figure 13, the first onsets are easily identifiable although the seismic background noise increases substantially at large offsets (RIN 73 to 90).

Picking first onsets for the entire dataset, i.e., for SIN-RIN pairs in all segments, will provide a large number of data points. Hence, the traveltime diagram might not be the most efficient way to review the traveltimes prior to the inversion. Alternatively, we can use a pseudosection, which illustrates the apparent seismic velocity determined from the picked traveltimes and the distance between shot and geophone. Those apparent velocity values are assigned to pseudolocations, i.e., the corresponding x- and z-coordinates are determined as the midpoint and as $1/3$ of the absolute offset between the shot and geophone, respectively. We can create a pseudosection through the `plot` method of the `SeismicRefractionManager`:

```
srm.plot('pseudosection')
```

A pseudosection, as presented in Figure 14, allows for the identification of outliers in the data, e.g., stark velocity

contrasts for adjacent points, or systematic errors, e.g., velocities erroneously influenced by a single shot or receiver. The main assumption here is that the pseudosection should reveal smooth transitions between lateral and vertical neighbors, considering that the data were collected with gradual changes in the position of the source (i.e., hammer blow) and the receiver (i.e., geophone). Of course the pseudosection will reveal large variations in case of abrupt changes in the topography or the geometry of the array. Such information is known by the user, and thus can be taken into account during the evaluation of the pseudosection aiming at the identification of outliers and possible systematic errors. For the Danube island dataset, the pseudosection suggests an increase in the seismic velocity along profile direction in deeper subsurface regions (i.e., larger pseudodepth). Such pattern could be related to the fault crossing the profile, thus indicating that the geophysical survey was conducted in an appropriate location to detect the fault. Moreover, the pseudosection presented in Figure 14 shows a low number of data points in the first segment of the Danube island survey. Such lack of data points at large pseudodepths is due to a low number of picked traveltimes at large offsets, and thus might indicate a low signal-noise ratio.

The `SeismicRefractionManager` also provides the possibility to review the data quality along the entire profile by visualizing the picking percentage, i.e., the ratio of actually picked traveltimes and total number of SIN-RIN pairs:

```
srm.plot('pickperc')
```

As can be seen from Figure 15, the picking percentage is visualized separately for each shot (SIN along the x-axis). In this way, we can easily identify shots affected by a low signal-to-noise ratio based on the correspondingly low picking percentage. For the Danube island dataset, we find a low picking percentage in the first segment, which confirms the lack of datapoints observed in the pseudosection. Besides the assessment of the data quality, the picking percentage plot can be used to track the picking progress. In case of a large dataset, for which the traveltimes cannot be determined in one session, the picking percentage plot provides information where to resume the first break picking in the next session. Moreover, it is possible to identify single shots that might have been forgotten during the first break picking. Hence, it is advisable to have a look at the picking percentage at least before exporting the corresponding traveltimes for the inversion.

3.3. Processing of a 3D seismic refraction dataset: the soda lake example

The `SeismicRefractionManager` allows also the visualization and processing of data collected with a 3D survey layout. To illustrate these capabilities, we present in Figure 16 the geometry of a 3D survey conducted in a soda lake located close to Vienna. The soda lake corresponds to quarternary sediments where capillary forces have developed a low permeable layer close to the surface (between 50 and 100 cm) with a high clay and salt content. The seismic survey aims to support the interpretation of the electrical and electromagnetic models obtained in a monitoring framework. Accordingly, the survey geometry shown in Figure 16 was specified by previously conducted electrical measurements

with electrodes arranged along two perpendicular lines. The seismic data were collected with 48 geophones deployed along the North-East to South-West oriented line, and 48 geophones along the North-West to South-East oriented line, with a spacing of 2 m between the geophones. Shots were generated with an 8 kg sledgehammer at the geophone positions as well as at positions along the diagonals to obtain a sufficient coverage, as indicated in Figure 16.

In the geometry file, we provide the 3D coordinates for each shot or receiver station, from which the `SeismicRefractionMan` infers the 3D survey layout and automatically configures the project for 3D processing. Figure 17 presents the seismic waveform data recorded for SIN 1, i.e., the shot position co-located with the first geophone (Station 01 in Figure 16). The data for RIN 1 to 48 appear familiar as the corresponding SIN-RIN layout refers to the one of conventional 2D profiles. In contrast, the seismic waveforms recorded at RIN 49 to 96 show an entirely different pattern. To understand such visualization, we need to take into account that RIN 49 to 96 are deployed perpendicular to the direction of propagation of the wavefront originating to the seismic wavefront propagating away from SIN 1. Accordingly, the observed curvature in the first onsets is due to the varying offset of the different SIN-RIN pairs. Irrespective of the survey geometry in the field, it is possible to determine the first break traveltimes for the SIN-RIN pairs shown in the seismogram plot. In this particular example, the first onsets are easy to identify, thus allowing to set first break picks for almost all traces.

Since we are processing a 3D dataset, we cannot use the conventional pseudosection to illustrate the apparent velocity values obtained from the picked first break traveltimes. However, the project is configured for 3D processing so the `SeismicRefractionManager` automatically switches to a 3D representation. In particular, the apparent velocity values are illustrated in an interactive 3D pseudosection, which can be rotated and the image section can be zoomed and panned; thus, permitting the user to easily investigate the data quality for 3D geometries. In Figure 18, we present a screenshot of the 3D pseudosection for the traveltimes corresponding to the salt lake dataset. However, such screenshot cannot reveal the full capabilities implemented in the `SeismicRefractionManger` for the interactive analysis and visualization of 3D pseudosections.

Once the first break picking is finished, the corresponding pickset can be exported as pck file for the inversion, e.g., in pyGIMLi. Prior to the exporting of the first break traveltimes, it is advisable to have a look at the picking percentage to check for any missing data. As can be seen from Figure 19, the visualization of the picking percentage plot does not change due to the 3D survey, and thus can be read in the same way as in case of a 2D survey geometry. The inversion results and their interpretation are not the scope of this manuscript, yet Figures 17 and 18 reveal the capabilities provided by the proposed framework for the visualization and processing of data collected in 3D survey geometries.

4. Conclusions and Outlook

We have presented formikoj, a flexible open-source library enabling the development of modeling and processing tools for geophysical data. Implemented in python and tested on all major operating systems (Linux/Unix, MacOS, Windows), formikoj is suitable for multi- and cross-platform applications; thus, allowing collaboration between users free from licensing costs and platform requirements.

We demonstrated the capabilities of the formikoj framework to develop versatile and easily scalable classes for the modeling and processing of waveforms in seismic refraction surveys. The required interaction with the user is reduced to a minimum as crucial processing steps are automatized within the `SeismicWaveformModeler` and `SeismicRefractionManager` based on efficient data input strategies, for instance the preparation and import of the geometry file or the keyboard-based interaction related to the first break picking. In this regard, the user controls the formikoj library by providing text-based commands preferably through an ipython shell to exploit the full interactive potential modeling and processing tools. However, applications of the formikoj library can also be automatized by implementing workflows in python scripts or jupyter notebooks.

Based on three exemplary use cases, we illustrated the applicability of both the `SeismicWaveformModeler` and the `SeismicRefractionManager`. In the first use case, we showed the possibility to forward model seismic waveform data based on custom subsurface models and survey geometries with the `SeismicWaveformModeler`. Additionally, the resulting waveforms can be subjected to systematic and random noise sources. The capabilities of the `SeismicRefractionManager` were demonstrated through the processing of field datasets collected in complex survey layout, namely a roll-along and a 3D geometry. Moreover, we showed how the different data visualization options can assist during the data processing to ensure consistency in the first break traveltimes. In particular, we developed a visualization of the traveltimes by means of pseuodsections illustrating the corresponding apparent seismic velocities. Such plots allow for a quick identification of systematic errors and outliers in both 2D and 3D datasets.

By making the source code of the formikoj library available under a permissive open-source license we intend to spark the development of further modeling and processing tools for various geophysical models based on this framework. Our further efforts will focus on implementing tools for other wave-based geophysical methods used in frame of our research activities, such as multi-channel analysis of (seismic) surface waves or magnetotelluric surveys.

5. Acknowledgments

The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme. The authors are grateful to Nathalie Roser and Lukas Aigner for benchmarking the formikoj library against established software packages in frame of their research activities. Furthermore, we would like to thank Clemens Moser, Martin Mayr, Vinzenz Schichl and Harald Pammer for their constructive comments during first tests of the formikoj framework

as well as for their help during the seismic surveys.

Code availability section

Name of the code/library: formikoj

Contact: matthias.steiner@geo.tuwien.ac.at

Hardware requirements: No specific requirements

Program language: Python

Software required: Anaconda/Miniconda recommended

The source codes and exemplary data sets are available for downloading at the link: https://git.geo.tuwien.ac.at/msteine1/form

References

Ahern, T., Casey, R., Barnes, D., Benson, R., Knight, T., Trabant, C., 2012. SEED Reference Manual, Version 2.4. URL: http://www.fdsn.org/pdf/SEEDManual_V2.4.pdf.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., Wassermann, J., 2010. Obspy: A python toolbox for seismology. Seismological Research Letters 81, 530–533. doi:10.1785/gssrl.81.3.530.

Blanchy, G., Saneiyan, S., Boyd, J., McLachlan, P., Binley, A., 2020. Resipy, an intuitive open source software for complex geoelectrical inversion/modeling. Computers & Geosciences 137, 104423. URL: https://www.sciencedirect.com/science/article/pii/S0098300419308192, doi:https://doi.org/10.1016/j.cageo.2020.104423.

Bücker, M., Flores Orozco, A., Gallistl, J., Steiner, M., Aigner, L., Hoppenbrock, J., Glebe, R., Morales Barrera, W., Pita de la Paz, C., García García, C.E., et al., 2021. Integrated land and water-borne geophysical surveys shed light on the sudden drying of large karst lakes in southern mexico. Solid Earth 12, 439–461. doi:10.5194/se-12-439-2021.

Cockett, R., Kang, S., Heagy, L.J., Pidlisecky, A., Oldenburg, D.W., 2015. Simpeg: An open source framework for simulation and gradient based parameter estimation in geophysical applications. Computers & Geosciences 85, 142–154. URL: https://www.sciencedirect.com/science/article/pii/S009830041530056X, doi:https://doi.org/10.1016/j.cageo.2015.09.015.

Dufour, J., Foltinek, D.S., 1996. The Plus-Minus time analysis method and its implementation. Technical Report. The CREWES Research Report.

Earle, P.S., Shearer, P.M., 1994. Characterization of global seismograms using an automatic-picking algorithm. Bulletin of the Seismological Society of America 84, 366–376.

Flores Orozco, A., Velimirovic, M., Tosco, T., Kemna, A., Sapion, H., Klaas, N., Sethi, R., Bastiaens, L., 2015. Monitoring the injection of microscale zerovalent iron particles for groundwater remediation by means of complex electrical conductivity imaging. Environmental science & technology 49, 5593–5600. doi:10.1021/acs.est.5b00208.

Guedes, V.J.C.B., Maciel, S.T.R., Rocha, M.P., 2022. Refrapy: A python program for seismic refraction data analysis. Computers & Geosciences 159, 105020. URL: https://www.sciencedirect.com/science/article/pii/S0098300421003022, doi:https://doi.org/10.1016/j.cageo.2021.105020.

Hagedoorn, J., 1959. The plus-minus method of interpreting seismic refraction sections. Geophysical prospecting 7, 158–182. doi:10.1111/j.1365-2478.1959.tb01460.x.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. URL: https://doi.org/10.1038/s41586-020-2649-2, doi:10.1038/s41586-020-2649-2.

Heimann, S., Kriegerowski, M., Isken, M., Cesca, S., Daout, S., Grigoli, F., Juretzek, C., Megies, T., Nooshiri, N., Steinberg, A., Sudhaus, H., Vasyura-Bathke, H., Willey, T., Dahm, T., 2017. Pyrocko - an open-source seismology toolbox and library doi:10.5880/GFZ.2.1.2017.001.

Hilbich, C., 2010. Time-lapse refraction seismic tomography for the detection of ground ice degradation. The Cryosphere 4, 243–259. doi:10.5194/tc-4-243-2010.

Hunter, J.D., 2007. Matplotlib: A 2d graphics environment. Computing in Science & Engineering 9, 90–95. doi:10.1109/MCSE.2007.55.

Krautblatter, M., Draebing, D., 2014. Pseudo 3-dp wave refraction seismic monitoring of permafrost in steep unstable bedrock. Journal of Geophysical Research: Earth Surface 119, 287–299. doi:10.1002/2012jf002638.

Lawton, D.C., 1989. Computation of refraction static corrections using first-break traveltime differences. Geophysics 54, 1289–1296.

McKinney, W., 2010. Data Structures for Statistical Computing in Python, in: Stéfan van der Walt, Jarrod Millman (Eds.), Proceedings of the 9th Python in Science Conference, pp. 56 – 61. doi:10.25080/Majora-92bf1922-00a.

Nguyen, F., Ghose, R., Isunza Manrique, I., Robert, T., Dumont, G., 2018. Managing past landfills for future site development: A review of the contribution of geophysical methods, in: Proceedings of the 4th International Symposium on Enhanced Landfill Mining, pp. 27–36.

Parsekian, A., Singha, K., Minsley, B.J., Holbrook, W.S., Slater, L., 2015. Multiscale geophysical imaging of the critical zone. Reviews of Geophysics 53, 1–26. doi:10.1002/2014RG000465.

Plattner, A.M., 2020. Gprpy: Open-source ground-penetrating radar processing and visualization software. The Leading Edge 39, 332–337. doi:10.1190/tle39050332.1.

Ramachandran, K., Bellefleur, G., Brent, T., Riedel, M., Dallimore, S., 2011. Imaging permafrost velocity structure using high resolution 3d seismic tomography. Geophysics 76, B187–B198. doi:10.1190/geo2010-0353.1.

Ringler, A.T., Evans, J.R., 2015. A quick seed tutorial. Seismological Research Letters 86, 1717–1725. doi:10.1785/0220150043.

Romero-Ruiz, A., Linde, N., Keller, T., Or, D., 2018. A review of geophysical methods for soil structure characterization. Reviews of Geophysics 56, 672–697. doi:10.1029/2018RG000611.

Ronczka, M., Wisén, R., Dahlin, T., 2018. Geophysical pre-investigation for a stockholm tunnel project: joint inversion and interpretation of geoelectric and seismic refraction data in an urban environment. Near Surface Geophysics 16, 258–268. doi:10.3997/1873-0604.2018009.

Rücker, C., Günther, T., Wagner, F.M., 2017. pygimli: An open-source library for modelling and inversion in geophysics. Computers & Geosciences 109, 106–123. doi:10.1016/j.cageo.2017.07.011.

Samyn, K., Travelletti, J., Bitri, A., Grandjean, G., Malet, J.P., 2012. Characterization of a landslide geometry using 3d seismic refraction traveltime tomography: The la valette landslide case history. Journal of Applied Geophysics 86, 120–132. doi:10.1016/j.jappgeo.2012.07.014.

Steiner, M., Katona, T., Fellner, J., Flores Orozco, A., 2022. Quantitative water content estimation in landfills through joint inversion of seismic refraction and electrical resistivity data considering surface conduction. Waste Management 149, 21–32. URL: https://www.sciencedirect.com/science/article/pii/S0956053X22002793, doi:https://doi.org/10.1016/j.wasman.2022.05.020.

Steiner, M., Wagner, F.M., Maierhofer, T., Schöner, W., Flores Orozco, A., 2021. Improved estimation of ice and water contents in alpine permafrost through constrained petrophysical joint inversion: The hoher sonnblick case study. Geophysics 86, 1–84. doi:10.1190/geo2020-0592.1.

Stockwell, J.W., 1999. The cwp/su: Seismic unix package. Computers & Geosciences 25, 415–419. URL: https://www.sciencedirect.com/science/article/pii/S0098300498001459, doi:10.1016/S0098-3004(98)00145-9.

Sullivan, C.B., Kaszynski, A., 2019. PyVista: 3d plotting and mesh analysis through a streamlined interface for the visualization toolkit (VTK). Journal of Open Source Software 4, 1450. URL: https://doi.org/10.21105/joss.01450, doi:10.21105/joss.01450.

Uhlemann, S., Hagedorn, S., Dashwood, B., Maurer, H., Gunn, D., Dijkstra, T., Chambers, J., 2016. Landslide characterization using p- and s-wave seismic refraction tomography — the importance of elastic moduli. Journal of Applied Geophysics 134, 64–76. URL: https://www.

sciencedirect.com/science/article/pii/S0926985116302439, doi:https://doi.org/10.1016/j.jappgeo.2016.08.014.

Uieda, L., Oliveira Jr, V.C., Barbosa, V.C., 2013. Modeling the earth with fatiando a terra, in: Proceedings of the 12th Python in Science Conference, pp. 96–103.

Wang, W., Cheadle, S., 1995. Branch point analysis in refraction interpretation, in: CSEG National Convention, Expanded abstracts.

Whiteley, J., Chambers, J., Uhlemann, S., Boyd, J., Cimpoiasu, M., Holmes, J., Inauen, C., Watlet, A., Hawley-Sibbett, L., Sujitapan, C., et al., 2020. Landslide monitoring using seismic refraction tomography–the importance of incorporating topographic variations. Engineering Geology 268, 105525. doi:10.1016/j.enggeo.2020.105525.

List of Figures

1 General architecture of the formikoj library comprising a utility, modeling and processing module. The base classes `DataModeler` and `MethodManager` can be used to build tools for specific geophysical methods, e.g., seismic refraction. . . . 29

2 The `SeismicRefractionManager` addresses the stations through consecutive station numbers based on the sort order in the geometry file. The shot index numbers (SIN) and receiver index numbers (RIN) are assigned to the shot and receiver stations separately to allow for an intuitive data handling for the user. . . . 30

3 The seismogram plot presents the currently selected traces along the x axis, where the sort order is determined through the geometry. The corresponding trace data are illustrated as a function of time along the y axis (solid curves), with positive and negative amplitudes depicted in blue and red, respectively. The selection criterion and the applied filter are shown in the upper left corner of the plot. Green crosses refer to the picked traveltimes stored in the currently active pickset. . . . 31

4 The status bar in the interactive seismogram plot displays the currently active processing and data scaling modes as well as the time (in seconds) at the current cursor position. By pressing the keys 'a', 'm', 'r', 'v' on the keyboard the different modes can be activated. . . . 32

5 The frequency spectrum illustrates the frequency content of the currently selected traces, which allows for the identification of frequency ranges associated to noise, which can be omitted through the application of corresponding frequency filtering. . . . 33

6 The traveltime diagram shows the first break traveltimes stored in the currently active pickset along the y axis (x symbols), where solid lines connect traveltimes assigned to a common shot. The sort order of the stations along the x axis reflects the geometry. Filled circles indicate stations with co-located shot and geophone (receiver), triangles refer to receiver stations (no shot) and stars refer to shot stations (no geophone). . . . 34

7 Synthetic seismic waveform data without random or systematic errors created with the `SeismicWaveformModeler()` class for a shot position in the center of the survey layout. . . . 35

8 Synthetic seismic waveform data with added noise created with the `SeismicWaveformModeler()` class for a shot position in the center of the survey layout. The random noise refers to an offset dependent decrease of the signal-to-noise ratio, while the systematic broken geophones and wrong polarity are systematic errors. . . . 36

9 Synthetic seismic datasets used for the performance evaluation of automatic first break picking algorithms. The scatter plot illustrates the correlation between the synthetic traveltimes and the traveltimes automatically picked in the synthetic seismic waveforms (after outlier removal based on the histogram). 37

10 Synthetic seismic datasets used for the evaluation of an implementation of the Plus-Minus method (Hagedoorn, 1959) based on a two-layered subsurface model without topography. The horizontal solid line illustrates the reconstructed interface depth; true and estimated seismic velocities are superimposed on the respective layers. . . . 38

11 The Danube island field dataset was collected along a single line with a roll-along survey layout; the filled triangles indicate the direction of the measurements. The five segments have an overlap of 50% to ensure an adequate data coverage along the entire profile. . . . 39

12 The common offset stack computed for the Danube island dataset clearly showing a two-layered subsurface. The seismic velocities within the layer can be estimated from the gradient of the lines drawn along the corresponding first onsets of the seismic waves. . . . 40

13 Examplary seismic waveforms from the Danube island dataset shown for shot index number (SIN) 99 with a 120 Hz lowpass filter applied to suppress high frequency noise. The receiver index numbers (RIN) start at 73, thus indicating that the data were collected with a roll-along survey geometry. . . . 41

14 Pseudosection showing the apparent seismic velocities determined from the first break traveltimes obtained from the Danube island dataset and the corresponding absolute offset between the shot and receiver stations. The apparent velocity for each shot-receiver pair is illustrated at the corresponding midpoint and pseudodepth (1/3 of the absolute offset). . . . 42

15 Picking percentage for each shot in the Danube island roll-along dataset. Low values in the first third of the dataset indicate a low signal-to-noise ratio in the seismograms hindering the picking of first break traveltimes for each shot-receiver pair. 43
16 The soda lakes field dataset was collected in a 3D survey layout with stations (co-located geophones and shots indicated by filled dots) deployed in form of a cross. Additional shots (yellow stars) were conducted in from of a cross rotated by 45° to increase the coverage of the dataset. 44
17 Examplary seismic waveforms from the soda lakes dataset shown for shot index number (SIN) 1 with a 100 Hz lowpass filter applied to suppress high frequency noise. The recorded seismic waveforms clearly reflect the geometry of the geophones with RIN 1 to 48 deployed along the direction of wave propagation, while RIN 49 to 96 are deployed perpendicular to the propagating wavefront. 45
18 3D pseudosection showing the apparent seismic velocities determined from the first break traveltimes obtained from the soda lakes dataset and the corresponding absolute offset between the shot and receiver stations. The apparent velocity for each shot-receiver pair is illustrated at the corresponding 2D midpoint and pseudodepth (1/3 of the absolute offset), thus yielding a 3D representation. 46
19 Picking percentage for each shot in the soda lakes dataset indicating a high signal-to-noise ratio allowing for the picking of first break traveltimes for nearly all shot-receiver pairs. 47

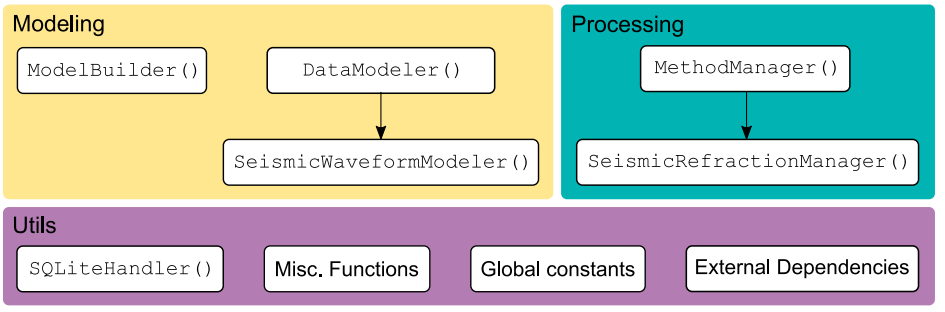

Figure 1: General architecture of the formikoj library comprising a utility, modeling and processing module. The base classes `DataModeler` and `MethodManager` can be used to build tools for specific geophysical methods, e.g., seismic refraction.

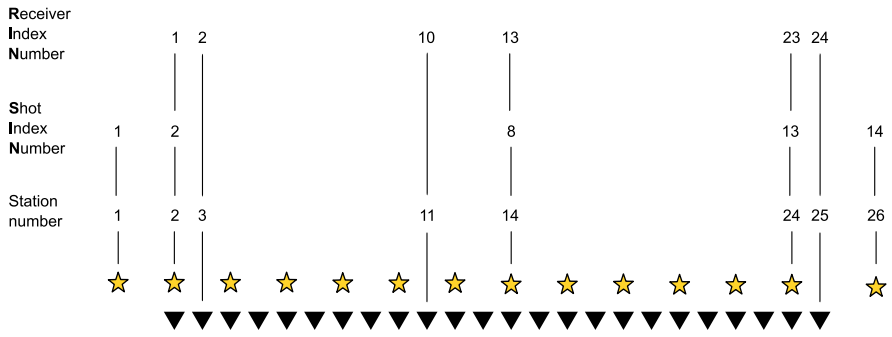

Figure 2: The `SeismicRefractionManager` addresses the stations through consecutive station numbers based on the sort order in the geometry file. The shot index numbers (SIN) and receiver index numbers (RIN) are assigned to the shot and receiver stations separately to allow for an intuitive data handling for the user.

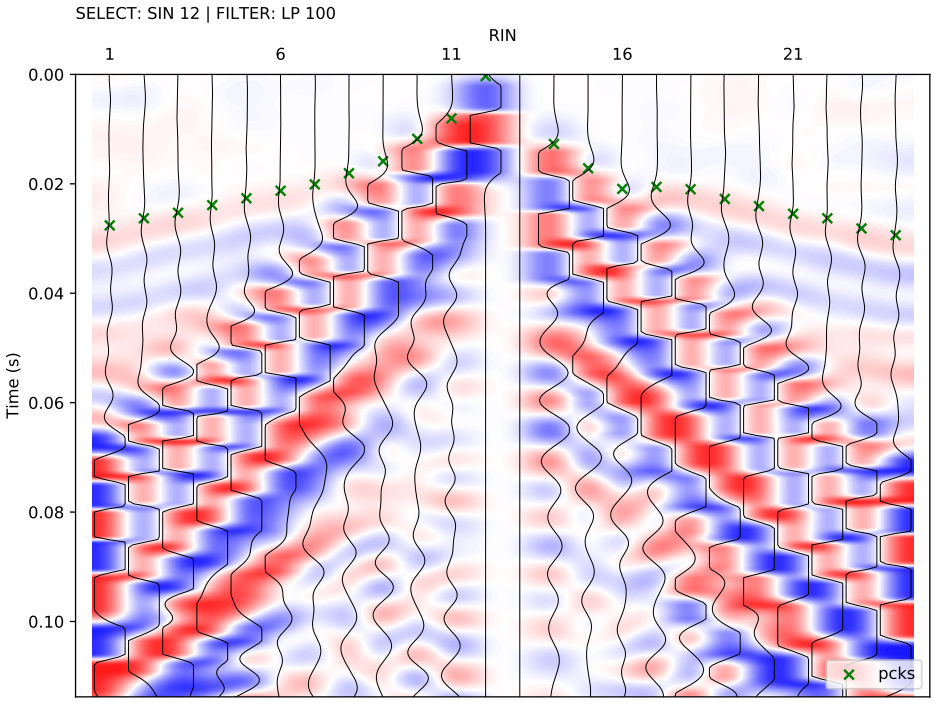

Figure 3: The seismogram plot presents the currently selected traces along the x axis, where the sort order is determined through the geometry. The corresponding trace data are illustrated as a function of time along the y axis (solid curves), with positive and negative amplitudes depicted in blue and red, respectively. The selection criterion and the applied filter are shown in the upper left corner of the plot. Green crosses refer to the picked traveltimes stored in the currently active pickset.

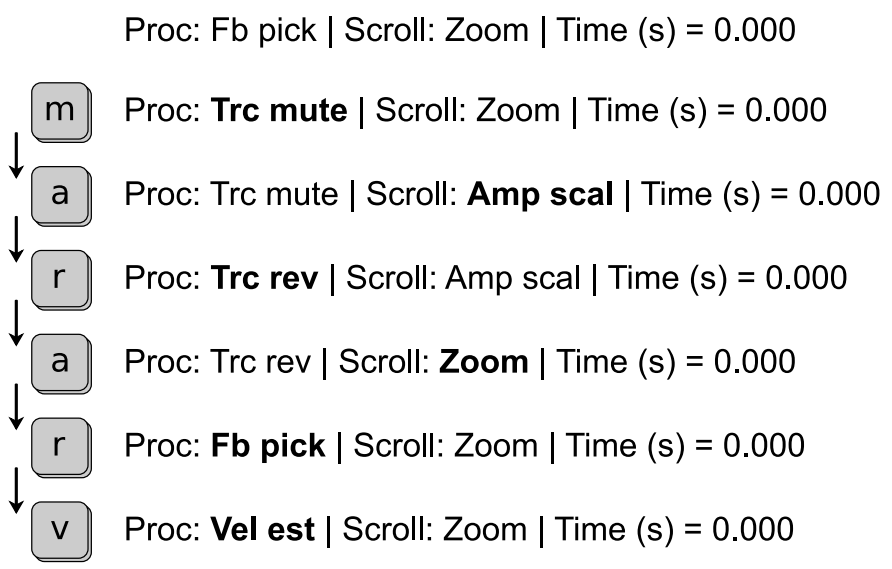

Figure 4: The status bar in the interactive seismogram plot displays the currently active processing and data scaling modes as well as the time (in seconds) at the current cursor position. By pressing the keys 'a', 'm', 'r', 'v' on the keyboard the different modes can be activated.

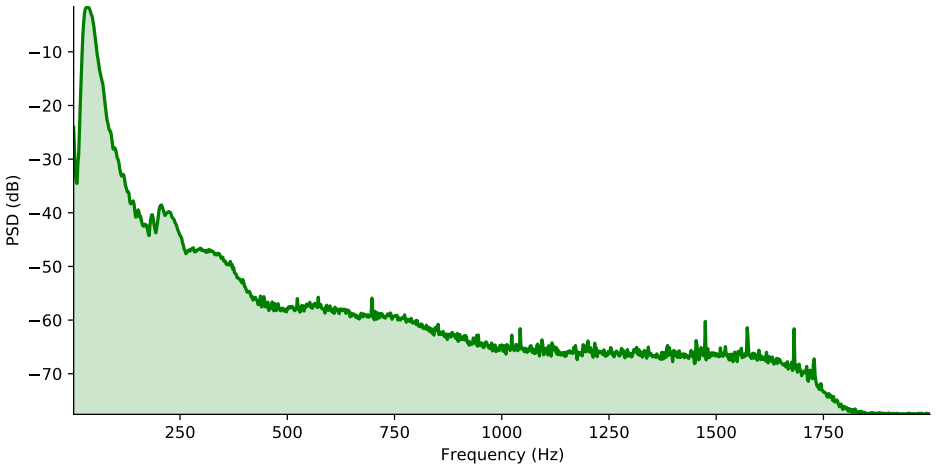

Figure 5: The frequency spectrum illustrates the frequency content of the currently selected traces, which allows for the identification of frequency ranges associated to noise, which can be omitted through the application of corresponding frequency filtering.

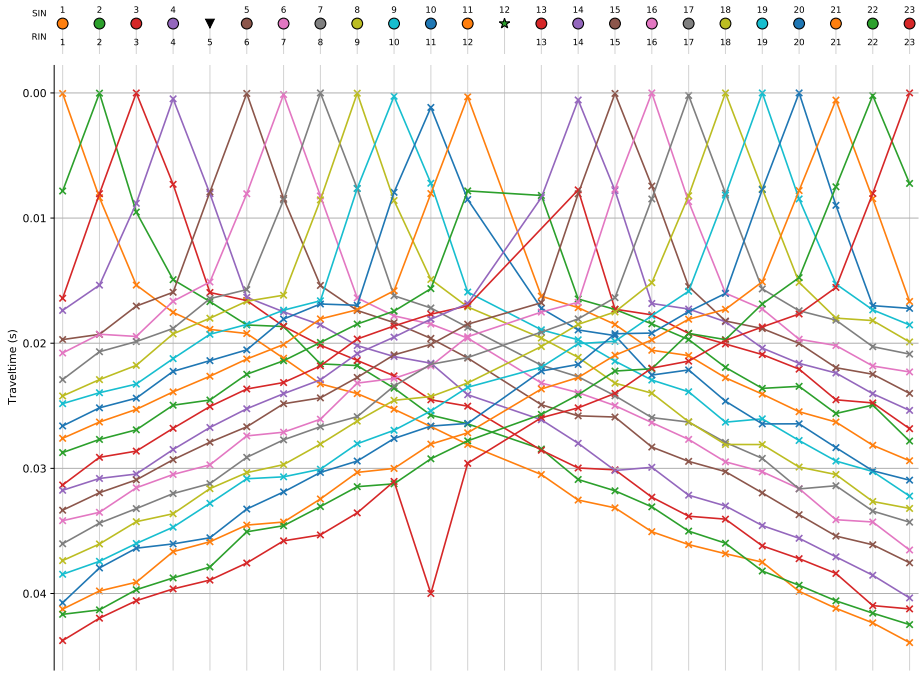

Figure 6: The traveltime diagram shows the first break traveltimes stored in the currently active pickset along the y axis (x symbols), where solid lines connect traveltimes assigned to a common shot. The sort order of the stations along the x axis reflects the geometry. Filled circles indicate stations with co-located shot and geophone (receiver), triangles refer to receiver stations (no shot) and stars refer to shot stations (no geophone).

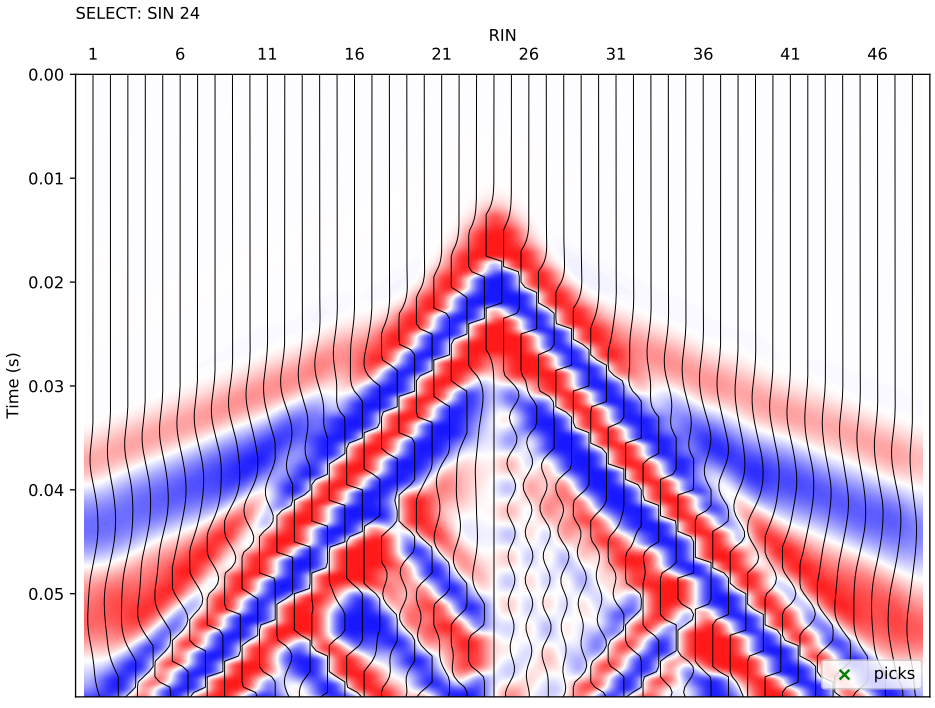

Figure 7: Synthetic seismic waveform data without random or systematic errors created with the `SeismicWaveformModeler()` class for a shot position in the center of the survey layout.

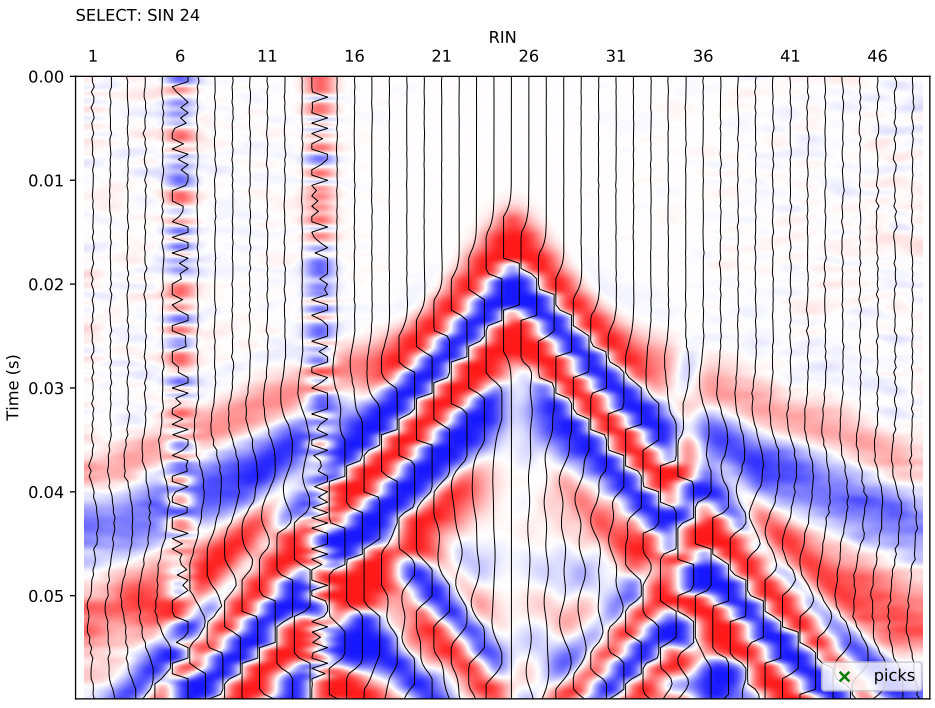

Figure 8: Synthetic seismic waveform data with added noise created with the `SeismicWaveformModeler()` class for a shot position in the center of the survey layout. The random noise refers to an offset dependent decrease of the signal-to-noise ratio, while the systematic broken geophones and wrong polarity are systematic errors.

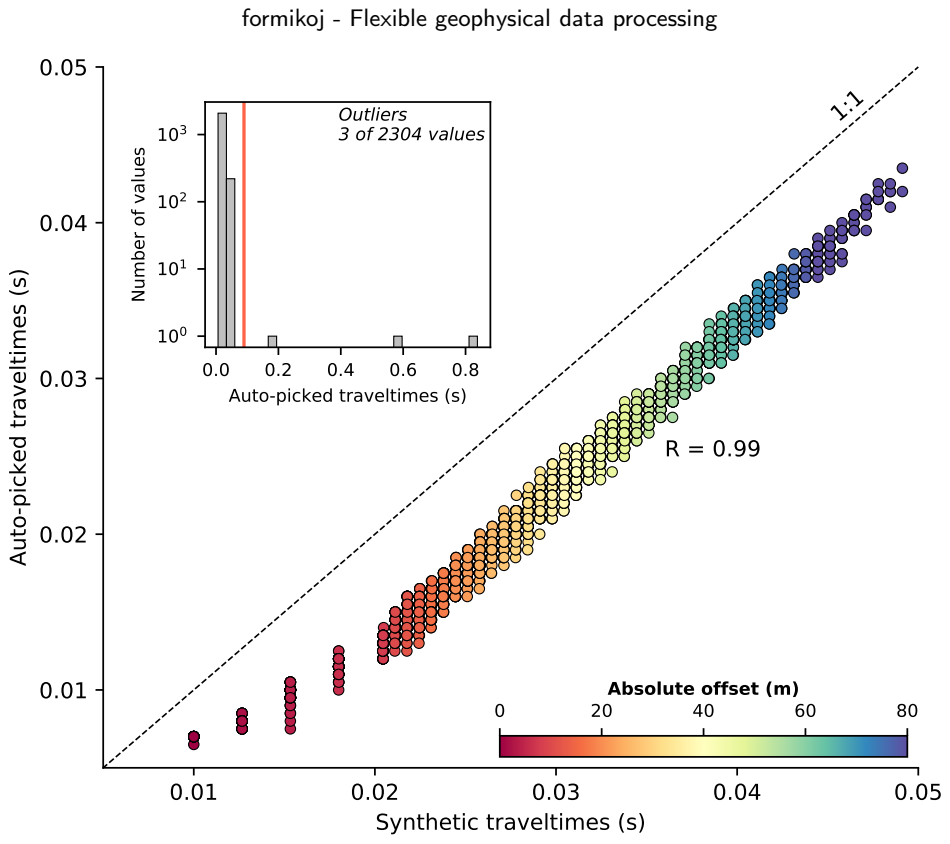

Figure 9: Synthetic seismic datasets used for the performance evaluation of automatic first break picking algorithms. The scatter plot illustrates the correlation between the synthetic traveltimes and the traveltimes automatically picked in the synthetic seismic waveforms (after outlier removal based on the histogram).

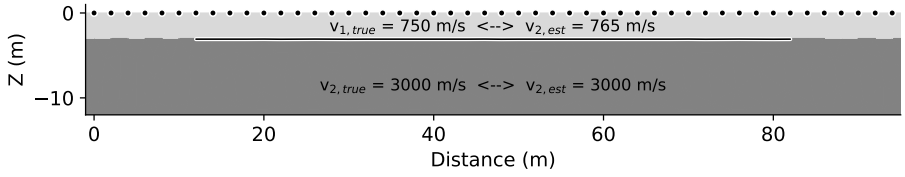

Figure 10: Synthetic seismic datasets used for the evaluation of an implementation of the Plus-Minus method (Hagedoorn, 1959) based on a two-layered subsurface model without topography. The horizontal solid line illustrates the reconstructed interface depth; true and estimated seismic velocities are superimposed on the respective layers.

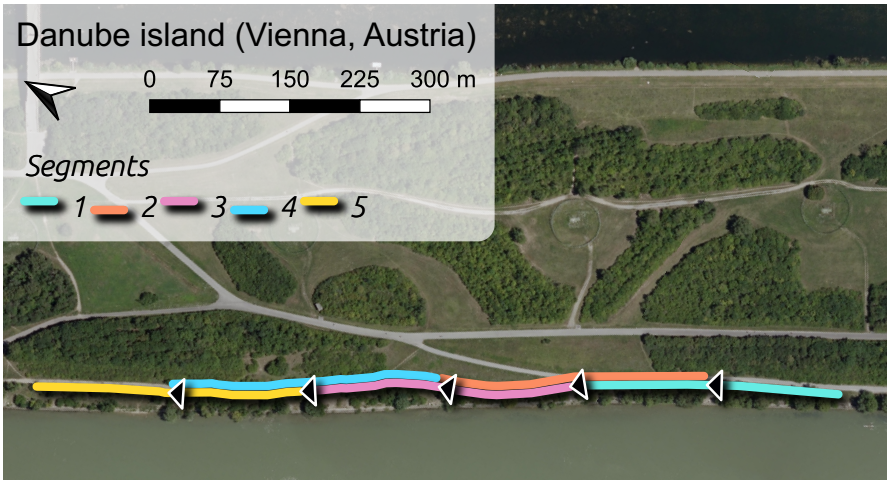

Figure 11: The Danube island field dataset was collected along a single line with a roll-along survey layout; the filled triangles indicate the direction of the measurements. The five segments have an overlap of 50% to ensure an adequate data coverage along the entire profile.

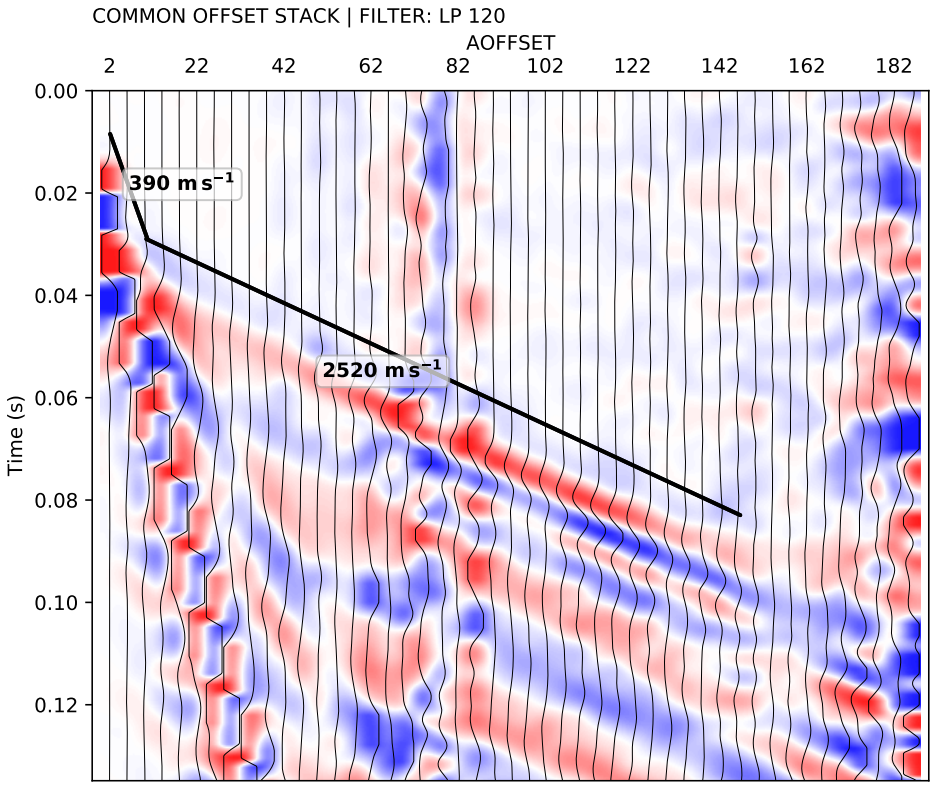

Figure 12: The common offset stack computed for the Danube island dataset clearly showing a two-layered subsurface. The seismic velocities within the layer can be estimated from the gradient of the lines drawn along the corresponding first onsets of the seismic waves.

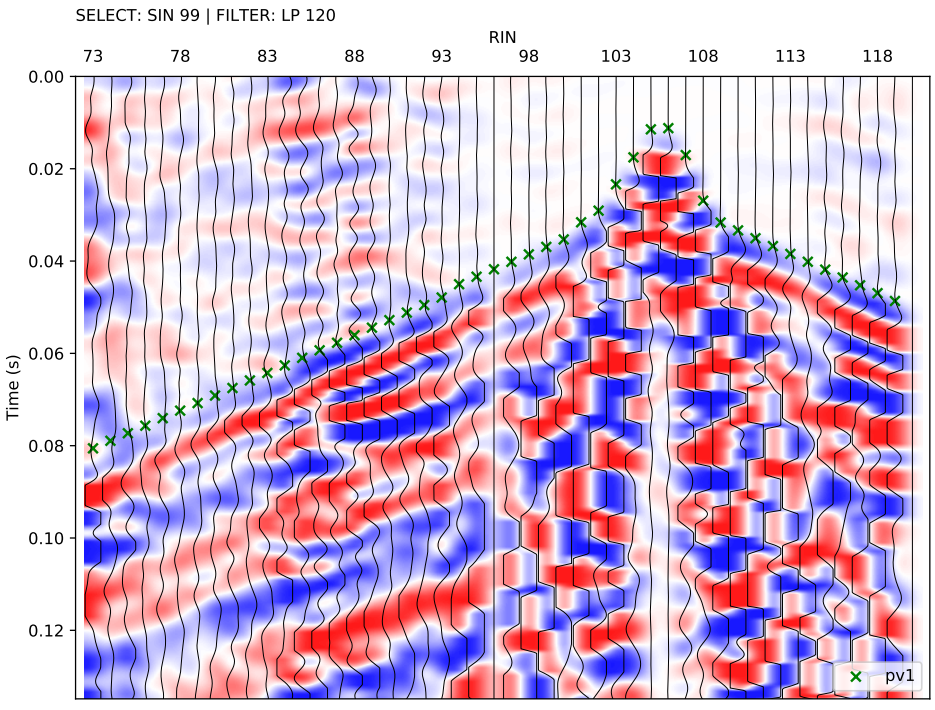

Figure 13: Examplary seismic waveforms from the Danube island dataset shown for shot index number (SIN) 99 with a 120 Hz lowpass filter applied to suppress high frequency noise. The receiver index numbers (RIN) start at 73, thus indicating that the data were collected with a roll-along survey geometry.

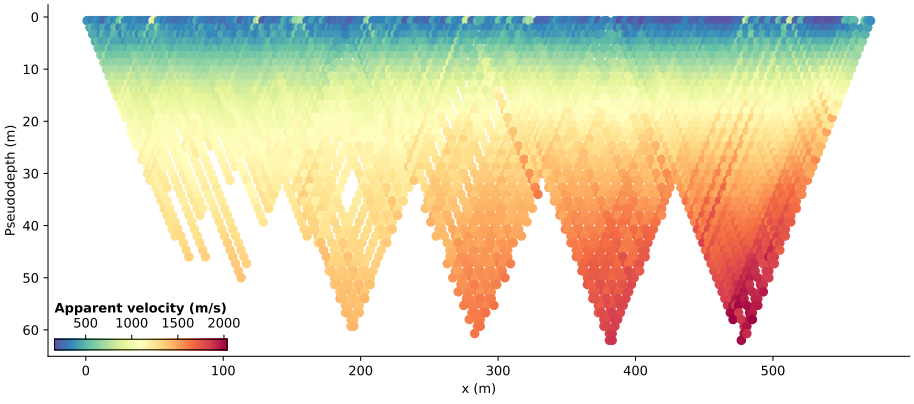

Figure 14: Pseudosection showing the apparent seismic velocities determined from the first break traveltimes obtained from the Danube island dataset and the corresponding absolute offset between the shot and receiver stations. The apparent velocity for each shot-receiver pair is illustrated at the corresponding midpoint and pseudodepth (1/3 of the absolute offset).

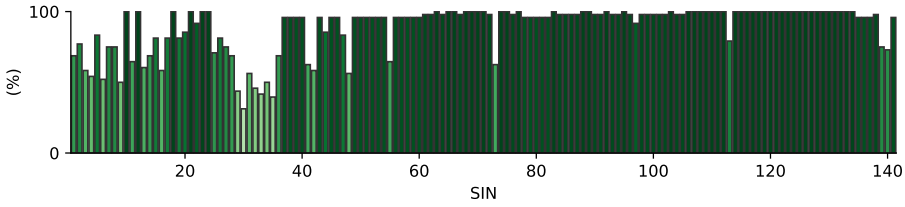

Figure 15: Picking percentage for each shot in the Danube island roll-along dataset. Low values in the first third of the dataset indicate a low signal-to-noise ratio in the seismograms hindering the picking of first break traveltimes for each shot-receiver pair.

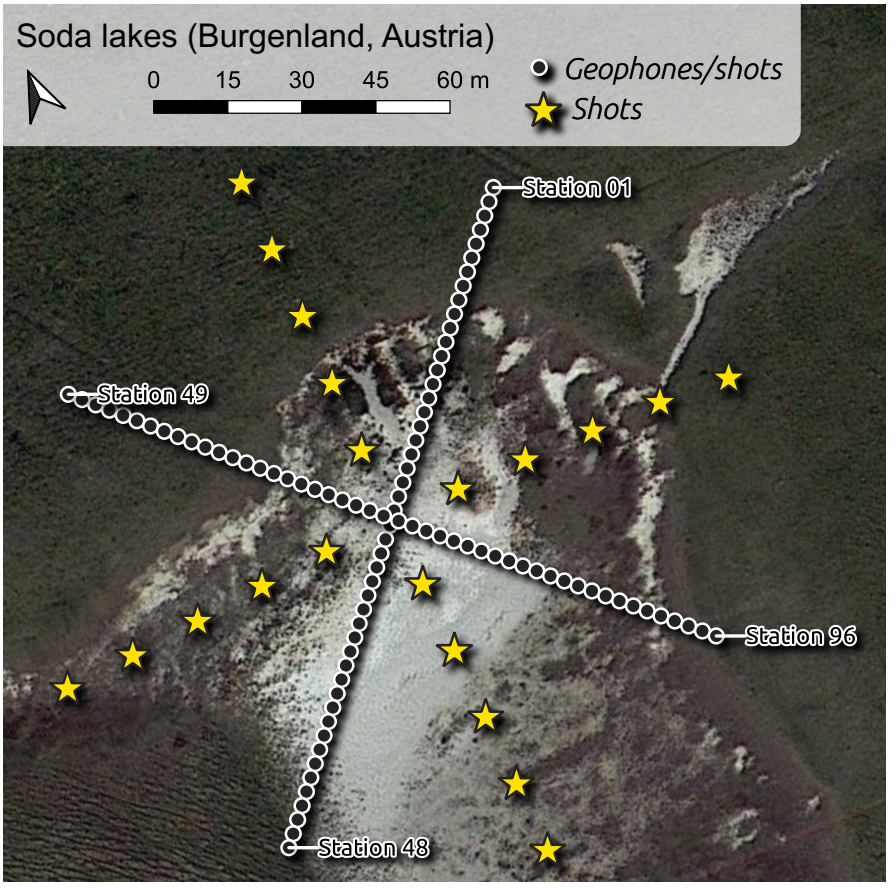

Figure 16: The soda lakes field dataset was collected in a 3D survey layout with stations (co-located geophones and shots indicated by filled dots) deployed in form of a cross. Additional shots (yellow stars) were conducted in from of a cross rotated by 45° to increase the coverage of the dataset.

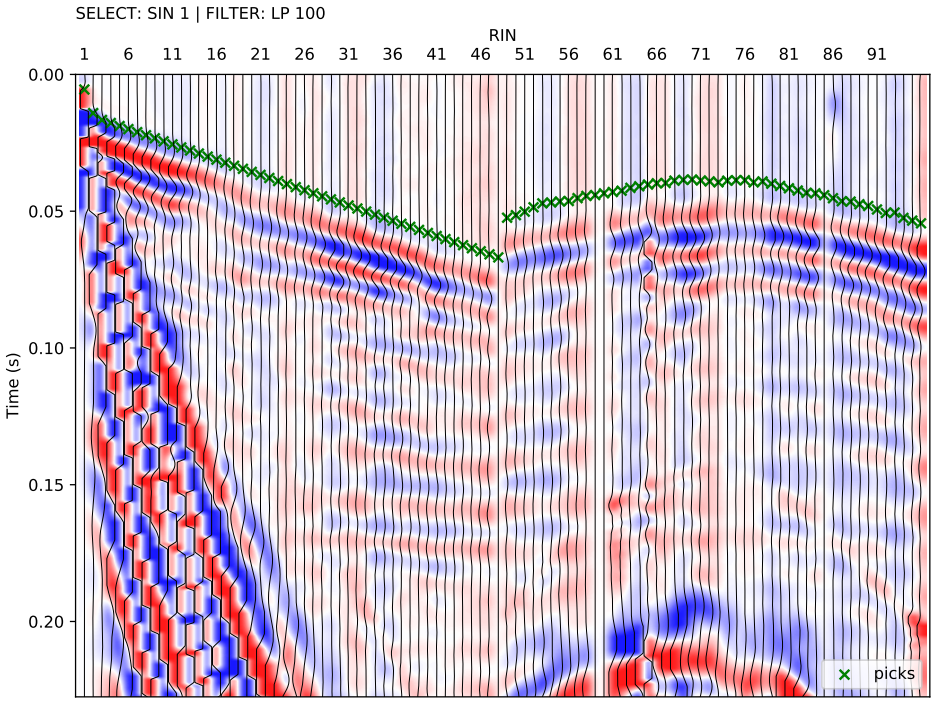

Figure 17: Examplary seismic waveforms from the soda lakes dataset shown for shot index number (SIN) 1 with a 100 Hz lowpass filter applied to suppress high frequency noise. The recorded seismic waveforms clearly reflect the geometry of the geophones with RIN 1 to 48 deployed along the direction of wave propagation, while RIN 49 to 96 are deployed perpendicular to the propagating wavefront.

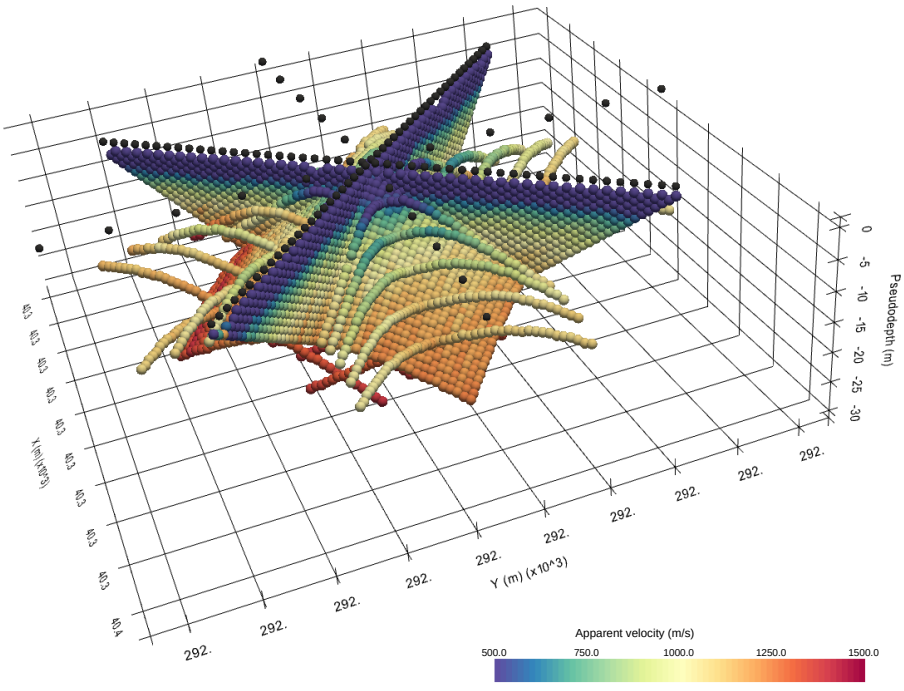

Figure 18: 3D pseudosection showing the apparent seismic velocities determined from the first break traveltimes obtained from the soda lakes dataset and the corresponding absolute offset between the shot and receiver stations. The apparent velocity for each shot-receiver pair is illustrated at the corresponding 2D midpoint and pseudodepth (1/3 of the absolute offset), thus yielding a 3D representation.

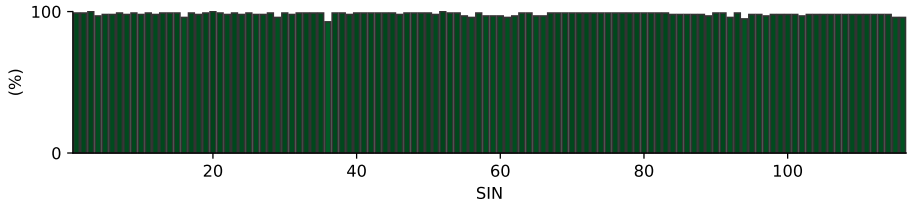

Figure 19: Picking percentage for each shot in the soda lakes dataset indicating a high signal-to-noise ratio allowing for the picking of first break traveltimes for nearly all shot-receiver pairs.