Front. Earth Sci. 2016, 10(3): 389–408
DOI 10.1007/s11707-016-0595-y

RESEARCH ARTICLE

The quest for conditional independence in prospectivity modeling: weights-of-evidence, boost weights-of-evidence, and logistic regression

Helmut SCHAEBEN (✉), Georg SEMMLER

Department of Geophysics and Geoinformatics, TU Bergakademie Freiberg, Freiberg 09596, Germany

Abstract The objective of prospectivity modeling is prediction of the conditional probability of the presence $\mathsf{T} = 1$ or absence $\mathsf{T} = 0$ of a target T given favorable or prohibitive predictors $\mathbf{B}$, or construction of a two classes $\{0,1\}$ classification of T. A special case of logistic regression called weights-of-evidence (WofE) is geologists' favorite method of prospectivity modeling due to its apparent simplicity. However, the numerical simplicity is deceiving as it is implied by the severe mathematical modeling assumption of joint conditional independence of all predictors given the target. General weights of evidence are explicitly introduced which are as simple to estimate as conventional weights, i.e., by counting, but do not require conditional independence. Complementary to the regression view is the classification view on prospectivity modeling. Boosting is the construction of a strong classifier from a set of weak classifiers. From the regression point of view it is closely related to logistic regression. Boost weights-of-evidence (BoostWofE) was introduced into prospectivity modeling to counterbalance violations of the assumption of conditional independence even though relaxation of modeling assumptions with respect to weak classifiers was not the (initial) purpose of boosting. In the original publication of BoostWofE a fabricated dataset was used to "validate" this approach. Using the same fabricated dataset it is shown that BoostWofE cannot generally compensate lacking conditional independence whatever the consecutively processing order of predictors. Thus the alleged features of BoostWofE are disproved by way of counterexamples, while theoretical findings are confirmed that logistic regression including interaction terms can exactly compensate violations of joint conditional independence if the predictors are indicators.

Received November 11, 2015; accepted April 7, 2016

E-mail: schaeben@tu-freiberg.de

Keywords general weights of evidence, joint conditional independence, naïve Bayes model, Hammersley–Clifford theorem, interaction terms, statistical significance

1 Introduction

The objective of prospectivity modeling is to identify locations (pixels, voxels) $x \subset D$ in some domain of definition D for which the conditional probability $P\Big(\mathsf{T}(x) = 1|\mathbf{B}(x)\Big)$ of the presence, $\mathsf{T}(x) = 1$, of a well defined target type of ore mineralization, given favorable or prohibitive factors $\mathbf{B}(x)$ is a relative maximum. Of course, the major prerequisite for such predictions is a proper conceptual model of the specified ore mineralization. A proper conceptual model may be turned into a regression-type model using the factors as spatially referenced predictors. Generally, a model considers the predictor $\mathbf{B}(x) = \Big(\mathsf{B}_0(x),\mathsf{B}_1(x),...,\mathsf{B}_m(x)\Big)^{\mathsf{T}}$, with $\mathsf{B}_0(x) \equiv 1$ for all $x \subset D$, and assigns a parameter $\boldsymbol{\theta} = (\theta_0,...,\theta_m)^{\mathsf{T}}$, which quantifies by means of a link function $\mathcal{L}$ the extent of dependence of the conditional probability $P\Big(\mathsf{T}(x) = 1|\mathbf{B}(x)\Big)$ on the predictors, i.e.,

$$\mathcal{L}P\Big(\mathsf{T}(x) = 1|\mathbf{B}(x)\Big) = \mathbf{B}(x)^{\mathsf{T}}\boldsymbol{\theta}. \tag{1}$$

The target $\mathsf{T}(x)$ as well as the predictor $\mathbf{B}(x)$ refer to locations $x \subset D$ with areal or volumetric extent, pixels or voxels, which are assumed to provide their physical support. Once computed, predicted conditional probabilities and the associated prediction errors, respectively, will be assigned to them as additional properties.

Recent surveys of mathematical models and numerical methods for prospectivity modeling have been compiled in

two special issues, *Mineral prospectivity analysis and quantitative resource estimation* of Ore Geology Reviews 38(3), 121–304, guest edited by Kreuzer and Porwal (2010), and *GIS-based mineral potential modelling and geological data analyses for mineral exploration* of Ore Geology Reviews 71, 477–881, guest edited by Porwal and Carranza (2015). Agterberg (2014) devotes a major chapter to this topic. ArcGIS by ESRI features ArcSDM providing tools for prospectivity modeling. Despite its ubiquity in geological prospection, the mathematical assumptions to authorize an approach do not seem to be well communicated. Of special concern is the role of conditional independence which is not yet another version of stochastic independence but a concept on its own.

Moreover, a preference to publish case studies of prospectivity modeling applying "novel" procedures or "novel" variants can be observed. Whatever new method of prediction or classification has been developed in statistics, fuzzy logic or machine learning, it is being applied to case studies of prospectivity modeling, cf. the references of (Kreuzer and Porwal, 2010; Porwal and Carranza, 2015). Subsequently, these case studies give rise to plenty empirical comparisons, for instance (Harris and Pan, 1999; Harris et al., 2003; Porwal et al., 2010; Rodriguez-Galiano et al., 2015; Ford et al., 2016) which often conclude superior or inferior performance in one way or another. Clarification of the origins and considerations of the mutual relationships of various methods may render some differences in the results of their practical applications less surprising than others. Eventually, mathematical models and corresponding methods cannot be validated, their properties cannot be derived by way of case studies.

As a result, the practitioner of prospectivity modeling often seems lost in a vast variety of procedures, especially when exposing the essentials of his/her method of choice. The threefold objective of this communication is (i) to clarify the mathematical relationship of logistic regression, weights-of-evidence, and boost weights-of-evidence, (ii) to disprove the major claims put forward by Cheng (2012, 2015) with respect to his boost weights-of-evidence by virtue of counterexamples including the fabricated training dataset provided by Cheng (2015), and (iii) to exemplify theoretical findings that interaction terms of logistic regression models compensate lack of conditional independence.

Boosting is the affirmative answer to the mathematical problem whether it is possible to construct a strong classifier (with superior properties) from a set of weak classifiers (with poor properties). Boosting was not meant to relax modeling assumptions possibly required by classifiers. Boost weights-of-evidence (BoostWofE) (Cheng, 2015) is the most recent attempt to relax the mathematical modeling assumption of weights-of-evidence (Good, 1950, 1960, 1968; Minsky and Selfridge, 1961; Agterberg et al., 1990; Bonham-Carter, 1994; Schaeben, 2014c) which is joint conditional independence of all predictors given the target. Applying weights-of-evidence despite lacking conditional independence corrupts not only the predicted conditional probabilities but also their rank transforms, and thus the spatial pattern of prospectivity (Schaeben, 2014a). Boosted weights of evidence differ from ordinary weights by additive terms (Cheng, 2015, Eqs. (26) and (27), p. 602–603), introduced as approximations of conditional weights (Cheng, 2015, Eqs. (11) and (12), p. 597) taken from the ν-model (Polyakova and Journel, 2007) without citing it. For reasons of completeness the ν-model is recalled in the Appendix A.

Weights-of-evidence is an application of Bayes theorem for several variables, and the special case of logistic regression if the predictors $\mathbf{B}$ are nominal (categorical) and jointly conditional independent given the target T (Schaeben, 2014b). In turn, logistic regression (Reed and Berkson, 1929; Berkson, 1944; Hosmer et al., 2013) is the canonical generalization of Bayesian weights-of-evidence allowing for deviations from conditional independence of a restricted form. Applying Hammersley–Clifford theorem, it was proven that logistic regression including interaction terms corresponding to violations of conditional independence compensates this lack exactly and is optimum, i.e., recovers the true conditional probability, if the joint distribution of predictors and target is of log-linear form (Schaeben, 2014c).

Weights-of-evidence, novel boost weights-of-evidence, and classical logistic regression are discussed in basic mathematical terms, and then applied to the fabricated training dataset used by Cheng (2015) for the purpose of empirical comparison. To gain additional insight, the methods are also applied to the fabricated training datasets RANKIT used in earlier communications, e.g. (Schaeben, 2014a, b).

2 Fundamentals: stochastic independence, conditional independence

2.1 Definition

For a set $\{0,...,m\}$ of indexes, the $\otimes$-product denotes both the product of random variables Z_ℓ defined as $\otimes_{\ell=0}^{m}\ \mathsf{Z}_\ell = (\mathsf{Z}_0,...,\mathsf{Z}_m)$, and the product of their probability measures P_{Z_ℓ}. If the random variables Z_ℓ, $\ell = 0,...,m$, are independent, then the joint probability of any subset of random variables Z_ℓ can be factorized into the product of the individual probabilities, i.e.,

$$P_{\otimes_{\ell\in M}\ \mathsf{z}_\ell} = \underset{\ell\in M}{\otimes} P_{\mathsf{Z}_\ell}, \tag{2}$$

where M denotes any non-empty subset of the set $\{0,...,m\}$. In particular

$$P_{\mathbf{Z}} = P_{\otimes_{\ell=0}^{m}\mathsf{Z}_\ell} = \overset{m}{\underset{\ell=0}{\otimes}} P_{\mathsf{Z}_\ell}.$$

If the random variables Z_ℓ, $\ell = 1,...,m$, are conditionally independent given Z_0, then the joint conditional probability of any subset of random variables Z_ℓ given Z_0 can be factorized into the product of the individual conditional probabilities, i.e.,

$$P_{\otimes_{\ell\in M}\mathsf{Z}_\ell|\mathsf{Z}_0} = \bigotimes_{\ell\in M} P_{\mathsf{Z}_\ell|\mathsf{Z}_0}, \tag{3}$$

and in particular

$$P_{\otimes_{\ell=1}^{m}\mathsf{Z}_\ell|\mathsf{Z}_0} = \bigotimes_{\ell=1}^{m} P_{\mathsf{Z}_\ell|\mathsf{Z}_0}.$$

Given Z_0, the observable variables $\mathsf{Z}_\ell, \ell \in M$, become independent. In practice, often the interpretation of Z_0 as common cause for Z_ℓ, $\ell \in M$, applies. Thus, conditional independence is a probabilistic approach to causality (Suppes, 1970; Dawid, 1979, 2004, 2007; Pearl, 2009; Chalak and White, 2012) while correlation, i.e., linear dependence, is not. Correlated random variables may be conditionally independent or not, conditionally independent random variables may be (significantly) correlated or not. Independence does not imply conditional independence and vice versa; pairwise conditional independence does not imply joint conditional independence.

Weak conditional independence was introduced by Wong and Butz (1999), and elaborated on by Butz and Sanscartier (2002). The definition of weak conditional independence by Cheng (2015) is irrelevant at this time as a statistical significance test is not provided.

2.2 Testing conditional independence

If the predictor variables B_ℓ, $\ell = 1,...,m$, and the target variable T are indicator variables, i.e., binary, then the joint probability is of log-linear form. If the predictor variables are jointly conditionally independent given the target variable, then by virtue of the Hammersley–Clifford theorem the log–linear model factorized into terms corresponding to the target variable T, the individual predictor variables B_ℓ, and the individual products $\mathsf{T}\otimes\mathsf{B}_\ell$, $\ell = 1,...,m$, is sufficiently large to represent the joint probability (Schaeben, 2014c). The sufficient size of such a factorized form of an appropriate log–linear model can be turned into the null-hypothesis of a statistical significance test. Thus, if the likelihood ratio test of this null-hypothesis leads to its possible rejection, then the assumption of joint conditional independence can be rejected, too.

3 Logistic regression, weights-of-evidence, boost weights-of-evidence

To fit a model, i.e., to estimate the model parameters $\boldsymbol{\theta}$ of model ansatz Eq. (1), data within a training region are required. However, in contrast to geostatistics (Chilès and Delfiner, 2012), and to their major detriment, the methods of prospectivity modeling considered here do not consider spatially induced dependencies between the target and the predictors nor between predictor variables themselves. Alternatives have been suggested by van den Boogaart and Schaeben (2012) and Tolosana-Delgado et al. (2014).

A proper definition of the notion of spatial association does not exist in the realm of potential mapping or prospectivity modeling. In fact, the classical assumption of independently identically distributed random variables applies, distributions do not depend on location. Therefore, any spatial reference can be dropped, and models of the form

$$\mathcal{L}P(\mathsf{T}=1|\mathbf{B}) = \mathbf{B}^{\mathsf{T}}\boldsymbol{\theta}, \tag{4}$$

are considered, only. Instead of an illusive spatial association, the ordinary correlation matrix may provide some instructive information how to choose the predictors of a proper regression model.

3.1 Logistic regression

Conditional expectation of a binary random target variable T given a $(m+1)$–variate random predictor variable $\mathbf{B} = (\mathsf{B}_0,\mathsf{B}_1,...,\mathsf{B}_m)^{\mathsf{T}}$ with $\mathsf{B}_0 \equiv 1$ is equal to a conditional probability, i.e.,

$$\mathrm{E}(\mathsf{T}|\mathbf{B}) = P(\mathsf{T}=1|\mathbf{B}).$$

Neglecting the error term as often the ordinary logistic regression model (without interaction terms) of a logit-transformed conditional probability in terms of a linear combination of predictors reads

$$\text{logit } P(\mathsf{T}=1|\mathbf{B}) = \beta_0 + \sum_\ell \beta_\ell \mathsf{B}_\ell, \tag{5}$$

which can be rewritten in terms of a conditional probability as

$$P(\mathsf{T}=1|\mathbf{B}) = \Lambda\left(\beta_0 + \sum_\ell \beta_\ell \mathsf{B}_\ell\right), \tag{6}$$

as the logistic function denoted Λ is the inverse of the logit-transform.

If the joint probability of the indicator target variable and the predictor variables is of log–linear form and all predictor variables are conditionally independent given the target variable, then the conditional probability of the target variable given the predictors is of the form of Eq. (6) of the ordinary logistic regression model. Thus, in this case the ordinary logistic regression model is optimum. In particular, it is optimum if the predictor variables are categorical or discrete and jointly conditionally independent given the target variable (Schaeben, 2014a).

The logistic regression model with interaction terms reads in terms of a logit

$$\text{logit } P(\mathsf{T}=1|\mathbf{B}) = \beta_0 + \sum_{\ell}\beta_\ell \mathsf{B}_\ell + \sum_{\ell_i,\ldots,\ell_j}\beta_{\ell_i,\ldots,\ell_j}\mathsf{B}_{\ell_i}\ldots\mathsf{B}_{\ell_j})$$

and in terms of a probability

$$P(\mathsf{T}=1|\mathbf{B}) = \Lambda\left(\beta_0 + \sum_{\ell}\beta_\ell \mathsf{B}_\ell + \sum_{\ell_i,\ldots,\ell_j}\beta_{\ell_i,\ldots,\ell_j}\mathsf{B}_{\ell_i}\ldots\mathsf{B}_{\ell_j}\right). \quad (7)$$

If the joint probability of the indicator target variable and the predictor variables is of log–linear form including interaction terms corresponding to lacking conditionally independence given the target variable, then the conditional probability of the target variable given the predictors is of the form of Eq. (7) of the logistic regression model enlarged by interaction terms. Thus, in this case the augmented logistic regression model is optimum. In particular, for categorical or discrete predictor variables, interaction terms can compensate any lack of conditional independence exactly, i.e., logistic regression with interaction terms is optimum in case of lacking conditional independence (Schaeben, 2014a).

Given the sample $b_{\ell,i}$, t_i, $i = 1,\ldots,n$, $\ell = 1,\ldots,m$, the parameters of the logistic regression model are estimated with well established, well understood methods based on probability, and encoded in any major statistical software package applying (i) the method of maximum likelihood estimation numerically (ii) realized with Fisher scoring algorithm (a form of Newton-Raphson, a special case of iteratively reweighted least squares algorithm) ensuring nice statistical properties of the estimates like consistency, asymptotic normality, and efficiency of the estimates.

The properties of a fitted logistic regression model are assessed by the significance of the estimated regression parameters, and by Akaike information criterion (AIC)

$$\text{AIC} = -\ln\,(L) + 2k,$$

where L is the maximized value of the likelihood function, and k be the total number of estimated parameters of the model. AIC provides a measure of the relative quality of a statistical model for a given training dataset. More measures to assess the fit of a logistic model including the receiver operating characteristic curve are discussed in (Hosmer et al., 2013). Despite a small AIC, a logistic regression model is discarded if at least one of its fitted parameters is not significant to prevent overfitting to the given training dataset and poor predicting performance for any other application dataset.

A fitted regression model is mathematically authorized for prediction if the model assumptions are satisfied, if it resembles the true relationship of the target and the predictors, and if their fitted regression parameters are significant. Large errors of predictions render a fitted model inappropriate for prediction in any case.

3.2 Weights-of-evidence

Weights-of-evidence is an application of Bayes theorem. Bayes theorem for several variables $\mathsf{B}_0,\mathsf{B}_1,\ldots,\mathsf{B}_m$, $\mathsf{B}_0 \equiv 1$, reads in terms of odds

$$\begin{aligned}&\mathrm{O}(\mathsf{T}=1|\mathbf{B}=\mathbf{b})\\ &= \mathrm{O}(\mathsf{T}=1)\frac{\prod_{\ell=1}^{m} P\left(\mathsf{B}_\ell|\otimes_{j=0}^{\ell-1}\mathsf{B}_j = (1,b_1,\ldots,b_{\ell-1})\wedge\mathsf{T}=1\right)}{\prod_{\ell=1}^{m} P\left(\mathsf{B}_\ell|\otimes_{j=0}^{\ell-1}\mathsf{B}_j = (1,b_1,\ldots,b_{\ell-1})\wedge\mathsf{T}=0\right)}\\ &= \mathrm{O}(\mathsf{T}=1)\prod_{\ell=1}^{m}F_\ell\end{aligned}$$

with Bayes factors (Good, 1968)

$$F_\ell = \frac{P\left(\mathsf{B}_\ell|\otimes_{j=0}^{\ell-1}\mathsf{B}_j = (1,b_1,\ldots,b_{\ell-1})\wedge\mathsf{T}=1\right)}{P\left(\mathsf{B}_\ell|\otimes_{j=0}^{\ell-1}\mathsf{B}_j = (1,b_1,\ldots,b_{\ell-1})\wedge\mathsf{T}=0\right)},$$

$$\ell = 1,\ldots,m. \quad (8)$$

Applying the logit transform yields

$$\text{logit}P(\mathsf{T}=1|\mathbf{B}) = \text{logit}P(\mathsf{T}=1) + \sum_{\ell=1}^{m}\ln\,F_\ell. \quad (9)$$

3.2.1 General weights of evidence

Let V_ℓ, $\ell = 1,\ldots, m$, denote the set of ℓ-variations of the set $\{0,1\}$, and $v_\ell = 2^\ell$ its total number of elements. Let $\mathbf{Z}_\ell = \otimes_{j=1}^{\ell}\mathsf{B}_j$, $\ell = 1,\ldots,m$. Given a realization $\mathbf{b}_{k,\ell} \in V_\ell$, $k = 1,\ldots,v_\ell$ of $\mathbf{Z}_\ell$, $\ell = 1,\ldots,m$, F_ℓ is rewritten for $\ell = 2,\ldots,m$ more detailed as

$$\begin{aligned}F_\ell &= F(\boldsymbol{b}_{k,\ell})\\ &= \frac{P(\mathbf{Z}_\ell = \mathbf{b}_{k,\ell}\wedge\mathsf{T}=1)}{P(\mathbf{Z}_\ell = \mathbf{b}_{k,\ell}\wedge\mathsf{T}=0)}\frac{P(\mathbf{Z}_{\ell-1} = \mathbf{b}_{k,\ell-1}\wedge\mathsf{T}=0)}{P(\mathbf{Z}_{\ell-1} = \mathbf{b}_{k,\ell-1}\wedge\mathsf{T}=1)},\end{aligned}$$

$$k = 1,\ldots,v_\ell,$$

where $\mathbf{b}_{k,\ell-1}$ agrees with $\mathbf{b}_{k,\ell}$ in the first $(\ell-1)$ entries. Then

$$F(\mathsf{Z}_\ell) = \sum_{k=1}^{v_\ell}F_\ell(\mathbf{b}_{k,\ell})\mathrm{II}_{\{\mathrm{b}_{k,\ell}\}}(\mathsf{Z}_\ell),$$

where $\mathrm{II}_{\{\mathbf{b}_{k,\ell}\}}$ denotes the indicator function with respect to the set $\{\mathbf{b}_{k,\ell}\}$ containing the vector $\mathbf{b}_{k,\ell}$. Then

$$\mathrm{O}(\mathsf{T}=1|\mathbf{B}) = \mathrm{O}(\mathsf{T}=1)\prod_{\ell=1}^{m}\sum_{k=1}^{v_\ell}F(\mathbf{b}_{k,\ell})\mathrm{II}_{\{\mathrm{b}_{k,\ell}\}}(\mathsf{Z}_\ell), \quad (10)$$

and

$$\text{logit } P(\mathsf{T}=1|\mathbf{B}) = \text{logit } P(\mathsf{T}=1) + \sum_{\ell=1}^{m}\sum_{k=1}^{v_\ell} \ln F(\mathbf{b}_{k,\ell}) \text{II}_{\{\mathbf{b}_{k,\ell}\}}(\mathsf{Z}_\ell), \quad (11)$$

where $\ln F(\mathbf{b}_{k,\ell})$ of Eq. (11) could be referred to as general weights of evidence, emphasizing that conditional independence is not required. Since the tacit assumption as usually is that the Bayes factors and their logarithms exist, i.e., that none of the involved probabilities is 0 or 1, they can be estimated elementarily by corresponding frequencies, i.e., by counting occurrences within a given training region.

3.2.2 Weights of evidence assuming conditional independence

Assuming joint conditional independence of all predictors $\mathsf{B}_1,...,\mathsf{B}_m$ given the target T simplifies F_ℓ to

$$F_\ell^{\text{CI}} = \frac{P(\mathsf{B}_\ell|\mathsf{T}=1)}{P(\mathsf{B}_\ell|\mathsf{T}=0)}, \ell = 1,...,m,$$

and results in

$$\text{logit } P(\mathsf{T}=1|\mathbf{B}) = \text{logit } P(\mathsf{T}=1) + \sum_{\ell=1}^{m} \ln \frac{P(\mathsf{B}_\ell \mid \mathsf{T}=1)}{P(\mathsf{B}_\ell \mid \mathsf{T}=0)},$$

$$= \text{logit } P(\mathsf{T}=1) + \sum_{\ell=1}^{m} W_\ell \quad (12)$$

with

$$W_\ell = \ln \frac{P(\mathsf{B}_\ell \mid \mathsf{T}=1)}{P(\mathsf{B}_\ell \mid \mathsf{T}=0)}, \quad (13)$$

provided that neither numerator nor denominator in the definition of W_ℓ, Eq. (13), vanish. Eq. (12) tells us how to update and improve the unconditional logit$P(\mathsf{T}=1)$ considering information provided by $\mathsf{B}_1,...,\mathsf{B}_m$ assuming their joint conditional independence given the target. Introducing

$$W_\ell^{(1)} = \ln \frac{P(\mathsf{B}_\ell = 1|\mathsf{T}=1)}{P(\mathsf{B}_\ell = 1|\mathsf{T}=0)},$$

$$W_\ell^{(0)} = \ln \frac{P(\mathsf{B}_\ell = 0|\mathsf{T}=1)}{P(\mathsf{B}_\ell = 0|\mathsf{T}=0)}. \quad (14)$$

Eq. (12) can be rewritten in terms of predictor variables as

$$\text{logit } P(\mathsf{T}=1|\mathbf{B}) = \text{logit } P(\mathsf{T}=1) + \sum_{\ell=1}^{m} \left(W_\ell^{(1)} \mathsf{B}_\ell + W_\ell^{(0)} (1-\mathsf{B}_\ell) \right)$$

$$= \text{logit } P(\mathsf{T}=1) + W^{(0)} + \sum_{\ell=1}^{m} C_\ell \mathsf{B}_\ell, \quad (15)$$

with contrasts

$$C_\ell = W_\ell^{(1)} - W_\ell^{(0)}, \ell = 1,...,m,$$

and $W^{(0)} = \sum_{\ell=1}^{m} W_\ell^{(0)}$. Besides less involved weights, the major difference between the two models, conventional weights-of-evidence assuming joint conditional independence, Eq. (12), and general weights-of-evidence, Eq. (11), is the presence of interaction terms, i.e., product terms of predictors, in the latter. Moreover, comparing Eq. (5) and Eq. (15) reveals that in case of joint conditional independence of all predictors given the target variable the regression coefficients simplify to

$$\beta_0 = \text{logit } P(\mathsf{T}=1) + W^{(0)}, \ \beta_\ell = C_\ell, \ell = 1,...,m,$$

(Schaeben, 2014a, b). Obviously the model parameters become independent of one another, and can be estimated by mere counting. This special case of a logistic regression model is usually referred to as the method of weights of evidence. Its practical application is restricted by the modeling assumption of joint conditional independence of all predictors given the target. The other way round, logistic regression is the canonical generalization of weights of evidence.

3.3 Boost weights-of-evidence (BoostWofE)

As with respect to weights-of-evidence, numerous attempts aim at relaxing its modeling assumption of joint conditional independence and subsequent corrections of the weights, e.g., the multiplicative τ- and the additive ν-correction of weights (Journel, 2002; Polyakova and Journel, 2007; Krishnan, 2008).

BoostWofE has recently been introduced by Cheng (2015) as combining elements of weights-of-evidence and AdaBoost (Freund and Schapire, 1997; Freund and Schapire, 1999; Hastie et al., 2009) to simplify estimation of the general weights of evidence, Eq. (11), and the additive ν correction term of Polyakova and Journel (2007), respectively (Cheng, 2015, p. 597).

In general, boosting turns several weak learners into a single strong learner. For the two-class problem, boosting can be viewed as an approximation to additive modeling on the logistic scale using maximum Bernoulli likelihood as a criterion (Friedman et al., 2000) used in logistic regression. AdaBoost is a linear classifier with all its desirable properties, its output converges to the logarithm of the likelihood ratio (Šochman and Matas, 2004). AdaBoost fits an additive logistic regression model, using a criterion similar to the binomial log-likelihood. LogitBoost directly optimizes the binomial log-likelihood (Friedman et al., 2000).

Weights-of-evidence is the special case of logistic regression characterized by the additional assumption of

jointly conditionally independent indicator predictors. Boosting weights of evidence appears to aim at improving especially weak learners depending on joint conditional independence while the first canonical improvement would be to proceed from weights-of-evidence to logistic regression.

Boost weights-of-evidence (Cheng, 2015) processes indicator predictors to update a prior unconditional probability sequentially with the conventional weight of evidence assigned to the first predictor, and successively boosted weights assigned to subsequent predictors (Cheng, 2015, Eq. (26), p. 602) resulting in

$$^{\text{boost}}W_{\ell}^{(i)} = W_{\ell}^{(i)} + Q_{\ell}^{(i)}, i = 0,1, \ell = 1,...,m, \qquad (16)$$

with $Q_1^{(i)} = 0$, and $Q_{\ell}^{(i)}$ presumably approximating ln $\nu_{\ell}^{(i)}$, the correction term provided by the ν-model (Polyakova and Journel, 2007), Eq. (A2) of the Appendix A, by a weighted mean of corresponding conditional probabilities (Cheng, 2015, p. 597). The boost terms Q_{ℓ} are explicitly given as sum of $(\ell-1)$ terms, where each term is given as logarithm of a ratio of sums of two ratios of conditional probabilities each (Cheng, 2015, Eqs. (26) and (27), pp. 602–603). For more details the reader is referred to Appendix B.

In a similar way as the additive ν modifications of conventional weights of evidence (Appendix A), the additive modifications by successive boosting (Cheng, 2015) of weights of evidence may allow for some small deviations from joint conditional independence in a very restricted form. They cannot emulate the effect of multiplicative interaction terms of predictors included in logistic regression models. Since $Q_1 = 0$, there are as many different boost weights-of-evidence models as permutations of m predictors B_{ℓ}, $\ell = 1,...,m$, i.e., $m!$ different boost models. Whether the procedure presented by Cheng (2015) to estimate the boost weights of evidence permutes correspondingly was not addressed; there is no obvious reason to assume that the order of predictors is irrelevant.

From an application with fabricated training data and a case study with observed training data, (Cheng, 2015, p. 618) concludes that his novel boost weights-of-evidence method can significantly reduce the effect of conditional dependence in a simple and intuitive way as derived by Cheng (2015, pp. 596–607), in fact in a more generic way than other approaches (Cheng, 2015, p. 620).

4 Empirical comparison

Since case studies cannot generally provide insight in the method applied, instructive examples with fabricated or simulated data are used to exemplify, expose and confirm theoretical findings. Here we primarily use the same dataset Q as Cheng (2015) for the particular purposes (i) to criticize and refute validation by case studies in general, and (ii) to disprove the major claims put forward by Cheng (2012, 2015) that BoostWofE significantly reduces the effect of lack of joint conditional independence (Cheng, 2015, p. 618). Since the assumption of conditional independence is only mildly violated for the dataset Q, we demonstrate the effect of its serious violation using once more the dataset RANKIT (Schaeben, 2014a, b).

4.1 Chengs's (2015) fabricated training dataset Q

The training dataset Q is taken from the publication (Cheng, 2015). The digital map images of spatial distributions of three indicator predictor variables B_{ℓ}, $\ell = 1,2,3$, and the indicator target variable T given their realizations $b_{\ell,i}, t_i, i = 1,...,100, \ell = 1,...,3$, are displayed in Fig. 1.

For three indicator predictor variables there are eight different combinations of their joint possible realizations. Then there are eight corresponding conditional frequencies $h\Big(\mathsf{T}=1|(\mathsf{B}_1,\mathsf{B}_2,\mathsf{B}_3)=(b_1,b_2,b_3)\Big), b_1,b_2,b_3 = 0,1$, referred to as ground truth of the training dataset Q because these conditional frequencies determined by counting are unbiased estimates of the corresponding conditional probabilities $P\Big(\mathsf{T}=1|(\mathsf{B}_1,\mathsf{B}_2,\mathsf{B}_3)=(b_1,b_2,b_3)\Big)$. The spatial distribution of the conditional frequencies is displayed in Fig. 2, the numbers are compiled in Table 15.

4.1.1 Pearson and Kendall correlation

A measure of spatial association like the variogram of geostatistics is unknown in all conventional methods of prospectivity modeling. Since the general assumption of identically independent distributed random variables applies to all conventional methods of prospectivity modeling, inspection of the correlation matrices is reasonable. In case of indicator predictors and data, respectively, it is sufficient to check any of the three conventional correlations, e.g. Table 1, as Pearson, Kendall, and Spearman correlation coefficients agree.

Statistical tests of Kendall correlation coefficients are summarized in Table 2 and Table 3.

Thus, B_1 and B_2 are inferred to be significantly correlated with T, while B_3 seems to be rather uncorrelated with T.

The tests suggest that B_1 is neither correlated with B_2 nor with B_3, and that B_2 and B_3 are significantly correlated for every level of significance α>3.261456e–06.

4.1.2 Conditional independence

Checking for instance for $(\mathsf{B}_1,\mathsf{B}_2,\mathsf{B}_3) = (1,1,1)$ whether the observed joint conditional frequencies interpreted as elementary estimates of the corresponding probabilities can be factorized given $\mathsf{T} = 1$

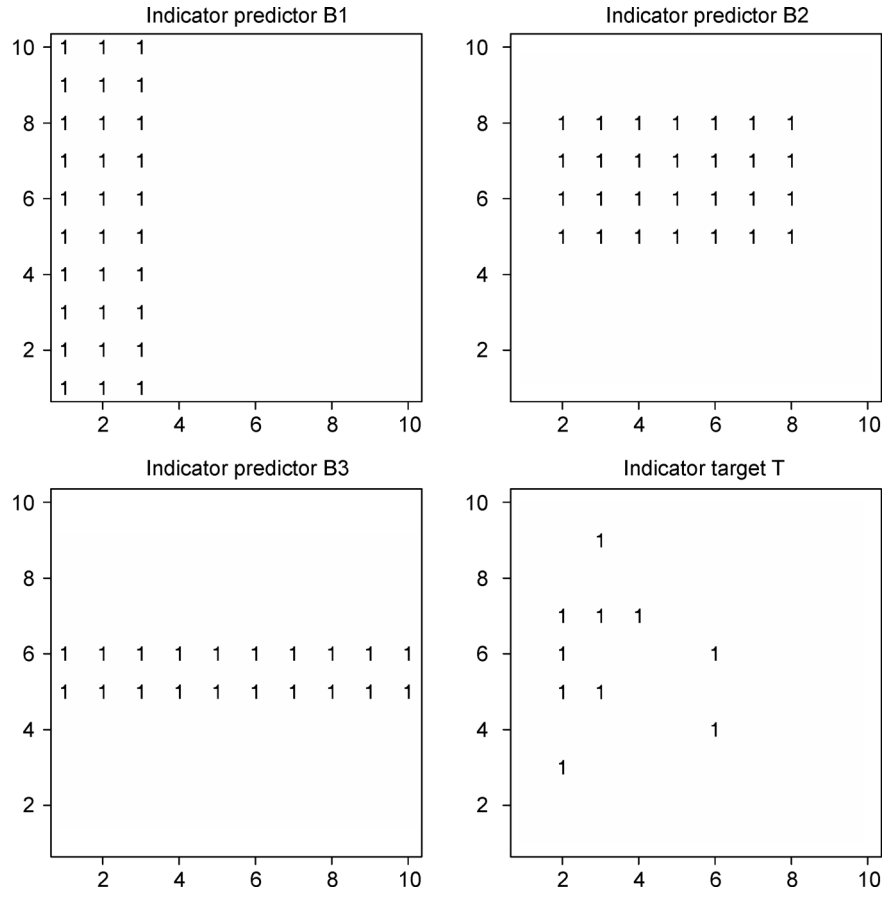

Fig. 1 Spatial distribution of three indicator predictor variables B_1, B_2, and B_3, and the indicator target variable T of training dataset Q. Blancs are used instead of displaying 0.

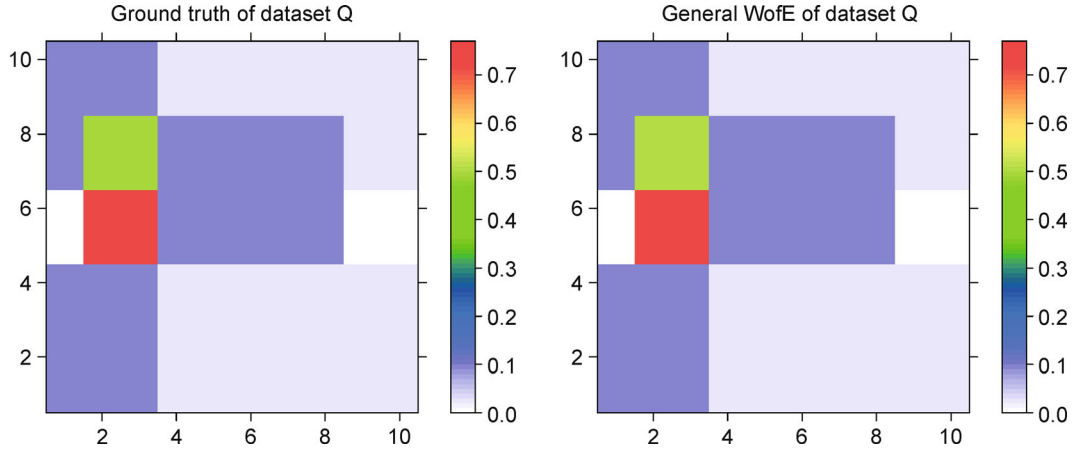

Fig. 2 Spatial distribution of the ground truth according to elementary estimation by counting in the training dataset Q (left), spatial distribution of conditional probabilities estimated with general weights of evidence without assuming conditional independence (right). The match is perfect.

Table 1 Pearson correlation matrix of the training dataset Q

	B_1	B_2	B_3	T
B_1	1.000	–0.019	0.000	0.291
B_2	–0.019	1.000	0.468	0.312
B_3	0.000	0.468	1.000	0.167
T	**0.291**	**0.312**	0.167	1.000

Table 2 Significance tests of Kendall correlations B_i, T, $i = 1,2,3$, for training dataset Q

Random variables B_i, T	*p*-value
B_1, T	0.003
B_2, T	0.001
B_3, T	0.097

Table 3 Significance tests of Kendall correlations B_i, B_j, $i, j = 1,2,3$, $i<j$, for training dataset Q

Random variables B_i, B_j	p-value
B_1, B_2	0.846
B_1, B_3	1.000
B_2, B_3	0.000

$$\hat{P}\Big((\mathsf{B}_1,\mathsf{B}_2,\mathsf{B}_3) = (1,1,1)|\mathsf{T} = 1\Big) = 0.300,$$

$$\hat{P}(\mathsf{B}_1=1|\mathsf{T}=1)*\hat{P}(\mathsf{B}_2=1|\mathsf{T}=1)*\hat{P}(\mathsf{B}_3=1|\mathsf{T}=1)=0.196,$$

or $\mathsf{T} = 0$

$$\hat{P}\Big((\mathsf{B}_1,\mathsf{B}_2,\mathsf{B}_3) = (1,1,1)|\mathsf{T} = 0\Big) = 0.011,$$

$$\hat{P}(\mathsf{B}_1=1|\mathsf{T}=0)*\hat{P}(\mathsf{B}_2=1|\mathsf{T}=0)*\hat{P}(\mathsf{B}_3=1|\mathsf{T}=0)=0.010,$$

suggests that the mathematical modeling assumption of joint conditional independence is violated to a minor extent only. Whether these minor deviations are statistically significant or not is the objective of a corresponding statistical test.

In terms of random variables joint conditional independence of indicator predictor variables given the indicator target variable can be tested with reference to a corresponding log–linear model. The full log–linear model is of course sufficiently large as all variables are indicator variables, and therefore the joint probability, i.e., all contingency tables, can be represented as log–linear model.

The statistical test of joint conditional independence referring to the three-terms log–linear model of Table 4 leads to infer that the null–hypothesis of joint conditional independence given T can reasonably be rejected for all levels of significance $\alpha>0.008$ with respect to the likelihood ratio or $\alpha>0.003$ with respect to Pearson statistic, i.e., the modeling assumption of joint conditional independence is significantly violated. However, the statistical tests of pairwise conditional independence referring to the corresponding two-terms log–linear models of Table 5, Table 6, and Table 7, respectively, reveal that only B_2 and B_3 are significantly not conditional independent given T, while conditional independence of B_1 and B_2, and B_1 and B_3, respectively, given T is violated to a statistically insignificant extent only.

In Cheng (2015) it is merely concluded that joint conditional independence does not apply as pairwise conditional independence of B_2 and B_3 given T does not apply.

4.1.3 Application of general weights-of-evidence

Applying the general weights of evidence, Eq. (11), estimated by counting as conventional weights of evidence results in

Table 4 Significance test of joint conditional independence referring to a log–linear model for training dataset Q

$$\text{loglm}\Big(\text{formula} = \sim \mathsf{B}_1*\mathsf{T} + \mathsf{B}_2*\mathsf{T} + \mathsf{B}_3*\mathsf{T},\ \text{data} = \text{xtabs}(.,\text{Q})\Big)$$

Statistics	χ^2	df	$P(>\chi^2)$
Likelihood ratio	20.7374	8	0.007
Pearson	23.6850	8	0.002

Table 5 Significance test of conditional independence of B_1 and B_2 given T

$$\text{loglm}\Big(\text{formula} = \sim \mathsf{B}_1*\mathsf{T} + \mathsf{B}_2*\mathsf{T},\ \text{data} = \text{xtabs}(.,\text{Q}[,-3])\Big)$$

Statistics	χ^2	df	$P(>\chi^2)$
Likelihood ratio	2.022720	2	0.363
Pearson	1.851326	2	0.396

Table 6 Significance test of conditional independence of B_1 and B_3 given T

$$\text{loglm}\Big(\text{formula} = \sim \mathsf{B}_1*\mathsf{T} + \mathsf{B}_3*\mathsf{T},\ \text{data} = \text{xtabs}(.,\text{Q}[,-2])\Big)$$

Statistics	χ^2	df	$P(>\chi^2)$
Likelihood ratio	0.5800827	2	0.748
Pearson	0.5531054	2	0.758

Table 7 Significance test of conditional independence of B_2 and B_3 given T

$$\text{loglm}\Big(\text{formula} = \sim \mathsf{B}_2{*}\mathsf{T} + \mathsf{B}_3{*}\mathsf{T},\ \text{data} = \text{xtabs}(.,\text{Q}[,-1])\Big)$$

Statistics	χ^2	df	$P(>\chi^2)$
Likelihood ratio	18.30563	2	0.000
Pearson	19.54423	2	0.000

$$\begin{aligned}
&O(\mathsf{T}=1|\mathsf{B}_1,\mathsf{B}_2,\mathsf{B}_3) = 0.111{*}[2.739\mathsf{B}_1 + 0.402(1-\mathsf{B}_1)]\\
&{*}[5.476\mathsf{B}_2\mathsf{B}_1 + 2.481\mathsf{B}_2(1-\mathsf{B}_1) + 0.328(1-\mathsf{B}_2){*}\mathsf{B}_1 + 0.455(1-\mathsf{B}_2){*}(1-\mathsf{B}_1)]\\
&{*}[1.800\mathsf{B}_3\mathsf{B}_1\mathsf{B}_2 + 1.000\mathsf{B}_3(1-\mathsf{B}_1)\mathsf{B}_2 + 0.600(1-\mathsf{B}_3)\mathsf{B}_1\mathsf{B}_2 + 1.111(1-\mathsf{B}_3)\mathsf{B}_1(1-\mathsf{B}_2)\\
&+1.000(1-\mathsf{B}_3)(1-\mathsf{B}_1)\mathsf{B}_2 + 1.088(1-\mathsf{B}_3)(1-\mathsf{B}_1)(1-\mathsf{B}_2)]
\end{aligned}$$

and perfectly recovers the ground truth, of course, cf. Fig. 2. It should be noted that the interaction terms $\mathsf{B}_3\mathsf{B}_1(1-\mathsf{B}_2)$ and $\mathsf{B}_3(1-\mathsf{B}_1)(1-\mathsf{B}_2)$ are not included as their weights vanish. The general weights of evidence could be used for the purpose of prediction in the same way as conventional weights. However, as with conventional weights of evidence, a measure for the predictive power or the reliability of the prediction is missing. Caution seems to be appropriate, as the general weights of evidence may result in overfitting.

4.1.4 Application of conventional weights-of-evidence assuming conditional independence

Despite the obvious violation of the required mathematical modeling assumption of joint conditional independence of all predictors given the target, weights-of-evidence is applied to the training dataset Q and results in weights and contrasts[1)] as compiled in Table 8.

Table 8 Numerical results of Wof3E applied to training dataset Q despite violation of joint conditional independence

	B_1	B_2	B_3	$\sum_\ell$
$W_\ell^{(1)}$	1.008	1.099	0.811	
$W_\ell^{(0)}$	–0.909	–0.938	–0.315	–2.162
C_ℓ	1.917	2.037	1.126	

With

$$P(\mathsf{T}=1) = 0.100,\ O(\mathsf{T}=1) = 0.111,$$

$$\ln\Big(O(\mathsf{T}=1)\Big) = -2.197$$

and

$$W^{(0)} = -2.162,\ \ln\Big(O(\mathsf{T}=1)\Big) + W^{(0)} = -4.359$$

the weights-of-evidence model reads explicitly

$$\begin{aligned}
&\hat{P}_{\text{Wof3E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3)\\
&= \Lambda(-4.359 + 1.917\mathsf{B}_1 + 2.037\mathsf{B}_2 + 1.126\mathsf{B}_3). \quad (17)
\end{aligned}$$

Comparing the total number of occurrences of the target event $\mathsf{T}=1$ given by the sum of the realizations t_i, $i = 1,...,100$, with the sum of all estimated conditional probabilities (interpreted as estimated total number of occurrences of the target event $\mathsf{T}=1$ (Cheng, 2015))

$$\begin{aligned}
&\sum t_i = 10, \sum \hat{P}_{\text{Wof3E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = 10.257,\\
&\sum \hat{P}_{\text{Wof3E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) - \sum t_i = 0.257, \quad (18)
\end{aligned}$$

which is the test statistic of the so-called "new omnibus test" of conditional independence (Agterberg and Cheng, 2002, p. 252), confirms that joint conditional independence is disturbed to a small extent only.

For the two-terms weights-of-evidence model

$$\hat{P}_{\text{Wof2E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2) = \Lambda(-4.359 + 1.917\mathsf{B}_1 + 2.037\mathsf{B}_2), \quad (19)$$

the difference

$$\sum \hat{P}_{\text{Wof2E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2) - \sum t_i = -0.611, \quad (20)$$

and reveals in particular a change of sign of the error compared to the three-terms weights-of-evidence model of Eq. (17).

Comparing digital map images of Fig. 3 displaying the ground truth estimated elementarily with conditional

1) The figures 0.898 and –0.477 for the conventional weights of B_2 given with the first and second equation of Eq. (49) by Cheng (2015, p. 612) are wrong, probably typos. Consequently, the figures for the conditional probabilities estimated with conventional weights of evidence given in the right column of Eq. (52) by Cheng (2015, p. 613) are erroneous, too.

frequencies, and the conditional probabilities $\hat{P}_{\mathrm{Wof3E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3)$ and $\hat{P}_{\mathrm{Wof2E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2)$, respectively, estimated with weights of evidence reveals that weights-of-evidence with three predictors yields a corrupted pattern of predicted conditional probabilities due to the violation of the modeling assumption. Weights-of-evidence with the two predictors B_1 and B_2, for which the null-hypothesis of conditional independence given T could not reasonably be rejected, cf. Table 5, results in a simplified pattern reflecting merely the spatial distribution of the two predictors used for prediction.

4.1.5 Application of BoostWofE

Processing the predictor variables in the same order B_1, B_2,B_3 as (Cheng, 2015) the weights and the contrast with respect to the predictor B_1 remain unchanged, while the weights and the contrasts with respect to predictors B_2 and B_3 are affected by boosting[1)], cf. Table 9.

With

$$^{\mathrm{boost123}}W^{(0)} = -1.876,$$

$$\ln\Big(O(\mathsf{T}=1)\Big)+{}^{\mathrm{boost123}}W^{(0)} = -4.073,$$

the Boost**123**WofE model by Cheng (2015) reads explicitly

$$^{\mathrm{boost123}}\hat{P}_{\mathrm{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3)$$

$$= \Lambda(-4.073 + 1.917\mathsf{B}_1 + 2.192\mathsf{B}_2 + 0.034\mathsf{B}_3), \quad (21)$$

and results in

$$\sum {}^{\mathrm{boost123}}\hat{P}_{\mathrm{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = 9.919,$$

$$\sum {}^{\mathrm{boost123}}\hat{P}_{\mathrm{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) - \sum t_i = -0.081. \quad (22)$$

Comparing WofE Eq. (17) and Boost**123**WofE Eq. (21), in this example boosting as introduced by Cheng (2015) puts less weight on B_3 and more weight on B_2 resulting in a largely improved error Eq.(22) of the same sign as the error Eq.(20) of the two-terms weights-of-evidence model Eq.(19) .

Comparing the map images of Fig. 4 leads to recognize that the result of Boost**123**WofE largely agrees with the result of Wof2E, the application of weights-of-evidence considering B_1 and B_2 only. The map images just confirm the comparison of the two models given explicitly with Eq. (19) and Eq. (21), respectively.

Next, the order of boosting is changed from $(\mathsf{B}_1,\mathsf{B}_2,\mathsf{B}_3)$ to $(\mathsf{B}_2,\mathsf{B}_1,\mathsf{B}_3)$ and $(\mathsf{B}_3,\mathsf{B}_1,\mathsf{B}_2)$, respectively.

For $(\mathsf{B}_2,\mathsf{B}_1,\mathsf{B}_3)$ Table 10 gives the corresponding figures.

With

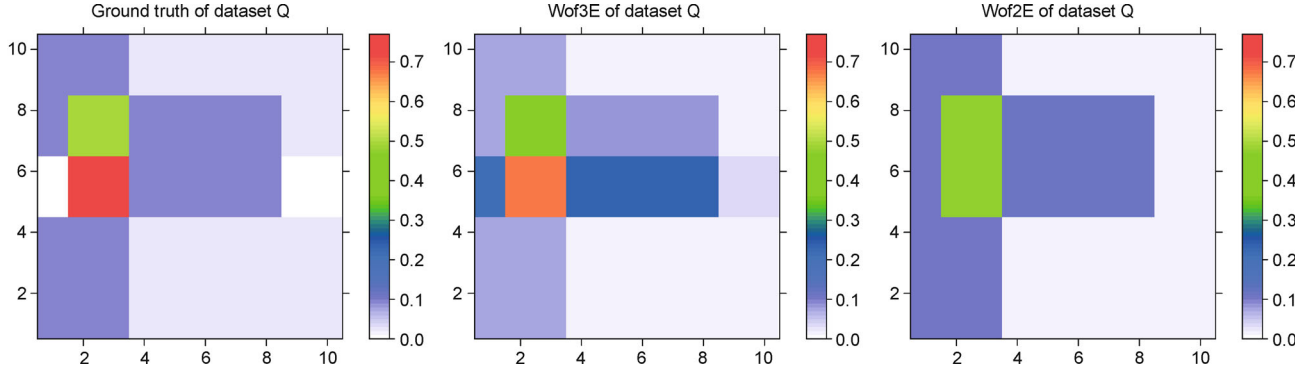

Fig. 3 Ground truth (left) and numerical results of weights-of-evidence with three predictors B_i, $i = 1,2,3$, despite violation of joint conditional independence (center), and two predictors B_i, $i = 1,2$ (right) applied to training dataset Q.

Table 9 Numerical results of Boost123WofE $(\mathsf{B}_1,\mathsf{B}_2,\mathsf{B}_3)$ formally applied to q taken from (Cheng, 2015, Eqs. (41), (45), (48), pp. 609–611). Boosted weights and contrast with respect to B_1 agree with the conventional ones

	B_1	B_2	B_3	$\sum_\ell$
$^{\mathrm{boost123}}W_\ell^{(1)}$	**0.008**	1.241	0.018	
$^{\mathrm{boost123}}W_\ell^{(0)}$	**–0.909**	–0.951	–0.016	–1.876
$^{\mathrm{boost123}}C_\ell$	**0.917**	2.192	0.034	

1) The figures 0.018 and –0.016 for the boosted weights of the third predictor B_3 given with Eq. (48) by Cheng (2015, p. 611) are likely to be typos. However their difference 0.034 is about right such that the computed predicted probabilities are not much affected.

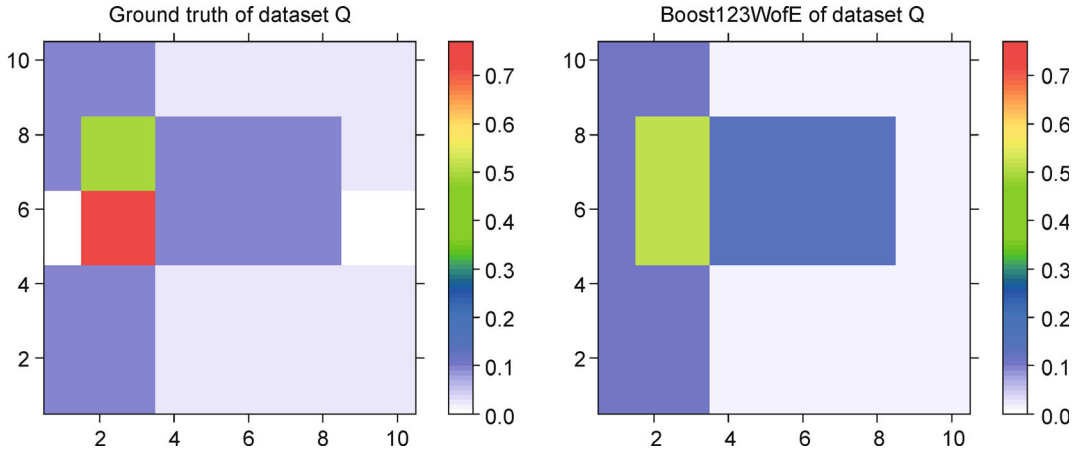

Fig. 4 Ground truth (left) and numerical results of Boost123WofE (right) formally applied to training dataset Q.

Table 10 Numerical results of Boost213WofE applied to training dataset Q. Boosted weights and contrast with respect to B_2 agree with the conventional ones

	B_2	B_1	B_3	$\sum_\ell$
${}^{\text{boost213}}W_\ell^{(1)}$	**0.098**	1.160	0.104	
${}^{\text{boost213}}W_\ell^{(0)}$	**–0.938**	–0.928	–0.036	–1.903
${}^{\text{boost213}}C_\ell$	**0.036**	2.088	0.140	

$$ {}^{\text{boost213}}W^{(0)} = -1.903,$$

$$\ln\Big(O(\mathsf{T}=1)\Big) + {}^{\text{boost213}}W^{(0)} = -4.100$$

the Boost**213**WofE model now reads explicitly

$$ {}^{\text{boost213}}\hat{P}_{\text{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3)$$

$$= \Lambda(-4.100 + 2.088\mathsf{B}_1 + 2.036\mathsf{B}_2 + 0.140\mathsf{B}_3), \quad (23)$$

and results in

$$\sum {}^{\text{boost213}}\hat{P}_{\text{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = 10.045,$$

$$\sum {}^{\text{boost213}}\hat{P}_{\text{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) - \sum t_i = 0.045. \quad (24)$$

For $(\mathsf{B}_3,\mathsf{B}_1,\mathsf{B}_2)$ Table 11 gives the corresponding figures.

With

$$ {}^{\text{boost312}}W^{(0)} = -2.247,$$

$$\ln\Big(O(\mathsf{T}=1)\Big) + {}^{\text{boost312}}W^{(0)} = -4.444,$$

the Boost**312**WofE model now reads explicitly

$$ {}^{\text{boost312}}\hat{P}_{\text{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3)$$

$$= \Lambda(-4.444 + 2.101\mathsf{B}_1 + 1.798\mathsf{B}_2 + 1.125\mathsf{B}_3), \quad (25)$$

and results in

$$\sum {}^{\text{boost312}}\hat{P}_{\text{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = 9.377$$

$$\sum {}^{\text{boost312}}\hat{P}_{\text{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) - \sum t_i = -0.623. \quad (26)$$

Table 11 Numerical results of Boost312WofE applied to training dataset q. Boosted weights and contrast with respect to B_3 agree with the conventional ones

	B_3	B_1	B_2	$\sum_\ell$
${}^{\text{boost312}}W_\ell^{(1)}$	**0.810**	1.129	0.838	
${}^{\text{boost312}}W_\ell^{(0)}$	**–0.315**	–0.972	–0.960	–2.247
${}^{\text{boost312}}C_\ell$	**0.125**	2.101	1.798	

The error Eq. (24) of the Boost**213**WofE model Eq. (23) is small and of the opposite sign as the error Eq. (22) of the Boost**123**WofE model Eq. (21), while the error Eq. (26) of the Boost**312**WofE model Eq. (25) is about the same as of the two-terms weights-of-evidence model, Eq. (20).

Visual inspection of the map images of Fig. 5 clearly indicates that the BoostWofE model (Cheng, 2015) depends on the choice of the first predictor variable to initiate the boost algorithm, and on the order of the subsequent predictors. Checking all permutations, cf. Fig. 5, the dataset designed by Cheng (2015) to "prove" properties of BoostWofE indicates that the relative order of predictors B_2 and B_3, which were inferred to be not conditionally independent given the target T, is particularly sensitive. When the predictor B_2 precedes the predictor B_3, the patterns of the predictors B_1 and B_2 are visible in the pattern of predicted conditional probability while the pattern of B_3 is not. Contrary, when the predictor B_3 precedes the predictor B_2, the pattern of the predictor B_3 becomes visible in the pattern of the predicted conditional probability and apparently contributes spatial resolution to the the pattern provided by B_2. Unfortunately, Cheng

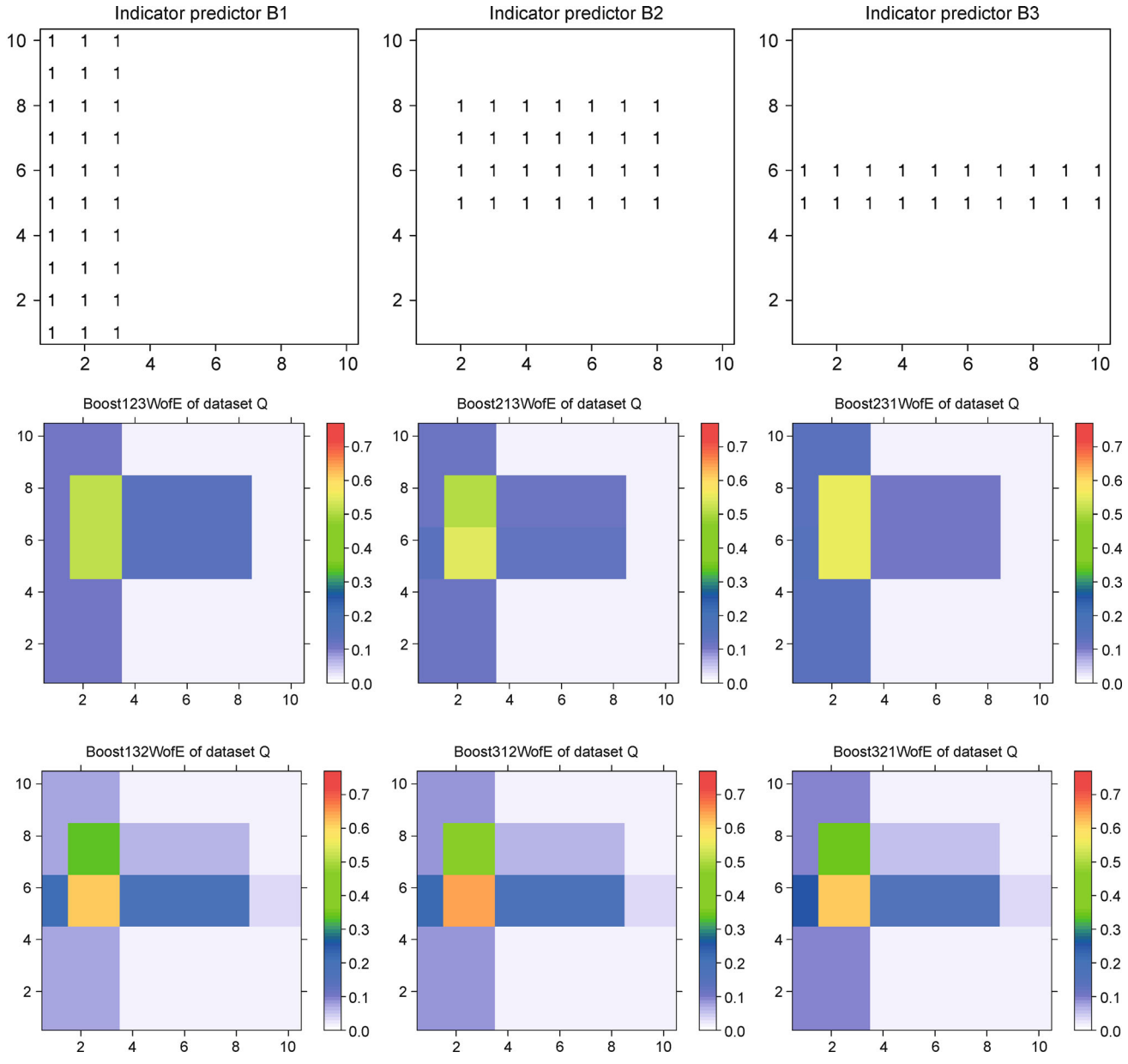

Fig. 5 Numerical result of BoostWofE as suggested by Cheng (2015) applied to training dataset Q comprising three predictors $B_i, i = 1,2,3$, (top) for all permutations of $\{1,2,3\}$. B_2 preceding B_3: Boost123WofE, Boost213WofE, and Boost231WofE (center); B_3 preceding B_2: Boost132WofE, Boost312WofE, and Boost321WofE (bottom).

(2015) does neither discuss how to choose the first predictor variable, nor the influence of the order of the subsequently one by one included predictors.

4.1.6 Application of logistic regression

Three logistic regression models are fitted to the dataset Q. The first uses B_1 and B_2 only, Table 12, the second adds B_3, Table 13, and the third adds the interaction term $\mathsf{B}_2 : \mathsf{B}_3$ to compensate for the lack of conditional independence, Table 14.

Table 12 Numerical results of two-terms logistic regression model applied to Q (AIC = 53.25)

	Estimate	Std. Error	z value	Pr ($>\|z\|$)
(Intercept)	–4.402	0.882	–4.990	0.000
B_1	2.299	0.821	2.800	0.005
B_2	2.407	0.820	2.930	0.003

Table 13 Numerical results of three-term logistic regression model applied to Q (AIC = 55.206)

	Estimate	Std. Error	z value	Pr ($>\|z\|$)
(Intercept)	–4.421	0.889	–4.970	0.000
B_1	2.301	0.822	2.800	0.005
B_2	2.330	0.896	2.600	0.009
B_3	0.187	0.890	0.210	0.833

Table 14 Numerical results of full logistic regression model applied to Q (AIC = 61.686)

	Estimate	Std. Error	z value	Pr ($>\|z\|$)
(Intercept)	–3.806	1.011	–3.770	0.000
B_1	1.609	1.256	1.280	0.200
B_2	1.609	1.460	1.100	0.270
B_3	–14.759	3261.319	–0.000	0.996
$\mathsf{B}_1 : \mathsf{B}_2$	0.587	1.920	0.310	0.759
$\mathsf{B}_1 : \mathsf{B}_3$	–1.609	5648.770	–0.000	0.999
$\mathsf{B}_2 : \mathsf{B}_3$	14.759	3261.319	0.000	0.996
$\mathsf{B}_1 : \mathsf{B}_2 : \mathsf{B}_3$	2.708	5648.771	0.000	0.999

For all logistic regression models $\sum \hat{P}_{\mathrm{lrM}}(\mathsf{T}{=}1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = \sum t_i$ as this is a constitutive equation.

Obviously, the full logistic regression model recovers the ground truth perfectly, cf. Fig. 6. However, its fitted parameters are not significant except the so-called intercept, and therefore it represents an instance of overfitting. Thus, for the training dataset Q the full logistic model is not at all appropriate for prediction.

Generally, for indicator predictors the general weights-of-evidence approach and the full logistic regression model will always recover the ground truth. As for the training dataset Q the full logistic regression model with $\sum_{\ell=1}^{m} \binom{m}{\ell} + 1 = 2^m = 8$ terms might be rendered parsimonious compared to the general weights-of-evidence model with $\sum_{\ell=1}^{m} 2^\ell + 1 = 2^{m+1} - 1 = 15$ terms. Moreover, only logistic regression provides statistical significance of its fitted parameters to judge whether a model is appropriate for prediction or not.

The only significant logistic regression model with predictors B_1 and B_2 results in a simplified pattern quite similar to the pattern of Wof2E, cf. Fig. 3, and Boost**123**WofE, cf. Fig. 4. However, the notion of significance does neither exist in WofE nor in BoostWofE by Cheng (2015).

If there was any doubt in the inappropriateness of the full logistic regression model for prospectivity modeling, the large standard errors as depicted in Fig. 7 confirm that the predicted conditional probabilities cannot be considered reasonably reliable.

4.1.7 Summary of training dataset Q

All fitted models are explicitly compiled in Eqs. (27) to (34).

$$\hat{P}_{\mathrm{Wof3E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = \Lambda(-4.359 + 1.917\mathsf{B}_1 + 2.037\mathsf{B}_2 + 1.126\mathsf{B}_3), \quad (27)$$

$$\hat{P}_{\mathrm{Wof2E}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = \Lambda(-4.044 + 1.917\mathsf{B}_1 + 2.037\mathsf{B}_2), \quad (28)$$

$${}^{\mathrm{boost123}}\hat{P}_{\mathrm{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = \Lambda(-4.073 + 1.917\mathsf{B}_1 + 2.192\mathsf{B}_2 + 0.034\mathsf{B}_3), \quad (29)$$

$${}^{\mathrm{boost213}}\hat{P}_{\mathrm{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = \Lambda(-4.100 + 2.088\mathsf{B}_1 + 2.036\mathsf{B}_2 + 0.140\mathsf{B}_3), \quad (30)$$

$${}^{\mathrm{boost312}}\hat{P}_{\mathrm{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = \Lambda(-4.444 + 2.101\mathsf{B}_1 + 1.798\mathsf{B}_2 + 1.125\mathsf{B}_3), \quad (31)$$

$$\hat{P}_{\mathrm{fit}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3) = \Lambda(-4.402 + 2.299\mathsf{B}_1 + 2.407\mathsf{B}_2), \quad (32)$$

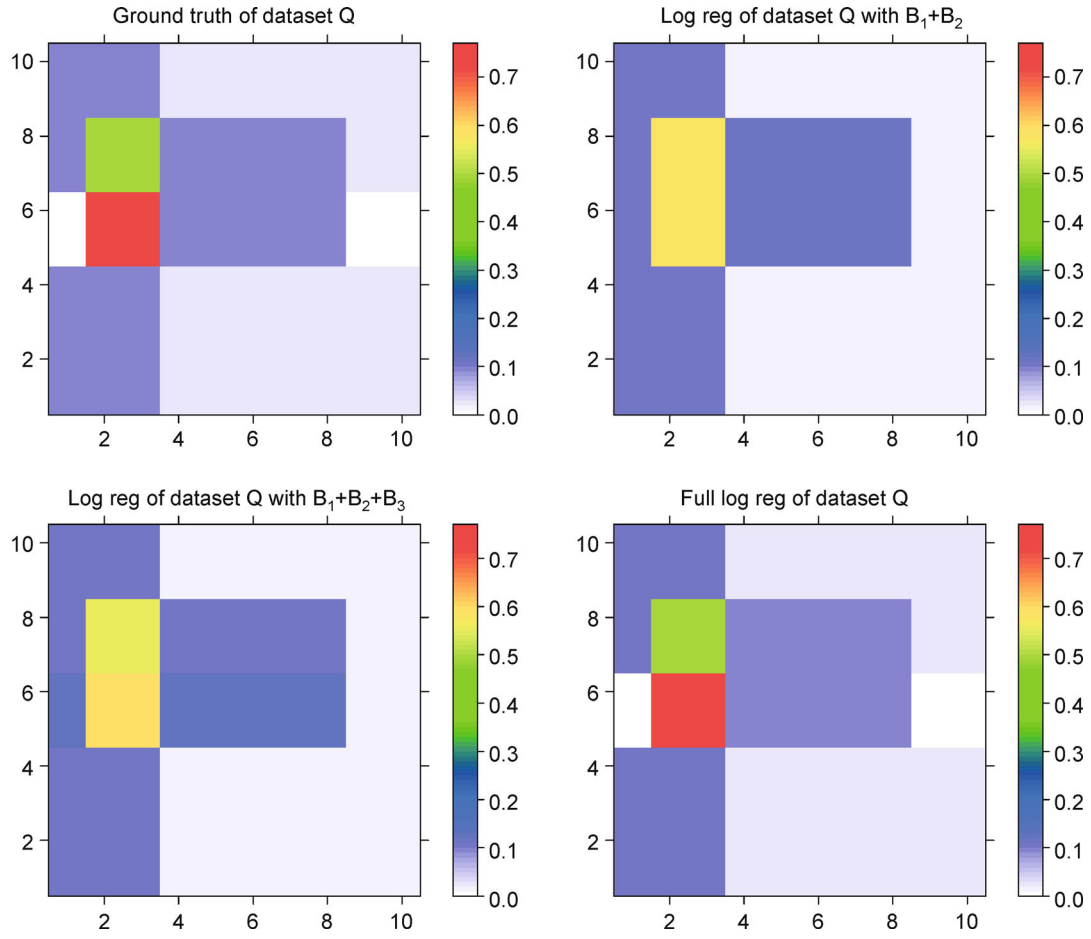

Fig. 6 Comparing the ground truth (top left) and numerical results of logistic regression models with two significant terms (top right), with three terms (bottom left), and with all interaction terms (bottom right) applied to training dataset Q.

$$\hat{P}_3(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3)$$

$$=\Lambda(-4.421+2.301\mathsf{B}_1+2.330\mathsf{B}_2+0.187\mathsf{B}_3), \quad (33)$$

$$\hat{P}_{\text{full}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3)$$

$$=\Lambda(-3.806+1.609\mathsf{B}_1+1.609\mathsf{B}_2-14.759\mathsf{B}_3$$

$$+0.587\mathsf{B}_1\mathsf{B}_2-1.609\mathsf{B}_1\mathsf{B}_3+14.759\mathsf{B}_2\mathsf{B}_3$$

$$+2.708\mathsf{B}_1\mathsf{B}_2\mathsf{B}_3). \quad (34)$$

Conditional probabilities $\hat{P}(\mathsf{T}=1|\mathbf{B})$ predicted with several fitted models are compiled in Table 15.

4.2 Fabricated training dataset RANKIT

The fabricated training dataset RANKIT, Fig. 8, was already used earlier to demonstrate the effects of a serious violation of the mathematical modeling assumption of conditional independence (Schaeben, 2014a, b). Here, applying boost weights-of-evidence as suggested by Cheng (2015) results in the fitted models explicitly given by

$$^{\text{boost12}}P_{\text{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2)$$

$$=\Lambda(-3.074+1.725\mathsf{B}_1+1.157\mathsf{B}_2), \quad (35)$$

$$^{\text{boost21}}P_{\text{WofE}}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2)$$

$$=\Lambda(-2.997+1.372\mathsf{B}_1+1.349\mathsf{B}_2). \quad (36)$$

First, Eqs. (35) and (36) confirm again that the fitted models depend on the order of processing the two predictors. However, here they are quite similar, but both of them do not reduce the effect of lacking conditional independence. In fact, Figure 9 clearly reveals that whatever the order boosting does not generally provide a

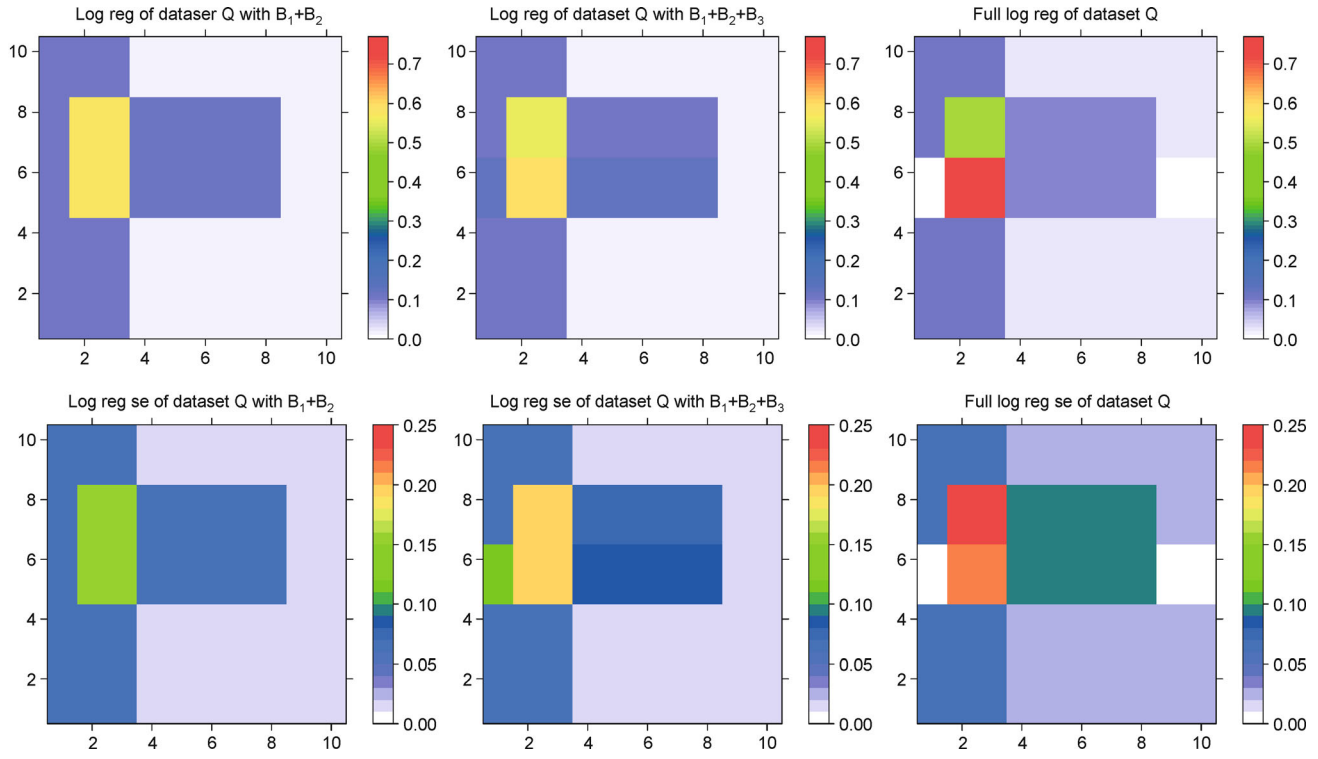

Fig. 7 Estimated conditional probabilities (top row) and estimation errors (bottom row) of significant two terms logistic regression model (left), three terms logistic regression model (center), and full logistic regression model with all interaction terms (right) applied to training dataset Q.

Table 15 Comparison of predicted conditional probabilities for various methods comprising the ground truth given in terms of conditional frequencies by counting (first column), numerical results of weights-of-evidence using all three predictors (second column), or the two predictors B_1 and B_2 only (third column), numerical results of the best significant logistic regression model using the same two predictors B_1 and B_2 (fourth column), original figures of weights-of-evidence from (Cheng, 2015) (sixth column), numerical results with Boost123WofE (seventh column), Boost213WofE (eighth column), (Boost312WofE (ninth column)

Predictors			$\hat{P}(T=1\|\mathbf{B})$								
B_1	B_2	B_3	Counting	Wof**3**E	Wof**2**E	2term lrM significant	WofE (Cheng, 2015)	BoostWoE (Cheng, 2015)	Boost**123** WofE	Boost**213** WofE	Boost**312** WofE
1	1	1	0.750	0.672	0.477	0.575	0.69	0.52	0.519	0.541	0.641
1	1	0	0.500	0.399	0.477	0.575	0.35	0.51	0.510	0.506	0.367
0	1	1	**0.100**	**0.232**	0.118	**0.119**	0.20	**0.14**	0.137	0.127	0.179
1	0	0	0.100	0.079	0.106	0.108	0.12	0.10	0.104	0.118	0.087
0	1	0	**0.100**	**0.089**	0.118	**0.119**	0.07	**0.13**	0.132	0.112	0.066
0	0	0	0.021	0.012	0.017	0.012	0.02	0.02	0.016	0.016	0.011
1	0	1	0.000	0.211	0.106	0.108	0.30	0.11	0.107	0.133	0.228
0	0	1	0.000	0.037	0.017	0.012	0.06	0.02	0.017	0.018	0.034

means to compensate the lack of conditional independence but yields similar patterns of predicted conditional probabilities as conventional weights-of-evidence or logistic regression without the interaction term $B_1 : B_2$. Including the interaction term results in the full logistic regression model which is significant (Schaeben, 2014b) and recovers the ground truth given in terms of conditional frequencies by counting.

5 Discussion of results

General weights of evidence are not "very difficult if not impossible" to estimate "given a limited number of training data" (Cheng, 2015, p. 597). In fact, if conditional probabilities being 0 or 1 can reasonably be excluded, they are estimated by counting frequencies of occurrences

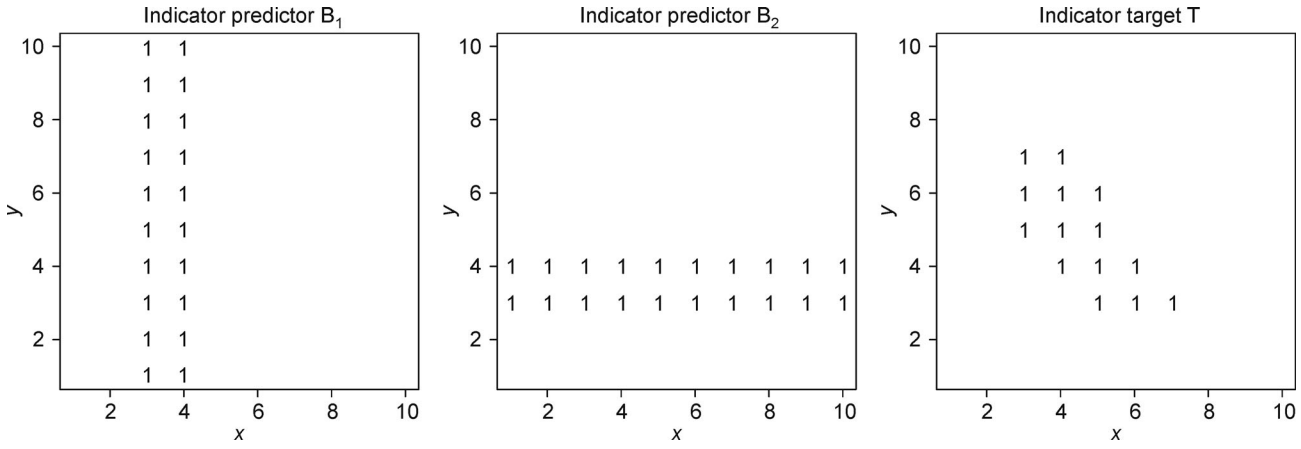

Fig. 8 Spatial distribution of two indicator predictor variables B_1,B_2 and the indicator target variable T of the training dataset rankit.

of several events like conventional weights. Since the target event $\mathsf{T}=1$ is usually a very rare event in prospectivity modeling, chances are fairly small that the denominators $P\Big((\mathsf{Z}_\ell,\mathsf{T})=(\mathbf{b}_{j,\ell},0)\Big)$ involved in the Bayes factors, Eq. (8), referring to $\mathsf{T}=0$ vanish. When applying general weights-of-evidence to the limited number of training data as provided by the dataset Q (Cheng, 2015) problems were not encountered. Thus, difficulties to be resolved by boost weights of evidence do not seem to exist.

Given the ten pages it takes (Cheng, 2015, pp. 596–607) to derive the procedure of boost weights-of-evidence relying on an ad hoc approximation which is not justified in any way, it cannot be confirmed that BoostWofE is simple and intuitively appealing as compared with logistic regression, Eq. (5), for instance. Like conventional weights-of-evidence and opposed to logistic regression, boost weights-of-evidence lacks the notion of significance of fitted model parameters.

As for the fabricated training data set Q used by Cheng (2015), the statistical test of the null–hypothesis of joint conditional independence of all three predictors B_ℓ, $\ell=1,2,3$, given the target T leads to infer to reasonably reject it. However, it can be rejected because B_2 and B_3 are not conditional independent given T, while the null hypothesis cannot reasonably be rejected for the other pairs of predictors.

Then, the fabricated training data set Q obviously exemplifies that the results of boost weights-of-evidence largely depend on the boosting sequence of predictors. It has to be noted that criteria for the user decision how to choose the sequence are not provided in (Cheng, 2015). Whatever the boosting sequence, boost weights-of-evidence does not reduce the effect of violated conditional independence but leads either to corrupted patterns of predicted conditional probabilities as conventional weights-of-evidence does, or to simplified patterns similar to those accomplished by omitting B_3 from the prediction.

As with respect to the training dataset RANKIT which is significantly lacking joint conditional independence, application of boost weights-of-evidence results in the same corrupted pattern of predicted conditional probabilities as conventional weights-of-evidence.

In fact, boost weights-of-evidence method as introduced by Cheng (2015) does not generally reduce the effect of lacking conditional independence, not to mention the simple, intuitive and more generic way as claimed by Cheng (2015, p. 620). Non of the claims by Cheng (2015) could be verified.

It is commonplace that examples cannot substitute mathematical proofs, while one counterexample is sufficient to disprove a mathematical statement. In the same way, it is not possible to derive properties of an ad hoc procedure nor to validate it by way of an example with fabricated or observed training data. It is impossible to judge the performance of an ad hoc procedure by way of its application to observed training data. It takes mathematical-statistical analysis to derive a method, to turn it into an algorithm, and to encode it in software.

6 Conclusions

Boost weights-of-evidence (Cheng, 2015) is another improper attempt to relax the mathematical modeling assumption of joint conditional independence of all predictors given the target hampering reliable predictions of prospectivity. Boost weights-of-evidence does not generally reduce not to mention *significantly* reduce the effect of lacking joint conditional independence. Its application yields corrupted predicted conditional probabilities and corrupted spatial patterns of prospectivity as weights-of-evidence does. In particular, its results depend on the sequential processing order of predictors. General weights of evidence do not require the modeling assumption of joint conditional independence of all predictors given the target, and can be estimated by mere counting like conventional weights. Logistic regression provides a

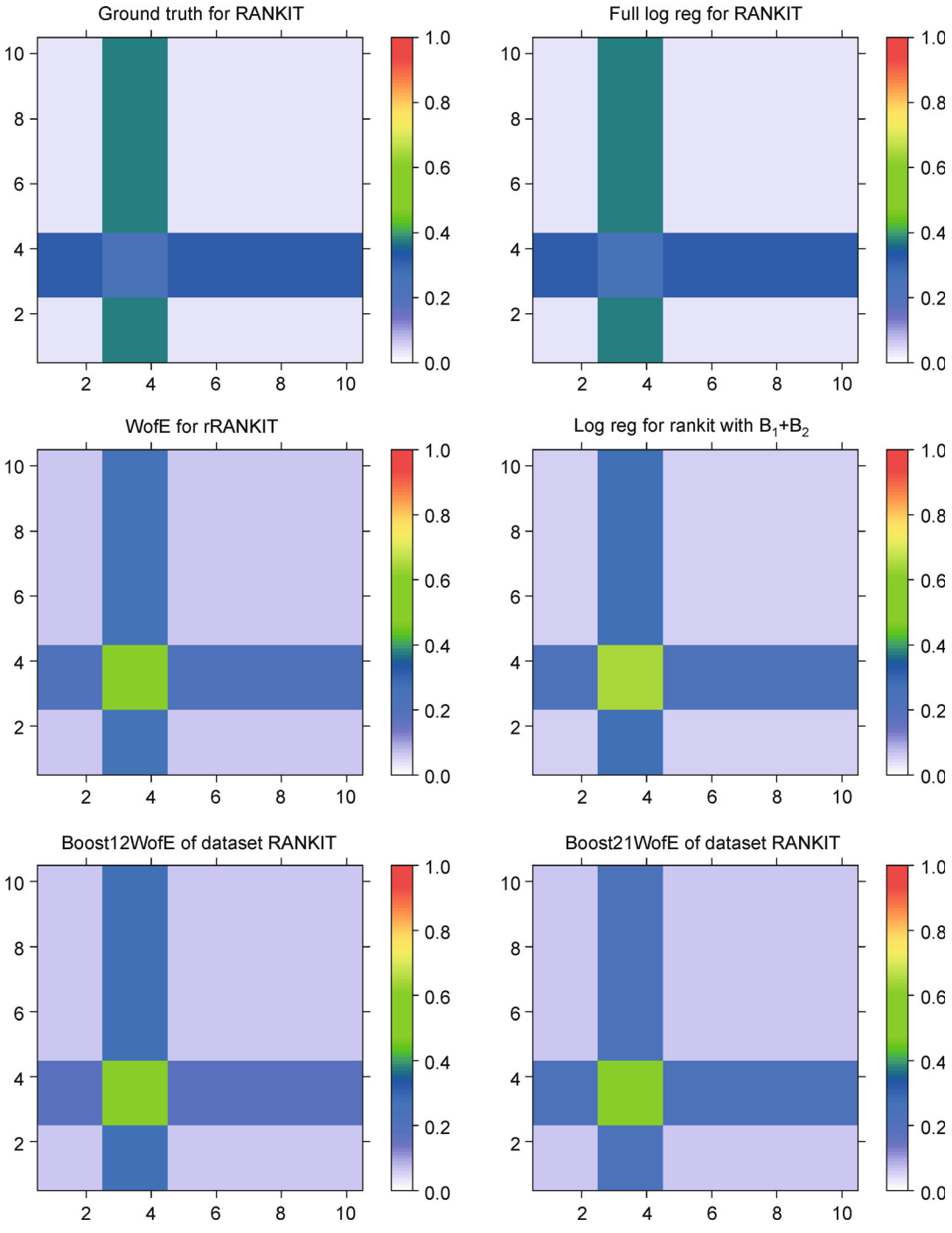

Fig. 9 Spatial distribution of predicted conditional probabilities $\hat{P}(\mathsf{T}=1|\mathsf{B}_1\mathsf{B}_2)$ for the training dataset RANKIT according to: elementary estimation by counting referred to as ground truth (top left); logistic regression with interaction term (top right); weights-of-evidence (middle left); logistic regression without interaction (middle right); boost12 weights-of-evidence (bottom left); boost21 weights-of-evidence (bottom right).

relative measure of fit, and can distinguish significantly fitted models appropriate for predictions.

Acknowledgements The authors would like to thank two anonymous reviewers for their thorough and constructive efforts to help us improve our manuscript. The authors gratefully acknowledge financial funding by the German Federal Ministry for Economic Affairs and Energy (BMWi) within the frame of "Zentrales Innovationsprogramm Mittelstand" (ZIM) on *Entwicklung eines Verfahrens zur dreidimensionalen Prognose von verdeckten Rohstofflagerstätten am Beispiel des Erzgebirges.* Last but not least the authors greatly appreciate H. Konstanze Zschoke's, MSc Geophysics, painstaking effort to convert our manuscript from the high-quality typesetting system LaTeX into the word processor MS Word.

Appendix A: the ν-approach

The ν-approach

$$F_\ell = \nu_\ell F_\ell^{\mathrm{CI}}, \tag{A1}$$

was introduced (Polyakova and Journel, 2007) as a heuristic alternative to conditional independence. It reads in greater detail

$$F_\ell^{(i)} = \nu_\ell^{(i)} \frac{P(\mathsf{B}_\ell = i|\mathsf{T}=1)}{P(\mathsf{B}_\ell = i|\mathsf{T}=0)},\ i = 0,1,$$

with

$$\nu_\ell^{(i)} = \frac{\dfrac{P(\mathsf{B}_\ell = i|\otimes_{j=0}^{\ell-1}\mathsf{B}_j \otimes \mathsf{T}=1)}{P(\mathsf{B}_\ell = i|\otimes_{j=0}^{\ell-1}\mathsf{B}_j \otimes \mathsf{T}=0)}}{\dfrac{P(\mathsf{B}_\ell = i|\mathsf{T}=1)}{P(\mathsf{B}_\ell = i|\mathsf{T}=0)}} = \frac{\dfrac{P(\mathsf{B}_\ell = i|\otimes_{j=0}^{\ell-1}\mathsf{B}_j \otimes \mathsf{T}=1)}{P(\mathsf{B}_\ell = i|\mathsf{T}=1)}}{\dfrac{P(\mathsf{B}_\ell = i|\otimes_{j=0}^{\ell-1}\mathsf{B}_j \otimes \mathsf{T}=0)}{P(\mathsf{B}_\ell = i|\mathsf{T}=0)}}, \tag{A2}$$

$$i = 0,1,\ \ell = 1,...,m,$$

which is actually not enlightening by itself. Formally, it leads to weights additively modified by $\alpha_\ell^{(i)} = \ln\ \nu_\ell^{(i)}$ and correspondingly modified contrasts

$$\tilde{W}_\ell^{(i)} = \alpha_\ell^{(i)} + W_\ell^{(i)},\ i = 0,1,\ \ell = 1,...,m,$$

$$\tilde{C}_\ell = \tilde{W}_\ell^{(1)} - \tilde{W}_\ell^{(0)} = \alpha_\ell^{(1)} - \alpha_\ell^{(0)} + C_\ell = \alpha_\ell + C_\ell,\ \ell = 1,...,m,$$

with $\alpha_\ell = \alpha_\ell^{(1)} - \alpha_\ell^{(0)}$, and with $\alpha^{(0)} = \sum_{\ell=1}^{m} \alpha_\ell^{(0)}$ eventually to

$$\begin{aligned}&\text{logit}\ P(\mathsf{T}=1|\mathbf{B})\\ &= \text{logit}\ P(\mathsf{T}=1) + \alpha^{(0)} + W^{(0)} + \sum_{\ell=1}^{m}(\alpha_\ell + C_\ell)\mathsf{B}_\ell.\end{aligned} \tag{A3}$$

The ν-approach is not equivalent to the τ-approach (Journel, 2002; Krishnan, 2008).

It was claimed to be "strictly" equivalent by its authors when they mistook logarithm for a linear function (Polyakova and Journel, 2007, p. 723), cf. Fig. A1.

Thus

$$\log\frac{x'}{x} = \tau_1 \log\frac{x_1'}{x_1} + \tau_2 \log\frac{x_2'}{x_2},$$

or equivalently

$$\frac{x'}{x} = \left(\frac{x_1'}{x_1}\right)^{\tau_1} + \left(\frac{x_2'}{x_2}\right)^{\tau_2}$$

Fig. A1 The logarithm function is mistaken for a linear function (Polyakova and Journel, 2007, Mathematical Geology 39, p. 723).

Despite that reasonable ways to estimate ν_ℓ do not seem to be known, there is obviously no way to emulate the effect of interaction terms included in logistic regression models by subsequently correcting the weights.

Appendix B: boost weights-of-evidence

To get the notation of Cheng (2015) straight, we introduce the random vector $\mathbf{Z}_\ell = (\mathsf{B}_1,...,\mathsf{B}_\ell)^{\mathsf{T}}$, $\ell = 1,...,m$, comprising the first ℓ predictors. Since all predictors are indicators, the vector $\mathbf{Z}_\ell$ possess 2^ℓ different realizations, each of which is an element of the set V_ℓ of ℓ-variations of the set $\{0,1\}$. The total number of different ℓ-variations is $v_\ell = 2^\ell$.

The boosted weights of evidence are defined with Eq. (13) of Cheng (2015, p. 598) as

$${}^{\text{boost}}W_1 = W_1,$$

$${}^{\text{boost}}W_\ell = \ln\frac{\sum_{j=1}^{v_{\ell-1}} \lambda_{j,\ell} P(\mathsf{B}_\ell|\mathbf{Z}_{\ell-1} = \mathbf{z}_j \wedge \mathsf{T}=1)}{\sum_{j=1}^{v_{\ell-1}} \lambda_{j,\ell} P(\mathsf{B}_\ell|\mathbf{Z}_{\ell-1} = \mathbf{z}_j \wedge \mathsf{T}=0)},$$

$$\ell = 2,...,m,$$

with $\mathbf{z}_j \in V_{\ell-1}$, $j = 1,...,v_{\ell-1}$ and coefficients $\lambda_{j,\ell}$ to be defined and determined. Boosted weights ${}^{\text{boost}}W_\ell$ refer to the predictor B_ℓ only, while the proper Bayes factors F_ℓ, Eq. (8), refer to $\mathsf{B}_1,...,\mathsf{B}_\ell$, i.e., to $\mathbf{Z}_\ell$. However, the approximation of the Bayes factors F_ℓ,

$$\frac{P\left(\mathsf{B}_\ell|\mathbf{Z}_{\ell-1}=(...)\wedge\mathsf{T}=1\right)}{P\left(\mathsf{B}_\ell|\mathbf{Z}_{\ell-1}=(...)\wedge\mathsf{T}=0\right)}$$

$$\sim\frac{\sum_{j=1}^{v_{\ell-1}}\lambda_{j,\ell}P(\mathsf{B}_\ell|\mathbf{Z}_{\ell-1}=\mathbf{z}_j\wedge\mathsf{T}=1)}{\sum_{j=1}^{v_{\ell-1}}\lambda_{j,\ell}P(\mathsf{B}_\ell|\mathbf{Z}_{\ell-1}=\mathbf{z}_j\wedge\mathsf{T}=0)}, \qquad \text{(B1)}$$

for $\ell=2,...,m$, is at best an ad hoc approximation as a mathematical justification in general is missing. The coefficients $\lambda_{j,\ell}$ are successively defined in a procedural way that takes 10 pages to elaborate on Cheng (2015, pp. 596–607).

References

Agterberg F P (2014). Geomathematics: Theoretical Foundations, Applications and Future Developments. Cham, Heidelberg, New York, Dordrecht, London: Springer

Agterberg F P, Bonham-Carter G F, Wright D F (1990). Statistical pattern integration for mineral exploration. In: Gaál G, Merriam D F, eds. Computer Applications in Resource Estimation Prediction and Assessment for Metals and Petroleum. Oxford, New York: Pergamon Press, 1–21

Agterberg F P, Cheng Q (2002). Conditional independence test for weights-of-evidence modeling. Nat Resour Res, 11(4): 249–255

Berkson J (1944). Application of the logistic function to bio-assay. J Am Stat Assoc, 39(227): 357–365

Bonham-Carter G (1994). Geographic Information Systems for Geoscientists: Modeling with GIS. New York: Pergamon, Elsevier Science

Butz C J, Sanscartier M J (2002). Properties of weak conditional independence. In: Alpigini J J, Peters J F, Skowron A, Zhong N, eds. Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science (Volume 2475). Berlin, Heidelberg: Springer, 349–356www2.cs.uregina.ca/ butz/publications/properties.ps.gz

Chalak K, White H (2012). Causality, conditional independence, and graphical separation in settable systems. Neural Comput, 24(7): 1611–1668

Cheng Q (2012). Application of a newly developed boost weights of evidence model (BoostWofE) for mineral resources quantitative assessments. Journal of Jilin University, Earth Sci Ed, 42(6): 1976–1985

Cheng Q (2015). BoostWofE: a new sequential weights of evidence model reducing the effect of conditional dependency. Math Geosci, 47(5): 591–621

Chilès J P, Delfiner P (2012). Geostatistics- Modeling Spatial Uncertainty (2nd ed). New York, Chichester, Weinheim, Brisbane, Singapore, Toronto: John Wiley & Sons

Dawid A P (1979). Conditional independence in statistical theory. J R Stat Soc, B, 41(1): 1–31

Dawid A P (2004). Probability, causality and the empirical world: a Bayes-de Finetti-Popper-Borel synthesis. Stat Sci, 19(1): 44–57

Dawid A P (2007). Fundamentals of Statistical Causality. Research Report 279, Department of Statistical Science, University College London ESRI, ArcGIS. http://www.esri.com/software/arcgis

Ford A, Miller J M, Mol A G (2016). A comparative analysis of weights of evidence, evidential belief functions, and fuzzy logic for mineral potential mapping using incomplete data at the scale of investigation. Nat Resour Res, 25(1): 19–33

Freund Y, Schapire R E (1997). A decision theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci, 55 (1): 119–139

Freund Y, Schapire R E (1999). A short introduction to boosting. Jinko Chino Gakkaishi, 14(5): 771–780

Friedman J, Hastie T, Tibshirani R (2000). Additive logistic regression: a statistical view of boosting. Ann Stat, 28(2): 337–407

Good I J (1950). Probability and the Weighing of Evidence. London: Griffin

Good I J (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. J R Stat Soc, B, 22(2): 319–331

Good I J (1968). The Estimation of Probabilities: An Essay on Modern Bayesian Methods. MIT Research Monograph No. 30, The MIT Press, Cambridge, MA, 109

Harris D P, Pan G C (1999). Mineral favorability mapping: a comparison of artificial neural networks, logistic regression and discriminant analysis. Nat Resour Res, 8(2): 93–109

Harris D P, Zurcher L, Stanley M, Marlow J, Pan G (2003). A comparative analysis of favorability mappings by weights of evidence, probabilistic neural networks, discriminant analysis, and logistic regression. Nat Resour Res, 12(4): 241–255

Hastie T, Tibshirani R, Friedman J (2009). The Elements of Statistical Learning (2nd ed). New York: Springer

Hosmer D W, Lemeshow S, Sturdivant R X (2013). Applied Logistic Regression (3rd ed). Hoboken, NJ: Wiley & Sons

Journel A G (2002). Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses. Math Geol, 34(5): 573–596

Kreuzer O, Porwal A, eds. (2010). Special Issue "Mineral Prospectivity Analysis and Quantitative Resource Estimation". Ore Geol Rev, 38 (3): 121–304

Krishnan S (2008). The τ-model for data redundancy and information combination in Earth sciences: theory and application. Math Geol, 40(6): 705–727

Minsky M, Selfridge O G (1961). Learning in random nets. In: Cherry C, ed. 4th London Symposium on Information Theory. London: Butterworths, 335–347

Pearl J (2009). Causality: Models, Reasoning, and Inference. 2nd ed. New York: Cambridge University Press

Polyakova E I, Journel A G (2007). The ν. Math Geol, 39(8): 715–733

Porwal A, Carranza E J M (2015). Introduction to the Special Issue: GIS-based mineral potential modelling and geological data analyses for mineral exploration. Ore Geol Rev, 71: 477–483

Porwal A, González-Álvarez I, Markwitz V, McCuaig T C, Mamuse A (2010). Weights of evidence and logistic regression modeling of magmatic nickel sulfide prospectivity in the Yilgarn Craton, Western Australia. Ore Geol Rev, 38(3): 184–196

Reed L J, Berkson J (1929). The application of the logistic function to experimental data. J Phys Chem, 33(5): 760–779

Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015). Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol Rev, 71: 804–818

Schaeben H (2014a). Targeting: logistic regression, special cases and extensions. ISPRS Int J Geoinf, 3(4): 1387–1411. Available at: http://www.mdpi.com/2220-9964/3/4/1387

Schaeben H (2014b). Potential modeling: conditional independence matters. GEM-International Journal on Geomathematics, 5(1): 99–116

Schaeben H (2014c). A mathematical view of weights-of-evidence, conditional independence, and logistic regression in terms of Markov random fields. Math Geosci, 46(6): 691–709

Šochman J, Matas J (2004). Adaboost with totally corrective updates for fast face detection. In: Proc. 6th IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, South Korea, 445–450

Suppes P (1970). A Probabilistic Theory of Causality. Amsterdam: North-Holland

Tolosana-Delgado R, van den Boogaart K G, Schaeben H (2014). Potential mapping from geochemical surveys using a Cox process. 10th Conference on Geostatistics for Environmental Applications, Paris, July 9–11, 2014

van den Boogaart K G, Schaeben H (2012). Mineral potential mapping using Cox–type regression for marked point processes. 34th IGC Brisbane, Australia

Wong M S K M , Butz C J (1999). Contextual weak independence in Bayesian networks. In: Proc. 15th Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 670–679