A review of probability aggregation methods in Earth sciences

Alessandro Comunian, Denis Allard and Philippe Renard

Abstract The need of combining in a probabilistic framework different sources of information is a frequent task in earth sciences. This can occur for example when modeling a reservoir using direct geological observations, geophysics, remote sensing, training images etc. For example, the probability of occurrence of a certain lithofacies at a certain location can easily be computed conditionally on the event observed at each source of information. The problem of aggregating these different conditional probability distributions into a single conditional distribution arises as an approximation to the inaccessible genuine conditional probability given all information. This paper makes a formal review of most aggregation methods proposed in the literature with a particular focus on their mathematical properties. Calibration of the aggregated probability distribution is of particular importance. It is known that linear aggregation operators are not calibrated. Here, we show that if a calibrated log-linear pooling exists, then it is the log-linear pooling with parameters estimated from maximum likelihood. Simulations in a spatial context illustrate the performance of these operators.

Alessandro Comunian
NCGRT, University of New South Wales, Sydney 2052 NSW, Australia
e-mail: a.comunian@unsw.edu.au

Denis Allard
INRA, UR546 BioSP, Site Agroparc 84914 Avignon, France
e-mail: allard@avignon.inra.fr

Philippe Renard
CHYN, University of Neuchâtel, Emile-Argand 11, 2000 Neuchâtel, Switzerland
e-mail: philippe.renard@unine.ch

Ninth International Geostatistics Congress, Oslo, Norway, June 11. – 15., 2012

1 Introduction

This extended abstract summarizes the work of Allard *et al.* (2012).We want to assess the probability of a given event A, conditional on the occurrence of a set of data events, D_i, $i = 1,\dots,n$, on the basis of the simultaneous knowledge of the corresponding conditional probabilities $P(A \mid D_i)$. Here we consider only the case where there is a finite number of possible outcomes for A. Let $\mathscr{A}$ be the finite set of events in Ω such that the events $A_1,\dots,A_K$ of $\mathscr{A}$ form a finite partition of Ω. The joint probability $P(A,D_1,\dots,D_n)$ is unknown and not accessible. We will approximate the full conditional probability $P(A|D_1,\dots,D_n)$ by means of a pooling operator P_G:

$$P_G(P(A|D_0),\dots,P(A|D_n)) \approx P(A|D_0,\dots,D_n), \qquad A \in \mathscr{A}. \tag{1}$$

Here $P(A|D_0) = P_0(A)$ is an *a priori* probability independent on all $P(A \mid D_i)$. It can be thought of as arising from an abstract and never specified information D_0. In the following, we will sometimes use the notation P_i in place of $P(A|D_i)$ and $P_G(A)$ in place of $P_G(P_0,P_1,\dots,P_n)(A)$.

2 Some mathematical properties

Definition 1 (Unanimity/Convexity). Given $P_i = p$ for $i = 1,\dots,n$, a pooling operator P_G preserves *unanimity* when $P_G = p$. Moreover, P_G is *convex* when it always verifies:

$$P_G \in [min\{P_1,\dots,P_n\}, max\{P_1,\dots,P_n\}] \tag{2}$$

Convexity is a sufficient condition for unanimity, but these two properties are not necessarily desirable. Indeed, consider for example two information $D_1 \neq D_2$ and an event $A \subset (D_1 \cap D_2)$. Then, $P(A \mid D_1) = P(A)/P(D_1)$, and $P(A \mid D_1 \cap D_2) = P(A)/P(D_1 \cap D_2)$. Now, $(D_1 \cap D_2) \subset D_1$ implies that $P(D_1 \cap D_2) < P(D_1)$. Hence $P(A \mid D_1 \cap D_2) > P(A \mid D_1)$. Therefore, the full conditional probability of A is larger than any partial conditional probability. A convex pooling operator cannot account for such kind of situations.

Definition 2 (External Bayesianity). Given a likelihood function $L(\cdot)$ on the events in $\mathscr{A}$, and let $P_i^L(A) = P_G^L(P_1,\dots,P_n)(A)$ define the Bayesian updating of P_i by L. Then, an aggregation operator is said to be *external Bayesian* if the operation of updating the probabilities with the likelihood L commutes with the aggregation operator, i.e. if

$$P_G(P_1^L,\dots,P_n^L)(A) = P_G^L(P_1,\dots,P_n)(A). \tag{3}$$

This interesting property is equivalent to the *weak likelihood ratio* [Bordley, 1982].

Definition 3 (0/1 forcing property). Let us suppose that there exists a source of information i such that $P(A|D_i) = 0$ and $P(A|D_j) \neq 1$ for $j \neq i$. In this case, an

aggregation operator that returns $P_G(A) = 0$ is said to enforce a certainty effect, a property also called the *0/1 forcing property* [Allard et al., 2011].

3 Pooling operators

Among the criteria that can be used for classifying pooling operators, the dichotomy between methods that combine probabilities using the addition and methods that use the multiplication is probably the most enlightening.

3.1 Linear pooling and Beta-transformed linear pooling

An intuitive way of aggregating the probabilities $P_1,\ldots,P_n$ is the *linear pooling*:

$$P_G(A) = \sum_{i=1}^{n} w_i P_i(A), \tag{4}$$

where the w_i are positive weights verifying $\sum_{i=1}^{n} w_i = 1$. This pooling operator is convex, but neither 0/1 forcing nor external Bayesianity is verified. The probability P_G is often multi-modal, since Eq. (4) corresponds to a mixture model or, in Boolean terms, to the operator "or". In a geosecience context, a pooling paradigm guided by the "and" logic appears more suitable for aggregating different information about the same object.

Ranjan and Gneiting (2010) proved that linear pooling is intrinsically suboptimal according to criteria presented in the next sections. In order to overcome this limitation they proposed the Beta-transformed Linear Pooling (BLP):

$$P_G(A) = H_{\alpha,\beta}\left(\sum_{i=1}^{n} w_i P_i(A)\right), \tag{5}$$

where $\sum_{i=1}^{n} w_i = 1$ and $H_{\alpha,\beta}$ is the cumulative density function of a Beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$. This pooling operator does not satisfy any of the properties listed in Section 2, unless in the case $\alpha = \beta = 1$, that is when it corresponds to a linear pooling. Ranjan and Gneiting (2010) showed on simulations and on real case studies that the Beta-transformed linear pool outperforms any linear pooling and that it presents very good performances.

3.2 Log-linear pooling

Genest and Zidek (1986) proved that any pooling operators depending explicitly on the events in $\mathscr{A}$, and verifying external Bayesianity must be of the form:

$$P_G(A) \propto H(A) P_0(A)^{1-\sum_{i=1}^n w_i} \prod_{i=1}^n P_i(A)^{w_i}, \tag{6}$$

with $H(A)$ being an arbitrary bounded function playing the role of a likelihood on the elements of $\mathscr{A}$. Probability distributions obtained with log-linear operators are in general unimodal and less dispersed than those obtained with linear operators. Clearly, log-linear pooling operators verify the 0/1 forcing property since they are based on a multiplication.

A particular log-linear operator is obtained from a maximum entropy principle. The maximum entropy pooling operator P_G verifying $P_G(P_0)(A) = P_0(A)$ and $P_G(P_0, P_i)(A) = P(A|D_i)$ for $i = 1, \ldots, n$ is of the form (6) with $w_i = 1$ for $i = 1, \ldots, n$ and $H(A) = c$. It can be shown to be equivalent to the conditional independence of all events D_i, given A [Allard et al., 2012]. The sum $S_{\mathbf{w}} = \sum_{i=1}^n w_i$ plays an important role in Eq. (6). If $S_{\mathbf{w}} = 1$, the prior term is filtered out ($w_0 = 0$) and unanimity is preserved. Now, suppose that $P_1 = \cdots = P_n = p$. In this case, if $S_{\mathbf{w}} > 1$, the prior term has a negative weight, and P_G will be further away from P_0 than p. The opposite holds when $S_{\mathbf{w}} < 1$.

3.3 Tau model

Journel (2002) derived a formula for aggregating probabilities that has been later named the Tau model. Let us define odds $O(A)$, with $O(A) = P(A)/(1-P(A))$. The Tau model aggregates the odds according to

$$O_G(A) = O_0(A)^{w_0} \prod_{i=1}^n \left(\frac{O_i(A)}{O_0(A)} \right)^{w_i} = O_0(A)^{w_0 - \sum_{i=1}^n w_i} \prod_{i=1}^n O_i(A) \tag{7}$$

where the weights w_i can vary in $[0, \infty)$. This pooling operator wa also proposed in Bordley (1982) in which it is shown that it is the only pooling operator verifying the weak likelihood ratio. In the case of a binary outcome, it can be shown that this pooling operator is mathematically equivalent to a log-linear pooling. In the more general case this equivalence is lost. A complete formulation of the Tau model in the general case is thus

$$P_G(A) \propto O_G(A)/(1+O_G(A)), \text{ with } O_G(A) = O_0(A)^{1-\sum_{i=1}^n w_i} \prod_{i=1}^n O_i(A)^{w_i},\ A \in \mathscr{A}. \tag{8}$$

4 Scores and calibration

A scoring rule $S(P_G,A_k)$ measures the discrepancy between $P(\cdot|D_1,\ldots,D_n)$ and $P_G(\cdot)$ by assigning a numerical value, a score, based on P_G and on the event A_k that materializes. $S(P_G,P)$ will denote the expected value of $S(P_G,A_k)$ under the true probability distribution P: $S(P_G,P)=\sum_{A_k\in\mathscr{A}} S(P_G,A_k)P(A_k)$. We will only consider scoring rules such that $S(P,P)\geq S(Q,P)$ for each distribution Q, with $S(P,P)=S(Q,P)$ if and only if $Q=P$. A divergence function $d(Q,P)=S(P,P)-S(Q,P)$ can be associated to the given scoring rule S.

Definition 4 (quadratic and logarithmic scores). The quadratic score, S_Q, and the logarithmic score, S_L, are defined by

$$S_Q(P,A_k)=-\sum_{j=1}^{K}(\delta_{jk}-p_j)^2; \qquad S_L(P,A_k)=\ln p_k \tag{9}$$

where δ_{jk} is the Kronecker delta.

The corresponding divergence functions are the Euclidean distance $d_Q(Q,P)=\sum_{k=1}^{K}(p_k-q_k)^2$ and the Kullback-Leibler divergence $d_L(Q,P)=\sum_{k=1}^{K}q_k=\ln(p_k/q_k)$. When $Q=P$ we have $S_Q(P,P)=0$, while $S_L(P,P)$ is the entropy of the distribution P.

Definition 5 (Calibration). Let us introduce $\mathbf{Y}=(Y_1,\ldots,Y_K)$ the random vector corresponding to the outcome, in which $Y_k=1$ if the outcome is A_k and $Y_k=0$ otherwise. The operator P_G is said to be *calibrated* if

$$P(Y_k|P_G(A_k))=P_G(A_k) \tag{10}$$

for each A_k in $\mathscr{A}$.

Ranjan and Gneiting (2010) proved that linear pooling operators are not calibrated. However, they showed on simulated and real cases that a Beta transformed linear pool can be calibrated. The calibration can be evaluated computing the squared difference between the empirical values of $P_G(A_k)$ and $P(Y_k|P_G(A_k))$.

4.1 Maximum likelihood estimation and calibration of log-linear pooling

At the exception of the maximum entropy pooling which is parameter free, all methods presented above have some parameters that need either to be estimated or set by the user.

When training data are available it is possible to estimate the optimum weights according to the optimization of some criterion. We will present the likelihood approach for estimating the parameters for methods based on the multiplication of

probabilities. A similar derivation for the linear opinion pool and its Beta transform can be found in Ranjan and Gneiting (2010). Maximum likelihood estimation is a special case of optimum score estimation, corresponding to maximizing the logarithmic score. Given a training data set of size M,

$$\{(\mathbf{Y}^{(m)}, P_0^{(m)}(A_k), \ldots, P_n^{(m)}(A_k))\}, \;\; m = 1, \ldots, M.$$

We will thus seek the vector of weights $\mathbf{w}= w_1, \ldots, w_K$ maximizing the score $\sum_{m=1}^{M} S(P_{G,\mathbf{w}}, A_k^m)$, which is equivalent to finding the maximum of the log-likelihood

$$L(\mathbf{w}) = \ln \prod_{m=1}^{M} \prod_{k=1}^{K} (P_{G,k}^{(m)})^{Y_k^{(m)}} = \sum_{m=1}^{M} \sum_{k=1}^{K} Y_k^{(m)} \ln P_{G,k}^{(m)}. \tag{11}$$

Concerning the calibration of log-linear pooling, Allard *et al.* (2012) proved the following result:

Theorem 1. *Suppose there exists a calibrated log-linear pooling. Then, asymptotically, it is the log-linear pooling with parameters estimated from maximum likelihood.*

5 Simulation example

Let us consider a Boolean model of spheres of radius $r = 0.07$, simulated in the unit cube $\mathscr{C}$. Let us denote $X(s), \;\; s \in \mathscr{C}$ its void indicator function and λ the mean number of spheres per unit volume. The void probability is $q = P(X(s) = 1) = \exp(-\lambda 4\pi r^3/3)$ [Lantuéjoul, 2002]. A prediction point s_0 is randomly located in the unit cube and information points s_i, $i = 1, \ldots, 4$ are randomly located around s_0. For this model, the conditional probabilities $P(X(s_0) = 1 | X(s_i) = 1)$ and $P(X(s_0) = 1 | X(s_i) = 0)$ for $i = 1, \ldots, 4$ can be computed explicitly.

A data set made of 50000 repetitions is built, and for this data set we computed the likelihood, the quadratic score S_Q and the calibration for the linear pool, the Beta-transformed linear pool, the maximum entropy (ME) and the log-linear pool (Table 1). For all methods except maximum entropy the weights were computed with the maximum likelihood approach described in the previous section. Remember that since the considered event is binary, the Tau model is equivalent to the log-linear pool. For comparison purpose, we also show the scores of the prior probability, $P_0 = P(X(s_0) = 1)$, and the (exact) conditional probability given only one data, $P_1 = P(X(s_0) = 1 \mid X(s_1))$.

On this example the linear pooling of the four data leads to scores only slightly better than considering only one data point. Beta-transformed linear pooling leads to a real improvement to linear pooling: the prediction is calibrated and the score S_Q is significantly improved.

Surprisingly good performances are obtained with Maximum Entropy, perhaps due to the Markovian nature of the Boolean model for which conditional expectation

Table 1 Likelihood, quadratic scoring rule and calibration for a Boolean model with four data points. Adapted from Allard *et al.*, (2012).

	− Loglik	S_Q	Calib.
P_0	29859.1	0.1981	0.0155
P_1	16042.0	0.0892	0.0120
Lin.	14443.3	0.0774	0.0206
BLP	9690.4	0.0575	**0.0008**
ME	7497.3	0.0433	0.0019
Log.lin	7178.0	**0.0416**	0.0010

is a not too poor approximation. The log-linear model leads to the lowest Likelihood, lowest score S_Q and very good calibration.

6 Conclusion and Discussion

When training are available, maximum likelihood provides an efficient method for estimating the parameters of any chosen model. On simulations, we were able to show that quadratic and logarithmic scores are efficient tools for determining the models leading to the best forecasts. They usually increase or decrease together. Our main result states that for log-linear poolings, calibration implies parameters estimated with maximum likelihood. The converse is not true in all generality. All simulated examples have shown that log-linear pooling formula with parameters estimated with maximum likelihood are very close to be calibrated.

A first conclusion, based on numerous simulations and examples, is that linear methods should not be used alone for aggregating probability distribution. They can be used if recalibrated with a Beta transformation whose parameters must be estimated. A second conclusion is that methods based on product of odds (Tau model) are not to be recommended. For binary events, they are equivalent to those based on product of probabilities. For non binary events they usually perform less well [Allard et al., 2012].

The main conclusion of this study is thus the following. For aggregating probability distributions, methods based on product of probabilities, in other words linear combinations of log-probabilities, should be preferred. First, they are easy to implement and to understand. Second, their parameters are easy to estimate using maximum likelihood. This has profound implications on the practice of spatial prediction and simulation of indicator functions. It implies that the kriging paradigm based on linear combinations of bivariate probabilities and its sequential indicator simulation (SIS) counterpart should probably be replaced by a different paradigm based on the product of probabilities as already proposed in Allard *et al.* (2011).

References

[Allard et al., 2011] Allard, D., D'Or, D., and Froidevaux, R. (2011). An efficient maximum entropy approach for categorical variable prediction. *European Journal of Soil Science*, 62(3):381–393.

[Allard et al., 2012] Allard, D., Comunian, A., and Renard, P. (2012). Probability aggregation methods in geoscience. *Mathematical Geosciences*, doi: 10.1007/s11004-012-9396-3.

[Bordley, 1982] Bordley, R. F. (1982). A multiplicative formula for aggregating probability assessments. *Management Science*, 28(10):1137–1148.

[Genest and Zidek, 1986] Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114148.

[Journel, 2002] Journel A (2002) Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. Math Geol 34:573-596

[Lantuéjoul, 2002] Lantuéjoul, C. (2002) Geostatistical Simulations. Springer, Berlin

[Ranjan and Gneiting, 2010] Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.