

R E S E A R C H P A P E R

GIS data sourcing, comparison, and preparation for environmental modelling and assessment: a comparative analysis

Andrea Paparini MSc MSc PhD^{1,2,3,4*}

¹*Faculty of Science, The University of Western Australia, Crawley WA 6009, Australia*

²*Econumerics Pty Ltd Consultants, Hilton, WA 6163, Australia*

³*College of Science, Health, Engineering and Education, Murdoch University, Murdoch, WA 6150, Australia*

⁴*School of Pharmacy and Biomedical Sciences, Faculty of Health Sciences, Curtin University of Technology, Bentley, WA, Australia*

**Corresponding author.*

E-mail address: info@econumerics.com.au (A. Paparini)

Background

For geographic information system (GIS)-based environmental modelling and assessment, proper selection, collection, and pre-processing of primary or secondary data is paramount. To this end, several preliminary requirements have been identified (Reed, Brown et al. 2002), such as the need for the data to be complete, up-to-date, and accurate, span multiple spatial and temporal scales, or be properly validated by field or laboratory based experimentation. Poor or conflicting resolution across datasets is another issue often encountered, like the limited availability of appropriate (i.e., viable, useful) information.

Accessibility of data has been considerably increasing during the last decades: a welcome trend favoured also by progresses in data acquisition, archival, management, and distribution. Nowadays, GIS scientists have often the choice of using one or more comparable data sets. Despite the similarity, however, this information may not be readily comparable or truly alternative to each other. As a result, understanding how to compare the information, what the differences are, and how its use could impact the model's results, is a crucial step in environmental modelling and assessment.

Introduction

Detailed and accurate geographical representation of the different rock and soil types in a region, is essential for various aspects of land-use planning and management (Smith and Ellison 1999).

Mapping the output of desktop-based investigations based on data mining, spatial modelling, and machine learning, for instance, have already proven usefulness for understanding the spatial patterns of soil carbon in Australia (Bui, Henderson et al. 2009). Machine learning and time series analyses have also been used to study land cover changes in the Darling Range (Western Australia - WA), and assess the impact of mining activities and land rehabilitation initiatives (Vasuki, Yu et al. 2019).

In a state like WA, where mining has shaped the economy since the gold finds in the 1890s, reliable geological maps are particularly important. In the present study two analogous maps were produced to represent the generalized geologic age and features, of surface outcrops of bedrock in a WA study site. The objective was to assess how efficacy of comparable geographical representations, can be curtailed by the lack of data consistency, accuracy and precision, across alternative providers.

Data source and Methods

Two shapefiles were obtained from the Australian Department of Primary Industries and Regional Development (DPIRD) and the U.S. Geological Survey (USGS) websites. The original datasets were the Soil landscape land quality - Zones (DPIRD-017) (Department of Primary Industries and Regional Development (DPIRD) 2015), and the older Generalized Geology of Australia and New Zealand (geo3cl) (U.S. Geological Survey (USGS) - Central Energy Resources Team 1999). Additionally, water bodies locations were obtained from the 2020 update of the shape file Medium Scale Topo Water Polygon (LGATE_016) (Landgate 2016).

In ArcGIS Pro 2.5 (ESRI, USA) a geodatabase was created, and the input files were imported using the "feature class to geodatabase" conversion tool. A new feature was created to identify an arbitrary study area, covering the coastal and near-coastal regions between approximately Geraldton and Manjimup, in Western Australia. All feature classes were projected using the coordinate reference system "WGS 1984 Web Mercator (auxiliary sphere)". Finally, the projected DPIRD-017 and geo3cl files were clipped using the study area vector feature.

A new field was created in the attribute table of each geology shapefile. This was populated using new *ad hoc* groupings for the polygons within the study area. Using the “select by attribute” and “calculate field” data management tools, whenever possible, polygons were binned into common categories, across the two datasets. These included unique or mixed geological eons and eras, to which the mapped geological formations could be associated with (Cenozoic, Mesozoic, Paleozoic, Precambrian, and Archaean).

As a guide, the 2016 International Chronostratigraphic Chart, produced by the International Commission on Stratigraphy (ICS) was used (Fig. 1).

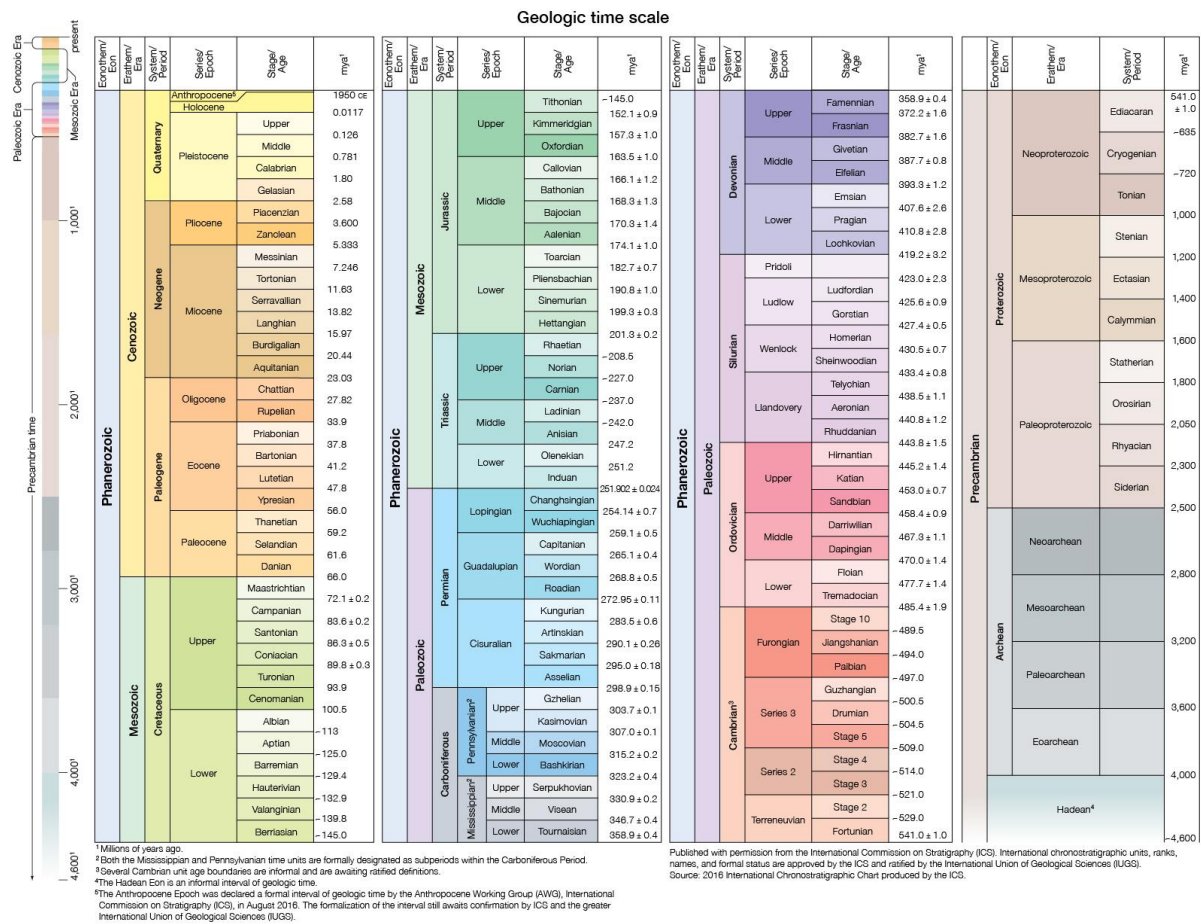


Figure 1. The 2016 International Chronostratigraphic Chart produced by the International Commission on Stratigraphy (ICS).

Results

Figures 2 and 3 show the map outputs of the present comparative analysis. Despite the apparent overall similarity, several slight variations or more important discrepancies can be noted.

A prominent feature of the DPIRD-017 map (Fig. 2) is the number of polygons forming the original dataset that could not be classified under the grouping scheme adopted here ($n=53$). These were records for which the attribute table contained no indication of the generalized geologic age of the surface outcrops of bedrock.

In the same dataset $n=6$ polygons, contained multiple land features associated with different geological periods that could not be separated. These records formed a unique class (Mesozoic/Paleozoic/Precambrian), that is not present in the geo3cl map (Fig. 3). Compared to the USGS-geo3cl map (Fig. 3) these are both important limitations, limiting the usefulness of the DPIRD-017 map (Fig. 2).

Another important difference across the maps is the large fragmentation into many polygons ($n=545$) of the original DPIRD-017 shapefile (i.e., prior to reclassification). The original count of records for each of the current classes, is given in bracket in figures 2 and 3. Remarkably, the Cenozoic class included $n=4$ and $n=434$ records, in the DPIRD-017 and geo3cl shapefiles, respectively.

The map created using the USGS data (Fig. 3) suggests a consistent longitudinal gradient of geological periods throughout the study area, with younger formations (Cenozoic) running along the coast. A notable exception is the large coastal polygon of Archean origin, south of Busselton (Fig. 3). In DPIRD-017, this was described as the Leeuwin Block (tectonic geology), moderately dissected lateritic plateau on granite, containing Tamala Limestone, some coastal dunes, and colluvial soils in the valleys. However, its geological origin was not specified in this latter dataset (Fig. 2). The longitudinal gradient somewhat evident in Fig. 3, is more difficult to appraise from Fig. 2, due to the large areas of unspecified geological periods.

More important discrepancies between the maps, were identified in smaller areas north of Geraldton, around Manjimup, and along the Avon River East north-east of Perth. Here while the USGS data (Fig. 3) identifies older formations from the Archean eon, the DPIRD survey assigns these features to younger periods (Mesozoic, near Geraldton; Cenozoic, elsewhere).

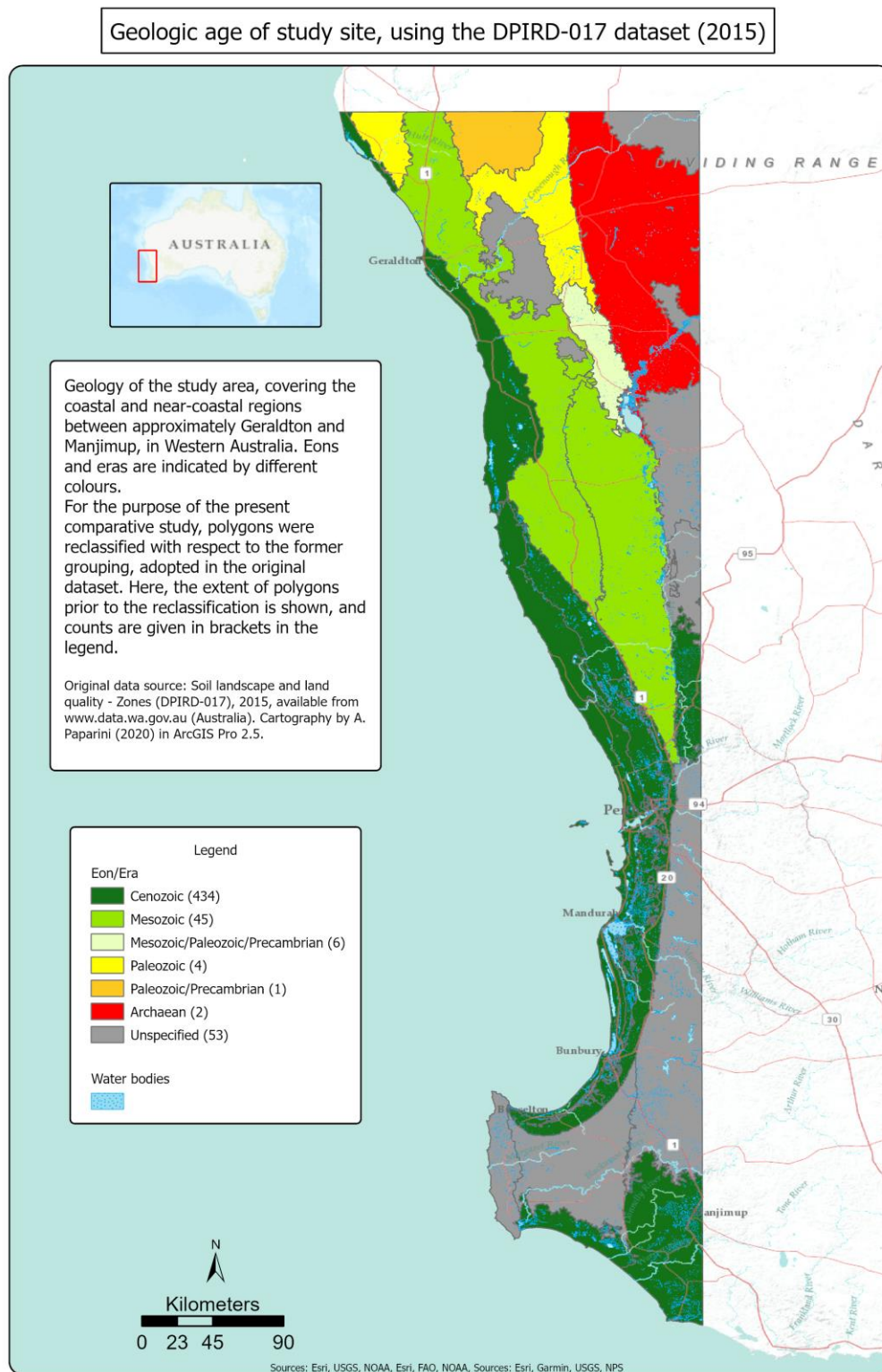


Figure 2. Generalized geologic age and features, of surface outcrops of bedrock in a WA study site, mapped using the shapefile Soil landscape land quality - Zones (DPIRD-017) (Department of Primary Industries and Regional Development (DPIRD) 2015).

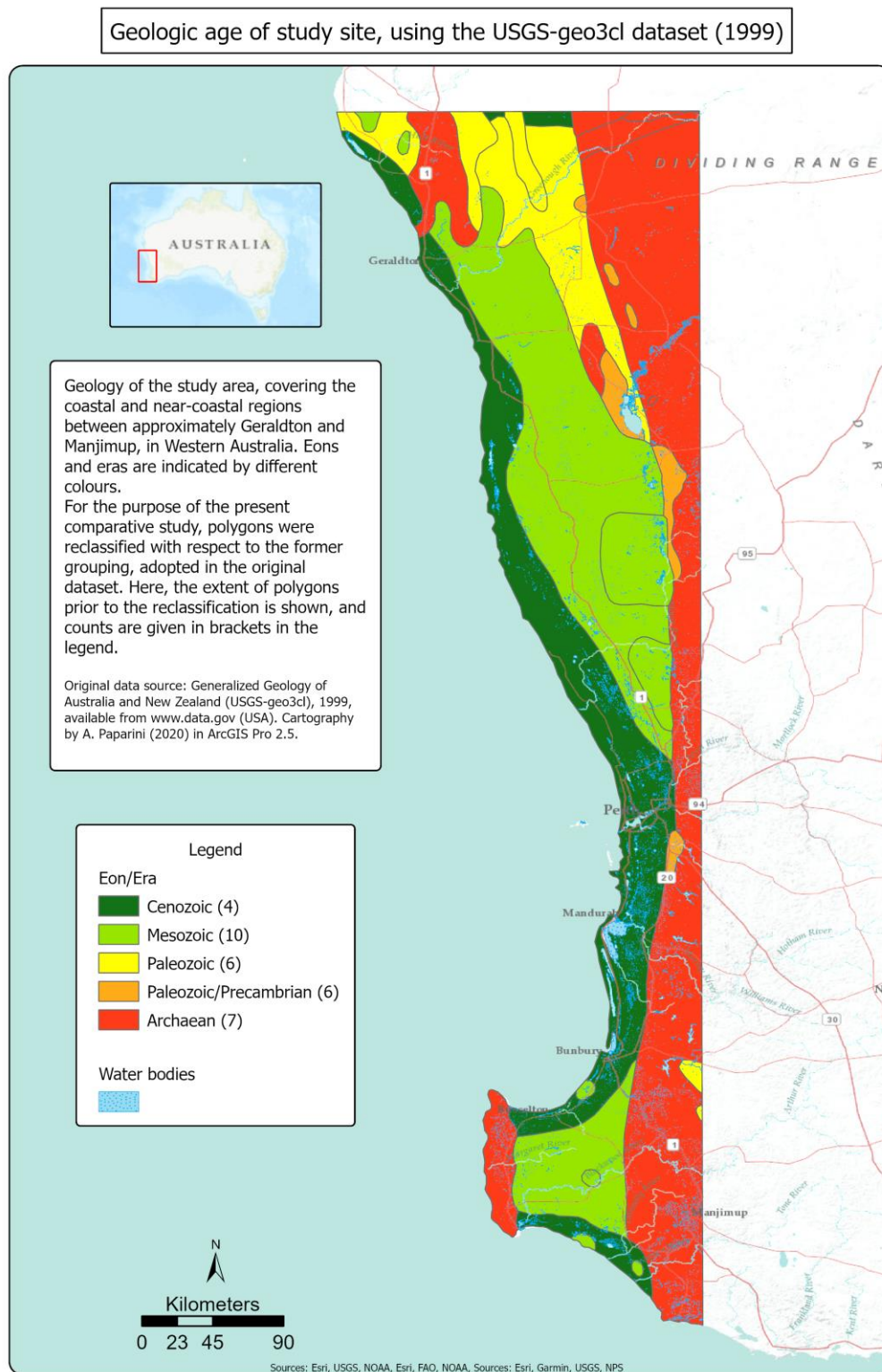


Figure 3. Generalized geologic age and features, of surface outcrops of bedrock in a WA study site, mapped using the shapefile Generalized Geology of Australia and New Zealand (geo3cl) (U.S. Geological Survey (USGS) - Central Energy Resources Team 1999).

Discussion and Conclusions

In the present study, a novel and opportunistic supervised classification was performed, in order to group polygons from two shapefiles, which had initially been created via independent surveys (1999 and 2015, then updated in 2020). It is important to note that these original surveys had likely different main objectives, clients, stakeholders, surveying technologies and techniques etc. It is little surprise that, despite showing a degree of spatial consistency (cf. longitudinal trend discussed above), the two map outputs are only partially corresponding. It is also important to note, that the speculative grouping adopted here (eons and eras; cf. Fig. 1), allowed successfully comparing the mapping outputs, but may have limited scientific validity or usefulness.

The huge number of features originally recorded in the DIPRD-017 data collection within the present study area (n=545) represents a priceless wealth of information. Recording such large number of topographical items at high resolution, is expensive and time consuming, but potentially allows answering more questions. Thus, for geospatial analyses at larger scale this map is likely to prove more useful than the USGS-geo3cl one. However, to justify this significant surveying effort, recorded items must be used for geospatial analyses and problem solving. Here, the novel and opportunistic classification created clearly overlooked the depth of the DIPRD-017 dataset, which, for the purpose of our analysis, was simply redundant.

The original unprocessed USGS geological shapefile (geo3cl), includes geology, geologic provinces, and oil and gas fields of a large portion of the Asia Pacific Region (Steinshouer, Qiang et al. 1999). Among other things, the project aims at assessing (and mapping) the undiscovered and technically recoverable oil and gas resources globally. Such maps are clearly particularly important for tenure, and explorations prior to commencing mining or other resource-exploitation operations. Yet, this information is valuable also in supporting remediation efforts of contaminated sites, for instance after mining closure or for managing oil and gas wastewater.

Figure 4 illustrates the risks of managing the wastes of oil and gas developments. While a ban on fracking currently exists over more than 98% of WA jurisdictions (Government of Western Australia 2020), the illustration allows appraising how useful geological maps such as those produced here could be in a hypothetical environmental protection plan.

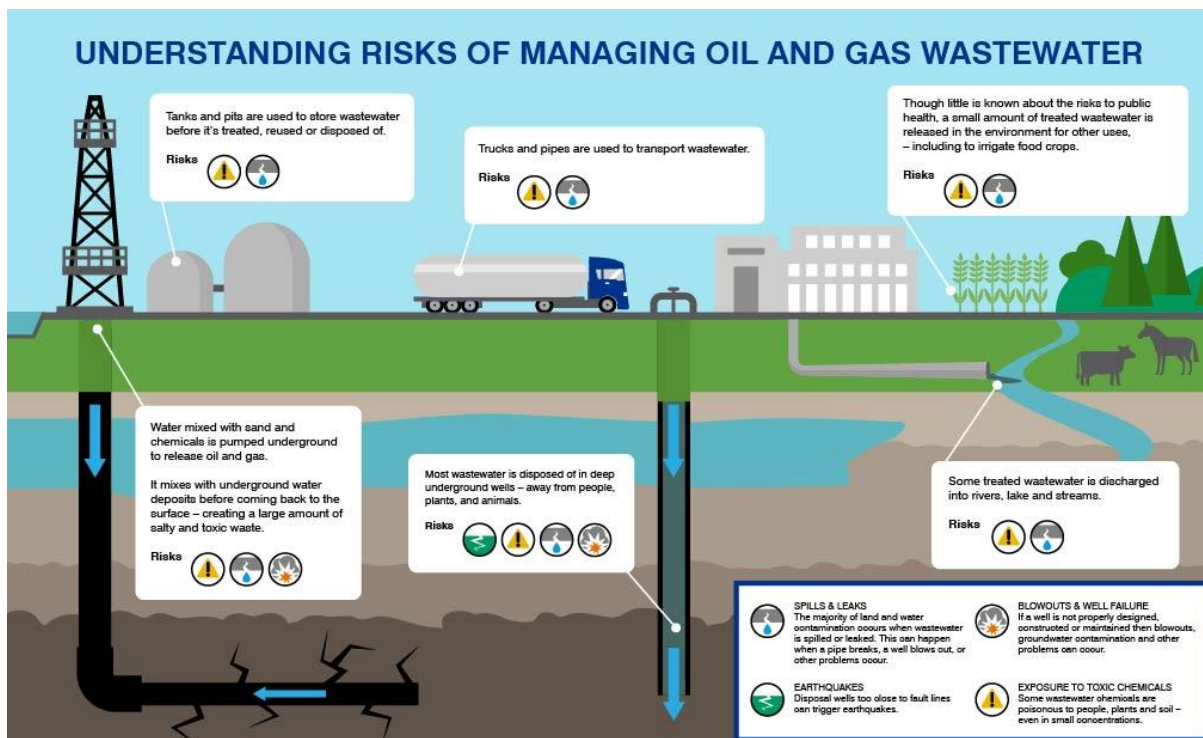


Figure 4. Risk of waste leaking to the environment, including briny wastewater generated from oil and gas development (<http://blogs.edf.org>).

The Medium Scale Topo Water (Polygon) (LGATE-016) used here (Landgate 2016), for instance, locates potentially vulnerable and precious water features, that must be protected from possible spills, disposal, or aeolian contaminants (e.g., mineral dust). On the other hand, water can also be the source of pollution, when (flowing) contaminated waterways act as vehicles of the pollutants. Penetration and dispersion of contaminants in the soil, can impact groundwater quality, with important repercussions on several human activities. The Darcy's Law modelling the behaviour of a liquid flowing through a porous medium, is one chief example of how mapping geological features can help predicting groundwater flows and in turn pollutants' spread (different media have different porosity indexes).

The USGS and DPIRD datasets could assist in identifying water catchments with different imperviousness and different runoff coefficients, due to different soil geological properties (further datasets are also required for this analysis). This has crucial implications in contaminated sites management.

Combinations of geological and groundwater maps are used to locate and understand aquitards and aquifers and assist in contamination management. Soil geology is important also for predicting soil hardness, stability, susceptibility to erosion or earthquakes (Fig. 2 and Fig. 3). From Fig. 4 it is evident how soil hardness can impact underground disposal of wastewater. Lack of stability and susceptibility to erosion or earthquakes could affect the safety of waste temporary or permanent storage sites. The coarse road network included in the base maps (Fig. 2 and Fig. 3), may also be used to plan waste disposal.

Geological provinces (Fig. 2 and Fig. 3) consists of rocks and outcrops of different (physical)chemical composition. This property influences how they chemically react to, release and sequester pollutants (e.g., chelation), and can have important implications in contamination sites management.

The present study highlighted the pitfalls posed by the direct comparison of multiple datasets, exhibiting diverse characteristics, sources, resolutions, surveying methods, or metadata formats. While comparisons are important for cross-validation, a desirable goal of many GIS analyses is to combine the information from multiple sources, in order to expand the value of individual datasets independently created.

References

- Bui, E., B. Henderson and K. Viergever (2009). "Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia." Global Biogeochemical Cycles **23**: 15.
- Department of Primary Industries and Regional Development (DPIRD) (2015). Soil landscape land quality - Zones (DPIRD-017). www.data.wa.gov.au.
- Government of Western Australia. (2020). "Implementation of the Government's response to the Independent Scientific Panel Inquiry into Hydraulic Fracture Stimulation in Western Australia." from <https://www.hydraulicfracturing.wa.gov.au/>.
- Landgate (2016). Medium Scale Topo Water (Polygon) (LGATE-016). www.data.wa.gov.au.
- Reed, B. C., J. F. Brown and T. R. Loveland (2002). Geographic data for environmental modelling and assessment. Environmental Modelling with GIS and Remote Sensing, Taylor and Francis London: 52-69.
- Smith, A. and R. A. Ellison (1999). "Applied geological maps for planning and development: a review of examples from England and Wales, 1983 to 1996." Quarterly Journal of Engineering Geology **32**: S1-S44.
- Steinshouer, D. W., J. Qiang, P. J. McCabe and R. T. Ryder (1999). Maps showing geology, oil and gas fields, and geologic provinces of the Asia Pacific region. Open-File Report. Reston, VA.
- U.S. Geological Survey (USGS) - Central Energy Resources Team (1999). Generalized Geology of Australia and New Zealand (geo3cl). www.data.gov.
- Vasuki, Y., L. Yu, E. J. Holden, P. Kovessi, D. Wedge and A. H. Grigg (2019). "The spatial-temporal patterns of land cover changes due to mining activities in the Darling Range, Western Australia: A Visual Analytics Approach." Ore Geology Reviews **108**: 23-32.