AEGC2019 Data to Discovery
Australasian Exploration Geoscience Conference
2-5 September 2019 • Perth, Western Australia
Incorporating the AIG, ASEG, PESA, and WABS

Supporting data-driven exploration in NSW

Keith Gates
Geological Survey of NSW
516 High Street, Maitland, NSW 2320
keith.gates@planning.nsw.gov.au

SUMMARY

NSW legislative changes in 2016 mean that from 1 June 2021, confidential mineral exploration company reports will become open file 5 years after submission. This will result in a large amount of previously confidential data being released on 1 June 2021.

In conjunction with the data release, substantial work is being undertaken to improve the quality, accessibility and usability of the datasets. To provide the most functional regional scale geochemical datasets possible, the usability of the datasets must be assessed by following a standard workflow of data validation > exploratory data analysis (EDA) > normalisation > interpolation.

Due to the large size of the datasets, programmatical solutions that expedite the validation and EDA processes are necessary. This presentation will give examples of methods used to condition the NSW geochemical data. An innovative data normalisation process will also be demonstrated, addressing the complexity of dealing with spatially overlapping and varied sample methods. An interpolation process that is suitable for use on large volume, large-scale datasets will also be demonstrated.

Key words: data validation, geochemistry, interpolation.

INTRODUCTION

NSW legislative changes in 2016 mean that from 1 June 2021, confidential mineral exploration company reports will become open file 5 years after submission. This will result in a large amount of previously confidential data being released on 1 June 2021. The Geological Survey of New South Wales (GSNSW) has commenced a project to review and revise its data storage model and to validate the submitted company data prior to public release. Tables 1 and 2 and Figure 1 show the volume of open file and closed file data currently held by GSNSW.

The required balance between maintaining the original data, efficient data storage and making modifications necessary to release functional datasets is a complex one. This is especially true when considering the wide and varied exploration activities undertaken by the mineral industry.

Surface geochemistry and drilling datasets are vast sources of critical information that can inform decision making processes, within GSNSW (e.g. land use management), for mineral exploration and for scientific research.

Table 1. Open file reports.

Type	Count
Drill holes	119,519
Drill hole assays	7,159,556
Drill hole samples	872,733
Surface samples	788,888
Surface sample assays	119,519

Table 2. Closed file reports.

Type	Count
Drill holes	117,803
Drill hole assays	48,700,392
Drill hole samples	4,155,640
Surface samples	716,569
Surface sample assays	13,843,622

METHODOLOGY

Interrogation of these datasets at a regional scale aims to identify geochemical anomalies via a generalised workflow of: data validation > EDA > normalisation > interpolation. The large volume surface sample dataset maintained by the GSNSW contains a variety of sampling collection (soil, stream sediment, auger, etc.) and analytical methodologies with overlapping spatial and temporal relationships.

The high volume of data necessitates that highly efficient solutions be implemented. This presentation demonstrates solutions used in the validation and EDA steps, and offers innovative solutions to the normalisation and interpolation processes.

Python™ was used as the principal programming language to perform the statistical and graphical workflows. Pandas™ and Scipy™ libraries were used for statistical analysis and Matplotlib™ for graphical representations. The workflows were incorporated into Jupyter™ notebooks to provide repeatability, versatility and user input as required.

VALIDATION AND EDA WORKFLOWS

Data validation used statistical analysis and semi-automated graphical visualisation. These semi-automated workflows expedited the validation process. Gold surface-geochemistry results (ppm) for surface samples and downhole assays have been significantly improved. Several sources of error are identified, and corrections made to the database. In the statistical workflow standard deviations, minimum and maximum values are used to identify erroneous results. The graphical visualisation used automated histogram generation to expeditiously identify distributions that do not conform to

the expectations for given sample types and elements. This helped identify erroneous results.

NORMALISATION WORKFLOW

Normalisation of geochemical data at regional scales is complicated by the spatially overlapping nature of the datasets and is compounded by varying sampling and analytical methods.

NSW geochemical dataset assay result ranges and detection limits are dependent on the numerous variations in sampling methodologies and analytical techniques. For regional scale interpretations, it is necessary to combine these subsets of data.

In company submission of the data, a single report can contain multiple sample surveys that are spatially distinct but are not flagged as such. Spatially distinct datasets will have varying underlying background levels due to changes in regolith and geology. To separate these datasets, individual survey points were spatially buffered and then unified to form distinct boundaries to form new spatial groups. A spatial intersection joined the new boundaries to the samples. Once combined, the normalisation scaling was applied, whereby the minimum and maximum values are scaled to between 0 and 1. Figure 2 shows the aggregated boundaries of the overlapping soil sampling surveys and gold assays coloured according to scaled values.

The normalisation implementation used feature scaling. This involves scaling each dataset between 0 and 1. The anomaly threshold was determined as 2 Median Absolute Deviations (MAD) from the median value (Tukey, 1977).

INTERPOLATION

Common interpolation algorithms for 2D geochemical datasets are inverse distance weighting, radial basis functions and, less frequently, kriging. The computation overhead on performing any of these techniques on a very large regional scale dataset is prohibitive. Beyond the computational complexity, the resolution of the resultant interpolations would be too fine for regional scale representations. Gridded representations that are scale appropriate have been generated using a simple maximum value within each cell. Several iterations at different cell sizes provide appropriate visualisation for different scales. Once interrogated at appropriate scale, a more robust interpolation method such as inverse distance weighting (IDW) can be applied.

CONCLUSIONS

The robustness and usability of the geochemical datasets is vastly improved by the validation process. Erroneous assay units and sample methods have been corrected. The efficient workflows presented for validation and spatial clustering are repeatable and may be applied to validate other large volume geochemical datasets.

The interpolations on normalised data provides a versatile dataset for end users. The process demonstrated a framework that can easily be applied to other large geochemical datasets.

Finally, the process itself and the tools used provides an excellent example of how simple programming procedures can offer massive gains in productivity.

REFERENCES

Tukey, J.W., 1977, Exploratory Data Analysis. Addison-Wesley, Reading, 688 pp.

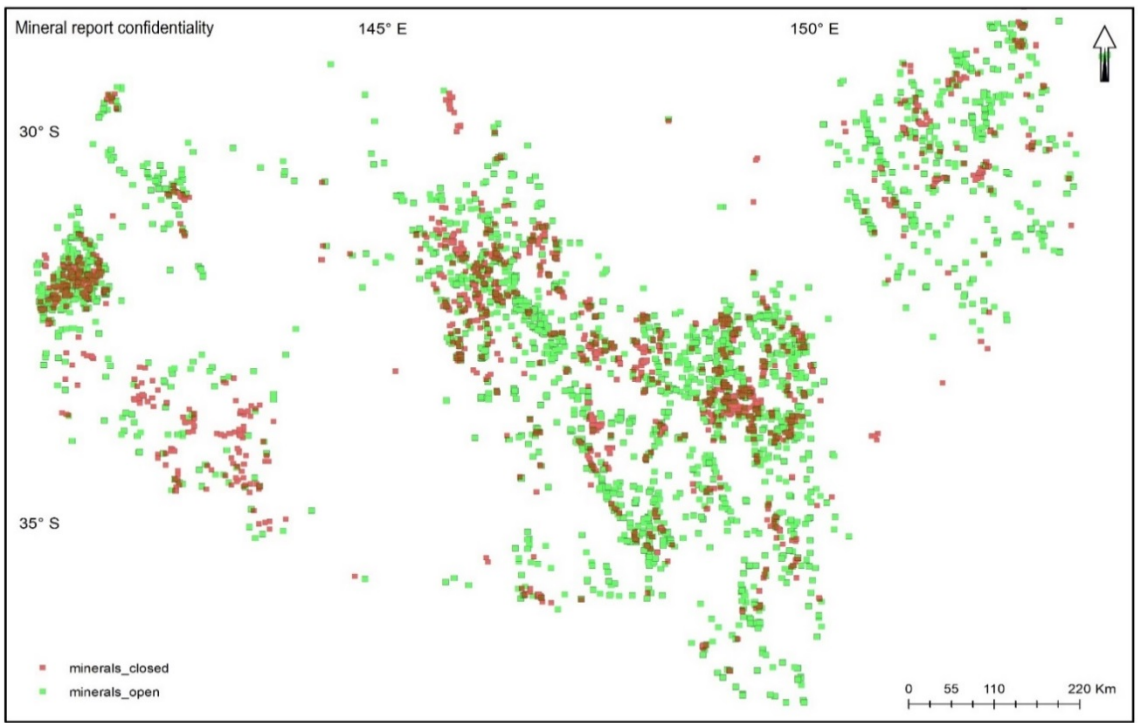

Figure 1. Open file (green) and closed file (red) reports.

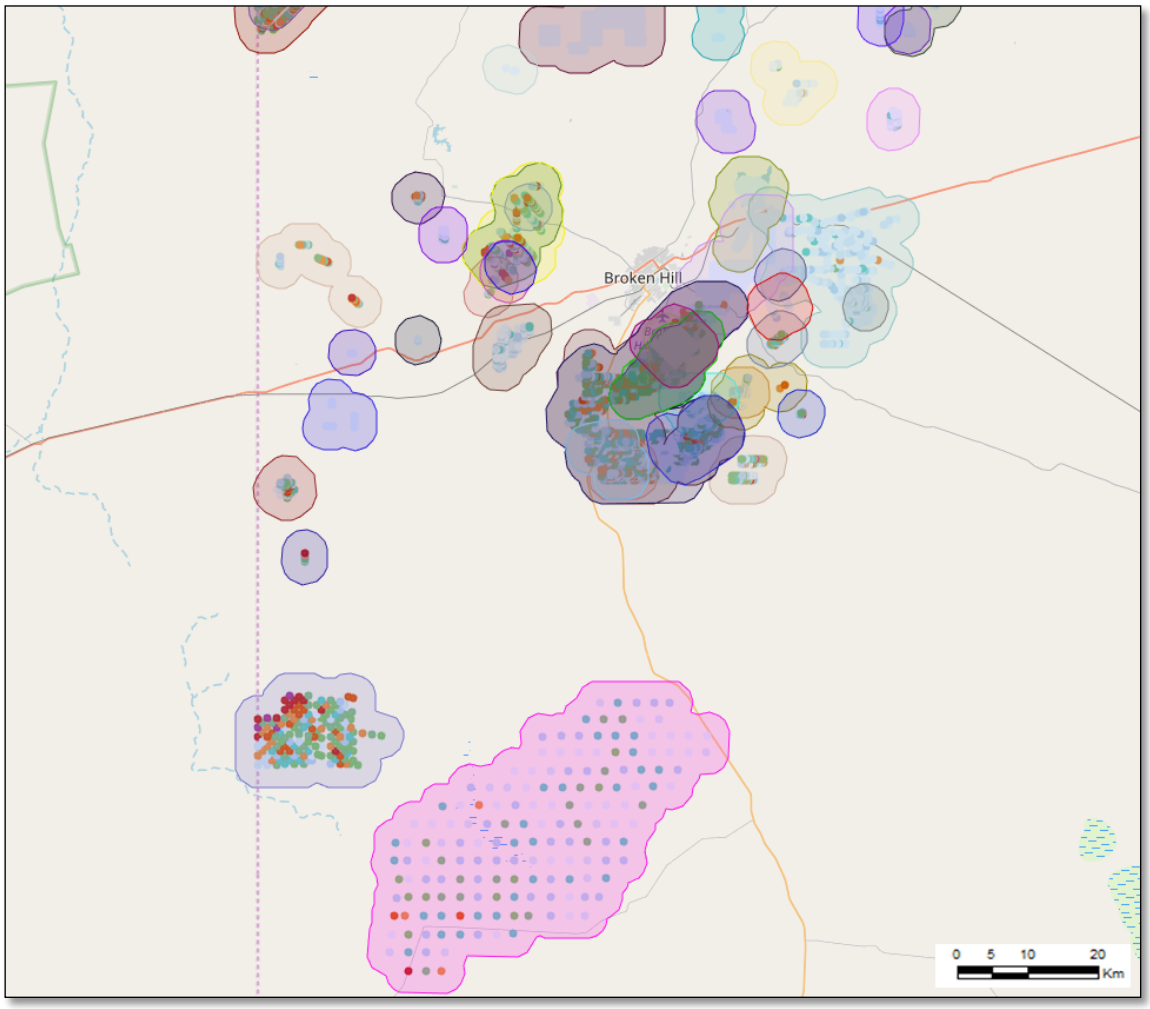

Figure 2. Open file surface geochemical samples analysed for gold near Broken Hill. The coloured areas represent the buffered extent of individual surveys. Coloured points range from low (cool colours) to high (red) gold assay values within each survey area.