Paper 34

The data mining approach to target generation in mature districts

Barnett, C. T. [1], Williams, P. M. [2]

1. BWMining, Boulder, Colorado, USA
2. BWMining, Brighton, Susssex, UK

ABSTRACT

There has been a recent explosion in the wealth of digital data available for exploration purposes. Whilst conventional methods of interpretation remain essential, they are not sufficient, in our view, for extracting the full information content from such data, especially from the multivariate relations between data sets. Supplementary computer-based Data Mining techniques are needed.

In mature districts, a supervised machine-learning approach offers new possibilities for improved target identification. Neural networks are particularly suitable since they can provide quantitative probabilistic target rankings. Care is needed, however, in devising suitable data representations, and in ensuring robust and reliable probability modeling.

Three case studies, from the USA, Australia and Canada, are presented to illustrate the approach. Each exhibits the sharply defined nature of the neural network targeting process. In the case where results of drilling subsequently became available, the statistical accuracy of the neural network model was corroborated.

INTRODUCTION

It is commonly observed that there has been a surge over recent years in the quantity of exploration data available. A similar explosion in the collection of digital data has taken place in bioinformatics, for example, as well as in fields such as particle physics and astronomy. It seems we are now collecting data almost faster than they can be absorbed. In our view, conventional methods of interpretation in exploration, though remaining important, are no longer sufficient to extract the full information content from this wealth of data, in particular from the multivariate relations between data sets. For that purpose, new methods are required, which collectively can be comprised under the heading of *Data Mining*. This is an umbrella term covering a range of techniques. Several of these have already been known for some time, including those we believe to be especially relevant to mineral exploration, namely *statistical pattern recognition*, *visualization* and *machine learning*. The need for computer-based statistical techniques was recognized by Bonham Carter (1994) and Groves et al. (2000), for example. The future importance of "data mining" was already noted in West (1997).

Machine learning

Machine learning broadly distinguishes between *supervised* and *unsupervised* learning. Supervised learning, or learning from examples, requires the existence of sufficiently many labeled cases. These form a set of known input-output pairs, usually called a *training set*, and the task is to learn the true input-output mapping from these examples. In the exploration case, the training set typically consists of a collection of known deposits and barren regions. For unsupervised learning, we know only the inputs and not the corresponding outputs. The aim is then to search for "interesting" features of the data, such as clusters or outliers, or some form of latent structure that would account for how they were generated.

In this paper we shall only discuss the case of supervised learning. This corresponds to the restriction in the title to target generation in mature districts. In such districts it can be assumed that outlines, or at least locations, of sufficiently many known deposits are available. The techniques described here are not immediately applicable to the search for new deposits in a relatively unexplored area. Data mining in such regions would need a different approach. References to some relevant unsupervised techniques, especially to powerful *visualization* methods, can be found in Barnett and Williams (2006).

There are a number of approaches to supervised learning, including neural networks, radial basis function networks, Gaussian process models, etc. Examples of different approaches and comparative studies can be found in Brown et al. (2000),

In "Proceedings of Exploration 07: Fifth Decennial International Conference on Mineral Exploration" edited by B. Milkereit, 2007, p. 513-524

Bougrain et al. (2003) and Harris et al. (2003). We believe, however, that the neural network approach holds particular promise for target identification, especially in its ability to provide quantitative probabilistic target rankings. We shall therefore concentrate exclusively on that approach in the present paper.

NEURAL NETWORK MODELLING

The neural network approach to target identification in mature districts can be seen as similar to a search for "look-alikes". It is not an attempt to target completely new types of deposit, only those which are sufficiently similar to known ones. However, this immediately invites the query "similar in what respects?" In practice there is an almost limitless variety of ways in which prospective locations can be compared. Decisions about similarity or difference will depend on which "features" are selected and how they are represented in the quantitative model. Furthermore two locations may be similar is some respects but differ in others, so that there can be any number of degrees of similarity or dissimilarity between two locations. How can such degrees be measured and how can they be translated into numerical probabilities? We identify these two issues, respectively, as those of the problem of feature selection and data representation, and the problem of robust and reliable probability modeling.

Feature selection and data representation

A wide range of geoscientific data sets is likely to be available in a mature district. This is increasingly true as government agencies seek to encourage exploration and mining. Many such data sets are now in the public domain. These may include gravity, magnetics, radiometrics, soil samples, stream sediments, lithology, structure, satellite imagery etc. and any of these data sets might be relevant to the targeting process. Some may be more relevant than others, depending on the minerals being targeted and the factors controlling the mineralizing system. There may be differing prior opinions about the likely ranking of data sets in terms of relevance, but it is better for that ranking to be a product of statistical analysis, rather than a presupposition. An even-handed approach should include all potentially relevant data sets and allow a statistical model to decide how much account to take of each.

Nonetheless, this is not a full solution to the problem of feature selection. The aim of targeting is to differentiate specific locations. How do two locations, ideally to be distinguished to the resolution needed for sitting a drill collar, differ with regard to their gravitational properties for instance? It is not sufficient to use just one numerical gravity value to distinguish locations. There is little meaning in a single-point geophysical reading, or even a topographical one. There is no reason why mineralization should occur at a single elevation, such as 3,000 meters. The same applies to a tranche of gravity readings. Clearly it is the pattern of data in the neighborhood of a given station that is important. It may be that flanks of gravity highs have particular significance, but it would be wrong to build this into the model as an assumption. It should only be necessary to provide the raw materials needed for the model to construct such features for itself, if the model determines that such features correlate with known deposits.

In the case of geophysical data sets, such as gravity or magnetics, a useful way of supplying this raw material is through a collection of derivatives of the primary data. This corresponds to representing a function by its Taylor series expansion in the neighborhood of a given point. For the approximation to be valid to a reasonable distance, it is necessary to include both first and second order terms, including cross derivatives. Together with primary data, this provides 10 numerical coefficients, including vertical derivatives, for characterizing a given grid location. Furthermore it is possible to view data at different scales, using upward continuation. The derivative process can then be repeated at a different elevation to provide a further 10 coefficients, making 20 in all if just two elevations are used. In all, these can provide a rich characterization of the local features of a given geophysical data set.

The same considerations apply to geochemical and remote sensing data. A single-point observation may not capture the full significance of the data; gradients and textures may also be important. In both cases the model needs information about neighboring values. These can again be provided through derivatives, though fewer may be needed if a shorter range is adequate.

Geology also needs special treatment. The geological map may record dozens of different formations, and each needs to be represented as a possible input to the model. But neighboring formations and contacts may also be important aspects of the geology so that, again, it is necessary to find a representation of how each location relates to its immediate neighborhood. Similar issues arise with respect to geological structure. There is a need to represent how any individual location relates, in terms of distance for example, to neighboring structures, both major and minor, and ideally to their strikes.

Equipped with suitable ways of representing the various individual data sets, a vector $\mathbf{x} = (x_1,...,x_M)$ can be associated with each geographical location, where each component of $\mathbf{x}$ is one of the quantities mentioned above. The number M of components may be large, of the order of several hundreds. The feature vector $\mathbf{x}$ then represents the raw exploration data, at a given location, in as complete and neutral a way as possible.

Probability modeling

The aim of the neural network approach is to determine a numerical probability for the occurrence of a mineral deposit at each grid point in the region of interest. This quantity can be written as $P(D|\mathbf{x})$. It is the conditional probability that a location with exploration features $\mathbf{x}$ hosts a deposit.

In order to achieve a reasonable degree of reliability in the modeling process, it is assumed in this paper that the region of interest is a mature district where there are sufficiently many known deposits. These deposits provide positive instances of mineralization, each with its own associated feature vector $\mathbf{x}_1,...,\mathbf{x}_N$. We also need negative instances whose feature vectors

$\mathbf{x}'_1,...,\mathbf{x}'_N$ are associated with an absence of mineralization. Ideally the latter will correspond to locations of existing exploration drill holes which have failed to intersect mineralization. Alternatively, we have it found equally satisfactory to pick N locations at random in the region of interest, and to label them conventionally as negative instances. [If there were a significant chance that a location picked at random would host a mineral deposit, the methods proposed in this paper would be largely unnecessary.]

The method used to fit the model is penalized maximum likelihood, where the parameters to be fitted are the connection strengths in a multi-layered feed-forward neural network. Details of the method can be found in Barnett and Williams (2006) and in papers referred to there. It is important to stress that the model used here is not a simple binary classifier. If it were, the final target map would consist of just two inferred types of region, prospective and barren, with no distinctions being made within prospective regions. By contrast the neural network model provides numerical probability estimates for each location, so that targets can be ranked in terms of favorability.

Fitting probability models of this type has to be undertaken with caution in view of the relatively small number of samples available. Even in a mature district, the number of known deposits of any significant size is likely to be measured in dozens rather than hundreds. It may be justifiable to extract multiple positive instances from large deposits, if outlines are known, but the total number of positive instances available is still unlikely to exceed several hundred. The number of parameters in the model, on the other hand, may be measured in thousands. The reason is that the exploration feature vector may need to contain hundreds of components, if full justice is to be done to the information contained in multiple data sets, since there is no knowing in advance which components are relevant, either individually or in combination. But even a linear model needs as many parameters as there are feature vector components. A non-linear model, which is certainly needed to model the complex non-linear relationships between feature vectors and target probabilities, will necessarily exceed that number by a sizeable factor. There is therefore cause for serious concern about over-fitting the data, which can be explained as follows.

Intelligence tests sometimes list a sequence of three or four integers and invite the subject to supply the next term, despite the fact that there are infinitely many legitimate answers. Since a rule exists to fit any finite sequence, including one obtained by adding a random integer to the original sequence, intelligence is presumably attributed to subjects whose responses intuitively conform to a "simple" rule. A similar, but more common, scientific problem is that of curve fitting. Even restricted to polynomials, if no bound is placed on the order of the polynomial, there are infinitely many curves fitting the data, but which differ arbitrarily when used for interpolation or extrapolation outside the sample. The problem faced by the statistical analyst is to match the complexity of the model to the information content of the data. The solution we adopt to this important problem is referenced in the Appendix to Barnett and Williams (2006). It involves the choice of a suitable penalty function in conjunction with maximum likelihood fitting.

A significant consequence of penalizing model complexity is that some known deposits may be fitted less well than others. A known deposit that is not well fitted by the model emerges as being untypical in terms of its exploration characteristics; exploration of locations with similar characteristics elsewhere should have relatively low priority. Conversely, some locations presented to the model as negative training instances may score relatively highly. These would be particularly interesting places to explore.

Target ranking

The neural network model provides conditional probabilities of deposits at each location. These are known as *posterior probabilities*. Numerically, however, they depend on the *prior probability* which, effectively, is an estimate of the total extent of mineralization over the region of interest, wherever it might be found. In practice, it is more reliable statistically, and it is sufficient for the purposes of exploration, to calculate a relative quantity, namely the ratio of the posterior odds to the prior odds. The logarithm of this quantity, which depends on the exploration characteristics at a given location, is known as the weight of evidence. [See Good (1950). This quantity is also used by the "weights-of-evidence" approach to targeting, as described for example in Bonham-Carter (1994), but the method of calculation and the assumptions underlying the calculations are quite different. The assumptions of the weights-of-evidence approach are that exploration data are conditionally independent, and that components of the feature vector x can only assume a few discrete values. By contrast, the present approach places no restrictions on the range of possible values of components of the feature vector, and deliberately seeks to exploit, rather than discard, the multivariate information in exploration data.] This can be imaged throughout the region, with contours of the weights of evidence being the same as contours of the corresponding conditional probabilities. High spots of this image are the neural network targets. The choice of a suitable contour then controls the desired extent and the resulting shape of a target area.

It is important to emphasize that any target ranking depends on the body of evidence employed in making the ranking. If the modeling process is repeated after further exploration data becomes available, exact locations and rankings of the targets may change. In practice, however, the process tends to converge, unless new data wholly uncorrelated with existing data are introduced. Nonetheless, even in a state of data saturation, the neural network ranking is only based on information that can be represented in the form of gridded data. Evidence that is not representable in this form, such as geological opinion or theory regarding the regional mineralizing system, needs to be taken into account alongside the empirical ranking provided by the neural network. Furthermore, the neural network can only hope to locate targets to a certain degree of accuracy. Site visits will still be required before any final decision is made whether to drill and, if so, where exactly to locate the holes.

Data relevance

The fact that target rankings depend on data input can be exploited to probe the rankings. There is no obstacle to training a network on a subset of the total data, where the subset comprises either an individual data set, or a partial combination. In this way it can seen which are the gravity targets, which are the geochemical targets, which are the structural targets, etc. The most reliable targeting will be based on the total evidence, which takes account of the full complex inter-relations between the data sets. But insight can be obtained through examining how targets, based on the combined data, match up with targets based on individual data sets.

Networks trained on individual data sets also permit the measurement of data relevance. The particular style of mineralization found in a given region is likely to determine which data sets are the most predictive. These differences can be investigated visually by comparing images of targets based on individual data sets and those based on combinations of data. It is possible to go beyond this, however, by using statistical information theory to provide a precise numerical measure of the information which an individual data set contributes, relative to the information contained in the full combination. This permits a ranking of the various data sets for their relevance over the area of interest.

CASE STUDIES

To illustrate the data mining approach described in the preceding sections of this paper, three case studies will be presented: one from the Western USA; one from Eastern Australia; and one from Eastern Canada.

Walker Lane

The Walker Lane shear zone, which straddles the border between Nevada and California in the western United States (see Fig. 1), has a long history of exploration and mining dating back to the discovery of the famous Comstock Lode in the late 1850s. Though perhaps not as well known and certainly not as productive as the neighbouring Carlin District, the Walker Lane is notable for its numerous occurrences of volcanic-hosted epithermal gold and silver deposits.

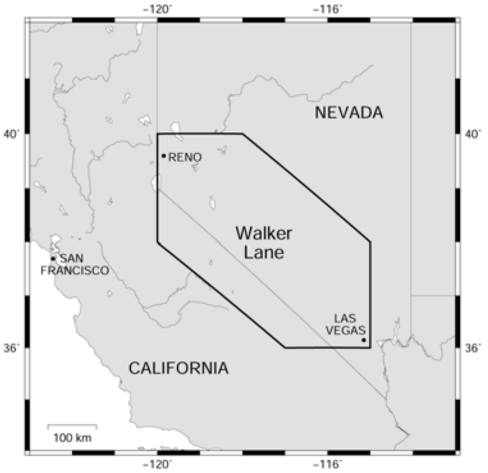

Figure 1: Map showing the location of the Walker Lane study area in the Western USA.

Aside from the Comstock Lode, which in its day produced nearly 200 million ounces of silver as well as 9 million ounces of gold, and Round Mountain, which has an estimated gold content of 14 million ounces, there are at least ten other deposits with established contents of over one million ounces of gold. If all the smaller (that is less than one million ounce) deposits are taken into account, then to date approximately 50 million ounces of gold have been discovered in the area.

The majority of these gold deposits occur in the Walker Lane shear zone, which is a 100-km wide, NW-trending structural corridor extending southeast from Reno towards Las Vegas (see Fig. 2). This strike-slip system contains a series of deep-seated, right-lateral shears and associated normal faults, which presumably provided the channel ways for magmatic and hydrothermal fluids.

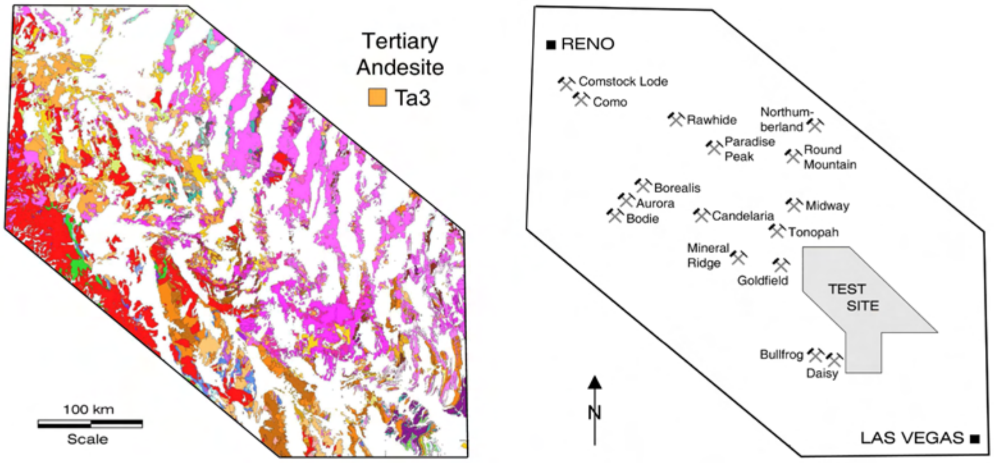

Figure 2: Left: Regional geologic map of the Walker Lane, in which the late-Caenozoic cover rocks have been left uncoloured. Right: Locations of known gold deposits. The area marked in gray contains the Nevada Test Site, and other military reserves that are off-limits to mineral exploration.

The Walker Lane is bounded to the southwest by the uplifted Sierra Nevada Mountains and to the northeast by the Basin and Range extensional terrain. The rock outcrop in this area is mostly Tertiary volcanics with an assortment of Mesozoic intrusions of mostly monzonitic or granodioritic composition. The basement is made up of strongly folded and thrusted Palaeozoic sedimentary rocks. These rocks are exposed in the southeast of the area around Las Vegas, where they contain a high proportion of carbonates.

Exploration data sets

The area covered by the Walker Lane study was about 140,000 square kilometres, as shown in Figure 1. Our data mining study was commissioned by Newmont Mining Corporation as part of a regional survey for gold in this area. The available exploration data sets used in this study were

Regional geology
Regional structure
Digital elevation model
Landsat thematic mapper
Airborne magnetics
Airborne radiometrics
Isostatic residual gravity
Stream sediment geochemistry
Known deposit footprints

The majority of these data sets are in the public domain and were obtained from sources like the Nevada Bureau of Mines and the National Geophysical Data Center. One exception was the regional structure, which was based on a proprietary in-house compilation (belonging to Newmont). This was not simply a tracing of linear features, but was a geodynamic interpretation which took into consideration regional tectonics such as the thrusting of the Basin and Range, the buttressing effect of the Sierra Nevada batholith, and the strike-slip movements of the San Andreas fault system (B. Davies, personal communication, 2003).

The regional geological base map (see Fig. 2) was created from a splice of the geology map of Nevada, drawn at a scale of 1:500,000, and the geology map of California, drawn at a scale of 1:750,000. Details of these maps, which are available in digital form from the respective state surveys, can be found in Stewart and Carlson (1978) and Jennings (1985). Since the Nevada map covered the larger area, both maps were unified to the legend of the Stewart and Carlson map.

The known deposits layer is the most critical layer, since it is used for training the network against all the other data sets. In producing this layer, all known gold deposits exceeding 50,000 ounces were carefully plotted from air photographs, detailed publications, or field visits with a global positioning system (GPS). About 150 separate deposits were located in this manner. Figure 2 shows the locations of some of the larger deposits in the Walker Lane, marked with traditional crossed hammers. In the data mining study, actual footprints were used wherever possible.

All these data sets have their own strengths and weaknesses. The geologists may feel that the state-wide geology is too broad brush and contains mapping errors; the geophysicists may be concerned about missing data due to the variable station interval of the gravity surveys; the geochemists may worry about the different sample and analytic procedures used to collect the stream sediment data. These are valid concerns. Certainly, the better the quality of the input data the better the results that will be achieved by an analytic process. Nonetheless, these are typical regional exploration data sets that we have learned to deal with in our normal manual interpretations. Furthermore, statistical correlations between multiple data sets provide some redundancy in overall information content, so that local deficiencies of an individual data set are compensated by the combined force of the others.

Data mining results

Since the Walker Lane is a competitive exploration area, the final, overall target map was deemed too sensitive to be shown in this publication. However, part of the survey area can be shown that falls inside the Nevada Test Site. Not surprisingly, after four decades of nuclear weapons testing, this part of the Walker Lane is permanently off-limits to gold mining.

Figure 3 shows the results that were obtained from an area which is about 50 km southeast of the historic mining district of Goldfield and covers about 2500 km^2 in the northwest corner of the Test Site. The target favorability map is based on all the exploration data sets described above except the geochemistry, which was not collected in the Test Site. The Landsat data and the geology are shown at the same scale for direct comparison to the target map in the center.

The data mining exercise produced two interesting targets, which would both certainly be followed up if access were permitted to this area. Target A can be seen to be coincident with a color anomaly in the TM data, indicating the presence of alteration associated with mineralization. Target B coincides with a circular feature that is visible in both the TM and the geology. Further interest is added by the small, orange-colored formation that crops out at the center of this probable caldera. This is a Tertiary andesite that is well known to Nevada geologists for hosting many of the larger volcanic-hosted gold deposits in the Great Basin. Remember, however, that these targets are based on all the available exploration data, for example the gravity, magnetics and structure, and not just the TM and geology.

These two favorability targets, which are typical of other targets produced by the data mining study elsewhere in the Walker Lane, are sufficiently small to be checked out in a matter of days by a preliminary field inspection and grab rock sampling to determine if further exploration is warranted. At this stage, detailed geological mapping, geophysical surveying and geochemical sampling would normally be carried out to refine the targets for drill testing.

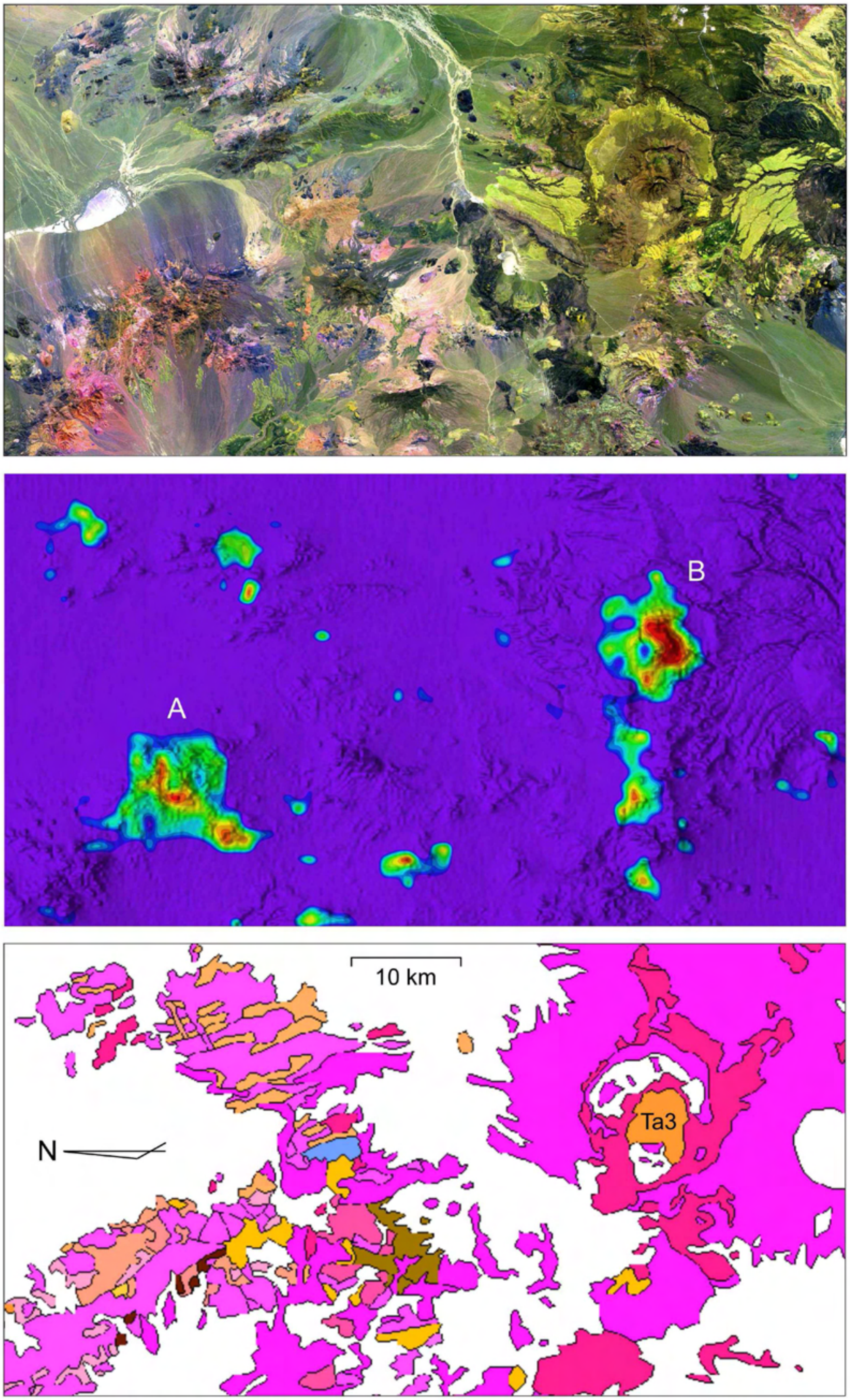

Figure 3: A close-up of part of the Nevada Test Site, an off-limits area within the Walker Lane. The target favorability map in the middle is compared to the TM image (bands 571) above and the geology below. The targets are superimposed on a shaded-relief image of the topography. Note the circular feature in both the TM and geology, and the outcrop of Tertiary andesite *Ta3*, in the vicinity of target B.

Lachlan Fold Belt

New South Wales has a history of exploration that goes back over 150 years. By official accounts, gold was first discovered near Bathurst, in 1851. There had been earlier discoveries in this area, only 200 km west of Sydney, but news of these had been suppressed by the government of the time, as gold finds were definitely not considered desirable in what was then still basically a convict society.

The 1851 announcement naturally sparked a gold rush, and the area soon swarmed with prospectors, and many other discoveries were made. None of these were on a grand scale, however, and the small army of prospectors eventually moved on to richer gold fields in Victoria and Western Australia. The Eastern Lachlan Fold Belt (LFB) was subsequently somewhat overlooked by mineral explorers until the discovery of a large base metal deposit at Woodlawn in 1972, followed by the rich gold-copper deposits at Cadia in 1992. This started a latter-day staking boom, and the LFB is now considered by geologists to be a highly prospective area for new base and precious metal discoveries.

The area selected for this study (see Fig. 4) spans 300 km E-W by 400 km N-S, making a total of some 120,000 square kilometres. The choice of this particular area was based on various factors: notably, the concentration of known deposits, the availability of modern data sets, the logistics and accessibility, and the general desirability of staying clear of government reserves or population centres where mining activities might be unwelcome.

The Lachlan Fold Belt forms a small portion of the Tasman Fold Belt, stretching from the Queensland to Tasmania. This is an orogenic belt, which developed during the Palaeozoic Era from the Cambrian through to the Carboniferous. The LFB has been affected by three principal orogenic events, namely the Benambran (Late Ordovician – Early Silurian), Tabberabberan (Middle Devonian) and Kanimblan (Early Carboniferous) orogenies. In the eastern part of the belt, Ordovician-Silurian volcanoes and intrusions were formed in a north-south trending island arc setting that is known as the Macquarie Arc.

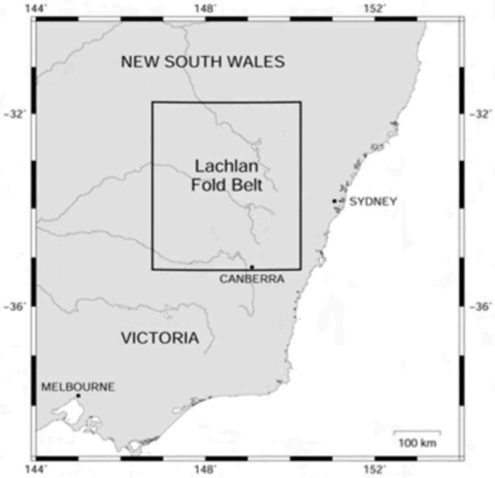

Figure 4: Map showing the location of the Lachlan Fold Belt study area in Eastern Australia.

The Ordovician and Silurian rocks which outcrop in the LFB study area consist of deep water sediments, e.g., turbidites, carbonates and shales, interspersed with intermediate to mafic volcanics, and intruded by shoshonitic intrusions. The best gold and copper mineralization, for example at Cadia, appears to be closely associated with Ordovician monzonites dated at around 440 Ma. The basement rocks are deduced to be Cambrian greenstones, which outcrop to the west of the study area. To the north and east, the Palaeozoic assemblages are overlapped by Jurassic and Triassic sediments.

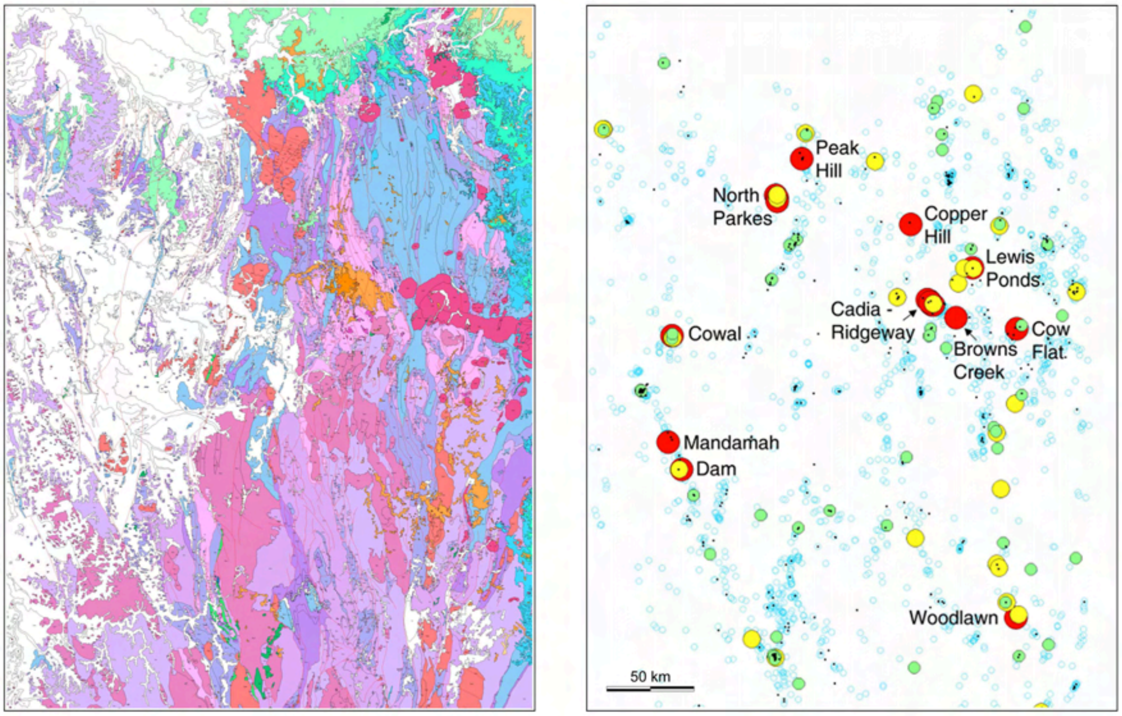

Figure 5: Left: Regional geologic map of the Lachlan Fold Belt, in which the late-Caenozoic cover rocks have been left uncoloured. Right: Locations of known gold occurrences. The larger deposits are marked with red, yellow and green circles.

The gold endowment of the study area is close to 50 million ounces (see Fig. 5). The best known deposits occur at Cadia-Ridgeway (over 30 million ounces), Lake Cowal (5 million ounces), and Northparkes (1.5 million ounces). However, within the greater Lachlan Fold Belt, which runs into Victoria and includes the huge deposits of Bendigo and Ballarat, the total endowment is over 150 million ounces.

Exploration data sets

The exploration data sets used in the LFB study were

- Regional geology
- Regional structure
- SRTM elevation data
- Landsat thematic mapper
- Airborne magnetics
- Airborne radiometrics
- Isostatic residual gravity
- Stream sediment geochemistry
- Known deposit footprints

The raw material for all these data sets are in the public domain, available partly from the New South Wales Department of Primary Industries, and partly from Geoscience Australia, an agency of the Australian federal government, so no proprietary data sets were used. The data mining study was carried out on our own initiative, but has subsequently been licensed to Rimfire Minerals Corporation.

The geological map that was used for this study came from Geoscience Australia and was just released a year ago. Details can be found in Liu et al. (2005). This is a seamless 1:1,000,000 scale digital map based largely on the 2003 version of the NSW state geology data set, which in turn was compiled from the historic 1:250,000 and 1:100,000 scale map sheets.

To establish the correlation of the various exploration data sets with gold mineralization in the LFB, the footprints of as many known gold deposits as possible were established. The primary information for this purpose came from the nationwide OZMIN database compiled and maintained by Geoscience Australia, and described by Ewers et al. (2002). This data set is based on literature search of old records, reviewed by geologists with specific knowledge of each deposit. Besides the deposit name and location, geological information such as grade, tonnage, strike and strike-length are often included. An estimate of the positional accuracy is also provided, which can range from very precise (e.g., under 50 m) to very approximate (over 1 km).

Figure 5 shows the distribution of known deposits in the LFB study area. Only the larger deposits were used to train the neural networks. To improve on the location information, we researched many of these deposits ourselves, by reviewing publications and company reports available online, and also by placing calls to companies controlling the deposits in the LFB. We succeeded in obtaining accurate footprints, or digitized outlines, for about 30 of the principal deposits in this manner. For another 60 or so smaller deposits, we simply used an oval outline based on the strike information in the OZMIN database.

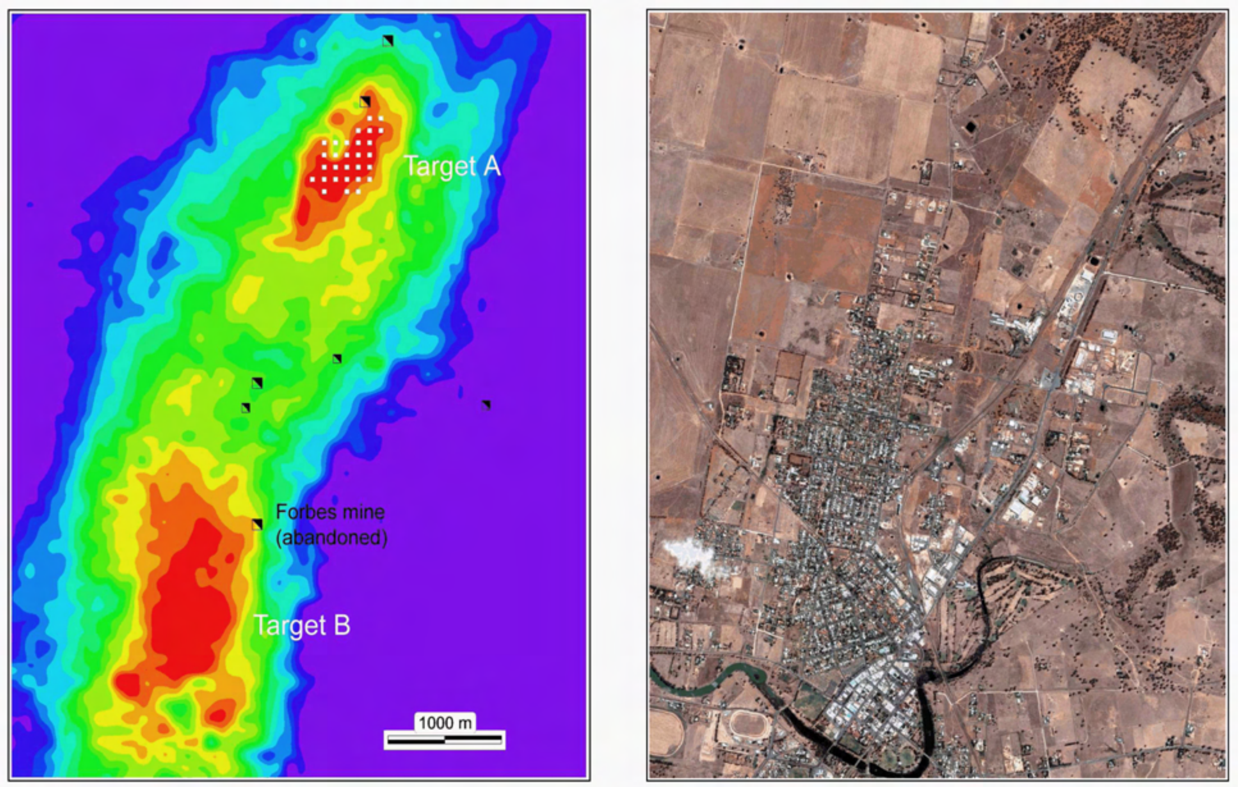

Figure 6: Left: Close-up of two targets in the Lachlan Fold Belt. The white squares mark a known deposit, which was one of several clusters of neural network training sites. Right: Satellite image of the area covered on the left. Unfortunately, target B lies under the historic gold mining town of Forbes.

Data mining results

The data mining process worked well in the LFB, where there is a first-class collection of modern data sets and a fine selection of known deposits that can be used for training neural networks. A strong, coherent statistical signal was consequently obtained from the exploration data as a whole. Of the eight exploration data sets that were incorporated in the study, structure, geology (i.e., lithology), and gravity were found to have the highest relevance. However, the other five data sets, geochemistry, radiometrics, magnetics, Landsat, and terrain all also made significant contributions.

Like the Walker Lane, the Lachlan Fold Belt is an active exploration area, so we can't show the overall target map that resulted from the data mining process. However, we can show an enlarged portion of the map. Figure 6 shows a pair of typical neural network targets. Target A in the north is actually a small known deposit, which was used as part of the neural network training set. The white squares are on a 100m grid, and represent the footprint of the known gold mineralization. The mineshaft symbols represent historic workings that produced minor amounts of gold.

The larger Target B in the south is typical of the targets that were generated by our data mining study. Note that this target is sharply defined and covers an area of roughly 2 square kilometres. A target like this would normally be quickly staked and followed up on the ground. Unfortunately, as can be seen in the satellite image on the right side of Figure 6, this target lies directly under the small country town of Forbes. This also happens to be where gold was first discovered by Harry Stephens, known as "German Harry", in June 1861. The discovery sparked a minor gold rush, which lasted for only a few years, reportedly because of difficult mining conditions. However, there is undoubtedly gold beneath the town of Forbes, which could possibly be extracted by modern mining methods were it not for the presence of the town with its population of 10,000 people.

Porcupine Gold Camp

The Porcupine Gold Camp, which is located in Eastern Canada (see Figure 7), is one of the most prolific gold mining districts in North America, with past production exceeding 60 million ounces of gold. Mining and exploration has been going on in this district since gold was first discovered near Timmins in 1907. Consequently, today there is a large number of known gold deposits and a huge accumulation of exploration data sets, ranging from surface and underground geology and geochemistry to airborne and satellite geophysics. For these reasons, this mature gold district was a natural candidate for a systematic data mining study.

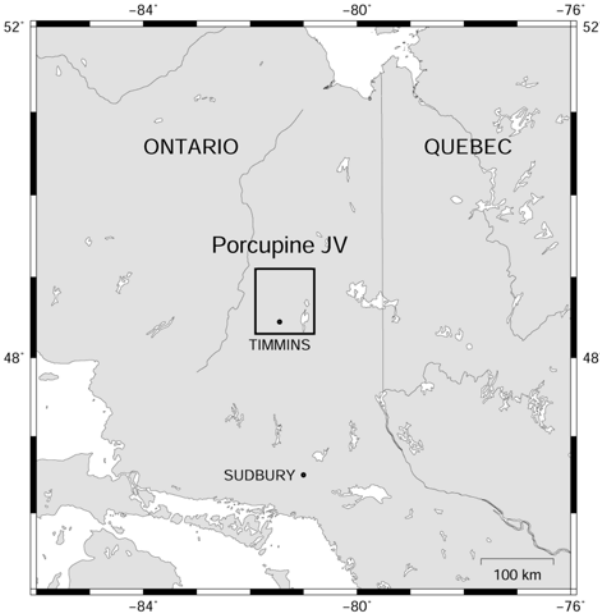

Figure 7: Map showing the location of the Porcupine Joint Venture study area in Eastern Canada.

The Abitibi greenstone belt, which contains the Porcupine Camp, is the largest greenstone belt in the world. The major gold camps within this greenstone belt are spatially associated with steeply dipping shear zones, such as the Destor-Porcupine Fault, a major east-west trending structure that extends for approximately 200 km. The Archaean rocks in the study area consist of mafic to felsic metavolcanics, metasediments, and a variety of granitoid intrusions.

The Abitibi greenstone belt is also is unique amongst greenstone belts of the Canadian Shield in that it has a high proportion of supracrustal rocks, has a generally low metamorphic grade, and contains a wide variety of mineral deposits, including volcanic-associated, massive sulphides (VMS) (e.g. Kidd Creek), komatiite-associated, Ni-Cu-platinum group metals (PGM), in addition to the large deposits of lode gold.

Most of the lode gold deposits in the Porcupine Camp are hosted by the Tisdale greenschist facies assemblage, which is bounded a few kilometres to the south by the Destor-Porcupine Fault. Some of the larger historic gold producers in the Camp include the Hollinger (19 Moz), Dome (16 Moz), McIntyre (11 Moz), Pamour (4 Moz) and the high-grade Hoyle Pond mine (2 Moz).

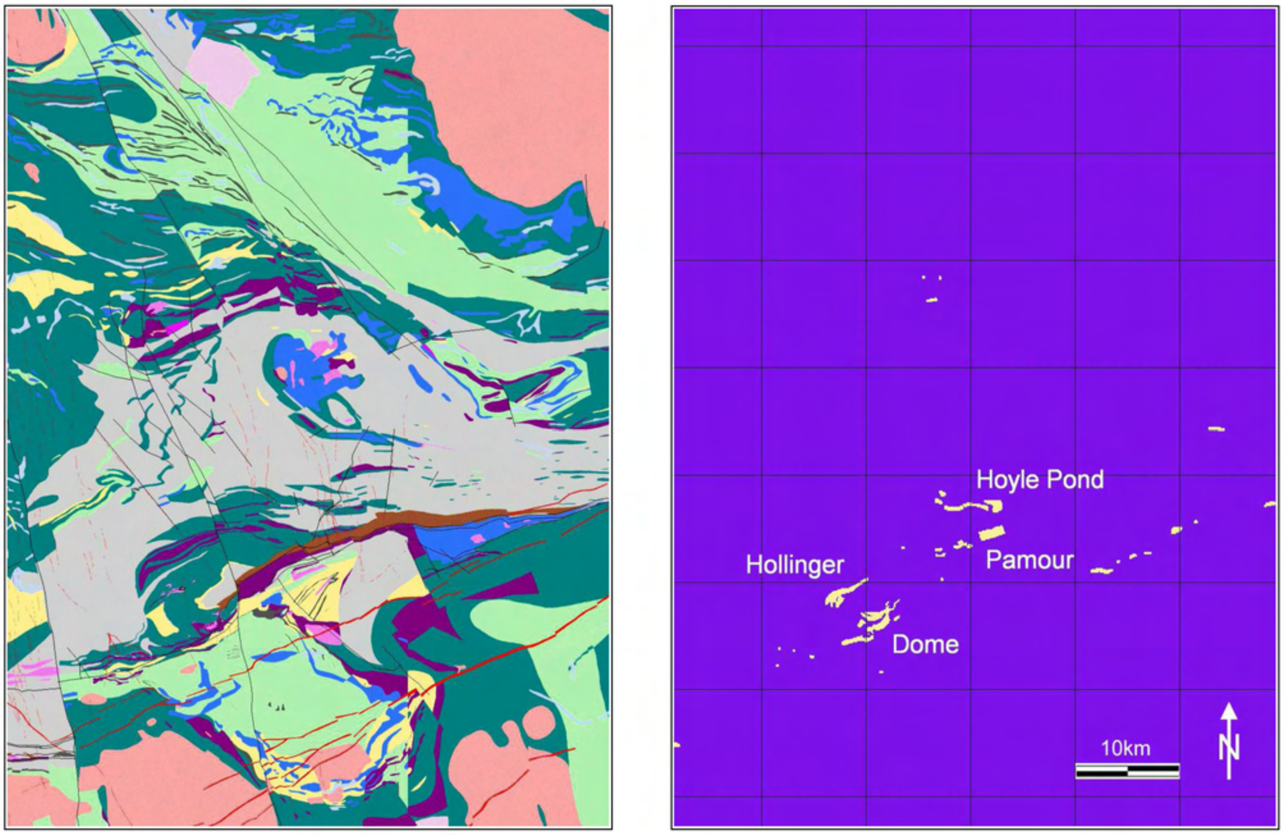

Figure 8: Left: Bedrock geology of the Porcupine study area, in which drift-covered areas to the north have been largely interpreted from the geophysics. Right: Footprints of the known gold deposits in this area.

Exploration data sets

As can be seen in Figure 7, the Porcupine study covered a much smaller area, roughly 8,000 square kilometres, than the first two case studies. Our data mining study of this area was commissioned by Placer Dome, who in 2005 were still managing what is known as the Porcupine Joint Venture (PJV). The PJV, which is currently a joint venture between Goldcorp Inc. and Kinross Gold Corporation, has landholdings of some 37,000 hectares around the major mines and is actively exploring for more gold. The exploration data sets that were made available to us in the Porcupine study were

Regional geology
Regional structure
SRTM elevation data
Airborne magnetics
Airborne Geotem survey
Ground and airborne gravity
Known deposit footprints

The geologic map of the Porcupine area was based on the 1:100,000 scale geological compilation of the Timmins area, released in digital form by the Ontario Geological Survey, and described by Ayer and Trowell (1998). Part of this map is shown in Figure 8. In this compilation, the rocks have been lumped into 15 principal lithologies. It is worth noting that this geological interpretation was heavily dependent on aeromagnetics and gravity, particularly in areas of thick glacial overburden to the north.

The magnetics were flown at a close line spacing and low terrain clearance, resulting in a high quality data set. The gravity map used for our data mining study was based on a blend of older ground gravity data with a recent airborne gravity survey. Before splicing these data, the ground gravity data were first upward continued to the mean flying height of the airborne survey, which was 253m AGL. Also available to us were the Shuttle Radar Topography Mission (SRTM) data, and a time-domain Geotem survey, which had originally been flown in 1987, and had been reprocessed more recently and made available by the Ontario Geological Survey.

In addition to these exploration data sets, the footprints of all the known gold deposits in the Porcupine camp were carefully established by the PJV staff. These footprints, which are shown on the right side of Figure 8, were based on historic mining records and geological interpretation, and were positioned as carefully as possible in plan view. It is important to note that these are not just showings, but significant gold occurrences with potentially economic grades.

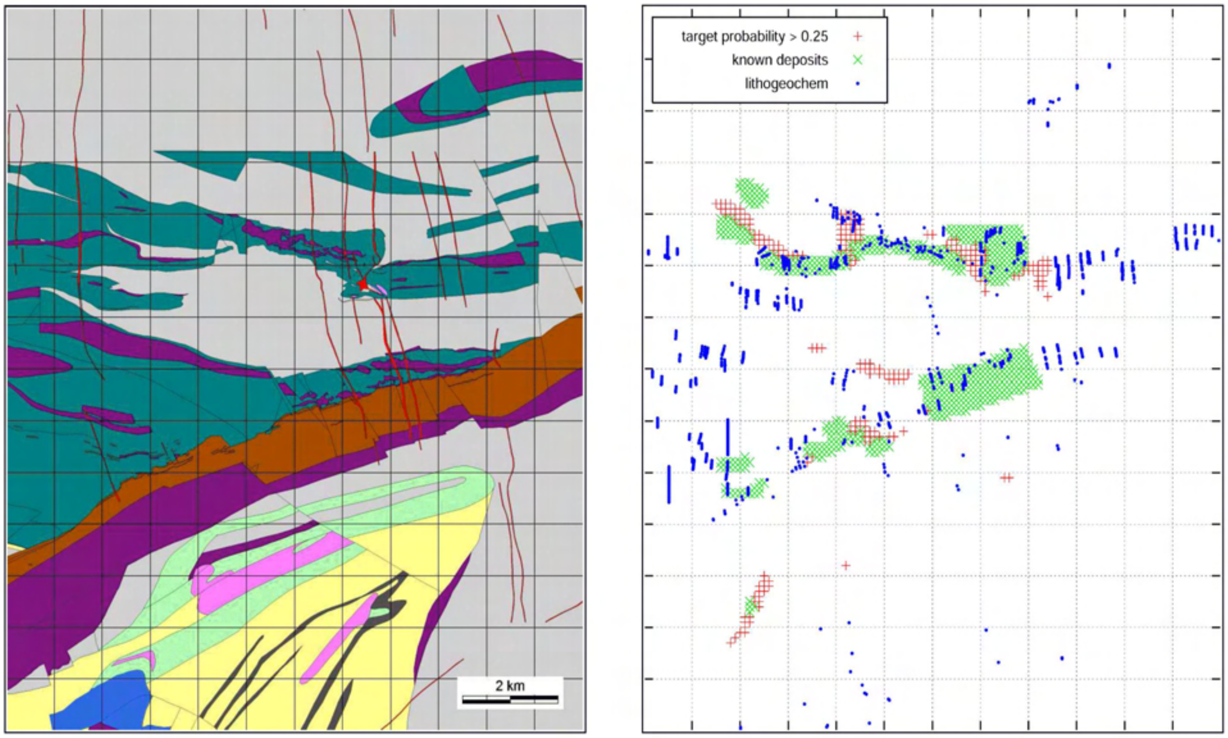

Figure 9: Left: Close-up of the geology in part of the Porcupine Camp. Right: Neural network target map corresponding to the same area. The green crosses mark known deposits, while the red crosses mark targets with a projected 1 in 4 probability of containing a gold deposit. The blue circles mark prior drill holes which have been analyzed for lithogeochemistry and gold concentration.

Data mining results

The Porcupine Gold Camp was found to be well suited to this type of objective analysis. Figure 9 shows a close-up of the results that were obtained in one part of the camp. The left side of this figure shows the geology, and the right side shows the corresponding target map resulting from our data mining study. On this map, the known deposits are marked in green, and targets exceeding a probability threshold of 0.25 (in other words, a projected one-in-four chance of containing a gold deposit) are marked in red. Similar maps can readily be produced for different probability thresholds.

The blue circles indicate drill holes and other sample sites that had previously been analyzed for their lithogeochemistry and gold concentrations. These data were not input to the data mining process and therefore provide a partial check on our results. It can be seen that there are several targets that do not appear to have been drill tested. It is also noteworthy that only a fraction of the grid nodes around the known deposits have received high scores. In other words, the data mining process is indicating where not to look as well as where the probability of finding gold is high.

An interesting experiment was conducted by Dr. Cliff Saunders, an independent consultant working for Kinross Gold Corporation, which has a (49%) minority interest in the Porcupine Joint Venture. As a statistical check on our results, he selected a number of test sites from the data mining target map, at some of which the data mining process predicted gold, and at others where the process predicted there should be no gold. Then, with the help of the PJV geological staff, he examined the drill logs and other geological information from these locations.

The results of this classical null hypothesis test can be briefly summarized as follows (C. Saunders, personal communication, 2006):

- At four sites where the networks predicted gold, and there was prior close-spaced drilling information, two were found to contain economic gold deposits. The other two sites did not contain economic gold; however the geologists felt both were good places to look.
- At five sites where the networks predicted gold, but there were no drill data, one site was judged to be prospective, and one not to be prospective, on the basis of the local stratigraphy. At a third site, sparse drilling had encountered a graphitic horizon but no gold. Unfortunately, little is known about the other two sites, one of which lies beneath a tailings pond and the second beneath a smelter plant. It would be ironic if these sites later prove to contain gold.
- At four sites where the network predicted no gold, and there had been some previous drilling, one site actually contained a small gold deposit, which we had not been previously told about. That must therefore be classed as a false negative. At all the other three sites, extensive drilling had encountered no gold, which confirmed the data mining predictions.

Overall, therefore, the data mining process performed well and gave satisfactory results. Apart from the one miss of a small economic deposit, and the couple of sites that were considered prospective by the network but were deemed unfavourable by the geologists, the process achieved a better than one-in-four success rate in Saunders' survey. This is a remarkably high score when one considers that, whilst all the right patterns and geological conditions may be present, in nature there still may be no economic gold found. In the words of a well-known, 17th-century bard: “All, as they say, that glitters is not gold.”

CONCLUSIONS

Modern digital data sets embody huge amounts of exploration data. Not only are such data sets now commonly available, each individually contains far more information than can be assimilated by the unaided human interpreter. Furthermore, critical indicators of prospectivity almost certainly lie in the subtle inter-relations between data sets, as much as in any single layer individually. These relations can only be extracted systematically by computer-based multivariate statistical techniques.

The neural network model described here is capable of analysing the full information content of the data, including its multivariate components. The output of the model is a precise numerical estimate of the probability of mineral occurrence at any given location. Target generation is highly specific, and capable of discriminating to within a few hundred meters in an area many hundreds of kilometres on a side. Target maps are bold and, consequently, easily testable.

The accompanying case studies illustrate these features. In particular, it has been shown that the approach can be applied successfully to search for a variety of ore deposit types, in a variety of geological environments, and at a variety of map scales. The examples we have presented range from Archaean lode gold and Ordovician porphyry gold to Tertiary volcanic-hosted gold deposits, and from regional to mining camp scales.

At these scales, the third dimension, depth, is barely significant, so these were all essentially 2-D studies. However, since the neural networks were not told any information about location and were simply given the patterns associated with the training sites, it would be a straightforward step to extend the data mining approach to an underground or 3-D situation.

ACKNOWLEDGEMENTS

The authors would like to thank Newmont Mining Corporation for supporting the early development of the methods described here and for permission to show some of the Walker Lane results. We should also like to thank Rimfire Minerals Corporation and Goldcorp Inc., respectively, for permission to show examples from the Lachlan and Porcupine data mining studies.

Finally, we should like to thank Kinross Gold Corporation and Dr. Cliff Saunders (www.tooserious.com) for their support, and for providing the results of his statistical test of the Porcupine case study.

REFERENCES

Ayer, J. A. and N. F. Trowell, 1998, Geological compilation of the Timmins area, Abitibi greenstone belt: Ontario Geological Survey, Preliminary Map P3379, scale 1:100,000.

Barnett, C. T. and P. M. Williams, 2006, Mineral exploration using modern data mining techniques, in Doggett, M. D. and J. R. Parry, eds., Wealth Creation in the Minerals Industry, Special Publication Number 12, chapter15, 295-310, Society of Economic Geologists, Inc.

Bonham-Carter, G. F., 1994, Geographic information systems for geoscientists: Modelling with GIS: Pergamon.

Bougrain, L., M. Gonzalez, V. Bouchot, D. Cassard, A. L. W. Lips, F. Alexandre, and G. Stein, 2003, Knowledge recovery for continental-scale mineral exploration by neural networks: Natural Resources Research, 12, 173-181.

Brown, W. M., T. D. Gedeon, D. I. Groves, and R. G. Barnes, 2000, Artificial neural networks: a new method for mineral prospectivity mapping: Australian Journal of Earth Sciences, 47, 757-770.

Ewers, G. R., N. Evans, M. Hazell, and B. Kilgour, 2002, OZMIN mineral deposits database [digital dataset]: Canberra, The Commonwealth of Australia, Geoscience Australia.

Good, I. J., 1950, Probability and the weighing of evidence: Charles Griffin & Company Limited.

Groves, D. I., R. J. Goldfarb, C. M. Knox-Robinson, J. Ojala, S. Gardoll, G. Y. Yun, and P. Holyland, 2000, Late kinematic timing of orogenic gold deposits and significance for computer-based exploration techniques with emphasis on the Yilgarn Block, Western Australia: Ore Geology Reviews, 17, 1-38.

Harris, D., L. Zurcher, M. Stanley, J. Marlow, and G. Pan, 2003, A comparative analysis of favorability mappings by weights of evidence, probabilistic neural networks, discriminant analysis, and logistic regression: Natural Resources Research, 12, 241-255.

Jennings, C. W., 1985, An explanatory text to accompany the 1:750,000 scale fault and geologic maps of California: California Division of Mines and Geology Bulletin 201, Appendix D, 125-197.

Liu, S. F., O. L. Raymond, A. J. Retter, D. Phillips, A. Kernich, C. Macgregor, and D. S. Percival, 2005, Surface geology of Australia 1:1,000,000 scale, New South Wales [digital dataset]: Canberra, The Commonwealth of Australia, Geoscience Australia.

Stewart, J. H. and J. E. Carlson, 1978, Geologic map of Nevada, 1:500,000 scale: U.S. Geological Survey, Nevada Bureau of Mines and Geology.

West, G. F., 1997, Progress in electrical and electromagnetic exploration techniques: Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 483-488.