



Converting scanned images of seismic reflection data into SEG-Y format

Daniel Sopher¹ 

Received: 18 July 2017 / Accepted: 24 October 2017 / Published online: 13 November 2017
© The Author(s) 2017. This article is an open access publication

Abstract Archives across the world contain vast amounts of old or “vintage” seismic reflection data, which are largely inaccessible for geo-scientific research, due to the out-dated media on which they are stored. Despite the age of these data, they often have great potential to be of use in modern day research. It is often the case that seismic reflection data within these archives are only available as a processed stacked section, printed on paper or film. In this study, a method for the conversion (vectorization) of scanned images of stacked reflection seismic data to standard SEG-Y format is presented. The method addresses data displayed with a line denoting the waveform, where areas on one side of the baseline are shaded (i.e. wiggle trace, variable fill). The method provides an improvement on other published methods utilized within currently available academic software. Unlike previous studies, the method used to detect trace baselines and to detect and remove timelines on the seismic image is described in detail. Furthermore, a quantitative analysis of the performance of the method is presented, showing that an average trace-to-trace correlation coefficient of between 0.8 and 0.95 can be achieved for typical plotting styles. Finally, a case study where the method is applied to vectorize over 1700 km of land seismic data from the island of Gotland (Sweden) is presented.

Keywords Vectorization / vectorisation · Digitization / digitisation · Vintage / historical data · Analogue to digital · OPAB dataset · TIFF to SEGY

Communicated by: H. A. Babaie

 Daniel Sopher
daniel.sopher@geo.uu.se

¹ Department of Earth Sciences, Uppsala University, Villavägen 16, SE-75236 Uppsala, Sweden

Introduction

Since the middle of the last century seismic reflection surveying has undergone rapid development and has become, arguably, the foremost method for imaging the structure of the Earth’s crust. Reflection seismic surveys are currently widely used in a range of disciplines in academia and industry (primarily hydrocarbon exploration), with many millions of kilometres of data acquired. Despite its extensive use, acquisition of new seismic reflection data is costly, especially offshore, where the operation of a seismic vessel can cost thousands to tens of thousands of euros per day, depending upon its size. These high costs can be prohibitive to the acquisition of new data, especially in academic studies. Over the past 70 years, dramatic advances in computing power and memory have also taken place, for example, the cost of digital data storage is currently lower than it was in 1980 by a factor of 10^7 (Griffin 2015). A consequence of these rapid developments in computer technology is that huge amounts of old or “vintage” seismic reflection data exist which are relatively inaccessible for use, largely due to their storage media. The presence of these large archives of historical data present a huge opportunity to the scientific community, for the following reasons:

1. It is often the case that reprocessing of historical data can yield an improved final image of the subsurface and hence, provide additional information, compared to the image obtained using the original processing flow.
2. By visualizing and interpreting historical data in modern software it is often possible to construct an improved/new interpretation of the data.
3. The availability of historical data can help to optimise the location and parameters of future data acquisition efforts.
4. When dealing with industry data, which has recently been de-classified and made available for academic study, it

can be the case that few or no results from the data have been published in scientific literature.

However, despite their value, data within these historical archives are under threat of being lost (Diviacco et al. 2015; Griffin 2015). Large amounts of vintage data are stored on outdated media, such as magnetic tapes. In many cases, no digital record of the data remains whatsoever and the seismic reflection data are only available as images of the final processed section, printed on film or paper. This makes the data relatively inaccessible for modern use, as specialist equipment and processes are required to read these data. If all of the necessary equipment can be found, the process can be time consuming and is often complicated by missing information such as tape labels, geometry information or acquisition reports. This makes the task of transcribing data from these old media expensive and hence presents those wishing to rescue their archives with a significant financial hurdle. Another important issue is the degradation of these storage media with time, for example “sticky shed syndrome” can occur where the tapes becomes sticky due to the humidity in the atmosphere, and hence are no longer readable (Ross and Gow 1999; Diviacco et al. 2015). Similarly paper sheets with images of stacked sections can become delicate and faded, making them difficult to scan. As a result, large volumes of data are at risk of becoming unreadable in the near future. Another risk to these data is that they may simply be discarded by their owners who no longer wish to pay to house the data (or simply cannot afford to).

This study focuses on the problem of vectorization or digitization of stacked reflection seismic data, which are only available as images printed on paper/film and presents a process for performing this task. Vectorization is defined here as the process of digitizing the amplitude vs. time information on each trace in the stacked seismic reflection data and outputting these data in a conventional digital format (SEG-Y). Furthermore, this study focuses on the problem of vectorizing variable fill wiggle trace data, i.e. where the waveform of the seismic trace is represented as a line on the image and deviations on one side of a chosen baseline are shaded (Fig. 1). The method presented in this study, initially reads in a binary image file generated by scanning the original hardcopy of the data. By applying a series of image processing procedures, the timelines on the stacked section and baselines of the seismic traces are then detected. The waveform of each trace is then extracted and the stacked seismic data are output as a SEG-Y file. Several other studies have presented methods for performing this task (Blake and Hewlett 1993; Miles et al. 2007; Diviacco et al. 2015; Owen et al. 2015 etc.). This study builds on previous work, but differs in the following ways: 1) unlike previous studies, a detailed description of the method for the detection of timelines and trace baselines is given. 2) The method presented here is an improvement upon those used by other freely available vectorization software packages (e.g. Miles et al. 2007). 3)

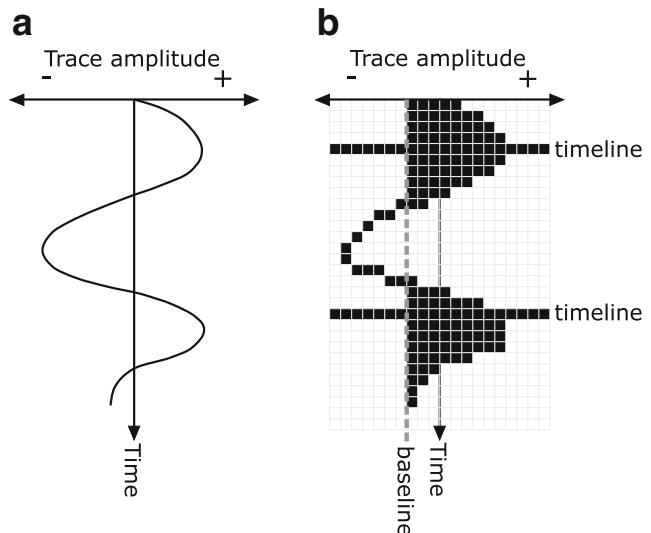


Fig. 1 A schematic diagram of a seismic trace. **a** A stacked seismic trace can be considered as a series of data values representing seismic amplitude vs. time. In this case the time series is plotted with amplitude as the x-axis and time as the y-axis. **b** Shows the same seismic trace as in a) after it has been plotted on paper using a variable fill wiggle format, scanned and saved as a raster format. Note that the image is now represented by pixels and that the area between the curve and a vertical line (baseline) is shaded. Furthermore, note that the baseline may not necessarily be located at the zero amplitude value. The gap between the y-axis and the baseline is sometimes referred to as the bias. For stacked seismic data it is common to display many traces side-by-side on the section, to describe the variation of the seismic response with surface location (Fig. 2)

Unlike other studies, the performance of the method in this study is assessed quantitatively. 4) A case study detailing the vectorization of over 1700 km of land data from the island of Gotland is presented. The script to perform the vectorization process for this study was written in Matlab and can be obtained for use in academic studies by contacting the author. Therefore, others can easily utilize the method described here to preserve and extract valuable information from large seismic datasets, which are currently at risk of being lost.

Background

Vectorization of stacked seismic reflection data

Seismic vectorization is a common practice in the hydrocarbon industry and there are many companies who specialize in the process of scanning and digitizing seismic data stored on paper sections. However, the costs associated with these services are tailored to the hydrocarbon industry and hence, are often prohibitively high for academic institutions. The methodologies utilized by these commercial vectorization companies are largely unpublished. Two US patents have been published which describe some aspects of two commercial vectorization algorithms (Chevion and Navon 2012, 2014). The first (Chevion

and Navon 2012) operates by translating the density of the waveform lines (wiggles) on the section into a grayscale image, which approximates the amplitude at all locations on the section. The amplitude information for each trace can then be extracted from the grayscale image at the baseline locations. In the second method (Chevion and Navon 2014), the image is split into two components. One component contains only the negative parts of the waveform, represented only by a line. All of the curves on this component are then tracked and associated with an appropriate baseline, based on a set of logical rules. The second component, containing only the positive amplitudes, is then used to fill in the missing parts of the trace information. Several articles describe case studies where data have been vectorized using commercial software and used for interpretation, as well as input into for seismic inversion (Al Mahdy and Sedek 2013; Owen et al. 2015).

Within academia and the public domain, several articles describe methodologies for vectorizing stacked images of reflection seismic data and a number of scripts / programs are available on-line. A widely used piece of software is IMAGE2SEGY (Farran 2008). However, this script is designed for the vectorization of seismic data displayed in variable density or colour and therefore is not well suited for digitizing variable fill wiggle trace seismic images. Other software such as Tif2seggy exists and utilizes Seismic Unix and Netpbm, however there is relatively little information available describing how the method works. Blake and Hewlett (1993) present a high level description of a method for vectorizing wiggle trace seismic images. The software SEISTRANS was developed by Caldera graphics as part of The European SEISCAN and SEISCANEX projects (Miles et al. 2007). One objective of these projects was to develop relatively cheap software, which could be used by academic and governmental institutions to vectorize their legacy data. The method described by Miles et al. (2007) begins by detecting trace baselines based on inputs defined by the user. Once the baselines have been defined, the program scans down each trace on the image and sums the amount of black pixels between the baseline and the adjacent trace. This is equivalent to extracting the horizontal thickness of the variable fill vs. time for the trace. This time vs. amplitude series is then bandpass filtered and the mean average value of the series subtracted, which gives an estimate of the seismic trace. A downside of this method is that only the amplitudes on one side of the baseline are considered, i.e. the sections of the trace represented only by a line are not included in the trace estimation process. Cooke and Bulat (2012) and Diviacco et al. (2015) both present case studies of vectorization projects where the SEISTRANS software was used. It should also be noted that a description of the process used to detect the baselines and timelines on the scanned image is not published as part of any of the aforementioned scripts or studies.

Limitations of vectorization

Although vectorization can be useful to extract valuable information from scanned images of seismic reflection data, there are a number of fundamental limitations and problems with the process, which are addressed in this section. Fig. 2a displays a more-or-less ideal dataset for vectorization, where it would be possible to recover the data from the image, almost perfectly. However, it is often the case that data of this type are displayed differently, reducing the accuracy of any vectorization efforts. Fig. 2b, c display data which exhibit some typical problems which reduce the accuracy of the data reconstruction and complicate the image processing problem. The most significant of these problems is that the waveform information is simply missing or is unrecoverable for parts of the trace. For example, it is common to display data with a certain amount of trace overlap, hence, in areas of high amplitude the shaded area of the traces will overlap entirely with the adjacent traces. The waveform information within these parts of the image is therefore not recoverable. Similarly, in some cases no line is used to display the data and only the variable shaded part of the trace is displayed (Fig. 2c). In these cases, waveform information from the negative amplitudes is not recoverable from the image. Annotations on the image such as timelines or interpretations which have been drawn on the section before scanning also obscure parts of the data in the image, making it unrecoverable. In an ideal case, the baselines of each trace are perfectly straight, equally spaced and perpendicular to the timelines on the image. This greatly simplifies the problem of locating each trace on the image accurately. However, it is often the case that the image has been distorted or warped during scanning or printing (Fig. 2b), which makes the process of baseline and timeline detection more challenging. Another important problem when vectorizing any seismic data from an image, is that the resolution of the image and the accuracy of the printing will limit the resolution of the waveform which is recovered from the image.

Vectorization of seismological data

In this section a brief summary of previous seismic vectorization studies within seismology is given. Within the field of seismology digital recording became standard in the 1970s and therefore, the majority of data collected earlier were stored only on paper or film. Kanamori et al. (2010) and An et al. (2015) provide examples of how re-analysis of historical data can provide insight into previous seismic events and improve the understanding of the Earth's structure. A range of initiatives have been set up to digitize and rescue seismological archives from around the world (Okal 2015). Several publications describe methodology and software, which have been developed to digitize historical seismological records (Baskoutas et al. 2000; Pintore et al. 2005; Xu and Xu 2014). Although there are some similarities between the digitization of seismological records and stacked sections of

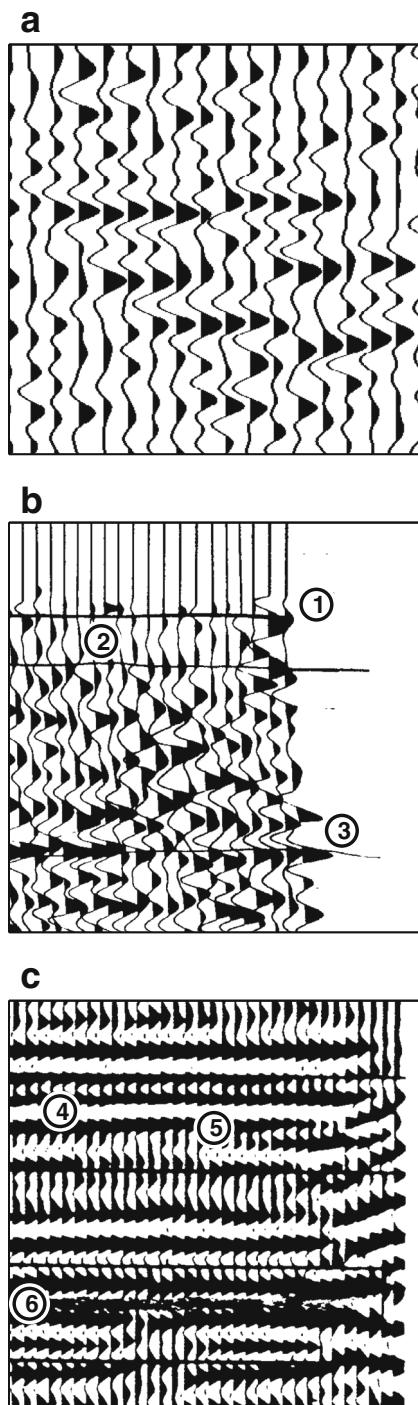


Fig. 2 Images of real seismic reflection data. **a** An ideal image for vectorization. **b** and **c** show examples of problems that make accurate vectorization of the seismic traces challenging. The numbers denote the following features: 1. horizontal warping of the image, which can be seen along the rightmost trace. 2. Vertical warping of the image. 3. Timelines. 4. No line visible in areas of high negative amplitude. 5. Overlap of filled areas when the amplitude is very high. 6. Interpretation that has been hand drawn on the hard copy and has not been removed before scanning

seismic reflection data, these methods are often not directly applicable due to differences in the way the data are displayed. i.e. the waveforms in seismological data are often stored as lines

only, with no variable fill and therefore represent a different image processing problem.

Method

Rotate image

The following sections provide a description of the key steps in the vectorization process (Table 1). The method described in this study is modified from the process presented briefly by Sopher (2016). The input image of the seismic data selected for vectorization should be in a binary format (i.e. only black and white pixels), for example a binary Tagged Image File Format (TIFF) format (Adobe Developers Association 1992). After selecting a range of input parameters three points on the image are specified, which define the vertical and horizontal axes of the image. When specifying these three points the seismic two-way-time of each point is also specified. The seismic-two-way time value for each pixel on the y-axis of the image can then be calculated assuming a linear relationship between pixel number and seismic two-way-time. The area of the image containing the seismic reflection data is then extracted and rotated so that the specified x-axis on the image is horizontal.

Detect time lines

The timelines on the image are then detected (lines denoting a constant time value, Fig. 2). Timelines are commonly annotated on hard copies of stacked seismic reflection data. As these timelines should in principle be perfectly straight and horizontal on the image, they can be used to correct vertical warping on the image, which may have arisen during the printing or scanning process. In order to detect the timelines on the image a series of morphological image processing steps are applied, namely erosion and dilation (Gonzalez and Woods 2008) (Fig. 3). If we consider an area of black pixels on a monochrome image, erosion can be viewed as a process which removes the outermost black pixels from this area. This has

Table 1 The key steps performed in the vectorization process detailed in this study

1. Rotate image
2. Detect timelines
3. Remove timelines
4. Correct image for vertical warping
5. Detect baseline positions
6. Extract amplitude information from each trace
7. Add the geometry to the trace headers
8. Output SEG-Y file

the effect of reducing the thickness of lines and shrinking the area of shapes on the image. Dilation has the opposite effect of erosion, where black pixels are added to the edges of any pre-existing areas of black pixels. In this study, erosion and dilation are only applied in one orientation, for example in the case of dilation, instead of adding pixels to all sides of a given area of black pixels on the image, black pixels are only added onto one side of this area (e.g. Fig. 3c and e). Some key input parameters for the process described here are: *HE* (Horizontal erode), *HLT* (Horizontal line thickness) and *TLT* (Trace line thickness). These parameters are quantities of pixels which are typically either added or removed from the image during the image processing routine described in the following section. Optimum values for these parameters must be chosen by testing. However, typical values for *HLT* and *TLT* would be approximately equal to (or slightly larger than) the average

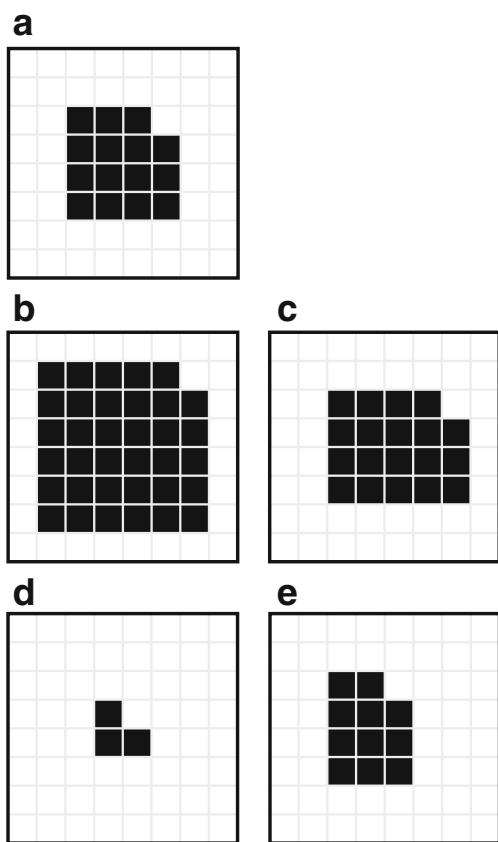


Fig. 3 A series of schematic diagrams detailing the process of erosion and dilation (Gonzalez and Woods 2008). **a** A subset of a raster image containing white and black pixels. **b** The result of applying dilation to the image in a) using a 3×3 structuring element, this is equivalent to adding 1 black pixel onto the edges of the area of black pixels in the original image. **c** The result of applying dilation only on the right hand side of the image in a), here 1 black pixel is added to the right hand side of the area of black pixels in the original image. **d** The result of applying erosion to the image in a) using a 3×3 structuring element, this is equivalent to removing 1 black pixel from the edges of the area of black pixels in the original image. **e** The result of applying erosion only from the right hand side of the image in a), this is equivalent to removing 1 black pixel from the right hand edge of pixels in the original image

thickness (in pixels) of the timelines and lines describing the waveforms of the seismic traces, respectively. Optimum values for *HE* are variable, however, 4–10 times the value of *HLT* or *TLT* would be somewhat typical.

The sequence of morphological image processing steps applied to detect timelines are described in Figs. 4 and 5, which provide examples of their application to both schematic and real data, respectively. Firstly, the image is eroded from the left hand side by *HE* pixels (which is equivalent to removing *HE* pixels from the left hand edge of all black parts of the image) (Fig. 5b). This is followed by a filter applied in the vertical direction, which removes any columns of black pixels with a vertical thickness greater than a factor *HLT* (Fig. 5c). The image is then eroded by *HE* pixels, this time from the right hand side of the image (Fig. 5d). This series of image processing steps is typically effective at removing everything from the image, with the exception of the near horizontal timelines. To detect timelines on the image a moving average is then calculated in the horizontal direction using the image shown in Fig. 5d. The location of the timelines can then be found by detecting maxima on the grid of average values.

Remove timelines

After detection of the timelines, efforts are made to remove these timelines from the image (Fig. 4). Firstly, a filter is applied to the original image which removes any black pixels which have a vertical thickness greater than *HLT* pixels (Fig. 5e). This filtered image contains the majority of the pixels describing the timelines, which we wish to remove, but also contains pixels describing the waveforms, which we wish to preserve. A second processed image is generated by dilating the image shown in Fig. 5d by *HLT*/2 pixels in the upwards and downwards directions (this is equivalent to adding *HLT*/2 pixels to the upper and lower edge of all black pixels on the image) (Fig. 5f). The image shown in Fig. 5e is then subtracted from the original image, but only in areas where black pixels are present in the image shown in Fig. 5f. This produces the image shown in Fig. 5g, which is a version of the original image where the timelines have been removed, with minimum damage to the waveform information.

Correct image for vertical warping

After removing the timelines from the image, the relative vertical position of the horizontal timelines (saved from a previous step) is used to apply a variable vertical shift to the image to correct for the effects of vertical warping (Fig. 6). Here the vertical difference between the leftmost point on a given timeline and every other point on that timeline is calculated. This process is repeated for each of the detected timelines. The mean average relative shift across all of the individual timelines is then calculated for each horizontal position on the

Fig. 4 A flow chart describing the morphological image processing steps applied in order to detect and remove timelines and detect baselines on the image of the stacked seismic data. A schematic close up image of the scanned seismic section is provided at each step. The letters displayed at the bottom right hand corner of each image, denote the corresponding image in Fig. 5. Note that the values used here for HE , HLT and TLT differ from those used in Fig. 5

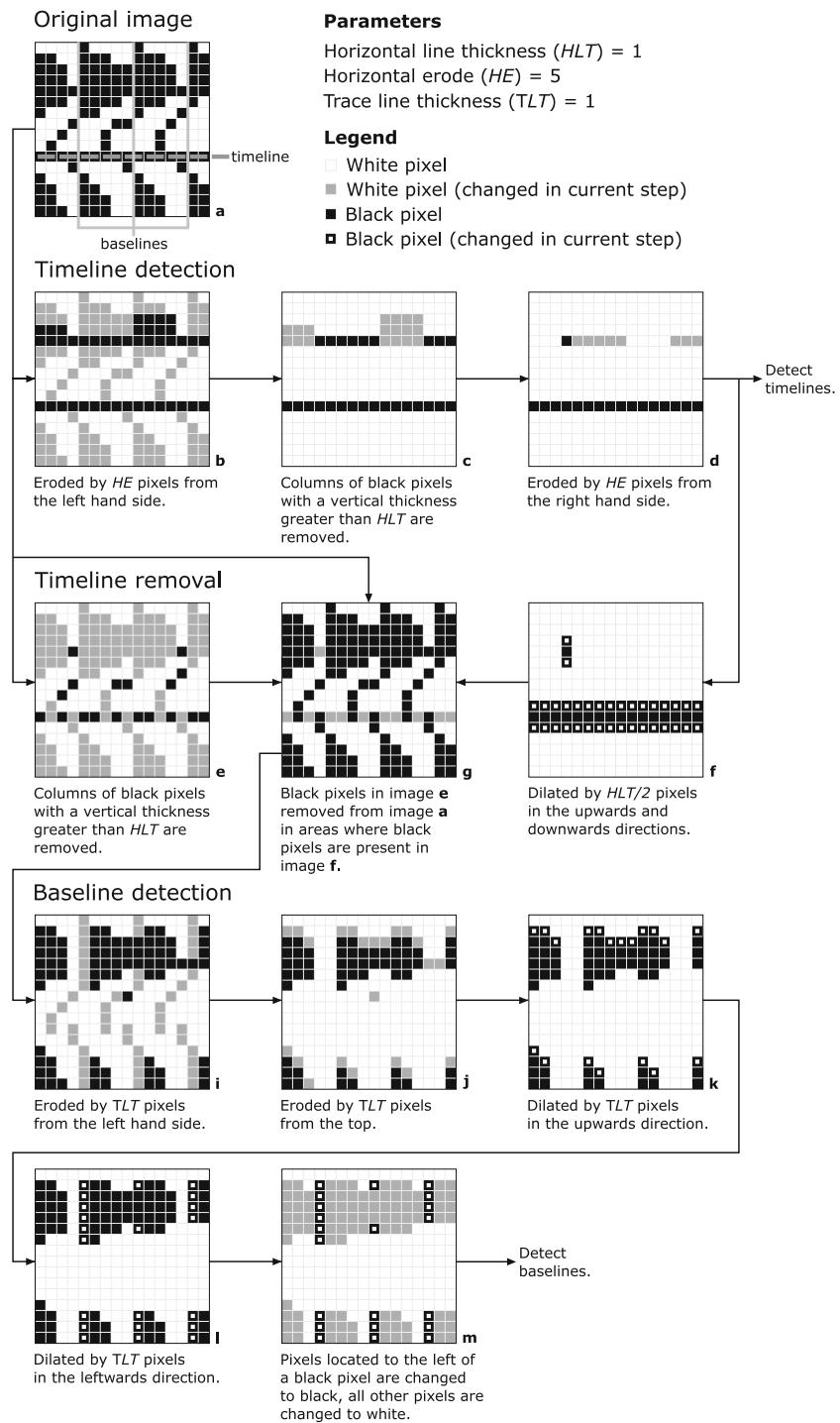


image to provide the variable vertical shift which is applied to correct the image (Fig. 6).

Detect baseline positions

The next step in the process is to detect the trace baselines on the image (Fig. 4). This is achieved by applying a series of image processing steps to the image of the data after timeline removal (Fig. 5h). Initially this image is eroded by TLT pixels

from the left hand side (Fig. 5i) and then eroded by TLT pixels from the top (Fig. 5j). The image is then dilated by TLT pixels in the upwards (Fig. 5k) and left (Fig. 5l) directions. This typically removes everything from the image with the exception of the shaded parts of the trace. Pixels located on the left hand side of a region of black pixels on image Fig. 5l are then detected to give the image shown in Fig. 5m. It is clear from the image shown in Fig. 5m that the black pixels provide a good indication of the trace baseline positions. The moving

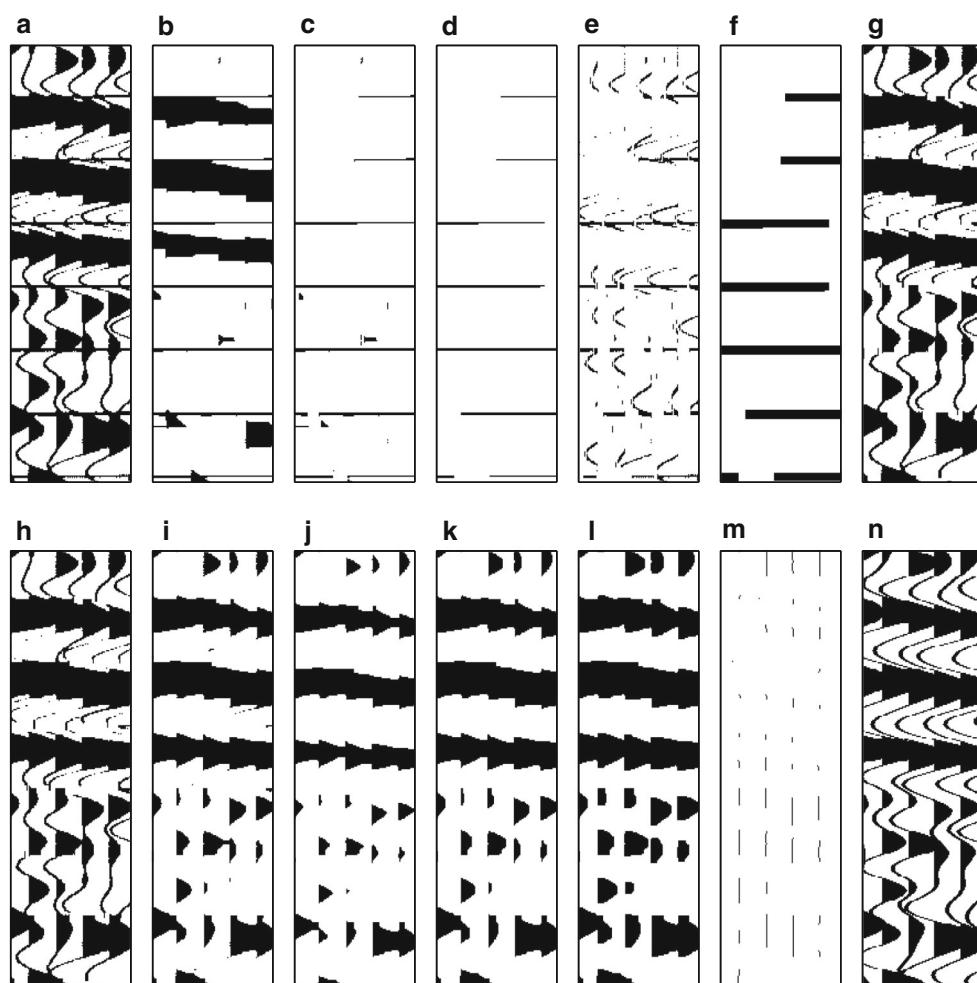


Fig. 5 Shows a subset of a real scanned image of seismic data at different points through the image processing workflow. **a** Original image. **b** Image a) after eroding from the left hand side by $HE(20)$ pixels. **c** Image b) after removing any black pixels with a vertical thickness greater than $HLT(5)$ pixels. **d** Image c) after eroding from the right hand side by $HE(20)$ pixels, this image is used for timeline detection. **e** Image a) after removing any pixels with a vertical thickness greater than $HLT(5)$ pixels. **f** Image d) after vertical dilation by $HLT(5)$ pixels (2 pixels upwards and 3 pixels downwards in this case). **g** Image a) after subtraction of Image e) in areas, where Image f) is black. This image is the original image after timeline

removal. **h** Image g) after applying a variable vertical shift to correct for vertical warping (no significant change in this case). **i** Image h) after eroding from the left by $TLT(6)$ pixels. **j** Image i) after eroding from the top by $TLT(6)$ pixels. **k** Image j) after dilation on top by $TLT(6)$ pixels. **l** Image k) after dilation on the left hand side by $TLT(6)$ pixels. **m** Image generated by detecting white pixels located on the left hand side of a black pixel on Image l). This image is used to detect the position of the trace baselines. **n** Shows the extracted SEG-Y file obtained by vectorization of image h) using method 4, plotted with a similar style as the original plot

average of the image shown in Fig. 5m is then calculated in the vertical direction, where the baseline locations can be detected as maxima.

Extract amplitude information from each trace

The amplitude information is then extracted from each trace individually. Initially, the polarity at every position along the trace is determined based on the presence of black pixels in a narrow window around the baseline. If the value is positive (black pixels present at the baseline) the amount of black pixels to the right of the baseline is calculated (i.e. the number of pixels between the baseline and the first white pixel). If the value is negative (white pixels present

at the baseline), the amount of white pixels to the left of the baseline is calculated (i.e. the number of pixels between the baseline and the first black pixel). This allows a somewhat noisy estimate of the trace amplitude vs time to be extracted from the image (Fig. 7a). The trace is then re-sampled in time, so that it has the desired output sampling rate. Stacked seismic reflection data are inherently limited to a certain frequency range due to the characteristics of the acquisition equipment used. It is also common to apply a bandpass frequency filter during the seismic processing sequence and to specify the parameters for this filter on the printed hard copy of the data. Noise and artefacts within the extracted trace, which contain frequencies outside of the bandwidth of the seismic data can therefore be removed

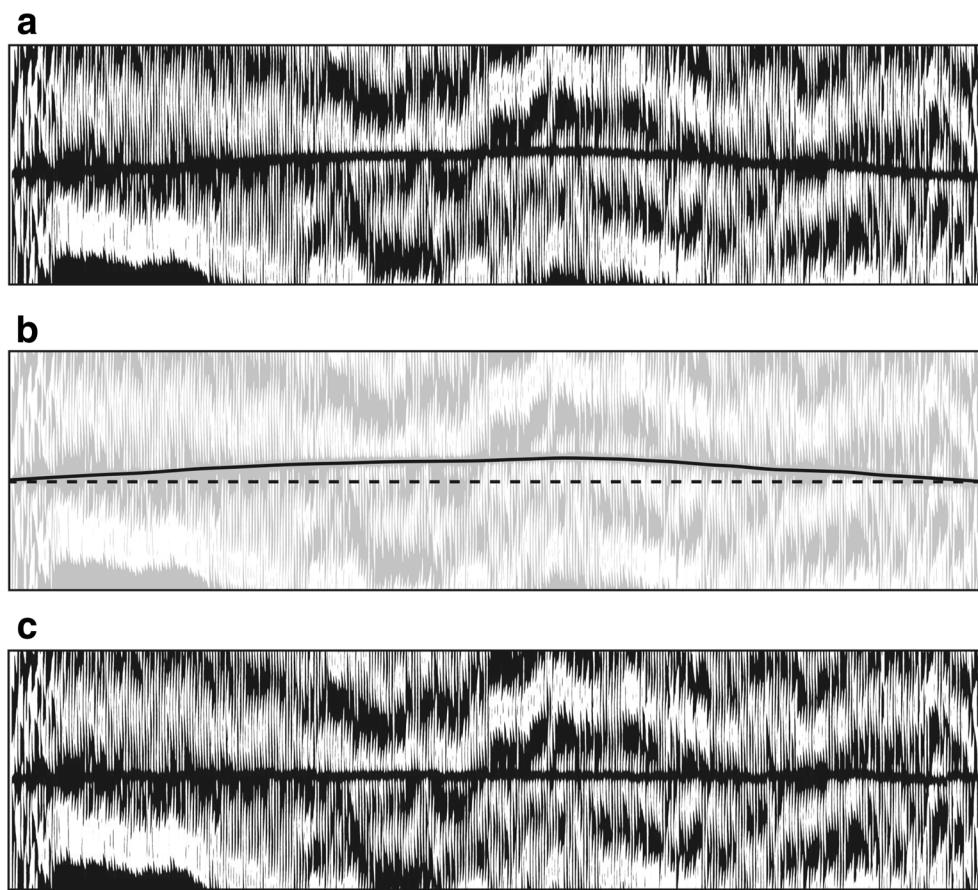


Fig. 6 A subset of a real dataset before and after applying the variable vertical shift to correct for vertical warping of the image during scanning or printing. **a** Scanned image of stacked seismic data after image rotation. A timeline on the image can be seen as a thin black line. Note that the line is not horizontal, i.e. the central part of the line is higher on the image than the ends of the line. **b** Shows a faded version of the image in a) which has been annotated. The dashed black line denotes a perfectly horizontal line. The solid black line shows the position of the timeline across the image.

Ideally, if the image was un-deformed, the timeline on the image would be located along the dashed line. The magnitude of the variable vertical shift applied to the image is described by the gap between the dotted and solid black lines, for all positions along the stacked section. **c** shows the image of the data in a) after applying the variable vertical shift to the data. Note that the timeline is now horizontal. All 3 images show a section of data spanning approximately 50 ms in time along the y-axis and 330 Common depth points (CDPs) along the x-axis

by frequency filtering. Miles et al. (2007) and Chevion and Navon (2012, 2014) also apply a bandpass filter during their vectorization procedures.

To achieve this filtering process a band limited linear inversion is performed as follows. The seismic trace x is expressed in terms of a Fourier transform below:

$$x = \sum_{k=0}^{N/2} a_k \cos(2\pi k i / N) + \sum_{k=0}^{N/2} b_k \sin(2\pi k i / N) \quad (1)$$

Where the trace has N discrete values, a_k and b_k represent the Fourier coefficients. The gradient (derivative) of the seismic trace x' can be expressed as follows:

$$x' = \sum_{k=0}^{N/2} -a_k (2\pi k / N) \sin(2\pi k i / N) + \sum_{k=0}^{N/2} b_k (2\pi k / N) \cos(2\pi k i / N) \quad (2)$$

The relationship between the trace amplitude (or gradient) and the Fourier coefficients is linear and can be expressed as:

$$x = Gm \quad (3)$$

Where x and m denote vectors containing the amplitude values and Fourier coefficients, respectively. The Fourier coefficients can be obtained using the following:

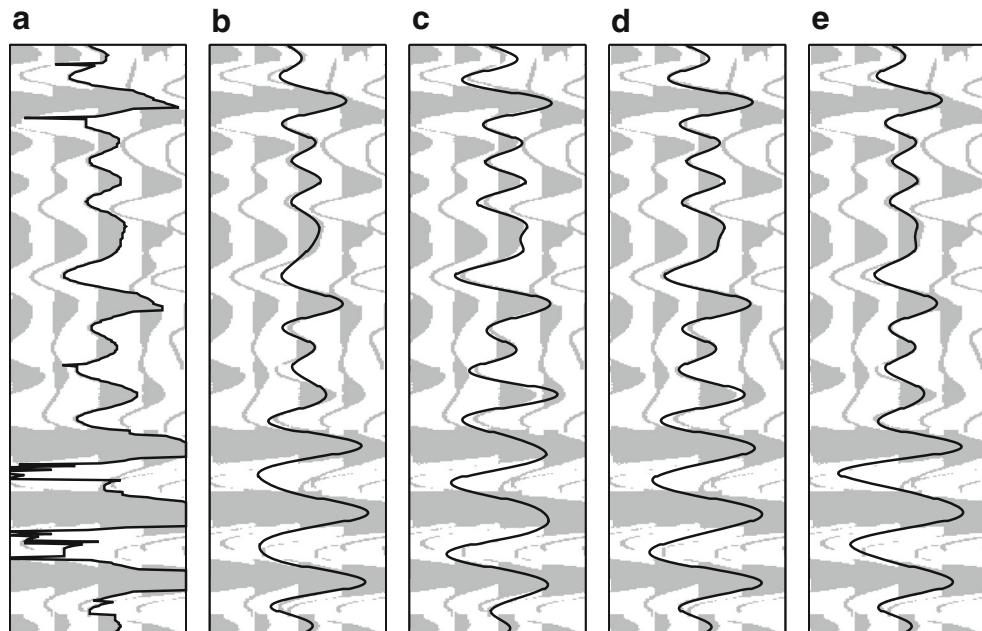
$$m = [G^T G]^{-1} G^T x \quad (4)$$

To limit the frequency band, unwanted frequencies are omitted by removing columns from the G matrix and the corresponding Fourier coefficients from the vector m . The expression below is used to calculate a set of band limited Fourier coefficients from the extracted trace. These coefficients are then used in eq. 3 to calculate a band limited estimate of the amplitude values of the trace:

$$m = [G^T G + \varepsilon + B\gamma]^{-1} G^T x \quad (5)$$

Here the diagonal matrix B provides tapering in the frequency domain at the ends of the frequency pass band being used and is scaled by the variable γ . The

Fig. 7 **a** Shows the raw trace extracted from the image. **b, c, d,** and **e** show trace **a**) after applying Methods 1, 2, 3 and 4 to the raw trace, respectively



diagonal matrix ε provides damping. The same method can be used to calculate a band limited seismic trace, from the gradient values.

In this study 4 different ways of applying the band limited inversion to the extracted trace in order to obtain the final estimate of the trace from the image are tested (Fig. 7), as different methods will achieve the best results with different input datasets. These different ways of applying the band limited inversion to the extracted trace will be referred to as Method 1, 2, 3 and 4 in the following sections. In Method 1 only the positive values are retained from the waveform and input into the band limited inversion (i.e. only the areas of the curve which are shaded), this is somewhat equivalent to the method used by Miles et al. (2007). In Method 2 the gradient of the entire extracted trace is used as input to the band limited inversion, instead of the amplitude. In Method 3 the mean average trace output from Methods 1 and 2 is calculated. In Method 4 the entire waveform extracted from the image is input into the band limited inversion (as in Method 1 but the negative parts of the trace are included). It is clear from Fig. 7 that all four methods achieve a reasonable result on this example dataset. However, as one might expect, Method 1 does not capture the negative waveform values well.

Add the geometry to the trace headers

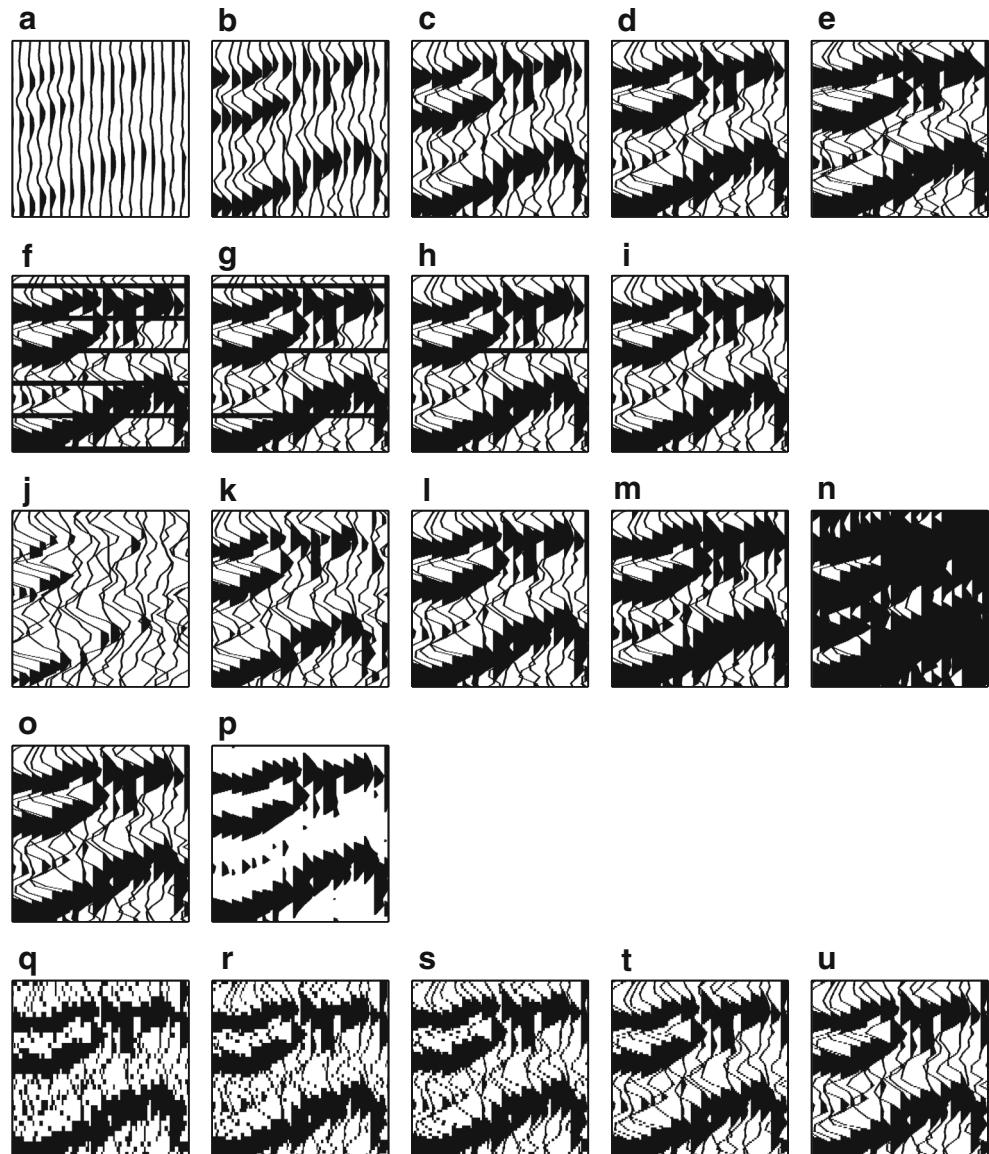
If geometry information is available it can be matched to the newly vectorized data and the location of each CDP on the image can be calculated. The location of each CDP can then be added to the trace header information. At the end of the process the extracted amplitude versus time information is then saved as a SEG-Y file (SEG 1994).

Vectorization results

To date, the vast majority of all published material describing the performance of different vectorization methods only provides subjective measures of performance. The most common form of performance assessment used in existing academic literature and on commercial websites is simply to display the original image and the vectorized data with similar display parameters side by side. In this section the performance of the method described in this study is quantitatively assessed using SEG-Y data from a real onshore seismic profile. The data in the assessment were recorded, processed and stored digitally (i.e. they were not generated from a scanned image). These data were then plotted and output as a series of TIFF images. In each TIFF image the plotting parameters were varied, namely, the trace deviation, the timeline interval, the bias and the resolution. An image where the lines defining the waveforms were omitted (i.e. only the shaded areas were retained) was also generated (Fig. 8). It should be noted that the series of TIFF images which were generated were used directly in the vectorization process (i.e. they were not printed out and then scanned). The mean average correlation coefficient between the SEG-Y data vectorized from the image and the original SEG-Y file was then calculated (Table 2). In some cases it was not possible to successfully detect the trace baseline positions on the image, in these situations the baseline positions were manually defined.

The highest average correlation coefficient achieved was 0.961, while values between 0.8 and 0.9 appear to be achievable with a large range of different display parameters. The highest correlation coefficient values are achieved with relatively low trace deviation and little trace overlap. A gradual decrease in the correlation coefficients is observed as the

Fig. 8 Shows the same section of seismic data, plotted with a range of display parameters. **a, b, c, d** and **e** show the data plotted with successively larger trace deviation. **f, g, h** and **i** show the data plotted with successively fewer timelines. **j, k, l, m** and **n** show the data plotted with varying bias. **o** and **p** show the data plotted with and without a line to define the waveform on the image. **q, r, s, t** and **u** show the data plotted with increased resolution. For more detail on the plotting parameters used to plot the different images see Table 2



amount of trace overlap increases. Although the increased frequency of timelines is detrimental to the result, they appear to have a relatively minor effect on the performance of the method. Trace bias refers to a shift of the baseline position relative to the zero amplitude value of the trace. Bias appears to have the largest effect on the correlation coefficients obtained. This is partially due to the somewhat extreme bias values tested and the fact that the method could not successfully detect baseline positions on some of these images. Both very high and very low bias values had a large detrimental effect on performance. The omission of a line to define the waveform of the trace appears to result in a decrease in the correlation coefficient (by a factor of approximately 0.1). Reducing the resolution of the image has very little effect on the correlation coefficients up until a certain point at which the values reduce fairly rapidly. This has implications for the selection of the resolution when scanning the paper hard copies and it is clear

that nothing is gained when scanning above a key threshold value. This dataset contained 167 traces and had a total recording time of 0.5 s, therefore if the data was printed with horizontal and vertical dimensions of 25 cm and 16 cm respectively, no improvement in the result would be obtained if the resolution of the scan was increased above 300 dots per inch (DPI). Scanning an image of the test dataset with this size at a resolution of 300 DPI would result in an image with approximately 17 pixels between baseline positions.

In terms of the relative performance of the four different methods of applying the band limited inversion to the extracted trace, it appears that Method 4, on average performs the best across the different display parameters, making it the best method to use in most situations. However, its performance is highly dependent on accurate location of the baseline positions, so in cases where the baseline positions are not successfully detected, the method performs poorly. Method 1

Table 2 Table showing average correlation coefficients between SEG-Y files extracted (vectorized) from a series of images generated using variable plotting parameters and the original SEG-Y file used to plot the images

Plotting parameter being varied	Figure	Baselines detected?	Correlation co-efficient			
			Method 1	Method 2	Method 3	Method 4
Trace deviation						
1	A	Yes	0.954	0.933	0.95	0.955
2		Yes	0.943	0.949	0.955	0.961
3	B	Yes	0.908	0.939	0.946	0.947
4		Yes	0.882	0.905	0.927	0.934
5	C	Yes	0.862	0.834	0.892	0.92
6		Yes	0.85	0.751	0.854	0.905
7	D	Yes	0.84	0.671	0.821	0.891
8		Yes	0.83	0.596	0.792	0.876
9	E	Yes	0.821	0.518	0.762	0.861
10		Yes	0.811	0.462	0.741	0.846
Time line interval (ms)						
5	F	Yes	0.825	0.534	0.76	0.873
10	G	Yes	0.832	0.605	0.793	0.881
20	H	Yes	0.836	0.64	0.808	0.886
50	I	Yes	0.839	0.658	0.816	0.889
100		Yes	0.84	0.665	0.819	0.89
None		Yes	0.84	0.671	0.821	0.891
Trace Bias						
No Fill		No	0.181	0.516	0.448	-0.054
-2	J	No	0.176	0.284	0.3	0.213
-1		No	0.526	0.465	0.567	0.59
-0.5	K	No	0.701	0.573	0.708	0.782
-0.25		Yes	0.799	0.644	0.789	0.866
0	L	Yes	0.84	0.671	0.821	0.891
0.25		Yes	0.864	0.672	0.837	0.899
0.5	M	Yes	0.873	0.651	0.84	0.896
1		Yes	0.853	0.573	0.812	0.863
2	N	No	0.694	0.359	0.608	0.695
Wiggle trace line						
With line	O	Yes	0.84	0.671	0.821	0.891
No line	P	Yes	0.841	0.418	0.714	0.801
Resolution DPI (PBB)						
100 (5.7)	Q	Yes	0.618	0.246	0.494	0.648
150 (8.6)	R	Yes	0.71	0.384	0.613	0.747
200 (11.5)	S	Yes	0.83	0.563	0.775	0.879
300 (17.2)	T	Yes	0.837	0.618	0.798	0.885
400 (22.9)	U	Yes	0.835	0.632	0.803	0.884
500 (28.6)		Yes	0.842	0.693	0.829	0.893
600 (34.4)		Yes	0.839	0.668	0.819	0.89

The letter listed in the Figure column refers to the corresponding image in Fig. 8. Note that in the Resolution results DPI refers to dots per inch and PBB refers to pixels between baselines

performs quite consistently across the different display parameters and is clearly the best method to select when no line is used to annotate the waveform and only shaded areas are

displayed (Fig. 8p). However, it clearly does not perform well when only a wiggle line is used to represent the traces, and no shaded areas are present. Method 2 performs the poorest of the

four methods, however, it performs relatively well when very little or no shaded areas are annotated on the image. Although Method 3 does not perform the best on average, it is possibly the most robust method, exhibiting the highest minimum correlation coefficient of the four methods. Method 4 is an improvement on the process utilized by Miles et al. (2007), which is equivalent to Method 1 in this study. The improved performance is likely due to the fact that Method 4 captures the entire waveform, rather than just the positive parts of the waveform.

Vectorization case study: Gotland, Sweden

In this section the vectorization process described in this study is applied to a vintage seismic dataset. The dataset is from the island of Gotland, which is located in the Baltic Sea in northern Europe. Geologically, Gotland lies on the north western flank of the Baltic Basin and is underlain by a sequence of Palaeozoic strata, which have a total thickness of about 350 m in the north of the island and about 700 m in the south. Deposition within the Baltic Basin began during the Cambrian, where shallow marine sandstones, siltstones and mudstones were deposited directly upon the heavily eroded basement (Nielsen and Schovsbo 2006). This was overlain by a sequence of laterally extensive Ordovician limestone units (Sopher et al. 2016). During the Late Silurian and Early Devonian the Baltic Basin developed as a flexural foreland basin, related to the Caledonian orogeny (Poprawa et al. 1999). This led to the deposition of a thick Silurian sequence, which constitutes the present day bedrock beneath Gotland. The Silurian strata on Gotland consist of a sequence of about 10 barrier reef complexes, which prograde to the SSE (Manten 1971; Calner et al. 2004). As the strata are very well preserved on Gotland, it has become a globally important locality for the study of Silurian geology and has been extensively documented (Calner and Säll 1999; Calner and Jeppsson 2003; Eriksson and Calner 2005; Calner 2005 etc.). However, despite this, a stratigraphic framework to describe the reef complexes on Gotland is only loosely defined and in general, the correlation between the surface geology and the deeper Silurian succession is poorly understood (Calner et al. 2004). This is largely because changes in lithology make the correlation between the shallow and deeper parts of the Silurian sequence challenging, and because the majority of previous studies focus on the surface geology. In recent years, declining groundwater levels and the potential to utilize a number of subsurface reservoirs in the basin for CO₂ and energy storage have highlighted the importance of gaining a better understanding of the subsurface geology of Gotland.

Recently, an extensive seismic reflection dataset, acquired between 1970 and 1990 by the Swedish Oil Exploration company (OPAB), has been made available for use by the Geological Survey of Sweden (SGU) (Sopher and Juhlin

2013). Reprocessing and interpretation of the data available digitally within this dataset has provided new insight into the structure and stratigraphy of the Baltic Basin and its CO₂ storage potential (Sopher et al. 2014; Sopher et al. 2016). As well as marine data, the dataset contains over 2300 km of land data from the island of Gotland. Some 600 km of this dataset were available digitally, however, the remaining part (some 1700 km) was only available as scanned images of the printed hard copies (Fig. 9). This dataset on Gotland is valuable as it has good spatial coverage of the island and provides a good image of the entire Palaeozoic sequence. To date, results from this dataset are almost entirely unpublished. Specifically, these data can be used to: (1) improve the understanding of the cyclic development of the prograding reef complexes on Gotland, (2) correlate the deep Silurian stratigraphy with the surface geology, and (3) provide valuable information on potential reservoirs for energy or CO₂ storage and associated cap rock formations, currently being investigated on Gotland.

In order to utilize this seismic dataset vectorization of the scanned sections was required. The dataset includes data acquired between 1974 and 1985, across 12 different surveys. The quality of the data and the way in which the data are plotted varies significantly. The data are typically 12 fold, with CDP spacing ranging between 10 and 30 m, vibroseis or minisource sources were typically used for the acquisition (Fig. 10).

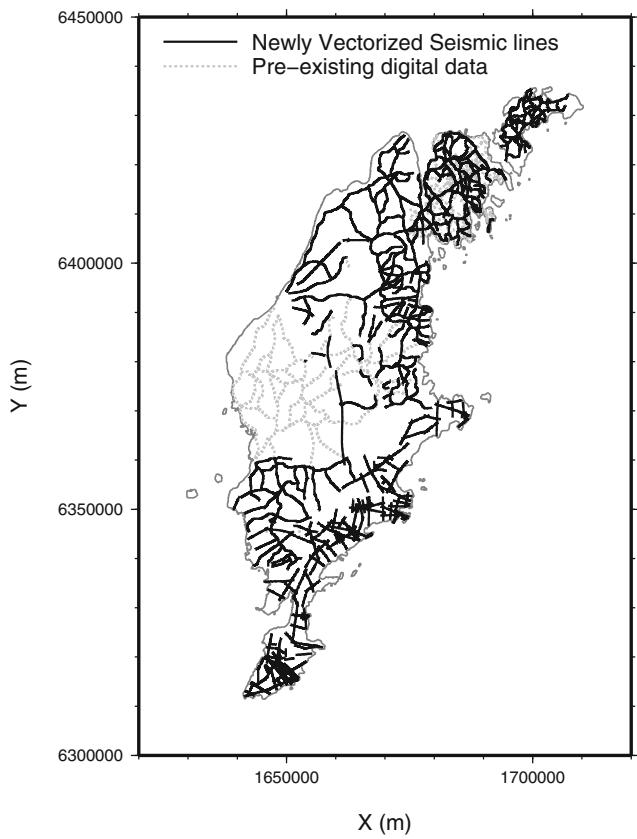


Fig. 9 Map showing the location of seismic reflection data that were only available as TIFF images of printed hard copies

The most time consuming part of the vectorization process (approximately 2/3 of the total time) was preparation of the input data. This involved organising the scanned sections into surveys and formatting the geometry information. It also involved checking all of the scanned hard copies for signs of annotation, or interpretation, which had been drawn on the hard copies before scanning. In many cases the most effective way to remove the old interpretation was by hand, using a graphic manipulation package (Fig. 10). The method described in this study was then used to vectorize the scanned seismic sections. Typically, all of the seismic sections in a given survey were plotted with the same plotting parameters and it was therefore most efficient to vectorize all lines from a given survey before moving onto the next. When beginning with a new survey, time was required to optimise the process for the given plotting style. In order to optimise the parameters and select the best method to use, a plot was generated using a plotting style as similar as possible to that of the survey being vectorized, using an available SEG-Y file. The correlation

coefficients between the vectorized and original SEG-Y files were then calculated for a range of different parameters and the four different methods, to select the optimum values. The corner frequencies for the band limited inversion were obtained from the processing information on the scanned hard copies. Fig. 10 shows two subsets of scanned seismic sections from the Gotland dataset alongside the extracted SEG-Y file plotted in both variable density and wiggle trace with variable fill format. Based on the performance information in Table 2 and the types of plotting styles used, it is likely that the average correlation coefficient between a given vectorized SEG-Y file and the original SEG-Y file (not available in this case) would have varied between 0.8 and 0.95 for the sections in the Gotland dataset.

Fig. 11 shows a regional seismic profile, which runs approximately north-south across the island of Gotland. Although the entire seismic dataset has been vectorized, further work is required to fully interpret the data. Therefore, only a preliminary interpretation of the regional seismic line is presented in

Fig. 10 Two examples of vectorization results from the Gotland dataset, using the process described in this study. **a** Scanned image of line MS-83-546, which has relatively good plotting parameters for vectorization. **b** SEG-Y file extracted from the scanned image in a), plotted with similar parameters to the original (Method 4 used). **c** same SEG-Y data as in b) plotted with a variable density display. **d** Scanned image of line MS-77-178, which has relatively poor plotting parameters for vectorization. This section contains interpretation that was manually removed before vectorization. **e** SEG-Y file extracted from the scanned image in d), plotted with similar parameters to the original (Method 4 used). **f** same SEG-Y data as in e) plotted with a variable density display

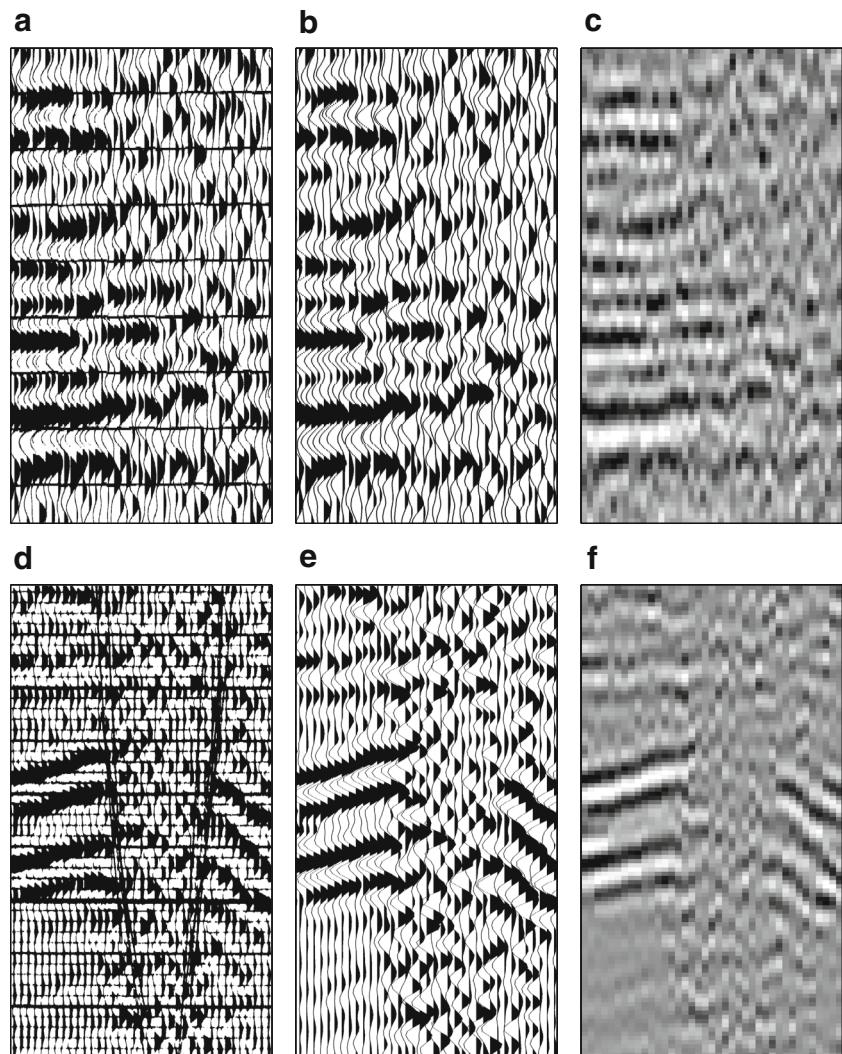


Fig. 11 Regional seismic profile constructed from 12 individual seismic lines. **a** shows the regional seismic profile with a variable density display. Note that approximately the first 45 km of the profile was acquired using a vibroseis source, while the remaining part of the profile was collected using a mini-sosie source. As a result the frequency content of the data changes significantly along the line. The location of the profile is shown in the small map in the lower right hand corner of the image. **b** shows a preliminary interpretation of the regional seismic line shown in a)

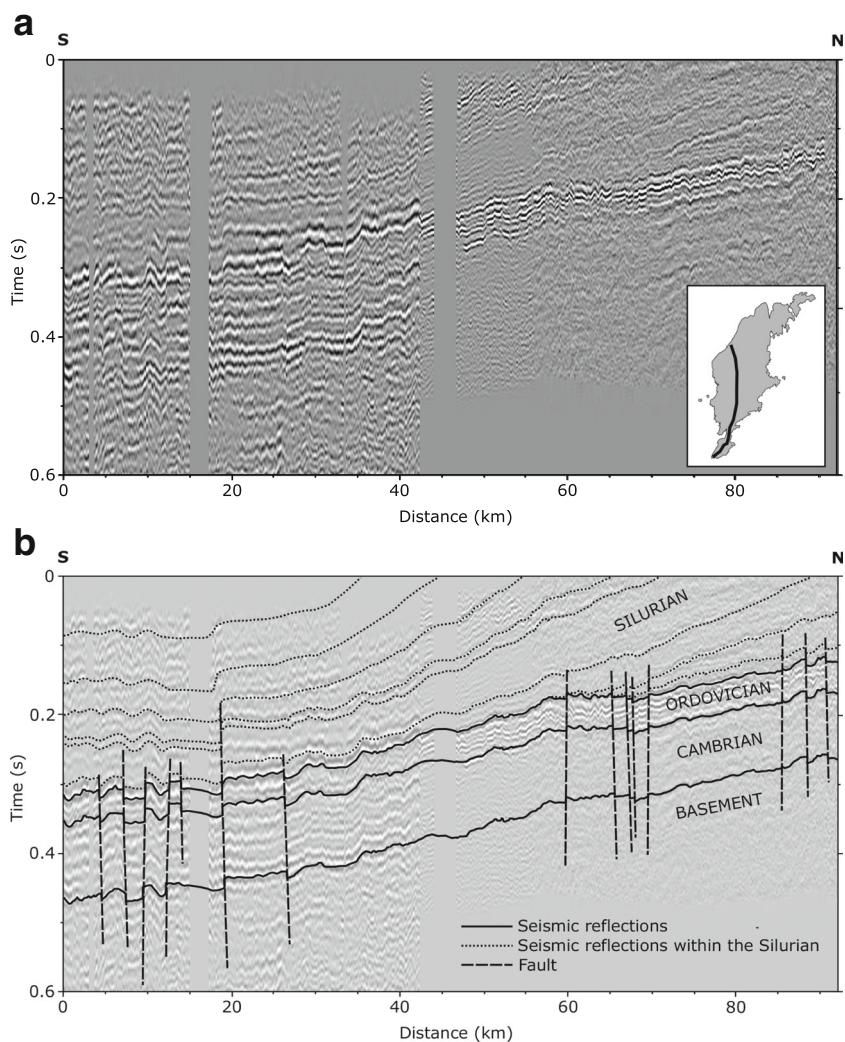


Fig. 11. This profile was constructed from 12 individual seismic lines acquired between 1974 and 1979. As the data were available in SEG-Y format it was possible to perform a series of post-stack seismic processing steps on the data to enhance the final image, these included migration, application of a horizontal median filter, amplitude scaling and bandpass filtering. A number of seismic reflections can be identified on the profile. These typically dip towards the south, which is consistent with the regional geological structure of the study area. Several high amplitude seismic events can be correlated across the entire section, which intersect the southern and northern ends of the profile at about 0.3 and 0.1 s, respectively. These strong seismic events are interpreted to be associated with the top and base of the laterally extensive Ordovician limestone units. Below the Ordovician sequence, lies the Middle Cambrian Faludden sandstone, which is one of the most prospective reservoirs in the Basin for energy and CO₂ storage. Currently, there are no maps describing the structure of this reservoir beneath Gotland. Therefore, it appears that these vintage data will allow the top of this reservoir to be mapped in detail. Within the relatively

thick Silurian sequence a number of continuous seismic reflections can be interpreted. These reflections appear to onlap onto the top of the Ordovician in places and to exhibit a forestepping pattern. It is likely that these reflections are associated with sequence boundaries within the Silurian and can therefore be used to describe the overall form of the prograding carbonate ramp system on Gotland and to correlate the surface geology with the deeper Silurian section. Therefore, based on these preliminary results, it appears that with further interpretation these vintage data can be utilized to significantly improve the understanding of the geology of Gotland.

Conclusions

A methodology for reconstructing SEG-Y files from scanned images of seismic data is presented. For the first time, a detailed explanation of the process used to detect and remove timelines and to detect trace baseline positions is described. The performance of the methodology has been quantitatively assessed and

it has been shown that average trace-to-trace correlation coefficients (between the extracted and original SEG-Y files) between 0.8 and 0.95 can be achieved in the majority of cases. The method presented here is implemented in the widely available Matlab language and provides a methodological improvement on that of Miles et al. (2007), the only other low-cost alternative software available for the vectorization of seismic data displayed in this way. A case study has also been presented which details the application of the vectorization process to an extensive dataset on Gotland, Sweden. Preliminary work on this dataset indicates that having available vectorised data from the area will allow significant improvements in the understanding of the Baltic Basin to be made. This study, therefore, also demonstrates the potential to solve long-standing problems in Earth sciences through rescue, recovery and re-interpretation of vintage datasets.

Acknowledgements The Swedish Research Council (VR) partly funded Daniel Sopher during this research (project number 2010-3657) and is gratefully acknowledged. GLOBE ClaritasTM under license from the Institute of Geological and Nuclear Sciences Limited, Lower Hutt, New Zealand was used to process the seismic data after vectorization. We thank Björn Bergman and Sverker Olsson at the Geological Survey for their input and providing data. Please contact the author directly if you wish to obtain a copy of the Matlab script used to perform the vectorization process in this study. I thank the three anonymous reviewers for their comments which were very helpful to improve the manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Adobe Developers Association (1992) TIFF 6.0. Adobe Systems incorporated, mountain view. URL <ftp://ftp.adobe.com/pub/adobe/DeveloperSupport/TechNotes/PDFfiles>
- Al Mahdy OMH, Sedek MS (2013) Abu Roash G dolomite reservoir characterization using seismic inversion technique, Horus field, Western Desert, Egypt. Arab J Geosci 6:1769–1797
- An VA, Ovtchinnikov VM, Kaazik PB, Adushkin VV, Sokolova IN, Aleschenko IB, Mikhailova NN, Kim W, Richards PG, Patton HJ, Phillips WS, Randall G, Baker D (2015) A digital seismogram archive of nuclear explosion signals, recorded at the Borovoye geophysical observatory, Kazakhstan, from 1966 to 1996. GeoResJ 6: 141–163
- Baskoutas IG, Kalogerias IS, Kourouzidis M, Panopoulou G (2000) A modern technique for the retrieval and processing of historical seismograms in Greece. Nat Hazards 21(1):55–64
- Blake N, Hewlett C (1993) Digital information recovery from paper seismic sections for work station loading. JEOFÍZÍK 7:3–14
- Calner M (2005) Silurian carbonate platforms and extinction events – ecosystem changes exemplified from the Silurian of Gotland. Facies 51:603–610
- Calner M, Jeppsson L (2003) Carbonate platform evolution and conodont stratigraphy during the middle Silurian Mulde event, Gotland, Sweden. Geol Mag 140:173–203
- Calner M, Säll E (1999) Transgressive oolites onlapping a Silurian rocky shoreline unconformity, Gotland, Sweden. GFF 121:91–100
- Calner M, Jeppsson L, Munnecke A (2004) The silurian of gotland – part I: review of the stratigraphic framework, event stratigraphy, and stable carbon and oxygen isotope development. Erlanger geologische Abhandlungen – Sonderband 5. Field Guide:113–131
- Chevion DS, Navon Y (2012) Method and system for retrieving seismic data from a seismic section in bitmap format. US 8326542 B2
- Chevion DS, Navon Y (2014) Tracing seismic sections to convert to digital format. US 8825409 B2
- Cooke R, Bulat J (2012) British geological survey internal report, IR/12/061 477pp
- Diviacco P, Wardell N, Forlin E, Sauli C, Burca M, Busato A, Centonze J, Pelos C (2015) Data rescue to extend the value of vintage seismic data: the OGS-SNAP experience. GeoResJ 6:44–52
- Eriksson ME, Calner M (2005) The dynamic Silurian earth: Subcommission on Silurian stratigraphy field meeting 2005. Sver Geol Unders Rapp Meddeland 121:6–99
- Farran M (2008) IMAGE2SEGY: Una aplicación informática para la conversión de imágenes de perfiles sísmicos a ficheros en formato SEG Y. Geo-Temas 10:1215–1218
- Gonzalez RC, Woods RE (2008) Digital image processing, 3rd edn. Pearson Prentice Hall, Upper Saddle River, p 954pp
- Griffin RE (2015) When are old data new data? GeoResJ 6:92–97
- Kanamori M, Rivera L, Lee WHK (2010) Historical seismograms for unravelling a mysterious earthquake: the 1907 Sumatra earthquake. Geophys J Int 183:358–337
- Manten AA (1971) Silurian reefs of Gotland Developments in Sedimentology 13, pp 539, Elsevier, Amsterdam
- Miles P, Schaming M, Lovera R (2007) Resurrecting vintage paper seismic records. Mar Geophys Res 28:319–329
- Nielsen A, Schovsbo N (2006) Cambrian to basal Ordovician lithostratigraphy in southern Scandinavia. Bull Geol Soc Den 53:47–92
- Okal EA (2015) Historical seismograms: preserving an endangered species. GeoResJ 6:53–64
- Owen MJ, Maslin MA, Day SJ, Long D (2015) Testing the reliability of paper seismic record to SEGY conversion on the surface and shallow sub-surface geology of the Barra fan (NE Atlantic Oceaan). Marine. Pet Geol 61:69–81
- Pintore S, Quintilani M, Franceschi D (2005) Teseo: a vectoriser of historical seismograms. Comput Geosci 31:1277–1285
- Poprawa P, Sliaupa S, Stephenson R, Lazauskiene J (1999) Late Vendian–early Palaeozoic tectonic evolution of the Baltic Basin: regional tectonic implications from subsidence analysis. Tectonophysics 314:219–239
- Ross S, Gow A (1999) Digital archaeology: rescuing neglected and damaged data resources. HATII- University of Glasgow 102pp
- SEG CFTS (1994) Digital field tape standards - SEG-D, revision 1 (special report). Geophysics 59(04):668–684
- Sopher D (2016) Characterization of the structure, stratigraphy and CO₂ storage potential of the Swedish sector of the Baltic and Hanö Bay basins using seismic reflection methods. Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1355, 85 pp. Uppsala: Acta Universitatis Upsaliensis
- Sopher D, Juhlin C (2013) Processing and interpretation of vintage 2D marine seismic data from the outer Hanö Bay area, Baltic Sea. J Appl Geophys 95:1–15
- Sopher D, Juhlin C, Erlström M (2014) A probabilistic assessment of the effective CO₂ storage capacity within the Swedish sector of the Baltic Basin. International Journal of Greenhouse Gas Control 30:148–170
- Sopher D, Erlström M, Bell N, Juhlin C (2016) The structure and stratigraphy of the sedimentary succession in the Swedish sector of the Baltic Basin: new insights from vintage 2D marine seismic data. Tectonophysics 676:90–111
- Xu Y, Xu T (2014) An interactive program on digitizing historical seismo-grams. Comput Geosci 63:88–95