A machine-learning benchmark for facies classification

Yazeed Alaudah[1], Patrycja Michałowicz[2], Motaz Alfarraj[1], and Ghassan AlRegib[1]

Abstract

The recent interest in using deep learning for seismic interpretation tasks, such as facies classification, has been facing a significant obstacle, namely, the absence of large publicly available annotated data sets for training and testing models. As a result, researchers have often resorted to annotating their own training and testing data. However, different researchers may annotate different classes or use different train and test splits. In addition, it is common for papers that apply machine learning for facies classification to not contain quantitative results, and rather rely solely on visual inspection of the results. All of these practices have led to subjective results and have greatly hindered our ability to compare different machine-learning models against each other and understand the advantages and disadvantages of each approach. To address these issues, we open source a fully annotated 3D geologic model of the Netherlands F3 block. This model is based on study of the 3D seismic data in addition to 26 well logs, and it is grounded on the careful study of the geology of the region. Furthermore, we have developed two baseline models for facies classification based on a deconvolution network architecture and make their codes publicly available. Finally, we have developed a scheme for evaluating different models on this data set, and we have evaluated the results of our baseline models. In addition to making the data set and the code publicly available, our work helps advance research in this area by creating an objective benchmark for comparing the results of different machine-learning approaches for facies classification.

Introduction

In recent years, there has been great interest in using fully supervised deep-learning models for seismic interpretation tasks such as facies classification (Araya-Polo et al., 2017; Huang et al., 2017; Rutherford Ildstad and Bormann, 2017; Waldeland and Solberg, 2017; Di et al., 2018; Dramsch and Lüthje, 2018; Shi et al., 2018; Zhao, 2018). Typically, deep-learning models — such as convolutional neural networks (CNNs) — have millions of free parameters and therefore require a large amount of annotated training data. Unfortunately, and unlike other areas of research such as computer vision, there is a lack of large publicly available annotated data sets for seismic interpretation that can be used to train and benchmark machine-learning models. To address this problem, some researchers resort to annotating their own training and testing data sets. For example, in the Netherlands F3 block, Zhao (2018) annotates 40 inlines, Di et al. (2018) annotate 12 inlines, whereas Rutherford Ildstad and Bormann (2017) only annotate a single inline for their model. The limited number of annotated sections is understandable given that the annotation process is time consuming, requires subject matter expertise, and can be quite subjective. Nevertheless, such limited annotations undermine the mass potential that machine learning could have when deployed in such a field.

Alternatively, there has been some research in attempting to avoid annotating large amounts of data by using weakly supervised learning approaches. Alaudah and AlRegib (2016) train a facies classification model using seismic images with image-level labels only. Later, Alaudah et al. (2019) propose a method for generating large amounts of training data using similarity-based retrieval and a weakly supervised label mapping algorithm. As few as one or two exemplar images per class were enough to automatically generate a large amount of training data. These automatically generated training data were then used to train a weakly supervised deconvolution network (Alaudah et al., 2018) for facies classification. Other researchers avoid supervision altogether by using traditional unsupervised machine-learning techniques such as principal component analysis or self-organizing maps. There is a very rich literature on traditional supervised and unsupervised methods for facies classification (Coléou et al., 2003; de Matos et al., 2006; Dubois et al., 2007). More recently, unsupervised techniques based on deep-learning models such as deep convolutional autoencoders

[1]Georgia Institute of Technology, Center for Energy and Geo Processing (CeGP), Atlanta, Georgia, USA. E-mail: alaudah@gatech.edu; motaz@gatech.edu; alregib@gatech.edu.

[2]University of Silesia, Faculty of Earth Sciences, Katowice, Poland. E-mail: pamichalowicz@us.edu.pl.

Manuscript received by the Editor 14 December 2018; revised manuscript received 10 April 2019; published ahead of production 15 May 2019; published online 17 July 2019. This paper appears in *Interpretation*, Vol. 7, No. 3 (August 2019); p. SE175–SE187, 10 FIGS., 3 TABLES.

http://dx.doi.org/10.1190/INT-2018-0249.1.

have been explored (Qian et al., 2018; Shafiq et al., 2018; Veillard et al., 2018).

Whether researchers annotate their own training data or use other techniques, there still remains a lack of large publicly available annotated data sets for seismic interpretation that can be used for training different models and comparing the performance of different approaches. Furthermore, it is common for papers that apply machine learning for facies classification, or other seismic interpretation tasks, to not contain quantitative results, but rather to rely solely on subjective visual inspection of the results. All of this leads to highly subjective results and greatly hinders the ability of researchers to compare different approaches against each other and understand the advantages and disadvantages of each approach.

To address these issues, and to help make machine-learning research in seismic interpretation more reproducible, we open source a fully annotated 3D geologic model of the Netherlands F3 block (dGB Earth Sciences, 1987). This model is grounded in the geology of the region, and it is based on the study of the 3D seismic data and 26 well logs located within the F3 block or its vicinity. The data also include fault planes that we have extracted from the F3 block. Although we do not use the fault data in this work, we do make the data publicly available for those who are interested in exploring fault detection within our model. Furthermore, we also present two baseline models for facies classification based on a deconvolution network architecture. The first baseline is a patch-based model that we train using a large number of small patches extracted from all the inlines and crosslines of the training set. The second baseline is a section-based model that we train directly on entire inlines and crosslines of the training set. In addition, we have open sourced all the codes that were used to train and test our baseline models using the PyTorch deep-learning library (Alaudah, 2019a). Finally, we propose a common procedure for evaluating different models on this data set, and we share the results for our baseline models. The next section provides a quick overview of the geology of the Netherlands F3 block and introduces our geologic model.

A 3D geologic model of the Netherlands F3 block

The North Sea is rich in hydrocarbon deposits, which is why this area is very well-studied in the literature (Doornenbal, 2014). The North Sea continental shelf, located off the shores of the Netherlands, is divided into geographical zones described by different letters of the alphabet; within these zones are smaller areas marked with numbers. One of these areas is a rectangle of dimensions 16 x 24 km known as the F3 block (see Figure 1).

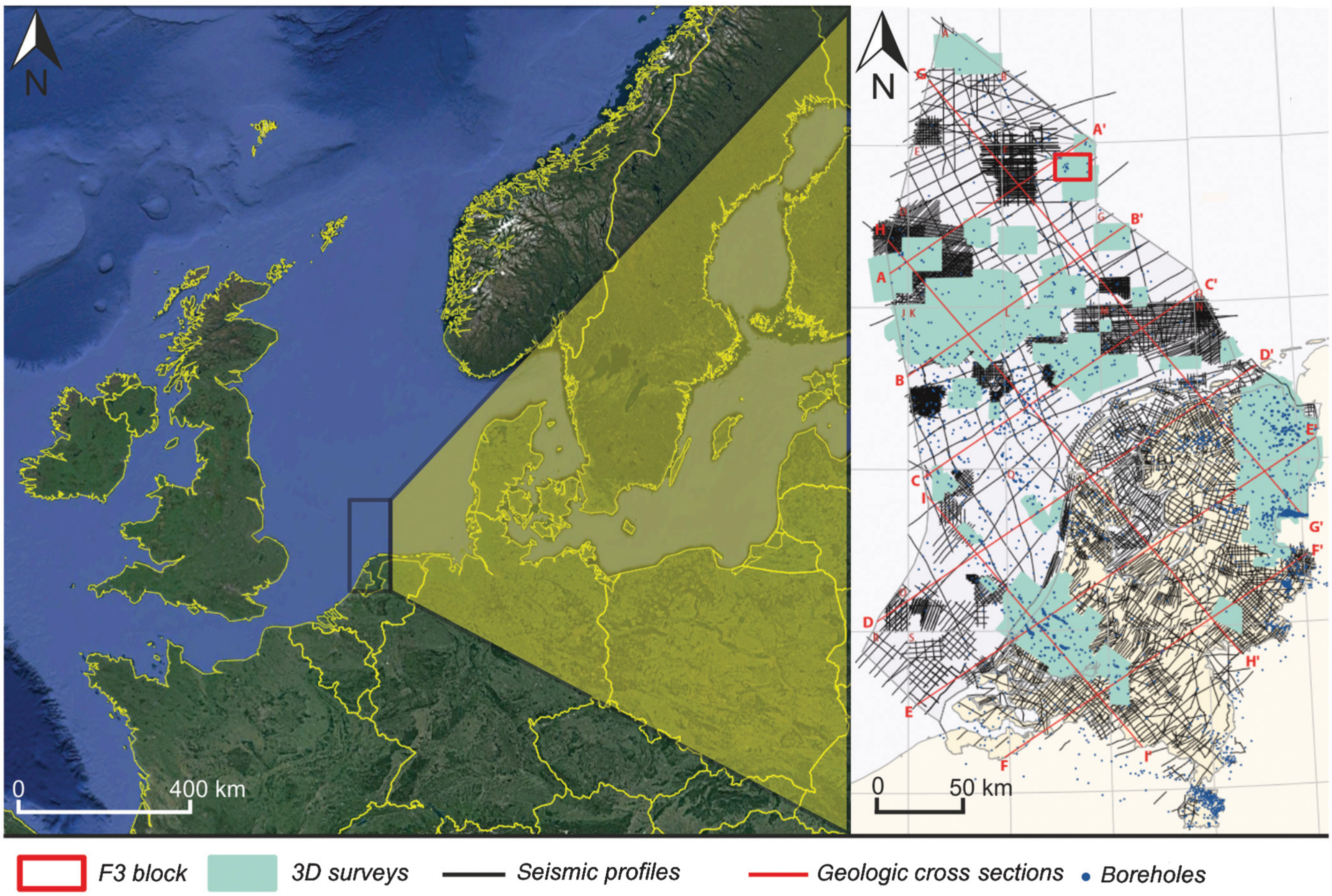

Figure 1. The location of the F3 block. Adapted from Duin et al. (2006).

In 1987, the F3 block 3D seismic survey was conducted to identify the geologic structures of this area and to search for hydrocarbon reservoirs. In addition, many boreholes were drilled within the F3 block throughout the years. The F3 block became one of the most widely known and studied seismic surveys after dGB Earth Sciences made the data obtained from the survey publicly available.

The aim of this section is to briefly describe the geology of the survey area and introduce the 3D geologic model that we have developed and how it was obtained.

The geology of the F3 block

Within the continental shelf of the North Sea, 10 groups of lithostratigraphic units have been identified in the literature (Van Adrichem Bogaert and Kouwe, 1993; Mijnlieff, 2002; Scheck-Wenderoth and Lamarche, 2005; Duin et al., 2006). These groups and their main lithostratigraphic features are listed below from newest to oldest:

1) Upper North Sea group: claystones and sandstones from Miocene to Quaternary
2) Lower and Middle North Sea groups: sands, sandstones, and claystones from Paleocene to Miocene
3) Chalk group: carbonates of Upper Cretaceous and Paleocene
4) Rijnland group: clay formations with sandstones of Upper Cretaceous
5) Schieland, Scruff, and Niedersachsen groups: claystones of Upper Jurassic and Lower Cretaceous
6) Altena group: claystones and carbonates of Lower and Middle Jurassic
7) Lower and Upper Germanic Trias groups: sandstones and claystones of Triassic
8) Zechstein group: evaporites and carbonates of Zechstein
9) Upper and Lower Rotliegend groups: siliceous rocks and basalts of the Lower Zechstein
10) Limburg group: Upper carboniferous siliceous rock, which are the bedrock for hydrocarbons.

The F3 block is located on the border of two tectonic structures: the Step Graben and the Dutch Central Graben (see Figure 2). These tectonic structures are characterized by different lithostratigraphic units of varying thickness. These varying thicknesses are a result of tectonic activity (Ziegler, 1988, 1990) that started in the Variscan orogeny (Schroot and De Haan, 2003). The area within the Step Graben is strongly disturbed by salt

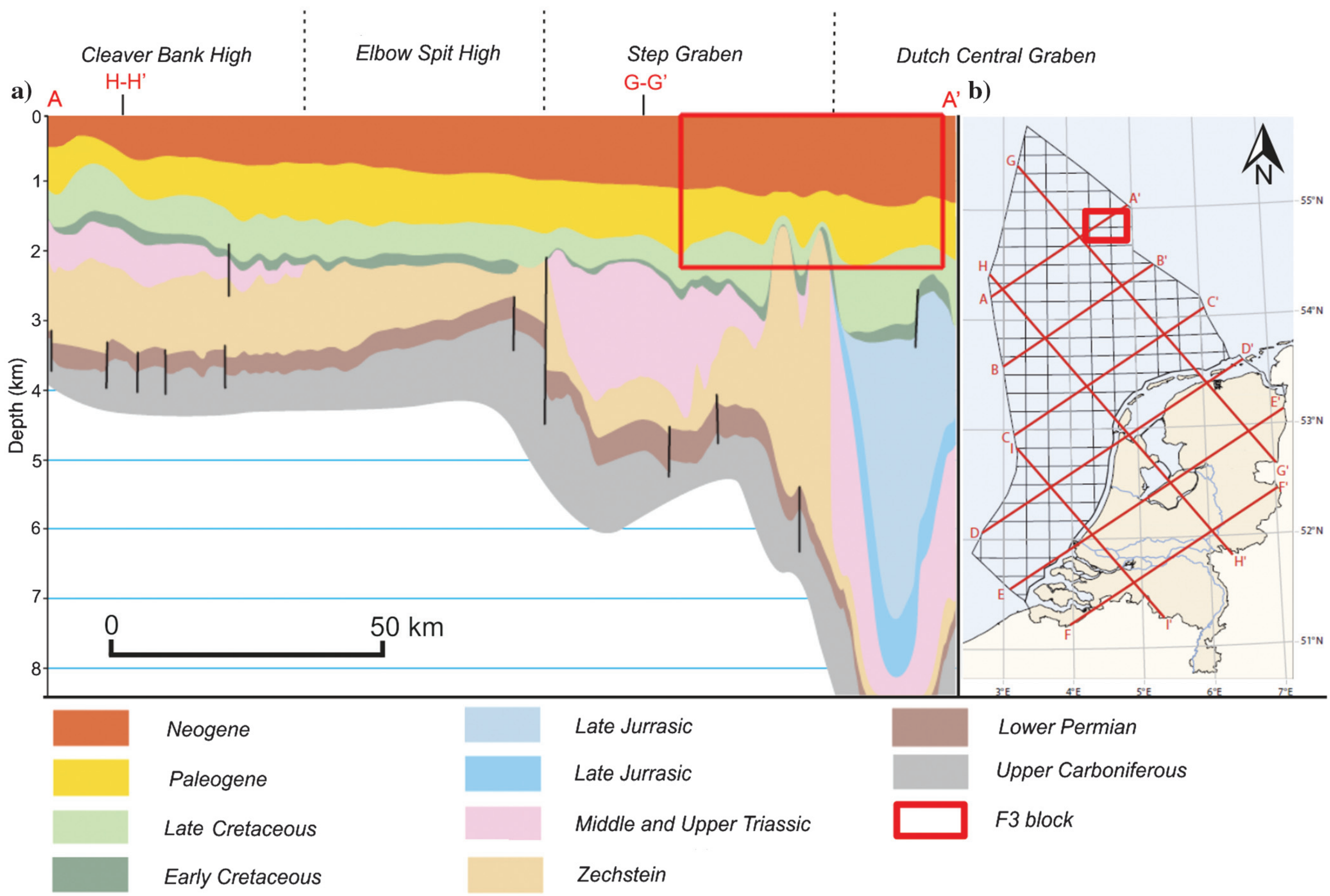

Figure 2. (a) A geologic cross section of the North Sea continental shelf along axis A-A' and (b) a map of the location of the cross section. Adapted from Duin et al. (2006).

diapirs, which were active several times, from the Zechstein to the Paleogene periods (Remmelts, 1996). On the other hand, and as a result of subsiding Jurassic rocks, the Altena, Scruff, Schieland, and Niedersachsen groups are observed only within the Dutch Central Graben (Duin et al., 2006).

The modeling process

To prepare our 3D geologic model of the F3 block, we relied on well logs and 3D seismic data. The next two subsections describe this process.

3D model building using well-log data

The well-log data were obtained from a website managed by the Geological Survey of the Netherlands (NLOG, 2019). The data (including information related to coordinates, true vertical depth, measured depth along the curvature, inclinations, and individual horizons) were collected for 26 boreholes located within the F3 block or its vicinity. The exact locations of these wells are shown in Figure 3.

Originally, the 26 wells contained 40 different horizons, so it was necessary to assign these different horizons to the various lithostratigraphic units that were adopted in the literature and were presented in the previous subsection. The next step was correlating wells with each other. After that, it was possible for us to create a preliminary 3D model based on the well-log data by using Petrel's *make/edit surface* tool. This process facilitated the preliminary visualization of the range of individual horizons, which was very helpful in the further interpretation of the 3D seismic data.

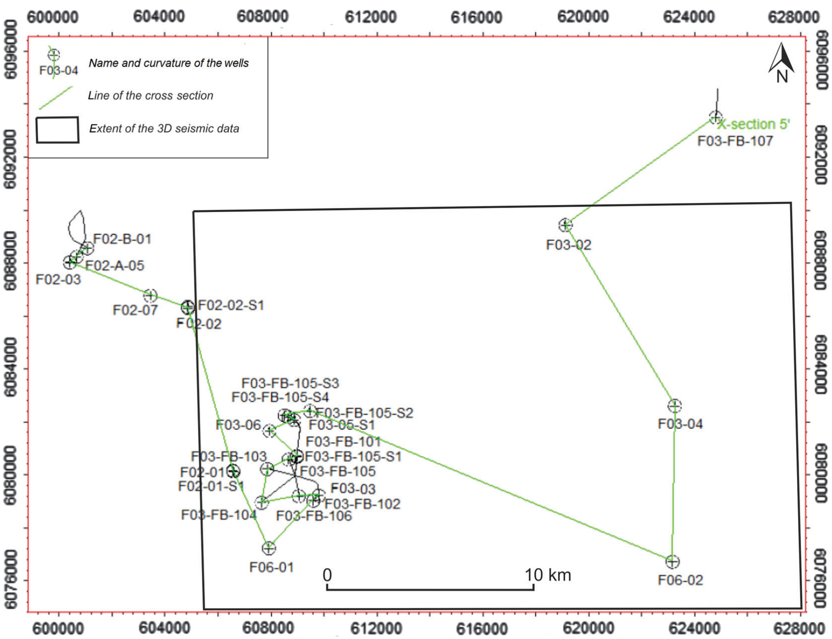

Figure 3. Locations of the boreholes that were used to create the geologic model.

3D model building using seismic data

The F3 block data were migrated in time, not depth, so it was necessary for us to do time-depth conversion because the structural model must be prepared in the depth domain. We used OpendTect 5.0 to perform the time-depth conversion using a velocity model that was provided with the F3 block data.

The next step in creating the model was fault-surface interpretation. Using Petrel's *polygon editing* tool, we interpreted the main fault surfaces and created the fault networks using the *fault framework modeling* tool. We interpreted horizons in a similar fashion, but by using the *seeded 3D autotracking* tool, which interpolated data automatically and took into account the faults networks modeled previously.

Based on the interpreted horizons and faults, we conducted preliminary modeling. We used the *horizon modeling* tool with *volume-based modeling*, which is an advanced method of isochronous geologic space modeling. The preliminary model included several imperfections in the interpretations of horizons and faults, so it was necessary for us to remodel several faults and conduct small corrections in the interpreted horizons.

After this, it was possible to create the final 3D model that highlights the regions between individual horizons. Here, we used Petrel's *structural modeling* module in the *horizon modeling* tool in addition to the *create zone model* function. The final 3D geologic model is shown in Figure 4.

The 3D geologic model

Within our 3D geologic model of the F3 block, we identified seven groups of lithostratigraphic units (see Figure 4) (from newest to oldest): the Upper North Sea group, the Middle North Sea group, the Lower North Sea group, the Chalk group, the Rijnland group, the Scruff group, and the Zechstein group.

These groups can be divided into three structural levels: *Cenozoic* (Lower, Middle, and Upper North Sea groups), *Mesozoic* (Scruff, Rijnland, and Chalk groups), and *Permian* (Zechstein group).

As is evident in Figure 4, the F3 block is characterized by highly variable geologic structures, in the horizontal and vertical ranges, which is manifested by the differential thicknesses of individual units and by the expanded fault network related to the salt tectonics. The area of the F3 block can be divided into two regions: eastern and western. The eastern region is disturbed by the occurrence of Zechstein diapirs and irregular fault networks. The western region is characterized by regular fault networks and a more uniform thickness of lithostratigraphic units.

The *Upper North Sea group* is the youngest and the flattest lithostratigraphic unit within our model. The top of the Upper North Sea group is the bottom of the North Sea at the same time, which is approximately −40 m above sea level (m a.s.l). Differences in the depth of the ocean floor are small, and they are maximally 6 m within the whole F3 block. It can be noted that the depth of this top decreases from the southwest to the northeast. The thickness of the Upper North Sea group varies from approximately 1000 m (in places deformed by Permian diapirs) to approximately 1320 m in the northern part of the research area (see Figure 4).

Below the Upper North Sea group lays the *Middle North Sea group*. The depth of the top of this unit ranges from −1000 m a.s.l. within the diapir in northeast part of the F3 block to approximately −1360 m a.s.l. in the northern part of this area, between diapirs. The thickness of the Middle North Sea group is from 20 to 150 m. As in the case of the Upper North Sea group, there is a clear relationship between the occurrence of Zechstein salts and the depth and thickness of this unit. Differences in the thickness of this unit between both sites of faults are also visible.

The next unit is the *Lower North Sea group*. This unit contains similar lithostratigraphic units to the Middle North Sea group, but it is visually distinct in the seismic data. The top is at a depth from −1100 m a.s.l., whereas the thickness is from approximately 180 to 750 m.

The top of the *Chalk group* is at a depth from −1300 m a.s.l. (above the diapirs in the northeast part of the survey) to −2100 m a.s.l. (in the eastern part of the survey, which is undisturbed by diapirs). The minimum thickness of this unit is 25 m, whereas above the salt diapirs in the northeast part of the F3 block, this substantially increases to 525 m.

The *Rijnland group* is submerged in the north–northeast direction, whereas it is the shallowest in the southwest part of the F3 block and above some Zechstein diapirs at the center of the survey. The maximum thickness of the Rijnland group is approximately 200 m (above some diapirs), whereas in the other parts of the F3 block it can be less than 20 m or it does not occur at all.

The *Scruff group*, similar to the Rijnland group, is thinned out in the north–northeast direction, more or less in the middle of the F3 block, in which the top of this layer has a depth of −2180 m a.s.l. This layer is shallowest (−1500 m a.s.l.) in the southwest part of the F3 block and above the Zechstein diapirs in

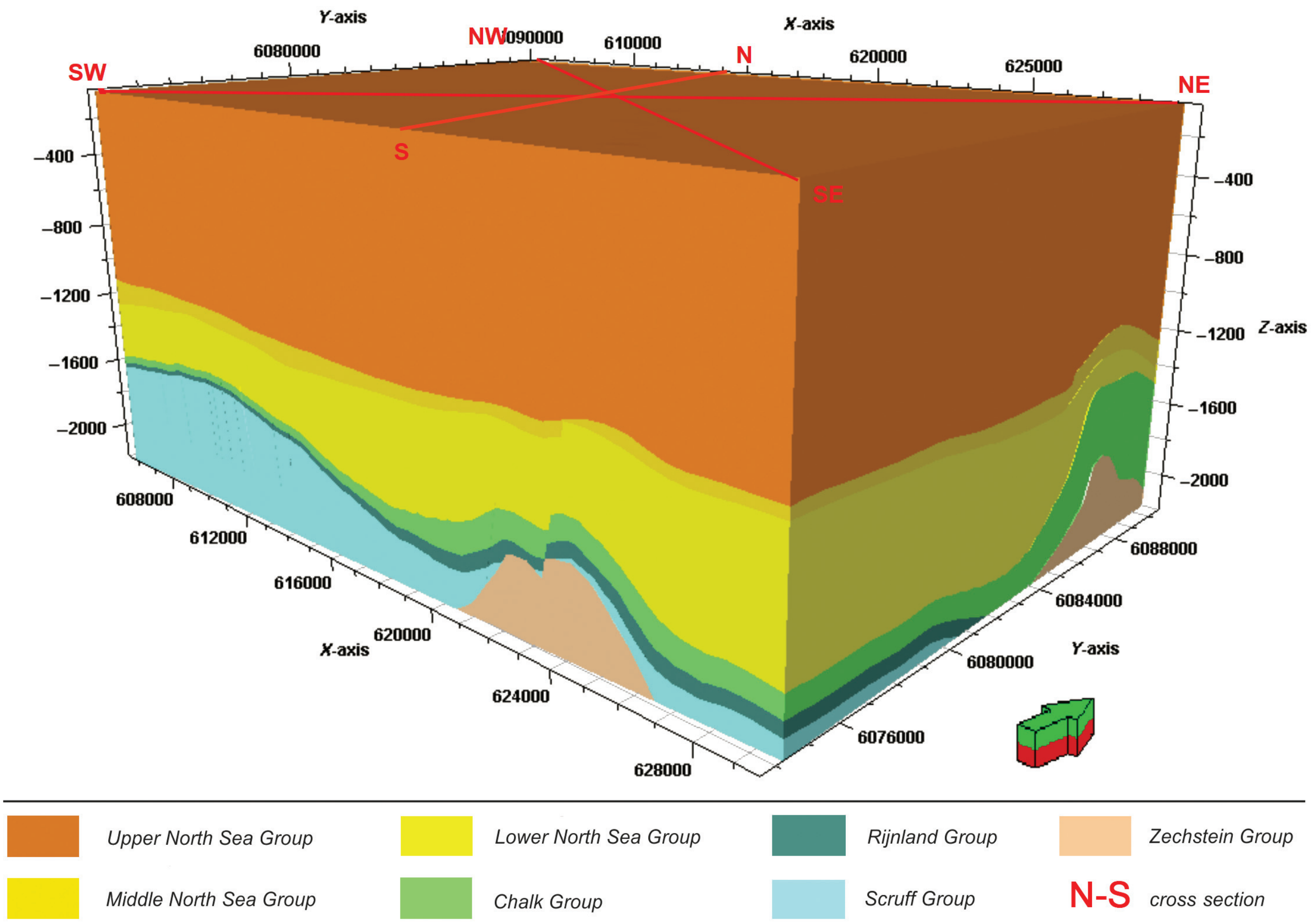

Figure 4. A 3D view of our geologic model of the F3 block.

the southern part of the survey. The thickness of the Scruff group within our model boundaries ranges from 100 m to almost 700 m, but it is much larger in reality and can reach several kilometers (Duin et al., 2006).

The *Zechstein group* occurs only in the eastern part of the survey as irregularly shaped salt diapirs. The shallowest part of the Zechstein group is at a depth of −1500 m a.s.l., whereas the maximum thickness of the Zechstein group within the research area is approximately 700 m. However, as in the case of the Scruff group, the depth is much bigger. According to the literature, it can reach several kilometers (Duin et al., 2006).

In addition to the identified groups of lithostratigraphic units mentioned above, we have also identified three generations of faults. The first generation are reverse, oblique-slip, sinistral faults with an south-southwest–north-northeast orientation. This direction is connected with the course of the tectonic axis of the Dutch Central Graben, which (similar to the whole Graben) has an south-southwest–north-northeast orientation. The second generation of faults are normal, oblique-slip, dextral faults with a west–east orientation. Finally, the third generation are faults that are genetically linked with faults from the first and second generations but were disturbed by the Permian halokinesis. Figure 5 shows an overhead view of the three generations of faults that we have identified. In addition, Figure 6 shows two diagonal cross sections along the southwest–northeast and northwest–southeast axes in our 3D model shown in Figure 4.

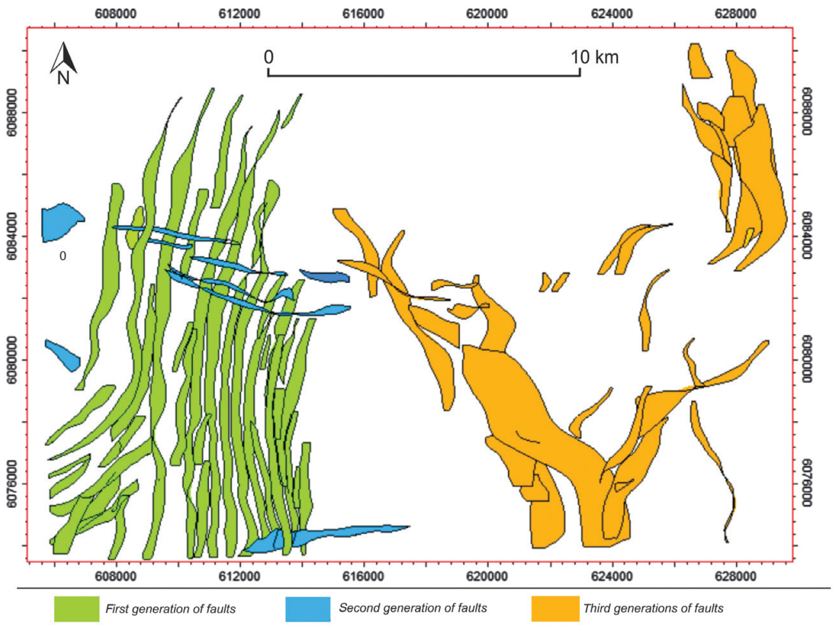

Figure 5. An overhead view of 3D fault planes from three different generations of faults that we have identified in the F3 block.

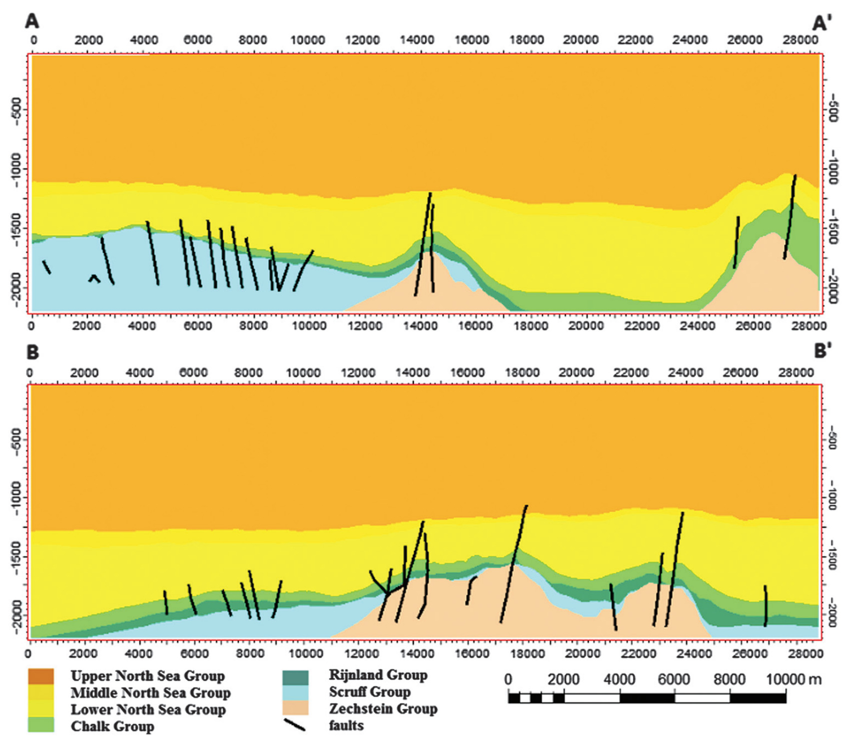

Figure 6. Two diagonal cross sections of our 3D geologic model in Figure 4.

Deconvolution network baseline

In addition to the geologic model we have introduced in the last section, we propose two baseline models for facies classification based on a deconvolution network architecture. In this section, we introduce deconvolution networks and describe our two baseline models.

Deconvolution networks

CNNs have seen great success in a wide number of visual applications, from image classification to semantic labeling. In the previous few years, it was not well-established that end-to-end CNN architectures can perform very well in semantic labeling tasks (such as facies classification). A major hurdle to the success of end-to-end CNN architectures in these tasks was what seemed like a trade-off between classification and localization accuracy. Deeper networks that have many convolution and pooling layers have proven to be the most successful models in image classification tasks. However, their large receptive fields and increased spatial invariance (due to pooling and convolutional layers) make it difficult to infer the locations of various objects within the image. In other words, the deeper we go into a network, the more it seems that we lose the location information of various objects within the image. Some researchers have attempted to overcome this hurdle by using various pre or postprocessing techniques. However, the introduction of fully convolutional network architectures, such as fully convolutional neural (FCN) network (Long et al., 2015) and DeconvNet (Noh et al., 2015) have shown that it is possible to achieve good semantic labeling results using a convolutional network only, with no pre or postprocessing steps required. FCN accomplish this by replacing the fully connected layers of the CNN with 1D convolutional layers that produce coarse feature maps. These coarse feature maps are then upsampled, and concatenated with the scores from intermediate feature maps in the network to generate the output. These upsampling steps, however, result in a blurred output that loses some of the resolution of the original image.

Deconvolution networks, on the other hand, overcome this problem by using a symmetric encoder-decoder style architecture composed of stacks of convolution and pooling layers in the encoder and stacks of deconvolution and unpooling layers in the decoder that mirror the encoders architecture. The role of the encoder can be seen as doing object detection and classification, whereas the decoder is used for accurate localization of these objects within the image. This architecture can achieve finer and more accurate results than those of the FCN, and therefore is adopted in our work.

A few recent papers have illustrated the successful application of deconvolution networks for seismic interpretation applications (Alaudah et al., 2018; Di et al., 2018). Figure 7 illustrates the architecture of the deconvolution network used for both of our baseline models. Every convolution or deconvolution layer (in white) is followed by a rectified linear unit nonlinearity. The layers in red perform 2×2 max pooling to select the maximum filter response within small windows. The indices of the maximum responses for every pooling layer are then shared with their respective unpooling layers (in green) to undo this pooling operation and get a higher resolution image.

Baseline models

In this work, we use two baseline models: a patch- and a section-based model. These two models use the exact same architecture, optimizer, and hyperparameters but differ in the way they are trained and the way they are used to label the seismic volume.

Patch-based model

We train the patch-based model on small patches extracted from the inlines and crosslines of the training data. For very large seismic volumes, this approach can be more feasible than using entire sections for training. During the training, the patches of seismic data and their associated labels are sampled randomly from the inlines and crosslines of the training set. During the testing, the model samples overlapping patches in the inline and crossline directions and averages the results to generate a 2D labeled version of the test inline or crossline. This is done for all inlines and crosslines in the test sets.

Section-based model

We train the section-based model on entire inline and crossline sections. The advantage of this approach is twofold. First, because the network is fed an entire section, it can easily learn the relationships between different lithostratigraphic units and can take the depth information into account when labeling the section. The second advantage is more practical. Training and testing entire sections at once means that the network can be trained or tested very quickly because there are only a relatively small number of seismic inlines and crosslines. (This is assuming the GPU memory is large enough to handle the size of the seismic sections. On our NVIDIA Titan X GPU, we trained the baseline section-based network — eight sections at a time — in approximately 70 min.) One advantage of using a fully convolutional architecture (such as the one we are using) is that the size of the network input does not have to be fixed. The size of the output of the network changes as the size of its input changes. Therefore, the different size of the inline and crossline sections does not pose any problem to the training of this network. (Although the sizes of the inlines and crosslines do not need to match, their resolutions [in terms of meters/pixel] should. In our case, pixels in the inline and crossline directions are 25×25 m.)

Other variations

In addition to the baseline patch- and section-based models, we have trained other variations of these models to test how they can be improved. We have tested the following variations:

- Baseline + data augmentation: Data augmentation is a technique that is used to increase the size of the training set artificially. Data augmentation applies different label-preserving transformations to the training data such as small rotations, random horizontal flipping, and the addition of Gaussian noise. This can help increase the training sample size and help the network generalize better to the test data.

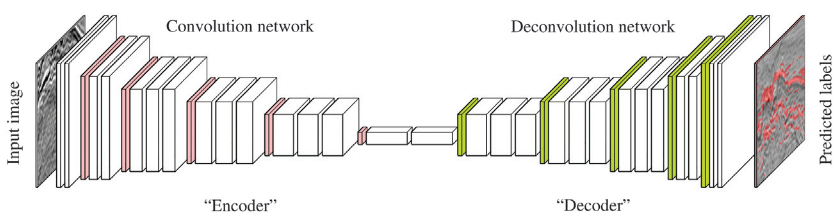

Figure 7. The architecture of the deconvolution network used in this work. The white layers are convolution or deconvolution layers. The red layers are max-pooling layers, whereas the green layers are unpooling layers.

- Baseline + data augmentation + skip connections: We further improve on the previous model by adding skip connections. In a deep neural network, the output of a layer is typically passed on as the input to the next layer in the network. Skip connections allow the output of a layer to be also passed as an input to a layer farther up the network, skipping intermediate layers in the process. These connections are implemented by directly adding the outputs of various layers in the encoder part of the deconvolution network to the outputs of the corresponding layers in the decoder. Skip connections help networks overcome the vanishing gradient problem (Hochreiter et al., 2001) by providing "shortcuts" for the computed gradients to propagate to the lower layers of the network.

Experimental setup

In this section, we will introduce the main elements of the experimental setup, including how the final geologic model was produced, how the model is split into training and testing sets, and what metrics are used to objectively evaluate the performance.

The geologic model

The final geologic model that we use to train and test our models is not the entire volume shown in Figure 4. The time-depth conversion process of the seismic data resulted in some artifacts. These artifacts were concentrated along the sigmoidal structure in the Upper North Sea group. Due to these artifacts, and missing data on the sides of the survey, we only use the data between inlines 100 and 701, crosslines 300 and 1201, and depth between 1005 and 1877 meters. Furthermore, we combine the Rijnland and the Chalk groups in our final model to a single class due to various issues with processing the Rijnland/Chalk boundary when generating the final model. Table 1 shows the percentage of different classes in our training set.

In addition to the final model labels and seismic data, we also release the original horizons for all the lithostratigraphic units, in addition to the extracted fault planes from all three generations.

The train/test split

Careful selection of the training and testing sets is crucial in any machine-learning application. This is especially true in seismic data, in which neighboring sections are highly correlated. Selecting the training and testing sections randomly will lead to artificially good test results that are not representative of the actual generalization performance of the tested models. Therefore, it is important to minimize the correlation between the training and testing sets as much as possible. It is also important to ensure that the training and testing sets have adequate representation of all the classes in the data set.

Therefore, we decide to split the data as shown in Figure 8. Namely, the data are split into the following three sets:

1) **Training set:** This includes all the data in the range of inlines [300,700] and crosslines [300,1000].
2) **Test set 1:** This set includes all the data in the range of inlines [100,299] and crosslines [300,1000].
3) **Test set 2:** This set includes all the data in the ranges of inlines [100,700] and crosslines [1001,1200]. This set includes a large Zechstein diapir in the northeast of the survey that is never seen in the training set.

For a fair comparison with others who might use this benchmark in the future, it is important to note that the test sets *should not be used more than once*. Testing a model on the test set, then retraining that model with different parameters means that the test set has been used for validation, which defeats its purpose.

Evaluation metrics

To objectively evaluate the performance of different models on our two test sets, we use the following metrics: pixel accuracy (PA), class accuracy (CA) for each individual class, mean CA (MCA) for all classes, and frequency-weighted intersection over union (FWIU). These metrics are detailed in Appendix A.

Results

After we created the final geologic model, we train each of the models described earlier on the training set until its training loss converges. We note that on our Nvidia Titan X GPU, the baseline patch-based model, and the augmented version, converged after 16 h of training. The patch-based model with skip connections required less than 5 h to converge. The section-based models required significantly less time, all of them converging in less than 90 min. We test these models by using them to label all inlines and crosslines in both test sets, and computing the performance metrics on the final result. Table 2 summarizes the objective results for all the models that we have tested on both test sets (the results for test set 1 and test set 2 are available from Alaudah, 2019b), whereas Figure 9 shows inline 200 of test set 1 labeled using the six different models that we have tested. In the remainder of this section, we will discuss these results, and we suggest various methods to improve upon them.

Table 1. The percentage of pixels from different classes in the training set.

Zechstein	Scruff	Rijnland/Chalk	Lower N. S.	Middle N. S.	Upper N. S.
1.48%	3.17%	6.53%	48.44%	11.89%	28.49%

Patch-based versus section-based models

Because the patch-based models are trained on patches from different depths in the data, they can easily confuse various classes that typically exist at different depths. In addition, the larger classes have a more diverse visual appearance compared with the smaller ones; therefore, it would be easier for the network to confuse features from smaller classes with those learned from larger ones (as shown, e.g., for the Lower North Sea group column in the confusion matrices of Figure 10); this indicates that, despite our careful choices, the intraclass variability of our classes is much different from one class to the other. Overall, as Figure 10 shows, although the patch-based model confuses many classes in our test sets with the Lower North Sea group, the section-based model performs better and does not confuse these classes as often.

Table 2 shows that the patch- and section-based models perform fairly well on the North Sea groups,

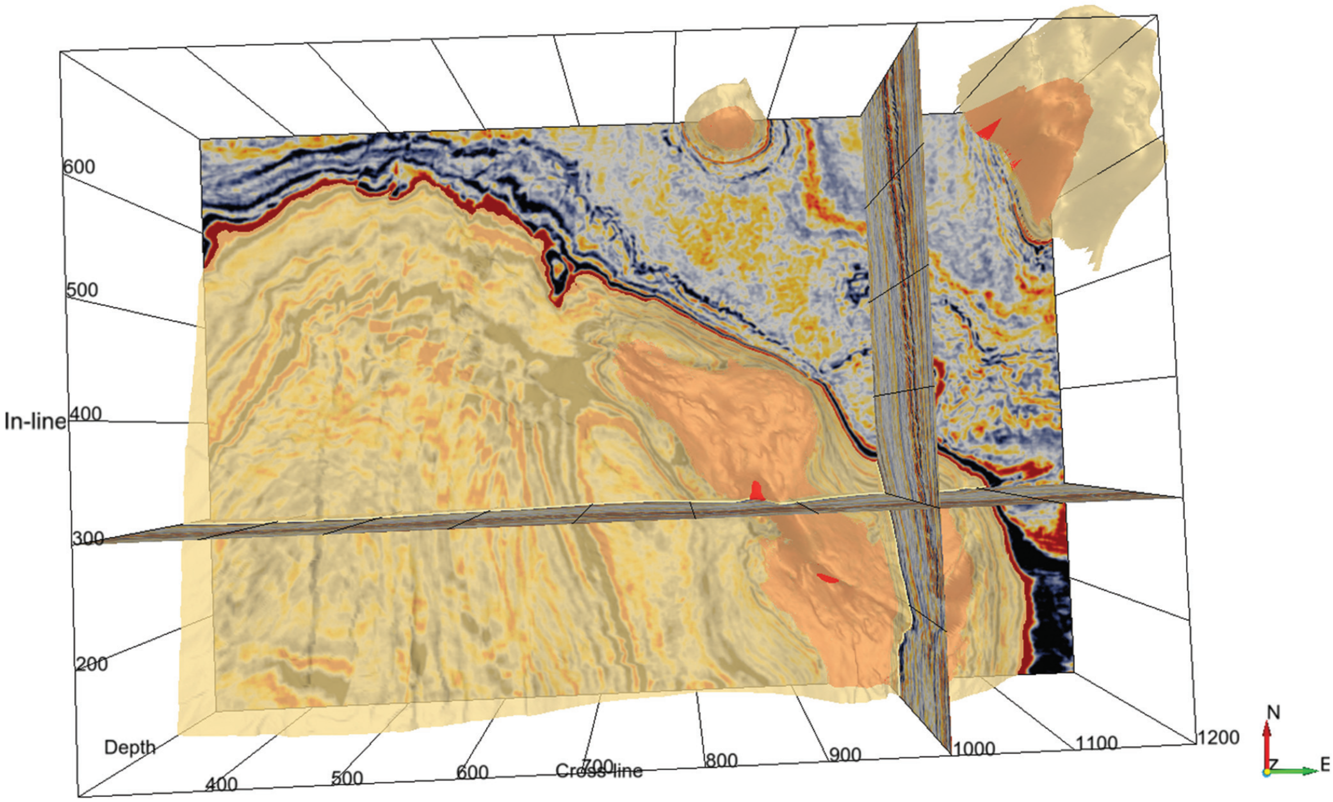

Figure 8. A 3D view of the F3 block from above with the Zechstein group shown in red, whereas the Rijnland/Chalk group is shown in a semitransparent beige color. Inline 300 and crossline 1000 divide the survey into four regions. The northwest region of the survey is used for training, whereas the southwest region constitutes test set 1. The remaining region east of crossline 1000 constitutes test set 2.

Table 2. Results of our two baseline models, with other variations, when tested on both test splits of our data set.

Model \ Metric	PA	Class Accuracy						MCA	FWIU
		Zechstein	Scruff	Rijnland/Chalk	Lower N. S.	Middle N. S.	Upper N. S.		
Patch-based model	0.788	0.264	0.074	0.499	**0.992**	0.804	0.754	0.565	0.640
Patch-based + aug.	0.852	0.434	0.221	0.707	0.974	0.884	0.916	0.689	0.743
Patch-based + aug + skip	0.862	0.458	0.286	0.673	0.974	0.912	0.926	0.705	0.757
Section-based model	0.879	0.219	0.539	0.744	0.951	0.872	0.973	0.716	0.789
Section-based + aug.	0.901	**0.714**	0.423	**0.812**	0.979	**0.940**	0.956	0.804	**0.844**
Section-based + aug + skip	**0.905**	0.602	**0.674**	0.772	0.941	0.938	**0.974**	**0.817**	0.832

Note: All metrics are in the range [0,1], with larger values being better. The best-performing model for every metric is highlighted in bold.

with the section-based models performing better. However, for smaller classes such as the Scruff and Zechstein groups, the section-based models show a clear advantage. The MCA score shows a 15% improvement of the section-based baseline model versus the patch-based model. Overall, section-based models are superior to patch-based models due to their ability to incorporate spatial and contextual information within each seismic section. They also have the advantage of being faster to train and test.

Imbalanced classes

As Table 1 shows, our data set is highly imbalanced. The Zechstein and Scruff groups are far smaller than the Lower or Upper North Sea groups. This means that while training, the network is trained on far more examples of Lower or Upper North Sea groups than the Zechstein or Scruff groups, for example. This leads to the networks being biased toward classifying pixels as Lower or Upper North Sea groups; therefore, we artificially achieve high CA scores for those classes. However, this is at the expense of the very poor performance for the smaller classes. In addition, the larger classes have a more diverse visual appearance compared with smaller ones; therefore, it would be easier for the network to confuse features from smaller classes with those learned from larger ones.

We do not make any changes to the baseline models to overcome this imbalance. However, using various techniques to overcome this class imbalance can significantly improve the results, especially for the smaller classes, such as the Zechstein and Scruff groups.

Data augmentation

Using data augmentation significantly improved the results for both baseline models, but especially for the patch-based model. The FWIU and MCA scores increased by more than 10% in the patch-based model, and it significantly improved the results for smaller classes such as the Zechstein and Scruff groups. The results of the section-based model

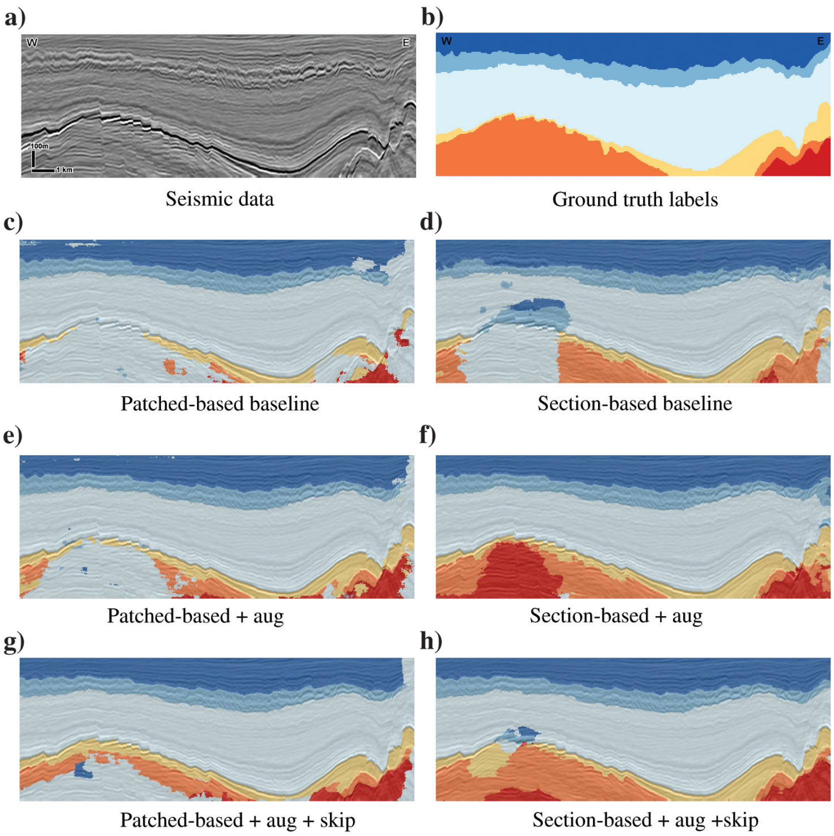

Figure 9. The results of the different models on inline 200 from test set 1. The color map is shown in Table 2.

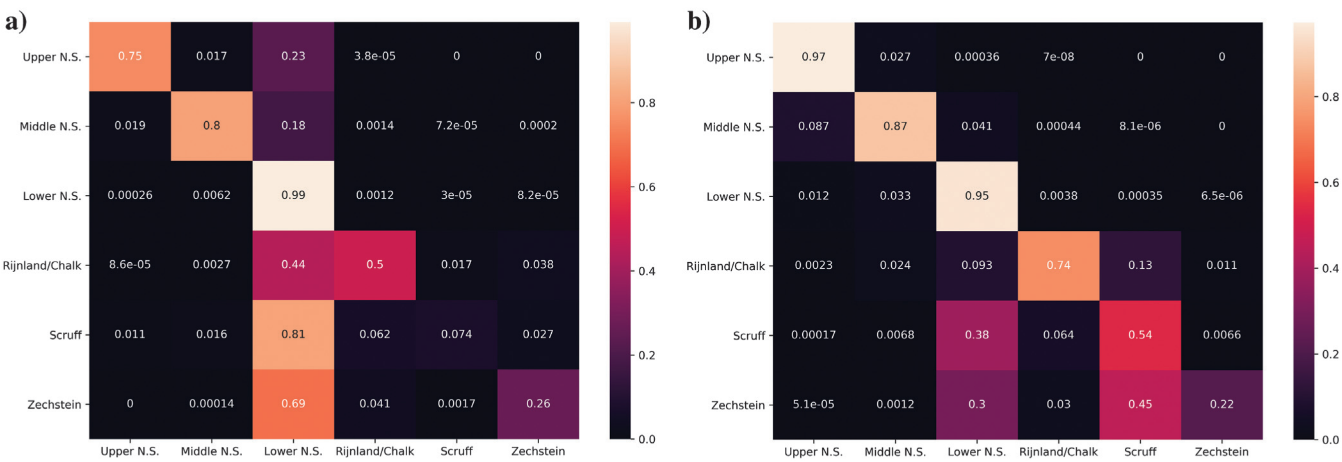

Figure 10. Confusion matrices for our two *baseline* models on test set 1 and 2. Each row shows the distribution of the model output for each class.

were also improved by using data augmentation, although to a lesser degree.

Skip connections

For patch- and section-based models, adding skip connections can improve the results, and speed up the training. This is especially noticeable in the patch-based model in which adding skip connections to the *baseline* + *aug* model improved the results by approximately 1% in the PA metric and approximately 1.5% in the MCA and FWIU metrics. The improvement in the results in the section-based models is more subtle because adding skip connections only improved the PA result by 0.1%. Interestingly, the Scruff group, which is the worst performing class in the patch- and section-based models, seemed to benefit the most from the addition of skip connections. The CA score for the Scruff group increased by 6.5% and 25% in the patch- and section-based models, respectively. Overall, adding skip connections seems to help improve the results. It can also speed up the training process. In the case of the patch-based model, the skip connection model converged four times faster than the baseline.

Conclusion

In conclusion, we have introduced and made publicly available a new annotated data set for facies classification. This data set includes six different lithostratigraphic classes based on the underlying geology of the Netherlands F3 block. The data also includes fault planes from three different generations that we have identified in the F3 block. In addition, we present two baseline deep-learning models for facies classification, a patch- and a section-based model, based on a deconvolution network architecture. We train these models using our data set, and we analyze their performance. Furthermore, we make the code for training and testing these models publicly available for others to use. It is our hope that this data set, and the code that we have released, will help facilitate more research in this area and help create an objective benchmark for comparing the results of different machine-learning approaches for facies classification.

Acknowledgments

We want to acknowledge the support of the Center for Energy and Geo Processing at the Georgia Institute of Technology and King Fahd University of Petroleum and Minerals. We would also like to thank dGB Earth Sciences for making the Netherlands F3 data publicly available and Schlumberger for providing educational licenses for the Petrel software.

Data and materials availability

Data associated with this research are available and can be accessed via the following URL: https://github.com/olivesgatech/facies_classification_benchmark.

Appendix A

Evaluation metrics

To objectively evaluate the performance of our models on this data set, we use a set of evaluation metrics that are commonly used in the computer vision literature. If we denote the set of pixels that belong to class i as G_i, and the set of pixels classified as class i as F_i, then, the set of correctly classified pixels is $G_i \cap F_i$. We use $|\cdot|$ to denote the number of elements in a set. Now, we can define the following metrics:

- *PA* is the percentage of pixels over all classes that are correctly classified,

$$\mathrm{PA} = \frac{\sum_i |F_i \cap G_i|}{\sum_i |G_i|}. \tag{A-1}$$

- *CA for class i* (CA_i) is the percentage of pixels that are correctly classified in a class i

$$\mathrm{CA}_i = \frac{|F_i \cap G_i|}{|G_i|}. \tag{A-2}$$

We will also define the *MCA* as the average of CA over all classes,

$$\mathrm{MCA} = \frac{1}{n_c}\sum_i \mathrm{CA}_i = \frac{1}{n_c}\sum_i \frac{|F_i \cap G_i|}{|G_i|}, \tag{A-3}$$

where n_c is the number of classes.

- *Intersection over union* (IU_i) is defined as the number of elements of the intersection of G_i and F_i over the number of elements of their union set,

$$\mathrm{IU}_i = \frac{|F_i \cap G_i|}{|F_i \cup G_i|}. \tag{A-4}$$

This metric measures the overlap between the two sets, and it should be one if and only if all pixels were correctly classified. Further, when we average IU over all classes, we arrive at the mean intersection over union (mean IU),

$$\text{Mean IU} = \frac{1}{n_c}\sum_i \mathrm{IU}_i = \frac{1}{n_c}\sum_i \frac{|F_i \cap G_i|}{|F_i \cup G_i|}. \tag{A-5}$$

To prevent this metric from being overly sensitive to small classes, it is common to weigh each class by its size. The resulting metric is known as *FWIU*:

$$\mathrm{FWIU} = \frac{1}{\sum_i |\mathcal{G}_i|} \cdot \sum_i |\mathcal{G}_i| \cdot \frac{|\mathcal{F}_i \cup \mathcal{G}_i|}{|\mathcal{F}_i \cap \mathcal{G}_i|}. \tag{A-6}$$

References

Alaudah, Y., 2019a, Georgia institute of technology, www.github.com/olivesgatech/facies_classification_benchmark, accessed 15 January 2019.

Alaudah, Y., 2019b, Georgia institute of technology, https://bit.ly/2VpRkIT, accessed 9 May 2019.

Alaudah, Y., M. Alfarraj, and G. AlRegib, 2019, Structure label prediction using similarity-based retrieval and weakly supervised label mapping: Geophysics, **84**, no. 1, V67–V79, doi: 10.1190/geo2018-0028.1.

Alaudah, Y., and G. AlRegib, 2016, Weakly-supervised labeling of seismic volumes using reference exemplars: IEEE International Conference on Image Processing (ICIP), 4373–4377.

Alaudah, Y., S. Gao, and G. AlRegib, 2018, Learning to label seismic structures with deconvolution networks and weak labels: 88th Annual International Meeting, SEG, Expanded Abstracts, 2121–2125, doi: 10.1190/segam2018-2997865.1.

Araya-Polo, M., T. Dahlke, C. Frogner, C. Zhang, T. Poggio, and D. Hohl, 2017, Automated fault detection without seismic processing: The Leading Edge, **36**, 208–214, doi: 10.1190/tle36030208.1.

Coléou, T., M. Poupon, and K. Azbel, 2003, Unsupervised seismic facies classification: A review and comparison of techniques and implementation: The Leading Edge, **22**, 942–953, doi: 10.1190/1.1623635.

de Matos, M. C., P. L. Osorio, and P. R. Johann, 2006, Unsupervised seismic facies analysis using wavelet transform and self-organizing maps: Geophysics, **72**, no. 1, P9–P21, doi: 10.1190/1.2392789.

dGB Earth Sciences, 1987, The Netherlands Offshore, The North Sea, F3 Block-Complete, https://bit.ly/2B44GO0, accessed 15 January 2019.

Di, H., Z. Wang, and G. AlRegib, 2018, Real-time seismic-image interpretation via deconvolutional neural network: 88th Annual International Meeting, SEG, Expanded Abstracts, 2051–2055, doi: 10.1190/segam2018-2997303.1.

Doornenbal, J. C., 2014, Kilka uwag o kartografii wglebnej i modelowaniu geologicznym wholandii: Przeglad Geologiczny, **62**, 806.

Dramsch, J. S., and M. Lüthje, 2018, Deep-learning seismic facies on state-of-the-art CNN architectures: 88th Annual International Meeting, SEG, Expanded Abstracts, 2036–2040, doi: 10.1190/segam2018-2996783.1.

Dubois, M. K., G. C. Bohling, and S. Chakrabarti, 2007, Comparison of four approaches to a rock facies classification problem: Computers and Geosciences, **33**, 599–617, doi: 10.1016/j.cageo.2006.08.011.

Duin, E., J. Doornenbal, R. Rijkers, J. Verbeek, and T. E. Wong, 2006, Subsurface structure of the netherlands — Results of recent onshore and offshore mapping: Netherlands Journal of Geosciences, **85**, 245–276, doi: 10.1017/S0016774600023064.

Hochreiter, S., Y. Bengio, P. Frasconi, and J. Schmidhuber, 2001, Gradient flow in recurrent nets: The difficulty of learning long-term dependencies, *in* J. F. Kolen and S. C. Kremer, eds., A field guide to dynamical recurrent networks: IEEE.

Huang, L., X. Dong, and T. E. Clee, 2017, A scalable deep learning platform for identifying geologic features from seismic attributes: The Leading Edge, **36**, 249–256, doi: 10.1190/tle36030249.1.

Long, J., E. Shelhamer, and T. Darrell, 2015, Fully convolutional networks for semantic segmentation: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440.

Mijnlieff, H., 2002, Top pre-Permian distribution map and some thematic regional geologic maps of the Netherlands: ICCP.

NLOG, 2019, Welkom bij NLOG, www.nlog.nl, accessed 15 December 2019.

Noh, H., S. Hong, and B. Han, 2015, Learning deconvolution network for semantic segmentation: Proceedings of the IEEE International Conference on Computer Vision, 1520–1528.

Qian, F., M. Yin, X.-Y. Liu, Y.-J. Wang, C. Lu, and G.-M. Hu, Unsupervised seismic facies analysis via deep convolutional autoencoders: Geophysics, **83**, no. 3, A39–A43, doi: https://doi.org/10.1190/geo2017-0524.1

Remmelts, G., 1996, Salt tectonics in the southern North Sea, the Netherlands, *in* H. E. Rondeel, D. A. J. Batjes, and W. H. Nieuwenhuijs, eds., Geology of gas and oil under the Netherlands: Springer, 143–158.

Rutherford Ildstad, C., and P. Bormann, 2017, MalenoV: Tool for training and classifying SEGY seismic facies using deep neural networks, https://github.com/bolgebrygg/MalenoV, accessed 10 September 2019.

Scheck-Wenderoth, M., and J. Lamarche, 2005, Crustal memory and basin evolution in the Central European Basin System — New insights from a 3D structural model: Tectonophysics, **397**, 143–165, doi: 10.1016/j.tecto.2004.10.007.

Schroot, B., and H. De Haan, 2003, An improved regional structural model of the upper carboniferous of the Cleaver Bank High based on 3D seismic interpretation: Geological Society, London, Special Publications 212, 23–37.

Shafiq, M. A., M. Prabhushankar, H. Di, and G. AlRegib, 2018, Towards understanding common features between natural and seismic images: 88th Annual International Meeting, SEG, Expanded Abstracts, 2076–2080, doi: 10.1190/segam2018-2996501.1.

Shi, Y., X. Wu, and S. Fomel, 2018, Automatic salt-body classification using a deep convolutional neural network: 88th Annual International Meeting, SEG, Expanded Abstracts, 1971–1975, doi: 10.1190/segam2018-2997304.1.

Van Adrichem Bogaert, H., and W. Kouwe, 1993, Stratigraphic nomenclature of the Netherlands, revision and update: Mededelingen Rijks Geologische Dienst, **50**.

Veillard, A., O. Morére, M. Grout, and J. Gruffeille, 2018, Fast 3D seismic interpretation with unsupervised deep learning: Application to a potash network in the North Sea: 80th Annual International Conference and Exhibition, EAGE, Extended Abstracts, doi: 10.3997/2214-4609.201800738.

Waldeland, A., and A. Solberg, 2017, Salt classification using deep learning: 79th Annual International Conference and Exhibition, EAGE, Extended Abstracts, doi: 10.3997/2214-4609.201700918.

Zhao, T., 2018, Seismic facies classification using different deep convolutional neural networks: 88th Annual International Meeting, SEG, Expanded Abstracts, 2046–2050, doi: 10.1190/segam2018-2997085.1.

Ziegler, P. A., 1988, Evolution of the Arctic-North Atlantic and the Western Tethys: A visual presentation of a series of paleogeographic-paleotectonic maps: AAPG Memoir, **43**, 164–196.

Ziegler, P. A., 1990, Geological atlas of western and Central Europe: Presented at theGeological Society of London, Shell International Petroleum Maatschappij.

Biographies and photographs of the authors are not available.