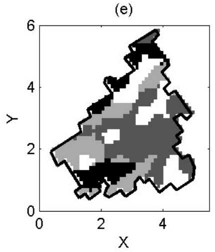
Combining probabilities with log-linear pooling : application to spatial data

Denis Allard[1], Philippe Renard[2], Alessandro Comunian[2,3], Dimitri D'Or[4]

[1] *Biostatistique et Processus Spatiaux (BioSP), INRA, Avignon*
[2] *CHYN, Université de Neuchâtel, Neuchâtel, Switzerland*
[3] *now at National Centre for Groundwater Research and Training, University of New South Wales, Sydney, Australia.*
[4] *Ephesia Consult, Geneva, Switzerland*

Réseau MSTGA
22 novembre 2012

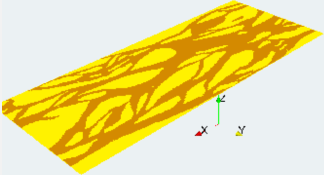
Motivation

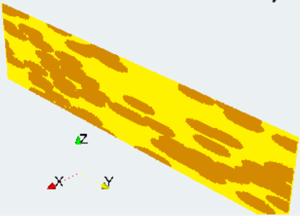


- No spatial model available for prediction, simulation
- But, there are models for covariance functions
- $\hookrightarrow$ How can we best use multiple bivariate probabilities ?

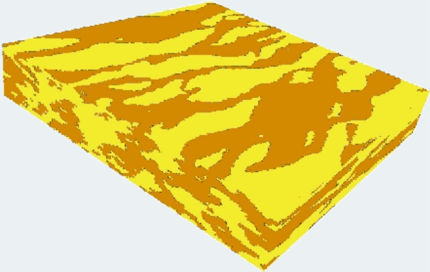
Motivation

3D simulations of geology when 2d cross-sections are known.





$\Rightarrow$



- Caers (2006)
- Okabe and Blunt (2004,2007)
- Comunian *et al.* (2011,2012)

Framework

- Consider discrete events : $A \in \mathcal{A} = \{A_1, \ldots, A_K\}$.
- We know conditional probabilities $P(A \mid D_i) = P_i(A)$, where the D_is come from different sources of information.
- We include the possibility of a prior probability, $P_0(A)$.
- Full model for $A, D_1, \ldots, D_n$ not available

Purpose

To provide an approximation of the probability $P(A \mid D_1, \ldots, D_n)$:

$$P(A|D_0, \ldots, D_n) \approx P_G(P(A|D_0), \ldots, P(A|D_n)). \qquad (1)$$

Framework

- Consider discrete events : $A \in \mathcal{A} = \{A_1, \ldots, A_K\}$.
- We know conditional probabilities $P(A \mid D_i) = P_i(A)$, where the D_is come from different sources of information.
- We include the possibility of a prior probability, $P_0(A)$.
- Full model for $A, D_1, \ldots, D_n$ not available

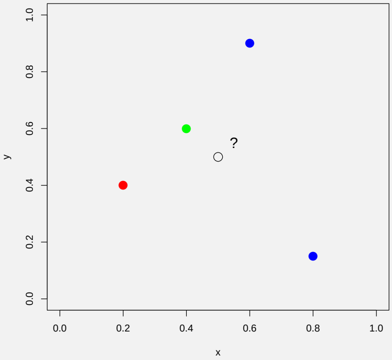
Purpose

To provide an approximation of the probability $P(A \mid D_1, \ldots, D_n)$:

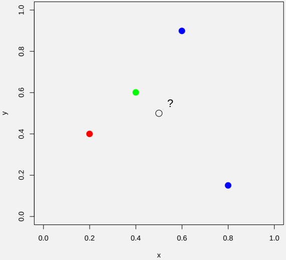
$$P(A|D_0, \ldots, D_n) \approx P_G(P(A|D_0), \ldots, P(A|D_n)). \tag{1}$$

An example : category for spatial data

- $A =$ category at a point s_0 of a domain $\mathcal{D}$
- $D_i =$ category at points s_i in domain $\mathcal{D}$
- Other possible D_i, not considered here : remote sensing information, a priori pattern



An example : category for spatial data



Typical data set :

Truth	A	$P_1(A) = P(A \mid D_1)$	$P_2(A) = P(A \mid D_2)$	$\cdots$
"blue"	"blue"	0.71	0.55	$\cdots$
	"red"	0.12	0.21	$\cdots$
	"green"	0.17	0.24	$\cdots$
"red"	"blue"	0.18	0.12	$\cdots$
	"red"	0.42	0.80	$\cdots$
	"green"	0.40	0.08	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Outline

- Mathematical properties
- Pooling formulas
- Scores and calibration
- Maximum likelihood
- Some results

Some mathematical properties

Convexity

An aggregation operator P_G verifying

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

is convex.

Unanimity preservation

An aggregation operator P_G verifying $P_G = p$ when $P_i = p$ for $i = 1, \ldots, n$ is said to preserve unanimity.

Convexity implies unanimity preservation.

In general, convexity is not necessarily a desirable property, see Example 1

Some mathematical properties

Convexity

An aggregation operator P_G verifying

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

is convex.

Unanimity preservation

An aggregation operator P_G verifying $P_G = p$ when $P_i = p$ for $i = 1, \ldots, n$ is said to preserve unanimity.

Convexity implies unanimity preservation.

In general, convexity is not necessarily a desirable property, see Example 1

Toy example

- Prior probability is 1/6 for each side
- D_1 : side is ≤ 3
- D_2 : side is even

We observe D_1 is true ; D_2 is not true

A_k	"1"	"2"	"3"	"4"	"5"	"6"
$P(A_k \mid D_1)$	1/3	1/3	1/3	0	0	0
$P(A_k \mid D_2)$	1/3	0	1/3	0	1/3	0

Toy example

- Prior probability is 1/6 for each side
- D_1 : side is ≤ 3
- D_2 : side is even

We observe D_1 is true ; D_2 is not true

A_k	"1"	"2"	"3"	"4"	"5"	"6"
$P(A_k \mid D_1)$	1/3	1/3	1/3	0	0	0
$P(A_k \mid D_2)$	1/3	0	1/3	0	1/3	0

Toy example

- Prior probability is 1/6 for each side
- D_1 : side is ≤ 3
- D_2 : side is even

Consider D_1 is true ; D_2 is not true

A_k	"1"	"2"	"3"	"4"	"5"	"6"
$P(A_k \mid D_1)$	1/3	1/3	1/3	0	0	0
$P(A_k \mid D_2)$	1/3	0	1/3	0	1/3	0
P_G = Average	4/12	2/12	4/12	0	2/12	0
$P_G \propto$ Product	1/2	0	1/2	0	0	0
$P(A_k \mid D_1, D_2)$	1/2	0	1/2	0	0	0

Similar for other combinations for D_1 and D_2

Multiplying the probabilities seems to provide sharper probabilities than adding them (here they are exact).

Toy example

- Prior probability is 1/6 for each side
- D_1 : side is ≤ 3
- D_2 : side is even

Consider D_1 is true ; D_2 is not true

A_k	"1"	"2"	"3"	"4"	"5"	"6"
$P(A_k \mid D_1)$	1/3	1/3	1/3	0	0	0
$P(A_k \mid D_2)$	1/3	0	1/3	0	1/3	0
P_G = Average	4/12	2/12	4/12	0	2/12	0
$P_G \propto$ Product	1/2	0	1/2	0	0	0
$P(A_k \mid D_1, D_2)$	1/2	0	1/2	0	0	0

Similar for other combinations for D_1 and D_2

Multiplying the probabilities seems to provide sharper probabilities than adding them (here they are exact).

Some mathematical properties

External Bayesianity

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, that is if

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

- It should not matter whether new information arrives before or after pooling
- Equivalent to the weak likelihood ratio property in Bordley (1982).
- Very compelling property, both from a theoretical point of view and from an algorithmic point of view.

Imposing this property leads to a very specific class of pooling operators.

Some mathematical properties

External Bayesianity

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, that is if

$$P_G(P_1^L, \dots, P_n^L)(A) = P_G^L(P_1, \dots, P_n)(A). \quad (3)$$

- It should not matter whether new information arrives before or after pooling
- Equivalent to the weak likelihood ratio property in Bordley (1982).
- Very compelling property, both from a theoretical point of view and from an algorithmic point of view.

Imposing this property leads to a very specific class of pooling operators.

Some mathematical properties

0/1 forcing

An aggregation operator which returns $P_G(A) = 0$ if $P_i(A) = 0$ for some $i = 1, \ldots, n$ is said to enforce a certainty effect, a property also called the 0/1 forcing property.

Linear pooling

Linear Pooling

$$P_G(A) = \sum_{i=0}^{n} w_i P_i(A), \tag{4}$$

where the w_i are positive weights verifying $\sum_{i=0}^{n} w_i = 1$

- Convex $\Rightarrow$ preserves unanimity.
- Neither verify external bayesianity, nor 0/1 forcing
- Cannot achieve calibration (Ranjan and Gneiting, 2010).

Ranjan and Gneiting (2010) proposed a Beta transformation of the linear pooling. Parameters are estimated via ML.

Log-linear pooling

Log-linear pooling

A log-linear pooling operator is a linear operator of the logarithms of the probabilities :

$$\ln P_G(A) = \ln Z + \sum_{i=0}^{n} w_i \ln P_i(A), \tag{5}$$

or equivalently

$$P_G(A) \propto \prod_{i=0}^{n} P_i(A)^{w_i}. \tag{6}$$

- Non Convex but preserves unanimity if $\sum_{i=0}^{n} w_i = 1$
- Verifies 0/1 forcing
- Verifies external bayesianity (Genest and Zidek, 1986)

Generalized log-linear pooling

Theorem (Genest and Zidek, 1986)

The only pooling operator P_G depending explicitly on A and verifying external Bayesianity is the log-linear pooling

$$P_G(A) \propto \nu(A) P_0(A)^{1-\sum_{i=1}^n w_i} \prod_{i=1}^n P_i(A)^{w_i}. \tag{7}$$

- Verifies external Bayesianity and 0/1 forcing
- $\nu(A)$ plays the role of an updating likelihood
- The sum $S_{\mathbf{w}} = \sum_{i=1}^n w_i$ plays an important role. Suppose that $P_i = p$ for each $i = 1, \dots, n$.
  - If $S_{\mathbf{w}} = 1$, the prior probability P_0 is filtered out. Then, $P_G = p$ and unanimity is preserved
  - if $S_{\mathbf{w}} > 1$, the prior probability has a negative weight and P_G will always be further from P_0 than p
  - $S_{\mathbf{w}} < 1$, the converse holds

Maximum entropy approach

Proposition

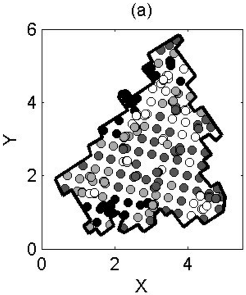
The pooling formula P_G maximizing the entropy subject to the following univariate and bivariate constraints $P_G(P_0)(A) = P_0(A)$ and $P_G(P_0, P_i)(A) = P(A \mid D_i)$ for $i = 1, \ldots, n$ is

$$P_G(P_1, \ldots, P_n)(A) = \frac{P_0(A)^{1-n} \prod_{i=1}^{n} P_i(A)}{\sum_{A \in \mathcal{A}} P_0(A)^{1-n} \prod_{i=1}^{n} P_i(A)}. \tag{8}$$

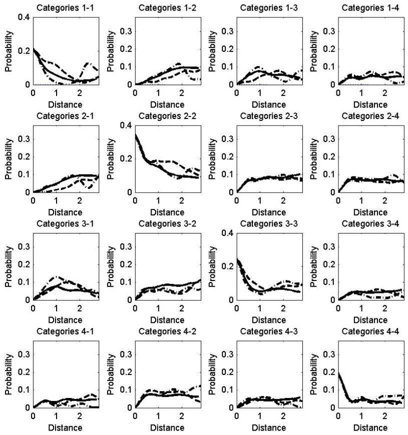
i.e. it is a log-linear formula with $w_i = 1$, for all $i = 1, \ldots, n$. Proposed in Allard (2011) for non parametric spatial prediction of soil type categories.

$$\{\text{Max. Ent.}\} \subset \{\text{Log linear pooling}\} \subset \{\text{Gen. log-linear pooling}\}.$$

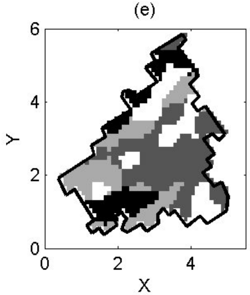
Maximum Entropy for spatial prediction



Maximum Entropy for spatial prediction



Maximum Entropy for spatial prediction



Estimating the weights for log-linear pooling

Maximum entropy is parameter free. For (generalized) log-linear pooling, how do we estimate the parameters ?

We will minimize scores

Quadratic or Brier score

The quadratic or Brier score (Brier, 1950) is defined by

$$S(P_G, A_k) = \sum_{j=1}^{K} (\delta_{jk} - P_G(j))^2 \tag{9}$$

Minimizing Brier score $\Leftrightarrow$ minimizing Euclidien distance.

Logarithmic score

The logarithmic score corresponds to

$$S(P_G, A_k) = \ln P_G(k) \tag{10}$$

Maximizing the logarithmic score $\Leftrightarrow$ minimizing KL distance.

Estimating the weights for log-linear pooling

Maximum entropy is parameter free. For (generalized) log-linear pooling, how do we estimate the parameters ?

We will minimize scores

Quadratic or Brier score

The quadratic or Brier score (Brier, 1950) is defined by

$$S(P_G, A_k) = \sum_{j=1}^{K} (\delta_{jk} - P_G(j))^2 \tag{9}$$

Minimizing Brier score $\Leftrightarrow$ minimizing Euclidien distance.

Logarithmic score

The logarithmic score corresponds to

$$S(P_G, A_k) = \ln P_G(k) \tag{10}$$

Maximizing the logarithmic score $\Leftrightarrow$ minimizing KL distance.

Estimating the weights for log-linear pooling

Maximum entropy is parameter free. For (generalized) log-linear pooling, how do we estimate the parameters ?

We will minimize scores

Quadratic or Brier score

The quadratic or Brier score (Brier, 1950) is defined by

$$S(P_G, A_k) = \sum_{j=1}^{K} (\delta_{jk} - P_G(j))^2 \tag{9}$$

Minimizing Brier score $\Leftrightarrow$ minimizing Euclidien distance.

Logarithmic score

The logarithmic score corresponds to

$$S(P_G, A_k) = \ln P_G(k) \tag{10}$$

Maximizing the logarithmic score $\Leftrightarrow$ minimizing KL distance.

Maximum likelihood estimation

Maximizing the logarithmic score $\Leftrightarrow$ maximizing the log-likelihood.

Let is consider M repetitions of a random experiment. For $m = 1, \ldots, M$:

- conditional probabilities $P_i^{(m)}(A_k)$
- aggregated probabilities $P_G^{(m)}(A_k)$
- $Y_k^{(m)} = 1$ if the outcome is A_k and $Y_k^{(m)} = 0$ otherwise

	Y_k^m	$P_1^m(A_k) = P^m(A_k \mid D_1)$	$P_2^m(A_k) = P^m(A_k \mid D_2)$	$\cdots$
$m = 1$	$Y_1^1 = 1$	0.71	0.55	$\cdots$
	$Y_2^1 = 0$	0.12	0.21	$\cdots$
	$Y_3^1 = 0$	0.17	0.24	$\cdots$
$m = 2$	$Y_1^2 = 0$	0.18	0.12	$\cdots$
	$Y_2^2 = 1$	0.42	0.80	$\cdots$
	$Y_3^2 = 0$	0.40	0.08	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Maximum likelihood estimation

Find $\mathbf{w} = (w_1, \ldots, w_n)$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$ maximazing

$$
\begin{aligned}
L(\mathbf{w}, \boldsymbol{\nu}) \quad = \quad & \sum_{m=1}^{M} \sum_{k=1}^{K} Y_k^{(m)} \left\{ \ln \nu_k + (1 - \sum_{i=1}^{n} w_i) \ln P_{0,k} + \sum_{i=1}^{n} w_i \ln P_{i,k}^{(m)} \right\} \\
& - \sum_{m=1}^{M} \ln \left\{ \sum_{k=1}^{K} \nu_k P_{0,k}^{1 - \sum_{i=1}^{n} w_i} \prod_{i=1}^{n} (P_{i,k}^{(m)})^{w_i} \right\}. \qquad (11)
\end{aligned}
$$

Calibration

Calibration

The aggregated probability $P_G(A)$ is said to be calibrated if

$$P(Y_k \mid P_G(A_k)) = P_G(A_k), \quad k = 1, \ldots, K \tag{12}$$

Theorem (Ranjan and Gneiting, 2010)

Linear pooling cannot be calibrated.

Theorem (Allard *et al.*, 2012)

Calibration $\Rightarrow$ maximum likelihood estimates.

"If P admits a log-linear pooling expression, the only calibrated parameters are those estimated from maximum likelihood."

Calibration

Calibration

The aggregated probability $P_G(A)$ is said to be calibrated if

$$P(Y_k \mid P_G(A_k)) = P_G(A_k), \quad k = 1, \ldots, K \tag{12}$$

Theorem (Ranjan and Gneiting, 2010)

Linear pooling cannot be calibrated.

Theorem (Allard *et al.*, 2012)

Calibration $\Rightarrow$ maximum likelihood estimates.

"If P admits a log-linear pooling expression, the only calibrated parameters are those estimated from maximum likelihood."

Calibration

Calibration

The aggregated probability $P_G(A)$ is said to be calibrated if

$$P(Y_k \mid P_G(A_k)) = P_G(A_k), \quad k = 1, \ldots, K \tag{12}$$

Theorem (Ranjan and Gneiting, 2010)

Linear pooling cannot be calibrated.

Theorem (Allard *et al.*, 2012)

Calibration $\Rightarrow$ maximum likelihood estimates.

"If P admits a log-linear pooling expression, the only calibrated parameters are those estimated from maximum likelihood."

Calibration

Theorem

Calibration $\Rightarrow$ maximum likelihood estimates.

Proof : A linear combination of the score functions yields

$$\sum_{m=1}^{M}\sum_{k=1}^{K} Y_k^{(m)} \ln P_{\hat{G},k}^{(m)} = \sum_{m=1}^{M}\sum_{k=1}^{K} P_{G,k}^{(m)} \ln P_{\hat{G},k}^{(m)}. \quad (13)$$

If the M experiments are drawn according to P

$$\mathbb{E}[\mathbf{Y}^t \ln \mathbf{P}_{\hat{G}}] = \mathbb{E}[\mathbf{P}_{\hat{G}}^t \ln \mathbf{P}_{\hat{G}}] \text{ as } M \to \infty. \quad (14)$$

From the conditional expectation formula :

$$\mathbb{E}[\mathbf{Y}^t \ln \mathbf{P}_{\hat{G}}] = \mathbb{E}\{\mathbb{E}[\mathbf{Y}^t \ln \mathbf{P}_{\hat{G}} \mid \mathbf{P}_{\hat{G}}]\} = \mathbb{E}\{\mathbb{E}[\mathbf{Y}^t \mid \mathbf{P}_{\hat{G}}] \ln \mathbf{P}_{\hat{G}}\}. \quad (15)$$

If $\mathbf{P}_{\hat{G}}$ is calibrated, i.e. if $\mathbb{E}[\mathbf{Y}^t \mid \mathbf{P}_{\hat{G}}] = \mathbf{P}_{\hat{G}}$, Eq. (14) is verified. Hence calibration $\Rightarrow$ weights in $P_{\hat{G}}$ are solution of the maximum likelihood.

Measure of calibration and sharpness

Recall Brier score

$$BS = \frac{1}{M} \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M} (P_G^{(m)}(A_k) - Y_k^{(m)})^2 \right\}, \qquad (16)$$

It can be decomposed in the following way :

$$BS = \text{calibration term} + \text{sharpness term} + Cte$$

- Calibration must be close to 0
- Conditional on calibration, sharpness must be as high as possible

First experiment : truncated Gaussian vector

- One prediction point s_0
- Three data s_1, s_2, s_3 defined by distances d_i and angles θ_i
- Random function $X(s)$ with exp. cov, parameter 1
- $D_i = \{X(s_i) \leq t\}$
- $A = \{X(s_0) \leq t - 1.35\}$
- 10,000 simulated thresholds so that $P(A)$ is almost uniformly sampled in $(0, 1)$

First case (symmetrical) : $d_1 = d_2 = d_3$; $\theta_1 = \theta_2 = \theta_3$

	Weight	Param.	$-$Loglik	BIC	BS	CALIB	SHARP
P_1	—	—	5782.2		0.1943	0.0019	0.0573
P_{12}	—	—	5686.8		0.1939	0.0006	0.0574
P_{123}	—	—	5650.0		0.1935	0.0007	0.0569
Lin.	—	—	5782.2	11564.4	0.1943	0.0019	0.0573
BLP	—	$\alpha = 0.67$	5704.7	11418.7	0.1932	**0.0006**	0.0570
ME	—	—	5720.1	11440.2	0.1974	0.0042	0.0564
Log.lin.	0.75	—	5651.4	**11312.0**	**0.1931**	**0.0006**	0.0571
Gen. Log.lin.	0.71	$\nu = 1.03$	**5650.0**	11318.3	0.1937	0.0008	0.0568

- Linear pooling very poor ; Beta transformation is an improvement
- Gen. Log. Lin : highest likelihood, but marginally
- Log linear pooling : lowest BIC and Brier Score
- Note that $S_{\mathbf{w}} = 2.25$

Second case (non symmetrical) : $(d_1, d_2, d_3) = (0.8, 1, 1.2)$; $\theta_1 = \theta_2 = \theta_3$

	Weight	Param.	$-$Loglik	BIC	BS	CALIB	SHARP
P_1	—	—	5786.6		0.1943	0.0022	0.0575
P_{12}	—	—	5730.8		0.1927	0.0007	0.0577
P_{123}	—	—	5641.4		0.1928	0.0009	0.0579
Lin.eq	$(1/3, 1/3, 1/3)$	—	5757.2	11514.4	0.1940	0.0018	0.0575
Lin.	$(1, 0, 0)$	—	5727.2	11482.0	0.1935	0.0015	0.0577
BLP	$(1, 0, 0)$	$\alpha = 0.66$	5680.5	11397.8	**0.1921**	**0.0004**	**0.0580**
ME	—	—	5727.7	11455.4	0.1972	0.0046	0.0571
Log.lin.eq.	$(0.72, 0.72, 0.72)$	—	5646.1	**11301.4**	0.1928	0.0006	0.0576
Log.lin.	$(1.87, 0, 0)$	—	5645.3	11318.3	0.1928	0.0007	0.0576
Gen. Log.lin.	$(1.28, 0.53, 0)$	$\nu = 1.04$	5643.1	11323.0	0.1930	0.0010	0.0576

- Optimal solution gives 100% weight to closest point
- BLP : lowest Brier score
- Log. linear pooling : lowest BIC ; almost calibrated

Simulated experiment : Boolean model

- Boolean model of spheres in 3D
- $A = \{s_0 \in \text{void}\}$
- 2 data points in horizontal plane + 2 data points in vertical plane (randomly located)
- $D_i = \{s_i \in \text{void}\}, \quad i = 1, \ldots, 4$
- $P(A \mid D_i)$ are easy to compute
- 50,000 repetitions
- $P(A)$ sampled in $(0.05, 0.95)$

Second experiment : Boolean model

	Weights	Param.	$-$ Loglik	BIC	BS	CALIB	SHARP
P_0	—	—	29859.1	59718.2	0.1981	0.0155	0.0479
P_i	—	—	16042.0	32084.0	0.0892	0.0120	0.1532
Lin.	$\simeq 0.25$	—	14443.3	28929.9	0.0774	0.0206	0.1736
BLP	$\simeq 0.25$	$(3.64, 4.91)$	9690.4	19445.7	0.0575	**0.0008**	0.1737
ME	—	—	7497.3	14994.6	0.0433	0.0019	0.1889
Log.lin	$\simeq 0.80$	—	7178.0	**14399.3**	**0.0416**	0.0010	0.1897
Gen. Log. lin	$\simeq 0.79$	$\nu = 1.04$	**7172.9**	14399.9	0.0417	0.0011	**0.1898**

- Log-linear has best scores
- Log-linear is sharper than BLP
- BS is significantly lower for Log. lin. than for BLP

Conclusions

New paradigm for spatial prediction of categorical variables :

use multiplication of probabilities instead of addition

- Demonstrated the usefulness of lig-linear pooling formula
- Optimality for parameters estimated by ML
- Very good performances on tested situations
- Outperforms BLP in some situations

To do

Implement Log-linear pooling for spatial prediction. Expected to outperform ME.

References

Allard D, Comunian A and Renard P (2012) Probability aggregation methods in geoscience Math Geosci DOI : 10.1007/s11004-012-9396-3

Allard D, D'Or D, Froidevaux R (2011) An efficient maximum entropy approach for categorical variable prediction. Eur J S Sci 62(3) :381-393

Genest C, Zidek JV (1986) Combining probability distributions : A critique and an annotated bibliography. Stat Sci 1 :114-148

Ranjan R, Gneiting T (2010) Combining probability forecasts. J Royal Stat Soc Ser B 72 :71-91