Applying Suitability Distributions in a Geological Context

Robin De Mol and Guy De Tré

Ghent University, Faculty of Engineering and Architecture
Department of Telecommunication and Information Processing
Databases, Documents and Content Management research group
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium,
robin.demol@ugent.be, guy.detre@ugent.be

Abstract. Some industrial purposes require specific marine resources. Companies rely on information from resource models to decide where to go and what the cost will be to perform the required extractions. Such models, however, are typical examples of imprecise data sets wherein most data is estimated rather than measured. This is especially true for marine resource models, for which acquiring real data samples is a long and costly endeavor. Consequently, such models are largely computed by interpolating data from a small set of measurements. In this paper, we discuss how we have applied fuzzy set theory on a real data set to deal with these issues. It is further explained how the resulting fuzzy model can be queried so it may be used in a decision making context. To evaluate queries, we use a novel preference modeling and evaluation technique specifically suited for dealing with uncertain data, based on *suitability distributions*. The technique is illustrated by evaluating an example query and discussing the results.

Keywords: Imperfect information, Decision support, Preference modeling, Suitability distributions

1 Introduction

Sand is one of the most important resources available to the industry. Towards long-term sustainable exploitation, the Transnational and Integrated Long-term Marine Exploitation Strategies (TILES)[1] project partners aim to create a geological knowledge base that can be consulted for a broad spectrum of applications. Creating such a knowledge base is difficult for multiple reasons, but it is particularly challenging because there is only little data available. This in turn is due to the fact that it is expensive and time-consuming to perform offshore measurements. As a result, the available data are sparse, (geographically) non-uniformly distributed across the region of interest and span multiple decades. Making decisions based off this data must be done with these things in mind to avoid arriving at false conclusions.

[1] https://odnature.naturalsciences.be/tiles/

In this paper, we discuss how these issues are treated in the TILES knowledge base. Fuzzy set theory is used to add data quality indicators, which are taken into account during querying. Query evaluation relies on a novel evaluation technique that is particularly well suited for such data sets. This is illustrated on a case study towards the extraction of sand.

The remainder of this paper is structured as follows. In section 2, the data set from the TILES project is introduced. It is shown how it deals with uncertainty regarding attribute values. Section 3 explains what suitability distributions are and how they can be used to evaluate fuzzy queries on data sets containing imperfect information. These are then applied to the TILES data set in section 4. A fictive, fuzzy query is evaluated on the model to show how different ways of dealing with the uncertainty in the model have an impact on the results, which are also visually represented. This shows the viability of the approach and the richness of the TILES data set. Section 5 concludes the paper by summarizing its content.

2 The TILES project and voxel modeling

TILES is a project in which the partners collaborate to create a state-of-the-art knowledge base of geological information of the subsurface of the Southern part of the North Sea, off the coast of Belgium and the Southern half of the Netherlands. The intention is to use the knowledge base for a multitude of purposes regarding resource availability, long-term ecological impact of exploitation, industrial decision making, and so on. Data is available in the form of borehole samples and multibeam echosounder information which both contain information of several geological properties such as lithology and lithostratigraphy. Throughout the paper we will focus only on lithology. Lithology is essentially the study of sediment and its characteristics. Sediment is typically classified based on its average grain size. There exist predefined classifications which map specific grain size ranges to named lithological classes (see Table 1). In TILES, a simplified Wentworth classification is used, combining some of the original Wentworth classes into fewer, larger classes, until only 6 remain: *clay*, *silt*, *fine sand*, *medium sand*, *coarse sand* and *gravel*.

To best meet the needs of the project partners and stakeholders, the TILES knowledge base is designed as a *voxel model*. A voxel model is essentially a spatial partitioning of a three dimensional region into a regular grid. Each cell in the grid is called a *voxel* (short for *volume element*, similar to the pixel in 2D) and represents a unique volume of space of the original region. From a data storage point of view, a voxel is a vector of attribute values indicating the (geological) properties in a specific area. It is assumed the property values are homogeneous per voxel. In TILES, each voxel represents a cuboid space of 200 by 200 by 1 meter. This cell size is deliberately chosen to strike a balance between computational requirements and model accuracy. Alternatively, a model with irregular cell sizes could be used but this discussion is outside the scope of the paper.

Table 1: Wentworth classification table

Size range	Aggregate name
>256 mm	Boulder
64 - 256 mm	Cobble
32 - 64 mm	Very coarse gravel
16 - 32 mm	Coarse gravel
8 - 16 mm	Medium gravel
4 - 8 mm	Fine gravel
2 - 4 mm	Very fine gravel
1 - 2 mm	Very coarse sand
0.5 - 1 mm	Coarse sand
0.25 - 0.5 mm	Medium sand
125 - 250 μm	Fine sand
62.5 - 125 μm	Very fine sand
3.9 - 62.5 μm	Silt
0.98 - 3.9 μm	Clay
0.95 - 977 nm	Colloid

To construct such a voxel model from the available data, a statistical approach called kriging [9] is applied. Kriging relies on *variograms*. Briefly put, the variogram of an attribute captures the directional trends between attribute values throughout a spatial data set. The kriging process then predicts attribute values voxel per voxel based on these variograms and on measurements that lie in the voxel's vicinity. In TILES, this technique is used to predict, among others, the lithological class of each voxel. A rendering of the voxel model showing the predicted lithological class for each voxel is given in Figure 1.

Due to the scarceness and non-uniformous distribution of the available data, the reliability of the predicted lithological class for many of the voxels is questionable. In an attempt to mitigate this, each voxel is enriched with data quality indicators during the prediction process. Example data quality indicators are *variability/entropy* (in case multiple prediction runs are made) and *borehole density* (amount of true measurements in a given vicinity of the voxel), but also imprecise indicators that are derived from metadata related to the measurements themselves are stored, such as *reliability of sample analysis method*, *reliability of the measurement vintage* and so on. These metadata are often available only in the form of free-text (*core descriptions*), if they are available at all. The interpretation of these descriptions are carried out by experienced geologists and recorded in the data set. Even then, their estimated level of quality is subjective, which calls for fuzzy logic. Some examples of such data quality indicators are: “the sample was analyzed with outdated techniques”, and “the positional information of the measurement is very reliable”. During the prediction process, these data quality indicators are taken into account and derived quality indicator values are added to the voxels. When predicting the lithological class based on measurements whose sample analysis quality are annotated as “outdated”,

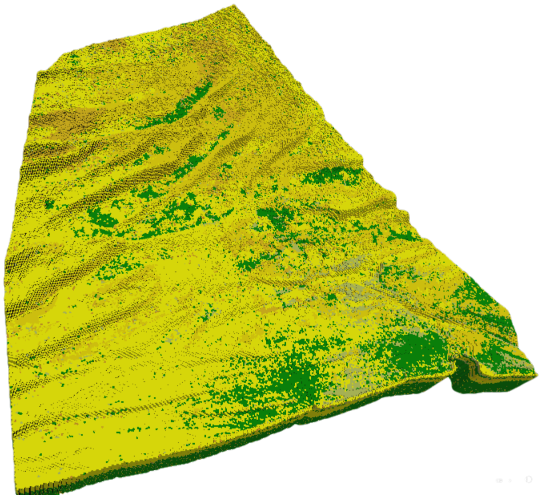


Fig. 1: An example visualization of the TILES voxel model. The voxels are colored based on their predicted lithological class.

“reliable” and “unknown”, the resulting voxel might receive a value of “possibly inaccurate” for that same data quality indicator. A voxel might look like this:

1. position: (51.1215768, 2.9187675)
2. coordinate reference system: WGS 84
3. depth from mean sea level: 37m
4. predicted lithological class: fine sand
5. prediction entropy: 0.57
6. amount of prediction runs: 100
7. borehole density: *very low*
8. vintage quality: *high*
9. sample analysis quality: *mediocre*
10. reliability description of original analyst: *unknown*
11. ...

Linguistic terms such as *high*, *mediocre* and *very low* are stored as possibility distributions over the unit interval. The inclusion of these data quality indicators essentially implies the TILES voxel model is a fuzzy database [1; 11].

3 Fuzzy Querying and Suitability Distributions

In this section we discuss how the TILES fuzzy voxel model can be queried. To that end, a special querying tool was implemented. The querying tool relies on concepts from fuzzy logic [5; 14] in order to allow decision makers to model their preferences using fuzzy sets and advanced aggregation operators [12; 13; 6; 7; 2; 10; 8]. The preference model expects that a criterion regarding a specific attribute is expressed by means of a fuzzy set that maps each value from the domain of this attribute to a number between 0 (unwanted) and 1 (preferred). Evaluating a datum is done by using this mapping to arrive at the degree associated to the datum's value. This is sometimes called the degree of suitability (of that datum for the query purpose). It is common to omit unwanted values from the mapping, and we will do so in this paper. Essentially, a fuzzy query specified by a preference model imposes a complete order on all values from the attribute's domain. The order may be partial, i.e. different values may be mapped to the same suitability degree to indicate they are considered equally suitable. Evaluating such a criterion on a set of data effectively corresponds to sorting the data according to the imposed order. It is easy to see that a regular query is a special case of a fuzzy query where all values are mapped to either 0 or 1.

Consider the following example fuzzy query preference model regarding lithological class:

- Coarse sand: 0.2
- Medium sand: 1.0
- Fine sand: 0.5
- Silt: 0.2

This can be interpreted as follows: for the purpose of the decision maker, medium sand is ideal, fine sand is good, and coarse sand and silt are equally poor, yet still acceptable. All other sediment types are considered unacceptable. The numbers used in the mapping might be based on external factors, such as a cost related to processing the resource before it can be used to serve its purpose, or arbitrarily chosen with the sole purpose of implying the desired order. In case of the latter, a different mapping that implies the same order would be equally expressive. Clearly, a preference model is subjective and is therefore best interpreted by the decision maker that created it.

In order to evaluate the fuzzy data quality indicators present in the TILES fuzzy voxel model, a novel evaluation technique specifically for dealing with imperfect information [4] was used. The premise of the technique is that evaluating a fuzzy value leads to a fuzzy degree of suitability, except in some specific cases. More precisely, evaluating a fuzzy preference model on a fuzzy datum results in a possibility distribution over degrees of suitability, called a *suitability distribution.* Computing the suitability distribution essentially comes down to evaluating all possible worlds and representing the results in a concise but complete way, without losing information. Essentially, when considering the evaluation of a datum as an operator, the computing of a suitability distribution comes down to applying Zadeh's extension principle.

Consider for example the data quality indicator *sample analysis quality.* Assume a decision maker wants to express a preference for voxels for which the sample analysis quality should be at least "reliable" such that higher quality levels are considered more preferable. This preference might be modelled by a

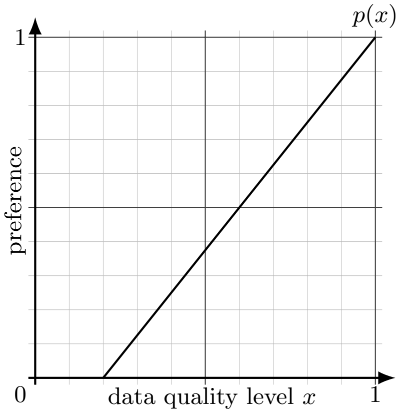


Fig. 2: An example fuzzy criterion regarding "sample analysis quality". The indicated preference model reflects that a higher value is considered more preferable.

fuzzy set with the following membership function (visualized in Figure 2):

$$p(x) = \begin{cases} 0 & x \leq 0.2 \\ (5x-1)/4 & x > 0.2 \end{cases}$$

To illustrate the evaluation process which leads to the construction of a suitability distribution, consider a voxel whose value for sample analysis quality is a fuzzy set labelled "inaccurate or possibly reliable". A possible membership function for this fuzzy set is given:

$$\pi(x) = \begin{cases} 1 & x \leq 1/2 \\ (8-10x)/3 & x > 1/2 \land x \leq 4/5 \\ 0 & x > 4/5 \end{cases}$$

This membership function and the suitability distribution that results from the evaluation of the previously defined fuzzy criterion are shown in Figure 3. The construction of the suitability distribution, $s(q)$, is now elaborated. First, the membership function of the fuzzy criterion is analyzed piecewise.

The first piece, over the range $[0, 1/5]$, denotes values that are considered equi-suitable for the purpose of the decision maker. In this case these values, denoting low quality, are all considered unacceptable. This piece, combined with the information regarding the voxel's value, can be used to derive information regarding the possibility that the voxel is not suitable for the decision maker. Therefore, we look at the fuzzy set "inaccurate or possibly reliable" in the range $[0, 1/5]$. The possibility that the voxel's quality level is such that it is not suitable regarding the criterion, is then equal to the maximal possibility that it takes a value from this range. For the voxel under evaluation, all values in this range are fully possible, hence it is fully possible that this voxel would be deemed unsuitable by the decision maker.

The second piece of the preference model implies a total order on the remaining quality values, reflecting that the decision maker wants to sort voxels by their quality from highest to lowest. This piece maps to the entire suitability range, so it will allow us to derive a possibility for each possible degree of suitability. By construction, suitability degrees not mapped by the preference model are considered impossible. Finding the possibility of each suitability degree is fairly straight forward due to the linear nature of the membership functions used in this example. A quality value of 4/5 (or higher) corresponds to a preference of $p(4/5) = 3/4$ (or higher). From $\pi(x)$ we know that all values of this quality or higher are impossible. Hence, $s(q) = 0, \forall q > 3/4$. For quality values between 1/5 and 1/2, the preference varies between $p(1/5) = 0$ and $p(1/2) = 3/8$ respectively. These values are fully possible, thus $s(q) = 1, \forall q \leq 3/8$. The remaining suitability degrees are linearly correlated to $\pi(x)$. Between 1/2 and 4/5, $s(q)$ declines linearly from 1 to 0.

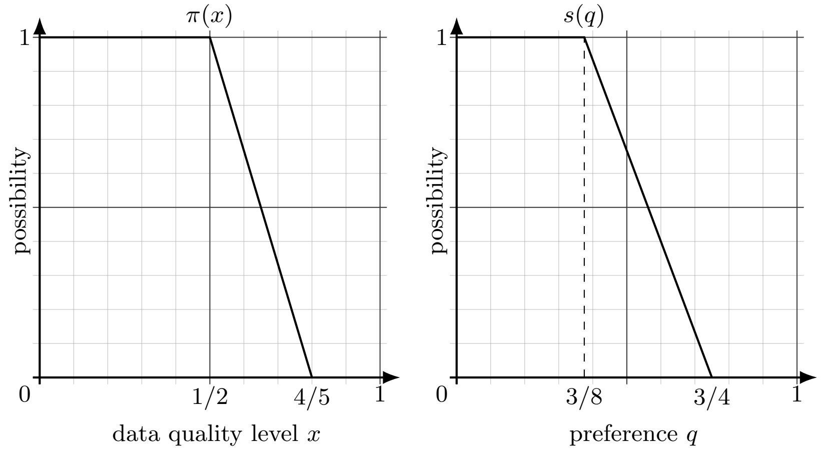
The results of the piecewise analysis is now combined by taking the pointwise maximum over the suitability range, with the understanding that the possibility is 0 where it is not defined. In this example, the overall suitability distribution is given by:

$$s(q) = \begin{cases} 1 & q \leq 3/8 \\ (6 - 8q)/3 & q > 3/8 \wedge q \leq 3/4 \\ 0 & q > 3/4 \end{cases}$$

The suitability distribution immediately conveys the following information. It is clear that this voxel can not be more suitable than to a degree of 3/4. However, it is mostly plausible that its suitability is less than 3/8. Moreover, it is fully possible that it is absolutely not suitable at all. Overall, the voxel is clearly not very suitable for the decision maker.

Note that the similarity between the suitability distribution and the voxel's fuzzy value for the data quality indicator can be explained by the shape of the preference function, which closely resembles the identity function. A fully detailed description of the suitability distribution technique can be found in [4].

Finally, one might wonder how these suitability distributions can be used in a decision support setting. Traditionally, in decision support, a preference model is used to evaluate a (possibly very large) set of systems by computing a suitability degree for each system so they can be sorted from “best” to “worst”. Using a suitability based approach, aggregating and sorting, two key concepts from decision support, is not as straight forward. There exist techniques to aggregate and compare possibility distributions, but a detailed analysis in the context of the semantics of suitability distributions has not yet been performed. Another approach, which has been studied in this context [3], is to use defuzzification. Defuzzifying a suitability distribution results in a suitability degree. In [3], it is established that different defuzzification strategies can lead to different representative suitability degrees. On the one hand, these strategies can be used to reflect a certain tolerance for uncertainty (or rather, mistakes in the suitability prediction), while on the other hand certain strategies have properties that render them particularly interesting for a specific purpose. Consider for example defuzzifica-



(a) "inaccurate or possibly reliable"

(b) suitability distribution

Fig. 3: A fictive voxel which has the value "inaccurate or possibly reliable" for the data quality indicator "sample analysis quality" on the left (3a) and the suitability distribution resulting from the evaluation of the fuzzy criterion from Figure 2 on this voxel on the right (3b).

tion through computing the area under the possibility distribution. Whereas, in general, the area under a possibility distribution may be arbitrarily large, the surface under a suitability distribution can always be scaled without modification of the semantics such that it is bound between 0 and 1. Note that this is naturally the case if the unit interval is used for representing possibility degrees and also for preference degrees, and that it is common to do so for both. This defuzzification strategy may be especially interesting to sort the distributions by uncertainty (a larger surface means a larger uncertainty) and consequently, in decision making, to identify the systems that require additional measurements. This strategy, however, would not be useful to find the "best" system. Using defuzzification, suitability distributions can be mapped to suitability degrees, thus making it possible to include them in existing decision support systems. Additionally, the decision maker can choose a different defuzzification strategies per criterion, providing an additional layer of control specifically for dealing with uncertainty.

4 A Case Study on the Extraction of Sand

In this section, we will apply the techniques described so far on real data in order to identify potential areas of interest towards medium sand extraction. Consider therefore the following fuzzy query:

> The preferred lithological class is medium sand, though fine sand and coarse sand are also acceptable, albeit to a lesser degree. Furthermore, most data quality indicators should indicate a high level of reliability.

We can break this query down hierarchically. In essence, it is a conjunction of...

- a fuzzy lithological class criterion
- the OWA aggregation (most) of...
  - sample analysis quality is high
  - vintage quality is high
  - positional accuracy is high
  - ...

The fuzzy lithological class criterion is implemented using a discrete mapping that associates medium sand to 1 and light and coarse sand to 0.5. The criteria on the data quality indicators are evaluated using the fuzzy set *high*, which is implemented by the membership function:

$$p(x) = \begin{cases} 0 & x < 1/2 \\ 2x - 1 & x \geq 1/2 \end{cases}$$

In order to be able to aggregate the suitability distributions resulting from the data quality criteria evaluations, they are defuzzified by using a "cautious optimism" strategy. This strategy takes the ordinate of the center of mass under the suitability distribution as representative suitability degree and has the property that its distinctive power is higher towards the bounds of the suitability degree domain. In other words, the distributions that are defuzzified onto suitability degrees near 0 (or 1) are reliably bad (respectively good). The downside of this strategy is that, for distributions that are defuzzified onto suitability degrees near 0.5, it is unclear whether the datum is uncertain or if it is known to be of mediocre suitability. We choose this defuzzification strategy because we are only interested in the top-k best voxels and thus do not care about the voxels with mediocre suitability degrees. A more detailed analysis of this approach can be found in [3].

The aggregation of the suitability degrees regarding data quality is translated into an OWA operator with the semantics of "most", which is implemented using the shape function $y = \sqrt{x}$. For each voxel, the final, global suitability degree is computed using a pure conjunction, corresponding to the mathematical *minimum* of the overall data quality suitability and lithological class suitability. This global suitability degree is representative of the overall degree to which the voxel is suitable for our purpose of finding and extracting medium sand with low risk of the model giving false information, as per the preference for high data quality indicators. For the TILES data, it makes most sense to explore the results visually, using a three dimensional *suitability map*. The suitability map is simply a rendering of the original voxel model wherein each voxel is colored based on its global suitability degree. In this example, the suitability map (shown

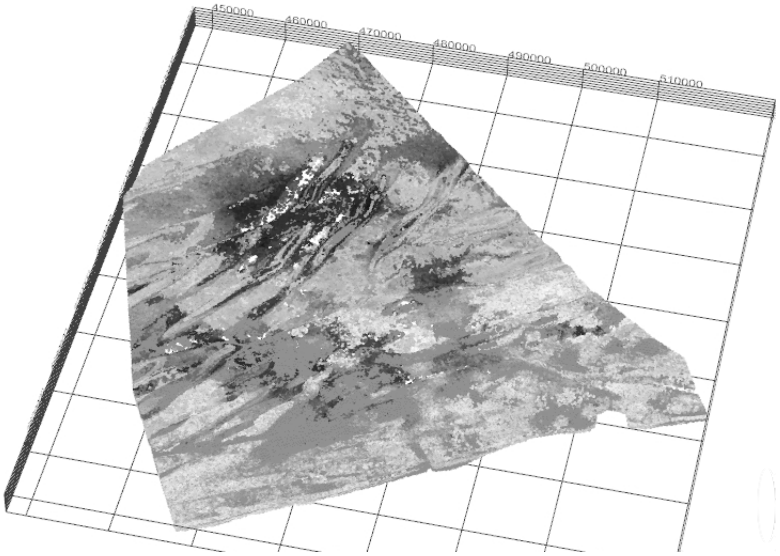


Fig. 4: The suitability map for the case study.

in Figure 4) uses a gray scale where white corresponds to unsuitable and black corresponds to maximally suitable.

Keeping in mind the chosen defuzzification strategies, we can interpret the suitability map as follows. The dark areas indicate regions where it is reliably certain that medium sand can be found. The white areas indicate regions where it is reliably certain that nor medium sand, fine sand nor coarse sand can be found. The remaining gray areas are either too uncertain for our purpose or they are reliably of lesser suitability, containing mostly fine or coarse sand rather than medium sand. These regions are ideal candidates to be examined first if the need for medium sand exceeds the amount that is available.

5 Conclusions

This paper illustrates a suitability distribution based data evaluation technique by applying it to a real data set of sediment information of the subsurface of the North sea off the coast of Belgium and the Southern half of the Netherlands. The data set contains fuzzy information, including data quality indicators, in a rasterized voxel model. It is shown how using suitability distributions makes it possible to evaluate uncertain data while still obtaining interpretive results that convey reliable information that can be important for decision makers.

Bibliography

[1] Patrick Bosc and Olivier Pivert. Fuzzy queries against regular and fuzzy databases. In *Flexible query answering systems*, pages 187–208. Springer, 1997.

[2] Gustave Choquet. Theory of capacities. In *Annales de l'institut Fourier*, volume 5, pages 131–295. Institut Fourier, 1954.

[3] Robin De Mol and Guy De Tré. Representing uncertainty regarding satisfaction degrees using possibility distributions. In *Advances in Fuzzy Logic and Technology 2017*, pages 597–604. Springer, 2017.

[4] Robin De Mol, Antoon Bronselaer, and Guy De Tré. Evaluating flexible criteria on uncertain data. *Fuzzy Sets and Systems*, 2017.

[5] Didier J Dubois. *Fuzzy sets and systems: theory and applications*, volume 144. Academic press, 1980.

[6] Jozo J. Dujmović. A comparison of andness/orness indicators. *Proceedings of the 11th Information Processing and Management of Uncertainty international conference (IPMU 2006)*, pages 691–698, 2006.

[7] Jozo J Dujmovic. Continuous preference logic for system evaluation. *IEEE Transactions on Fuzzy Systems*, 15(6):1082–1099, 2007.

[8] Michel Grabisch. The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89(3):445–456, 1996. ISSN 03772217. doi: 10.1016/0377-2217(95)00176-X.

[9] Daniel G Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

[10] Michio Sugeno. Theory of fuzzy integrals and its applications. *Theory of Fuzzy Integrals and Its Applications*, 1975.

[11] Motohide Umano. Freedom-0: a fuzzy database system. *Fuzzy information and decision processes*, pages 339–347, 1982.

[12] Ronald R Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on systems, Man, and Cybernetics*, 18(1):183–190, 1988.

[13] Ronald R. Yager. Families of OWA operators. *Fuzzy Sets and Systems*, 59: 125–148, 1993. ISSN 01650114. doi: 10.1016/0165-0114(93)90194-M.

[14] Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.