Optimisation d'un plan d'échantillonnage selon le paradigme bayésien : application au cas du saumon fumé en France

Natalie Commeau, Marie Cornu, Eric Parent

Optimisation d'un plan d'échantillonnage selon le paradigme bayésien : application au cas du saumon fumé

Natalie Commeau [1,2,3] & Marie Cornu [1] & Eric Parent [2]

1 : Agence Français de Sécurité Sanitaire des Aliments, Microbiologie quantitative et estimation des risques, Maisons-Alfort, France
2 : UMR 518 AgroParisTech / INRA Mathématiques et informatique appliquée Paris, France
3 : AgroParisTech ENGREF Paris, France

Résumé

Prenant l'exemple de la surveillance de *L.monocytogenes* dans le saumon fumé en France, nous proposons d'optimiser le plan d'échantillonnage adopté (autocontrôle industriel). Pour cela, une modélisation de ce plan est présentée en se plaçant dans le cadre de la théorie de la décision bayésienne. Le travail est poursuivi par l'utilisation d'un algorithme particulaire afin trouver le plan optimal. Cet algorithme permet d'explorer des espaces de dimension élevée et de concentrer les particules sur le mode d'une distribution grâce à la technique du recuit simulé.

Abstract

Relying on the Bayesian approach, we model a sampling plan to assess *Listeria monocytogenes* contamination in French cold smoked salmon and try to determine the best plan. To solve this optimization problem we will use a particle sampler combined with simulated annealing to concentrate the particles near the modes of a target probability density function.

Mots-clés : plans d'échantillonnage, théorie bayésienne de la décision, algorithme particulaire, *Listeria monocytogenes*

Introduction

Le contrôle de qualité est l'un des outils utilisé par les entreprises afin de déterminer si un produit possède les propriétés requises pour être mis en vente. La plupart des méthodes de contrôle utilisent l'approche classique, or celle-ci n'est pas très facile à mettre en place lorsqu'il existe plusieurs points d'échantillonnage ou que le modèle comprend des paramètres de dimension élevée (Amzal et al., 2006). Pour optimiser un plan d'échantillonnage complexe, on se place ici dans le cadre de la théorie de la décision d'un

point de vue bayésien (Berger, 1985). L'optimum est déterminé grace à algorithme stochastique de simulation basée sur les chaînes de Markov Monte-Carlo (MCMC).
Le contrôle de qualité choisi porte sur la surveillance du pathogène *Listeria monocytogenes* dans une usine de fabrication de saumon fumé en France. *L.monocytogenes* peut provoquer la listeriose, une maladie fatale dans environ 20% des cas. Le nombre de malades augmente à nouveau depuis le début des années 2000 en Europe, la France compte environ 220 cas par an. Le saumon fumé est un aliment dit "*ready-to-eat*" car il n'y a pas d'étape de cuisson entre la sortie usine et la consommation, de plus, il permet la croissance de *L.monocytogenes*. Il a donc fait l'objet de plusieurs évaluations des risques (Pouillot et al., 2009 et FAO/WHO, 2004) et apparaît comme un vecteur à ne pas négliger. Il est donc essentiel de maîtriser la contamination du saumon fumé.
L'objet du travail présenté ici est de modéliser une stratégie d'échantillonnage efficace, prenant en charge les conséquences économiques, puis de déterminer le plan optimal et son budget.

Règle de décision

Une règle de décision d est une fonction définie de l'ensemble des données vers l'ensemble $\mathfrak{D}$ des actions : cette stratégie fait correspondre à chaque donnée observée y une action $d(y)$, également appelée décision. La fonction d'utilité $u(d(y), \theta)$ quantifie la préférence du décideur pour la décision d, sachant que le résultat observé est y et le paramètre θ. Le décideur doit prendre une décision d en univers incertain puisqu'il ne connaît pas le paramètre θ. La décision optimale d^* est celle qui minimise l'espérance de u prise sur la distribution du couple (y, θ). Afin de déterminer ce minimum, une approche bayésienne est adoptée (au lieu d'une approche classique qui consisterait à évaluer l'espérance de u pour différentes valeurs de θ).
On se place dans le cadre d'une analyse normale (ou prédictive) : les données y n'ont pas encore été observées. On cherche alors la règle d^* qui minimise :

$$R(d) = \int \int u(d(y), \theta)[y, \theta] dy d\theta$$

Les distributions de probabilité sont notées entre [ ].

Application au cas du saumon fumé : règle de décision et coûts

Dans le cas présent, la règle de décision est représentée sur la figure 1. On se place à l'échelle d'un lot de saumon fumé dont la concentration moyenne en *L.monocytogenes* vaut θ (en ufc/g [1]). La distribution *a priori* sur la concentration est une loi gamma inter-lots ajustée sur des données recueillies dans l'article de Beaufort et al., 2007 (usine 8). Un nombre n_1 d'échantillons de masse m_1 sont prélevés par tirage au sort. On note y_1 le nombre de produits positifs à l'issue des analyses.

[1] ufc : unité formant colonie

– si $y_1 \leq s$, le lot est accepté (s est un seuil prédéfini) ;
– si $y_1 > s$, on procède à des analyses de dénombrement.

Dans ce cas, n_2 échantillons supplémentaires de masse m_2 sont prélevés dans le lot, également par tirage au sort. Le résultat des concentrations en ufc/g pour chaque échantillon j, $j = 1, \ldots, n_2$ est noté y_2.

– si aucun y_{2j} n'atteint la concentration seuil A, le lot est accepté ;
– si au moins une concentration y_{2j} atteint ou dépasse A, le lot est rejeté.

Ici, la fonction d'utilité u associe un coût à la décision prise. On note k_1 le coût d'une détection et k_2 celui d'un dénombrement. Quels que soit la décision et le chemin pour y parvenir, l'entreprise paie les analyses de détection et, le cas échéant, celles de dénombrement. Si l'entreprise vend le lot, il est possible qu'elle mette sur le marché des produits contaminés dont certains peuvent provoquer des listérioses. Si cela arrive et si l'enquête parvient à remonter à l'usine responsable, les dommages subis par celle-ci peuvent être très importants en terme d'image de marque et de part de marché. Pour l'instant, ce coût est noté $f(\theta)$ car il dépend de la concentration moyenne en *L.monocytogenes* dans le lot et sera explicité plus bas dans le document. Si le lot est rejeté, l'entreprise ne reçoit pas le prix de la vente, montant fixe que l'on note V. Les coûts s'écrivent donc en résumé :

$$\begin{cases} u(d(y_1, y_2), \theta) = k_1 n_1 + f(\theta), & \text{si } y_1 \leq s \\ u(d(y_1, y_2), \theta) = k_1 n_1 + k_2 n_2 + f(\theta), & \text{si } y_1 > s \text{ et } max_j(y_{2j} < A) \\ u(d(y_1, y_2), \theta) = k_1 n_1 + k_2 n_2 + V, & \text{si } y_1 > s \text{ et } max_j(y_{2j} \geq A) \end{cases}$$

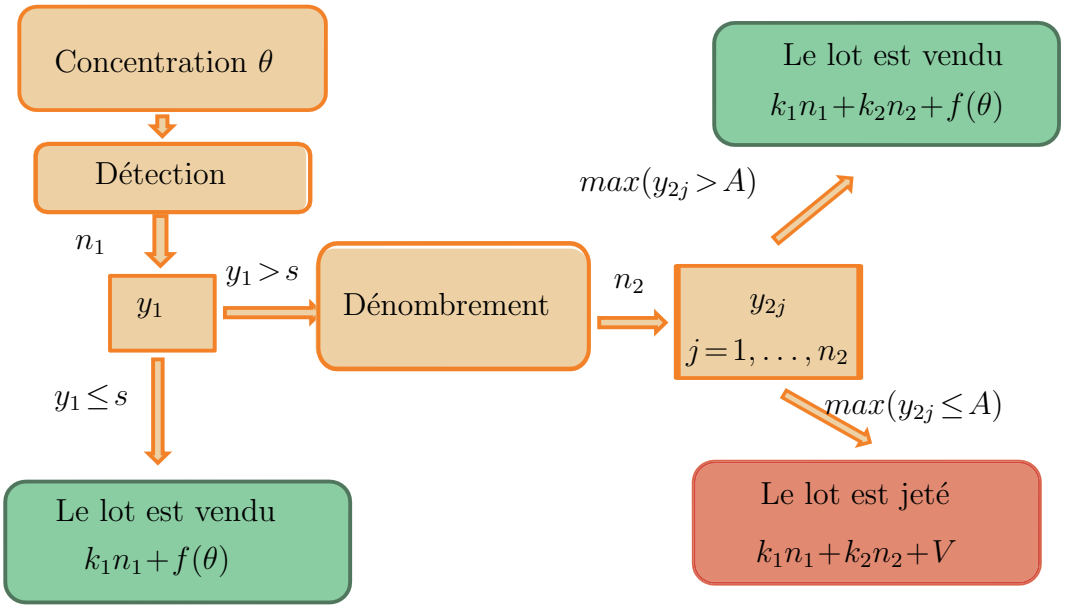

FIG. 1 – Décisions possibles et coûts engendrés

Incertitude de mesure

Les analyses de détection et de dénombrement fournissent des résultats entachés d'incertitude de plusieurs ordres. Considérons une prise d'essai contenant z *L.monocytogenes*. Si on considère que ce pathogène est réparti de manière homogène dans la matrice, pour une masse m_1 d'échantillon prélevé on peut imaginer une répartition Poissonienne : $z \sim \mathcal{P}ois(\theta * m_1)$. C'est une approximation car le saumon fumé est un aliment solide. Cependant, à la fin de la fabrication, les colonies n'ont pas encore eu le temps de se former (ou alors elles sont très petites). Les différences de concentration entre les tranches sont peu élevées, donc l'hypothèse Poissonienne est réaliste.

Détection Wilrich, se sont penchés sur la probabilité d'avoir un résultat positif de détection en fonction de la dose de *L.monocytogenes*. Aucune différence significative n'a été mise en évidence entre cette courbe et le résultat obtenu si la méthode avait une sensibilité égale à 1. On considère donc la mesure "parfaite" *i.e.* le résultat est positif si $z > 0$ et négatif sinon, autrement dit $y_1 \sim Bin(n_1, 1 - e^{-\theta m_1})$.

Dénombrement Un dénombrement se déroule de la manière suivante : une prise d'essai de m_2 g est prélevée sur une tranche de saumon fumé. Elle est placée dans une solution de $9 * m_2$ mL d'eau peptonée puis stomaché (un stomacher est un appareil muni de deux pales qui appuient successivement sur la solution de manière à broyer le saumon afin d'homogénéiser la solution). 1 mL de cette solution est prélevé et déposé sur une boîte de Pétri. Les colonies sont décomptées au bout de 24h. Pour plus de détail sur la méthode, se référer à la norme ISO 112902 (Anon., 1998). Pour la prise d'essai j ($j = 1, \ldots, n_2$), le nombre de *L. monocytogenes* est $z_{2j} \sim \mathcal{P}ois(\theta m_2)$. Le nombre de colonies présentes sur la boîte de Pétri est $x_{2j} \sim \mathcal{P}ois(z_{2j} d)$ avec d la dilution. La concentration en *L. monocytogenes* est donc : $y_{2j} = x_{2j}/(m_2 * d)$.

Forme de la fonction f

Comme évoqué plus haut, $f(\theta)$ quantifie les coûts de perte d'image de marque et de part de marché consécutifs à l'identification d'une causalité entre listériose(s) et saumon fumé contaminé issu de l'entreprise considérée.
Pour obtenir la probabilité d'être malade P_{inf} à partir de la concentration en sortie usine θ, on utilise le modèle de Pouillot et al., 2007. Celui-ci permet de passer de θ à θ_{conso} la concentration moyenne du lot au moment de la consommation, puis de θ_{conso} à P_{inf}.
Pour obtenir la probabilité d'être malade pour le lot entier, il faut multiplier P_{inf} par le nombre de portions typiques contenues dans un lot. Enfin, il reste à multiplier le résultat obtenu par le coût de perte d'image de marque et de parts de marché si une personne ayant consommé du saumon fumé contracte une listériose. Une manière d'estimer ce coût est d'examiner le montant de la couverture d'assurance que l'entreprise a souscrit afin se

prémunir contre ce genre de situation.

Quel plan adopter ?

Dans cet exemple, la décision est caractérisée par le quadruplet $\mathbf{d} = (n_1, n_2, s, A)$ qui détermine entièrement le plan d'échantillonnage. Le but recherché est de déterminer le quadruplet tel que $U(\mathbf{d}) = \int u(\omega, \mathbf{d})[\omega|\mathbf{d}]d\omega$ soit minimale, avec $\omega = (y_1, y_2, \theta)$. Transformer u en $max(u) - u$ revient à rechercher le maximum de la nouvelle fonction. Pour trouver le maximum, l'idée, développée initialement par Müller, 1999 est de considérer une densité de probabilité $f \propto u(\omega, \mathbf{d})[\omega|\mathbf{d}]$. La densité marginale de f selon $\mathbf{d}$ est proportionnelle à U. Déterminer le maximum de U revient à chercher le mode de cette densité marginale. Comme f n'est connue qu'à une coefficient multiplicatif près, il faut un algorithme pour tirer des valeurs selon cette loi. On transforme donc un problème d'optimisation en un problème de simulation. Le mode de f peut ne pas se détacher beaucoup des autres points. Une manière d'accentuer ce mode est de faire du recuit-simulé en considéant f élevée à une puissance J.
L'algorithme de simulation utilisé est l'algorithme particulaire (Parent et al., 2008). Pour trouver le mode de f^J en augmentant J progressivement, on utilise un apprentissage statistique séquentiel qui concentre les particules sur le mode.

Intérêt du modèle

Cet algorithme a déjà été utilisé pour trouver le mode d'une distribution même lorsque l'espace des décisions et des paramètres est élevé mais lorsque le paramètre θ a la même valeur lorsque la décision est prise et lorsque le produit est utilisé. L'originalité introduite ici est double : les résultats observés par les analyses ne correspondent pas exactement à la réalité à cause de l'incertitude de mesure ; la décision prise se fait en fonction de la concentration moyenne θ au moment de la sortie d'usine alors qu'un consommateur mangera du saumon fumé avec une concentration moyenne θ_{conso} en *L.monocytogenes* différente dont on ne peut avoir une idée très précise, lorsque le lot quitte l'entreprise de fabrication. En effet, d'après Pouillot et al., 2009 le lien entre la contamination du saumon fumé en sortie usine et le risque d'être malade lors de sa consommation n'est pas très fort : un saumon faiblement contaminé juste après sa fabrication peut aboutir à un produit fortement contaminé et donc conduire à un risque de maladie élevé. De plus, l'enchaînement entre les observations et la décision prise est ici relativement complexe puisque les types de contrôle varient en fonction de la valeur prise par y_1.
L'approche développée ici est très flexible puisque les distributions sur les paramètres et la fonction u peuvent être modifiés. Elle devrait pouvoir être facilement transposée à des cas pratiques très variés.

Bibliographie

Amzal, B., Bois, F., Parent, E., and Robert, C. (2006). Bayesian optimal design via interacting particle systems. *Journal of the American Statistical Association*, 101(474) :773–785.

Anon. (1998). Méthode horizontale pour la recherche et le dénombrement de *Listeria monocytogenes*. *AFNOR*.

Beaufort, A., Rudelle, S., Gnanou-Besse, N., Toquin, M.-T., Kerouanton, A., Bergis, H., Salvat, G., and Cornu, M. (2007). Prevalence and growth of *Listeria monocytogenes* in naturally contaminated cold-smoked salmon. *Letters in applied microbiology*, 44(4) :406-411.

Berger, J. (1985). *Statistical decision theory and bayesian analysis*. Springer Verlag, New York, second edition.

FAO/WHO (2004). *Risk assessment of Listeria monocytogenes in ready-to-eat food*. Microbiological risk assessment series. Roma. 269.

Müller, P. (1999). Simulation-based optimal design. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics*, volume 6, pages 459–474. Oxford University Press.

Parent, E., Amzal, B., and Girard, P. (2008). Investigations particulaires pour l'inférence statistique et l'optimisation de plan d'expériences. *Journal de la Sfds*, 149 (1).

Pouillot, R., Goullet, V., Delignette, M.-L., Mahé, A., and Cornu, M. (2009). Quantitative risk assessment of *Listeria monocytogenes* in French cold smoked salmon : II. Risk characterization (in press). *Risk analysis*.

Pouillot, R., Miconnet, N., Afchain, A.-L., Delignette-Luller, M.-L., Beaufort, A., R.-L., Denis, J.-B., and Cornu, M. (2007). Quantitative risk assessment of *Listeria monocytogenes* in French cold smoked salmon : I. Quantitative exposure assessment. *Risk analysis*, 27(3) :683–700.

Wilrich, C. et Wilrich, P. Estimation of the pod function and the lod of a qualitative microbiological measurement method. *Document de travail du groupe 'ISO/TC 34/SC 9/WG 2 Statistics*.