Building a national scale offshore soils database from both word-based and numeric datasets

New software converts word-based descriptions of geologic materials into numerics for use in GIS and Numerical Modelling.
The first application has produced detailed seabed materials maps for the Australian EEZ

Chris Jenkins
Ocean Sciences Institute (OSI),
University of Sydney (F05), Australia

Back to dbSEABED

An overwhelming amount of data on seabed materials is in the form of word-based descriptions. With the rapid growth of mapping systems, relational databases and numerical modelling tools, a strong need has arisen for ways to transform that data into numerics which are much better suited to computerised display and analysis systems. A new technique for doing this draws on the concepts and processes of parsing, thesaurus, syntax and fuzzy set theory.

Reference: Jenkins, C.J. 1997. Building Offshore Soils Databases. Sea Technology, **38(12)**, p. 25-28.

Word based data

Seafloor materials have been described linguistically rather than numerically for good reason - they are extremely complex in their composition, structure and properties. Composition, for example, involves chemistry and grain type; grain type is complex in itself, embracing mineral composition, grain structure, alteration, shape and origin.

A large number of numbers is needed in order to describe a sediment effectively. In practice sedimentologists make a compromise and precision is traded for brevity. Descriptions are made which convey characteristics to a given level of accuracy by assigning each sample to a number of sediment classifications , like "muddy", "phosphatic" or "cemented". It is worth noting that even in new acoustic seafloor classification systems outputs are provided as descriptive word data.

Offshore scientists and surveyors look set to continue using word-based descriptions, not only for brevity, but because the investments in equipment and time for detailed numerical measurement of seabed properties over large regions are unlikely to occur. A vast store of descriptive seafloor data has been amassed over the decades and continues to be amassed. Considering the urgent need for seafloor datasets to support decisions in environmental management, offshore industry and fisheries, we need to use all this data

How can word-based data be used in modern computational mapping and modelling systems that prefer numerical types of data? Some GIS and RDB systems do handle word data but not flexibly. Classically, it is done by keyword or synonym searches and the results may be ranked by frequency of occurrence (as in World Wide Web searches).

New approach

These standard techniques do not recognise meanings / inferences which occur in word-based data and are obvious and important to a sedimentologist - for instance that a "bryozoan sand" is physically a gravelly sand composed of low sphericity, highly porous and crushable carbonate grains. Neither do they make use of abundance weightings built into the terms and description syntax such as terms like "slightly" or "abundant".

Our approach allows the implied and explicit meanings in word-based descriptions to be used. The method combines a parser, a thesaurus, syntax recognition, some symbology in the database and also fuzzy set theory. It is probably widely applicable in word-rich geological and ecological sciences.

Most sediment descriptions contain constructions like: Ö [quantity] (modifier) object Ö . In the example, "slightly muddy relict bryozoan sand with rare green shark teeth" we recognize 4 objects and their modifiers: [slight] ( ) mud, [ ] (relict) bryozoa, [ ] ( ) sand, [rare] (green) shark_teeth. In this way the description is analogous to a linear equation of the form: n * a x + m * b y + ... = total .

Recognition of the objects, modifiers and quantities is assisted using a simple and easily implemented symbology in the database: for example, "rare/ A" and "A /rare" signify that A is rare; "coarse-" and "soupy-" are modifiers; also terms like "bryozoan sand" can be welded into one, "bryozoan_sand", here so that instances of compositionally ambiguous "sand" are minimised.

The objects, modifiers and quantifiers are identified during parsing by reference to a look-up table which is both a dictionary and thesaurus. Synonyms - words or numerics - can be assigned from the table. For example "beachrock" is simply "rock" in terms of texture, "carbonate" in composition and "weakly cemented" in consolidation. Synonyms also allow a table of grain type abundances to be created so a GIS can display geographic abundance patterns of sediment constituents - such as heavy minerals, bryozoa or phosphate.

Many objects (with their modifiers) have a characteristic grain size and composition. "Bryozoa" for example are 100% carbonate and about -1phi in average grainsize. From a description a weighted sum is formed of the carbonate, grainsize and other attributes of each object where that data is available (an 'unknown' component is also accumulated and is used to accept/reject a parsing). Weightings can be derived from quantity factors internal to the term (e.g., "marl" ~50% carbonate) and/or attached in syntax as in the case of terms like "/rare". Where a syntax obeys a rule of most or least significant component last, objects are (gaussian) weighted according to their number and position in the description.

The software also handles petrological grain count data but differently than for descriptions.

Outputs

The end result of the parsing operation is a tabulated set of numerical aliases - approximations - for the description: component abundances, average grainsize, gravel:sand:mud ratio, carbonate percent, indexes of rock and weed abundance and other outputs which are very well suited to use in GIS and modelling systems.

Successfully parsed results are output to tables separate from those reporting actual measured grainsizes, etc. The accuracy of the parser is then tested using samples/locations having available both described and measured data. The results (Fig. 1) show a statistically good correspondence with R2=0.77 on a linear fit and the difference <+2 phi in 96% cases. For context, grainsizes of repeated samplings within a m2 area of seabed often vary by 2phi.

The reliability performance of each parsing is monitored; if the unknowns exceed a set limit (usually 10%) then no result is returned. If a word does not occur in the dictionary the parsing is aborted. Parser performance could be improved further. Separate adjustments could be made for individual datasets, regions or facies. As more combined numeric / descriptive data comes to hand the power of the 'calibrating / teaching' dataset will grow.

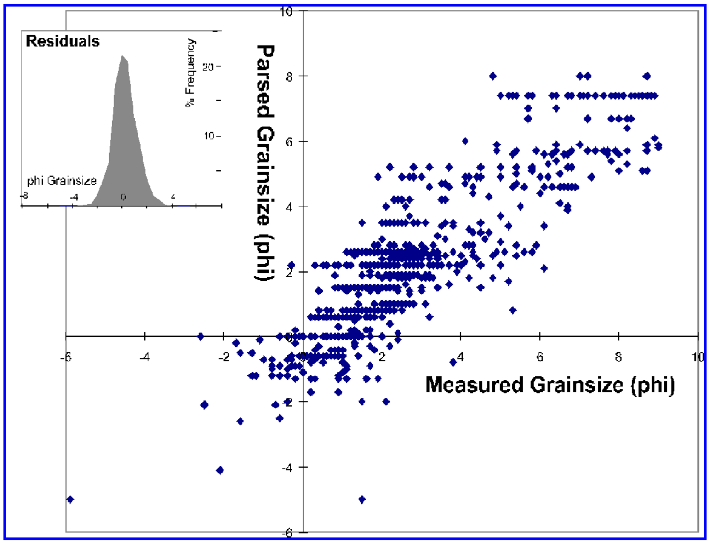

Figure 1: Parsed vs measured grainsizes
for samples where both word and numerical data are available.
(Click for detail)

Applications

Having produced numerical data from the text descriptions, and combining that with actual measurements of parameters like grainsize, it is possible - even for sparsely researched offshore Australia - to make relatively detailed maps of offshore soil types. Significantly, the use of the text-based descriptive data has most impact on the coverages in inshore zones where the call for data to assist environmental management and engineering feasibility studies are most frequent.

A key demonstration product from the parsing process is the mapping of average grainsizes for the Australian maritime area (Fig. 2). This mapping and others like it is now finding wide application in studies of seabed stability, naval acoustics, biological habitat diversity, nutrient budgets and the ground truthing of seabed swath mapping. As more data is added resolution and reliability of the data will increase.

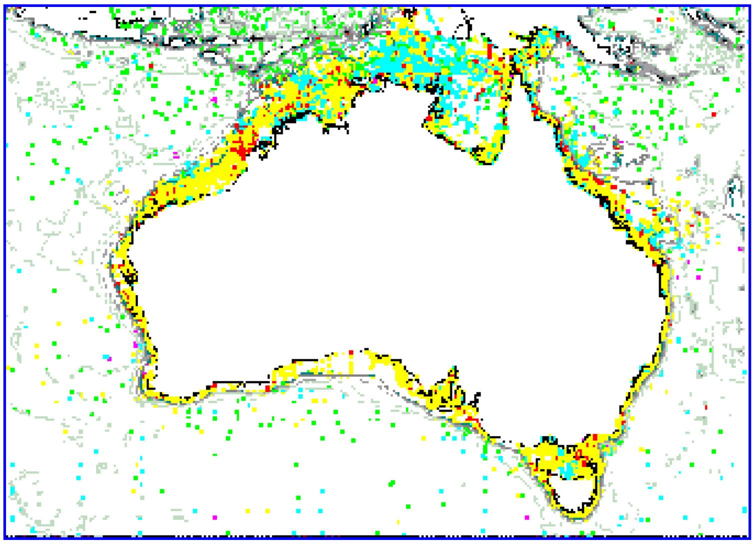

Figure 2: Grainsizes of Australian offshore sediments
(red to blue -6 to 6 phi).
(Click for detail)

Fuzzy Facies Recognition

The concept of Fuzzy Membership is invaluable as a tool for handling the linguistics of sediment descriptions. At a basic level, Fuzzy Set Theory provides a formalism for handling the component weightings within the sediment descriptions; for example, word-described grainsize classes are classic fuzzy set elements with distinct membership functions.

After the parsing Fuzzy Set Theory plays a strong role in sediment facies recognition - a most important step in dividing the offshore into zones where different processes act. A sediment facies is first defined with memberships attached, again using the [quantity] (modifier) object structure - for example, "[0.5] () mud AND [0.5] () Halimeda" for the mud-Halimeda facies. A sediment's membership of the facies is the sum of the minima of memberships (abundances) of each component - i.e., of the mud and the oatmeal-like alga Halimeda. Fuzzy OR and other fuzzy functions like CONCENTRATION can be used.

In this way each sample in the offshore database can be assigned a membership of each conceivable facies and GIS mappings of the intensity of a facies can be mapped around Australia (Fig. 3).

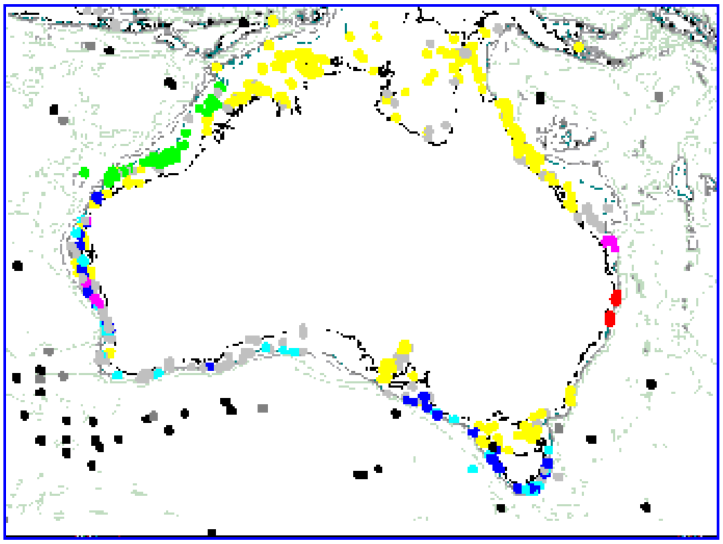

*Figure 3: Some important sediment facies of the Australian EEZ (blue-grey - bryozoan, purple - rhodolith,
green - oolith, yellow - shell, red - phosphates,
black - Mn nodules/crusts). (Click for detail).*

Summary

Word-based sediment descriptions dominate our data holdings on offshore soil types and a new method of comprehensively bringing them into numerical formats for use in GIS and numerical modelling is now available.

Chris Jenkins (1997) was a Senior Research Fellow in marine geophysics at the Ocean Sciences Institute, Sydney University, Australia. Activities include science in seabed swath mapping, seafloor physical-acoustic properties and seafloor stability. The Ocean Sciences Institute projects university marine science research into industry and government.

Back to auSEABED

(Web site update: 26-9-1999, CJJ)

Email: cjenkins@es.su.oz.au