Surficial geochemical data extracted from assessment reports: Development and initial release of the database

Yury Klyukin[*]
Yukon Geological Survey

Klyukin, Y., 2024. Surficial geochemical data extracted from assessment reports: Development and initial release of the database. *In*: Yukon Exploration and Geology Technical Papers 2023, L.H. Weston and Purple Rock Inc. (eds.), Yukon Geological Survey, p. 51–56.

Abstract

The Yukon archive of assessment reports contains a significant amount of data; however, accessing these data is challenging due to the absence of a centralized storage system. This paper introduces the Geochemical Assessment Report Data Extracted database (GARDEd), which is specifically designed to store these data. Initially, GARDEd was constructed using data from assessment reports from a 75 km radius of the Casino deposit in the Yukon, but this spatial limit was subsequently removed. The initial release of the database contains geochemical data describing more than 300 000 surficial samples from more than 300 assessment reports submitted after 2004.

The structure of GARDEd follows the data model developed by the British Columbia Geological Survey for storing surficial geochemical sample data acquired from their assessment reports. This paper outlines the structure of the database and discusses the workflow for extracting the data. The paper also describes a custom Python tool developed to automate data extraction from digital assessment reports.

Introduction

The Yukon Geological Survey (YGS) stores a significant amount of geological information gathered by exploration companies in compliance with the Quartz Mining Act. This information is primarily contained in assessment reports (AR). Since 1920 (Fig. 1), more than 8000 hardrock assessment reports have been collected, with more than 6000 of these containing geochemical data on surficial samples (e.g., soil, rock, stream sediment, and vegetation). The sheer volume and historical depth of these data provide valuable insight into the mineral potential of the Yukon.

Exploration companies rely on data from ARs to search for exploration targets. Furthermore, regional and local studies such as those conducted by Grunsky and Caritat (2020) and Wang and Zuo (2022) rely on similar datasets to test and develop models to identify geochemical anomalies in their areas of interest. To aid exploration companies, YGS has initiated the development of a database that will consolidate geochemical data extracted from assessment reports.

The primary objective of this paper is to introduce the Geochemical Assessment Report Data Extracted database (GARDEd), which was created by extracting and compiling data from more than 300 assessment reports (Fig. 1). This paper details the structure of GARDEd and provides instructions for accessing the database when it is released in early 2024. The paper also introduces a toolset that was used to extract data from the reports. The toolset is still in development but is available upon request from geology@yukon.ca.

[*] yury.klyukin@yukon.ca

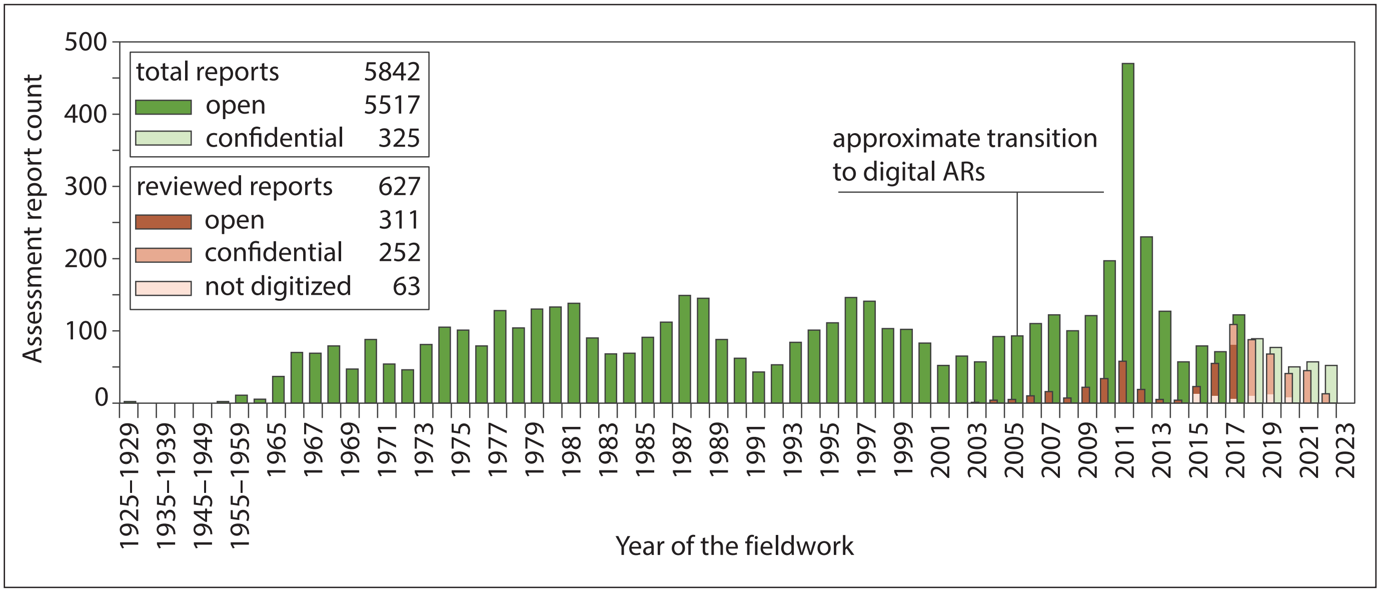

***Figure 1.** Histogram showing the number of assessment reports (AR) that contain surficial geochemical data, plotted against the year of submission. The bin size for ARs between 1925 and 1964 is 5 years, while the rest have a bin size of 1 year. Prior to 2005, most reports are scanned copies. After 2005, reports began to be submitted in electronic formats, which has facilitated data extraction efforts.*

Database structure

The structure of GARDEd is based on the work of Han et al. (2019) and Norris and Fortin (2019), who developed a data model for storing surficial geochemical sample data from assessment reports submitted to the British Columbia Geological Survey. The database described in this document closely follows the original data model by Norris and Fortin (2019), with minor modifications, which are outlined below. GARDEd is released as a Geopackage, an open standard for storing geospatial data in the SQLite file format.

GARDEd consists of tables with the prefixes "code_" and "data_". The tables with the data_ prefix store sample data and assay certificates, whereas the tables with the code_ prefix store metadata about analytical techniques and assay certificates (Fig. 2). Detailed descriptions of all fields are provided in the GARDEd manual, which will be included in the release of the database in early 2024.

The data_sample table is a central table in the database. It stores information about the sample type, description, original coordinates, and additional metadata as described in the assessment report. A notable modification made to Norris and Fortin's (2019) database model is the direct storage of the coordinates, which have been converted to decimal degrees (WGS 84) in the data_sample table. This differs from the dynamically generated table described in Norris and Fortin (2019). The converted coordinates are stored in two formats: as a Geopackage point geometry column, and as separate latitude and longitude fields (Fig. 2). The data_source_info table contains details on how the data were captured from the assessment reports. Analytical data are stored in the data_analyte table, whereas metadata extracted from assay certificates are stored in the data_cert table.

Tables with the code_ prefix store details on how samples were analyzed. The code_lab table records the names and acronyms of laboratories. The code_method table stores analytical method details described in assay certificates. The code_prep table reports on the sample preparation techniques used during the analytical process. The unit associated with each analytical result is stored as a text value in the unit_id field. In cases where assessment reports did not specify units, the units have been inferred by the author of this paper. In the Geopackage, these units are marked with the 'infer_' prefix.

The initial release of GARDEd is further complemented by the inclusion of comma-separated value (CSV) files, which can be used directly in GIS software. These files

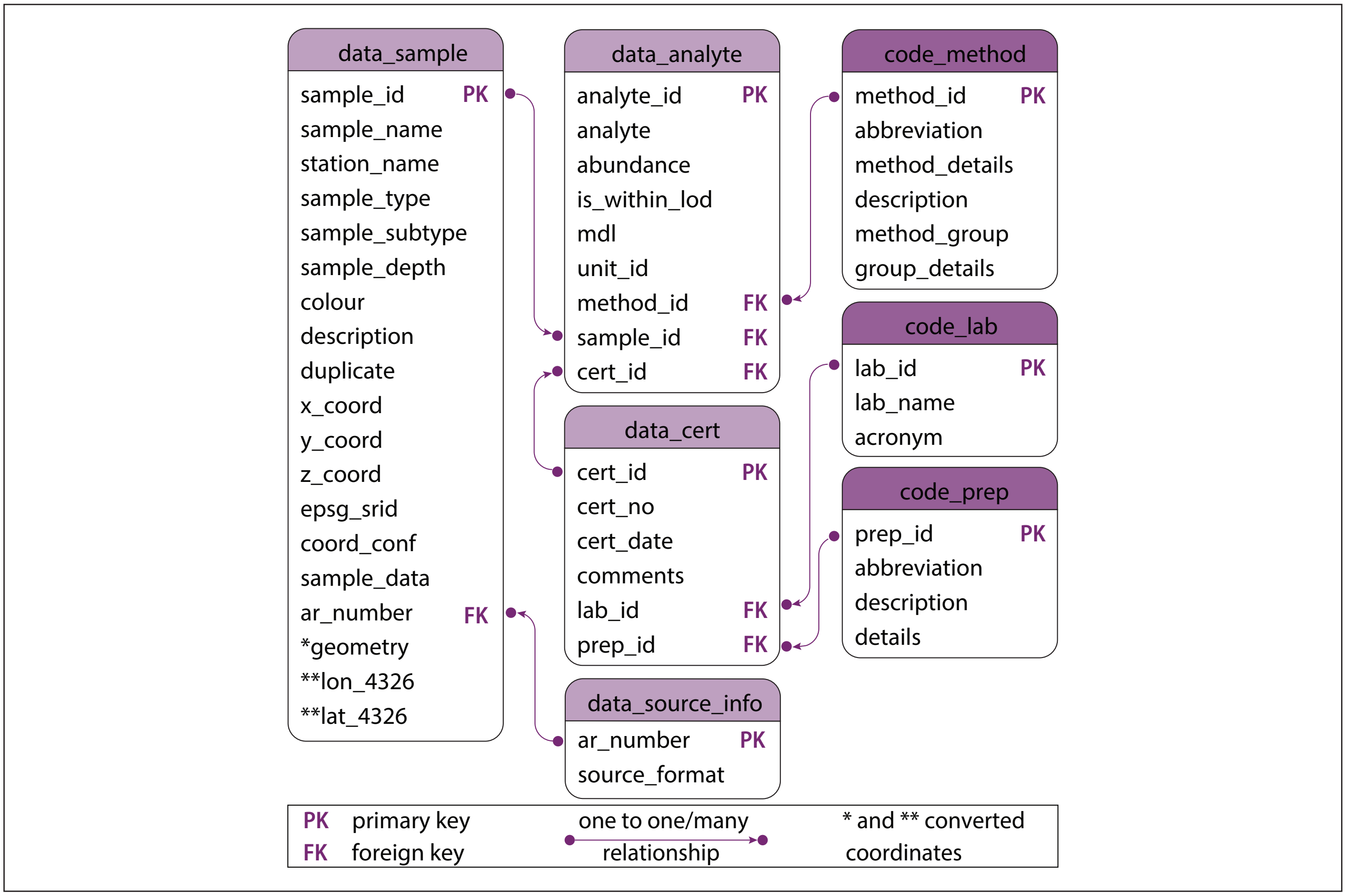

Figure 2. *Entity-relationship diagram of the Geopackage instance of GARDEd, adapted from Norris and Fortin (2019). Fields in data_sample table marked by an asterisk (*) contain longitude and latitude for sample locations converted to WGS84 (EPSG 4326) from those originally reported in assessment reports, compliant with the Geopackage specification for geospatial fields. The fields marked by two asterisks (**) are longitude and latitude stored in separate columns, facilitating their use in GIS software.*

provide a simplified, flat table version of the database, with certain data modified or removed compared to the Geopackage. For example, the CSV tables do not contain the original sample location coordinates and have element concentrations and units changed to match those specified in Table 1. In cases where abundance is reported as oxide concentrations, the measurements were recalculated into elemental concentrations. These modifications eliminate the need for an extra column to accommodate the same element reported in an oxide form or with a different unit, minimizing the width of the flat table by storing each measured element in a single column. This, however, may introduce data errors, especially in cases where the analytical unit was not explicitly stated in the assessment report and had to be inferred.

Table 1. *Standard units used to store analytical results within CSV flat tables. The originally reported analytical values and units are stored in the Geopackage version of GARDEd. Conversion to standardized units simplifies mapping of data from multiple sources; however, it may introduce errors if the original unit was incorrectly inferred during data extraction in reports that did not explicitly specify units.*

Unit	Elements
%	Al, Ca, Fe, K, Mg, Mn, Na, P, S, Si, Ti, TC (total carbon), TS (total sulfur), LOI (loss on ignition)
ppb	Au
ppm	Ag, As, B, Ba, Be, Bi, Br, Cd, Ce, Cl, Co, Cr, Cs, Cu, Dy, Er, Eu, Ga, Gd, Ge, Hf, Hg, Ho, In, Ir, La, Li, Lu, Mo, Nb, Nd, Ni, Pb, Pd, Pr, Pt, Rb, Re, Rh, Sb, Sc, Se, Sm, Sn, Sr, Ta, Tb, Te, Th, Tl, Tm, U, V, W, Y, Yb, Zn, Zr

The Geopackage GARDEd retains data as in the original assessment report and includes a view (a virtual table based on the result-set of a SQL query) named “wide_table” that can be used to generate CSV exports.

Database coverage

GARDEd was initially populated with data from samples within a 75 km radius of the Casino deposit to evaluate the data extraction procedures and assess the feasibility of the database. Data capture focused on recent ARs because these were available in a digital format. In general, ARs filed before 2005 were submitted in hard copy and subsequently scanned to PDF, so the digital file required optical character recognition (OCR), which posed data capture challenges. Once the initial test was completed and the Casino-area digital data were captured, efforts shifted to expanding sample coverage across the entire territory, while maintaining focus on digital data from relatively recent reports. The current release includes surficial geochemical data from 311 digital AR files, which taken together, is a dataset of 307 311 samples (Fig. 3). Most of the samples are located in the Dawson Range area. This concentration resulted from the initial focus on the 75 km radius around the Casino deposit, coupled with significant exploration work in the region between 2005 and 2018.

Data extraction and processing

Submitted assessment reports are stored in PDF format and may include spatial sample data and/or assay certificates as appendices in separate digital files of varying formats, making the data easily accessible; this is more common with newer reports. Reports created between the late 2000s and mid-2010s typically consist of a single PDF with all the data included in appendices within the report. Reports submitted prior to 2005 are almost always scans of hard copy reports; the quality of the scans generally decreasing in older reports. Extracting data from scanned reports requires more effort. The document must undergo OCR and the results must be reviewed before further processing is possible, and prior to adding it to the GARDEd database. Only a few scanned reports were included in the database to evaluate the required workload. As a result, the author focused on data capture from digital tables in newer ARs because these data could be captured relatively quickly and with fewer errors.

The author developed a custom tool to assist with data extraction and review. The tool is written in Python and is designed to capture tabulated data from custom tables and assay certificates generated by the laboratories. The tool heavily relies on libraries that can manipulate PDF, CSV and Microsoft Excel files; operate and analyze tables and interact with SQLite databases; and work with spatial data and operate with different coordinate systems to create dynamic maps. The tool expedites data extraction and applies quality assurance/quality control procedures to the extracted data. It automatically corrects minor inaccuracies (strips extra spaces in descriptive fields), creates dynamic sample maps showing the AR footprint overlain with the sample results that have been included or skipped in the geochemical data, and flags suspected errors for manual review.

Next steps

It is anticipated that GARDEd will be released in early 2024. It will be accessible under the Spatial Data and Compilations section of Yukon Geological Survey’s Integrated Data System, available at https://data.geology.gov.yk.ca. The YGS encourages users to send the feedback to geology@yukon.ca to assess the usefulness of the dataset and determine how much effort should be invested to capture older, non-digital ARs.

Confidential assessment reports submitted to YGS have already had their data extracted. These data will be added to future releases of GARDEd following the expiration of the confidentiality period. The YGS encourages the inclusion of digital data with future AR submissions to streamline the integration of surficial geochemistry data into GARDEd.

Summary

This paper presents the development and upcoming release of the Geochemical Assessment Report Data Extracted Database (GARDEd), which contains geochemical data extracted from ARs based in the Yukon. The paper outlines the database structure and provides a snapshot of the data captured to date.

Although the initial release of GARDEd only contains approximately 10% of the geochemical data from

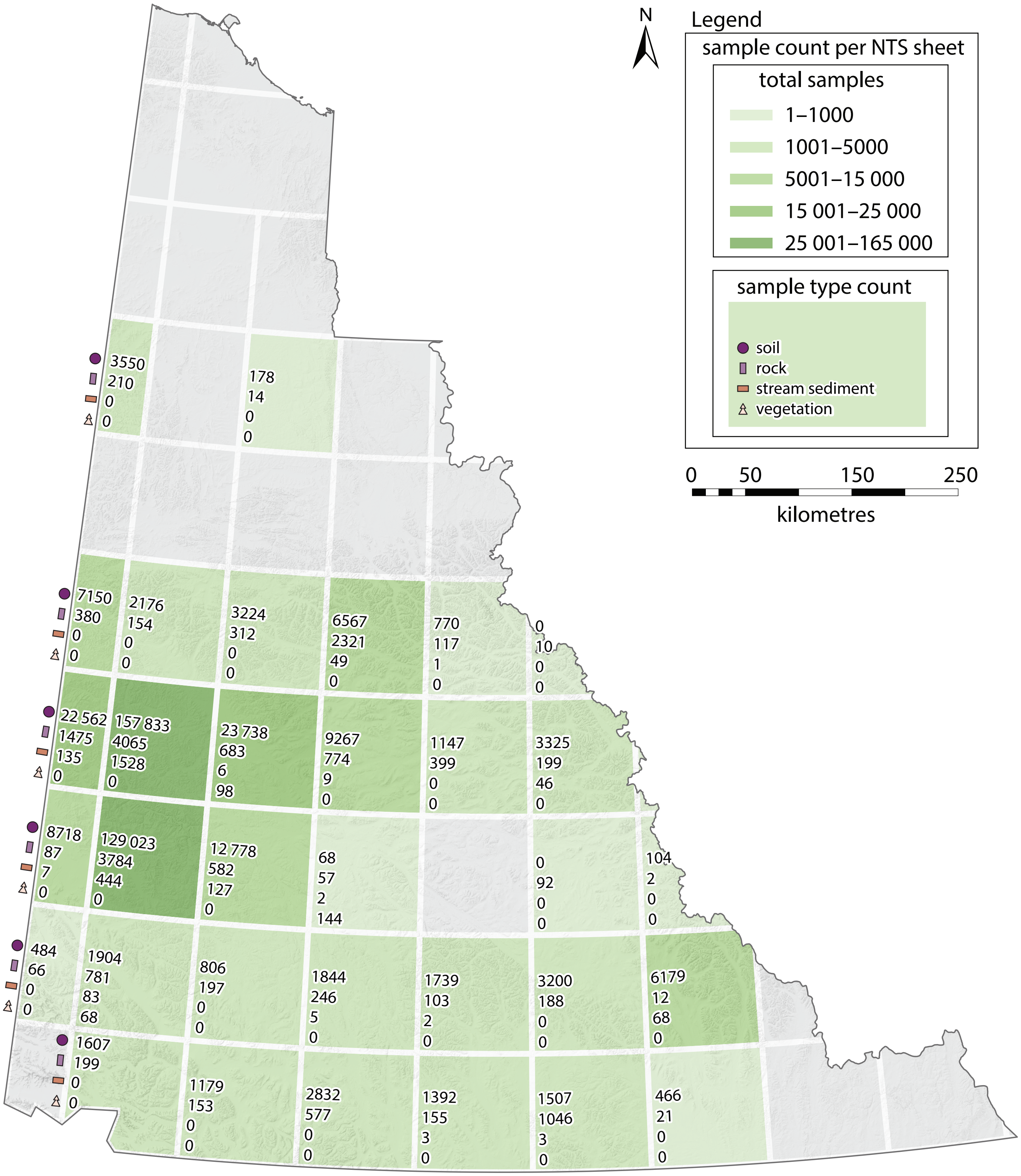

Figure 3. *Count of extracted samples per NTS map sheet. The colour gradation applies to total count of samples. Labels within each NTS map sheet indicate the count for each sample type: soil, rock, stream sediment and vegetation.*

surface samples within Yukon ARs, it serves as a valuable resource for researchers and exploration companies interested in the mineral potential of the Yukon. It provides centralized, standardized and internally consistent geochemical data for analysis and modelling. These data offer users valuable information for mineral exploration programs, allowing them to design custom maps and test different approaches to identify geochemical anomalies linked with known or potential mineralization. GARDEd is an evolving product that will be updated regularly with new data. Users are encouraged to provide feedback on their experience.

Acknowledgments

I express sincere gratitude to Brett Elliot and Chad Cote for their diligent efforts in assuring data integrity and conducting thorough database checks. Their expertise and attention to detail greatly contributed to the reliability and accuracy of the Geochemical Assessment Report Data Extracted database. Sarah Ellis helped with data extraction from half of the processed assessment reports and provided important feedback, helping with developing the front end of the data extraction tool. Thank you also to Scott Casselman for his invaluable support and encouragement in initiating this project.

References

Grunsky, E.C. and de Caritat, P., 2020. State-of-the-art analysis of geochemical data for mineral exploration. Geochemistry: Exploration, Environment, Analysis, vol. 20, no. 2, p. 217–232. https://doi.org/10.1144/geochem2019-031.

Han, T., Rukhlov, A., Riddell, J. and Ferbey, T., 2019. A skeleton data model for geochemical databases at the British Columbia Geological Survey. *In*: Geological Fieldwork 2018, British Columbia Ministry of Energy, Mines and Low Carbon Innovation, British Columbia Geological Survey Paper 2019-01, p. 125–135.

Norris, J. and Fortin, G., 2019. Assessment report-sourced surface sediment geochemical database: Development and initial data release from the Interior Plateau. British Columbia Ministry of Energy, Mines and Low Carbon Innovation, British Columbia Geological Survey GeoFile 2019-04, p. 1–10.

Wang, J. and Zuo, R., 2022. Model averaging for identification of geochemical anomalies linked to mineralization. Ore Geology Reviews, vol. 146, article 104955, p. 1–12. https://doi.org/10.1016/j.oregeorev.2022.104955.