

## Review Article

# Modified Weights-of-Evidence Modeling with Example of Missing Geochemical Data

Daojun Zhang <sup>1,2</sup> and Frits Agterberg<sup>3</sup>

<sup>1</sup>College of Economics and Management, Northwest A&F University, Yangling 712100, China

<sup>2</sup>Center for Resource Economics and Environment Management, Northwest A&F University, Yangling 712100, China

<sup>3</sup>Department of Earth and Space Science and Engineering, York University, Toronto, Canada M3J1P3

Correspondence should be addressed to Daojun Zhang; [cugzdj@gmail.com](mailto:cugzdj@gmail.com)

Received 15 January 2018; Accepted 28 August 2018; Published 1 November 2018

Academic Editor: Yan-Ling Wei

Copyright © 2018 Daojun Zhang and Frits Agterberg. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Weights of evidence (WofE) and logistic regression (LR) are two loglinear methods for mineral potential mapping. Both models are limited by their respective basic assumptions in application. Ideally, WofE indicator patterns have the property of conditional independence (CI) with respect to the point pattern of mineral deposits to be predicted; in LR, there supposedly are no interactions between the point pattern and two or more of the indicator patterns. If the CI assumption is satisfied, estimated LR coefficients become approximately equal to WofE contrasts and the two methods produce similar results; additionally, bias then is avoided in that the sum of all estimated posterior probabilities becomes approximately equal to the number of observed discrete events. WofE allows construction of input layers that have missing data as a separate category in addition to known presence-absence type input, while logistic regression as such is not capable of handling missing data. As an improved WofE model based on LR, modified weights of evidence (MWofE) inherit the advantages of both LR and WofE, i.e., eliminates bias due to lack of CI and can handle missing data as well. Pixel or unit area input for MWofE consists of positive and negative weights for presence and absence of a pattern plus zeros for missing data. MWofE first is illustrated by application to simple examples. Next, it is applied to a study area with 20 known gold occurrences in southwestern Nova Scotia in relation to four input layers based on geological and lake geochemical data. Assuming that geochemical data were missing for the northern part of the study area, MWofE, like WofE but unlike LR, provides posterior probabilities for the entire area.

## 1. Introduction

Introductions to basic principles of weights of evidence (WofE) can be found in Agterberg [1], Bonham-Carter et al. [2], and Bonham-Carter [3]. The method has been applied widely in different fields. The reader is referred to Lindsay et al. [4], Chen et al. [5], and Qin and Liu [6] for recent examples of application. Besides, as a powerful tool to deal with missing data, WofE was also used to develop a spatially weighted logistic regression model for mineral prospectivity mapping [7]. A prerequisite for WofE modeling is that there is approximately conditional independence (CI) between the evidential layers with respect to the target layer [8]. It can be difficult to fully satisfy the CI hypothesis in practice. Various methods have been developed (1) to test for CI and (2) to overcome its effects if it exists. Pairwise

$G^2$ - and  $\chi^2$ -tests for CI were proposed by Bonham-Carter et al. [2]. Later, the omnibus test [3], new omnibus test [9, 10], and Kolmogorov-Smirnov test [11] were introduced for testing the CI hypothesis. In general, lack of CI can be reduced by combining interdependent explanatory variables with one another, or the conditional dependence (CD) problem can be circumvented by using a different statistical model. Agterberg [12] introduced the use of logistic regression (LR); Journel [13] proposed the Tau model which was refined by Krishnan et al. [14], Caumon et al. [15], and Polyakova and Journel [16]. These methods have been discussed in the comprehensive review paper by Schaabben [17]. Additionally, several weighted WofE and stepwise WofE models have been developed by Zhang et al. [18], Deng [19], Agterberg [20, 21], and Cheng [22–24]. It should be kept in

mind that the purpose of WofE and LR is prediction of probable locations of point events that have not been observed to exist. In general, the true total number of events in a study area remains unknown. If unknown events would occur in the same geological environments as known events, only the prior probability in WofE is affected by this lack of information. In general, WofE predictive maps are only valid in a relative sense, because true event occurrences patterns are either overestimated or underestimated. They are used to delineate prospective target areas with no or few known events within a study area. This approach of using data from the entire study area differs from the one in which the well-explored part of a study area is taken as training area, for prediction in a target area (cf. [25]).

Weights of evidence and logistic regression both belong to the family of loglinear models (e.g., [26–31]). They are based on different assumptions: (1) ideally, WofE indicator patterns are conditionally independent with respect to the point pattern, and (2) in LR there are no interactions between the point pattern and two or more of the indicator patterns [12]. Compared to WofE, LR can be applied to avoid or weaken the effect of CD although it can remain affected by interactions between groups of three or more variables; LR necessarily results in a total number of estimated events (=sum of posterior probabilities for all predicted event occurrences) that is exactly equal to the number of known events [21, 32]. Because of these advantages, Agterberg [20, 21] and Deng [19] independently proposed to use LR to improve upon WofE results with lack of CI. Deng's model does not result in unbiased estimates but it significantly reduces bias; in an example used for illustration by Agterberg [21], total bias in comparison with WofE results was reduced from 92% to 31% of the estimated total number of events when Deng's method was used. Later, the theoretical error in Deng's original assumption of unbiasedness was pointed out and described in detail by Schaeben and van den Boogaart [33]. Deng's model diminishes lack of CI bias in WofE results but better results generally can be obtained by other methods such as boosting [22–24] or by modified weights of evidence (MWofE) in which WofE and LR are combined with one another [21, 34]. MWofE was firstly proposed by Agterberg [21] to overcome the lack of CI in WofE modeling. It used positive and negative weights instead of 1 and 0 to estimate the coefficients in LR and then used these coefficients as correction factors for the weights of evidential layers in WofE. Therefore, it belongs to the weighted WofE models [34]. When the explanatory variables are presence-absence type and quantified from maps, the weighted logistic regression (WLR) generally is to be preferred over logistic regression because numerous "observations" with equal values for the explanatory variables must be combined with one another to form "unique conditions," which are weighted according to their map areas [12]. The term "unique condition" describes the set of unit areas or pixels with identical values for all explanatory variables. The map areas for a unique condition either can be measured directly or approximated by counting pixels with the same values for all map patterns. The advantage of WLR is that the number of unique conditions ( $n$ ) generally is much smaller than the number of

pixels or unit areas ( $n'$ ). For example, in a square study area,  $n' = m^2$  where  $m$  can be as large as several millions so that  $n'$  becomes much too large for practical applications. On the other hand,  $n$  is derived from the number of different input patterns ( $p$ ) with  $n = 2^p$ . In general,  $(n \times n) \ll (n' \times n')$ , thus facilitating estimation of logistic regression coefficients in applications using an iterative process that involves successive inversions of  $(n \times n)$  instead of  $(n' \times n')$  matrices assuming that all variables have been corrected for their arithmetic means before matrix inversions are applied. For example, if  $n' = 10^6$  for 7 binary map patterns digitized for a square array with  $1000 \times 1000$  pixels, the  $(256 \times 256)$  WLR input matrix for the explanatory variables has  $65 \cdot 10^6$  fewer elements than the corresponding LR input matrix with  $10^{12}$  elements. Some software packages (e.g., recent versions of IBM SPSS) allow weighting of observations in a manner that is equivalent to WLR. However, in applications to very large data sets, it can be advantageous to perform a final convergence check by comparing the sum of all posterior probabilities with the number of known events. These two quantities should be equal to one another. If there is a significant discrepancy, iterations should be continued until full convergence is reached [35]. WLR should obtain the same results as LR if afronding-precision is not considered.

WofE and LR produce the same coefficients when a point pattern is related to a single map layer [36, 37] but standard deviations of these coefficients as obtained by the two methods are different. When there are relatively few point events, LR is likely to produce smaller variances of the coefficients than WofE contrasts provided that the map patterns are nearly conditionally independent. This is because WofE variances of weights and contrasts are based on an assumption of asymptotic normality of maximum likelihood estimators [38], a condition that is unlikely to be satisfied when the number of observed point events is relatively small. It is noted here that the WofE contrast measures the strength of spatial correlation between a point pattern and an indicator pattern. If a WofE contrast would be standardized to the  $[-1, 1]$  interval, it becomes Yule's original coefficient of association for binary variables [38].

In general, differences in results obtained by two different loglinear models can be tested for statistical significance using the  $G^2$ -test (cf. [12]). If there are  $p$  map layers,  $G^2$  for the difference in results obtained by WofE and LR is approximately distributed as chi-square with  $k$  degrees of freedom. For example,  $k = 1$  if  $p = 2$ , and  $k = 4$  if  $p = 3$ . When  $p$  is large,  $k$  rapidly becomes very large, and other strategies for model comparison may be adopted. For example, Agterberg et al. [11] related a gold deposit pattern in Nova Scotia to seven map patterns that were approximately conditionally independent of the point pattern. In this situation, the resulting logistic regression coefficients were close to the WofE contrasts ([11], Table 1).

Logistic regression prevents lack of CI bias in situations where WofE does not. However, if the CI condition is not satisfied, individual regression coefficients normally acquire large variances because of near singularities in the matrices to be inverted when the iterative scoring (or any other) method is used for LR. Contrary to WofE weights, LR

TABLE 1: Probabilities estimated for unique conditions in missing data variant of artificial example of Schaeben ([42], Figure 1). Information on  $B_2$  was assumed to be missing (m in column 2) for 30% of the area on the left side of the study area for this example. Area,  $n(D)$ ,  $P_1$ , and  $P_2$  as in Table 4.

$B_1$	$B_2$	Area	$n(D)$	$P_1$	$P_2$
1	1	2	1	0.6842	0.7679
1	0	8	3	0.2207	0.2143
1	m	10	3	0.3500	0.3750
0	1	12	5	0.2784	0.3122
0	0	48	2	0.0480	0.0361
0	m	20	0	0.0875	0.0761

TABLE 2: Probabilities estimated for Schaeben and van den Boogaart's [33] fabricated data set. Binary predictors  $B_1$ ,  $B_2$ , and  $B_3$  are the explanatory variables and  $D$  represents the dependent variable.  $P_1$  is predicted WofE probability and  $P_2$  is predicted modified WofE probability.

$B_1$	$B_2$	$B_3$	$D$	$P_1$	$P_2$
1	1	1	1	0.7805	0.7458
1	1	1	1	0.7805	0.7458
1	1	1	1	0.7805	0.7458
1	1	1	1	0.7805	0.7458
1	1	1	1	0.7805	0.7458
1	1	1	0	0.7805	0.7458
1	1	0	1	0.6400	0.6314
1	1	0	0	0.6400	0.6314
1	0	1	1	0.6400	0.6314
1	0	1	0	0.6400	0.6314
1	0	0	1	0.4706	0.5000
1	0	0	0	0.4706	0.5000
0	1	1	1	0.6400	0.6314
0	1	1	0	0.6400	0.6314
0	1	0	1	0.4706	0.5000
0	1	0	0	0.4706	0.5000
0	0	1	1	0.4706	0.5000
0	0	1	0	0.4706	0.5000
0	0	0	1	0.3077	0.3686
0	0	0	0	0.3077	0.3686

coefficients then cannot be used individually for interpretation or prediction purposes. This in itself can be a good reason to prefer usage of WofE. There are many different strategies for obtaining approximate CI. For example, patterns based on geochemical elements measured across a whole study area generally have strong mutual interactions with respect to occurrences of mineral deposits. Approximate CI then usually can be obtained by combining all elements into a single index either by multiple regression analysis (e.g., [39]) or by using scores for the first principal component of the correlation matrix (e.g., [40]; also see Section 5 in this paper). Even if CD exists, the new AdaBoost

TABLE 3: Probabilities estimated for Agterberg et al. [11] seafloor example. Five binary predictors  $B_i$  ( $i=1,2,\dots,5$ ) are for age, topography, contact between youngest volcanics, rock type, and fissures, respectively. Area of unique condition is given in units of  $10\text{ m} \times 10\text{ m}$ ;  $n(D)$  represents the number of vents per unique condition.  $P_1$  is predicted WofE probability and  $P_2$  is predicted modified WofE probability.

$B_1$	$B_2$	$B_3$	$B_4$	$B_5$	Area	$n(D)$	$P_1$	$P_2$
0	0	0	0	0	10,052	0	0.0000	0.0000
0	0	0	0	1	3363	1	0.0000	0.0000
0	0	0	1	0	3268	0	0.0000	0.0000
0	0	0	1	1	1074	0	0.0000	0.0000
0	1	0	0	0	5455	0	0.0000	0.0001
0	1	0	0	1	25	0	0.0000	0.0000
0	0	1	0	0	3482	0	0.0001	0.0002
0	1	0	1	0	2518	0	0.0001	0.0001
0	0	1	0	1	1474	0	0.0001	0.0001
0	1	0	1	1	1371	0	0.0000	0.0000
1	0	0	0	0	5	0	0.0002	0.0004
1	0	0	0	1	705	0	0.0002	0.0002
0	0	1	1	0	5	0	0.0001	0.0002
0	0	1	1	1	744	0	0.0001	0.0001
1	0	0	1	0	422	0	0.0004	0.0004
1	0	0	1	1	58	0	0.0004	0.0002
0	1	1	0	0	12	0	0.0003	0.0004
0	1	1	0	1	179	0	0.0003	0.0002
1	1	0	0	0	1766	2	0.0009	0.0009
1	1	0	0	1	119	0	0.0008	0.0004
0	1	1	1	0	1055	1	0.0006	0.0004
0	1	1	1	1	33	0	0.0005	0.0002
1	0	1	0	0	10	0	0.0024	0.0033
1	1	0	1	0	146	0	0.0018	0.0009
1	0	1	0	1	623	1	0.0020	0.0014
1	1	0	1	1	145	0	0.0015	0.0004
1	0	1	1	0	504	2	0.0048	0.0033
1	0	1	1	1	1	0	0.0040	0.0014
1	1	1	0	0	317	2	0.0098	0.0069
1	1	1	0	1	277	1	0.0083	0.0029
1	1	1	1	0	348	3	0.0194	0.0069
1	1	1	1	1	295	0	0.0164	0.0029

TABLE 4: Probabilities estimated for unique conditions in the artificial example of Schaeben ([42], Figure 1) with two binary predictors. Area is the number of unit cells;  $n(D)$  represents number of events per unique condition.  $P_1$  is predicted WofE probability and  $P_2$  is predicted modified WofE probability.

$B_1$	$B_2$	Area	$n(D)$	$P_1$	$P_2$
1	1	4	1	0.586364	0.640555
1	0	16	6	0.26875	0.277361
0	1	16	5	0.201563	0.214861
0	0	64	2	0.061429	0.05566

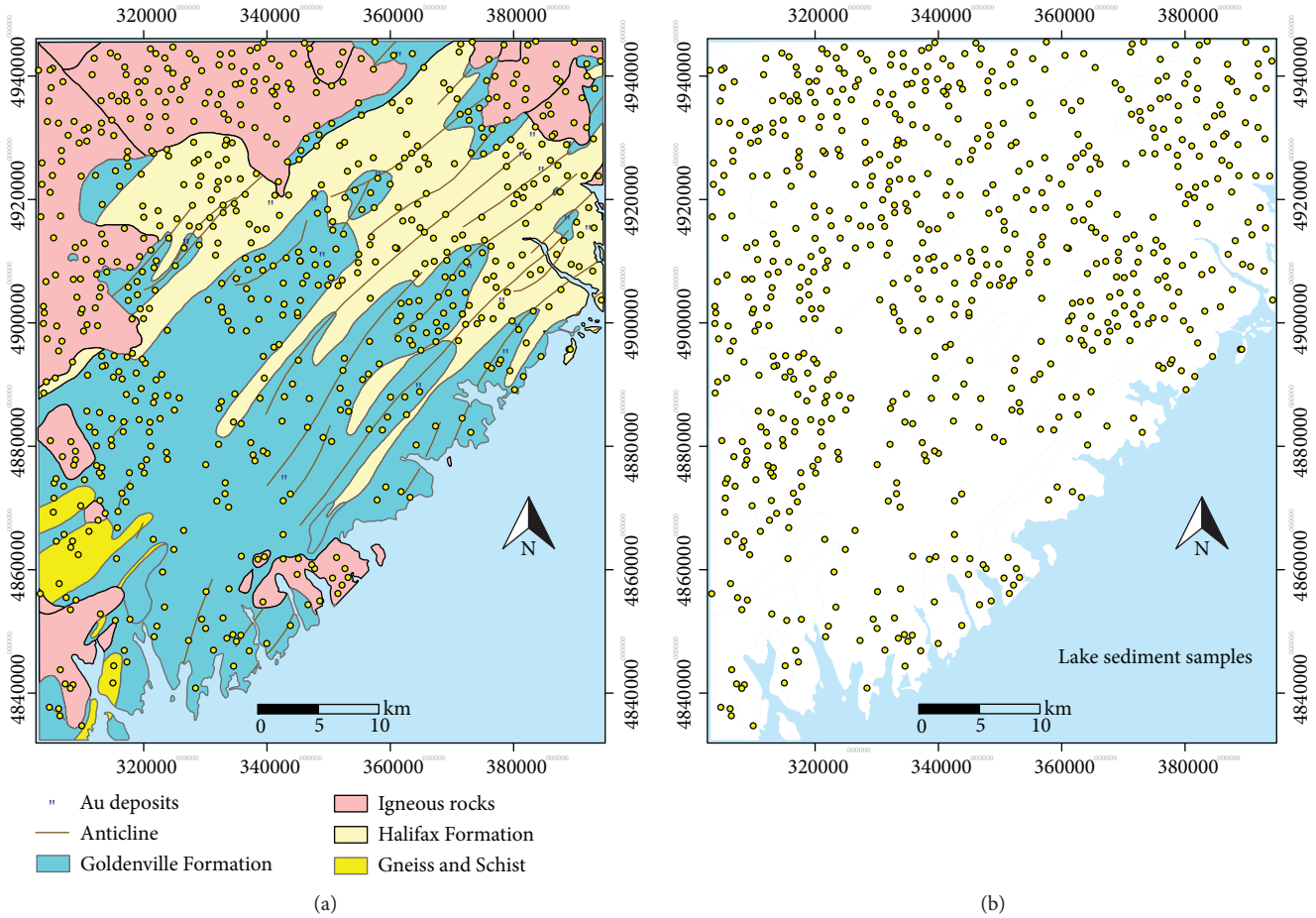


FIGURE 1: (a) The simplified geology and distribution of known gold deposits (black dots) in southern Nova Scotia. Solid black lines are traces of Devonian anticlines. (b) The locations of 671 lake sediment samples in the area (source: [8], Figure 1).

WofE method still can give good approximations provided that the number of patterns is not very large [24].

In the following section, it will be discussed that MWofE is equivalent to LR. A major advantage of MWofE is that, like WofE, it can handle missing data better than LR. This subject will be discussed in the remainder of this paper.

## 2. Why Modified Weights of Evidence Is Equivalent to General Logistic Regression

Equivalence of LR and MWofE will be reviewed and illustrated by two examples of application. In Agterberg's study [20], it was pointed out that MWofE is based on a method originally introduced into the GLADYS medical expert system [41]. Weights of evidence in GLADYS usually resulted in biased posterior probabilities. For this reason, Spiegelhalter and Knill-Jones used logistic regression of their presence-absence data with explanatory variables for which the presence/absence data were replaced by positive and negative weights in order to eliminate this bias.

Schaeben and van den Boogaart [33] illustrated the fact that Deng's [19] adjusted WofE model does not produce unbiased results by using a "fabricated" data set in which there are three explanatory variables ( $B_1$ ,  $B_2$ , and  $B_3$ ) and a

binary dependent variable ( $D$ ). Table 2 shows the values for these variables in the first four columns. Probabilities ( $P_1$ ) obtained by WofE are shown in the next column. They differ from probabilities ( $P_2$ ) in the last column that were obtained by LR or MWofE, which produce identical results. In Section 4, it will be illustrated that similar results are obtained for a "practical example with fabricated training data set results" used by Schaeben [42].

Equivalence of LR and MWofE already had been illustrated in Agterberg [20, 21] for a relatively small rectangular study area of about  $4.0 \text{ km}^2$  on the East Pacific Rise with 13 volcanic vents that were related to depth below sea level, fissures, relative age, and composition of volcanic rocks. Input for this seafloor example is shown in the first five columns of Table 3. The rows in this table are for the  $2^5 = 32$  unique conditions of which the total areas (in units of  $10 \text{ m} \times 10 \text{ m}$ ) are given in the sixth column. The dependent variable in the seafloor example of Table 3 is for the presence or absence of volcanic vents. The number of volcanic vents per unique condition is given in the column for  $n(D)$ . In total, there were 39,851 unit areas in this example, and WLR is to be preferred over LR for the reasons given in Section 1 when the scoring method is used to estimate the logistic regression coefficients. As in Table 2, the last two columns of Table 3 show WofE

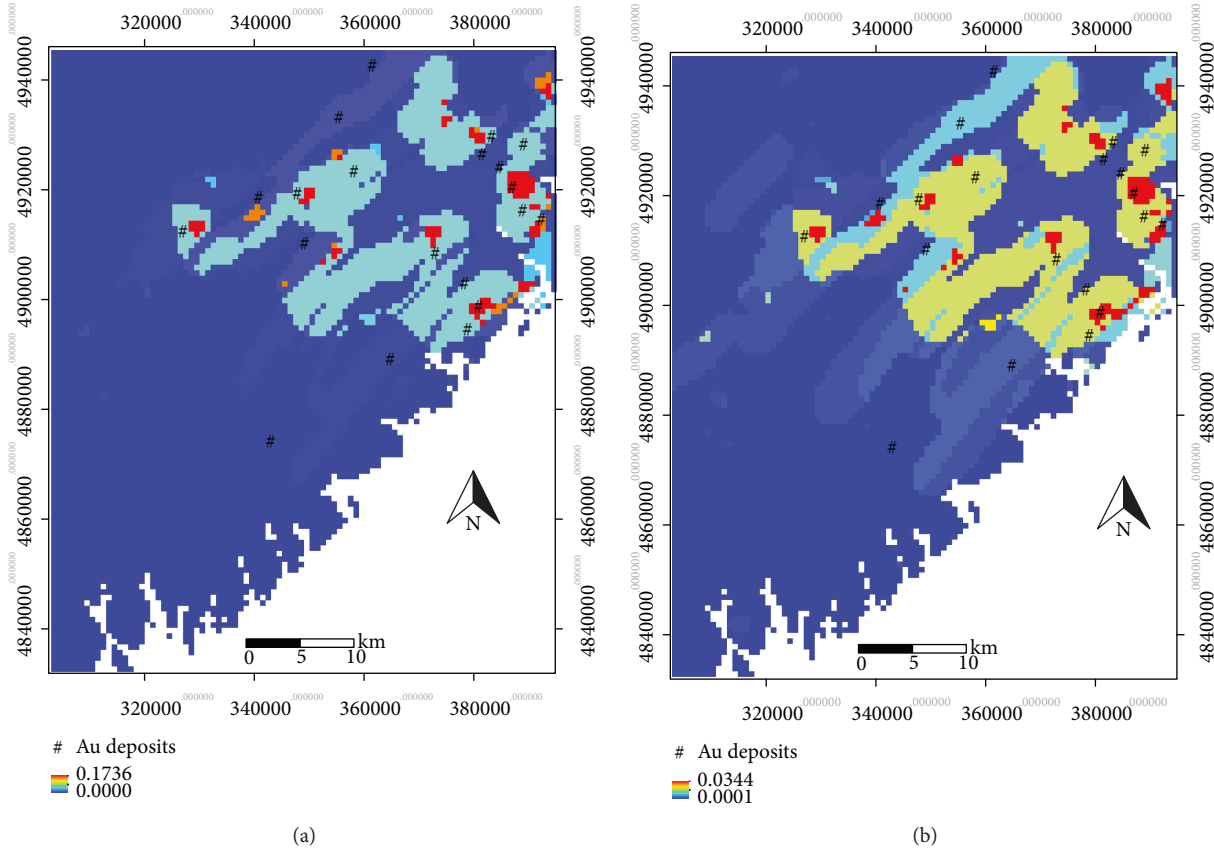


FIGURE 2: Posterior probability maps obtained by (a) weights of evidence and (b) logistic regression in experiment 3.

( $P_1$ ) and MWofE ( $P_2$ ) probabilities for ( $10\text{ m} \times 10\text{ m}$ ) unit areas with the same unique conditions. The latter are equal to posterior probabilities obtained by WLR. In this example, there is strong conditional dependence of the five map layers [11]. This can be seen from the fact that the sum of all ( $=39,851$ ) WofE posterior probabilities ( $=24.8$ ) in Table 3 significantly exceeds the corresponding sum of MWofE probabilities ( $=13.0$ ). Lack of CI can be assessed by estimating the variance of the sum of all WofE posterior probabilities followed by application of a simple  $z$ -test or  $t$ -test to determine whether or not the difference between the sum of all posterior probabilities and total number of known events is statistically significant [9].

### 3. Lack of Constancy of Statistical Parameters in Large Areas

One of the drawbacks of LR is that estimated regression coefficients may have large variances unless there is approximate CI. There are several other potentially serious disadvantages as well, especially if the method is applied naively to continuous explanatory variables for prediction purposes. Agterberg and Bonham-Carter [43] have systematically compared WofE and WLR predictions with one another in a number of experiments on the relationship between gold occurrences with geological, geochemical, and geophysical map data in the Gowganda area on the Canadian Shield in east-central Ontario. The results of these experiments

clearly showed that discretization (reducing map patterns to binary or ternary form, before using them as explanatory variables) and integration of training and testing areas in mineral potential evaluation studies result in better predictions when logistic regression is used. The reason for this is that a primary requirement for validity of application of a mathematical model in nature is constancy of statistical parameters within the entire study area. Extrapolation from a training area to a testing area that is geographically distinct from the training area is possible but special arrangements have to be made to ensure approximate equality of statistical parameters for the explanatory variables (cf. [25]).

Because of the widespread uniqueness of geological phenomena that can rapidly change within the same study area, a single target population with constant parameters such as means, variances, and covariances often does not exist. The parameters of interest can change systematically from place to place but methods are available to reduce or eliminate the effects of systematic changes in the values of the parameters of interest. Some geological properties such as lithological composition or anomalies versus background are intrinsically binary. Discretization by Heaviside (0 or 1) transformation of nonbinary geochemical or geophysical variables can stabilize statistical parameters within a region. Another example of stabilization of statistical parameters is to replace observed data by residuals after regional trend elimination.

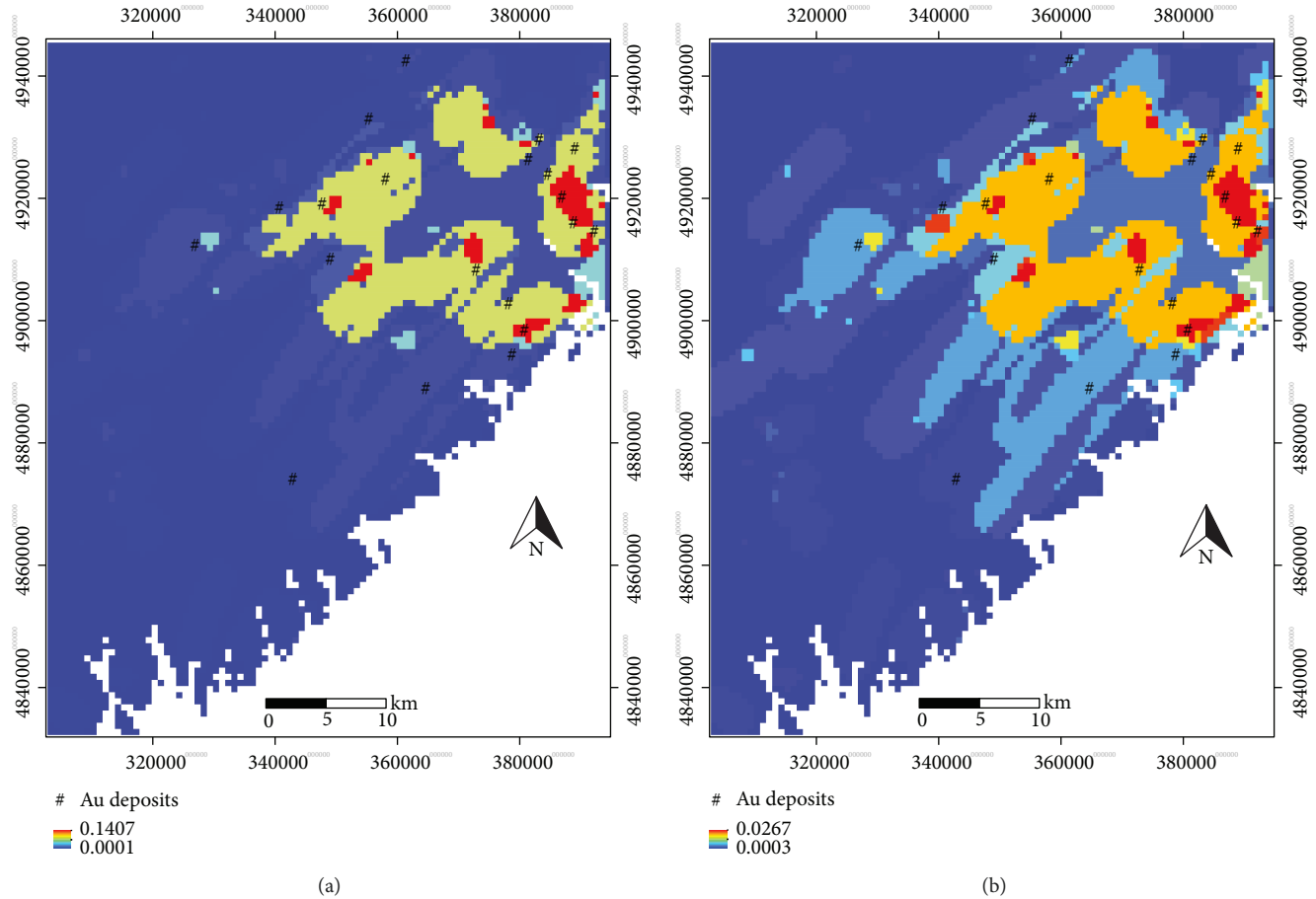


FIGURE 3: Posterior probability maps obtained by (a) weights of evidence and (b) logistic regression in experiment 1.

#### 4. Map Patterns with Missing Data

Weights of evidence as well as logistic regression were originally developed by Good [44], Spiegelhalter and Knill-Jones [41], and many others for independent identically distributed observations. Mosaic-type map data that are quantities for vanishingly small unit areas or pixels constitute observations of a very different type. Nevertheless, WofE and WLR, including Spiegelhalter's [45] refinement in GLADYS to handle missing data can be applied successfully to mosaic data. Geoscience patterns often have gaps for different reasons: geochemical lake or stream sediment data may be available for parts of a study area only (obviously, they are missing in subareas where there are no lakes or rivers). Biogeochemical data are restricted to areas where specific plants or trees occur. It may be that only a limited part of an area was mapped to obtain a particular kind of data; e.g., a geochemical survey may have been restricted to parts of an area. Also, exposure of bedrock can be uneven. Locally, it may not be known whether or not a rock type is present. In situations of this type, the binary presence-absence pattern of one or more explanatory variables has gaps that can be quantified separately so that the binary pattern becomes a ternary pattern with separate states for presence, absence, and unknown. WofE and MWofE can cope with missing or unknown data and will first be illustrated in the following simple example.

Tables 1 and 4 show results for an artificial example with a square study area ([42], Figure 1) that contains  $(10 \times 10)$  regularly spaced data points located at the centers of square cells for which presence or absence is assumed to be known for two rock types ( $B_1$  and  $B_2$ ) and for the point event of interest. These data are shown in Table 4 together with probabilities of occurrence estimated by WofE and MWofE. As in Tables 2 and 3, MWofE results are identical to those obtained by LR. These probabilities are the same as those previously reported by Schaeben [42].

As a variant on this artificial example, we assumed that information for  $B_2$  was incomplete in that it was not available for the 30% of the study area on the left side of Schaeben's Figure 1. This means that the binary pattern for  $B_2$  becomes a ternary pattern with three states (yes, no, and missing). Both WofE and modified WofE were applied to this new artificial data set with the results shown in Table 1. In this application, Spiegelhalter's [45] method for dealing with missing data was applied twice: initially to obtain the WofE weights and again when MWofE was applied.

The example of Table 1 illustrates how missing data problems can be solved by recognizing "missing" as a separate state that requires special consideration. Other possible methods to apply LR when there are missing data for a predictor variable were discussed in detail by Agterberg and Bonham-Carter [46]. In their paper, four methods are

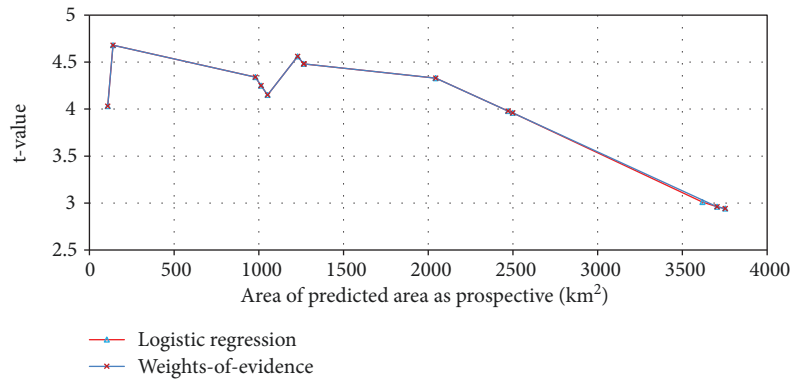


FIGURE 4: *t* values for weights of evidence and logistic regression in experiment 1.

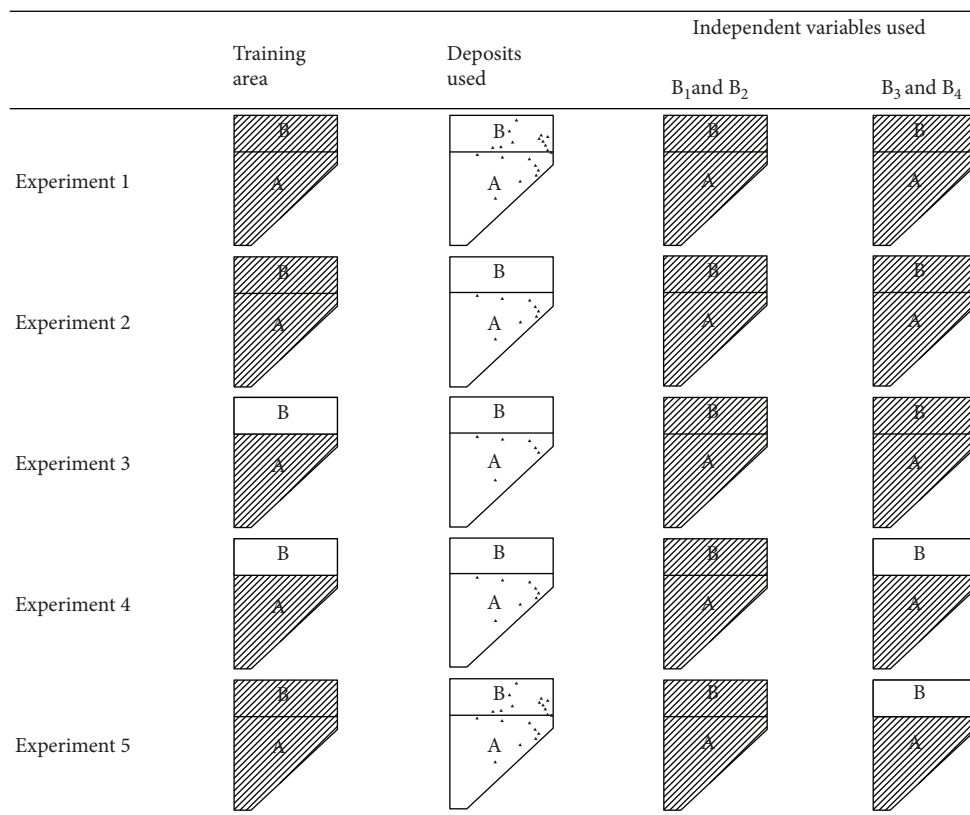


FIGURE 5: Four experiments in which the 20 gold deposits were related to the four binary input layers for the study area shown in Figure 1. Results for experiment 1 were shown in Figure 3. For the other three experiments, the study area was divided into two parts according to the east-west line in the middle while deposits and pairs of four binary layers were used as shown.

applied to the previously mentioned example of gold deposits in Nova Scotia: (1) the variable with missing data is deleted; (2) the missing data are replaced by zeros for absence; (3) all polygons or subareas with missing data are deleted; and (4) the missing data are replaced by a mean value for the same variable estimated in the part of the study area where it is known. It was concluded that patterns of posterior probabilities resulting from methods (1) and (3) were relatively poor in comparison with patterns resulting from methods

(2) and (4). A motivation for the development of MWofE was to provide an even better method that can remedy missing data problems in the same way as WofE [21].

### 5. Case History Study of Gold Deposits in Southwestern Nova Scotia, Canada

This larger-scale example builds on a study of weights-of-evidence modeling originally published by Cheng [8].

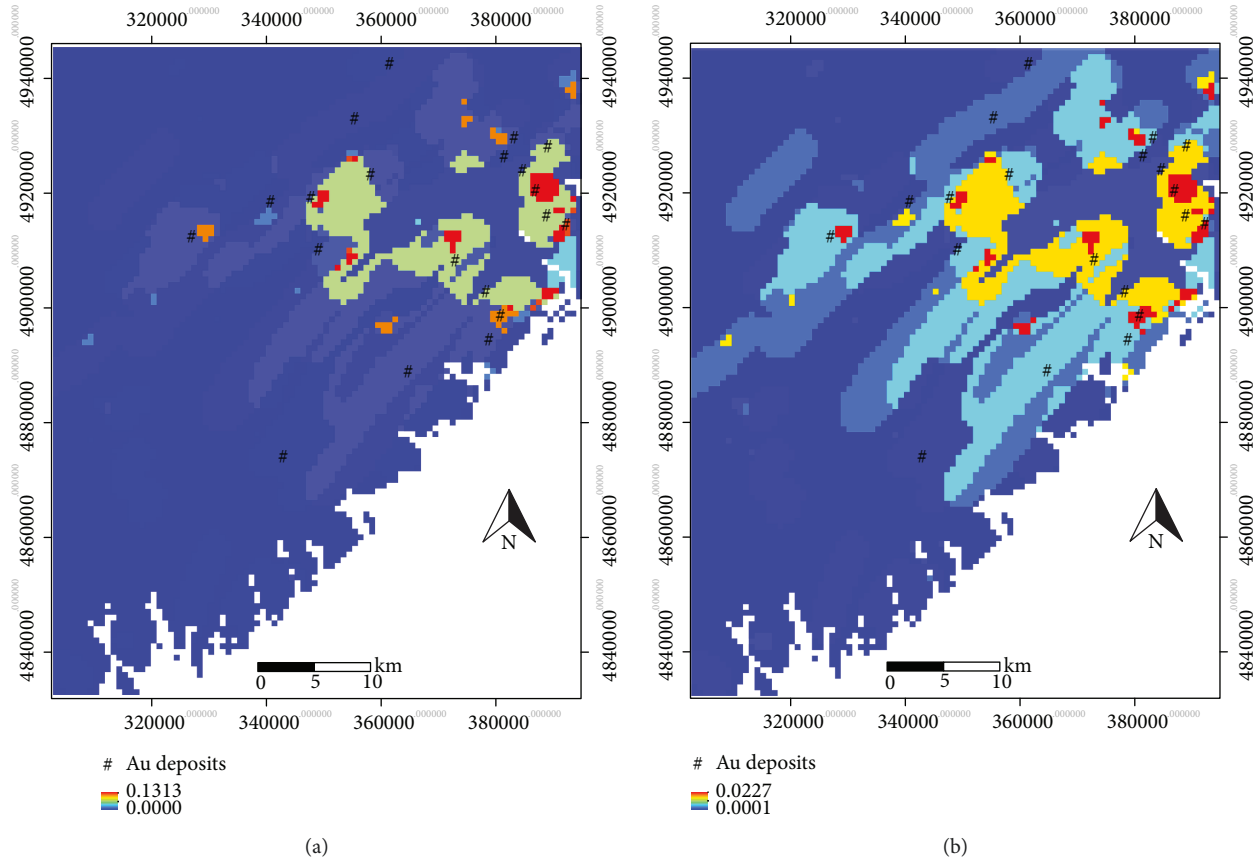


FIGURE 6: Posterior probability maps obtained by (a) weights of evidence and (b) logistic regression in experiment 2.

Geology, occurrences of gold deposits, and locations of lake sediment samples are given in Figure 1 (after [8], Figure 1). The geochemical lake sediment data consisted of 671 samples with concentration values for 16 elements (Cu, Pb, Zn, Ag, F, Li, Nb, Rb, Sn, Zr, Ti, Au, Sb, As, Th, and W) originally published by Rogers et al. (1987). Sampling density was about 1 sample per 5 km<sup>2</sup>. Results of principal component analysis and application of high-pass and low-pass filtering to principal component scores are described in detail by Cheng [8].

Originally, four binary maps (cf. [8], Figure 2; Table 2) were used as input for WofE. The same input data were used by us for the current study, although there were slight differences in choice of binary maps and resolution. Here, we used: (A) proximity to traces of anticlines; (B) proximity to contact between Goldenville and Halifax Formations; (C) high-pass filter applied to scores of first principal component; and (D) low-pass filter applied to scores of first principal component. Figure 3 shows WofE results in comparison with LR results for the same four input layers. These results are for experiment 1 in a series of five experiments of which the other four experiments will be described later. Our Figure 3(a) differs slightly (but not significantly) from the earlier WofE map of Cheng [8]. The two maps in Figure 3 for the 20 gold deposits are similar in appearance. However, from the legends, it can be seen that estimated posterior probabilities of gold deposits are much larger in Figure 3(a) than in Figure 3(b). On average, the WofE probabilities are more than twice as large as the LR probabilities. The sums of all

posterior probabilities were 46.9 (WofE) and 20 (LR), respectively. LR results are unbiased in the sense that this sum is exactly equal to the number of gold deposits used. The WofE results are systematically too large due to violations of the conditional independence (CI) assumption. It is noted that, even if the WofE would be corrected for bias due lack of CI, both totals would remain too low because of undiscovered gold deposits. Such bias is likely because intensity of exploration varied greatly across the area. Also, undiscovered gold deposits may not only occur near the surface of bedrock in relatively unexplored parts of the area, but probably everywhere in favorable environments at greater depths.

Student's  $t$ -test (cf. [38]) can be used to measure the strength of spatial correlation between known deposits and prognostic contours as shown in Figure 4. Every  $t$ -value in Figure 4 is for a comparison of the contrast  $C$  in two subareas as follows. One of these subareas consists of all points within a given distance of one or more gold deposits. Its area is shown along the horizontal axis in Figure 4. The other sub-area is simply for the complement representing the remainder of the study area. Due to spatial correlation effects, the number of degrees of freedom for  $t$ -tests of this type is not known exactly. However, for a one-tailed  $t$ -test with significance level set at 0.05, values greater than approximately 1.645 would indicate positive spatial correlation between the point pattern for deposits and the patterns of posterior probabilities for WofE or LR. Obviously, there is a relatively strong spatial correlation between the pattern of gold



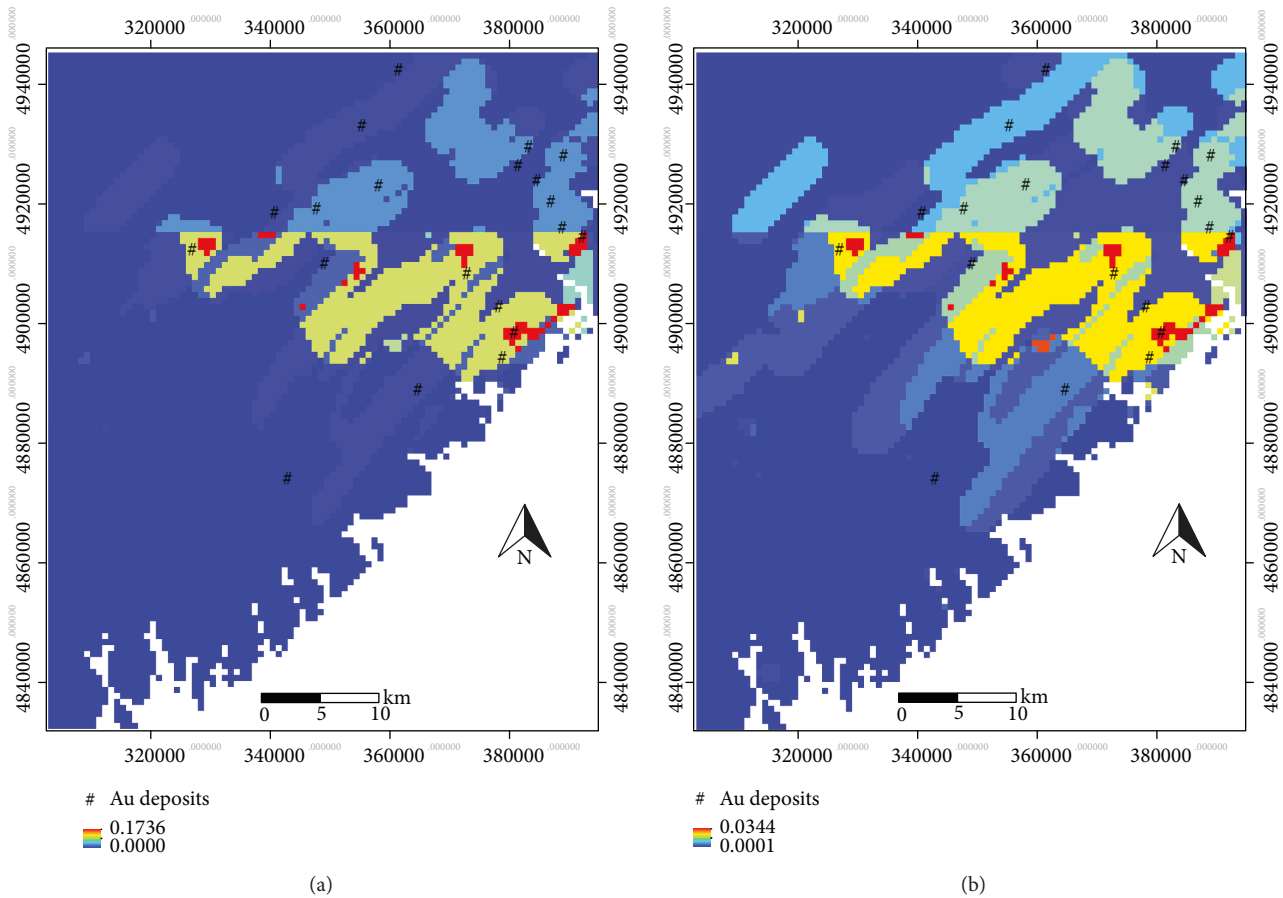


FIGURE 7: Posterior probability maps obtained by (a) weights of evidence and (b) modified weights of evidence in experiment 4.

deposits and either one of the posterior probability maps shown in Figure 3.

The results shown in Figure 3 are for experiment 1 from a series of five experiments conducted to compare WofE, LR, and modified weights-of-evidence (MWofE) for the 20 gold deposits in this study area. In the other four experiments, the study area was divided into two parts according to the east-west line in Figure 5 showing gold deposits and pairs of binary layers as were used in the four experiments. Experiments 2 to 4 used deposits in the southern part of the area (A) only for prediction in the northern part (B), and all deposits were used in experiment 5.

Results for experiments 2 and 3 are shown in Figures 2 and 6. The difference between these two experiments is that four binary layers were used as input from the entire study area (A + B) for Figure 6 and from subarea A only for Figure 2. Both WofE and LR could be used for experiments 2 and 3. It could not be used for LR in experiments 4 and 5 because geochemical information in northern half of the study area was assumed to be missing. Contrary to LR, MWofE could be used in experiments 4 and 5 because, like WofE, MWofE can cope with missing data. The geochemical input layers in these two experiments were assumed to be ternary instead of binary. The difference in modeling then is that WofE immediately produces a result for the two binary and the two ternary input layers, but for MWofE the

ones and zeros used in the WofE input layers are replaced by weights for all four independent variables using as much information as was available for each of them. As in WofE, zeros were used as evidential weights in MWofE for subarea B where the two geochemical input layers were assumed to be missing. The difference between experiments 4 and 5 is that both areas A and B were used as training area in experiment 5 whereas only deposits in area A were used in experiment 4. Results for these two experiments are shown in Figures 7 and 8, respectively.

Table 1 shows the expected number of deposits (EDN) for all models in the five experiments. If there are no missing data, the sum of posterior probabilities for LR is exactly equal to the number of deposits, whereas the WofE sum is larger due to violation of the conditional independence assumption. On the other hand, for MWofE (experiments 4 and 5), the expected number of deposits in the training area is equal to the number of deposits in such area that were assumed to be known.

Tables 5 and 6 show the weights and contrasts with standard deviations for WofE results as shown in Figures 7(a) and 8(a) followed by MWofE regression coefficients with their standard deviations. As in experiments 1–3, the WofE weights in Tables 5 and 6 are too large. The corresponding contrasts with standard deviations also are too large because of the lack of conditional independence. Moreover, the

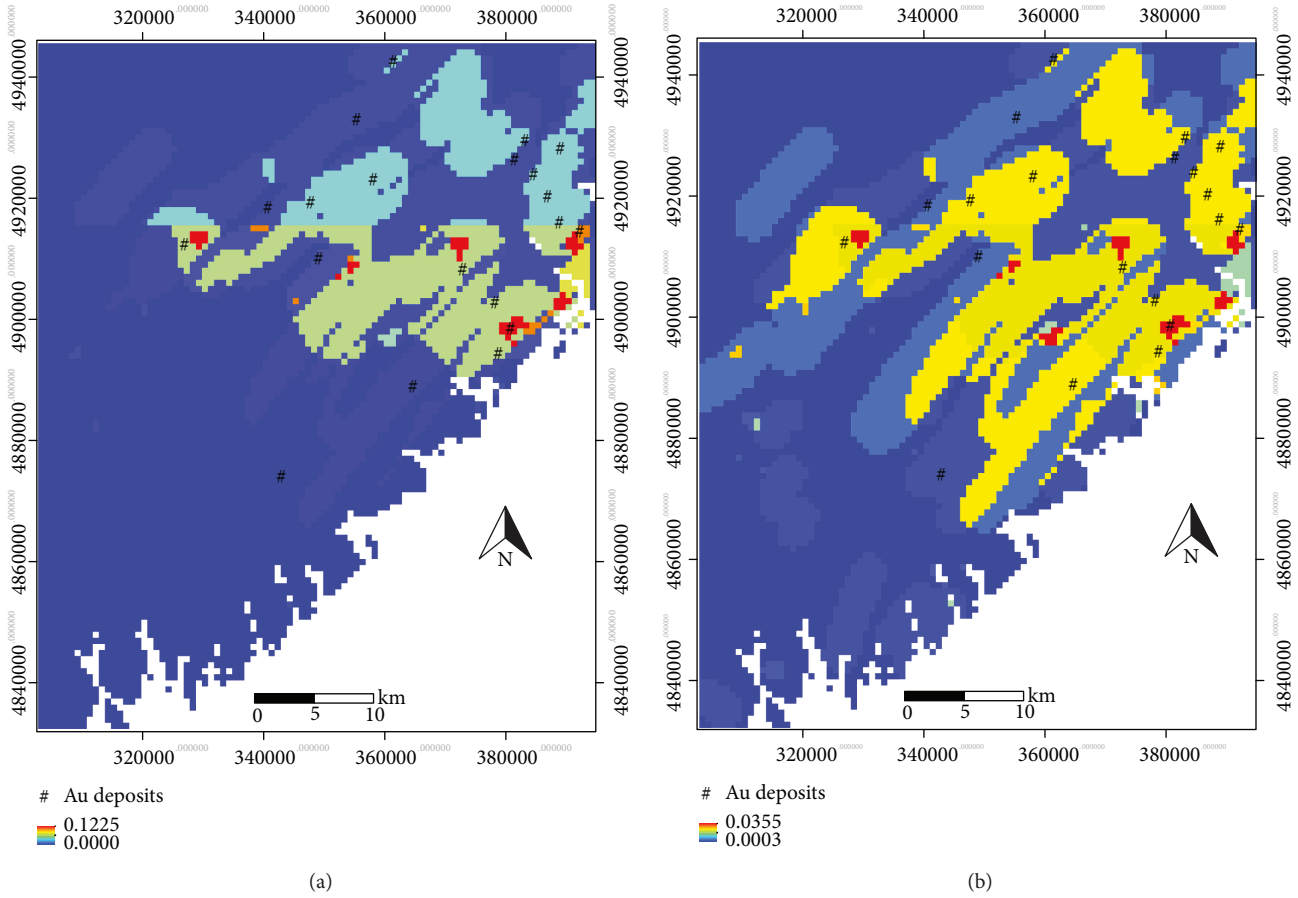


FIGURE 8: Posterior probability maps obtained by (a) weights of evidence and (b) modified weights of evidence in experiment 5.

TABLE 5: Estimated deposit number (EDN) for three methods in the 5 experiments; modified weight-of-evidence (MWofE) for experiments 4 and 5 in the last two columns.

	Experiment 1		Experiment 2		Experiment 3		Experiment 4		Experiment 5	
	WofE	LR	WofE	LR	WofE	LR	WofE	MWofE	WofE	MWofE
EDN in area A (9 deposits)	25.2	11.7	11.4	5.6	22.7	9	22.7	9	16.9	13.9
EDN in area B (11 deposits)	21.8	8.3	8.7	3.4	19.9	7.1	5.1	4.0	7.8	6.1
Total EDM	46.9	20	20.1	9	42.6	16.1	27.8	13.0	24.7	20

TABLE 6: Comparison of WofE weights and contrasts with MWofE coefficients and their standard deviations (Sdev) in experiment 4.

Independent variables	WofE				MWofE	
	$W^+$	$W^-$	C	Sdev	Coefficient	Sdev
$B_1$	0.56	-0.92	1.48	0.80	0.34	0.58
$B_2$	0.90	-1.75	2.65	1.06	0.75	0.42
$B_3$	2.10	-0.22	2.32	0.81	0.70	0.36
$B_4$	1.20	-1.24	2.44	0.80	0.56	0.37

standard deviations of the contrasts were derived by using the standard asymptotic formula which is not necessarily satisfied. The MWofE coefficients as well as their standard deviations are probably more realistic.

## 6. Concluding Remarks

Equivalence of the modified WofE and logistic regression was discussed in Section 2. However, the question can be asked about whether or not there are advantages in using modified WofE given that this method and logistic regression produced the same results for the three examples (Tables 2–4). There are three potential advantages:

- (1) Most existing LR computer software/programs including those written in *R* do not offer satisfactory remedies for missing data, because they require input values for the explanatory variables at all data points and do not allow for data gaps. The problem is that if data are missing for one or more patterns at a given

TABLE 7: Comparison of WofE weights and contrasts with MWofE coefficients and their standard deviations (Sdev) in experiment 5.

Independent variables	WofE				MWofE	
	$W^+$	$W^-$	C	Sdev	Coefficient	Sdev
$B_1$	0.66	-1.08	1.74	0.56	0.70	0.34
$B_2$	0.75	-1.38	2.13	0.63	0.81	0.30
$B_3$	1.74	-0.09	1.83	0.75	0.81	0.42
$B_4$	0.84	-0.27	1.11	0.47	0.30	0.50

point, other data may have to be deleted as well before LR can be applied. This is because the unit areas or pixels used for quantification of map patterns are exceedingly small. They are subject to strong spatial correlation simply because they can be made arbitrarily small so that neighboring pixels for any pixel probably belong to the same unique condition. An example of missing data with predictions for the entire study area without deletion of data blocks was given in Figures 7 and 8 and Tables 1, 5–7.

- (2) Variances of logistic regression coefficients and posterior probabilities are likely to become more precise than variances of WofE weights, contrasts, and posterior probabilities, because the latter are based on an assumption of normality for maximum likelihood estimators that can provide poor results, especially if there are relatively few discrete events for the dependent variable. However, a requirement for this advantage would be that the input map patterns are approximately conditionally independent with respect to the discrete event pattern. Otherwise, the variances of individual logistic regression coefficients can become very large, although the corresponding variances of posterior probabilities would not suffer from this drawback. This potential advantage as well as the previous one will have to be tested further in other experiments and applications.
- (3) WofE is easy to understand and MWofE retains this feature. Although the coefficients in LR obtained through maximum likelihood estimation are used in MWofE to adjust the positive and negative weights for all evidential layers [34], the MWofE almost maintains the framework of WofE, which confirms to the geologists' intuitive understanding in mineral prospectivity mapping.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This study benefited from a joint financial support by the Programs of the National Natural Science Foundation of China (nos. 41602336 and 71503200), Postdoctoral Science Foundation of China (nos. 2016M592840 and 2017T100773), Natural Science Foundation of Shaanxi

Province (no. 2017JQ7010), and the Fundamental Research Funds for the Central Universities (no. 2017RWYB08). The first author thanks former supervisor Dr. Qiuming Cheng for the discussions about spatial weights and for providing constructive suggestions.

## References

- [1] F. P. Agterberg, "Computer programs for mineral exploration," *Science*, vol. 245, no. 4913, pp. 76–81, 1989.
- [2] G. F. Bonham-Carter, F. P. Agterberg, and D. F. Wright, "Weights of evidence modeling : a new approach to mapping mineral potential," in *Statistical Applications in the Earth Sciences, Geological Survey, Canada Paper 89-9*, F. P. Agterberg and G. F. Bonham-Carter, Eds., pp. 171–183, 1989.
- [3] G. F. Bonham-Carter, *Geographic Information Systems for Geoscientists: Modeling with GIS*, Pergamon, Oxford, 1994.
- [4] M. D. Lindsay, P. G. Betts, and L. Ailleres, "Data fusion and porphyry copper prospectivity models, southeastern Arizona," *Ore Geology Reviews*, vol. 61, pp. 120–140, 2014.
- [5] W. Chen, H. Li, E. Hou et al., "GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models," *Science of the Total Environment*, vol. 634, pp. 853–867, 2018.
- [6] Y. Qin and L. Liu, "Quantitative 3D association of geological factors and geophysical fields with mineralization and its significance for ore prediction: an example from Anqing orefield, China," *Minerals*, vol. 8, no. 7, p. 300, 2018.
- [7] D. Zhang, N. Ren, and X. Hou, "An improved logistic regression model based on a spatially weighted technique (ILRBSWT v1. 0) and its application to mineral prospectivity mapping," *Geoscientific Model Development*, vol. 11, no. 6, pp. 2525–2539, 2018.
- [8] Q. Cheng, "Non-linear theory and power-law models for information integration and mineral resources quantitative assessments," *Mathematical Geosciences*, vol. 40, no. 5, pp. 503–532, 2008.
- [9] F. P. Agterberg and Q. Cheng, "Conditional independence test for weights-of-evidence modeling," *Natural Resources Research*, vol. 11, no. 4, pp. 249–255, 2002.
- [10] C. Thiart, G. B. Bonham-Carter, F. P. Agterberg, Q. Cheng, and A. Panahi, "An application of the new omnibus test for conditional independence in weights-of-evidence modeling," *GIS in the Earth Sciences, Geological Association of Canada Special Volume*, J. Harris, Ed., pp. 131–142, 2006.
- [11] F. P. Agterberg, G. F. Bonham-Carter, Q. Cheng, and D. F. Wright, "Weights of evidence modeling and weighted logistic regression for mineral potential mapping," in *Computers in Geology – 25 Years of Progress*, J. C. Davis and U. C. Herzfeld, Eds., pp. 13–32, Oxford University Press, New York, NY, USA, 1993.
- [12] F. P. Agterberg, "Combining indicator patterns in weights of evidence modeling for resource evaluation," *Nonrenewable Resources*, vol. 1, no. 1, pp. 39–50, 1992.
- [13] A. G. Journel, "Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses," *Mathematical Geosciences*, vol. 34, no. 5, pp. 573–596, 2002.
- [14] S. Krishnan, A. Boucher, and A. G. Journel, "Evaluating information redundancy through the tau model," *Quantitative Geology and Geostatistics*, vol. 14, pp. 1037–1046, 2005.

- [15] G. Caumon, J. M. Ortiz, and O. Rabeau, "A comparative study of three data-driven mineral potential mapping techniques," in *Proceedings CD of the 10th Conference of IAMG*, pp. S13–S05, Liège, Belgium, 2006.
- [16] E. I. Polyakova and A. G. Journal, "The nu expression for probabilistic data integration," *Mathematical Geosciences*, vol. 39, no. 8, pp. 715–733, 2007.
- [17] H. Schaeben, "Comparison of mathematical methods of potential modeling," *Mathematical Geosciences*, vol. 44, no. 1, pp. 101–129, 2012.
- [18] S. Zhang, Q. Cheng, S. Zhang, and Q. Xia, "Weighted weights of evidence and stepwise weights of evidence and their applications in Sn–Cu mineral potential mapping in Gejiu, Yunnan Province, China," *Journal of China University of Geosciences*, vol. 34, pp. 281–286, 2009.
- [19] M. Deng, "A conditional dependence adjusted weights of evidence model," *Natural Resources Research*, vol. 18, no. 4, pp. 249–258, 2009.
- [20] F. P. Agterberg, "A modified weights-of-evidence method for regional mineral resource estimation," in *Proceedings CD of the 11th Annual Conference of IAMG*, Stanford, CA, USA, August 2009.
- [21] F. Agterberg, "A modified weights-of-evidence method for regional mineral resource estimation," *Natural Resources Research*, vol. 20, no. 2, pp. 95–101, 2011.
- [22] Q. Cheng, "Application of a newly developed boost weights of evidence model(BoostWofE) for mineral resources quantitative assessments," *Journal of Jilin University (Earth Science Edition)*, vol. 42, pp. 1976–1984, 2012.
- [23] Q. Cheng, "Sequential weights of evidence as a machine learning model for mineral deposits prediction," in *Proceedings of the 15th Conference of IAMG*, pp. 157–161, Madrid, Spain, 2013.
- [24] Q. Cheng, "BoostWofE: a new sequential weights of evidence model reducing the effect of conditional dependency," *Mathematical Geosciences*, vol. 47, no. 5, pp. 591–621, 2015.
- [25] Q. Cheng, F. P. Agterberg, G. F. Bonham-Carter, and J. Sun, "Artificial intelligence modelling for integrating spatial patterns in mineral potential estimation with incomplete information," in *Proceedings of the 6th Canadian Conference of Surveying and Mapping*, vol. 1, pp. 206–219, Ottawa, Canada, 1994.
- [26] E. B. Andersen, *The Statistical Analysis of Categorical Data*, Springer-Verlag, Berlin, 3rd edition, 1994.
- [27] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA, USA, 1975.
- [28] R. Christensen, *Loglinear Models*, Springer-Verlag, Berlin, 1990.
- [29] H. Schaeben, "A mathematical view of weights-of-evidence, conditional independence, and logistic regression in terms of Markov random fields," *Mathematical Geosciences*, vol. 46, no. 6, pp. 691–709, 2014.
- [30] Y. Wei, J. Qiu, and H. K. Lam, "A novel approach to reliable output feedback control of fuzzy-affine systems with time delays and sensor faults," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 6, pp. 1808–1823, 2017.
- [31] Y. Wei, J. Qiu, P. Shi, and M. Chadli, "Fixed-order piecewise-affine output feedback controller for fuzzy-affine-model-based nonlinear systems with time-varying delay," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 4, pp. 945–958, 2017.
- [32] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, Wiley, New York, NY, USA, 3rd edition, 2013.
- [33] H. Schaeben and K. G. van den Boogaart, "Comment on "A conditional dependence adjusted weights of evidence model" by Minfeng Deng in Natural Resources Research 18 (2009), 249–258," *Natural Resources Research*, vol. 20, no. 4, pp. 401–406, 2011.
- [34] D. Zhang, F. Agterberg, Q. Cheng, and R. Zuo, "A comparison of modified fuzzy weights of evidence, fuzzy weights of evidence, and logistic regression for mapping mineral prospectivity," *Mathematical Geosciences*, vol. 46, no. 7, pp. 869–885, 2014.
- [35] F. P. Agterberg, *Geomathematics: Theoretical Foundations, Applications and Future Developments*, Springer, Heidelberg, 2014.
- [36] F. P. Agterberg, "Principles of probabilistic regional mineral resource estimation," *Journal of China University of Geosciences*, vol. 36, no. 2, pp. 189–200, 2011.
- [37] H. Schaeben, "Comment on "A comparison of modified fuzzy weights of evidence, fuzzy weights of evidence, and logistic regression for mapping mineral prospectivity" by Daojun Zhang, Frits Agterberg, Qiuming Cheng, and Renguang Zuo Math Geosci DOI 10.1007/s11004-013-9496-8," *Mathematical Geosciences*, vol. 46, no. 7, pp. 887–893, 2014.
- [38] F. P. Agterberg, G. F. Bonham-Carter, and D. F. Wright, "Statistical pattern integration for mineral exploration," in *Computer Applications in Resource Estimation*, G. Gaal and D. F. Merriam, Eds., pp. 1–21, Pergamon, Oxford, 1990.
- [39] G. F. Bonham-Carter, F. P. Agterberg, and D. F. Wright, "Integration of geological data sets for gold exploration in Nova Scotia," *Photogrammetric Engineering and Remote Sensing*, vol. 54, pp. 1585–1592, 1988.
- [40] W. Wang, J. Zhao, Q. Cheng, and J. Liu, "Tectonic-geochemical exploration modeling for characterizing geo-anomalies in southeastern Yunnan District, China," *Journal of Geochemical Exploration*, vol. 122, pp. 71–80, 2012.
- [41] D. J. Spiegelhalter and R. P. Knill-Jones, "Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology," *Journal of the Royal Statistical Society. Series A (General)*, vol. 147, no. 1, pp. 35–77, 1984.
- [42] H. Schaeben, "Potential modeling: conditional independence matters," *GEM - International Journal on Geomathematics*, vol. 5, no. 1, pp. 99–116, 2014.
- [43] F. P. Agterberg and G. F. Bonham-Carter, "Measuring the performance of mineral potential maps," *Natural Resources Research*, vol. 14, no. 1, pp. 1–17, 2005.
- [44] H. Good, *Probability and the Weighing of Evidence*, Griffin, London, 1950.
- [45] D. J. Spiegelhalter, "Uncertainty in expert systems," in *Artificial Intelligence and Statistics*, W. A. Gale, Ed., pp. 17–55, Addison-Wesley, Reading, MA, USA, 1986.
- [46] F. P. Agterberg and G. F. Bonham-Carter, "Logistic regression and weights of evidence modeling in mineral exploration," in *Proceedings of the 28th International Symposium on Computer Applications in the Mineral Industries*, pp. 483–490, Golden, CO, USA, 1999.

