LITHOLOGICAL UNCERTAINTY EXPRESSED BY NORMALIZED COMPRESSION DISTANCE

Jānis JĀTNIEKS, Tomas SAKS, Aija DĒLIŅA, Konrāds POPOVS

Faculty of Geography and Earth Sciences, University of Latvia, janis.jatnieks@lu.lv, puma.lu.lv

Acknowledgments. This work was supported by the European Social Fund project "Establishment of interdisciplinary scientist group and modelling system for ground-water research", 2009/0212/1DP/1.1.1.2.0/09/APIA/VIAA/060

Introduction

Lithological composition and structure of the Quaternary deposits is highly complex and heterogeneous in nature, especially as described in borehole log data.

This work aims to develop a universal solution for quantifying uncertainty based on mutual information shared between the borehole logs. Such an approach presents tangible information directly useful in generalization of the geometry and lithology of the Quaternary sediments. Such generalization can be of use in regional groundwater flow models as a qualitative estimate of lithological uncertainty involving thousands of borehole logs would be humanly impossible due to the amount of raw data involved. Our aim is to improve parametrization of recharge in the Quaternary strata. This research however holds appeal for other areas of reservoir modelling, as demonstrated in the 2011 paper by Wellmann & Regenauer-Lieb.

Data

For our experiments we used extracts of the Quaternary strata from general-purpose geological borehole log database, maintained by the Latvian Environment, Geology and Meteorology Centre, spanning the territory of Latvia (Takčidi 1998). Lithological codes were manually generalised into 20 rock types using a lithology classifier for all lithology codes in the database **(Table 1).**

For generating the results in this poster presentation, 3210 borehole logs were used. These logs had to have at least 4 distinct lithology layers, according to classifier in Table 1 and pass through Quaternary deposit cover.

class	lithology
1	sandstone
2	sandstone-low conductivity
3	sandstone-high conductivity
4	limestone
5	limestone-low conductivity
6	dolomite-high conductivity
7	dolomite
8	domerite (clayey dolomite)
9	clay
10	loam
11	sandy loam
12	silt
13	sand
14	gravel
15	gypsum
16	peat
17	gyttja and other biogenic sediments
18	soil
19	other

Table 1. Main lithology classifier table. Quaternary deposit logs consist mainly of classes 9-14,16-19.

Methods

Our calculation of borehole log similarity relies on the concept of information distance, proposed by Bennet et al. in 1998.

This was developed into a practical data mining application by Cilibrasi in the 2007 dissertation. The resulting implementation called CompLearn utilities provide a calculation of the Normalized Compression Distance (NCD) metric.

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

The **C** in this formula denotes a length of output by a universal data compression program such as zlib for generating the popular zip archives, bzip2, 7zip, PPMd or, in principle, any other data compression library. The NCD is dissimilarity metric, expressing dissimilarity in range between 0 and 1 (**Figure 3**).

The universal data compression algorithms, used in this way, estimate the mutual information content in the data. This approach has proven to be universally successful for parameter free data mining in disciplines ranging from molecular biology, handwriting recognition, creation of language trees and a multitude of other surprisingly different applications (Cilibrasi, 2007).

To improve this approach for use in geology, it is beneficial to apply a transformation of borehole log data as a pre-processing step. Text stream compressors, such as prediction by partial matching (PPM), used for computing the NCD metric in our experiments, is highly dependant on context. We assign unique symbols for aggregate lithology types and serialize the borehole logs into text strings **(**such as in **Fig. 2)**, where the string length represents normalized borehole depth **(Fig. 1).** This encoding ensures that lithology types as well as thickness structure and sequence of strata is comparable in a form, most native to the universal data compression software, that calculates the pairwise NCD dissimilarity matrix **(Fig. 3).**

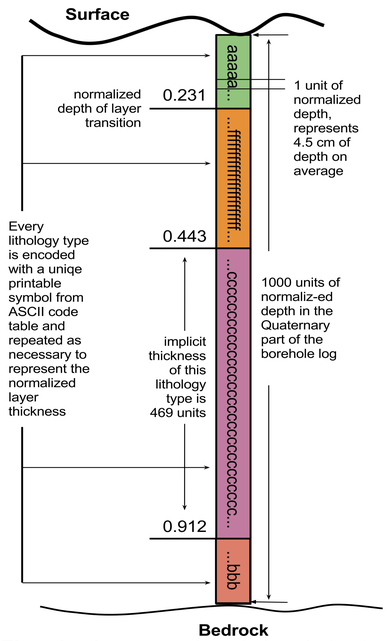


Fig. 1. An example of borehole log serialization transformation for a synthetic borehole with 3 layer transitions and 4 layers with different aggregate lithology types.

Borehole ID	Depth, m	Depth normalized to thickness of Quaternary cover	Lithology Class
11390	8	0.1	sand
11390	12	0.151	sand
11390	15	0.188	silt
11390	24	0.301	clay
11390	28.8	0.361	silt
11390	56	0.703	loam
11390	58.8	0.738	gravel
11390	71	0.891	sandy loam
11390	78.7	0.987	gravel
11390	79.7	1	gravel

Table 2. Sample borehole log with lithology classes and depth, normalized to Quaternary thickness.

Fig. 2. Normalized borhole log from Table 2 in serialized text form. This encodes the thickness, sequence and lithology type of layers in a form well suited for text compression programs such as PPMd (Shkarin 2002) that are used for calculating the NCD metric.

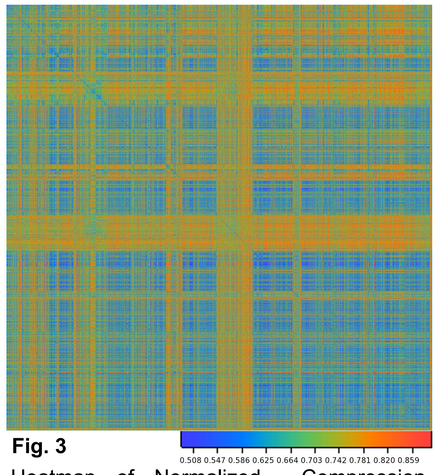


Fig. 3 Heatmap of Normalized Compression Distance matrix calculated using Shkarin's **PPMd** compression algorithm on 3120 serialized borehole logs.

The NCD results can be used for studying the structure of Quaternary sediments from the perspective of similarity, according to universal entropy coding copression algorithms (Cilibrasi 2007).

We used complete-link hierarchical clustering, as implemented in C Clustering library by Hoon et al. on two experiment matrices – NCD and NCD + range normalized Euclidean matrix of borehole locations (NCD+E). The differene of spatial clustering results is shown in **Figure 6** and structure of NCD+E clustering shown in dendrogram in **Figure 5.**

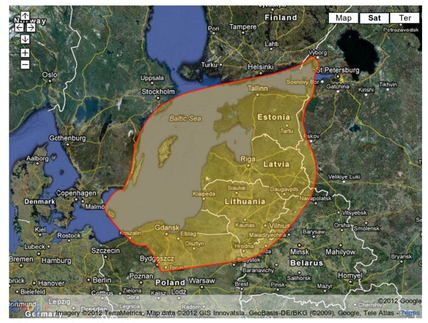


Fig. 4 (left). Spatial context of modelling territory. Spatial cluster solutions in Figure 6 cover the territory of Latvia.

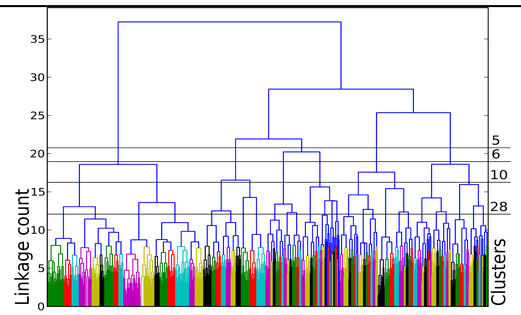
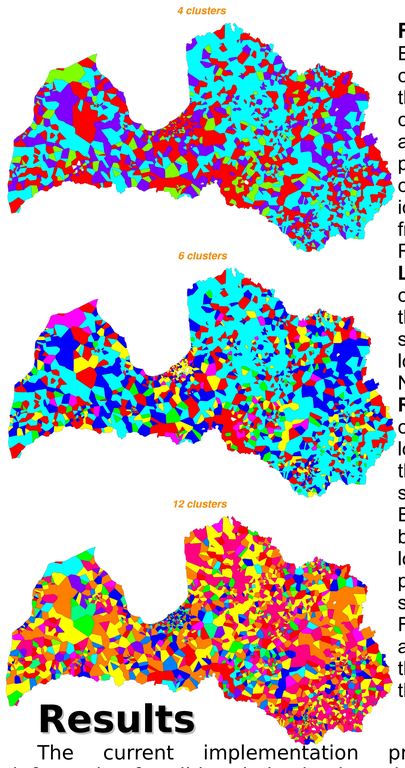


Fig. 5. Complete-link agglomerative hierarchical clustering dendrogram of Normalized Compression Distance matrix + range normalized borehole Euclidean distance matrix.



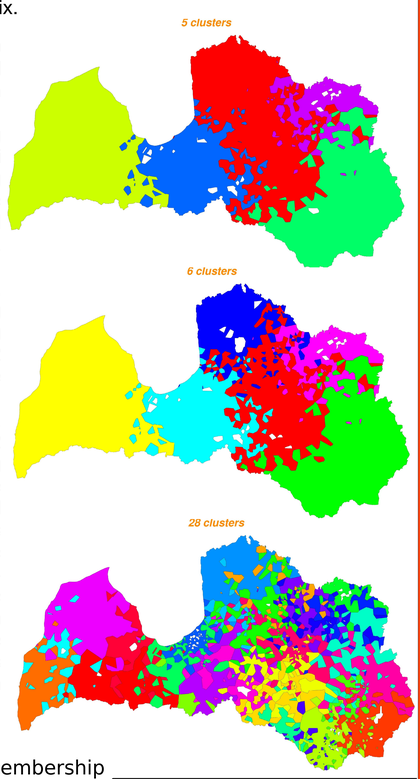


Fig. 6 (left & right) Examples of spatial clustering solutions in the territory of Latvia, created by dissolving adjacent Voronoi polygons with matching cluster membership identifiers, generated from borehole locations. Random colors.
Left - Spatial clusters, created from flattening the hierarhical clustering solution into 4,6 and 12 logical clusers, using the NCD matrix distances.
Right - Spatial clusters, created from 4,6 and 12 logical clusters, using the NCD matrix distance sum with normalized Euclidean distances between borehole locations. EPSG:25884 projected coordinate system. Dendrogram in Fig. 5. Note that there are more spatial clusters than logical clusters in the dendrogram.

Results

The current implementation provides cluster membership information for all boreholes in clustering solution as well as numerical determination of the most representative borehole section for each cluster. This information can be used for generalization of 3D lithological structure in hydrogeological modelling as it allows for minimization of uncertainty in the Quaternary strata.

Output for delineation of regions of similar lithological structure **minimizes uncertainty** by delineating regions with quantifiably more similar Quaternary lithological structure using Voronoi tesselation and NCD as a measure of dissimilarity.

This work paves way for practical application of NCD metric to basin modelling, where these results can be used as input in hydrogeological models for building a better representation of Quaternary lithological structure.

References

1. Bennett, C. H., Gacs P., Li M., Vitanyi P., Zurek W. 1998, Information Distance, IEEE Transactions on Information Theory, 44(4), 1407-1423., IEEE.
2. Cilibrasi, R. 2007., Statistical Inference Through Data Compression, ILLC Dissertation Series DS-2007-01, Institute for Logic, Language and Computation, Universiteit van Amsterdam.
3. Hoon M., Imoto, S., Miyano, S., 2010, The C Clustering Library, The University of Tokyo, Institute of Medical Science, Human Genome Center.
4. Shkarin D., 2002, PPM: one step to practicality, Proceedings of the Data Compression Conference 2002, IEEE.
5. Takčidi, E. 1999. Datu bāzes "Urbumi" dokumentācija [Documentation of the database "Boreholes"]. Valsts ģeoloģijas dienests, Rīga. [In Latvian].
6. Wellmann J.F., Regenauer-Lieb K., 2011, Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models, Tectonophysics, Elsevier (in press).