*Nonrenewable Resources, Vol. 6, No. 1, 1997*

# Classification of Mineral Deposits into Types Using Mineralogy with a Probabilistic Neural Network

**Donald A. Singer[1,3] and Ryoichi Kouda[2]**

*Received August 15, 1996; accepted September 9, 1996*

In order to determine whether it is desirable to quantify mineral-deposit models further, a test of the ability of a probabilistic neural network to classify deposits into types based on mineralogy was conducted. Presence or absence of ore and alteration mineralogy in well-typed deposits were used to train the network. To reduce the number of minerals considered, the analyzed data were restricted to minerals present in at least 20% of at least one deposit type. An advantage of this restriction is that single or rare occurrences of minerals did not dominate the results. Probabilistic neural networks can provide mathematically sound confidence measures based on Bayes theorem and are relatively insensitive to outliers. Founded on Parzen density estimation, they require no assumptions about distributions of random variables used for classification, even handling multimodal distributions. They train quickly and work as well as, or better than, multiple-layer feedforward networks. Tests were performed with a probabilistic neural network employing a Gaussian kernel and separate sigma weights for each class and each variable. The training set was reduced to the presence or absence of 58 reported minerals in eight deposit types. The training set included: 49 Cyprus massive sulfide deposits; 200 kuroko massive sulfide deposits; 59 Comstock epithermal vein gold districts; 17 quartz-alunite epithermal gold deposits; 25 Creede epithermal gold deposits; 28 sedimentary-exhalative zinc-lead deposits; 28 Sado epithermal vein gold deposits; and 100 porphyry copper deposits. The most common training problem was the error of classifying about 27% of Cyprus-type deposits in the training set as kuroko. In independent tests with deposits not used in the training set, 88% of 224 kuroko massive sulfide deposits were classed correctly, 92% of 25 porphyry copper deposits, 78% of 9 Comstock epithermal gold-silver districts, and 83% of six quartz-alunite epithermal gold deposits were classed correctly. Across all deposit types, 88% of deposits in the validation dataset were correctly classed. Misclassifications were most common if a deposit was characterized by only a few minerals, e.g., pyrite, chalcopyrite,and sphalerite. The success rate jumped to 98% correctly classed deposits when just two rock types were added. Such a high success rate of the probabilistic neural network suggests that not only should this preliminary test be expanded to include other deposit types, but that other deposit features should be added.

**KEY WORDS:** Probabilistic neural network; mineral deposit models; mineralogy; Bayes.

## INTRODUCTION

At present the only parts of mineral deposit models not based on subjective estimates are the grades and tonnages (Cox and Singer, 1986). Removal of the subjective element of estimating grade and tonnage distributions has resulted in significant benefits to resource assessments and exploration planning. These benefits exceed the costs even after considering the great effort required to construct proper grade and

[1] U.S. Geological Survey, 345 Middlefield Road, Menlo Park, California 94025. (e-mail: singer@mojave.wr.usgs.gov)

[2] Geological Survey of Japan, 1-1-3 Higashi, Tsukuba, Ibaraki-ken, 305 Japan. (e-mail: roy@gsj.go.jp)

[3] Correspondence should be addressed to Donald A. Singer, U.S. Geological Survey, 345 Middlefield Road, Menlo Park, California 94025. (e-mail: singer@mojave.wr.usgs.gov)

tonnage models (Singer, 1993). If it were possible to correctly classify a large proportion of deposits and occurrences into deposit types based on the kinds of information commonly available in the geologic literature, then a system could be built that would automatically screen large data files. In such a system, the necessary and sufficient information would exist to discriminate among deposit types. Extensions to this kind of system might serve as a basis for integrating geological, geophysical, and geochemical information for estimating and managing risk.

Barton (1986) provided estimates of the frequency of mineral occurrence by deposit type. His subjective estimates for over 150 minerals in about 80 deposit types were used by McCammon (1992) with subjective estimates of frequencies of rock types, ages, alteration, geophysical, and geochemical signatures in an attempt to classify deposits with a system called Prospector II. McCammon's test of this system (1992) resulted in 83% of 124 deposits correctly classed.

In expert systems like Prospector II, a human expert's knowledge, in the form of qualitative principals as perceived by the expert, is encoded. Performance of these systems depends on the quality of the expert's knowledge and the care taken in the representation of that knowledge. Such expert systems are desirable where the underlying model relationships or information are not known. Expert systems have difficulties where the experts are internally inconsistent or rely on inconsistent information.

Where information is available, inductive learning systems exist that can use preclassified samples as a training set to learn the appropriate classification rule. These learning systems can be very good classifying previously unseen samples, that is, at generalization. Examples of inductive learning systems are decision trees (Quinlan, 1986), artificial neural networks (Masters, 1995), and statistical pattern recognition (Fukunaga, 1990). Features of statistical pattern recognition such as probabilistic estimates of class membership and ability to handle contradictory examples are integral to probabilistic neural networks.

In order to determine whether it is desirable to quantify mineral deposit models further, a test of the ability of a probabilistic neural network to classify deposits into types based on a simple representation of mineralogy is conducted here. The study is relatively small in scale in that only the mineralogy in eight deposit types and 773 deposits are employed. The nature and sources of these data are discussed first. Following this, probabilistic neural networks and their implementation in this study are discussed. Classification of deposits into types by the neural network is tested in the next section. Finally, classification errors are examined and possible improvements identified.

## THE DATA

Information on the mineralogy of mineral deposits varies widely in quantity and quality. Depending on the purpose of a study and its researcher's interest, a report on a mineral deposit might contain a detailed listing of alteration minerals and a mention of unnamed sulfide and sulfosalt minerals, a detailed list of ore minerals and mention of alteration in broad terms, a complete list of all minerals, or a sparse list of minerals. In some studies, the author attempted to list the relative or absolute amounts of each mineral. Unfortunately, these attempts were not common and frequently not comparable with many other reports because of different standards. Thus, it was decided to use only the presence or absence of minerals in our study.

Both ore and alteration minerals were recorded for this study. Rock forming minerals such as varieties of quartz, feldspars (except adularia), and amphiboles were not recorded, even if they locally represent alteration. General statements about mineralogy such as "clays," "carbonates," or "phyllic alteration" present were ignored because multiple minerals were possible. These decisions were made to keep minerals not related to the mineral deposit type out of the analysis, to reduce the number of minerals considered, and to keep the data objective. Even with these restrictions and the exclusion of clearly single case listings, the presence or absence of 132 minerals was recorded. Closely related minerals such as the tellurides, manganese oxide minerals, anhydrite-gypsum, and enargite-luzonite were combined to reduce the number of minerals to 109. To further reduce the number of minerals considered, the analyzed data were restricted to minerals present in at least 20% of at least one deposit type used in the study. An advantage of this restriction is that rare occurrences of minerals cannot dominate the results. The data were reduced to the presence or absence of 58 reported minerals in eight deposit types (table 1).

The data from eight deposit types were collected and divided into a training set containing 506 deposits and a validation set consisting of 267 deposits. The training set contained: 49 Cyprus-type massive sulfide deposits (Singer, 1986 a); 200 kuroko massive sulfide deposits (Singer, 1986 b); 59 Comstock epithermal

**Table 1.** List of Minerals Used in the Training and Validation Data

| | | | |
|---|---|---|---|
| adularia | alunite | anhydrite/gypsum | ankerite |
| apatite | argentite | arsenopyrite | azurite/malachite |
| barite | biotite | bornite | cerargyrite |
| cerussite | chalcocite/digenite | chalcopyrite | chlorite |
| chrysocolla | copper | covellite | cuprite/tenorite |
| electrum | engarite | epidote | famatinite |
| fluorite | galena | garnet | goethite/limonite |
| gold | graphite/organics | hematite/specularite | jarosite |
| jasper | kaolinite/illite | luzonite | magnetite |
| manganite/psilomelane/pyrolucite/wad | | marcasite | molybdenite |
| muscovite/sericite | pearceite | polybasite | proustite |
| pyrargyrite | pyrite | pyrophyllite | pyrrhotite |
| rhodochrosite | rhodonite | siderite | silver |
| sphalerite/wurtzite | stephanite | stibnite | sulfur |
| tellurides/calverite/hessite/petzite/sylvanite | | tennanite | tetrahedrite |

vein gold deposits (Mosier and others, 1986 c); 17 quartz-alunite epithermal gold deposits (Berger, 1986); 25 Creede epithermal gold deposits (Mosier and others, 1986 d); 28 sedimentary-exhalative zinc-lead deposits (Briskey, 1986); 28 Sado epithermal vein gold deposits (Mosier and others, 1986 a); and 100 porphyry copper deposits (Cox, 1986). All Cyprus-type and kuroko training and validation data were from Mosier and others (1983), the Comstock, Creede, quartz-alunite, and Sado data were compiled by Mosier and others (1986 b). The sedimentary-exhalative Zn–Pb, porphyry copper, and some quartz-alunite data were compiled for this study.

## THE PROBABILISTIC NEURAL NETWORK

The goal here is to be able to make an estimate of the probability that an unknown mineral deposit belongs to a given deposit type. Standard statistical classification methods assume some knowledge of the distribution of the variables used to classify. Typically a multivariate normal distribution is assumed and the training data are used to estimate the means and variances. Large deviations from normality or multimodal distributions cause these methods to fail. Neural networks can typically handle very complex distributions. The three-layer feedforward network (Singer and Kouda, 1996) is an excellent classifier (Masters, 1995); however, it trains very slowly and does not produce probabilities.

Probabilistic neural networks were designed to be classifiers. If we know the true probability density function, $f_k(x)$, for all populations, then there is a Bayes optimal decision rule for classifying unknown sample $x$ into population $i$:

$$p_i c_i f_i(x) > p_j c_j f_j(x) \qquad (1)$$

for all populations $j$ not equal to $i$. Where $p_k$ is the prior probability of the general class $k$, and $c_k$ is the cost associated with misclassification of population $k$. Under these conditions, a Bayes decision rule will minimize the expected cost of misclassification. The problem is that we do not know the true probability density function, $f_k(x)$. Standard statistical classification methods, such as discriminant analysis, typically assume that the variables follow a multivariate normal distribution or that the nearest neighbor is the appropriate class regardless of the density of other samples near the unknown.

The development by Parzen (1962) of a general way to estimate a univariate probability density function from a random sample, even when the parent density function is unknown, provides a necessary tool to free us from these constraints. Parzen's estimator is essentially a sphere-of-influence weighting function, commonly called a kernel, and the scaling parameter, $\sigma$, controls the width of the area of influence. The weighting function has its largest values at sample points and decreases toward zero as the distance increases. For a single population of $x$ which has sample size $n$, the estimated density function for the population is:

$$f_k(x) = \frac{1}{n\sigma} \sum_{i=0}^{n-1} W\left(\frac{x - x_i}{\sigma}\right) \qquad (2)$$

For this study separate sigma weights ($\sigma$) were used for each class and each variable and a Gaussian kernel was used for the weighting ($W$) function (Masters, 1995). The choice of the Gaussian function is based on its excellent performance and has nothing to do with assumptions of normal distributions. Specht (1990) constructed a neural network form of Parzen's estimation procedure. In this study, the algorithms for a probabilistic neural network developed by Masters (1995) were employed. Masters' algorithms find the scale factors, $\sigma$, that minimize the error of misclassification of the training data using the standard statistical technique called jackknifing in which every case is sequentially held back from training.

Probabilistic neural networks require no assumptions about distributions of random variables used to classify; they even can handle multimodal distributions. They train quickly and as well as, or better than, multiple-layer feedforward networks. They have the ability to provide mathematically sound confidence levels and are relatively insensitive to outliers. Mathematically sound Bayesian confidence levels require that the classes are mutually exclusive and exhaustive (i.e., no case can possibly fall into more than one population and the training set encompasses all populations fairly). When these conditions exist, Bayes' Theorem can be used to compute the probability that an observation $X$ was the product of population $A$.

$$P[A|X] = \frac{f_a(X)}{\sum_k f_k(X)} \tag{3}$$

Each density estimate, $f_k(X)$, in the numerator and denominator of equation 3 could be multiplied by prior probabilities and cost constants, if desired. These features are not used in this study, however.

In many practical cases, the mutually exclusive and exhaustive class conditions might not exist. The unknown sample used in testing might be from a population different from any of the training classes. For example, if the mineralogy of a polymetallic replacement or polymetallic vein deposit were tested in the network developed in this study, Bayesian confidence estimates could not be properly computed. The neural network program will estimate the probabilities that the unknown deposit belongs to the deposit classes it has been taught; thus, careless , use of a neural network could lead to mistaken classifications.

## TESTING THE NEURAL NETWORK

Two datasets exist to test the neural network. The validation data, because it was not used in any training, is the proper dataset to test the efficiency of classification. Failures of proper classification of deposits used in training can provide important information about problems in the data or in the class identification. About 97% of the 506 deposits in the training data were classed properly.

The most common training problem was the incorrect classification of 27% of Cyprus-type massive sulfide deposits as kuroko massive sulfide deposits. If deposits having less than four minerals reported are excluded from the training and validation data, this misclassification rate drops to 17% of Cyprus-type massive sulfide deposits. In general, misclassifications were more common where the deposit had only a few minerals listed such as pyrite, chalcopyrite, and sphalerite; for several deposits only pyrite and chalcopyrite were listed.

Independent tests of previously unseen samples from the validation set gave the following results: 88% of 224 kuroko massive sulfide deposits were classed correctly (table 2); 92% of 25 porphyry copper deposits correct; 78% of nine Comstock epithermal gold-silver correct; and 83% of six quartz-alunite epithermal gold deposits correct. The one "misclassification" of a quartz-alunite epithermal Au deposit was Recsk, Hungary, which has parts that are porphyry copper, quartz-alunite epithermal gold, and skarn. The neural network classed Recsk as a porphyry copper with a probability of 0.56, the remaining probability (0.44) was assigned to quartz-alunite. Had the neural network been taught to recognize skarns, it might have distributed some of the probability to that class also. Although the misclassification of Recsk lowered the success rate statistics, it also suggests that the probabilistic neural network might be able to recognize mixed deposit types and various positions between end member deposit types.

Many classification errors are of kuroko deposits that were classed as epithermal or other deposit types that would not be confused with kuroko deposits if the geologic setting of the deposits were known. For example, just knowing whether there are marine mafic volcanic rocks or marine felsic to intermediate volcanic rocks near the deposits increases the correctly classified kuroko deposits in the validation set from 88% to 98% and decreased the training error rate for Cyprus-type deposits from 27% to 4%. The overall correct classification rate in the validation set increased from 88% to 98%.

Table 2. Confusion Matrix Showing the Number of Mineral Deposits Correctly (in **bold**) and Incorrectly Classified from the Validation Set

| | | True class | | | | |
|---|---|---|---|---|---|---|
| | Deposit type | Kuroko ms | Comstock Au–Ag | Qtz. alunite Au–Ag | Sed.-exhal. Zn–Pb | Porphyry Cu |
| | Cyprus ms | 10 | 0 | 0 | 0 | 0 |
| | Kuroko ms | **196** | 1 | 0 | 0 | 0 |
| | Comstock Au–Ag | 2 | **7** | 0 | 0 | 1 |
| Predicted | Qtz. alunite Au–Ag | 0 | 0 | **5** | 0 | 0 |
| Class | Creede Au–Ag | 4 | 1 | 0 | 0 | 1 |
| | Sed.-exhal. Zn–Pb | 8 | 0 | 0 | **3** | 0 |
| | Sado. Au–Ag | 1 | 0 | 0 | 0 | 0 |
| | Porphyry Cu | 3 | 0 | 1 | 0 | **23** |
| | Total | 224 | 9 | 6 | 3 | 25 |

## CONCLUSIONS

Correctly classifying 88% of the validation set deposits using only presence or absence of reported mineralogy is a remarkably good outcome. This is in comparison to 83% classified correctly by Prospector II (McCammon, 1992) which used mineralogy, rock types, ages, alteration, geophysical, and geochemical signatures. Because the original mineral deposit models were based on lithologies, mineral abundances, geochemistry, and other attributes, one might expect that a system using these features would have a very high success rate in classifying deposits into types. It is important to remember that the original estimates of frequencies of these features and the estimation of their importance in Prospector II were made subjectively. The high success rate of the neural network suggests that there is a clear improvement in correct classification when objective data are used in the deposit models. It also suggests a clear advantage of probabilistic neural networks over expert systems when information is available. In the authors' view, not only should this preliminary test be expanded to include other deposit types, but other deposit features should be added. Just knowing whether there are marine mafic volcanic rocks or marine felsic to intermediate volcanic rocks near the deposits increases the correctly classified deposits in the validation set from 88% to 98%, a significant improvement over the Prospector II results.

These results suggest that it is possible to correctly classify a large proportion of deposits and occurrences into deposit types based on the kinds of information commonly available in literature. Extensions to this kind of system might serve as a basis for integrating geological, geophysical, and geochemical information for estimating the probabilities of specific deposit types existing within a given area.

## ACKNOWLEDGMENTS

B. R. Berger provided data on a number of epithermal quartz-alunite Au deposits which allowed more thorough testing of the system.

## REFERENCES

Barton, P. B., 1986, Mineralogical index, *in* Cox, D. P., and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 318–348.

Berger, B. R., 1986, Descriptive model of epithermal quartz-alunite Au, *in* Cox, D. P., and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 158.

Briskey, J. A., 1986, Descriptive model of sedimentary exhalative Zn-Pb, *in* Cox, D. P., and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 211.

Cox, D. P., 1986, Descriptive model of porphyry Cu, *in* Cox, D. P., and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 76.

Cox, D. P., and Singer, D. A., eds., 1986, Mineral deposit models: U.S. Geological Survey Bulletin 1693, 379 p.

Fukunaga, Keinosuke, 1990, Introduction to statistical pattern recognition: second edition, Academic Press, Inc., San Diego, CA, 591 p.

McCammon, R. B., 1992, Numerical mineral deposit models, *in* Bliss, J. D., ed., Developments in deposit modeling U.S. Geological Survey Bulletin 2004, p. 6–12.

Masters, Timothy, 1995, Advanced algorithms for neural networks: a C++ sourcebook: John Wiley & Sons, New York, 425 p.

Mosier, D. L., Singer, D. A., and Salem, B. B., 1983, Geologic and grade-tonnage information on volcanic-hosted copper-zinc-lead massive sulfide deposits: U.S. Geological Survey Open-File Report 83–89, 78 p.

Mosier, D. L., Berger, B. R., and Singer, D. A., 1986 a, Descriptive model of Sado epithermal veins, *in* Cox, D. P. and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 154.

Mosier, D. L., Menzie, W. D., and Kleinhampl, F. J., 1986 b, Geologic and grade-tonnage information on Tertiary epithermal precious- and base-metal vein districts associated with volcanic rocks: U.S. Geological Survey Bulletin 1666, 39 p.

Mosier, D. L., Singer, D. A., and Berger, B. R., 1986 c Descriptive model of Comstock epithermal veins, *in* Cox, D. P. and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 150.

Mosier, D. L., Sato, Takeo, Page, N. J., Singer, D. A., and Berger, B. R., 1986 d, Descriptive model of Creede epithermal veins, *in* Cox, D. P., and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 145–146.

Parzen, E., 1962, On estimation of a probability density function and mode: Annals of Mathematical Statistics, v. 33, p. 1065–1076.

Quinlan, J. R., 1986, Induction of decision trees: Machine Learning, v. 1, n. 1, p. 81–106.

Singer, D. A., 1986 a, Descriptive model of Cyprus massive sulfide, *in* Cox, D. P. and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 131–133.

Singer, D. A., 1986 b, Descriptive model of kuroko massive sulfide, *in* Cox, D. P. and Singer, D. A., eds., Mineral deposit models: U.S. Geological Survey Bulletin 1693, p. 189–194.

Singer, D. A., and Kouda, Ryoichi, 1996, Application of a feedforward neural network in the search for kuroko deposits in the Hokuroku District, Japan: Mathematical Geology, p. 1017–1023.

Singer, D. A., 1993, Development of grade and tonnage models for different deposit types, *in* Kirkham, R. V., Sinclair, R. V., Thorpe, W. D., and Duke, J. M., eds., Mineral deposit modeling: Geological Association Canada Special Paper 40, p. 21–30.

Specht, Donald, 1990, Probabilistic neural networks: Neural Networks, v. 3, p. 109–118.