



Original Paper

# Data Mining of a Geoscience Database Containing Key Features of Gold Deposits and Occurrences in Southwestern Uganda: A Pilot Study

Tsehaie Woldai<sup>1,3</sup> and Andrea G. Fabbri<sup>2</sup>

Received 30 September 2021; accepted 16 April 2022  
Published online: 25 May 2022

Data mining is a promising new tool in mineral exploration. Here, we combined data-mining procedures with spatial prediction modeling for gold exploration targeting in the Buhweju area in southwestern Uganda. It was employed in a data-rich context of unavoidably partly redundant and correlated information that offered challenges in extracting significant relationships. Our study utilized a database of co-registered digital maps related to gold mineralization. It comprised Landsat TM, Shuttle Radar Topographic Mission (SRTM), and geophysical (radiometric and magnetic) datasets for geological and structural mapping. The locations of 15 orogenic gold deposits and 87 gold occurrences were obtained from the Geological Survey of Uganda database. These were considered direct evidence of the presence of gold mineralization. The geological and geophysical settings at the gold deposit/occurrences locations were based on geological units as host rocks, contacts, and structural elements, together with continuous field values of geophysics, radiometry, and other remotely sensed imagery. A gold exploration targeting proposition ( $T_p$ ) was defined as: “That a point  $p$  within the study area contains a gold deposit given the presence of spatial evidence.” All outstanding combinations of spatial evidence were obtained using empirical likelihood ratios. With a data-mining strategy, the ratios were filtered and modeled to identify stronger spatial associations, to rank the study area according to the likelihood of future discoveries, to represent ranking quality, to estimate associated uncertainty, and to select prospective target areas. The empirical likelihood ratios facilitated a transparent strategy for generating *prediction patterns* and extracting small prospective target areas with higher likelihood of discovery and lower-ranking uncertainty. Conclusions are provided on the knowledge extraction for prospectivity with further data and the challenges of reducing the arbitrariness of decisional steps.

**KEY WORDS:** Data mining, Empirical likelihood ratio, Prediction pattern, Target pattern, Uncertainty pattern, Gold prospectivity, Uganda.

## INTRODUCTION

Geological, geochemical, geophysical, and remote sensing digital data are critical ingredients in mineral exploration targeting for new mineral resources (Chung, 2003). Commonly, within a study area, the context of known mineral deposits of dis-

<sup>1</sup>School of Geosciences, University of Witwatersrand, Johannesburg, South Africa.

<sup>2</sup>DISAT, University of Milano-Bicocca, Milan, Italy.

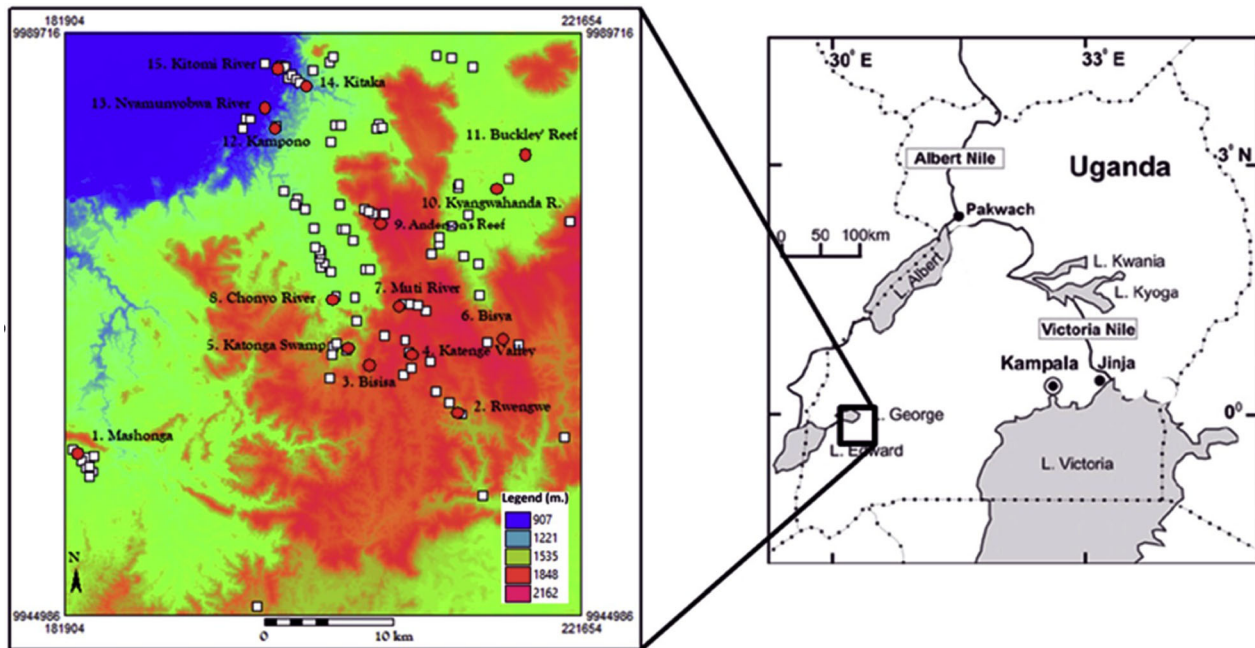
<sup>3</sup>To whom correspondence should be addressed; e-mail: tsehaie.woldai@wits.ac.za

tinct genetic types is characterized by spatial modeling to identify the setting of future occurrences at locations at some distance from the known ones (Chung et al., 2002; Fabbri and Chung, 2008; Agterberg, 2014; Chung and Fabbri, 2020). In a data-driven task, the detail and resolution of the digital data collected for the purpose are assumed sufficient for modeling. When the data are deemed incomplete or unsatisfactory, an expert's knowledge is used to complement the data and the spatial relationships to obtain model- or knowledge-driven maps of the likelihood of future deposit discovery (Harris and Sanborn-Barrie, 2006; Harris et al., 2015; Ford et al., 2019; Montsion et al. 2019). A vision of a future exploration information system (EIS) was proposed by Yousefi et al. (2019). It is based on a mineral-system concept of connecting incompletely known critical mineralization processes to mappable criteria in a workflow. More recently, Yousefi et al. (2021) pointed out the risks of overfitting in data-driven applications or the risk of biases in knowledge-driven applications.

Herbert (2012) and Herbert et al. (2014) integrated regional geoscience data to generate a new geological interpretation of a poorly mapped part of southwestern Uganda. The purpose of the data integration was to model prospectivity for orogenic gold (Groves et al., 1998) using historical geological maps, field observations, digital terrain models, Landsat TM data, and airborne geophysical data. Given the scarcity of geological information and the absence of significant gold deposits, a mineral-system approach (McCuaig et al., 2010) based on expert knowledge was applied to link conceptual mineralization models with data available at a regional-level scale. Evidence was used for the source of gold, active mineralization pathways, and physical traps at a district scale (1:100,000) for a study area of roughly 46,875 km<sup>2</sup>. Their approach implied a workflow similar to the one envisaged by Yousefi et al. (2019, 2021). A total of 15 gold deposits and 87 gold occurrences were recorded in the study area. A knowledge-driven multi-class index overlay method was employed to model prospectivity, and eight sub-areas covering 2,500 km<sup>2</sup> were identified and recommended for follow-up exploration (Herbert et al., 2014). However, using expert opinion is potentially biased toward the expert's view, and so it becomes hard to separate it from the data-driven numerical support provided by data.

Likewise, Bahiru and Woldai (2016) integrated geologic mapping and updated the geology of a much smaller study sub-area (1,801 km<sup>2</sup>) in southwestern Uganda that contains the gold deposits and occurrences previously studied by Herbert et al. (2014). They tested the relationships between lithology and structures in the sub-area using Landsat TM, shuttle radar topographic mission (SRTM), and potential field datasets (radiometry and magnetics), which were subsequently validated directly in the field, generating a richer and more detailed database. They obtained better lithologic differentiation, structural data, and more detailed relationships with gold deposits and occurrences. The locations of the sub-area, the gold deposits, and the occurrences are shown in Figure 1.

Our project's goal was to use data-driven modeling in a data-rich and much smaller sub-area in southwestern Uganda. For this, we used a large part of the database created by Bahiru (2011), Herbert (2012), Herbert et al. (2014), and Bahiru and Woldai (2016), converted into co-registered digital images related to geology, structure, and gold mineralization in the Buhweju area (see Figs. 1, 2 and 3). Having at our disposal a data-rich database focused mainly on lithology mapping, however, represented a challenge for navigating through many different numerical relationships and extracting the most relevant and interpretable ones. Therefore, our data-mining exercise was meant to test the relevance of the various digital maps of the database to identify the ones more closely related to the presence of orogenic gold deposits and gold occurrences using empirical likelihood ratios (ELRs). After describing the geology and metallogeny of the Buhweju area and its database, we introduce a mathematical model of spatial relationships. The modeling is to identify and measure the numerical support of relationships with the presence of gold occurrences and deposits and integrate them into digital images. These are termed *prediction patterns* and express the likelihood of future deposit discovery. Next, a strategy of data-mining experiments is proposed within a workflow leading to the identification of the most supportive data and some of the possible *prediction patterns* of undiscovered gold deposits. We discuss the data-mining experiments for applying that strategy. We finally develop an example of using top ranks of *prediction patterns* in prospectivity mapping.



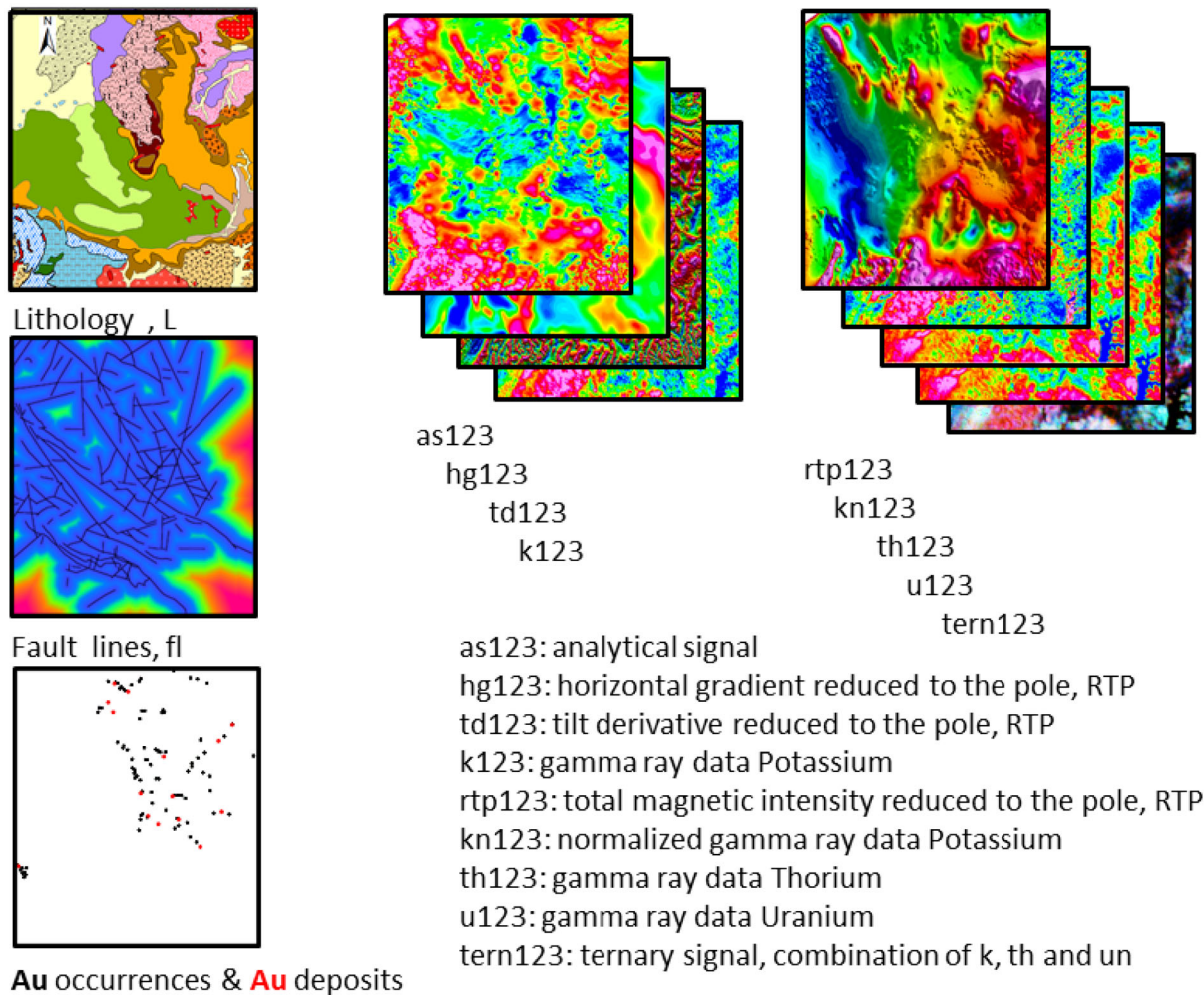
**Figure 1.** Buhweju study area and location in southwestern Uganda. Gold deposits are shown as 15 red dots and gold occurrences as 87 white squares, with a digital elevation model as background. The presence of gold deposits and occurrences is represented as digital images in Figure 2 to match the digital database.

### THE BUHWEJU STUDY AREA: REGIONAL AND LOCAL GEOLOGY, STRUCTURE, MINERALIZATION, AND DIGITAL DATABASE

Three major Proterozoic belts underlie central and west Uganda: (1) Paleo-Proterozoic Buganda-Toro metasediments (Westerhof et al., 2014); (2) Meso-Proterozoic Karagwe-Ankolean (Kibaran) series in the southwest of the country consisting of metamorphosed rocks varying from slightly cleaved mudstone, shales to mica schists; and (3) Neo-Proterozoic sediments (Gabert, 1990). Paleozoic, Mesozoic, and Cenozoic rocks related to the rift activities cover the country's eastern and western borders, whereas Tertiary and recent sediments fill the down-faulted western rift.

The area covered comprises various lithological units, ranging from Archaean to Paleo-Proterozoic metamorphic complexes. The rock units (Fig. 3) are represented by tonalite-trondhjemite-granodiorite rocks (or TTG) in the northern part and plutonic rocks and dykes belonging to the 'Toro Complex', assigned to the Rukungiri Suite (Westerhof et al., 2014) in the southern part. The TTG comprises

medium-grained, equigranular-to-porphyroblastic rocks with variable granitoid compositions. Unconformably overlying the Rukungiri Suite is the Buganda Group, representing a sedimentary sequence comprising mica schists with quartzite and conglomerate interbeds metamorphosed and migmatized. Its sedimentary features reflect shallow-water deposition with argillites, shales, and sandstones in a smooth succession. The thinner sandstones and quartzites are lenticular. Unconformably overlying the Buganda Group is the Buhweju Group, consisting of psammitic and pelitic rocks. The latter group is a sedimentary sequence comprising parashist with occasional quartzite that has been metamorphosed and migmatized. The Buhweju Group comprises six formations: Isingiro, Lubare, Nsika, Kasyoha and Munyoni Formations, in the order of succession from bottom to top (Fig. 3). Quartz veins and doleritic dykes represent intrusive rocks older than Pleistocene but younger than the Buhweju Group (Combe, 1934; Reece, 1959, 1961; Bahiru, 2011). The beddings, cleavages, and linear structures mapped on various lithological units of the Buhweju Group led to an interpretation that it was folded with WNW-trending axes and the intensity of fold-



**Figure 2.** Components of the Buhweju area database: digital images used in the analyses to establish spatial relationships with the presence of gold occurrences or deposits. The numbers 1, 2, and 3 next to short names indicate the respective 3 Landsat TM bands, here combined and displayed with pseudo-colors (value for each band 0–255). Only lithology is categorical (see Fig. 3), and all other images are continuous fields. The fault lines in black were converted to distance functions ranging from 0 to 9648 m. All images consist of arrays (rasters) of pixels of  $30 \times 30$  m resolution. The raster dimension was 1327 pixels  $\times$  1508 lines. The short names were used to list the individual bands input to each analysis. Occurrences and deposits were two separate images here combined into one and used as direct evidence of mineralization, while the remaining images were treated as indirect evidence expressing the local settings of gold mineralization.

ing increasing to the south. Pleistocene rocks occupying the rift valley floor are poorly consolidated and comprise fine-grained sand, silt, clay, and highly calcareous volcanic tuffs and conglomerates with occasional lavas (Combe, 1934; Reece, 1959, 1961).

Our investigation showed that most gold occurrences were restricted to specific orientations and host rocks (Bahiru, 2011; Bahiru & Woldai, 2016). Density analysis of surface and subsurface lineaments also reveal that the Buhweju Group rocks and the undifferentiated schist and amphibolites are highly impacted by surface lineaments,

whereas the basement rocks are affected mainly by subsurface lineaments. Two large gold occurrence clusters were identified in the area. A low magnetic anomaly and high lineament density characterize the gold cluster at Kitaka, while the gold occurrence at the Katonga mine is hosted by mudstone and Lubare quartzite and underlain by a high magnetic anomaly (Bahiru, 2011; Bahiru & Woldai, 2016). The NNW- and NNE-trending faults spatially control these gold clusters, and their intersection determines the localization of most Buhweju gold occurrences. The two structural trends in the study area could



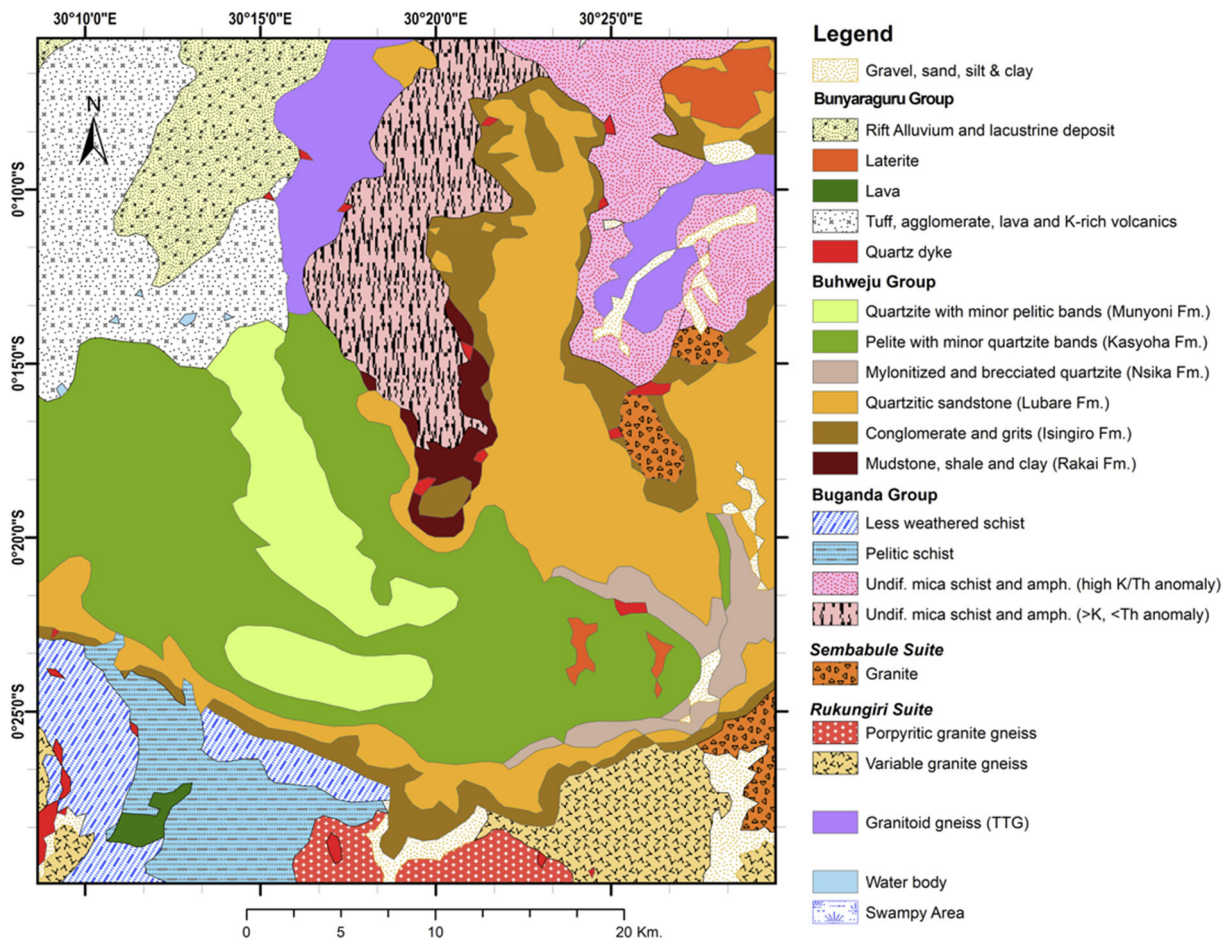


Figure 3. Geological map of the Buhweju area (modified after Bahiru (2011), Lehto et al. (2014a), Bahiru and Woldai (2016).

potentially be viewed in terms of the NNW-trending Geita (in Tanzania) and the NNE-trending Twangiza (in the Democratic Republic of Congo) controls on gold mineralization (Magnus, 2003; Banro Corporation, 2006).

Data concerning known mineral occurrences were obtained from the database, maps, and publications of the Department of Geological Survey and Mines, DGSM, in Entebbe, Uganda. With this database in hand, Bahiru (2011), Bahiru and Woldai (2016), and Herbert (2012) extracted 102 known gold occurrences in the study area. Very few scientific and/or company reports are available for all these occurrences. Therefore, in subsequent fieldwork, they observed and recorded all interesting indications of various mineral commodities, including structural elements and potential rock types hosting mineralization. As shown in Figure 4, most gold occurrences are located north of the NW-SE-

striking fault A-B. Specifically for the current work, 15 sites were assigned as deposits and 87 as occurrences, as outlined in the DGSM database. The term deposit in this paper refers to an established orogenic gold deposit where the host mineral is known and has been or is currently in production. The deposits are represented as red stars in Figure 4.

Much of the present land surface of Uganda has suffered intense weathering over geological time. These processes have led to some materials being concentrated in ores, most notably gold, tin, pyrochlore (niobium, tantalum, and rare earth), and apatite (African Mining, 2000; Lehto et al., 2014a). For example, in the Buhweju study area, fresh bedrock is only apparent on the tops of hills and along streams and deep valleys. Much of the so-called alluvials or gold occurrences referred in the current work are soft, mica- and kaolinite-rich material and laterite or formations molded by

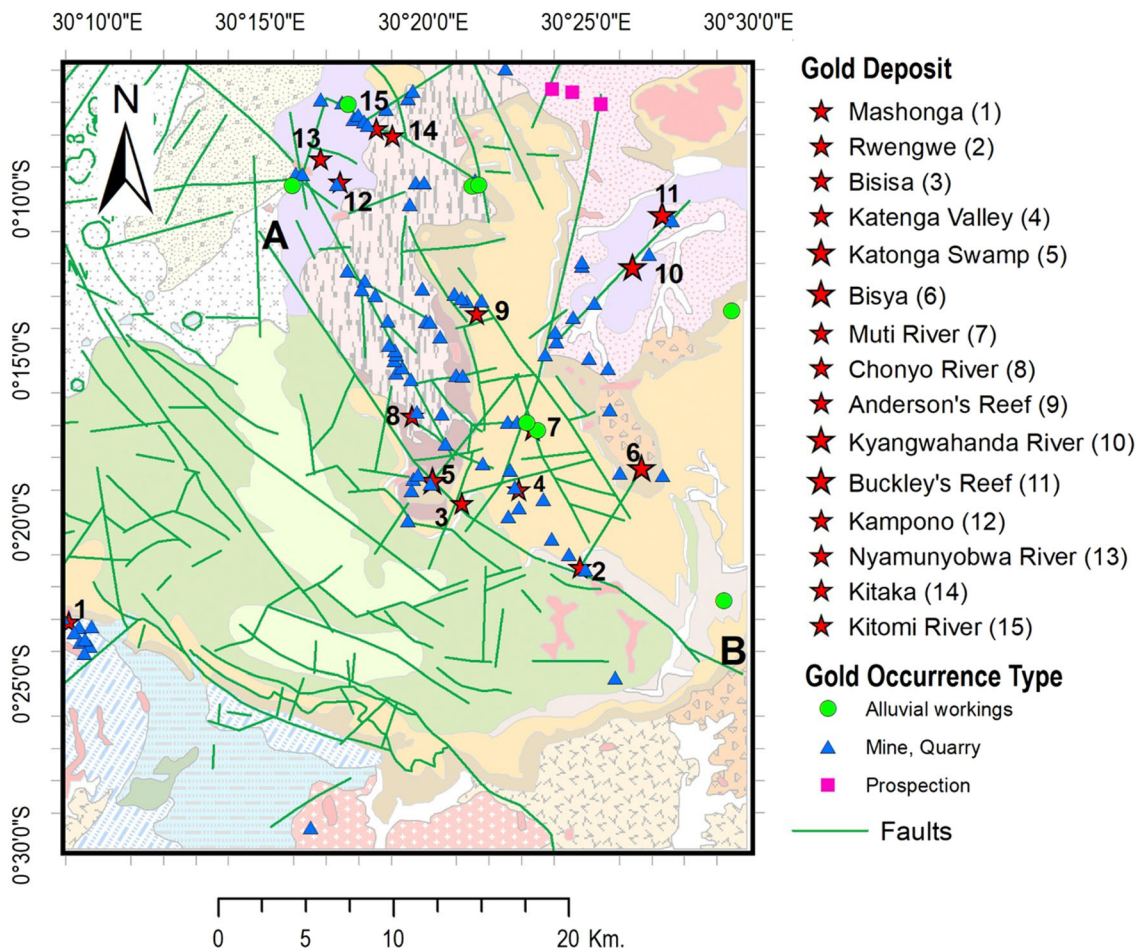
weathering of vast volumes of the underlying Archaean greenstones that can reach depths of 15–20 m and extends over tens of square kilometers (Wayland, 1936; The Republic of Uganda, 2006; Lehto and Kato, 2014). The 87 occurrences correspond mostly to current or historically exploited small, high-grade alluvial gold findings (Wayland, 1936; Herbert 2012; Lehto et al., 2014a, b) and in the database of the DGSM are evident in the form of shallow underground adits or quarries (88.5%), alluvial diggings (8%) and prospection (3.5%). On the Buhweju plateau, gold-bearing solutions were mostly diffuse, resulting in general impregnations rather than definite veins formation. According to Combe (1939), the absence of veins carrying economic quantities and gold concentrations indicates that the alluvial deposits were significantly clustered due to weathering. Gold production has been by small producers, including licensed and illegal miners or artisans. However, production statistics from these mines are only indicative, given that most operators are not licensed, and even the licensed ones tend to under-declare; hence, most of the gold is transacted through dubious channels (Nagudi, 2011).

Information on the gold deposits in the Buhweju–Mashonga study area was derived from the work of various investigators: Barnes (1961), Reece (1961), Pekkala et al. (1995), Bahiru (2011), Herbert (2012), Lehto (2014a, b), and the final Geological map of Uganda (GTK, 2012). The 15 gold deposits were excavated in-depth so that their locations correspond to gold mines for which the host rocks are known to represent orogenic mineralization. These include (1) Mashonga, (2) Rwengwe, (3) Bisisa, (4) Katenge Valley, (5) Katonga Swamp, (6) Bisya, (7) Muti River, (8) Chonyo River, (9) Anderson’s Reef, (10) Kyangwahanda River, (11) Buckley’s Reef, (12) Kampono, (13) Nyamunyobwa River, (14) Kitaka, and (15) Kitomi River. Limited scientific and company reports are available for all these localities. Even the database of the Department of Geological Survey and Mines, DGSM, in Entebbe, Uganda, from which the source for the 15 gold deposits was extracted, does not shed light on host rocks, mineralogy, structural controls, size, and grade. However, recent field exploration by Bahiru (2011) and Bahiru & Woldai (2016), and lithological interpretations reveal the Lubare quartzitic sandstone and mudstone, shale, and clay of the Rakai Formation, as the dominant country rocks hosting gold deposits in the Buhweju plateau. Gold workings below the plateau

are hosted by different bedrocks, such as schist, gneiss, granites, and amphibolites (Bahiru, 2011). For example, in the surveyed mine areas, Kitaka and Kampona Pb–Zn mines in quartz veins have gold disseminated throughout the host rock (Nyakecho and Hagemann, 2014; Bahiru and Woldai, 2016). Recent exploration in the structurally controlled intrusion-hosted mesozonal Mashonga gold district, SW of the study area, shows much of the gold extracted by artisans derived from lateritic gravel and conglomerate lying on kaolinized bedrock (Fig. 5a). The thickness of the weathered profile ranges from 12 to 20 m (Herbert, 2012). Quartz veins do exist, and some are being reworked (Taylor, 2007).

According to Wayland (1934), reef gold occurs in the structurally controlled sandstone-hosted and Meso-zonal Muti mine (Fig. 5b), forming stockwork of quartz stringers transecting a quartzite bore of pyrite and fine gold. Wayland (1934) also indicated gold occurrences in the sandy transition of the quartzite and pelites. Three sulfide veins have been identified (Reece, 1961) at Kitaka and Kampono. For example, in the Kitaka mine, galena, chalcopyrite, gold, and Fe-sulfide-bearing quartz veins are quarried in a medium-grained meta-dolerite, emplaced in schist and gneiss. In these mines, coarse crystalline epithermal gold occurs in vugs lined by quartz crystals (Reece, 1961). In addition to the observations described above, specific indications may be considered “favorable”. The first being provided by granite-quartzite contacts, which are not always visible; they are also more common in higher than lower valleys. The second is a concentrate obtained by panning (from the bottom of the gravels and higher parts); this is rich in pegmatite minerals, particularly zircon, monazite, and frequently rutile are conspicuous (Wayland, 1936), and a small percentage of tiny cassiterite crystals, sometimes of relatively long, bi-pyramidal form, is generally present.

Table 1 summarizes the gold deposits in the Buhweju–Mashonga area derived from the work of various investigators: Barnes (1961), Reece (1961), Pekkala et al. (1995), Bahiru (2011), Herbert (2012), Lehto et al. (2014a), Magnus (2003), and the final Geological map of Uganda. The spatial distribution of the 87 gold occurrences identifies just the mapped rock units at those locations. The source of mineralization is not well known. Therefore, to some extent, their distribution around the 15 gold deposits reflects the intensity of exploration. For this reason, modeling of undiscovered gold occurrence will be



**Figure 4.** Map showing known gold deposits and gold occurrences in the Buhweju–Mashonga area. The 87 gold occurrences consist of shallow underground adits or quarries (blue triangles), alluvial workings (green circles), and prospectings (purple squares). The geology of the area and the structure are in the background. Lines A–B indicate the NW–SE-striking fault separating the gold mineralized area to the north.

used as a descriptor of gold occurrences of presently uncertifiable origin but representing geologic settings similar to those of the gold deposits. The corresponding *prediction patterns* and their characteristics were initially suspected of being of potential support in interpreting the *prediction patterns* of the 15 gold deposits. They have been used in the analyses under the assumption of potential support.

The digital database containing the 15 gold deposits and the 87 gold occurrences is portrayed in Figs. 1, 2, and 4. It was digitized at spatial resolution of 30 m. Occurrences and deposit locations were represented as 3 × 3 pixel neighborhoods within a study area of raster 1327 × 1508 = 2,001,116 pixels, i.e., ~ 1801 km<sup>2</sup>. They were considered direct evidence of the presence of gold mineralization. The

geologic settings of their neighborhoods in the study area were represented by geologic units, lithologic contacts, structural elements, and continuous field values of geophysics, radiometry, and other remotely sensed imagery, all digitized at the same resolution of 30 m (Fig. 2). All digital images of direct and indirect evidence of gold mineralization have the same spatial resolution, raster dimension, and pixel-to-pixel correspondence so that the spatial relationships are established and quantified.

**THE EMPIRICAL LIKELIHOOD RATIO FUNCTION**

The ELR is one of the well-known mathematical tools used together with other Bayesian models





**Figure 5.** Artisanal gold miners working in the gold mine areas of **a** Mashonga and **b** Muti. At Mashonga, gold is in a steep NE–SW-trending shear zone composed of soft schistose material and kaolinized granite veins with only minor quartz veins. The soft mica-rich rock in the shear zone has resulted from strong deformation, hydrothermal alteration, and weathering into clay-rich material (Lehto et al., 2014b). The panning sites are located downstream from an old gold reef composed of a quartz stockwork in a light-colored sericite quartzite (or sandstone) matrix at Muti River workings. The river cuts this reef, which is most likely the gold source.

for spatial prediction modeling in mineral exploration and natural hazard assessment (Bonham-Carter, 1994; Chung et al., 2002; Chung, 2006; Chung and Fabbri, 2020). It is based on joint conditional probability to establish spatial relationships. For instance, in gold exploration, the likelihood ratio is the ratio of the normalized frequency of a map unit or value in the presence of gold deposits to the normalized frequency in their absence within the study area. A corresponding ratio is computed for the presence and absence of gold occurrences. The function is used to classify a study area into classes of similarity with areas with the presence of mineralization.

Chung and Fabbri (1993) developed a mathematical framework for various spatial prediction models, termed “favorability function.” As discussed by those authors and Chung (2003), several representations within well-established mathematical frameworks can be used as a favorability function: the conditional probability function, the likelihood ratio (also the certainty factor function or the weights of evidence function as special cases of the likelihood ratio function), the Dempster–Shafer belief functions, and the fuzzy set membership functions (Chung and Fabbri, 1993, 2001). We used only one representation in this contribution, namely the ELR, to express spatial relationships and the favorability function. The reason for using it is the easiness of interpretation of likelihood ratios and the

combination rules of the function (Chung and Fabbri, 1998). Chung (2006) discussed the function and its assumptions in detail in the context of landslide hazards. Here, we summarize the function and its use in mineral exploration.

A fundamental assumption in spatial prediction modeling is that what is modeled is the occurrence of future events given the known characteristics of past events through time and space. In practice, this is done for a study area, of given dimensions, which is hopefully large enough to represent those characteristics numerically. Over a study area, besides the distribution of known events in point-form or of relatively small sizes, the distribution of corresponding categorical map units or continuous fields is used to capture the characteristics of the events’ typical conditions or settings. Past and future events (known and expected, respectively) are assumed to be relatively rare in a study area and to belong to a distinct, recognizable genetic type.

For instance, in mineral exploration, known events can be occurrences of gold mineralization or gold mines whose spatial distribution is related to the geological, geochemical, and geophysical maps and images, including those from various remote sensors. Their units or their value ranges are thought by experts to characterize or support the distribution of the location of mineralization events. It means that two types of spatial distributions are established and compared: that of the events and those of the



**Table 1.** Attribute data of the 15 gold deposits used in the study area

FID	E_WGS 84UTM	N_WGS 84UTM	Deposit_Na	Type	Comment
1	182,934	– 42,564	Mashonga	Au	Quartz vein stockwork in muscovite schists of the Igara Formation adjacent to granitic intrusion. Proximal to thrust contact of Meso-Proterozoic conglomerate (Nyakecho & Hagemann, 2014)
2	212,081	– 39,452	Rwengwe	Au	Gold deposit associated with the gently west-dipping mylonitized and brecciated quartzite
3	205,331	– 35,802	Bisisa	Au	The host rocks of the gold are dominantly sandstone belonging to the Lubare Quartzitic sandstone Formation
4	208,581	– 35,002	Katenga Valley	Au	Katenga Valley gold is found a few kilometers SW of Muti River. According to Hooper (1964), the hydrothermal alteration was confirmed by the Ugandan Mines Department records. The host rock comprises Quartzitic sandstone belonging to the Lubare Formation
5	203,681	– 34,502	Katonga Swamp	Au	Old mine striking N20W; Swampy area. The gold occurrences at Katonga Swamp are hosted by Lubare quartzitic sandstone and mudstone of Rakai Fm. It is underlain by relatively high magnetic material (Bahiru, 2011; Bahiru & Woldai, 2016)
6	215,581	– 33,802	Bisya	Au	Weathered granite on top. The host rock comprises conglomerate and grit of the Isingiro Formation
7	209,384	– 31,588	Muti River	Au	According to Nyakecho and Hagemann (2014), the gold in the Muti River is hosted by poorly sorted, fine- to medium-grained quartzitic sandstone conglomerate and grit overlying granite exposed 4 km East of the Muti River working
8	202,481	– 30,802	Chonyo River	Au	Alluvial gold was found in the downstream and valleys of the Chonyo River. A large part of the weathered sediments observed are by-products of the Mica schist and amphibolite higher in the Buhweju plateau
9	206,181	– 24,952	Anderson's Reef	Au, Sn, Ta	Quartz vein stockwork in Paleo-Proterozoic schists, conglomerate and grit of the Isingiro Formation. Cassiterite, tantalite, and tin are found associated with gold. The mineralization covers a wide area
10	215,081	– 22,302	Kyangwahanda River	Au	According to Bahiru (2011), the rock type associated with this area's gold deposit is granite gneiss
11	216,769	– 19,324	Buckley's Reef	Au	Several small quartzite reefs and stringers penetrated Quartz vein stockwork in Paleo-Proterozoic mica schists with psammite and gneiss interbeds derived from sandstone
12	198,388	– 17,424	Kampono	Au, Pb, Cu	Gold occurs in quartz veins associated with Pb and Cu
13	197,281	– 16,102	Nyamunio-bwa River	Au	Unknown source of gold, but it is found in the area covered by granitoid gneiss (TTG)
14	200,474	– 14,393	Kitaka	Au, Ag, Pb, Au	Coarse crystalline gold occurs in vugs lined by quartz crystals and was the last (epithermal) phase of complicated mesothermal sulfide mineralization. Gold is hosted by meta-granodiorite and meta-dolerite. It is associated with a predominantly Pb–Zn–Ag deposit
15	201,370	– 14,800	Kitomi River	Au	Coarse crystalline gold occurs in vugs lined by quartz crystals in sulfide veins (Nyakecho and Hagemann, 2014)

supporting map units and continuous value ranges. The basic idea is that one should know enough past events in a study area to extract their distinct characteristics. It is expected within a study area to have future events (undiscovered gold occurrences or deposits) of the same type and setting and at some distance from the known ones.

The ELR function model is then applied to compare and integrate the normalized frequencies of categorical map units and the density functions of

continuous fields in the presence of the events with those in their absence within the remainder of a study area. To do this, a proposition is constructed so that a mathematical expression can be supported to be true using the data available in a study area. In our case, the events are the set of locations of gold deposits or gold occurrences (i.e., their spatial distribution in the study area), which we term direct supporting pattern (DSP) of the proposition. An example of the proposition ( $T_p$ ) is: “That a point  $p$  in

*the study area contains gold occurrences | the presence of additional spatial evidence*". Here  $T_p$  stands for true,  $p$  refers to a point or pixel, and the symbol "|" indicates "given." We termed the spatial evidence, additional to the presence of the occurrences, indirect supporting patterns (ISPs) of the proposition. Such spatial evidence is the presence at the location of the occurrences of categorical map units and continuous field values that indicate, are related to, and support the presence of the gold occurrences (deposits, mines) whose location they share.

The computation of ELRs is simplified by having a database for the study area in which both the DSP and the ISPs are digitized at the same spatial resolution and are in point-to-point correspondence (although this is not a restriction). Figure 2 shows the digital images used to model future gold discoveries in the Buhweju study area. Lithology (**L**) is a categorical image with 21 units. Fault line (**fl**) is a continuous distance image of values ranging from 0 to 9648 m, and the other continuous field images, from **as1** to **tern3**, have values ranging from 0 to 255. The remaining two images are locations of the 15 gold deposits and the 87 gold occurrences represented as index numbers from 1 to 15 or 1 to 87, respectively.

Tables 2 and 3 show the likelihood ratio values for the individual lithology units in correspondence of the 15 gold deposit pixels and the 87 gold occurrence pixels, respectively. Table 4 shows the likelihood ratio values for the 15 gold deposits for all the different ISPs in the study area analyzed. We used the short name **Au15dp** to indicate the image of locations of the 15 gold deposits within the study area. For the modeling analyses, they represent a DSP. Similarly, the short name **Au87oc** indicates the image of locations of the 87 gold occurrences as another DSP. Table 5 shows the ratio values for the **Au87oc**.

Examples of normalized frequencies, density functions, and ELR functions for gold deposits are shown in Figure 7. Besides empirically generating the ratios DSP/ISPs for each point of the study area, the ELR is also used as a modeling function in several conditional independence assumptions to combine, by its rules, the ratios into integrated values (Chung, 2006). ELR values can range from 0 to infinity. For instance, an ELR value of 2 indicates a frequency in the presence of occurrences that is twice the one in their absence in the remainder of the study area, and thus, it supports the proposition. A value close to 1 indicates the same frequency in

the presence and the absence of occurrences, therefore not supporting the proposition (i.e., no contrast between the two conditions). Integrated ELR values, also ranging from 0 to  $\infty$ , are then used to separate areas likely to have deposits from areas not known to have them. Even if we expect more deposits and occurrences in the study area, being relatively rare, they are considered not to affect the remainder of the study area numerically.

As Chung (2006) discussed, the modeled absolute values of integrated ELRs are not interpretable beyond their ranks being simply relative measures of dissimilarity. For this reason, it is recommended to convert the ordered sequence of ELR values (from maximum to minimum) into a conveniently small number of equal area ranks, such as 200 ranks, corresponding each to 0.5% of the study area. The 200 ranks can then be visualized as *prediction patterns* by grouping them into more comprehensive wider ranges for lower ranks and narrower ranges for higher ranks, i.e., of gradually greater relative support.

Obtaining *prediction patterns*, however, is only the initial step of the process of prediction modeling. The generation of a *prediction pattern* is based on the assumption that the characteristics of undiscovered gold deposits or occurrences are the same or are similar to the known ones in the study area. Several strategies can then be thought of to assess the predictive quality and stability of the *prediction patterns*, including their uncertainty. The ELR function model of spatial prediction can also be used as a tool for data mining to extract from the study area's database the most supporting ISPs that generate geologically interpretable *prediction patterns*. Recent works by Chung and Fabbri (2020) discussed unresolved research issues in prediction modeling for mineral occurrence target mapping. As demonstrated earlier by Chung and Fabbri (1998), the results of applying the three functions termed weights-of-evidence (WoE), certainty factor (CF), and ELR provide an identical equal-area ranking of a study area when using the same input data.

The modeling applied established spatial relationships between pixels representing the deposit or occurrence locations (in our case,  $3 \times 3$  pixel neighborhoods) and the corresponding locations of the categorical map (lithology, **L**) and continuous field values (distance from fault lines, **fl**, and the 28 bands from **as1** to **tern3**) listed in Figure 2. The tentative numerical strategy used for the Buhweju area is discussed next.

**Table 2.** Lithology units, their number of pixels, normalized frequencies within the **Au15dp** ( $15 \times 3 \times 3 = 135$  pixels), within the remainder of the area (2,000,981 pixels), and the corresponding ELRs, in the Buhweju study area

ID	Lithology unit	Pixels	Freq-occurrence	Freq-non-occurrence	ELR
<b>L<sub>1</sub></b>	<b>Gravel, sand, silt and clay</b>	56,369	0.0667	0.0282	<b>2.3669</b>
L <sub>2</sub>	Rift Alluvium & lacustrine deposit	89,695	0.0000	0.0448	0.0000
L <sub>3</sub>	Tuff, agglomerate, lava and K-rich Volcanics	162,310	0.0000	0.0811	0.0000
L <sub>4</sub>	Laterite	23,295	0.0000	0.0116	0.0000
L <sub>5</sub>	Lava	6482	0.0000	0.0032	0.0000
L <sub>6</sub>	Quartzite with minor pelitic bands	115,971	0.0000	0.0580	0.0000
L <sub>7</sub>	Pelite with minor quartzite bands	391,850	0.0000	0.1959	0.0000
<b>L<sub>8</sub></b>	<b>Mylonitized and brecciated quartzite</b>	40,046	0.0667	0.0200	<b>3.3319</b>
L <sub>9</sub>	Quartzitic sandstone	310,854	0.2667	0.1553	1.7167
L <sub>10</sub>	Conglomerate & grits	121,680	0.0667	0.0608	1.0964
L <sub>11</sub>	Mudstone, shale, clay	22,661	0.0000	0.0113	0.0000
L <sub>12</sub>	Less weathered schist	72,822	0.0593	0.0364	1.6284
L <sub>13</sub>	Pelitic schist	83,824	0.0000	0.0419	0.0000
L <sub>14</sub>	Undif. mica schist and amph. (high K/Th anomaly)	108,933	0.0667	0.0544	1.2247
<b>L<sub>15</sub></b>	<b>Undif. mica schist and amph. (&gt; K, &lt; Th anomaly)</b>	130,370	0.1333	0.0651	<b>2.0467</b>
L <sub>16</sub>	Quartz dyke	23,621	0.0074	0.0118	0.6275
<b>L<sub>17</sub></b>	<b>Granitoid gneiss (TTG)</b>	86,914	0.2000	0.0434	<b>4.6059</b>
<b>L<sub>18</sub></b>	<b>Granite (1987 A <math>\pm</math> 5 MA)</b>	15,912	0.0667	0.0080	<b>8.3883</b>
L <sub>19</sub>	Porphyritic granite gneiss	98,974	0.0000	0.0495	0.0000
L <sub>20</sub>	Variable granite Gneiss	36,241	0.0000	0.0181	0.0000
L <sub>21</sub>	Water body	2292	0.0000	0.0012	0.0000

In **bold** are ELRs  $\geq 2$  and their corresponding lithology units

**Table 3.** Lithology units, their number of pixels, normalized frequencies within the **Au87oc** ( $87 \times 3 \times 3 = 783$  pixels), within the remainder of the area (2,000,333 pixels), and the corresponding ELRs, in the Buhweju study area

ID	Lithology unit	Pixels	Freq-occurrence	Freq-non-occurrence	ELR
L <sub>1</sub>	Gravel, sand, silt, and clay	56,369	0.0230	0.0282	0.8160
L <sub>2</sub>	Rift Alluvium and lacustrine deposit	89,695	0.0038	0.0448	0.0854
L <sub>3</sub>	Tuff, agglomerate, lava, and K-rich Volcanics	162,310	0.0038	0.0811	0.0472
L <sub>4</sub>	Laterite	23,295	0.0000	0.0116	0.0000
L <sub>5</sub>	Lava	6482	0.0000	0.0032	0.0000
L <sub>6</sub>	Quartzite with minor pelitic bands	115,971	0.0000	0.0580	0.0000
L <sub>7</sub>	Pelite with minor quartzite bands	391,850	0.0000	0.1959	0.0000
L <sub>8</sub>	Mylonitized and brecciated quartzite	40,046	0.0000	0.0200	0.0000
L <sub>9</sub>	Quartzitic sandstone	310,854	0.1483	0.1553	0.9537
L <sub>10</sub>	Conglomerate and grits	121,680	0.1162	0.0608	1.9120
<b>L<sub>11</sub></b>	<b>Mudstone, shale, clay</b>	22,661	0.0575	0.0113	<b>5.0832</b>
<b>L<sub>12</sub></b>	<b>Less weathered schist</b>	72,822	0.0920	0.0364	<b>2.5284</b>
L <sub>13</sub>	Pelitic schist	83,824	0.0000	0.0419	0.0000
<b>L<sub>14</sub></b>	<b>Undif. mica schist and amphibolite (high K, Th anomaly)</b>	108,933	0.1379	0.0544	<b>2.5353</b>
<b>L<sub>15</sub></b>	<b>Undif. mica schist and amphibolite (&gt; K, &lt; Th anomaly)</b>	130,370	0.2644	0.0651	<b>4.0628</b>
L <sub>16</sub>	Quartz dyke	23,621	0.0230	0.0118	1.9483
<b>L<sub>17</sub></b>	<b>Granitoid gneiss (TTG)</b>	86,914	0.1188	0.0434	<b>2.7365</b>
<b>L<sub>18</sub></b>	<b>Granite (1987 A <math>\pm</math> 5 MA)</b>	15,912	0.0115	0.0080	1.4458
L <sub>19</sub>	Porphyritic granite gneiss	98,974	0.0000	0.0495	0.0000
L <sub>20</sub>	Variable granite Gneiss	36,241	0.0000	0.0181	0.0000
L <sub>21</sub>	Water body	2292	0.0000	0.0012	0.0000

In **bold** are ELRs  $\geq 2$  and their corresponding lithology units



**Table 4.** Main ELR values for predictions with the distribution of the 15 gold deposits, **Au15dp** as DSP and as **ISPs**, the complete set listed in Figure 2: **L\_flas123hgm123td123k123rtp123kn123th123u123tern123**

<b>Au15dp</b> as DSP with ISPs	ELR values
Lithology (21)	$L_1$ <b>2.37</b> , $L_8$ <b>3.33</b> , $L_9$ 1.72, $L_{10}$ 1.10, $L_{11}$ 0.00, $L_{12}$ 1.63, $L_{14}$ 1.22, $L_{15}$ <b>2.05</b> , $L_{16}$ 0.63, $L_{17}$ <b>4.61</b> , $L_{18}$ <b>8.39</b> ;
Fault lines, <b>fl</b> distance (0–11064 m)	<b>fl</b> $\geq 2$ 0.00–260.00 (0.00 max <b>2.46</b> );
Analytical signal (0–255, bands 1, 2, 3)	$as1 \geq 1$ , 0.00–24.48 (0.00 max 1.66), 90.78–157.34 (134.38 max 1.99), 176.71–209.36 (195.84 max 1.44); <b>as2</b> $\geq 2$ , 35.95–63.24 (53.30 max <b>2.48</b> ); $as3 \geq 1$ , 0.00–14.53 (6.63 max 1.02), 148.15–175.70 (161.16 max 1.78), 201.20–250.67 (232.56 max 1.66);
Horizontal gradient (0–255, bands 1, 2, 3)	<b>hg1</b> $\geq 2$ , 85.17–107.86 (93.58 max <b>3.09</b> ), 168.05–224.91 (211.14 max <b>2.36</b> ); <b>hg2</b> $\geq 2$ , 97.15–118.32 (105.31 max <b>3.13</b> ); <b>hg3</b> $\geq 2$ , 58.65–87.46 (79.30 max <b>4.37</b> );
Tilt derivative from reduced to the pole, RTP magnetics (0–255, bands 1,2,3)	$td1 \geq 1$ , 0.00–28.56 (0.00 max 1.40), 59.16–163.71 (121.63 max 1.64); <b>td2</b> $\geq 2$ , 213.44–234.35 (226.19 max <b>2.21</b> ); <b>td3</b> $\geq 2$ , 83.13–106.59 (94.09 max <b>2.97</b> );
Gamma-ray data potassium (0–255, bands 1,2,3)	$k1 \geq 2$ , 125.46–168.56 (145.09 max <b>3.53</b> ); $k2 \geq 1$ , 73.44–89.25 (80.32 max 1.10), 115.00–153.25 (125.71 max 1.26), 162.95–174.68 (169.58 max 1.04), 230.78–255.00 (255.00 max 1.92); $k3 \geq 1$ , 0.00–84.15 (43.61 max 1.31);
Total magnetic intensity RTP (0–255, bands 1,2,3)	<b>rtp1</b> $\geq 2$ 27.54–48.20 (37.43 max <b>4.30</b> ); <b>rtp2</b> $\geq 2$ 181.82–200.94 (192.02 max <b>2.62</b> ); <b>rtp3</b> $\geq 2$ 188.87–27.54 (23.46 max <b>2.07</b> );
Normalized gamma-ray data potassium (0–255, bands 1, 2, 3)	$kn1 \geq 1$ , 0.00–159.38 (90.27 max 1.53), 203.24–226.44 (214.97 max 1.15); $kn2 \geq 1$ , 42.59–66.56 (58.81 max 1.14), 113.98–203.49 (193.55 max 1.24), 225.68–255.00 (234.60 max 1.08); <b>kn3</b> $\geq 2$ , 185.39–202.73 (195.59 max <b>2.14</b> );
Gamma-ray data thorium (0–255, bands 1, 2, 3)	$th1 \geq 1$ , 0.00–162.69 (51.77 max 1.67); $th2 \geq 1$ , 118.06–136.68 (127.75 max 1.21), 210.12–255.00 (255.00 max 1.98); <b>th3</b> $\geq 2$ , 26.26–68.60 (37.74 max <b>3.24</b> ), 133.87–157.08 (146.37 max <b>2.18</b> );
Gamma-ray data uranium (0–255, bands 1, 2, 3)	$u1 \geq 1$ 37.99–88.74 (50.23 max 1.65), 99.45–164.73 (137.95 max 1.68), 234.60–255.00 (255.00 max 1.07); $u2 \geq 1$ 10.20–37.48 (29.32 max 1.61), 69.62–128.26 (105.31 max 1.95), 191.25–225.68 (210.63 max 1.52); $u3 \geq 1$ 0.00–23.46 (0.00 max 1.31), 64.52–124.18 (99.45 max 1.25), 235.37–255.00 (255.00 max 1.13);
Ternary signal comb. of K, Th and U (0–255, bands 1, 2, 3)	<b>tern1</b> $\geq 2$ , 104.80–134.38 (120.10 max <b>2.88</b> ); <b>tern2</b> $\geq 2$ , 25.24–41.31 (38.89 max <b>2.27</b> ); $tern3 \geq 1$ , 0.00–84.66 (71.15 max 1.35), 121.13–167.03 (135.66 max 1.77), 186.41–203.75 (194.31 max 1.14)

The long name is the list of lithology, fault lineament distance and all the bands of continuous field values maps. Abbreviations for categorical ISPs are as in Table 2:  $L_{1-21}$ . Next to the name of continuous ISPs in column 1 is a numeric range within brackets. Values are bold if ELR  $\geq 2.00$ . The corresponding range of classes is in *italics*, with the maximum class and ratio in brackets

## DATA-MINING QUESTIONS AND STRATEGIES

Data mining is the process of discovering patterns in large datasets involving methods at the intersection of machine learning, statistics, and database systems. It is the analytical steps of the so-called knowledge discovery in databases (KDD) process. It has been applied to mineral exploration in areas ranging from mineral prospectivity prediction (e.g., Barnett and Williams, 2006) to geologic terminology interpretation of mining reports (Holden et al., 2019). Examples of challenges addressed by data mining are: sparse, imbalanced labels of mineral occurrences, varied label reliability, and a wide range of data uncertainty (Granek and Haber, 2015); massive size, multiple sources, multiple types, multi-temporality, non-stationarity, and heterogeneity of geological prospecting big data (Zuo,

2020); lack of transparency of “black box” approaches that use deep learning networks for which exploration data are poorly suited (Dasharnais et al., 2017); comparison of data-mining algorithms in terms of prediction and processing efficiency (Chen and Wu, 2019) or the determination of predictive variables based on mineral-deposit models (Wang et al., 2022).

In our case, the Buhweju database resulted from efforts to improve the geologic knowledge of the corresponding study area to assist in mineral exploration eventually. Its focus on the compilation, documentation, and fieldwork verification of Bahiru and Woldai (2016) was to obtain explicitly a detailed geological map by integrating previous maps, direct field observations, and remotely sensed imagery. It means that all spatial data captured provided information strictly related to the presence of different lithologies that could only in part be mapped in the

**Table 5.** Main ELRs for predictions with the distribution of the 87 gold occurrences, **Au87oc** as DSP and as ISPs, the complete set listed in Figure 2: **L\_flas123hg123td123k123rtp123kn123th123u123tern123**

<b>Au87oc</b> as DSP with ISPs	ELR values
Lithology (21)	L <sub>1</sub> 0.82, L <sub>8</sub> 0.00, L <sub>9</sub> 0.95, L <sub>10</sub> 1.91, <b>L<sub>11</sub> 5.08</b> , <b>L<sub>12</sub> 2.53</b> , <b>L<sub>14</sub> 2.54</b> , <b>L<sub>15</sub> 4.06</b> , L <sub>16</sub> 1.95, <b>L<sub>17</sub> 2.74</b> , L <sub>18</sub> 1.44;
Fault lines, fl distance (0–11064 m)	fl ≥ 1, 0.00–1507.90 (0.00 max 1.86);
Analytical signal (0–255, bands 1, 2, 3)	<b>as1</b> ≥ 2, 37.70–57.63 (47.68 max <b>2.90</b> ); as2 ≥ 1, 194.82–251.43 (226.95 max 1.55); as3 ≥ 1, 0.00–29.98 (19.89 max 1.17), 152.49–179.52 (164.98 max 1.24), 211.40–243.78 (230.01 max 1.31);
Horizontal gradient (0–255, bands 1, 2, 3)	<b>hg1</b> ≥ 2, 132.34–159.12 (146.88 max <b>2.19</b> ); hg2 ≥ 1, 65.79–95.11 (77.52 max 1.15), 113.22–144.84(130.81 max 1.39), 170.59–225.17 (206.30 max 1.10), 233.84–255.00 (255.00 max 1.08); hg3 ≥ 1, 0.00–127.75 (100.98 max 1.82);
Tilt derivative from reduced to the pole, RTP magnetics (0–255, bands 1,2,3)	td1 ≥ 1, 0.00–27.79 (0.00 max 1.41), 143.56–178.25 (162.43 max 1.29); td2 ≥ 1, 0.00–21.42 (3.82 max 1.07), 64.77–139.74 (75.99 max 1.15), 164.73–195.84 (181.30 max 1.09); td3 ≥ 1, 57.38–64.26 (60.69 max 1.03), 87.72–142.03 (122.91 max 1.42), 170.79–239.45 (225.42 max 1.42);
Gamma-ray data potassium (0–255, bands 1,2,3)	k1 ≥ 1, 0.00–42.07 (0.00 max 1.49), 90.01–228.23 (127.24 max 1.91); k2 ≥ 1, 170.34–255.00 (203.49 max 1.60); k3 ≥ 1, 0.00–37.99 (29.58 max 1.05), 64.26–84.40 (79.05 max 1.27), 128.26–162.44 (144.58 max 1.68), 232.56–255.00 (254.49 max 1.46);
Total magnetic intensity RTP (0–255, bands 1,2,3)	rtp1 ≥ 1, 27.54–41.56 (34.17 max 1.37), 85.17–229.25 (209.35 max 1.50); <b>rtp2</b> ≥ 2, 189.98–207.83 (200.43 max <b>2.10</b> ); rtp3 ≥ 1, 0.00–32.38 (6.63 max 1.94);
Normalized gamma- ray data potassium (0–255, bands 1, 2, 3)	kn1 ≥ 1, 0.00–118.83 (91.29 max 1.36), 199.41–230.27 (213.69 max 1.34); kn2 ≥ 1, 124.18–166.52 (147.39 max 1.30), 217.26–255.00 (255.00 max 1.21); <b>kn3</b> ≥ 2, 127.24–162.18 (144.58 max <b>2.61</b> );
Gamma-ray data thorium (0–255, bands 1, 2, 3)	th1 ≥ 1, 0.00–108.37 (0.00 max 1.99), 131.83–201.20 (179.52 max 1.33); <b>th2</b> ≥ 2, 73.69–88.48 (80.32 max <b>2.13</b> ); th3 ≥ 1, 24.48–85.42 (78.29 max 1.51), 128.52–162.18 (145.10 max 1.58), 196.35–203.75 (199.92 max 1.07), 230.78–255.00 (255.00 max 1.89);
Gamma-ray data uranium (0–255, bands 1, 2, 3)	u1 ≥ 1, 225.42–255.00 (255.00 max 1.66); <b>u2</b> ≥ 2, 26.01–30.85 (28.56 max <b>2.01</b> ); <b>u3</b> ≥ 2, 92.82–119.08 (106.59 max <b>2.11</b> )
Ternary signal comb. of K, Th and U (0–255, bands 1, 2, 3)	<b>tern1</b> ≥ 2, 0.00–23.71 (9.69 max <b>2.81</b> ); <b>tern2</b> ≥ 2, 0.00–24.22 (0.00 max <b>2.35</b> ); tern3 ≥ 1, 0.00–100.21 (53.55 max 1.86), 159.38–173.15 (166.26 max 1.07)

The long name is the list of lithology, fault lineament distance and all the bands of continuous field values maps. All the abbreviations for categorical ISP lithology units are as in Table 1: L<sub>1-21</sub>. Next to the name of continuous ISPs in column 1 is a numeric range within brackets. Values are bold if ELR ≥ 2.00. The corresponding range of classes is in *italics*, with the maximum class and ratio in brackets

field. Because it contains information on the distribution and characteristics of a few gold deposits and several gold occurrences (15 and 87, respectively) and 29 digital maps of unavoidably partly redundant and correlated data, the database offers the opportunity of applying data-mining strategies using spatial prediction models and fast processing steps. In particular, it allows considering label reliability of mineral occurrences, data uncertainty, multiple sources of information, and transparency of processing.

We have considered just a part of the database to be modeled by a favorability function. Chung and Fabbri (1993) discussed a variety of such functions. In particular, the ELR function (Chung et al., 2002; Chung, 2006) was used here to identify and model strategic sequences of computational steps leading to knowledge discovery, e.g., the suitability of databases for gold mineralization prospecting. In par-

ticular, answers to questions like the following can be considered forms of knowledge discovery from our Buhweju study area.

1. Which maps, here termed ISPs, once converted into ELR images, are more supportive of the proposition so that we can separate them from the redundant or ineffective ones?
2. How “good” are the more supported *prediction patterns*?
3. How different are the gold deposit *prediction patterns* from those of the gold occurrences?
4. Can we use the gold occurrence distribution, as DSP, to predict the location of the known gold deposits and vice versa?
5. Can we partition a DSP, such as the gold deposits, using *prediction pattern* ranks and interpret the subgroups’ distribution differences and characteristics?

The following are strategic steps that have tentatively been selected and followed. They imply a workflow (e.g., Ford et al., 2019, Figure 1) of data compilation and preparation (Step 1), mapping of spatial supports (Steps 2–3), testing the usefulness and uncertainty of prediction maps (Steps 4–11), production of mineral potential maps and assessment of results (Step 12). Figure 6 shows a chart of these strategic steps.

- (1) Use in turns the DSPs (deposits, **Au15dp** or occurrences, **Au87oc**) and all ISPs to compute the ELR values representing separation power between the presence of mineralization and presumed absence in the study area. ELR values for the two types of DSPs are given in Tables 4 and 5. They were obtained using normalized frequencies of the categorical ISP, with 21 lithology units, **L<sub>1-21</sub>**, and the density functions for all 28 continuous field ISPs, from **fl**, **as1**, **as2**, **as3**, ..., to **tern3** (short names listed in Fig. 2). They link the DSPs and ISPs to the *prediction patterns*. Systematically use lithology units and fault lineament distances (**L<sub>fl</sub>**) together with nine sets of three bands for all other continuous ISPs (e.g., **as1**, **as2**, **as3** separately and **as123** together) to generate *prediction patterns* to interpret and compare them visually. This is to familiarize with supportive and not supportive ISPs generating, respectively, higher and lower ELR values and isolate the supportive ones.
- (2) Set an ELR value threshold for supporting ISPs, e.g., “all ISPs with ELR values close to or higher than 2.00” (as a rule of thumb). Discard, because ineffective, all ISPs with lower ELR values and retain all the others.
- (3) Generate *prediction patterns* using as ISPs only those with ratio values above the threshold. They are **L**, **fl**, ..., **tern3**, listed in Tables 4 and 5 in bold fonts. Note that, because of the strength of support shown by **L** and **fl**, if necessary for interpretation, use these two and one or more of the more supporting ones. Repeat also without **L** and **fl** but then show the geologic boundaries overlaid as vectors. The boundaries are a visual help in interpreting noisy patterns.
- (4) Use the iterative cross-validation of sequential exclusion of 1, **Au15m1dp** (**m** stands for minus) and of 5, **Au87m5oc**, to obtain and compare the resulting prediction-rate curves.
- (5) Compare predictions within the DSP **Au15dp**, or **Au87oc**, and between the two DSPs, **Au15dp** and **Au87oc**, by cross-validation. Based on numerical results and interpretation of prediction patterns using **Au15dp** as DSP or **Au87oc**, decide which *prediction patterns* are preferable and significant.
- (6) Reconsider sequentially the *prediction patterns* obtained using one-by-one all accepted ISPs ( $ELR \approx 2$  or  $ELR \geq 2$ ) for visual inspection. Why a visual inspection? This is because, not knowing, initially, how to measure similarities and differences between relative rankings, visual interpretation allows associating *prediction patterns* (fixed relative ranks of equal area groupings) with geologic knowledge and detecting “redundancies, similarities, and deterioration of area continuity” at a glance in pairwise comparisons. Exclude all ISPs that generate redundancy or disturbance of pattern. Later, we can easily automate the process.
- (7) Aggregate accepted ISPs to generate new *prediction patterns* for visual evaluation (recall that two ISPs appear indispensable: **L** and **fl**).
- (8) Once acceptable subsets of ISPs to generate **Au15dp** or **Au87oc** *prediction patterns* are obtained, use again the sequential cross-validation procedures applied earlier, namely **Au15m1dp** and **Au87m5oc**. Compare the benefits of using the reduced set of ISPs to generate new *prediction patterns* and associated prediction-rate curves. Are these curves better (steeper) than the ones obtained in Step 5? Use the **Au15dp** to cross-validate the *prediction pattern* generated with **Au87oc** as DSP, and vice versa. Is the prediction-rate curve better than the one obtained in Step 5? This is to assess whether it is possible to improve the prediction quality. If not, it will mean that we have finished demonstrating database’s predicting capability and, consequently, its usefulness in mineral prospectivity.
- (9) As done in step 5, compare predictions within the DSP **Au15dp**, or **Au87oc**, and also be-



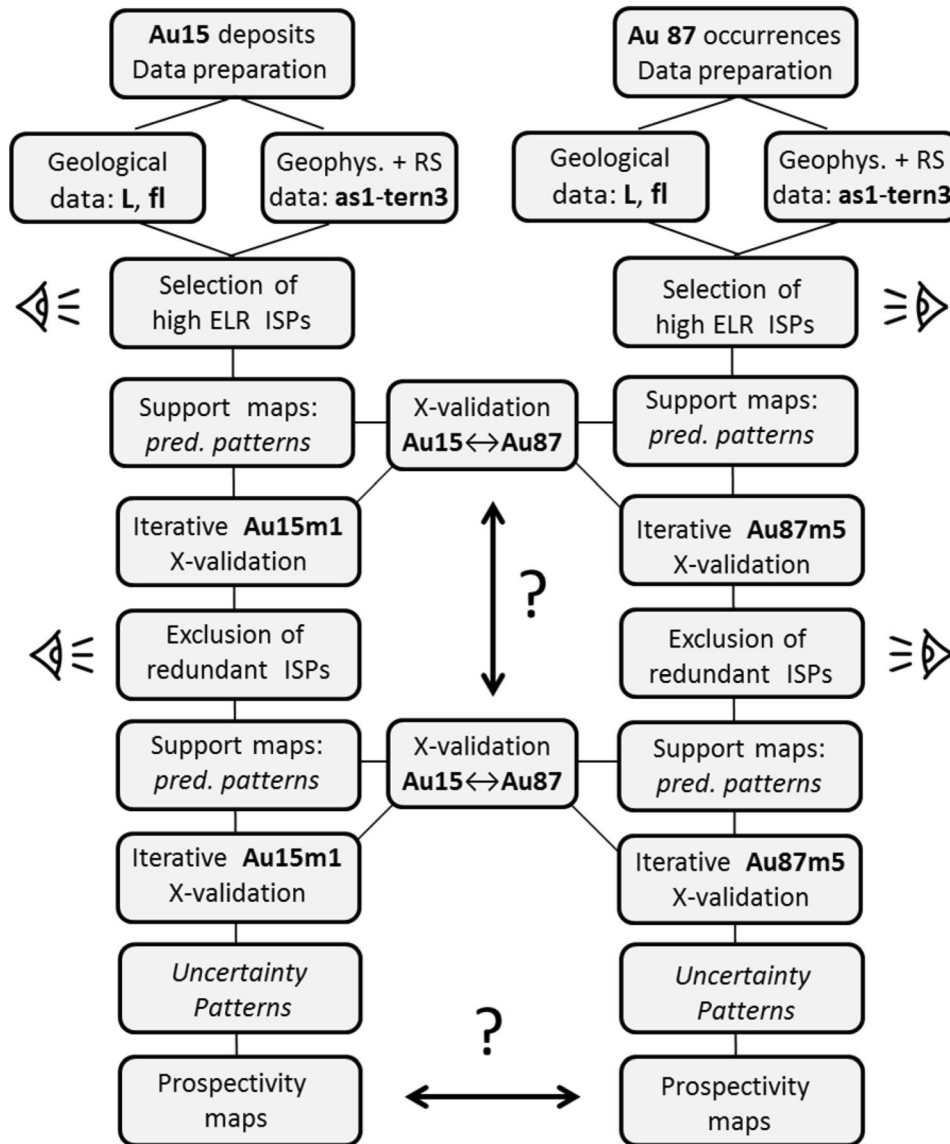


Figure 6. Analytical steps in gold prospectivity modeling in the Buhweju study area. The vertical arrow indicates the comparison of prediction-rate curves, the horizontal one the integration of ranks.

tween the two DSPs, **Au15dp** and **Au87oc**, by cross-validation. Based on numerical results and interpretation of prediction patterns using **Au15dp** as DSP or **Au87oc**, decide which *prediction patterns* are preferable and significant to predict the **Au15dp**.

- (10) Obtain *Target Patterns* and *Uncertainty Patterns*, after Step 8, as median rank and range of ranks of the *prediction patterns* from the cross-validation processes. Combine them

into *Combination Patterns* to identify areas of lower uncertainty in the study area.

- (11) Explore further the congruity of the 15 gold deposit settings. Separate **Au15dp** into well-predicted and poorly predicted groups using **Au15m1dp** procedures. Get the ELR values for the separate groups and display their distribution for interpretation. What is causing the separation into two groups? From what we know from the database, is it justi-

fiable to consider them as one group or split into two or more and repeat the analyses?

- (12) Extract the convenient part of the study area with the higher ranks and the lower uncertainty to guide in prospecting for gold. Use cost/benefit criteria to delimit the high-priority prospective areas with lesser uncertainty of ranking.

**L** and **fl** are of fundamental support to the proposition, and patterns without them are harder to interpret and appear very noisy. The data-mining steps were used to evaluate all ISPs using DSP, either the deposits, **Au15dp**, or the occurrences, **Au87oc**, which obviously have a different geologic significance in terms of prospectivity, i.e., characteristic signatures in the database. The specific suitability of the database has been explored via the ELR modeling procedure leading to information extraction. The process can be fine-tuned and automated for larger datasets, but the essential steps have been identified. An important aspect of the process area is its transparency, simplicity, interpretability and general applicability beyond the present study. The following section provides the results of the various experiments on the Buhweju study area.

## DATA-MINING EXPERIMENTS

Prediction modeling with ELR functions was considered a functional means of navigating through a complex database unavoidably containing many conditionally dependent digital images, the ISPs, poorly congruent 15 gold deposits, and 87 gold occurrences, the DSPs. The computation of ELR values, the generation of *prediction patterns* using 29 ISPs, and a fixed set of rank classes were found to be fast and practical. Visual comparison of patterns in organized sequences was also found conveniently fast. As to the prediction-rate curves and histograms, use was made of a spreadsheet. The procedure followed intends to be detailed, transparent, and widely applicable.

The following subsections describe the main results leading to the extraction of the suitability of the database of supporting the proposition mentioned in “[The Empirical Likelihood Ratio Function](#)” section. They document the characterization of the ISPs, the generation and validation of the *prediction patterns*, the subsequent computation of

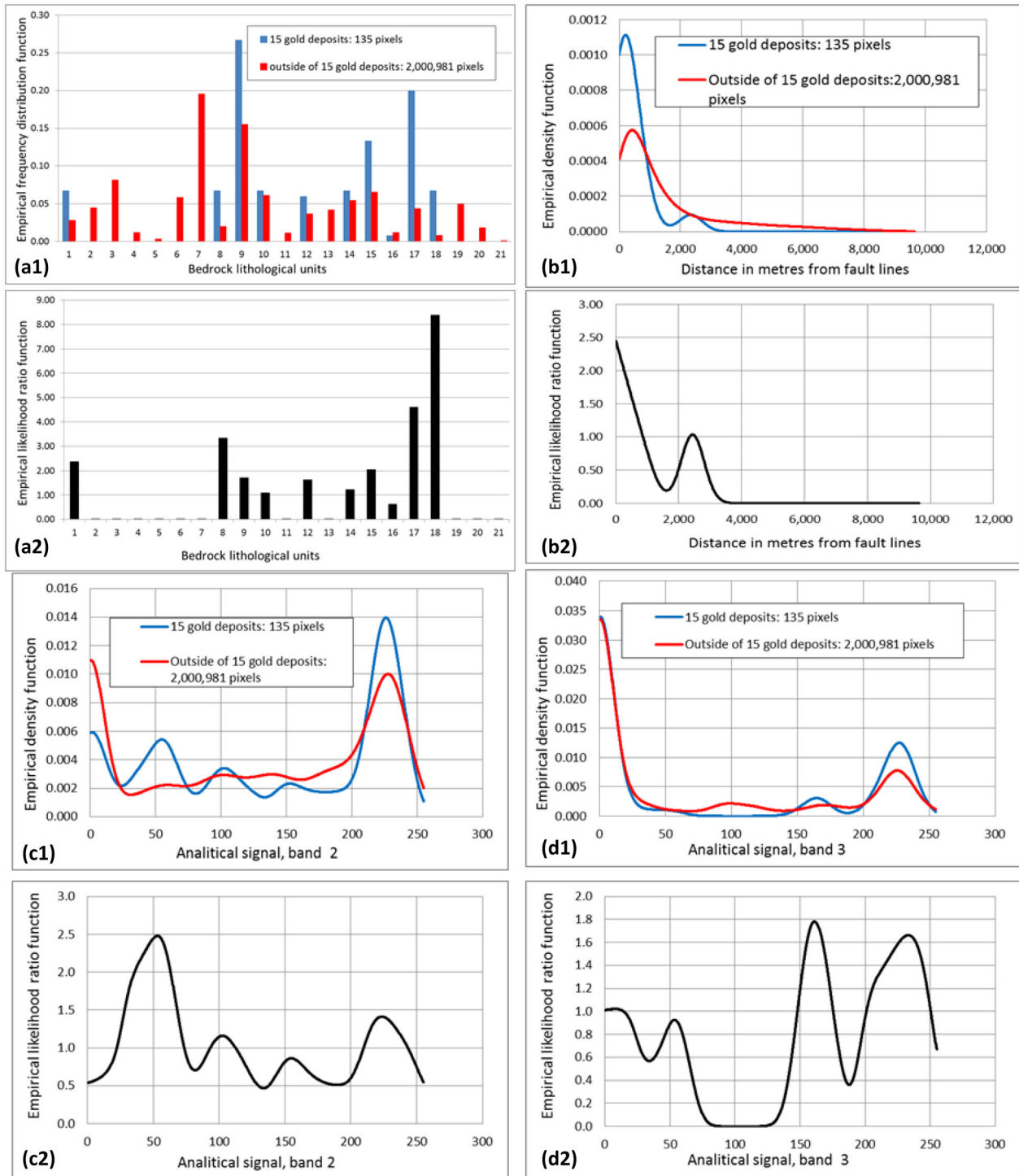
*Target and Uncertainty Patterns*, and an example of congruity analysis of the gold deposit inspired by those patterns. Finally, an attempt was made to extract prospective areas.

## Characterization of Indirect Supporting Patterns

Having at our disposal in the database several digital maps converted into supporting spatial evidence, the ISPs, i.e., images of spatial relationships as ELR values, became the first priority to measure their usefulness to the ELR prediction modeling. A first step in the analysis was to obtain the ELR values for the 15 gold deposits from all lithologic units (Table 2). Lithology is the only categorical ISP in the database. The table shows the unit ID, the lithology unit name, the corresponding number of pixels occupied in the study area, the frequency in correspondence of the occurrence pixels of the 15 gold deposits, the frequency in the remainder of the study area, and the likelihood ratio, ELR, in the last column to the right.

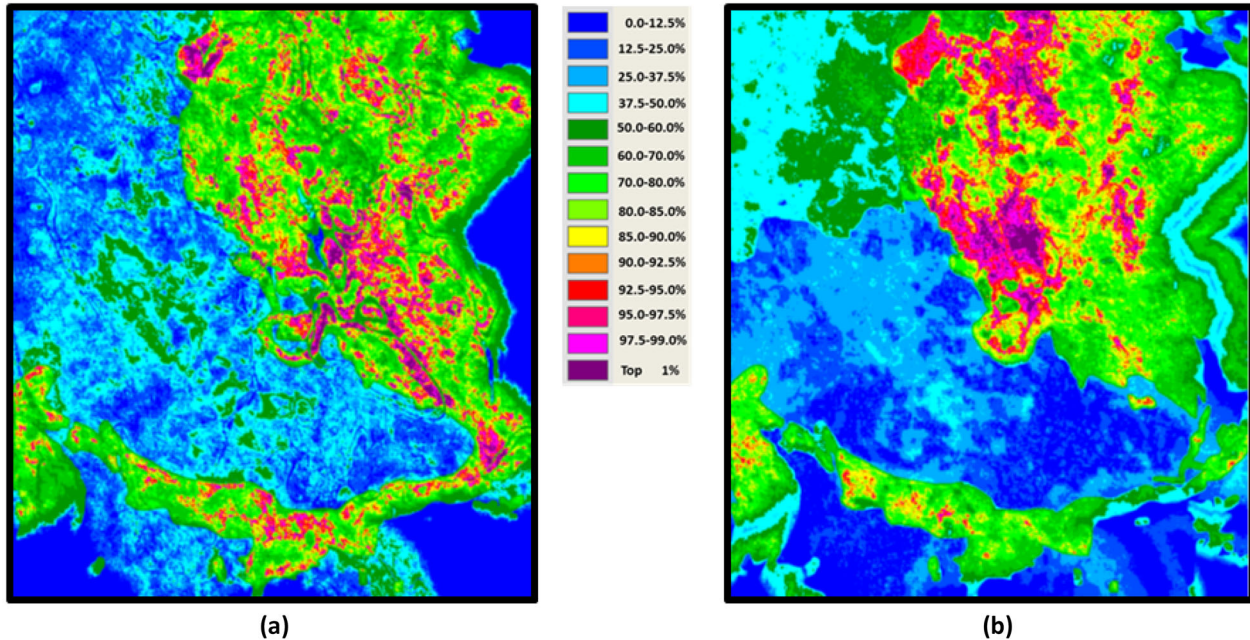
As we can see in Table 2 for the 15 gold deposits, lithology units **L<sub>1</sub>**, **L<sub>8</sub>**, **L<sub>15</sub>**, **L<sub>17</sub>**, and **L<sub>18</sub>** most support the proposition. For instance, the frequency value of 0.0667 for unit **L<sub>8</sub>** corresponds to 9 out of the 135 pixels representing the position of one gold deposit. The remaining  $40,046 - 135 = 39,911$  pixels are the frequency 0.0200 of  $2,000,116 - 135 = 2,000,981$  pixels in the remainder of the study area. The resulting ratio is  $0.0667 / 0.0200 = 3.3319$ . Of course, the statistics, when comparing areas of 135 pixels with areas of over two million pixels, are coarse. Nevertheless, this represents what we know from the database. The area occupied by the gold occurrences is 783 pixels and in Table 3 for the 87 gold occurrences, it is the units **L<sub>11</sub>**, **L<sub>12</sub>**, **L<sub>14</sub>**, **L<sub>15</sub>** and **L<sub>17</sub>** that most support the proposition. In addition, **L<sub>10</sub>** and **L<sub>16</sub>** show ELR values rather close to 2 (**1.91** and **1.95**, respectively).

Let us now consider the statistics relative to all the ISPs. Table 4 shows ELR values of the 29 ISPs of the database for the 15 gold deposits, **Au15dp**. Note the selected ratios for the units of the lithology ISP, also listed in Table 2. Table 5 shows the ratios obtained for the gold occurrences, **Au87oc**, ISPs. Observe the main differences and similarities in ratio values for lithology units in Tables 2 and 4. In particular note the high values for **Au15dp** of **L<sub>1</sub>** (**2.37**), **L<sub>8</sub>** (**3.33**), and **L<sub>18</sub>** (**8.39**) in Table 4 for the deposits, and the low ratios for **Au87oc** of **L<sub>1</sub>** (0.82), **L<sub>8</sub>** (0.00),



**Figure 7.** Empirical frequency function, density functions, and likelihood ratio functions for selected ISPs using **Au15dp** as DSP: **a** lithologic units, **b** fault line distance, **c** band 2 of analytical signal, and **d** band 3. ELR values  $\geq 2$  are good discriminators. See explanations in the text.





**Figure 8.** ELR prediction patterns using only the ISPs with ratios  $\geq 2$ . **a** Pattern for **Au15dp\_L\_flas2hg1td3k1rtp1kn3th2tern1**, whose ratios are listed in Table 4. **b** Pattern for **Au87oc\_L\_flas1hg1rtp2kn3th3u3tern1**, whose ratios are listed in Table 5. The long names are sequences of the DSP and the ISPs being used in the analysis, from lithology, fault lines and so on, listed in Figure 2.

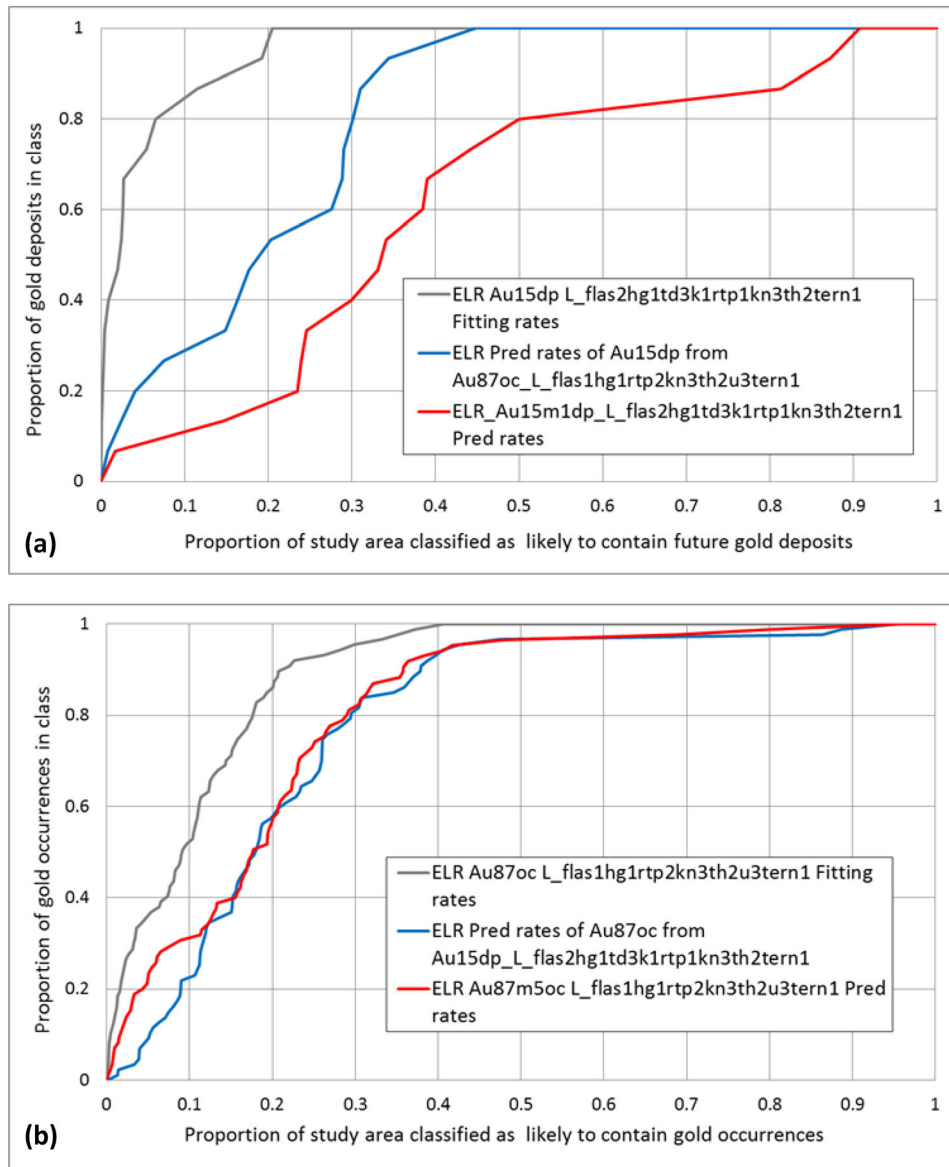
and  $L_{18}$  (1.44) in Table 5 for the occurrences. High lithology unit ratios in common for deposits and occurrences are:  $L_{15}$  (2.05 and 4.06) and  $L_{17}$  (4.61 and 2.74), respectively. The ELR values of the continuous-field ISPs identify the numeric ranges corresponding to the ratios  $\geq 2$ , or when such ratio is not reached, ELR values  $\geq 1$  listing also the maximum observed. Such simplifications in the table avoid displaying the diagrams of frequency distribution and likelihood ratios like the ones in Figure 7.

Figure 7 shows some of the ELR functions, whose characteristics are synthesized in Tables 2 and 4 for the gold deposits, **Au15dp**. The histogram in Figure 7a1 compares the normalized frequency of each of the 21 lithology units in the presence of the 15 deposit pixels ( $15 \times 9 = 135$ ), blue columns, with the corresponding one of the remainder 2,000,981 pixels in the study area, red columns. In Table 4, the corresponding values for the lithology units,  $L_1$ - $L_{21}$ , the fault lineament distance, **fl**, and analytical signal bands 2 and 3, **as2-as3**.

The histogram in Figure 7a2 shows the corresponding ratio values. Units  $L_1$ ,  $L_8$ ,  $L_{17}$ ,  $L_{15}$  and  $L_{18}$ , have ratios close to or higher than 2.00. They provide strong support to the proposition in (1) and separate the areas with presence from those with the

absence of deposits. Figure 7b1 shows the fault lineament distance density function in the presence of the deposits, blue curve, and in the remainder of the study area, red curve. As we can see in Figure 7b2, the corresponding ratios show higher values up to 2.46 only for short distances  $< 260$  m, showing the separation of the pixels in the presence of occurrences from the ones in their absence. Such separation by fault lineament distance is weaker for the gold occurrences that reach a maximum ratio value of 1.86 for distances  $< 1507$  m when comparing the respective values in Tables 4 and 5.

Figures 7c and d show the analytical signal's density functions and likelihood ratio functions of bands 2 and 3. Band 2 in Figure 7c2 shows good support in separating the presence and absence of occurrences with a peak at 2.48, while band 3 in Figure 7d2 shows lower support with peaks at 1.78 and 1.66. A complete list of such numerical support is shown in Tables 4 and 5 for the 15 gold deposits and the 87 gold occurrences, respectively. The selection of well-supporting ISPs was made considering the values in Tables 4 and 5. All ISPs providing ratios close to, or above 2 were used to obtain ELR prediction patterns (Fig. 8) for the deposits and the occurrences, **Au15dp** and **Au87oc**, separately.



**Figure 9.** Fitting- and prediction-rate curves. In **a** they are for **Au15dp**, using all the ISPs with  $ELR \geq 2$  with cross-validation **Au15m1dp** of the *prediction pattern* in Figure 8a (gray and red curves), and with the *prediction pattern* of **Au87oc** shown in Figure 8b (blue curve). In **B** they are for **Au87oc**, using all the respective ISPs with  $ELR \geq 2$  with cross-validation of **Au87m5oc** in Figure 8b (gray and red curves), and with the *prediction pattern* of **Au15dp** shown in Figure 8a (blue curve).

Tables 4 and 5 were considered decision tables for proceeding further in the data-mining exercise. They were used to group the relevant ISPs and generate the *prediction patterns* mapping their support (Fig. 8). The patterns in the illustration were obtained by equal area ranking of the prediction scores, initially computed as real numbers ranging

from zero to infinity. They were sequenced in decreasing order, from maximum to minimum, and then converted into 200 equal-area ranks, each corresponding to the 0.5% of the study area, i.e.  $2,001,116/200 \approx 10,006$  pixels. This is a general robust procedure of relative ranking visualization and interpretation.

The ranks of interest in the two *prediction patterns* (colored using the pseudo-color lookup table of the legend to indicate the groups of ranks) are the top 30% (from 70.0 to the top 1% class, i.e. from light green to violet) and in particular the top 10% (i.e., from 90.0 to the top 1%, i.e., from orange to violet). The two patterns reveal sharp differences in the spatial configuration of the top ranks as a consequence of the differences in support. Note that the ISPs selected for the generation of Figure 8a differ from those selected for Figure 8b. Figure 8a, for **Au15dp**, shows a dispersed distribution of high ranks that extends to the southeastern and southern parts as a narrow belt, also reflecting, only in part, the distribution of the 15 gold deposits (see Figs. 1 and 4). Figure 8b, for **Au87oc**, shows large patches of high ranks from the center area to the north-northeasterly areas where the gold occurrences are concentrated (see Figs. 1 and 4).

In addition, the distributions or clusters of the deposits and occurrences used to integrate the ISP support throughout the patterns are represented as gray fitting-rate curves in Figure 9. The horizontal axes represent cumulative proportions of study area pixels classified as having relatively higher ranks vs., on the vertical axes, the corresponding cumulative proportions of gold deposits and occurrences. For this, the 200 equal-area ranks were placed in decreasing order on the horizontal axis, and the cumulative proportions of deposits or occurrences in the corresponding ranks are on the vertical axes. For instance, the top 10% ranks in the gray curve of Figure 9a indicates that the corresponding area contains 85% of the deposits. The top 20% instead contains 95%, and the top 30% contains 100%. As to the occurrences, the gray curve in Figure 9b, the corresponding proportions are 52%, 86% and 95%. The fitting rates represent only how well the DSP aggregates in the equal area ranks that classify the study area.

However, having obtained the prediction patterns does not yet indicate the “goodness” as predictors of gold deposits or occurrences. For this, we need some form of cross-validation, in which we use a subset of deposits or occurrences to model the pattern and another subset to verify whether the remainder of deposits or occurrences happens to be located within relatively high ranking classes in the pattern. This is discussed next.

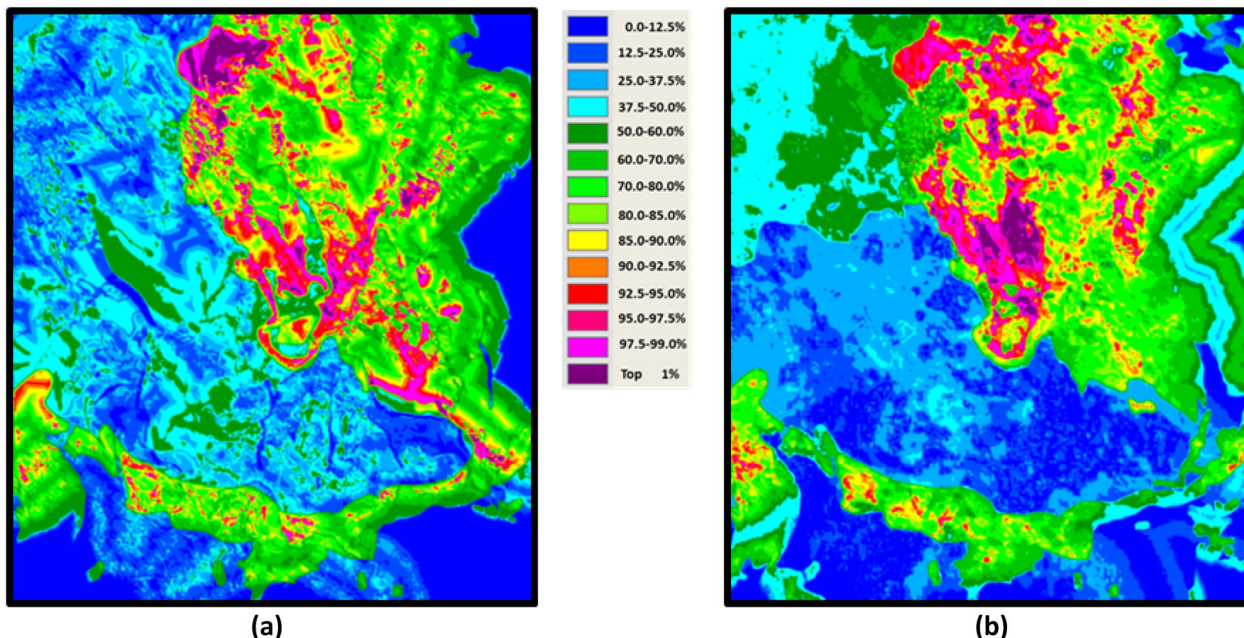
## Validation of Prediction Patterns

Strategies of cross-validation were applied to obtain prediction-rate curves and assess the prediction capability of the patterns. Such curves are different from the fitting-rate curves, although constructed similarly. Not having the time distribution of discoveries but needing to test such capability (i.e., “goodness”) of the selected set of DSP and ISPs, we pretended not to know the presence of some of the occurrences or deposits. We used the remainder to generate the *prediction patterns* and then used the “unknown” ones to validate those patterns. For the Buhweju study area, we had 15 gold deposits, termed **Au15dp**, and 87 gold occurrences, termed **Au87oc**. We decided to use a sequential exclusion strategy of 1 gold deposit and of 5 gold occurrences, respectively. Other strategies could also be used.

The sequential cross-validation process provided a prediction rate for each gold deposit and occurrence. They have been displayed as red prediction-rate curves in Fig. 9a and b, naming the processes as **Au15m1dp** (**m** stays for minus) and **Au87m5oc**, respectively. The red prediction-rate curve in Figure 9a tells us how well the sequences of 15 prediction patterns from the iterations classify (predict) each “next” deposit in the validation process, i.e., how well the 15 deposits are classified in the 15 corresponding patterns. The red prediction-rate curve in Figure 9b tells us how well the sequences of 17 prediction patterns from the iterations classify sets of 5 “next” occurrences. They were generated iteratively, each using the remaining 83 occurrences. Clearly, the red prediction-rate curve for deposits in Figure 9a is less efficient than that for occurrences in Figure 9b and is also far less steep than the corresponding fitting-rate curve. The concavity and convexity of the red curve in Figure 9a might indicate that the database signature of the gold deposits reflects the presence of two or more distinct settings. We will consider this aspect in the Congruity Analysis of Gold Deposit Database subsection.

Note that in Figure 9b the red curve is above or touching the blue one, telling us that **Au87m5oc**, red curve, is a better curve than that for **Au15m1dp**, blue curve, which indicates how well the *prediction pattern* of the gold deposits predicts the gold occurrences. However, in Figure 9a we have the reverse situation in which the occurrence *prediction pattern*, the blue curve, is a better predictor of the





**Figure 10.** ELR *prediction patterns* using reduced numbers of ISPs with ratios  $\geq 2$ , **a** for **Au15dp\_L\_flas2rtp2**, and **b** for **Au87oc\_L\_flas1rtp2tern1**. Compare the corresponding *prediction patterns* in Figure 8, where all ISPs with  $ELR \geq 2$  were used. The long names list the DSP, and the ISPs, used in the analysis, from lithology, fault lines, and so on, are listed in Figure 2.

gold deposits than that of the red curve deposits. It must be anticipated that only the higher part of the prediction-rate curves, say the top 30% or even 20%, indicates relevant support to the modeling’s proposition. This is because the uncertainty of the ranks tends to increase for the lower ranks. This is observed in the uncertainty patterns discussed in the next section.

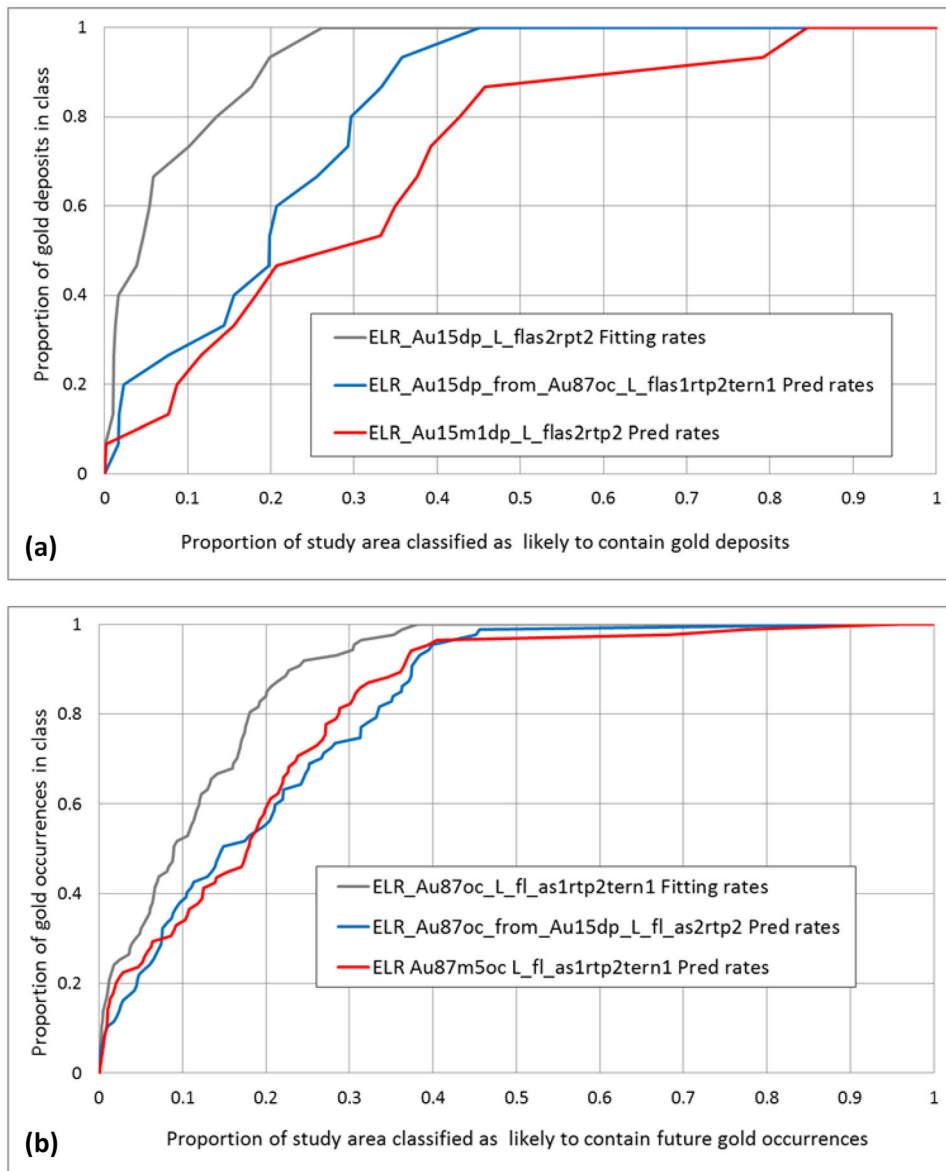
We tentatively used the *prediction pattern* of the 87 gold occurrences and validated it with the set of 15 gold deposits and, vice versa, used the *prediction pattern* of the 15 gold deposits and validated it with the set of the 87 gold occurrences. The corresponding blue prediction-rate curves are shown in Fig. 9a and b. They have approximately the same inclination, telling us that: the occurrence pattern predicts the deposits “and” the deposit pattern predicts the occurrences. What is remarkable is that the blue prediction-rate curve in Figure 9a is more efficient than the red curve.

One may ask a natural question at this point: can we do any better at generating *prediction patterns* with this database and DSPs and ISPs? As a result of our data-mining procedure, we have identified a reduced set of ISPs to model *prediction patterns*. Figure 10a shows *prediction pattern* **Au15dp\_L\_flas2rtp2**, and Figure 10b shows *predic-*

*tion pattern* **Au87oc\_L\_flas1rtp2tern1**. Compared with the patterns in Figure 8, the ones in Figure 10 show better compactness of ranks. A wider high ranking area is visible to the North in Figure 10a, and the disappearance of the high ranks to the North-East, South-East, and South. The pattern in Figure 10b shows only a minor difference from that in Figure 8b.

We can validate these patterns as was done previously when using a larger set of ISPs. Figures 11a and b show the corresponding fitting-rate and prediction-rate curves. Note the similarity of gray fitting-rate curves with the previous ones in Figure 9, the overlap of the red and blue curves in Figure 11b and some minor improvements for the blue and red curves in Figure 11a. The blue prediction-rate curve for **Au15dp** using the prediction pattern **Au87oc\_L\_flas1rtp2tern1** in Figure 11a is an improvement and probably the best we can do with the database at present. If we compare the blue and red curves at the 10, 20, and 30% top ranks in the prediction of deposits, **Au15dp**, in Figure 9a and 11a, we can see that the blue hardly improved (from 28%, 53%, 80% to 28%, 53%, 82%), but the red curves improved (from 11%, 17%, 40% to 23%, 45%, 52%).

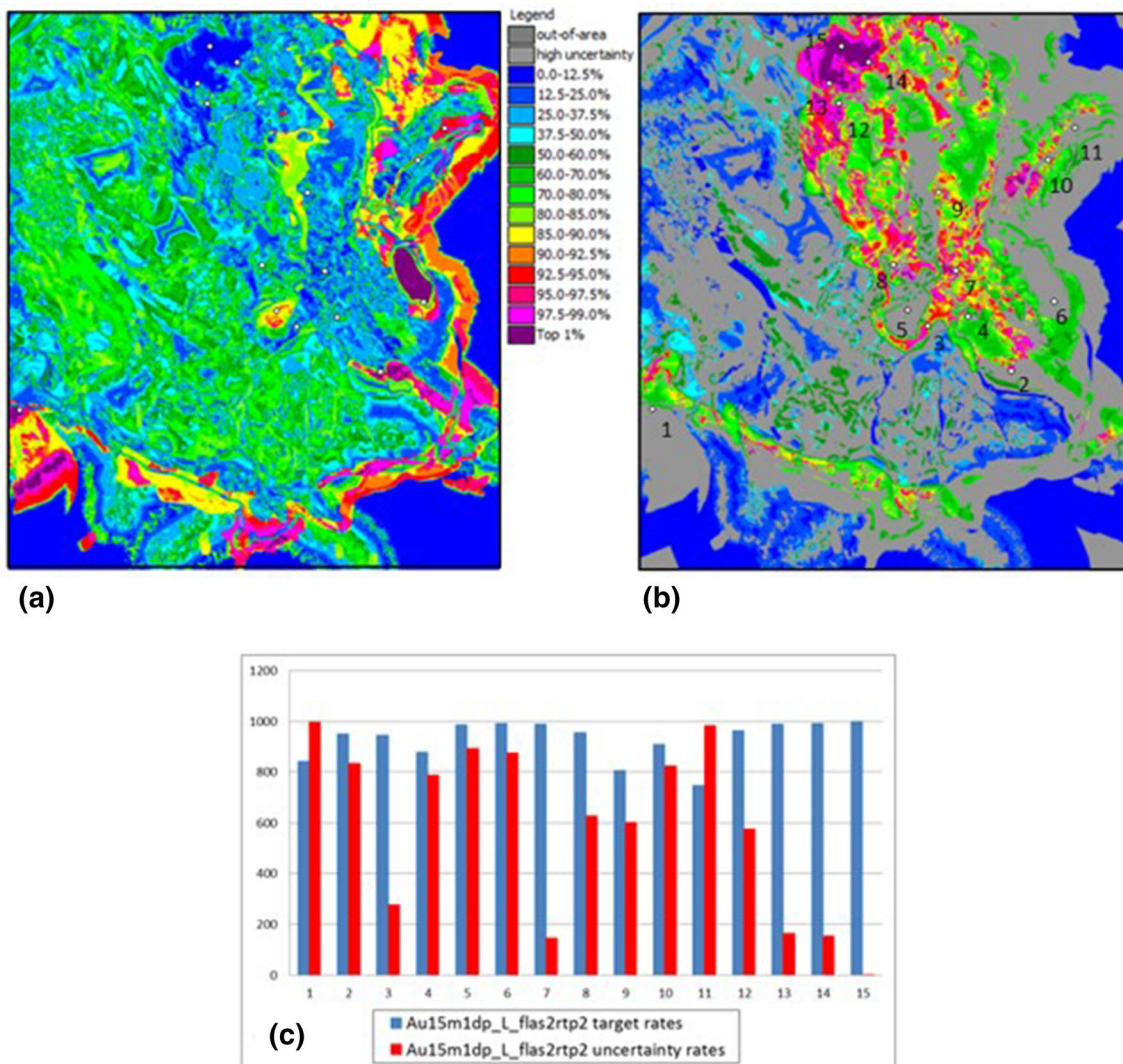




**Figure 11.** Fitting-rate and prediction-rate curves. In **a** they are for **Au15dp**, with the reduced set of ISPs with  $ELR \geq 2.00$ , using cross-validation **Au15m1dp** of the *prediction pattern* in Figure 10a, and with the prediction pattern of **Au87oc** shown in Figure 10b. In **b** they are for **Au87oc**, with the respective reduced set of ISPs with  $ELR \geq 2.00$  with cross-validation of **Au87m5oc** in Figure 10b, and with the *prediction pattern* of **Au15dp** shown in Figure 10a.

We can conclude that the 87 gold occurrences are better predictors of the 15 gold deposits than are the 15 gold deposits taken 14 each time in the iterative-cross-validation process. In addition, the fact that the strategy of sequential exclusion has provided sets of 15 and 17 *prediction patterns* offers us a

way of estimating the uncertainty associated with the ranks of the prediction. One instance is for the 87 occurrence *prediction pattern* in Figure 10b when used to predict the 15 gold deposits, and this aspect is explored in the following subsection.



**Figure 12.** Uncertainty Pattern of **Au15m1dp\_L\_flas2rtp2** in **a** and corresponding 50% *Combination Pattern* in **b**. The circles indicate the distribution of **Au15dp** over these patterns. The numbers 1 to 15 in **b** indicate the individual deposits (see list in Fig. 4 and Table 1). In **c** are the histograms of corresponding target and uncertainty ranks (vertical axis in \*1000 units) for **Au15dp** (numbers 1 to 15 on the horizontal axis). Note that 10 deposits 1, 2, 4, 5, 6, 8, 9, 10, 11, and 12 are located in areas with uncertainty ranks higher than 500. The long names are listing the DSP and the ISPs being used in the analyses, from lithology, fault lines, and so on, indicated in Figure 2. See explanations in the text. *Uncertainty Pattern* of **Au87m5oc\_L\_flas1rtp2tern1** in **d** and corresponding lower 50% *Uncertainty-Target Combination Pattern* in **(e)**. The circles indicate the distribution of **Au15dp** over these patterns. The numbers 1 to 15 in **(e)** indicate the individual deposits (see list in Fig. 4 and Table 1). In **f** are the histograms of target and uncertainty ranks (vertical axis in \*1000 units) for **Au15dp**, (numbers 1 to 15 on the horizontal axis). Note that 4 deposits 1, 5, 6, and 13 are located in areas with uncertainty ranks higher than 500, i.e., above the lower 50%.

### Generation of Target and Uncertainty Patterns

The red prediction-rate curve in Figure 11a resulted from 15 sequential predictions, each of which used 14 gold deposits for modeling and the remaining 1 for validating the ranks. We can extend the analysis by integrating the 15 patterns into what

is termed a *Target Pattern* (Chung and Fabbri, 2020). We used rank-based statistics to compute the median ranks for each pixel out of the 15 patterns. The *Uncertainty Pattern* in Figure 12a consists of such median ranks and its image, **Au15m1dp\_L\_flas2rtp2**, is almost indistinguishable from the original *prediction pattern*, **Au15dp\_L\_flas2rtp2**, in Figure 10a,

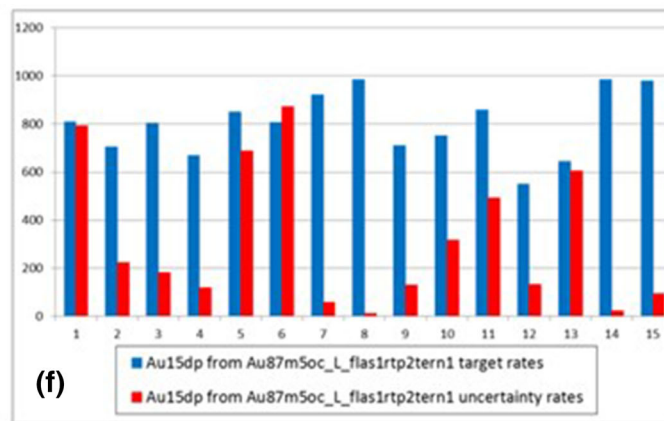
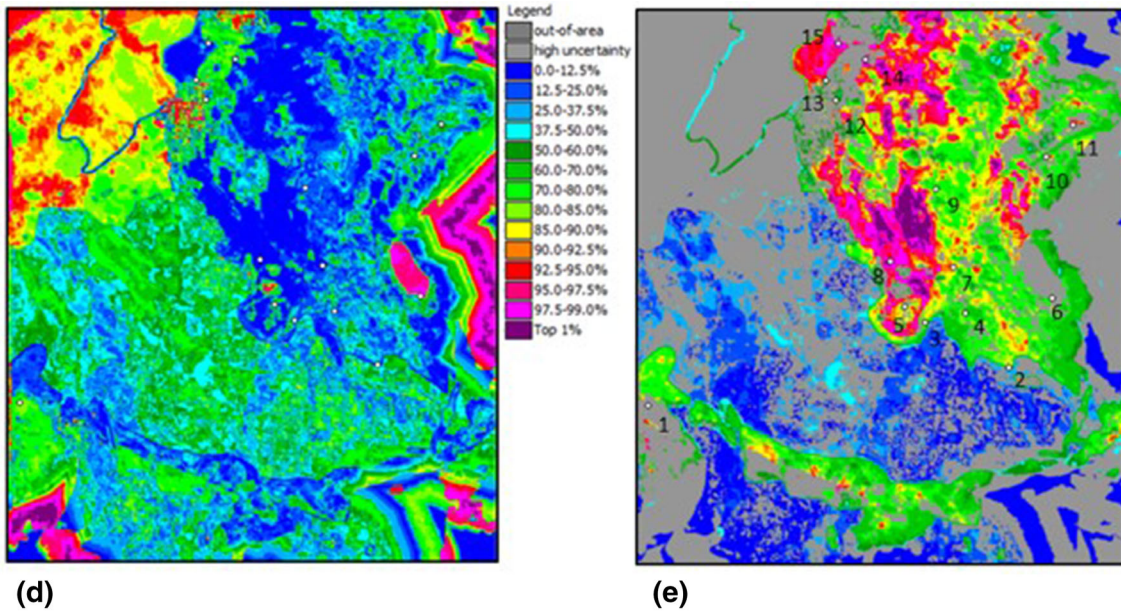


Figure 12. continued.

when using the same color legend. For this reason, it is not shown here. However, the *Target Pattern* allows computing for each pixel the range around the median ranks (the range of the 15 iteration ranks). The image ranking the ranges was termed the *Uncertainty Pattern*. The wider the range, the higher the uncertainty associated with the median ranks in the *Target Pattern*. We used the same legend for *prediction patterns*, *Target Patterns*, and *Uncertainty Patterns*. However, for uncertainty ranks, the significance is the reverse: the lower the ranks are, the lower the uncertainty is; this is what we aim at. The distribution of the 15 gold deposits, displayed as circles, in the *Uncertainty Pattern* of Figure 12a,

indicates that only 5 out of 15 gold deposits are located mostly in areas of relatively low uncertainty, i.e., areas with ranks blue to dark green.

To interpret the *Uncertainty Pattern* relationships with the *Target Pattern*, we set a convenient threshold for the uncertainty ranks in Figure 12a as follows. In the *Uncertainty Pattern*, we selected all the uncertainty ranks corresponding to the pixels with the lower 50% of the ranks. They were then used, in turn, to select all the corresponding pixels in the *Target Pattern*. The remaining target rank pixels were then considered as having uncertainty above the threshold. In the 50% *Combination Pattern* in Figure 12b, such pixels are displayed with a gray



color. In practice, we filtered the *Target Pattern* for the corresponding lower uncertainty ranks. Note, comparing the *prediction pattern* in Figure 10a with the *Uncertainty Pattern* of Figure 12a, that large parts of the intermediate ranks correspond to higher uncertainty areas, as visible in the *50% Combination Pattern* of Figure 12b.

The histogram in Figure 12c compares the target and the uncertainty ranks for the 15 gold deposits. The ranks are in \*1000 units, i.e., values of 900 correspond to the top 10% rank. The value 500 for the uncertainty ranks corresponds to the threshold value selected for the 50% lower ranks of the *Combination Pattern*, shown in Figure 12b, where the pixels corresponding to uncertainty ranks above 500 were set to gray color. The blue columns in the histogram show the target ranks, while the red columns show the uncertainty ranks for each deposit. Only deposits 3, 7, 13, 14, and 15 are located in areas of lower uncertainty, i.e., 5 out of 15.

*Uncertainty pattern* and *50% Combination Patterns* for the gold occurrences **Au87m5oc\_L\_flas1rtp2tern1** and the corresponding histogram with target and uncertainty ranks are shown in Figs. 11d, 12d, and e, respectively. Note that the uncertainty here is in the ranking of the occurrence *Target Pattern*. We have superimposed the location of the 15 gold deposits on the occurrence *Target* and *Uncertainty Patterns* to obtain the corresponding ranks for the gold deposits **Au15dp** in Figure 12d. Observe the difference in the distribution of higher uncertainty values between the *Uncertainty Patterns* in Fig. 12a and d and the similarities/differences between the two *50% Combination Patterns* in Fig. 12b and e. The *Target Pattern* of **Au15m1\_L\_flas2rtp2** is affected by the relatively higher uncertainty as shown in the histogram of Figure 12c, where 10 gold deposits fall on higher uncertainty ranking areas, much higher than in the histogram of Figure 12f, where only 4 deposits fall on higher uncertainty areas of the *50% Combination Pattern* of **Au87m5oc\_L\_flas1rtp2tern1**, 1, 5, 6, and 13. This means that their validation of prediction rank is doubtful.

We can conclude that the occurrence *Target Pattern* shows relatively lower uncertainty ranks in correspondence with the location of most of the gold deposits. This is in addition to the observation made earlier that the occurrence prediction pattern **Au87oc\_L\_flas1rtp2tern1** is a better predictor of the 15 gold deposits than of **Au15m1dp\_L\_flas2rtp2**. Observe the red prediction-rate curves in Figs. 9a

and 11a. They show concavities or irregularities that might indicate differences in the settings of deposits. This would lead to a separation into two or more groups. We considered such aspects in the following subsection. Then, we exemplify the aggregation of supports from the two types of *Target* and *50% Combination Patterns* for deposits and occurrences.

### Congruity Analysis of Gold Deposit Database Signatures

The irregularities of the red prediction-rate curves in Figs. 9a and 11a, generated by the iterative cross-validations **Au15m1dp\_L\_fl\_as2hg1td3k1rt-p1kn3th2tern1** and **Au15m1dp\_L\_flas2rtp2** indicate some discontinuity of ranking of the deposits, i.e., of their characterization in the database. Can we separate the deposits into two or more groups with better congruity and start new modeling processes?

We tentatively used the ranks of each deposit from the iterative process to separate them into two groups. Previously, we computed the 200 ranks to generate the prediction-rate curves, but in this case, we may prefer to use 1000 ranks to get a finer numerical sequencing. That many ranks would make it easier to separate many more deposits should we have hundreds more. A simple strategy is then to list such \*1000 ranks and set some arbitrary threshold values to separate groups of high and low ranks. For instance, if an arbitrary threshold of  $\geq 800$  is used, it separates the deposits predicted within the top 20% ranks from the rest located within lower ranks. Next, we could use the separate groups to obtain their relative ELR values from the database. This we did to generate Table 6 by iterative cross-validations **Au15m1dp\_L\_fl** and **Au15m1dp\_L\_flas2rtp2**. The table lists the deposit ID numbers in column 1, the \*1000 ranks in columns 2 and 3, two subsets of the 15 gold deposits after selecting an arbitrary threshold of  $\geq 750$  for the values in column 2, and the corresponding ELR values for **L** and **fl** for the two groups in column 5. In column 2, the ranks range from a minimum of 155 to a maximum of 992, and in column 3, from 156 to 999.

Tentative names of **Au9dpHr**, 9 high-ranking deposits, and **Au6dpLr**, 6 low-ranking deposits, were also used for ranks  $\geq 750$  and for ranks  $< 750$ , respectively. In column 3 of the table, where also **as2** and **rtp2** were used, the ranking differs for deposits 4 and 9. However, when lowering the threshold to  $\geq 650$  we get the same groups of 9 and 6 deposits. We



**Table 6.** Identification, \*1000 ranking of the gold deposits, **Au15dp**, and their partition into two groups with the corresponding ELRs

ID	*1000 rank <b>L<sub>fl</sub></b>	*1000 rank <b>L<sub>fl</sub>_as2rtp2</b>	Subset DSP <b>L<sub>fl</sub></b>	ELR <b>L<sub>fl</sub></b>
1	155	156	<b>Au9dpHr</b> ≥ 750	L <sub>1</sub> 0.00, L <sub>8</sub> 0.00, L <sub>9</sub> <b>2.86</b> , L <sub>10</sub> 0.00, L <sub>12</sub> 0.00, L <sub>14</sub> 0.00, L <sub>15</sub> <b>3.41</b> , L <sub>16</sub> 0.00, L <sub>17</sub> <b>7.68</b> , L <sub>18</sub> 0.00;
2	686	609		
3	<b>864</b>	<b>845</b>		<b>fl</b> ≥ 2 0.00–202.00 (0.00 max <b>2.09</b> );
4	<b>769</b>	652		
5	581	544		
6	576	574	<b>Au6dpLr</b> < 750	L <sub>1</sub> <b>5.92</b> , L <sub>8</sub> <b>8.33</b> , L <sub>9</sub> 0.00, L <sub>10</sub> <b>2.74</b> , L <sub>12</sub> <b>4.07</b> , L <sub>14</sub> <b>3.06</b> , L <sub>15</sub> 0.00, L <sub>16</sub> 1.57, L <sub>17</sub> 0.00, L <sub>18</sub> <b>20.97</b> ;
7	<b>883</b>	<b>885</b>		
8	<b>794</b>	<b>794</b>		<b>fl</b> ≥ 2 0.00–290.00 (0.00 max <b>3.11</b> )
9	<b>786</b>	669		
10	652	625		
11	565	209		
12	<b>831</b>	<b>819</b>		
13	<b>968</b>	<b>914</b>		
14	<b>899</b>	<b>924</b>		
15	<b>992</b>	<b>999</b>		

Only ISPs **L** and **fl** were used in column 2 to compute the ELR ranks using the cross-validation strategy **Au15m1\_L<sub>fl</sub>**. In column 3, **Au15m1\_L<sub>fl</sub>\_as3rtp2** was used instead. An arbitrary threshold value for the rank was selected as 750, corresponding to the top 25% ranks for the “higher ranking” group, **Au9dpHr**, and below that rank for what we have tentatively termed “lower-ranking” group, **Au6dpLr**

can identify in column 5 the different supports of the lithology units and distances from fault lines from the two groups. **Au9dpHr** is supported by **L<sub>9</sub>**, **L<sub>15</sub>** and **L<sub>17</sub>** and by **fl** at distances close to 0 m and less than 202 m. From Tables 2 and 4, we know they are the following: Quartzitic sandstone, **L<sub>9</sub>**; Undiff. Schist & amphibolite (> K, < Th anomaly), **L<sub>15</sub>**; and Granitoid gneiss (TTG) **L<sub>17</sub>**, respectively.

In contrast, **Au6dpLr** is supported by **L<sub>1</sub>**, **L<sub>8</sub>**, **L<sub>10</sub>**, **L<sub>12</sub>**, **L<sub>14</sub>**, and **L<sub>18</sub>**, in addition to **fl** at distances close to 0 m but less than 290 m. Again from Tables 2 and 4, we know that the supporting lithology units are as follows: gravel, sand, silt, and clay, **L<sub>1</sub>**; Mylonitized & brecciated quartzite, **L<sub>8</sub>**; Conglomerate and grits, **L<sub>10</sub>**; Less weathered schist, **L<sub>12</sub>**; Undiff. Mica schists and amphibolite (high K, Th anomaly), **L<sub>14</sub>**; and granite, **L<sub>18</sub>**. Note the high value for **L<sub>1</sub>** (**5.92**), gravel, sand, silt, and clay. It should not be surprising because one deposit (number 10) falls entirely on this unit, i.e. 1/6 = 16.7% of the **Au6dpLr**. This is the only gold deposit apparently located in alluvial units in the database. The separation of the two groups explains the shallow part of the red prediction-rate curves in Figs. 9a and 11a, caused by the mixing of deposits with different settings within the same DSP, **Au15dp**. Clearly, the separation is due to the different lithology units in which the gold deposits are located, possibly indi-

cating, not only different settings but perhaps different genetic types of mineralization. This cannot be concluded here and would have to be confirmed in further studies. However, the spatial distribution of the two groups of gold deposits in the study area shows an inner NW-trending elongated cluster of deposits **Au9dpHr** (3, 4, 7, 8, 9, 12, 13, 14 and 15), surrounded to the South and the East by an arch of deposits, **Au6dpLr** (1, 2, 5, 6, 10 and 11).

### Example of Interpretation of Prediction Patterns for Prospectivity

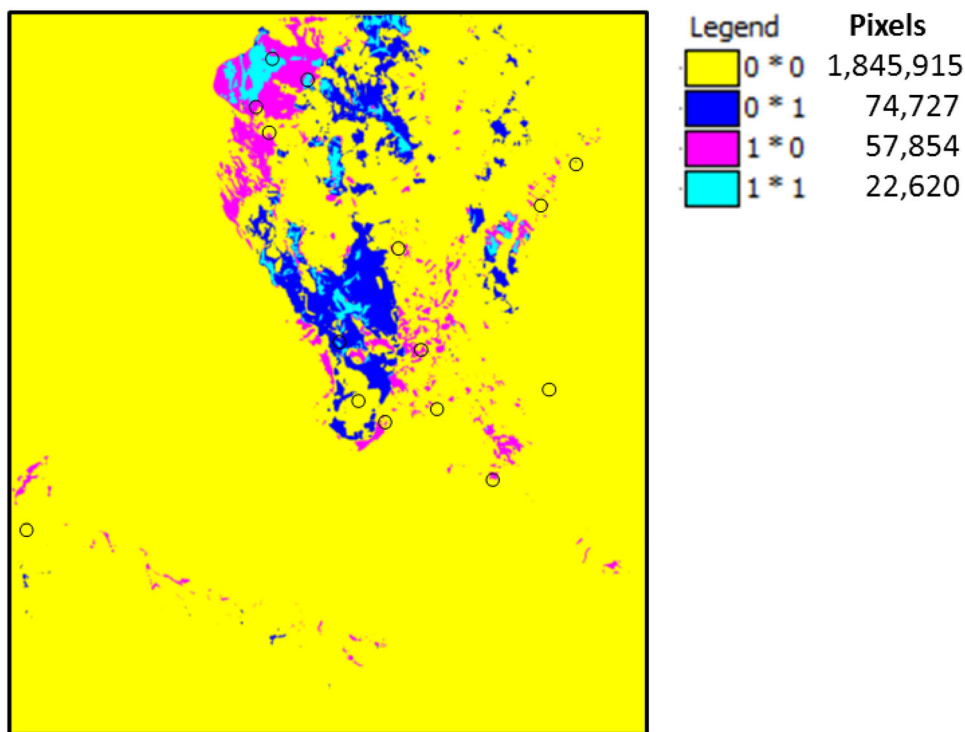
We now consider situations in which the *prediction patterns* of gold occurrences and deposits can be used for decision making in prospecting. The red and blue prediction-rate curves in Figure 11a are used as guidance in terms of costs/benefits in further exploration.

Let us select as tentative prospective areas the top 5% ranks of the 50% *Combination Patterns* in Figure 12b and e, and the corresponding *Target Patterns* (very similar to the *prediction patterns* in Figure 10a and b. Assume that we have limited means for exploration, and we would like to extract from our database the 5% of the study area with the

**Table 7.** Areas of overlap of 50% *Combination* and *Target Patterns*: top 5% and 10% ranks binarized into 1 s and 0 s

Cross	ELR_Au15m1_L_flas2rtp2 VERSUS ELR_Au87m5_L_fl_as1rtp2tern1			
X-labels	Top 5%		Top 10%	
	Target (km <sup>2</sup> )	50% Combination (km <sup>2</sup> )	Target (km <sup>2</sup> )	50% Combination (km <sup>2</sup> )
0-0	<b>1,828,508</b> (1,645.66)	<b>1,845,915</b> (1661.43)	<b>1,684,686</b> (1516.22)	<b>1,735,668</b> (1562.10)
0-1	<b>72,546</b> (65.29)	<b>74,727</b> (67.25)	<b>116,328</b> (104.69)	<b>124,809</b> (112.33)
1-0	<b>72,557</b> (66.30)	<b>57,854</b> (52.07)	<b>116,317</b> (104.69)	<b>82,088</b> (73.88)
1-1	<b>27,505</b> (24.75)	<b>22,620</b> (20.36)	<b>83,785</b> (75.41)	<b>58,551</b> (52.70)

Pixel numbers are in bold fonts and corresponding km<sup>2</sup> in gray within brackets



**Figure 13.** Overlap of top 5% ranks of 50% *Combination Patterns* **Au15m1dp\_L\_flas2rtp2** and **Au87m5oc\_L\_flas1rtp2tern1**. The top rank pixels were relabeled as 1 s and the remaining pixels as 0 s. The areas of overlap are in numbers of pixels of resolution 30 × 30 m. Of the 15 gold deposits, shown as circles, two are located in sub-area 1–1, two more on 1–0, and one on 0–1. In yellow are the remaining of the study area, outside the top 5% of the two *Combination Patterns*.

highest likelihood and lower uncertainty of discovering “the next gold deposit.”

We extracted all the pixels with ranks > 95%, i.e., the top 5%, in the *Combination* and *Target Patterns*, and converted the pixel values ranging

from 190 to 200 to the value 1, while all other pixels were converted to the value 0. Then, we overlaid the resulting binary images to obtain the respective areas of value overlaps. These overlaps are listed in Table 7 for the top 5% and for the top 10% ranks for

comparison of their respective areas. Figure 13 shows the image resulting from the overlaying (crossing) of the top 5% ranks of the two 50% *Combination Patterns* of deposits and occurrences, namely **Au15m1dp\_L\_flas2rtp2** and **Au87m5oc\_L\_flas1rtp2tern1**. The image provides us with three sub-areas of importance: 1–1 with 22,620 pixels ( $\sim 20 \text{ km}^2$ , i.e., 1.13% of study area with two deposits); 1–0 with 57,854 pixels ( $\sim 52 \text{ km}^2$ , i.e., 2.89% of study area with two deposits); and 0–1 with 74,727 pixels ( $\sim 57 \text{ km}^2$ , i.e., 3.73% of study area, with one deposit). Subareas 1–1 and 1–0 are the results from the 15 gold deposits as the DSP. Sub-area 0–1 is due to the gold occurrences. The corresponding pixel numbers for the top 5% overlaps of the *Target Patterns* and the top 10% are shown in Table 7. There we can see the corresponding numbers of target pixels for the top 5% and 10% ranks. For instance, at the bottom of column 2, we have a first priority target area of 27,505 pixels, larger than the corresponding 5% *Combination area* ( $27,505 - 22,620 = 4,885$  pixels or  $4.39 \text{ km}^2$

In other words, we identified  $20.36 \text{ km}^2$ , at the bottom of the third column, to explore as the first priority disregarding the areas of higher uncertainty and the  $52.07 \text{ km}^2$  to explore as a second priority (altogether 4.2% of the study area). In essence, the top 1.13% of study area contains 13.4% of the deposits, the next 2.89% contains another 13.4%, and the next 3.73% contains 6.7% of the deposits (1 divided by 15 = 6.7%). In terms of costs/benefits, we considered area 1–1 as the highest benefit, then 1–0 the second-highest, and 0–1 the third-highest likelihood of discovery of the “next gold deposit.” We may want to contemplate area 0–1, the less uncertain part of the top 5% of **Au87m5oc**,  $67.25 \text{ km}^2$ , not overlapping with the less uncertain part of the top 5% **Au15m1dp**, as part for integrating the supports of two 50% *Combination Patterns*.

It should be stressed that all selections made in this example are arbitrary and commonsensical. They are the result of considerations from the data-mining experiments out of the available database. We are just applying cost/benefit criteria to the statistics obtained from our analyses. In addition to these, only geological/metallogenic and economic considerations and interpretations can help further in selecting an acceptable size of prospective area for gold in the study area. The Bushweju study area has to be considered a depository of information for planning further prospection. Our results are just providing a decisional platform.

## CONCLUDING REMARKS

In modeling the spatial prediction of gold deposits, we used a database from a study area in southern Uganda. Our data-mining strategy applied ELRs to capture spatial relationships. Furthermore, the ELR function was used to integrate those relationships into *prediction patterns*. The locations of 15 gold deposits and 87 gold occurrences were instrumental in portraying their relationships with various corresponding categorical map units and continuous-value maps in the database: lithology, distance from fault lines, and geophysical and remotely sensed images. All data did focus on accurate mapping of lithology after field verification and contained unavoidable redundancy of spatial information.

The analyses led to a reduced set of digital maps that resulted effective in modeling. They represent the support of the modeling proposition to obtain and eventually improve the relative ranks of gold deposits in *prediction patterns*. Complementary support of the propositions was also observed when using the more numerous gold occurrences. The prediction capability of the database, its *Uncertainty Pattern* and the low congruity of the deposit settings were the knowledge extracted from the experiments. Using the gold occurrences and a reduced set of digital maps, the deposits were predicted better than using the entire set of images and just the deposits as direct support. For this, we used a sequential-exclusion iterative process of cross-validation. The most geologically interpretable and stable relationships were extracted: the lithology units showed conveniently higher ELR values  $\geq 2$  for the 15 gold deposits, most geologically interpretable and stable spatial relationships were: gravel, sand, silt, and clay, **L<sub>1</sub>**; Mylonitized and Brecciated quartzite, **L<sub>8</sub>**; Undif. Mica schist & amph. ( $> K$ ,  $< Th$  anomaly), **L<sub>15</sub>**; Granitoid gneiss (TTG), **L<sub>17</sub>**; and Granite ( $1987 A \pm 5 MA$ ), **L<sub>18</sub>**; distances from fault lines, **fl**, close to 0 m and not greater than 260 m; band 2 of analytical signal, **as2**; and band 2 of total magnetic intensity reduced to the pole, RTP, **rtp2**.

As for the 87 gold occurrences, mudstone, shale, and clay, **L<sub>11</sub>**; less weathered Mica schist, **L<sub>12</sub>**; Undif. Mica Schist & amphibolite (high K, Th anomaly), **L<sub>14</sub>**; Undif. Mica Schist & Amphibolite ( $> K$ ,  $< Th$  anomaly), **L<sub>15</sub>**; and granitoid gneiss (TTG), **L<sub>17</sub>**; distances from fault lines, **fl**, close to 0 m and not greater than 1507 m, band 1 of analytical signal, **as1**; band 2 of total magnetic intensity

reduced to the pole, RTP, **rtp2**; and band 1 of ternary signal, **tern1**, a combination of normalized K, Th and U.

An example of prospectivity considerations is provided using *Target*, and *Combination Pattern* top rank-threshold overlaps. It implies the application of cost/benefit criteria and, most of all exposes the difficulties of making an informed choice out of the knowledge from the data mining and spatial prediction modeling. The top ranks of the occurrence *prediction pattern* were found to partly overlap and complement the top ranks of the deposit *prediction patterns*. It raises many questions in need of new answers. For instance, using a threshold of the ELRs, setting 50% lower uncertainty or using rank-based statistics are decisions that can be objected to or modified or just fine-tuned. These are issues worth further attention and research work. The transparency attempted in our study wants to point at the question: “when will more sophisticated tools be called for by the complexities of the database to justify deep learning methodologies?”

In its present condition, the Buhweju study area database can be considered a tentative baseline against which future spatial data, such as stream-sediment geochemistry related to gold mineral deposits and occurrences, could be further integrated to improve the quality of the *prediction patterns* and the associated prediction-rate curves. Furthermore, the results obtained are of guidance in the planning of more detailed mapping and exploration.

While an application has been made to a particular study area, the authors believe that the transparent and detailed proposed procedure has wide applicability in spatial prediction modeling in situations where “black-box” models are undesirable.

## ACKNOWLEDGMENTS

Between 2008 and 2012, Uganda was remapped by the Geological Survey of Finland, GTK, in a consortium as part of the World Bank-funded Sustainable Management of Mineral Resources Project SMMRP on “Geological mapping, geochemical surveys, and mineral resources assessment in selected areas of Uganda.” The consortium led by GTK involved the collaboration of the Entebbe Department of Geological Survey and Mines, DGSM in Uganda, and the International Institute

for Geoinformation Sciences and Earth Observation, ITC, in The Netherlands. Various MSc study projects were conducted under this collaboration. This research is a product and continuation of it. Our warm appreciation goes to all the researchers who, in one way or other, allowed us to use some of their manuscripts and datasets that gave us the ideas presented in this paper. Most, if not all, have been cited in this paper. We are grateful for the support and cooperation of Mr. Tapio Lehto of GTK, who allowed us to use the geological field geo-datasets collected for the Mapping Project. We thank Mr. John Odida, the former Director of the Ugandan Geological Survey, for granting access to the field-work sites. We also are grateful to three reviewers for their constructive comments, which helped us improve the readability of our paper: Jeff R. Harris (Harris Geoscience, Canada), Greg A. Partington (Kenex Ltd., New Zealand), and Oliver P. Kreuzer (James Cook University, Australia).

## DECLARATIONS

**Conflict of Interest** The authors declare that there has been no significant financial support for this work that could have influenced its outcome, nor any known conflicts of interest associated with this publication.

## REFERENCES

- Africa Mining. (2000). *Simplified geology of Uganda*, Extracts from Mining Journal Uganda Supplement April 1996.
- Agterberg, F. P. (2014). Chapter 5. Prediction of occurrences of discrete events. In: Agterberg, F. P. (ed.) *Geomathematics: theoretical foundations, applications and future developments. Quantitative Geology and Geostatistics 18*. Springer International Publishing, Switzerland, pp. 139–188. [https://doi.org/10.1007/978-3-319-06874-9\\_5](https://doi.org/10.1007/978-3-319-06874-9_5).
- Bahiru, E. A., & Woldai, T. (2016). Integrated geological mapping approach and gold mineralization in Buhweju Area, Uganda. *Ore Geology Reviews*, 72, 777–793.
- Bahiru, E. A. (2011). *Interrelationship between lithology and structure and its control on gold mineralization in Buhweju area, SW of Uganda*. Unpublished M.Sc. Thesis, University of Twente Faculty of Geo-Information and Earth Observation (ITC), Enschede, The Netherlands.
- Banro Corporation. (2006). *Twangiza project*, South Kivu Province, Democratic Republic of Congo (DRC), NI43-101 Technical Report. [https://www.sec.gov/Archives/edgar/data/1286597/000091228206000402/banro6k\\_twangiza-techrpt.htm](https://www.sec.gov/Archives/edgar/data/1286597/000091228206000402/banro6k_twangiza-techrpt.htm). Accessed 7 April 2022.
- Barnes, J. W. (1961). *The mineral resources of Uganda*. Department of Geological Survey and Mines, Uganda, Bulletin No. 4.



- Barnett, C., & Williams, P. M. (2006). Mineral exploration using modern data mining techniques. *Society of Economic Geologists, Special Publication, 12*, 295–310.
- Bonham-Carter, G. F. (1994). *Geographic information systems for geoscientists: Modelling with GIS* (p. 398). Pergamon Press.
- Chen, Y., & Wu, W. (2019). Isolation forest as an alternative data-driven mineral prospectivity mapping method with a higher data-processing efficiency. *Natural Resources Research, 28*, 31–46.
- Chung, C. F. (2006). Using likelihood ratio functions for modeling the conditional probability of occurrence of future landslides for risk assessment. *Computers & Geosciences, 32*, 1052–1068.
- Chung, C. F., & Fabbri, A. G. (1993). The Representation of geoscience information for data integration. *Non-renewable Resources, 2*, 122–139.
- Chung, C. F., & Fabbri, A. G. (2001). Prediction models for landslide hazard using fuzzy set approach. In M. Marchetti & V. Rivas (Eds.), *Geomorphology and environmental impact assessment* (pp. 31–47). Balkema.
- Chung, C. F., & Fabbri, A. G. (2020). Mineral occurrence target mapping: A general iterative strategy in prediction modeling for mineral exploration. *Natural Resources Research, 29*, 115–134.
- Chung, C.-J., & Fabbri, A. G. (1998). Three Bayesian prediction models for landslide hazard. In: A. Buccianti, G. Nardi & R. Potenza (Eds.), *Proceedings of IAMG '98, 4–9 October 1998, Ischia Island, Naples, Italy* (pp. 204–211). De Frede Editore, Naples.
- Chung C. F., Fabbri, A. G., & Chi, K. H. (2002). A strategy for sustainable development of non-renewable resources using spatial prediction models. In: A. G. Fabbri, G. Gàal, & R. B. McCammon (Eds.), *Deposit and geoenvironmental models for resource exploitation and environmental security*. NATO Science Series 2. Environmental Security Vol. 80, Dordrecht, Kluwer Academic Publishing (pp. 101–118).
- Chung, C. F. (2003). Use of airborne geophysical surveys for constructing mineral potential maps. In: W. D. Goodfellow, S. R. McCutcheon & J. M. Peter (Eds.), *Massive sulfide deposits of the bathurst mining camp, and northern maine. Economic Geology Monograph* (vol. 11, pp. 879–891).
- Combe, A. D. (1934). Notes on the geology of the northern half of Igara, western Ankole. *Geol. Surv. Uganda. Ann. Rept. Bull., 1933*, 16–19.
- Combe, A. D. (1939). Geology of the Lubare area (western Ankole). *Uganda Geol. Surv. Bull., 3*, 1–22.
- Desharnais, G., Paiement, J. P., Hatfield, D., & Poupart, N. (2017). Mining BIG data: The future of exploration targeting using machine learning. In: V. Tschirhart, & M. D. Thomas (Eds.), *Proceedings of exploration 17: Sixtz decennial international conference on mineral exploration, Toronto, Canada, October 22–25* (pp. 319–323).
- Fabbri, A. G., & Chung, C.-J. (2008). On blind tests and spatial prediction models. *Natural Resources Research, 17*(2), 107–118.
- Ford, A., Peters, K. J., Partington, G. A., Blevin, P. L., Downes, P. M., Fitzherbert, J. A., & Greenfield, J. E. (2019). Translating expressions of intrusion-related mineral systems into mappable spatial proxies for mineral potential mapping: Case studies from the Southern New England Orogen. *Australia. Ore Geology Reviews, 111*, 102943.
- Gabert, G. (1990). Lithostratigraphic and tectonic setting of gold mineralization in the Archean cratons of Tanzania and Uganda East Africa. *Precambrian Research, 46*(1–2), 59–69.
- Granek, J. & Haber E. (2015). Data mining for real mining: A robust algorithm for prospectivity mapping with uncertainties. In *Proceedings of the 2015 SIAM international conference on data mining (SMD)*, pp. 145–153.
- Groves, D. I., Goldfarb, R. J., Gebre-Mariam, M., Hagemann, S. G., & Robert, F. (1998). Orogenic gold deposits: A proposed classification in the context of their crustal distribution and relationship to other gold deposit types. *Ore Geology Reviews, 13*, 7–27.
- GTK. (2012). Geological survey of Finland, geological map of Uganda, 1:100,000 SHEET N° 76, sustainable management of mineral resources project: Geological mapping, geochemical surveys and mineral resources assessment in selected areas of Uganda, Contract No.: MEMD/SMMRP/services/2006/000011 IDA.
- Harris, J. R., Grunsky, E., Behnia, P., & Corrigan, D. (2015). Data- and knowledge-driven mineral prospectivity maps for Canada's north. *Ore Geology Reviews, 71*, 788–803.
- Harris, J. R., & Sanborn-Barrie, M. (2006). Mineral potential mapping: examples from the Red Lake Greenstone Belt, Northwest Ontario. In J. R. Harris (Ed.), *GIS for the earth sciences*. Geological association of Canada, Special Publication 44, p. 1–21, Geological Survey of Canada Contribution #2005076.
- Herbert, S., Woldai, T., Carranza, E. J. M., & Van Ruitenbeek, F. J. A. (2014). Predictive mapping of prospectivity for orogenic gold in Uganda. *Journal of African Earth Sciences, 99*, 666–693.
- Herbert, S. (2012). *Predictive mapping for orogenic gold prospectivity in Uganda*. Unpublished M.Sc. Thesis, University of Twente, Faculty of Geo-Information and Earth Observation (ITC), Enschede, Netherlands.
- Holden, E.-J., Liu, W., Horrocks, T., Wang, R., Wedge, D., Douring, P., & Beardmore, T. (2019). GeoDocA-fast analysis of geological content in mineral exploration reports: A text mining approach. *Ore Geology Reviews, 111*, 102919.
- Hooper, M. J. (1964). *Proposal for the extension of the Buhweju gold project*. Geological Survey of Uganda, Report No. MJH/2, 31<sup>st</sup> July 1964.
- KDD. [https://Wikipedia.org/wiki/Data\\_mining](https://Wikipedia.org/wiki/Data_mining).
- Lehto, T. and Katto, E. (2014) (eds) GTK consortium geological surveys in Uganda 2008–2012. Geological Survey of Finland, Special Paper 56, 384 pages.
- Lehto, T., Westerhof, A. B., Lehtonen, M. I., Manninen, T., Makitie, H., Virransalo, P., Pokki, J., Harma, P., Koistinen, T., Saalman, K., Kuosmanen, E., Manttari, I., Katto, E., Baguma, Z., Kock, G. de & Elepu, D. (2014a). *Geological map of Uganda, scale 1:1 000 000*. Geological Survey of Finland, Espoo, Finland.
- Lehto, T., Pekkala, Y., Kärkkäinen, N., Data, G., & Kato, V. (2014b). *Mineral resources in Uganda*. Geological Survey of Finland, Special Paper 56, 165–231, 61 figures.
- Magnus. (2003). Uganda project retrieved 13/02/2011 from (12/2009) [http://www.magnusresources.com/africa\\_uganda\\_projects.htm](http://www.magnusresources.com/africa_uganda_projects.htm).
- Mänttari, I. (2014). Mesoarchean to neoproterozoic U-Pb and Sm-Nd ages from Uganda. In T. Lehto, & E. Katto, (Eds). GTK consortium geological surveys in Uganda 2008–2012. Geological Survey of Finland, Special Paper 56.
- McCuaig, T. C., Beresford, S., & Hronsky, J. (2010). Translating the mineral systems approach into an effective exploration targeting system. *Ore Geology Reviews, 38*, 128–138.
- Montsion, R. M., Saumar, B. M., Acosta-Gongora, P., Gadd, M. G., Tschirhart, P., & Tschirhart, V. (2019). Knowledge-driven mineral prospectivity modelling in areas with glacial overburden: Porphyry Cu exploration in Quesnellia, British Columbia Canada. *Applied Earth Science, 128*(4), 181–196.
- Nagudi, B., (2011). *Status of geological resources in Uganda*, Report for the Embassy of the Republic of Korea in Uganda, 35pp.
- Nyakecho, C., & Hagemann, S. G. (2014). An overview of gold systems in Uganda. *Australian Journal of Earth Sciences, 61*(1), 59–88.
- Pekkala, Y., Baguma, Z., Byamugisha, S., & Turyasingura P. (1995). *Geochemical exploration programme on gold and*

- base metals, Buhweju, SW-Uganda. UNDP Project UGA/89/001, Department of Geological Survey and Mines, Uganda.
- Reece, A. W. (1959). The stratigraphy, structure and metamorphism of the Precambrian rocks of North-West Ankole Uganda. *Quarterly Journal of the Geological Society*, 115, 389–420.
- Reece, A. W. (1961). *Explanation of the geology of sheet 76 (Buhweju)*. Geol. Surv. Uganda, Entebbe, Rept. 4.
- Roberts, R. O. (1942). *Notes on a visit to the Munyonyo mica deposit*. Unpubl. Report. Geol. Surv.
- Taylor, M. J. (2007). Report on the Mashonga exploration property of African mineral fields Inc., a subsidiary of Magnus International Resources Inc. in the Republic of Uganda, East Africa, Exhibit 99.3, Report 10.1007, pp. 59. [https://www.sec.gov/Archives/edgar/data/1163003/000112785507000585/magnu\\_scxh99\\_3.htm](https://www.sec.gov/Archives/edgar/data/1163003/000112785507000585/magnu_scxh99_3.htm).
- The Republic of Uganda (2006). *Opportunities for mining investment*, Print Book, National government publication, OCLC number: 70668986, 68 pages.
- Wang, C., Chen, J., & Ouyang, Y. (2022). Determination of predictive variables in mineral prospectivity mapping using supervised and unsupervised methods. *Natural Resources Research*. <https://doi.org/10.1007/s11053-021-09982-7>.
- Wayland, E. J. (1934). Crystalline gold and the possibilities of auriferous lodes in the vicinity of the Muti stream, Buhweju-Ankole. Department of Geological Survey and Mines, Uganda. *Ann. Rept. and Bulletin for, 1933*, 34–38.
- Wayland, B. J. (1936). *Gold of the Buhwezu plateau area*. reprinted from bulletin No. 2, published by the Geological Survey of Uganda, 1935, printed by the Government Printer, Uganda.
- Westerhof, P., Paavo Härmä, A.B., Isabirye, E., Katto, E., Koistinen, T., Kuosmanen, E., Lehto, T., Lehtonen, M. I., Mäkitie, H., Manninen, T., Mänttari, I., Pekkala, Y., Pokki, J., Saalman K. & Petri Virransalo, P. (2014). Geology and geodynamic development of Uganda with explanation of the 1:1,000,000 Scale Geological Map, *Geological Survey of Finland, Special Paper 55*, 387 p., 329 Fig, 29 tables and 2 appendices + appended map.
- Yousefi, M., Kreuzer, O. P., Nykänen, V., & Hronsky, J. M. A. (2019). Exploration information systems – A proposal for the future use of GIS in mineral exploration targeting. *Ore Geology Reviews*, 111, 103005.
- Yousefi, M., Carranza, E. J. M., Kreuzer, O. P., Nykänen, V., Hronsky, J. M. A., & Mihalasky, M. J. (2021). Data analysis methods for prospectivity modelling as applied and to mineral exploration targeting: State-of-the-art and outlook. *Journal of Geochemical Exploration*, 229, 106839.
- Zuo, R. (2020). Geodata science-based mineral prospectivity mapping: A review. *Natural Resources Research*, 29(6), 3415–3424.