

## A comparative study of three data-driven Mineral Potential Mapping techniques

G. Caumon<sup>1</sup>, J.M. Ortiz<sup>2</sup>, O. Rabeau<sup>3</sup>

<sup>1</sup> CRPG-CNRS / Nancy Université, Nancy, France

<sup>2</sup> Departamento de Ingeniería de Minas, Universidad de Chile, Santiago

<sup>3</sup> UR Technologie Minérale, UQAT, Val d'Or, Qc, Canada

Corresponding author: [Guillaume.Caumon@ensg.inpl-nancy.fr](mailto:Guillaume.Caumon@ensg.inpl-nancy.fr)

**ABSTRACT** : High prices of mineral resources and progresses in models of mineral deposits maintain a high activity in mining exploration. Many data integration techniques exist to assess the mining potential of a region, and it may be difficult for the practitioner to select the most appropriate. This paper compares two simple data-driven techniques: logistic regression and weights of evidence (WoE), and proposes a cross-validation technique to depart from conditional independence in WoE to better manage data redundancy. All three methods are applied to gold potential mapping in a 3D geo-model of the Duparquet region located on the Porcupine Destor fault in the Abitibi subprovince, Canada.

**KEYWORDS** : Probability Theory, Weights of Evidence, Logistic Regression, Conditional Independence.

### 1. Introduction

The challenge of integrating data from different sources is very high in subsurface characterization efforts. Many techniques have been described to assess the probability of occurrence of an economic mineralization at some location given several observations (or layer maps in GIS terms). These techniques are classically split into data-driven methods, wherein the relationships between variables are calibrated from the available observations, and expert-driven methods, wherein expert knowledge is incorporated in the prediction process (Bonham-Carter, 1994). In this paper, we discuss the underlying assumptions of two classical techniques used in mineral potential mapping (Agterberg et al, 1989; Bonham-Carter 1994; Agterberg and Bonham-Carter, 1999): the Logistic regression, and the Weights of Evidence methods. We then propose a means to account for redundancy in the Weights of Evidence (Section 2), and apply these methods to a 3D gold exploration problem in the Duparquet area, Abitibi, Canada (Section 3).

### 2. Overview of the considered methods

All predictive methods in mineral potential mapping aim at evaluating the probability  $P(Y | X_1 = x_1, \dots, X_K = x_K)$  of a mineralization  $Y$  occurring at some location, given  $K$  secondary variables (or covariates)  $\{X_1, X_2, \dots, X_K\}$  representing the observations.

#### 2.1. Logistic regression (LR)

This probability can be estimated from a discrete set of events  $X_k = x_k$  (noted  $X_k$  for conciseness) using the logistic regression equation:

$$P(Y | X_1, X_2, \dots, X_K) = \frac{e^{a_0 + a_1 X_1 + \dots + a_K X_K}}{1 + e^{a_0 + a_1 X_1 + \dots + a_K X_K}} = \frac{e^{a^T x}}{1 + e^{a^T x}} \quad (1)$$

The  $K$  coefficients  $\mathbf{a}^T = [a_1, \dots, a_K]$  of the linear combination are evaluated from  $N$  observed data points  $\{y_n, x_{k,n}\}$ ,  $n=1, \dots, N$  and  $k=1, \dots, K$ . The values  $\mathbf{y} = \{y_1, \dots, y_N\}$  correspond to the observed binary outcomes (0 or 1) of the  $N \times K$  observations  $\mathbf{X} = \mathbf{x}_n^T = [x_{n1}, \dots, x_{nK}]$ ,  $n=1, \dots, N$ . Because the variance of the regression error depends on the values of the data  $x_n$ , classical least-square estimation cannot be applied to obtain the logistic regression parameters  $\mathbf{a}$ . Instead, the unknown coefficients  $\mathbf{a}$  are chosen so as to maximize the data likelihood. Under the (strong, and questionable) hypothesis that all data points  $\{y_n, x_{k,n}\}$ , are independent from each other, this likelihood is expressed as:

$$l(\mathbf{a}) = \prod_{n=1}^N \left( \frac{e^{\mathbf{a}^T \mathbf{x}_n}}{1 + e^{\mathbf{a}^T \mathbf{x}_n}} \right)^{y_n} \left( \frac{1}{1 + e^{\mathbf{a}^T \mathbf{x}_n}} \right)^{1-y_n} \quad (2)$$

Finding the maximum likelihood is a non-linear optimization problem, which can be solved numerically by setting the derivatives of the log-likelihood to 0. Komarek (2004) recently investigated several numerical methods to find the maximum likelihood in an efficient manner; his code, freely available at <http://komarix.org/ac/lr/lrtrirls>, was used in this paper. As shown by Figure 1, the logistic regression model is parsimonious, and does not include cross terms between variables, meaning that iso-probability (hyper)lines on the regression (hyper)surface are defined by a linear equation on  $\mathbf{X}_1, \dots, \mathbf{X}_K$ . A better fit to experimental data could probably be achieved using cross terms in the regression expression. For instance, in the bivariate case, the regression equation would become:

$$P(Y | X_1, X_2, \dots, X_K) = \frac{e^{a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2}}{1 + e^{a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2}} \quad (3)$$

Generalizing equation (3) to  $K$  predictor variables could be interesting, but would also increase exponentially the computational cost of logistic regression, since the number of parameters would increase from  $K$  to  $2K - 1$ .

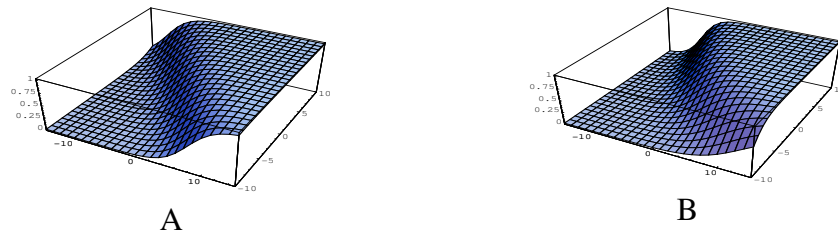


Fig. 1. A: example of a logistic regression function in the bivariate case. Note that iso-probability curves are straight lines in parameter space. B: an alternative logistic regression function that includes a cross term ( $x_1 \times x_2$ ) could provide more flexibility in the logistic regression surface.

## 2.2. Weights of Evidence (WoE)

Instead of using an analytical model to derive probabilities, another possible approach is to directly get probabilities from some observed data set. Such a computation relies on the definition of classes for all variables to compute frequency tables, which are later interpreted in terms of probabilities. The Weights of Evidence (WoE) is one such method; it has become a standard in mineral potential mapping application [Agterberg et al, 1989; Bonham-Carter, 1994; Agterberg and Cheng, 2002].

The WoE is a Bayesian inversion method based on the definition of conditional probabilities of the presence ( $Y$ ) or absence ( $\bar{Y}$ ) of mineralization given some *binary* indicator variables  $X_k$ . Continuous evidential variables can fit into this framework using several binary classes using a set of thresholds (indicator approach in geostatistics). The Weights of evidence is based on Bayes formula:

$$P(Y | X_1, \dots, X_K) = \frac{P(X_1 | Y, X_2, \dots, X_K) P(X_2 | Y, X_3, \dots, X_K) \dots P(X_K | Y) P(Y)}{P(X_1, X_2, \dots, X_K)} \quad (4)$$

Eq. (4) can be simplified by assuming conditional independence of all variables  $X_k$  given  $Y$ :

$$P(Y | X_1, \dots, X_K) = \frac{P(X_1 | Y) P(X_2 | Y) \dots P(X_K | Y) P(Y)}{P(X_1, X_2, \dots, X_K)} \quad (5)$$

The hypothesis of conditional independence assumes that the redundancy between variables is ignored. Assume we have a map  $X_1$  of altered rocks and a map  $X_2$  of locations close to faults. Conditional independence means that given the presence of a mineralization  $Y$ , alteration  $X_1$  and distance to faults  $X_2$  can be assumed to be independent. This is true if, over all mineralized locations, the probability  $P(X_1 | Y)$  of being in an altered rock is the same as the probability  $P(X_1 | Y, X_2)$  of being in an altered rock, given that we are also in a fault neighborhood. This hypothesis may be violated, e.g. if alteration fluids responsible for the mineralization circulated away from faults.

### 2.3. Weights of Evidence accounting for data redundancy

One way to depart from the conditional independence hypothesis is to model data redundancy in a similar fashion as Journel (2002) and Krishnan and Journel (2004): the conditional probability  $P(X_1 | Y, X_2)$  can be written, for some real  $\mu$  value:

$$P(X_1 | Y, X_2) = P(X_1 | Y)^\mu \quad (6)$$

Then, Equation (4) can be strictly written as:

$$P(Y | X_1, \dots, X_K) = \frac{P(X_1 | Y)^{\mu_1} P(X_2 | Y)^{\mu_2} \dots P(X_K | Y) P(Y)}{P(X_1, X_2, \dots, X_K)} \quad (7)$$

The parameter  $\mu$  could be computed directly from the calibration data set using eq. (6). However, such an evaluation would simply amount estimating  $P(Y | X_1, \dots, X_K)$  directly from the observed frequency tables. Instead, we propose to adjust this  $\mu$  parameter through cross validation on the available data. The  $\mu$  value is computed so as to minimize, at all observed mineralizations, the mean squared error  $\{1 - P(Y | X_1, \dots, X_K)\}^2$  obtained with eq. (7).

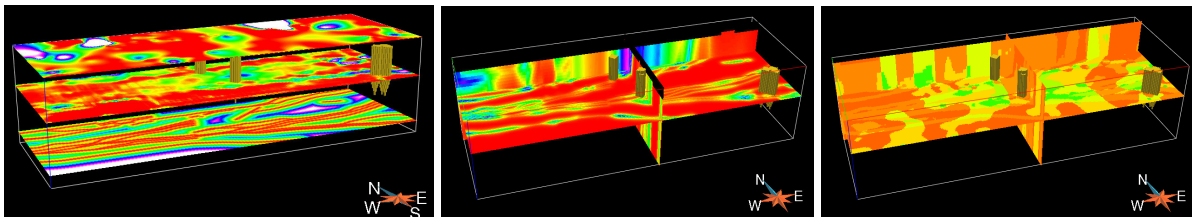


Fig. 2. Left: input data set (top section: mud lake As grade; middle section: magnetic susceptibility; bottom section: distance to faults). Center: probability cube obtained by logistic regression. Right: evidence map obtained with the WoE implementation of Marcus Apel ([www.geo.tu-freiberg.de/~apelm](http://www.geo.tu-freiberg.de/~apelm)). Vertical scaling  $\times 3$ .

### 3. Application to the Dupartquet Region, Abitibi.

The described methods have been applied to gold potential mapping in the Porcupine-Destor area in the Abitibi Greenstone Belt, Canada (Fallara et al., 2004, 2006). Three evidential variables were selected for prediction on a  $11.4 \times 4 \times 1$  km volume, according to the metallogenic processes described by Legault et al. (2003, 2004): the distance to the E/W family of faults, the As grade in lake sediments and the magnetic susceptibility obtained after geophysical inversion. The As grade was smoothly interpolated (Mallet, 1997) on the topography, then propagated in depth by vertical translation. Data and results obtained are summarized in Fig. 2.

#### 4. Discussion

While LR makes a direct estimation from the data, the WoE method allows for some modulation of the estimates through prior information, or through letting experts directly input the weights / likelihoods. Therefore, the Weights of Evidence method gives more *subjective control* to correct for bias.

LR considers the calibration data at once in the  $(K + 1)$ -dimensional space, while the WoE method consider  $K$  bivariate relations, which only provides a partial view of the problem. A global view of all data points as in the logistic regression is very interesting because it does not raise the problem of recombining the bivariate relations into one multivariate model. The down side is that locations where some variables are missing should be rigorously excluded from the calibration set (see Agterberg and Bonham-Carter (1999) for possible strategies).

In the practice of mineral potential mapping, it is rare to account rigorously for spatial redundancy when integrating several evidence maps. Ideally, these techniques should be applied on declustered data only. More research is required to account for multi-point statistics in declustering. The effects of volume support should also be investigated, since the resolution of evidence variables is generally variable from one map to another.

*Acknowledgments : The authors would like to thank the Codelco Chair on Ore Reserve Estimation at the University of Chile, the members of the Gocad Consortium and the Ministry of Natural Resources of Québec for their support. We also thank Marcus Apel for providing his implementation of the Weights of evidence and EarthDecision for the Gocad software.*

#### REFERENCES

- F. P. Agterberg, G. F. Bonham-Carter, D. F. Wright and Q. Cheng, 1989. Weights of evidence modeling and weighted logistic regression for mineral potential mapping. In *J.C. Davis and U.C Herzfel, eds., Computers in Geology 25 years of progress*. Oxford University Press, New York. 13–32.
- F. P. Agterberg and G. F. Bonham-Carter, 1999. Logistic regression and weights of evidence modeling in mineral exploration. In *Proc. APCOM'99, Golden, Colorado*. 583-590.
- F. Agterberg and Q. Cheng, 2002. Conditional independence test for weights-of-evidence modeling. *Natural Resources Research*, 11(4):249-255.
- G. F. Bonham-Carter, 1994. *Geographic Information Systems for Geoscientists: Modelling with GIS*. Computer Methods in the Geosciences. Pergamon Press, New York. 414 p.
- F. Fallara, M. Legault, L. Cheng, O. Rabeau, and J. Goutier, 2004. Modèle 3d géo-intégré d'un segment de la faille de porcupine-destor, synthèse métallogénique de duparquet (phase 2/2). *Technical Report 3D 2004-01, MRNF Québec*.
- F. Fallara, M. Legault, L. Cheng, and O. Rabeau, 2006. 3-d integrated geological modeling in the abitibi subprovince: Techniques and applications. *J. of mining and exploration geology*, ICM, in press.
- A. G. Journel, 2002. Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. *Mathematical Geology*, 34(5):573-596.
- P. Komarek, 2004. Logistic regression for data mining and high-dimensional classification. *Technical Report CMU-RI-TR-04-34, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA*.
- M. Legault, F. Fallara, L. Cheng, G. Baudoin, J. Goutier, G. Perron, O. Rabeau, and M. Aucoin, 2003. A new look at an old mining camp: The Destor-Porcupine fault, Abitibi subprovince (phase 2). *Technical Report DV 2003-08, MRNF Québec*.
- M. Legault, F. Fallara, L. Cheng, G. Baudoin, J. Goutier, G. Perron, O. Rabeau, and M. Aucoin, 2004. Synthèse métallogénique et modélisation 3-D de la faille de Porcupine-Destor dans le secteur de Duparquet, sous-province de l'Abitibi (phase 2/3). *Technical Report RP 2004-07, MRNF Québec*.
- J.-L. Mallet, 1997. Discrete modeling for natural objects. *Mathematical Geology* 29(2): 199-209.
- S. Krishnan and A. Journel, 2004. Evaluating information redundancy through the tau model. In *O. Leuangthong and C. V. Deutsch, eds., Proc. Geostatistics Banff*. Kluwer, Dordrecht.