Computers & Geosciences 141 (2020) 104516

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

# 3D lithological mapping of borehole descriptions using word embeddings

Ignacio Fuentes[a,*], José Padarian[a], Takuya Iwanaga[b], R. Willem Vervoort[a]

[a] School of Life and Environmental Sciences, The University of Sydney, Sydney, 2006, Australia
[b] Fenner School of Environment & Society, Australian National University, Canberra, 0200, Australia

ABSTRACT

In recent years the exponential growth in digital data and the expansion of machine learning have fostered the development of new applications in geosciences. Natural Language Processing (NLP) tackles various issues that arise from using human language data. In this study, NLP is applied to classify and map lithological descriptions in a three dimensional space. The data originates from the Australian Groundwater Explorer dataset of the Bureau of Meteorology, which contains the description and geolocation of bores drilled in New South Wales (NSW), Australia. A GloVe model trained with scientific journal articles and Wikipedia contents related to geosciences was used to obtain embeddings (vectors) from borehole descriptions. In parallel, and as a baseline, the descriptions were classified combining regular expressions and expert criterion. The description embeddings were subsequently classified using a multilayer perceptron neural network (MLP). The performance was evaluated using different accuracy metrics. The embeddings were triangulated and the resulting embeddings were classified using the trained MLP and compared against a nearest neighbour (NN) interpolation of lithological classes. The mapping of the descriptions was carried out by using 3D voxels. Coupling NLP with supervised classification alternatives and interpolation methods resulted in reasonable 3D representation of lithologies. This methodology is a first step in demonstrating the applicability of NLP to the geosciences, which also allows for an uncertainty quantification in the different steps of the process, such as classification and interpolation. Interpolation techniques, although acceptable, might be replaced by machine learning techniques to improve the performance of 3D models.

## 1. Introduction

The digital era of the last decades has led to an exponential growth of information (Maarala et al., 2015). This has resulted in a change from analog to digital sources of information, which has been accompanied by increases in the storage capacity and computation power (Hilbert and López, 2011). Geosciences is one of the interdisciplinary fields of sciences that has suddenly changed since the beginning of the digital revolution. From being a poor data field it became a rich data field, integrating different sources of information such as remote sensing and geophysical surveys (Karpatne et al., 2017; Nativi et al., 2015).

Currently, sub-disciplines of geosciences, like geology, are reaching a stage of synthesis in relation to the information gathered. As pointed out by Culshaw (2005), traditional geological data collection from field survey campaigns is in decline. However, a new stage is emerging, in which new technologies allow the digitalization, storage, processing, synthesis and analysis of big (legacy) datasets.

The increasing amount of digital data has not only promoted the development of the different branches of science, but has also led to the formation of an entire new subfield of sciences, whose sole purpose is the detection of patterns in the data to solve problems (Boulton, 2018). This field, known as machine learning, has been widely applied to overcome difficulties caused by the use of big data.

While machine learning techniques have been used extensively in geosciences (Smirnoff et al., 2008; O'Brien et al., 2015; Lary et al., 2016), Natural Language Processing (NLP), which includes handling and analysing the relationships between words (Nadkarni et al., 2011; Jain et al., 2018), has seldom been applied. This is caused by a bias of scientific knowledge towards numerical data (McBratney et al., 2018). However, large amounts of geological and pedological information have been recorded as descriptions to categorise the materials and provide qualitative information to scientists. Neglecting this information due to the above mentioned bias lacks practicality. Furthermore, the advances in NLP and machine learning mean that the subjectivity and ambiguity introduced by language might be removed by text processing and analysis (Escudero, 2006; Recasens et al., 2013).

From the most common uses of NLP, dimensionality reduction, classification, and clustering of text are the most important, which have mainly been applied to advertising and the analysis of social media (Aggarwal and Zhai, 2012). More recently, text mining tasks have expanded to other research areas, including the fields of medicine and psychiatry, and it is expected to expand to many different fields of science that deal with textual descriptions of reality (Pestian et al., 2010; Perlis et al., 2011).

* Corresponding author.
*E-mail address:* ignacio.fuentes@sydney.edu.au (I. Fuentes).

https://doi.org/10.1016/j.cageo.2020.104516
Received 6 July 2019; Received in revised form 16 March 2020; Accepted 1 May 2020
Available online 23 May 2020

Geologic datasets contain many textual descriptions. For example, borehole drill descriptions also contain textual information of the characteristics of the underlying materials, apart from geospatial information. Due to the quantity of wells drilled, bore logs are usually one of the main data sources used for the synthesis of geological information in geologic models (Kaufmann and Martin, 2008). However, it is difficult to classify the amount of heterogeneous lithological descriptions which tend to be contained in short sentences, and are highly focused towards geologists. The semantic analysis of these descriptions is therefore constrained by the specialised geological lexicon. This makes NLP an interesting alternative to test the classification of bore log descriptions for the development of 3D geologic models.

The main difficulty of text mining is around the way in which text can be processed and analysed. It is clear that a collection of characters - referred to as strings - which make up the textual framework, cannot be used in isolation. Instead, NLP is used to pre-process and transform text into a numerical or network representation (Srivastava and Sahami, 2009). Several representation schemes have been proposed, but these have been mostly focussed on general domain text (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017), which usually perform poorly in domain-specific tasks. Padarian and Fuentes (2019) generated word embeddings specifically trained for geosciences, which will be used in this study.

The main objective of this study is the production of three-dimensional lithological maps developed by applying NLP techniques on a big dataset with borehole drilling descriptions.

## 2. Materials and methods

### 2.1. Study area

The focus of the study is New South Wales (NSW), Australia. NSW is one of the six states of Australia geologically characterized by several sedimentary basins, overlying the Ordovician to Early Cretaceous basement (Fig. 1). In the eastern part of Australia, orogens and fold belts resulting from multiple deformation events were eroded during exhumation and were the main source of sediments that filled the main sedimentary basins, such as the Sydney-Gunnedah-Bowen, Surat, Murray, Gloucester and Clarence Moreton basins (Fig. 2; O'Neill and Danis, 2013; Welsh et al., 2014).

### 2.2. Data and pre-processing steps

The main source of lithological data used in this study is the groundwater database obtained from the Australian Groundwater Explorer of the Bureau of Meteorology (refer to "Code and data availability" section at the end of the article). It contains the geolocation and the bore logs of all the boreholes drilled in NSW. The dataset contains 100,582 boreholes (Fig. 3) and 835,411 descriptions. Each borehole can have several descriptions associated to the different underlying lithologies. In addition, each description in the dataset is classified into a MajorLithCode, which corresponds to the major lithological classification in the dataset.

There are 549 different lithological classes in the dataset, several of which aggregate descriptions into ordinal categories, or aggregate descriptions based on adjectives which were less relevant for this study, such as colour or weathering state. For instance, different clay categories based on the colour of sediments can be found in the dataset, and even though these may inform the presence of different mineralogies, they would further complicate the lithological interpretability of descriptions. Another dataset characteristic is the heterogeneity of descriptions and the imbalance between lithological classes. Thus, 10 classes with multiple descriptions contain around 82% of the dataset descriptions, and the distribution of the length of sentences within the descriptions is highly right skewed (Fig. 4). Therefore, a reclassification of this dataset is required.

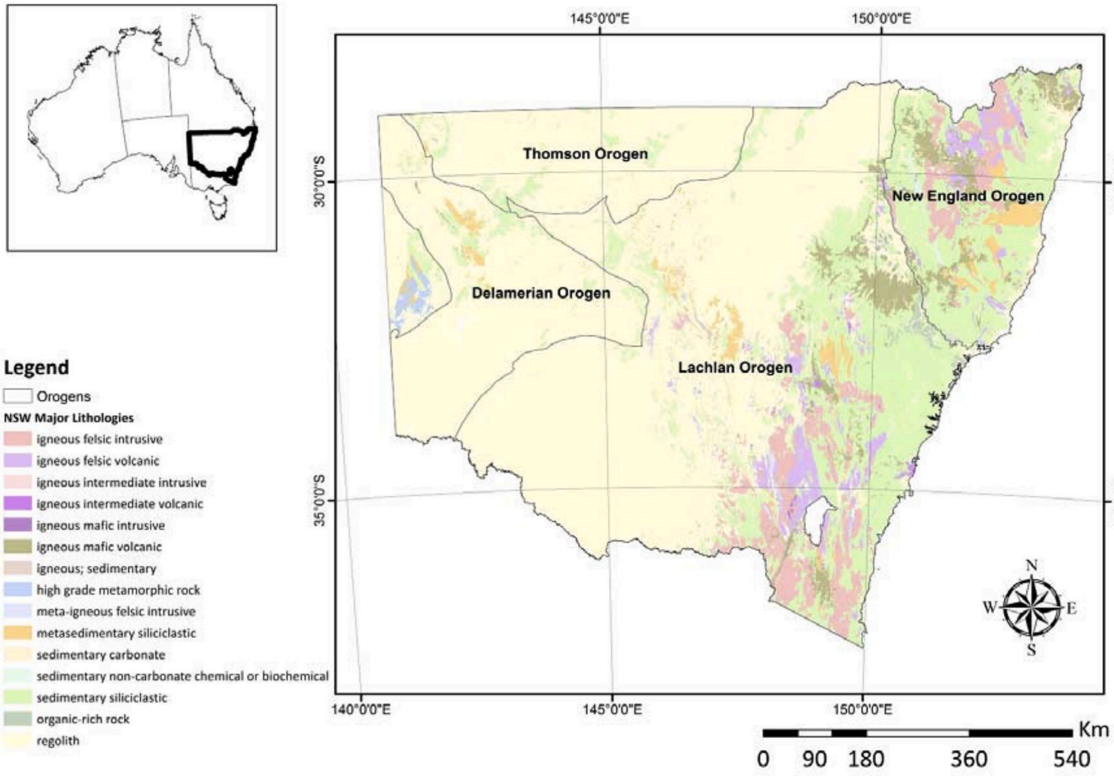

**Fig. 1.** Study area and lithologic framework depicting the different orogens that are the main source of sediments that filled the sedimentary basins in NSW. Different colours represent the different surface lithologies. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

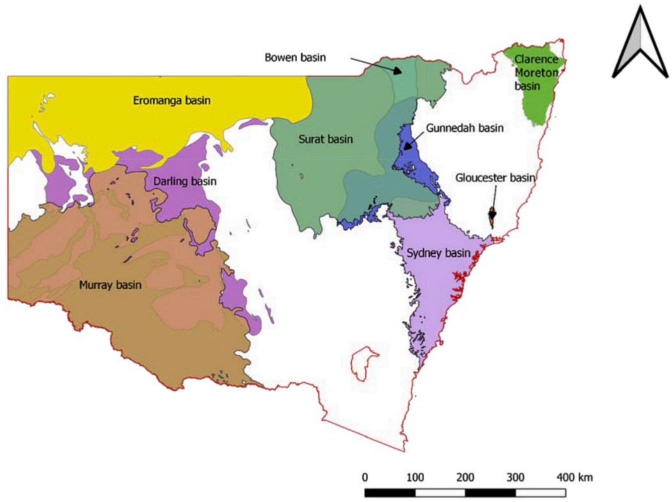

**Fig. 2.** Main sedimentary basins located in NSW. In the case of Murray and the Surat basins, these overlie the Darling and Sydney-Gunnedah-Bowen basins in some areas, respectively. The sedimentary basins were obtained from Stewart et al. (2013).

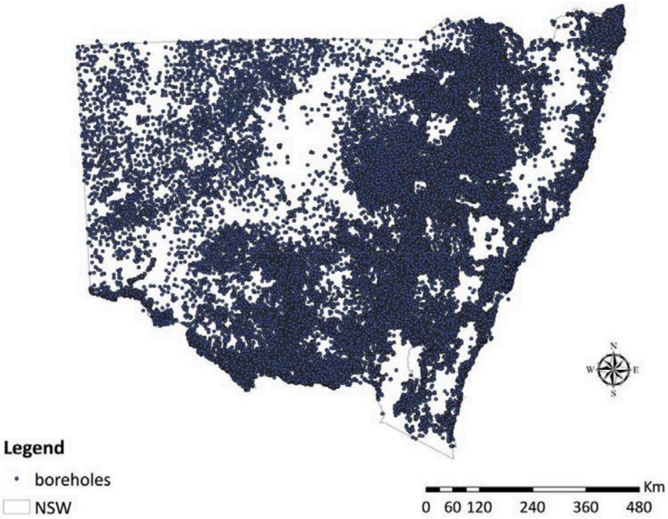

**Fig. 3.** Boreholes distribution in NSW.

The descriptions contained in the bore logs were pre-processed such that:

1. The set of descriptions was tokenized (divided into words) and lemmatization was applied to all nouns. This simply involves removing inflections at the end of words in order to get the “lemma” or root of the words. In this step, lists of tokens were obtained.
2. All tokens (words) with non-alphabetic characters and tokens with less than three characters were removed.
3. The remaining tokens were converted to its lowercase form.
4. Stopwords (a set of words frequently used in language which are irrelevant for text mining purposes) were removed.

A flowchart of the general methodology applied in this study to generate three-dimensional lithologic maps from borehole descriptions is presented in Fig. 5. The different steps in this study are discussed further in the following sections.

As additional supporting data, the Shuttle Radar Topography Mission (SRTM) digital elevation data was obtained for NSW at 30 m resolution. It was used in the mapping stage to clip the interpolated lithologies to the terrain surface.

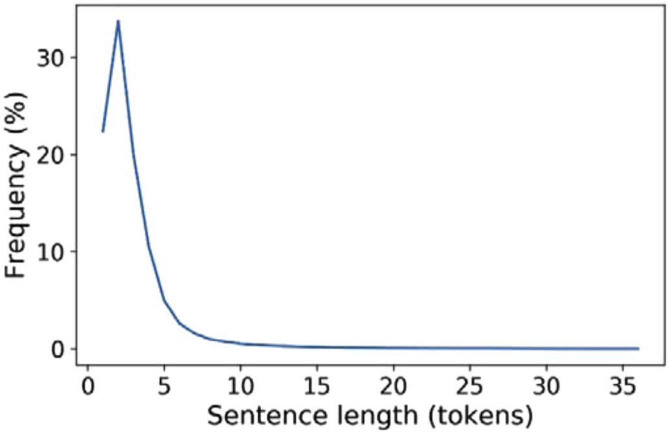

**Fig. 4.** Frequency of the descriptions length in the dataset used.

Models, machine learning techniques and text processing tools used in this study were developed through the GloVe (Pennington et al., 2014), scikit-learn (sklearn; Pedregosa et al., 2011), and the natural language toolkit (NLTK; Bird et al., 2009) libraries implemented in the Python programming language.

### 2.3. Descriptions to vectors (embeddings)

In order to obtain a vectorial (numerical) representation of the words contained in a description (i.e., an embedding), the GeoVec model developed by Padarian and Fuentes (2019) was applied to the words from the descriptions. The model consist of an application of the GloVe model (Pennington et al., 2014) to the geosciences domain, and was trained with a corpus of over 300,000 scientific full-text articles and over 1000 Wikipedia articles related to geosciences.

Since the unsupervised model relies on a matrix of co-occurrence between words, the distance and angles between words in the vectorial space generated from the corpus identify different relationships. Fig. 6 shows examples of how GeoVec depicts the semantic relationship between words, linking either specific lithologies to their corresponding rock type (Fig. 6 left panel), or minerals and their corresponding mineral group (Fig. 6 right panel). These groups can then be used to evaluate the numerical vectorial representation of words, which is an “embedding”.

Another property that can be obtained from this vectorial space is the interpolation between concepts. This means words between concepts can be found, which may have an empirical meaning. In Fig. 7 (left panel) it can be observed how different particle sizes are found and sorted between the “clay” and “boulder” words, leading to the final increasing sequence: clay < silt < sand < gravel < cobble < boulder. In Fig. 7 (right panel) a scale of metamorphic grade was found between two extreme terms, “slate” and “migmatite”, with the final increasing sequence: slate < phyllite < schist < gneiss < migmatite. In this study, we further explore this type of word embedding interpolation with an application to 3D spatial modelling.

However, since each description contains a list of words and since the GeoVec model creates one embedding for each word within a description, an average of the constituting word embeddings that comprise each description was calculated, yielding a single vector of length 300 for each lithological description. This is a simple and commonly used alternative to create representative texts (Pagliardini et al., 2018).

### 2.4. Classification of embeddings

For the supervised classification of embeddings, a semi-manual classification of the descriptions, combining the use of regular expressions and expert criteria, was carried out on a subset of over 700,000

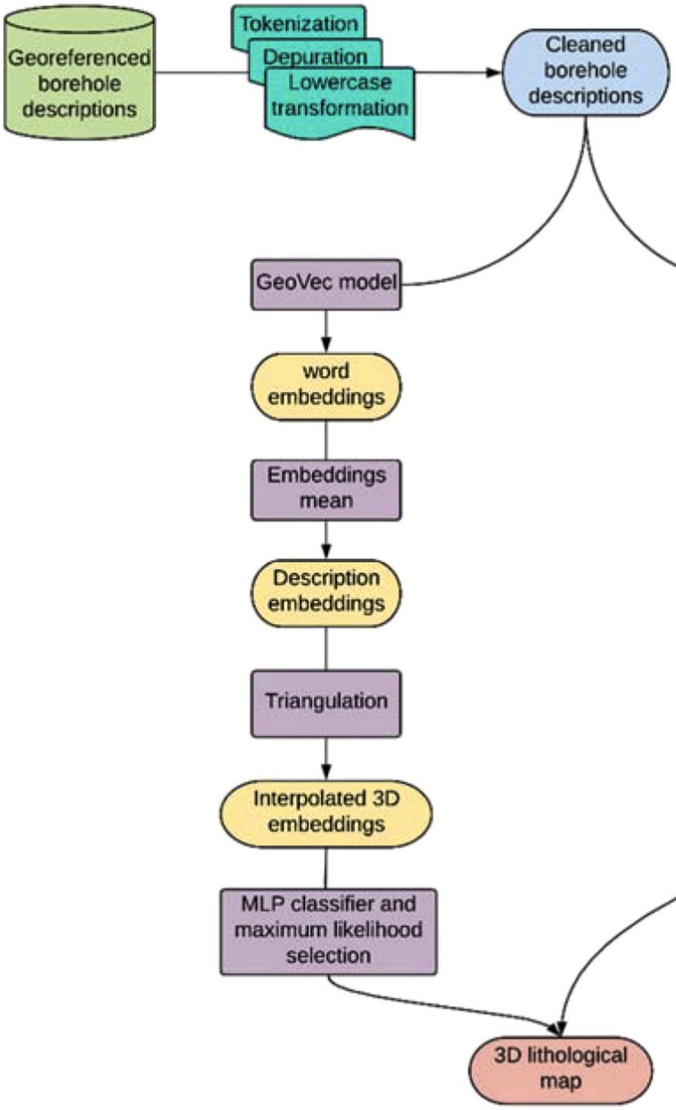

**Fig. 5.** Flowchart of the 3D lithologic map generation.

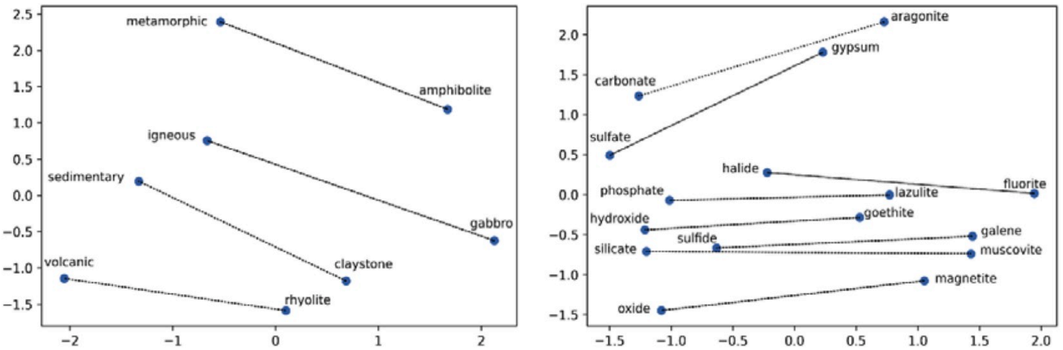

**Fig. 6.** Example of semantic relationships between types of rocks and lithologies (left) and between mineral groups and minerals (right) in a two dimensional PCA projection obtained by using a domain-specific GloVe model for geosciences (Padarian and Fuentes, 2019).

points. This consisted in a two-step procedure. In the manual first step, lithologies were assigned based on the descriptions, while the second step corresponded to an aggregation of specific lithologies into a series of major lithological groups (Table 1) based on our interpretation of the detailed lithologies. For instance, intrusive igneous rocks are aggregated into a single class, whilst sedimentary rocks are divided in several major groups such as sandstones, shales, and limestones. The same applies for sediments, where two groups are identified based on the granulometry of sediments.

While the presented grouping (Table 1) is qualitative, it captures the major variation in the original dataset. In the future, a natural language processing algorithm could be envisioned that would do this

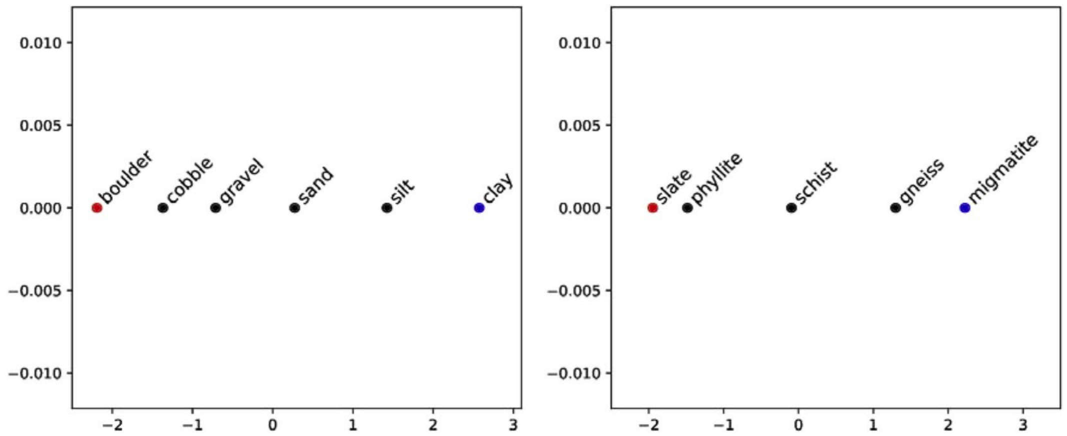

**Fig. 7.** Set of interpolation tokens obtained between two extreme terms showing scales of particle size (left) and metamorphic grade (right; Padarian and Fuentes, 2019). The PCA sign is arbitrary and hence, particle sizes increase from right to left, and metamorphic grades increase from left to right.

**Table 1**
Aggregation of lithologies into Major classes.

| Major Lithology | Lithologies included |
|---|---|
| Volcanic | basalt; volcanic; lava; tuff; breccia; rhyolite; agglomerate; ignimbrite; zeolite; andesite; latite; trachyte; scoria; dacite; pyroclastic |
| Intrusive | granite; diorite; porphyry; dolerite; igneous; feldspar; granodiorite; syenite; monzonite; pyroxenite; quartz |
| Metamorphic | slate; phyllite; schist; soapstone; gneiss; serpentine; mica; amphibolite; hornfels; pegmatite; metamorphic; marble; quartzite; biotite |
| Sandstone | sandstone; greywacke; arkose; wacke |
| Conglomerate | conglomerate |
| Shale | mudstone; claystone; siltstone; shale; argillite |
| Limestone | limestone; dolomite; calcrete; siderite; chalk; marl; calcite |
| Carbonaceous | carbonaceous; coal; lignite; wood; charcoal; bitumen |
| Chemical | silcrete; laterite; bauxite; ironstone; cement; chert; jasper; gypsum; apatite; pyrite; opal |
| Soil | soil; topsoil; subsoil; earth; |
| Fine sediments | clay; mud; pug; bentonite; kaolinite; silty clay; loam; sandy loam; silty loam; clay loam; sandy clay loam; drift; stones clay; clay gravel; mud gravel; clay boulders; silty; sandy silt; silty gravel; gritty clay; sandy clay; silty sandy clay; mud sand; clay sand; silty clay sand |
| Coarse sediments | sand; silty sand; gravel; stones gravel; stones sand; sand gravel; sand boulders; clay sand gravel; pebbles, boulders, stones; blue metal |
| Bedrock | bedrock |
| Alluvium | alluvium |
| Sedimentary | sedimentary |
| Water | water |
| Cavity | cavity |
| Peat | peat |

classification.

One of the difficulties of the semi-manual classification is the high number of descriptions and the ambiguity of some. These can lead to the misclassification of some of the descriptions, which must be considered when evaluating any supervised classification. For example, the description "Sand some clay trace gravel brown damp loose medium poorly sorted" was classified as "Coarse sediments". However, the description is ambiguous, which can be further explored by using description embeddings.

For the supervised classification, a stratified random sampling was implemented such that a 10% of the dataset was included into a test subset. The remaining 90% of the dataset was subsequently divided randomly into a training and a validation subsets, which accounted for the 90% and 10% of the remaining subset, respectively.

A Multi-layer perceptron (MLP) neural network was trained to classify the resulting embeddings. MLP networks are effective for classification tasks (Gardner and Dorling, 1998), taking advantage of all the combinations of features of the layer sequences. Here we present the results of the best model obtained by performing a grid search of the hyper-parameters, varying the number of fully-connected layers, and batch size (training samples simultaneously propagated through the network during training). The hyper-parameters were evaluated using the validation dataset. The number of epochs (times that the neural network passes through the entire training data) was set to 30 to avoid over-fitting based on the learning curves of the training and validation sets. The network consisted of three fully-connected layers with 100 neurons each (Fig. 8) and a 'ReLU' activation function (Nair and Hinton, 2010). The network was trained using the Adam optimizer (Kingma and Ba, 2014), with a batch size of 100. The output layer of the network yields the probability of the embeddings belonging to the different lithological classes found in the dataset, which, unlike the semi-manual classification, allows evaluation of the ambiguity of the descriptions.

Three metrics were used to evaluate the performance of the MLP classification, taking into account that the classes in the dataset are imbalanced. For the classification assessment, lithological classes with the highest probabilities from the classifier were used. The performance metrics included the accuracy, f1 (weighted mean of precision and recall), and balanced accuracy scores. All methods range from 0 to 1, with 1 being the exact match. In addition, a confusion matrix was constructed to evaluate the classification for the different lithological classes (Congalton, 1991).

### 2.5. *Classification uncertainty evaluation*

One of the advantages of using embeddings in conjunction with the MLP classifier is that, in contrast to a manual classification, it allows quantification of the uncertainty of the classification. In this case, the normalised Shannon entropy was used as a measure of uncertainty using Eq. (1) (Saco et al., 2010):

$$H[P] = \frac{\left[-\sum_{i=1}^{n} p_i \ln p_i\right]}{S_{max}} \tag{1}$$

where $n$ corresponds to the number of classes, $p_i$ is the probability of each class, and $S_{\max}$ is equal to ln $n$. In this case, 0 represents a value of null uncertainty in the classification, and 1 a very high uncertainty.

Additionally, an assessment of the ambiguity of the classification was also carried out based on an estimated Confusion Index (CI; Burrough et al., 1997) by selecting the two predicted classes with highest probabilities at each point as shown in Eq. (2):

$$CI = \left[1 - \left(\mu_{max_i} - \mu_{(max-1)_i}\right)\right] \tag{2}$$

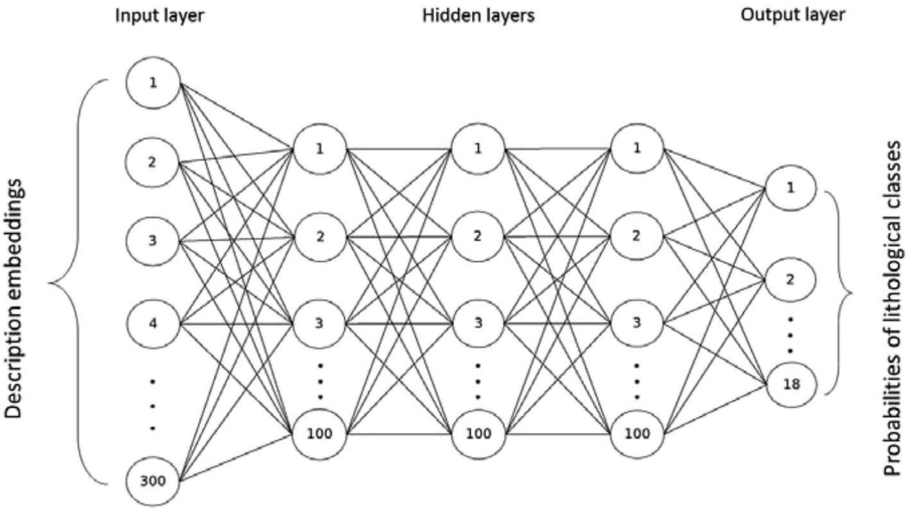

**Fig. 8.** Architecture of the MLP neural network used for the lithological classification.

where $\mu_{\max\ i}$ corresponds to the probability of the maximum likelihood class at site *i* while $\mu_{(max\ -\ 1)i}$ is the value of the second largest likelihood membership at site *i*. CI values range from 0 to 1, being 0 a value of null ambiguity in the classification, and 1 an index that implies high confusion, and therefore, high ambiguity.

### 2.6. *Interpolation and modelling of classes/embeddings*

Using the borehole dataset, a bore density map was developed prior to the interpolation and mapping of the lithological classes (Fig. 9). Based on the sectors with high bore density, two areas of interest (AOI) located near the towns of Moree and Coleambally were selected to develop lithological 3D models due to the higher borehole density, which may facilitate the evaluation of results.

Instead of using a regular grid sampling design, the lithologies (classes and vectors) were extracted based on the Gallerini and Donatis (2009) methodology, which considers a maximum depth interval to extract lithological sample points, depending on the thickness of the strata or layers. This approach was used because it more accurately represents stratigraphic sequences, without missing geological information in thin strata.

Since the generation of 3D models requires continuous spatial data, the embeddings need to be interpolated. Therefore, a Delaunay triangulation followed by a barycentric linear interpolation of the

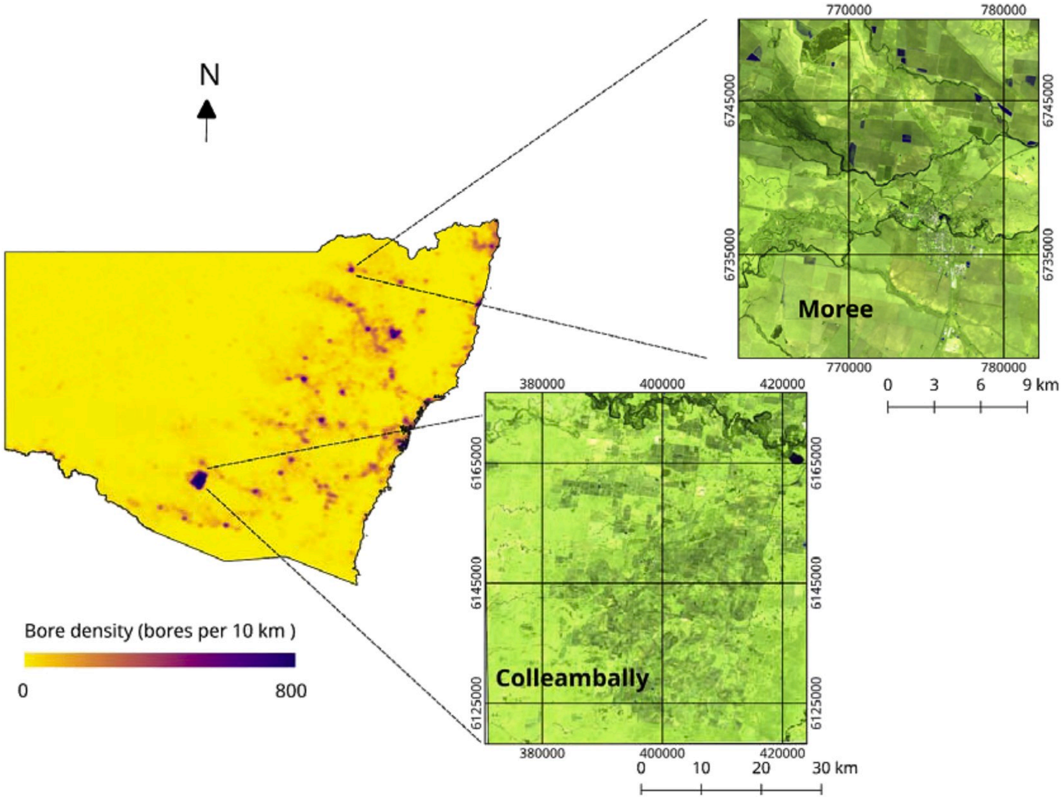

**Fig. 9.** Bore density map and selected of areas of interest (AOIs). The zoomed satellite images correspond to false colour (bands 3-4-1) Landsat 5 images. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

embeddings (referred to as triangulation in the following sections) was carried out. The interpolated vectors were subsequently classified using the trained MLP classifier. In order to compare the interpolation of the embeddings, a nearest neighbour (NN) interpolator, which is commonly used for interpolation of categorical data (Baboo and Devi, 2010; Babak, 2013; Li and Heap, 2014), was applied to the lithological classes derived using the semi-manual classification.

The interpolations were performed sequentially in depth through a 2.5D approach by layering the lithologies in the dataset using a depth interval of one m (Falivene et al., 2007). Training and validation of the interpolations were done by splitting the bores in the dataset into test, training and validation bores in the same proportion as described in the classification step. The interpolations were trained using the training dataset, and were evaluated using the f1, accuracy and balanced accuracy scores. The NN interpolation gives directly interpolated lithological classes, while the triangulation retrieves interpolated embeddings, which were passed to the MLP classifier to obtain the final lithological classes using NLP. A mapping of the interpolated lithologies was finally carried out using voxels in a 3D space. Even though the interpolation of embeddings was carried out in the two selected AOI, the methodology may be extrapolated to different areas. Likewise, this methodology for three-dimensional mapping of lithologies may be applied to any georeferenced lithological description dataset written in English.

Uncertainty maps were obtained through bootstrapping, using 100 iterations. This number of iterations was chosen because it allows a good representation of the distribution of lithologies/embeddings interpolated without making the process too expensive in terms of time and computational requirements. In the NN interpolation, a lithological class probability was obtained for each class in each voxel by counting the number of occurrences and dividing it by the 100 iterations. In the triangulation, the probabilities obtained from running the classifier were averaged in each voxel. The uncertainties for both methodologies were then estimated through the CI and the normalised Shannon entropy as described in section 2.5.

## 3. Results

### 3.1. Classification performance

The confusion matrix from the MLP classifier (Table 2) indicates the imbalance between the different lithological classes in the borehole descriptions, where alluvial sediments (coarse and fine grained) dominate the NSW landscape. These are followed by sedimentary lithofacies, with a predominance of sandstones, followed by shales.

Lithological classes that had lower accuracies were those with a fewer number of descriptions. The class water had the worst classification, and it was mostly misclassified as volcanic, shale and sandstone. The misclassification indicates the main water bearing lithologies, or in some other cases indicates overlap between classes. For instance, bedrock is a broad concept that can refer to different lithologies, but had no further lithological detail in the descriptions. Other example is the soil and fine sediments classes. Most soils are composed of a mix of different grain sizes mineral particles and organic matter, but can also be described as the textural class of those materials, which are classified as fine grained sediments for the classifier instead of soils. This leads to a confusion which, if not reflected in the classification step, will affect the interpolation results.

In addition, as mentioned in the methodology, the semi-manual classification also created misclassification due to the ambiguity of descriptions. In some cases, semi-manually misclassified descriptions were correctly classified by the MLP classifier, but not reflected in the metrics used, since the semi-manual classification is used as reference. For instance, in the aforementioned example the MLP classified the description “Sand some clay trace gravel brown damp loose medium poorly sorted” as “Fine sediments”, which includes the “sand clay” lithology.

Overall, the MLP classification with the averaged word embeddings achieved good results based on the given inputs, with 0.958 accuracy and a balanced accuracy index of 0.864. Based on this overall performance (shown in Table 3) the MLP classifier was used for subsequent statistical and spatial analysis.

Since the semi-manual classification is a time-consuming task, the training performance as a function of the size of the training dataset was further explored (Fig. 10). As expected, more training samples lead to higher accuracy of the validation and a slight decrease in the training accuracy, which results in minimum error between the curves at around 500,000 samples. Regardless of the increase in accuracy, even with a

**Table 3**
Classification performance of MLP neural network for the sentence embeddings obtained by averaging the word embeddings of the descriptions.

| Embeddings | Accuracy | F1 | Balanced accuracy |
|---|---|---|---|
| Training | 0.965 | 0.937 | 0.939 |
| Validation | 0.958 | 0.868 | 0.864 |

**Table 2**
Confusion matrix generated comparing the semi manual and the automated classification using description embeddings. Columns correspond to the predicted classes using description embeddings and rows to the lithological classes semi-manually classified (assumed as the actual lithological classes).

| | 1* | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 829 | 3 | 0 | 0 | 0 | 11 | 1 | 30 | 4 | 0 | 1 | 0 | 6 | 0 | 0 | 0 | 0 | 2 | 887 |
| **2** | 0 | 527 | 0 | 0 | 0 | 1 | 2 | 19 | 1 | 2 | 12 | 0 | 6 | 0 | 6 | 1 | 0 | 0 | 577 |
| **3** | 1 | 1 | 3,546 | 0 | 1 | 81 | 2 | 235 | 4 | 1 | 1 | 0 | 17 | 0 | 32 | 1 | 2 | 0 | 3,925 |
| **4** | 0 | 0 | 1 | 81 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 88 |
| **5** | 0 | 0 | 0 | 1 | 2,106 | 54 | 17 | 112 | 2 | 7 | 9 | 0 | 70 | 0 | 22 | 4 | 8 | 0 | 2,412 |
| **6** | 0 | 2 | 92 | 0 | 12 | 135,606 | 73 | 6,898 | 78 | 22 | 56 | 1 | 693 | 0 | 265 | 121 | 274 | 41 | 144,234 |
| **7** | 0 | 0 | 0 | 0 | 0 | 124 | 2,702 | 143 | 3 | 2 | 1 | 0 | 71 | 0 | 27 | 6 | 13 | 0 | 3,092 |
| **8** | 3 | 3 | 92 | 1 | 13 | 6,396 | 76 | 304,989 | 105 | 54 | 93 | 4 | 780 | 3 | 453 | 604 | 305 | 2 | 313,976 |
| **9** | 0 | 0 | 2 | 0 | 16 | 49 | 3 | 179 | 18,659 | 1 | 47 | 0 | 42 | 2 | 51 | 2 | 65 | 1 | 19,119 |
| **10** | 0 | 1 | 0 | 3 | 1 | 16 | 1 | 150 | 3 | 2,687 | 6 | 0 | 10 | 0 | 22 | 1 | 10 | 1 | 2,912 |
| **11** | 0 | 0 | 0 | 0 | 0 | 21 | 0 | 83 | 20 | 0 | 7,760 | 0 | 41 | 2 | 148 | 28 | 172 | 2 | 8,277 |
| **12** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 26 | 0 | 0 | 0 | 0 | 0 | 0 | 28 |
| **13** | 0 | 6 | 47 | 2 | 14 | 146 | 15 | 417 | 73 | 4 | 20 | 0 | 71,820 | 3 | 603 | 1 | 49 | 30 | 73,250 |
| **14** | 0 | 0 | 3 | 0 | 0 | 3 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 123 | 0 | 0 | 0 | 0 | 135 |
| **15** | 0 | 3 | 105 | 0 | 1 | 96 | 4 | 1,510 | 61 | 11 | 38 | 0 | 1,122 | 1 | 79,804 | 19 | 110 | 57 | 82,942 |
| **16** | 0 | 0 | 2 | 0 | 2 | 63 | 1 | 255 | 13 | 0 | 3 | 0 | 10 | 0 | 12 | 43,135 | 14 | 0 | 43,510 |
| **17** | 1 | 3 | 2 | 0 | 10 | 145 | 15 | 424 | 33 | 1 | 20 | 0 | 52 | 4 | 116 | 9 | 29,870 | 71 | 30,776 |
| **18** | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 7 | 6 | 2 | 1 | 0 | 25 | 0 | 1 | 0 | 2 | 393 | 447 |
| **Total** | 834 | 549 | 3,893 | 88 | 2,176 | 142,822 | 2,912 | 315,460 | 19,065 | 2,795 | 8,068 | 31 | 74,767 | 138 | 81,562 | 43,932 | 30,895 | 600 | 730,587 |

*1: alluvium; 2: bedrock; 3: carbonaceous; 4: cavity; 5: chemical; 6: coarse sediments; 7: conglomerate; 8: fine sediments; 9: intrusive; 10: limestone; 11: metamorphic; 12: peat; 13: sandstone; 14: sedimentary; 15: shale; 16: soil; 17: volcanic; 18: water.

small fraction of the dataset used for training, the classification accuracy is greater than 0.9. This means that with any data set with more than 50,000 training samples, the classifier leads to good results, independent of any further increase in training sample size. This means a threshold in the resources used for the semi-manual classification can be based on this training size.

### 3.2. Classification uncertainty

Analysing the mean normalised entropy per lithological class and the lithological class percentages in the different AOIs, it is clear that the highest entropies are associated with lithologies with a small occurrence, and except for the water and sedimentary classes, the entropies are mainly low (0–0.1; Fig. 11). Again, sediments dominate because NSW is mainly composed of sedimentary basins and the boreholes are mainly drilled in near-surface alluvial deposits to access water supply and rarely reach basement rocks. Obviously, the distribution of lithologies has a higher spread for the entire NSW dataset, compared with the two AOIs.

### 3.3. Interpolation and 3D modelling

An overall assessment of the two interpolation methods used in both AOIs is in Table 4. Both interpolation methods have results comparable to other three-dimensional geological/soil studies applied at different scales (Falivene et al., 2007; Hengl et al., 2014). However, the Coleambally AOI, on average, has better performance than Moree.

Fig. 12 shows the performance of the interpolation in depth (left) in relation to the number of samples (right). In both areas of study the number of training and validation points increases in the first meters depth, but diminishes rapidly at deeper depths. In Coleambally, the maximum number of sampled points reaches around 10,000, whilst in Moree there are always less than 1000. The decrease of samples with depth is more gradual in Moree. However, after 55 m depth a sudden reduction in samples is observed (from 746 boreholes to only 122), mainly due to reaching the bottom of the alluvial productive aquifer (several drilling logs describe shale and sandstone sedimentary rocks between 55 and 65 m depth), which leads to a sudden drop in the interpolation performance.

Accuracies fluctuate with depth due to the 2.5D interpolation scheme applied to a one m interval. Even though a very general trend may be detected suggesting that the interpolation of shallow lithologies leads to a better performance, this does not relate clearly with the number of training samples. Thus, while at shallow depths there is a slight performance increase, this is followed by a steady decrease with depth in Coleambally, which relates to the peak in the number of training samples (Fig. 12). A completely different picture emerges in Moree, where surface lithologies have high accuracies, which diminish rapidly with depth. A Pearson correlation test was used to assess the relationship, indicating that both AOIs had a positive correlation between accuracies and the number of training and validation samples ($0.38 < r < 0.51$, $p$-values $< 0.05$). Therefore, the number of samples affects the performance of the interpolator, which must be considered prior to building of geologic models. However, as the relationships were not strong, other factors affecting the performance must be taken into account, such as the geologic complexity.

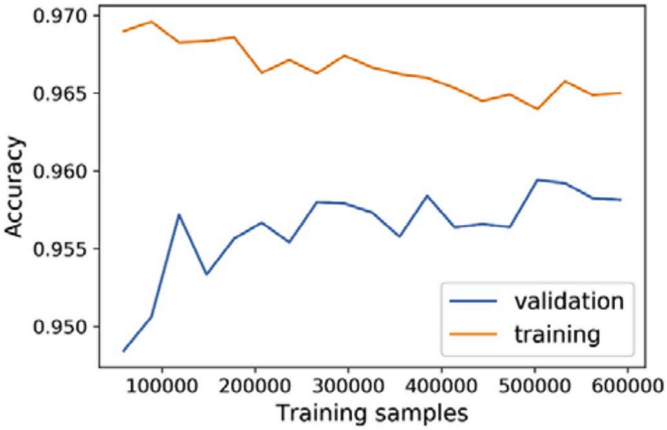

**Fig. 10.** Training size and its effect on the classification accuracies for the training and validation datasets and on the classifier training time.

Overall, there was little difference between the two interpolation methods for each AOI (Fig. 12).

The voxel maps of the lithologies in the Moree study area based on the two different interpolation methods are in Fig. 13. While the flexibility of the nearest neighbour interpolation means the interpolation can go beyond the training points in the map, the results in Fig. 13 were masked based on a convex hull border, to make comparison of the results of both interpolation methods easier. This might reduce the uncertainty of the results outside the training samples in the interpolation area.

The lithologies at Moree are mostly sequences of sediments that alternate between fine and coarse grain sizes, except for the northeast of the AOI. Here sedimentary sequences can be observed in depth that include conglomerates and sandstones. In the area, the Mehi and the Gwydir Rivers dominate the developed alluvial landscape and lead to Cenozoic alluvial deposits, which form the Lower Gwydir alluvium, mainly comprised by clay, silt, sand and gravel (NSW Department of Industry, 2018). This presents two alluvial formations, known as the Narrabri formation (shallower and composed of medium-coarse grained sediments) and the Gunnedah formation (deeper and composed by coarse grained sediments), which contain the two main alluvial aquifers (Welsh et al., 2014). These are overlain by soils and fine grained sediments, yet coarse sediments are found at shallow depths in the middle of the AOI, where the rivers are located (Fig. 9). Lamontagne et al. (2011), studying the interconnection of the surface and groundwater systems in the region, generated some geologic cross sections that show the heterogeneity between fine and coarse sediment layers characterising the alluvium in the area. This hydrogeologic setting can be distinguished in Fig. 13, where soil and fine sediments can be found overlying deeper sediments/strata. Additionally, both alluvial formations and the heterogeneity in the distribution of sediments can be easily recognised. On the north eastern and southern extremes of the AOI, surface colluvial sediments are found (Geoscience Australia, 2012). The depth of the alluvium tends to increase to the west.

In the Coleambally AOI sediments are deeper and dominated by fine-medium grain sizes in the lithology, with some patches of coarse grained sediments at depth (Fig. 14). According to Prathapar et al. (1997), the alluvial deposits in the area range in thickness between 100 and 200 m, which can be observed in the generated maps. In the northern area a strip of surface coarse sediments can be found, which follows the main course of the Murrumbidgee River (Fig. 9). This agrees with a strip of alluvial sediments surrounding the main channel within dominant Cenozoic alluvial clays mapped in the Groundwater Geofabric product (Bureau of Meteorology, 2012). This surface strip of coarse sediments might correspond to point bar deposits associated with the meandering river system (Nanson and Page, 1983). With depth, layers and lenses of coarse sediments indicate where palaeochannels occur (Page et al., 1996). These paleochannels act as shallow aquifers in the region (Prathapar et al., 1997). The dominant surface lithologies are part of the Shepparton Formation which forms extensive alluvial floodplains. This Formation is described as unconsolidated to poorly consolidated fine sediments, including clay or silty clay grain sizes, having lenses of polymictic sand and gravel (Geoscience Australia, 2012). The deepest extensive coarse sediment layers that can be observed in Fig. 14 represent the Calivil Formation, which is the most transmissive strata in the region, and therefore it has been highly exploited for irrigation purposes (Page, 1994).

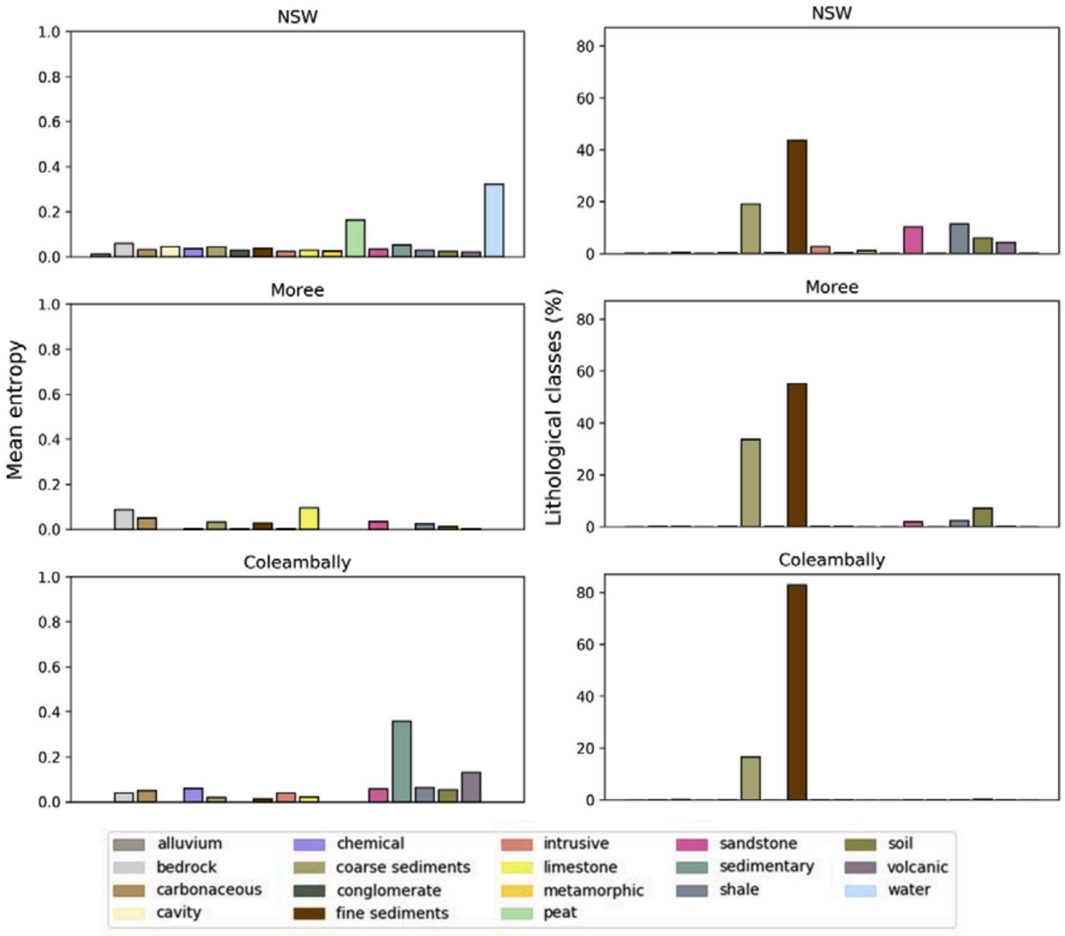

**Fig. 11.** Mean normalised Shannon entropy per lithological class (left panel) and the percentage of each lithological class (right panel) in the different study areas.

**Table 4**
Mean performance metrics at Moree up to 60 m depth, and in Coleambally up to 80 m depth.

| AOI | Interpolation | Accuracy | F1 | Balanced accuracy |
|---|---|---|---|---|
| Moree | NN training | 0.984 | 0.970 | 0.970 |
| | NN validation | 0.600 | 0.415 | 0.487 |
| | Triangulation training | 0.964 | 0.922 | 0.954 |
| | Triangulation validation | 0.600 | 0.446 | 0.494 |
| Coleambally | NN training | 0.988 | 0.984 | 0.987 |
| | NN validation | 0.720 | 0.569 | 0.579 |
| | Triangulation training | 0.961 | 0.659 | 0.954 |
| | Triangulation validation | 0.705 | 0.564 | 0.571 |

### 3.4. Uncertainty of 3D models

Clearly, the uncertainty for the Moree AOI (Fig. 15) increases for small lithological patches surrounded by dominant lithological classes, indicating that the interpolation alternatives do not perfectly match the actual terrain distribution of rocks and sediments. In the case of uncertainties obtained through triangulation, the boundary between lithologies tends to indicate high uncertainty values. The overall uncertainties using both interpolation alternatives are relatively low. The mean CI value for the NN interpolation was 0.095 while it was 0.275 using the triangulation. Mean entropies were 0.052 and 0.093 using the NN interpolation and the triangulation, respectively. Additionally, a two-sample Kolmogorov-Smirnov test indicated that the NN interpolation (Fig. 15 above) has significantly lower uncertainties (for both CI and entropies) than the triangulation ($p$-value $< 0.05$; Fig. 15 below).

The uncertainty of the Coleambally lithological maps is in Fig. 16. Again, uncertainty values are mostly low. However, some small patches of higher uncertainty can be observed. Again the NN interpolation (Fig. 16 above) has significantly lower uncertainties ($p$-value $< 0.05$; mean CI of 0.069 and mean entropy of 0.037) than the triangulation (mean CI of 0.248 and mean entropy of 0.093; Fig. 16 below), yet in both cases the uncertainties are lower than in the Moree AOI.

The distribution of the mean confusion index (CI) also fluctuates with depth (Fig. S1 and Fig. S2, supplement). A moderate negative correlation ($-0.44 \leq r \leq -0.55$; $p$-values $< 0.05$) was found between CIs and triangulation accuracies in depth for both AOIs, which means that the ambiguity of embeddings also affects the interpolation performance.

## 4. Discussion

There are limited studies applying text mining techniques to geosciences (Padarian and Fuentes, 2019). For example, Pollock et al. (2012) used regular expressions and geospatial data to map lithological features by using the interpolation of match scores (from the regular expressions). However, this had several limitations, such as a small area of application, the use of only three different lithologies, and the lack of automation in the methodology.

In this study, we move a step further in the use of text mining to address geosciences problems. By combining NLP (with the simplest conversion from word embeddings to sentence embeddings), a supervised classification method (MLP neural network) and the simplest of

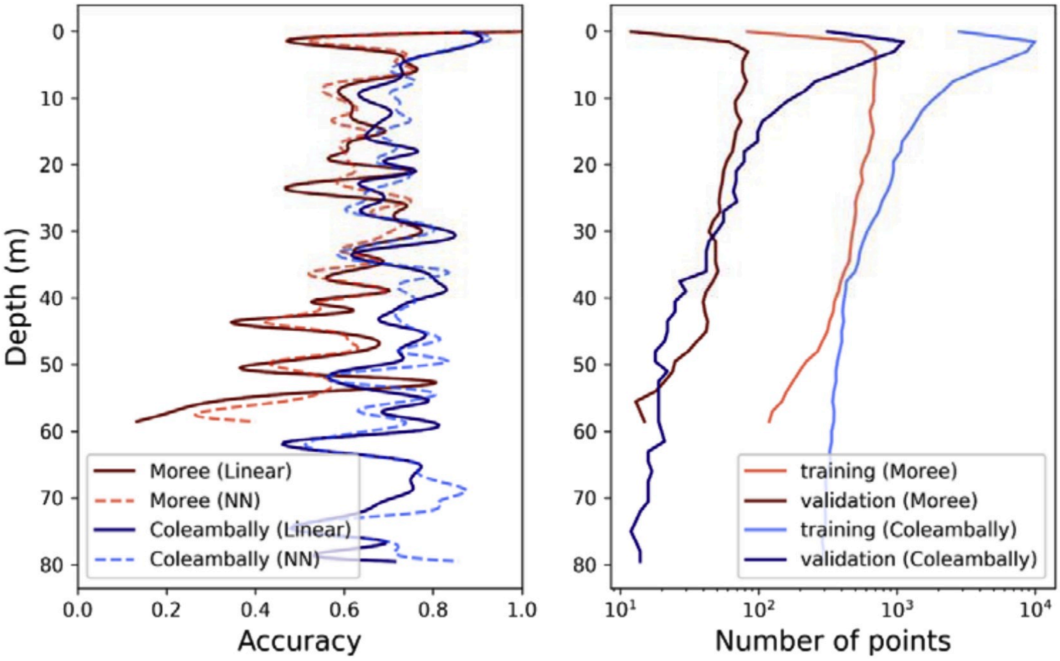

**Fig. 12.** Interpolation accuracies in depth (left) and number of points used in the interpolation (right).

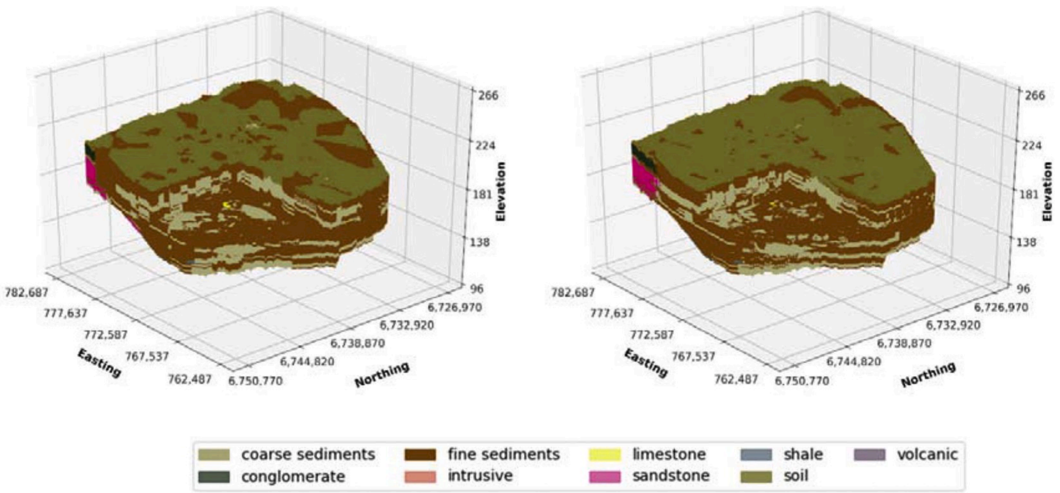

**Fig. 13.** Moree 3D lithological maps obtained using a NN interpolation (left) and a triangulation of embeddings (right).

interpolation techniques (a triangulation of embeddings) we were able to build 3D lithological maps from borehole descriptions, and assess prediction uncertainties. Even though NLP might indicate areas with geologic structures due to the distribution and location of associated lithologies, these should be complemented with the detection of surface geological lineaments and other machine learning algorithms that could capture lithological patterns associated with these. However, this is an area for future research.

The use of word embeddings led to very accurate results in the classification step compared with a semi-manual classification. However, since the description embeddings are built up based on a specialised lexicon, other alternatives rather than a simple averaging of words embeddings, such as sequential denoising autoencoders or neural networks with complex architectures (LSTM) might lead to an improvement of the classification (Wieting et al., 2015).

Overall, the results obtained were comparable with those derived using a semi-manual expert criteria classification, and even slightly better for the interpolation results in Moree, and is therefore suggested to be useful for such applications. Another advantage of the automated method is in the performance of the classification step. Even a small dataset (≈50,000 samples; Fig. 10) leads to relatively good performance.

Even though the current implementation is not fully automated, it can be considered as a first step in the automation of these tasks, and a first step demonstrating the applicability of NLP to the geosciences. For a full automation, unsupervised classification alternatives will need to be coupled to NLP. This is currently part of our ongoing research.

Most geological studies that address three dimensional modelling using borehole and other data sources do not provide an evaluation of accuracies, but generally assume a good performance (Kaufmann and Martin, 2008; Kessler et al., 2009; Høyer et al., 2015). For instance, Gallerini and Donatis (2009) did not carry out an evaluation of their model performance arguing that fluvial environments present heavy facies variations, and therefore, used the entire dataset in the model. Even though this argument seems reasonable, this precludes evaluation of the performance of the final model. However, in some studies that do present performance metrics (Falivene et al., 2007; He et al., 2010;

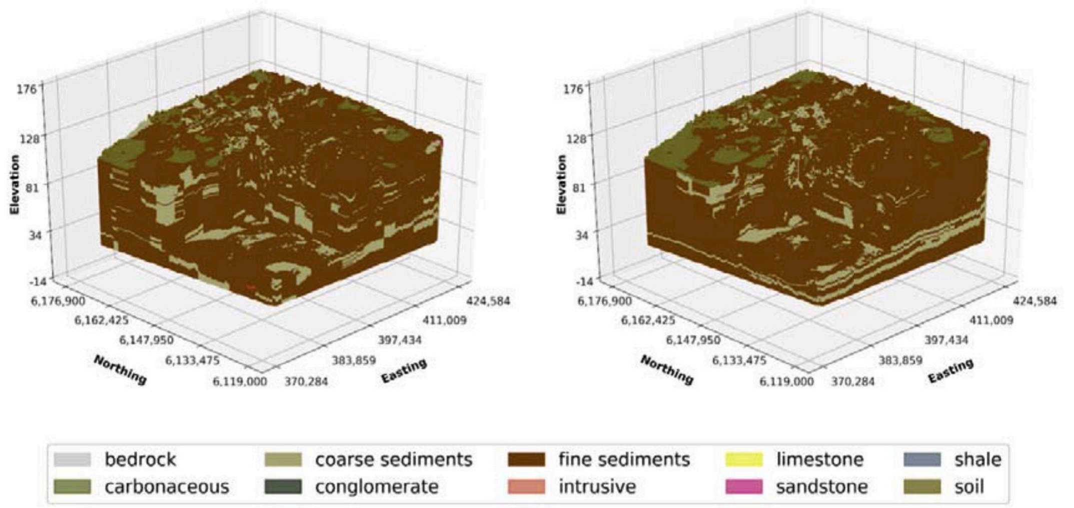

**Fig. 14.** Coleambally 3D lithological maps obtained using a NN interpolation (left) and a triangulation of embeddings (right).

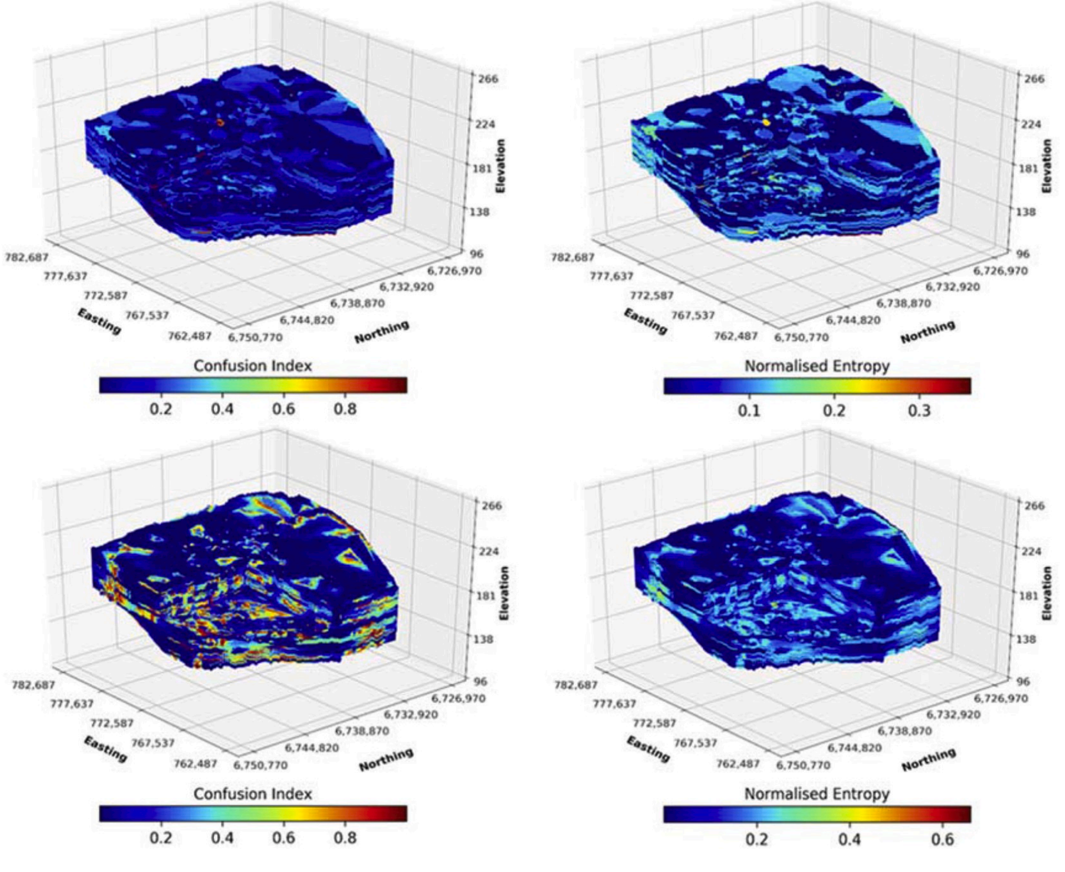

**Fig. 15.** Uncertainty mapping for the Moree AOI using the NN interpolation (above) and the triangulation of embeddings (below). Entropy scales were normalised to the range of values obtained.

Hengl et al., 2014), 3D models usually have limitations due to the moderate performance of interpolation techniques in the 3D space. As pointed out by Kumar et al. (2000), the interpolation between boreholes is a hard task for humans and leads to only moderate reliability.

This study indicates acceptable quantitative results for the presented 3D lithological models at the regional scale, taking into account the results obtained by Falivene et al. (2007) at the local scale. Additionally, these coincide qualitatively with the geological expert knowledge gained from the studied regions (Prathapar et al., 1997; Geoscience Australia, 2012; Welsh et al., 2014). These results may be useful for groundwater modelling at the regional scale, considering that the methodology used provides uncertainty estimations which may be used and propagated in the models, and may also guide future geological/geophysical explorations.

The 2.5D modelling scheme used precludes inclusion of vertical information in the interpolation though. An interpolation in a 3D space does not necessarily lead to better results (Wu et al., 2005), which may be related to the variance of the data (Sahlin et al., 2014), and the type of

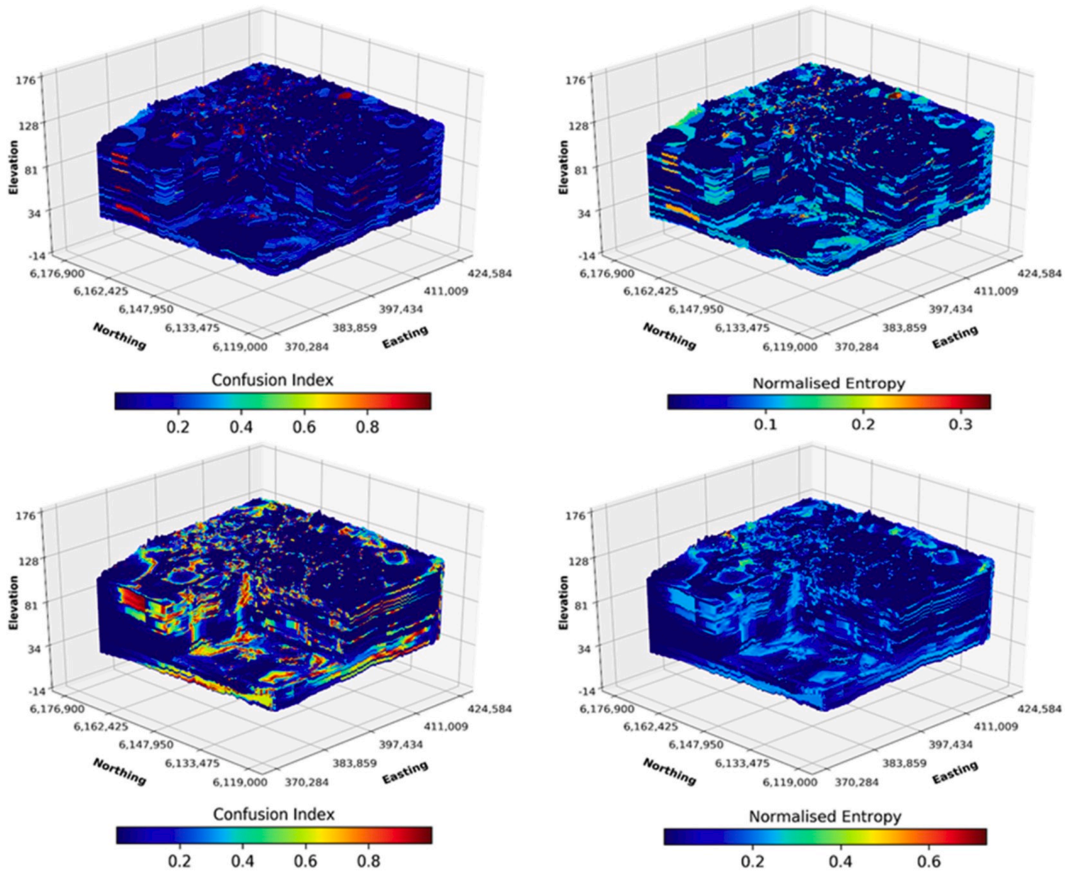

**Fig. 16.** Uncertainty mapping for the Coleambally AOI using a NN interpolation (above) and a triangulation of embeddings (below).

lithological classes. Abrupt transitions between different sediments/facies can be accurately captured by the use of 2.5D schemes, while smooth transitions might be better represented using 3D schemes. In this case, the 2.5D scheme led to slight fluctuations in the interpolation performance in depth, which depended moderately on the number of samples and the ambiguity of embeddings, and possibly on the geological complexity. Since traditional interpolation methods show a moderate performance when working on 3D arrays (Zhou et al., 2005; Sahlin et al., 2014; Hengl et al., 2014), this opens the door for the use of machine learning techniques to address these tasks.

Uncertainty is usually discussed in different 3D geologic modelling studies; however, most of them avoid its quantification (Wu et al., 2005; Gallerini and Donatis, 2009; Zhu et al., 2012). Even though both methodologies produce 3D models that are not a perfect representation of reality (Table 4), only a proper uncertainty analysis is possible to identify where the performance of models is unsatisfactory (Lindsay et al., 2012). The 3D lithological models derived from word embeddings show a higher uncertainty than the models obtained from the semi-manual classification and the NN interpolation. This is because the uncertainty in the semi-manual classification cannot be estimated. Therefore, even though the word embeddings lead to higher uncertainties in the lithological models, these estimates must be interpreted as more realistic, since they correspond to the propagation of uncertainties in the classification and interpolation stages (Jones et al., 2004). Thus, the proposed thorough evaluation and mapping of uncertainty can be used to guide future explorations to collect additional data for more accurate results, which is a clear advantage of using word embeddings.

While several geologic software packages have been developed to build 3D models (RockWare®, Leapfrog®, Georeka, GSI3D), most of them have a high economic cost. In this case, using open source modules implemented in Python allowed the development of 3D lithological maps using voxels, capturing the lithological setting of large areas.

Different applications can be found for the resulting 3D lithological models. They can be included as input in more complex geological models, and these can also be used in hydrogeological modelling, ore prospecting, territorial ordering and environmental studies, indicating valuable data synthesis for the geosciences field.

## 5. Conclusions

Applying NLP is useful for geoscience applications. By integrating NLP, machine learning algorithms and spatial interpolation techniques, lithological 3D maps could be obtained. NLP and machine learning can automate 3D geological mapping from text input bore descriptions, which might be further explored using unsupervised classification algorithms. This allows the otherwise qualitative and manually interpreted data to be applied quantitatively, opening a new information source for geoscience applications.

Firstly, the classification of lithological description embeddings through MLP neural networks is very accurate and just a small fraction of samples used for training the classifier results in high accuracy. Secondly, NLP allows for an uncertainty quantification in the different stages of the 3D model generation. Simple interpolation techniques using a 2.5D approach give an acceptable performance, demonstrating that the triangulation of embeddings and their subsequent classification gives equivalent results, and in some cases even slightly better, than by using a NN interpolation of lithological classes obtained manually.

The 3D lithological models generated using voxels through Python public libraries correspond to reasonable representations of complex lithological settings in relatively large areas, which can be further improved using uncertainty maps.

## Code and data availability

Code associated to the GeoVec model used will be available at https://github.com/spadarian/GeoVec

The original dataset used corresponds to the Australian Groundwater Explorer of the Bureau of Meteorology, which is publicly available at http://www.bom.gov.au/water/groundwater/

The code that applies the GeoVec model to the dataset will be available at https://github.com/IFuentesSR/GeoVectoLitho

## Author contribution

I. Fuentes and R.W. Vervoort contributed with the original idea, which was further developed by I. Fuentes and J. Padarian. I. Fuentes and J. Padarian developed the model and the code to evaluate it. I. Fuentes prepared the manuscript with contributions from all co-authors. Lastly, R.W. Vervoort supervised the entire process.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The first author was supported through a scholarship by the Chilean Government ("Becas Chile", Agencia Nacional de Investigación y Desarrollo, ANID).

This research was supported by the Sydney Informatics Hub, funded by the University of Sydney.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2020.104516.

## References

Aggarwal, C., Zhai, C., 2012. Mining Text Data. Springer Science + Business Media, p. 521pp.

Babak, O., 2013. Inverse distance interpolation for facies modeling. Stoch. Environ. Res. Risk Assess. 28 (6), 1373–1382. https://doi.org/10.1007/s00477-013-0833-8.

Baboo, S., Devi, M.R., 2010. An analysis of different resampling methods in Coimbatore, District. Global J. Comput. Sci. Technol. 10, 61–66.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. O'Reilly, p. 479pp.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Ling. 5, 135–146. https://doi.org/10.1162/tacl_a_00051.

Boulton, G., 2018. The challenges of a big data Earth. Big Earth Data 2 (1), 1–7. https://doi.org/10.1080/20964471.2017.1397411.

Bureau of Meteorology, 2012. Australian Hydrological Geospatial Fabric (Geofabric) Data Product Specification: Groundwater Cartography. Bureau of Meteorology, Canberra, Australia, Version 2.1.

Burrough, P., Gaans, P.V., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77 (2–4), 115–135. https://doi.org/10.1016/s0016-7061(97)00018-9.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37 (1), 35–46. https://doi.org/10.1016/0034-4257(91)90048-b.

Culshaw, M., 2005. From concept towards reality: developing the attributed 3D geological model of the shallow subsurface. Q. J. Eng. Geol. Hydrogeol. 38 (3), 231–284. https://doi.org/10.1144/1470-9236/04-072.

Escudero, G., 2006. Machine Learning Techniques for Word Sense Disambiguation. Ph.D. thesis, Universitat Politècnica de Catalunya, España, p. 162.

Falivene, O., Cabrera, L., Sáez, A., 2007. Optimum and robust 3D facies interpolation strategies in a heterogeneous coal zone (Tertiary as Pontes basin, NW Spain). Int. J. Coal Geol. 71 (2–3), 185–208. https://doi.org/10.1016/j.coal.2006.08.008.

Gallerini, G., Donatis, M.D., 2009. 3D modeling using geognostic data: the case of the low valley of Foglia river (Italy). Comput. Geosci. 35 (1), 146–164. https://doi.org/10.1016/j.cageo.2007.09.012.

Gardner, M., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. 32 (14–15), 2627–2636. https://doi.org/10.1016/s1352-2310(97)00447-0.

Geoscience Australia, 2012. Surface Geology of Australia, 1:1,000,000 Scale, 2012 edition. Bioregional Assessment Source Dataset.

He, Y., Hu, K., Chen, D., Suter, H., Li, Y., Li, B., Yuan, X., Huang, Y., 2010. Three dimensional spatial distribution modeling of soil texture under agricultural systems using a sequence indicator simulation algorithm. Comput. Electron. Agric. 71 https://doi.org/10.1016/j.compag.2009.06.012.

Hengl, T., Jesus, J.M.D., Macmillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — global soil information based on automated mapping. PloS One 9 (8). https://doi.org/10.1371/journal.pone.0105992.

Hilbert, M., Lopez, P., 2011. The world's technological capacity to store, communicate, and compute information. Science 332 (6025), 60–65. https://doi.org/10.1126/science.1200970.

Høyer, A.-S., Jørgensen, F., Sandersen, P., Viezzoli, A., Møller, I., 2015. 3D geological modelling of a complex buried-valley network delineated from borehole and AEM data. J. Appl. Geophys. 122, 94–102. https://doi.org/10.1016/j.jappgeo.2015.09.004.

Jain, A., Kulkarni, G., Shah, V., 2018. Natural language processing. Int. J. Comput. Sci. Eng. 6, 161–167.

Jones, R.R., Mccaffrey, K.J.W., Wilson, R.W., Holdsworth, R.E., 2004. Digital field data acquisition: towards increased quantification of uncertainty during geological mapping. Geol. Soc., London, Special Publications 239 (1), 43–56. https://doi.org/10.1144/gsl.sp.2004.239.01.04.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2017. Machine Learning for the Geosciences: Challenges and Opportunities arXiv preprint arXiv:1711.04708.

Kaufmann, O., Martin, T., 2008. 3D geological modelling from boreholes, cross-sections and geological maps, application over former natural gas storages in coal mines. Comput. Geosci. 34 (3), 278–290. https://doi.org/10.1016/j.cageo.2007.09.005.

Kessler, H., Mathers, S., Sobisch, H.-G., 2009. The capture and dissemination of integrated 3D geospatial knowledge at the British Geological Survey using GSI3D software and methodology. Comput. Geosci. 35 (6), 1311–1321. https://doi.org/10.1016/j.cageo.2008.04.005.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization arXiv:1412.6980 [cs].

Kumar, J.K., Konno, M., Yasuda, N., 2000. Subsurface soil-geology interpolation using fuzzy neural network. J. Geotech. Geoenviron. Eng. 126 (7), 632–639. https://doi.org/10.1061/(asce)1090-0241 (2000)126:7(632).

Lamontagne, S., Taylor, A., Cook, P., Barrett, C., 2011. Interconnection of Surface and Groundwater Systems - River Losses from Losing/disconnected Streams. Gwydir River Site Report. CSIRO: Water for a Healthy Country National Research Flagship, Adelaide, Australia, p. 31. https://doi.org/10.4225/08/58518c1bcb3d0.

Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. Geosc. Front. 7 (1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003.

Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Model. Software 53, 173–189. https://doi.org/10.1016/j.envsoft.2013.12.008.

Lindsay, M.D., Aillères, L., Jessell, M.W., Kemp, E.A.D., Betts, P.G., 2012. Locating and quantifying geological uncertainty in three-dimensional models: analysis of the gippsland basin, southeastern Australia. Tectonophysics 546–547, 10–27. https://doi.org/10.1016/j.tecto.2012.04.007.

Maarala, A.I., Rautiainen, M., Salmi, M., Pirttikangas, S., Riekki, J., 2015. Low latency analytics for streaming traffic data with Apache Spark, 2015. IEEE Int. Conf. Big Data (Big Data). https://doi.org/10.1109/bigdata.2015.7364101.

McBratney, A.B., Minasny, B., Stockmann, U., 2018. Pedometrics. Springer, p. 720pp.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space arXiv:1301.3781 [cs].

Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W., 2011. Natural language processing: an introduction. J. Am. Med. Inf. Assoc. 18 (5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning. ICML-10), pp. 807–814.

Nanson, G., Page, K., 1983. Lateral accretion of fine-grained concave benches on meandering rivers. In: Collinson, J.D., Lewin, J. (Eds.), Modern and Ancient Fluvial Systems. Blackwell Scientific Publications, pp. 133–143.

Nativi, S., Mazzetti, P., Santoro, M., Papeschi, F., Craglia, M., Ochiai, O., 2015. Big data challenges in building the global Earth observation system of systems. Environ. Model. Software 68, 1–26. https://doi.org/10.1016/j.envsoft.2015.01.017.

NSW Department of Industry, 2018. Gwydir Alluvium Water Resource Plan – Groundwater Resource Description. NSW Government, p. 61pp.

O'Brien, J.J., Spry, P.G., Nettleton, D., Xu, R., Teale, G.S., 2015. Using Random Forests to distinguish gahnite compositions as an exploration guide to Broken Hill-type Pb–Zn–Ag deposits in the Broken Hill domain, Australia. J. Geochem. Explor. 149, 74–86. https://doi.org/10.1016/j.gexplo.2014.11.010.

O'Neill, C., Danis, C., 2013. The Geology of NSW. The Geological Characteristics and History of NSW with a Focus on Coal Seam Gas (CSG) Resources. A Report Commissioned for the NSW Chief Scientist's Office. Macquarie University, Sydney, Australia, p. 114pp.

Padarian, J., Fuentes, I., 2019. Word embeddings for application in geosciences: development, evaluation and examples of soil-related concepts. SOIL 5, 177–187. https://doi.org/10.5194/soil-5-177-2019.

Page, K., 1994. Late Quaternary Stratigraphy and Chronology of the Riverine Plain, Southeastern Australia. Doctor of Philosophy Thesis. University of Wollongong, Australia.

Page, K., Nanson, G., Price, D., 1996. Chronology of Murrumbidgee River palaeochannels on the riverine plain, southeastern Australia. J. Quat. Sci. 11 (4), 311–326. https://doi.org/10.1002/(sici)1099-1417(199607/08)11:4<311::aid-jqs256>3.0.co, 2-1.

Pagliardini, M., Gupta, P., Jaggi, M., 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1. https://doi.org/10.18653/v1/n18-1049 (Long Papers).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/d14-1162.

Perlis, R.H., Iosifescu, D.V., Castro, V.M., Murphy, S.N., Gainer, V.S., Minnier, J., Cai, T., Goryachev, S., Zeng, Q., Gallagher, P.J., Fava, M., Weilburg, J.B., Churchill, S.E., Kohane, I.S., Smoller, J.W., 2011. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol. Med. 42, 41–50. https://doi.org/10.1017/s0033291711000997, 01.

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., Leenaars, A., 2010. Suicide note classification using natural language processing: a content analysis. Biomed. Inf. Insights 3. https://doi.org/10.4137/bii.s4706.

Pollock, D.W., Barron, O.V., Donn, M.J., 2012. 3D exploratory analysis of descriptive lithology records using regular expressions. Comput. Geosci. 39, 111–119. https://doi.org/10.1016/j.cageo.2011.06.018.

Prathapar, S.A., Lawson, S., Enever, D.J., 1997. Hydrogeology of the Coleambally Irrigation Area: A Brief Description for Use with a Groundwater Simulation Model. CSIRO, Australia. Technical report 3/97, June 1997.

Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D., 2013. Linguistic models for analyzing and detecting biased language. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, vol. 4–9, pp. 1650–1659. August 2013.

Saco, P.M., Carpi, L.C., Figliola, A., Serrano, E., Rosso, O.A., 2010. Entropy analysis of the dynamics of El niño/southern oscillation during the holocene. Phys. Stat. Mech. Appl. 389 (21), 5022–5027. https://doi.org/10.1016/j.physa.2010.07.006.

Sahlin, J., Mostafavi, M.A., Forest, A., Babin, M., 2014. Assessment of 3D spatial interpolation methods for study of the marine pelagic environment. Mar. Geodes. 37 (2), 238–266. https://doi.org/10.1080/01490419.2014.902883.

Smirnoff, A., Boisvert, E., Paradis, S.J., 2008. Support vector machine for 3D modelling from sparse geological information of various origins. Comput. Geosci. 34 (2), 127–143. https://doi.org/10.1016/j.cageo.2006.12.008.

Srivastava, A.N., Sahami, M., 2009. Text Mining: Classification, Clustering, and Applications. CRC Press, p. 328pp.

Stewart, A.J., Raymond, O.L., Totterdell, J.M., Zhang, W., Gallagher, R., 2013. Australian Geological Provinces, 2013.01 Edition [Digital Dataset]. Commonwealth of Australia, Geoscience Australia. Canberra.

Welsh, W., Herron, N., Rohead-O'Brien, H., Cook, S., Aryal, S., Mitchell, P., Ransley, T., Cassel, R., 2014. Context Statement for the Gwydir Subregion. Product 1.1 for the Northern Inland Catchments Bioregional Assessment. Department of the Environment, Bureau of Meteorology, CSIRO and Geoscience Australia, Australia.

Wieting, J., Bansal, M., Gimpel, K., Livescu, K., 2015. Towards Universal Paraphrastic Sentence Embeddings arXiv:1511.08198 [cs].

Wu, Q., Xu, H., Zou, X., 2005. An effective method for 3D geological modeling with multi-source data integration. Comput. Geosci. 31 (1), 35–43. https://doi.org/10.1016/j.cageo.2004.09.005.

Zhou, X.W., Fu, H., Wu, C.Y., 2005. Application study of spatial interpolation method in geological random field. J. Rock. Soil Mech. 26 (2), 221–224.

Zhu, L., Zhang, C., Li, M., Pan, X., Sun, J., 2012. Building 3D solid models of sedimentary stratigraphic systems from borehole data: an automatic method and case studies. Eng. Geol. 127, 1–13. https://doi.org/10.1016/j.enggeo.2011.12.001.