Recherche

Modélisation de la contamination par *Listeria monocytogenes* pour l'amélioration de la surveillance dans les industries agro-alimentaires

Natalie Commeau (natalie.commeau@gmail.com)
Inra/AgroParisTech, UMR 518 MIA, Paris, France
Anses, Laboratoire de sécurité des aliments, Maisons-Alfort, France

Les industriels du secteur agro-alimentaire sont responsables de la qualité des produits mis sur le marché. Un moyen de vérifier cette qualité consiste à déterminer la distribution de la contamination. Un outil utile est le plan d'échantillonnage. Nous proposons une approche basée sur la théorie de la décision en sortie usine pour déterminer la taille optimale de l'échantillon à prélever par lot de manière à minimiser le coût moyen supporté par le fabricant. Dans cet article, nous utilisons des données portant sur *L. monocytogenes* durant la fabrication de lardons. Nous avons élaboré des modèles pour décrire la concentration en prenant ou non en compte diverses variabilités, nous avons estimé les paramètres par inférence bayésienne, puis comparé leur capacité à simuler des données proches des observations. Enfin, nous présentons une application de minimisation des coûts moyens de l'entreprise pour le couple *L. monocytogenes*/lardons.

Échantillonnage intra- et inter- lots

La connaissance des différentes contaminations par des pathogènes dans une usine est nécessaire à l'industriel afin qu'il puisse mettre en place des actions adaptées pour réduire la contamination. Pour acquérir cette connaissance, des analyses sont nécessaires (par exemple dénombrement ou recherche) sur des produits alimentaires ou sur des surfaces. Comment effectuer les prélèvements à une étape donnée de la production : en tirant au sort dans la production ou en tirant au sort dans des lots ? Cette question n'est pas anodine car, selon le produit fabriqué et le procédé, la variabilité de contamination au sein d'un lot ou entre plusieurs lots peut être très différente. Ainsi, la **figure 1** présente les distributions de contamination au sein de plusieurs lots. La variabilité inter-lots (1a) est bien plus faible que la variabilité intra-lot donc échantillonner en tirant au sort dans la production totale sans se préoccuper d'une appartenance à un lot suffit. En revanche, si les distributions sont celles de la figure 1b, alors l'échantillonnage par lot est déterminant pour savoir si un lot est peu contaminé ou non. La variabilité inter et intra-lot a fait l'objet de publications récentes (ILSI, 2010 ou Gonzales-Barron et Butler, 2011).

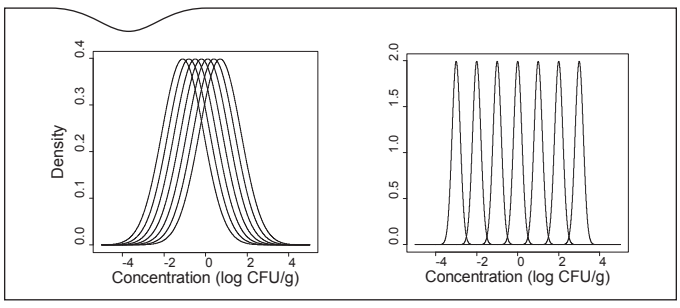

Figure 1 : Représentation de la variabilité inter et intra-lots. Chaque courbe représente la distribution de contamination d'un lot (log ufc/g). Pour la figure 1a (gauche), l'écart-type (e.t.) de la contamination dans un lot est de 1 log ufc/g et l'e.t. inter-lot est de 0,3 log ufc/g/. Pour la figure 1b (droite) l'e.t. inter-lots est égal à 1 log ufc/g et l'e.t. intra-lot est de 0,2 log d'ufc/g.

Avant de poursuivre, définissons le terme de « lot ». C'est un terme fréquemment utilisé dans le langage courant mais pas si simple à définir. D'après le règlement N° 2073/2005 (CE) (article 2), un lot est « un groupe ou une série de produits identifiables obtenus par un procédé donné dans des conditions pratiquement identiques et produits dans un endroit donné et au cours d'une période de production déterminée ». La définition donnée par l'ICMSF (ICMSF, 2002) commence par énoncer que c'est une quantité d'aliments fabriqués et manipulés dans des conditions uniformes, mais elle va plus loin car il est précisé que cette définition implique qu'il y a une homogénéité au sein du lot, par exemple que le logarithme de la concentration suit une loi normale. Cependant, les auteurs remarquent que cette condition d'homogénéité n'est pas toujours vérifiée dans un lot pour la concentration microbienne car celle-ci peut être très hétérogène. Il est donc suggéré d'adapter la taille du lot en fonction du procédé. L'homogénéité est indispensable au statisticien pour qu'il puisse concevoir une distribution décrivant la contamination de la production. Le lot défini par l'industriel repose sur des contraintes de traçabilité et d'organisation.

Détermination de la structure de la contamination dans une usine de lardons

Afin de déterminer la structure de la contamination, nous avons effectué des prélèvements de poitrines de porc après malaxage dans une usine fabriquant des lardons nature et non avons analysé la présence et la concentration en *L. monocytogenes*. Le lot a été défini comme l'ensemble des poitrines de porc contenues dans un malaxeur, première étape du procédé de fabrication. Au total, 8 ou 9 poitrines de porc ont été prélevées sur 12 lots différents. Pour chaque poitrine, 100 cm² de chair étaient prélevés et analysés pour détection et dénombrement de *L. monocytogenes*. Avec les protocoles utilisés, la limite de détection était de 0.01 UFC/cm² (unités formant colonie) et la limite de quantification de 0.2 UFC/cm². Les données brutes (présence ou absence pour la détection, nombre de colonie dénombrées pour le dénombrement) sont présentées au **Tableau 1**.

Recherche

Tableau 1 : données brutes de détection (0=absence, 1=présence) et de dénombrement (nombre d'UFC comptées sur boite de Pétri) de *L. monocytogenes* sur les échantillons de 100 cm² de poitrine de porc après malaxage, c.a.d. la première étape du processus de production de lardons, durant laquelle les poitrines de porc sont malaxées avec de la saumure pendant plusieurs heures. Le même échantillon était utilisé pour la détection et le dénombrement.

Nombre de lots	Détection	Dénombrement (UFC)
1	0-1-1-1-0-1-1-1-1	0-0-0-0-0-0-0-0-0
2, 8, 9 & 10	0-0-0-0-0-0-0-0-0	0-0-0-0-0-0-0-0-0
3	0-0-0-0-1-0-0-0-1	0-0-0-0-1-0-0-0-0
4	0-1-0-0-1-0-0-0-0	0-0-0-0-0-0-0-0-0
5	0-0-0-0-1-0-0-0-0	0-0-0-0-0-0-0-0-0
6	0-0-0-0-0-0-0-0	0-0-0-0-0-0-0-0
7	0-0-0-1-1-0-0-0-1	0-0-0-0-0-0-0-0-0
11	1-1-1-1-1-1-1-1	11-9-6-5-12-29-16-3
12	0-0-1-0-0-0-0-0-0	0-0-0-0-0-0-0-0-0

Les résultats ont été injectés dans quatre modèles de contaminations :
- avec une structure d'unités et de lots (modèle REF) ;
- avec une structure de lots (modèle B) ;
- avec une structure d'unités (modèle U) ;
- sans structure (modèle NS).

L'unité est ici la poitrine de porc, car nous nous demandions si les variabilités inter et intra-unités existent au même titre que les variabilités inter et intra-lots. L'approche retenue est l'approche bayésienne, qui présente l'avantage d'incorporer de l'information autre que les données dans le modèle. Tous les modèles combinent les distributions binomiales, Poisson et normales. Les modèles NS, B et U sont inclus dans le modèle REF.
Le modèle REF est le suivant : soit x_{ijk} le résultat de détection (1 si le résultat est positif, 0 s'il est négatif) du lot *i*, de la poitrine de porc *j* et de la portion test *k*, et y_{ijkl} le résultat de dénombrement du lot *i*, de la poitrine de porc *j*, de la prise d'essai *k* et une fraction *l*. Une prise d'essai est l'échantillon de chair sur laquelle les expériences sont réalisées (ici 100 cm²). La fraction est le volume de la solution formée par la portion diluée dans le liquide de culture approprié qui est versée sur la boîte de Pétri pour dénombrement de *L. monocytogenes*.
La variable x_{ijk} suit une distribution binomiale et la variable y_{ijk} suit une distribution de Poisson :

$$x_{ijk} \sim Bin(1, 1 - exp(-10^{\theta_{ij}} S_k))$$
$$y_{ijkl} \sim P(10^{\theta_{ij}} S_k d_l)$$

où θ_{ij} est le logarithme en base 10 de la concentration de *L. monocytogenes* dans une poitrine de porc *j* appartenant au lot *i*, S_k est la surface de la prise d'essai *k* et *d* est la dilution de la fraction *l*. Le logarithme de la concentration θ_{ij} suit une distribution normale:

$$z_i \sim N(\mu, \sigma^2)$$

où z_i est le log de la concentration de *L. monocytogenes* dans un lot *i* et λ^2 est la variance des logarithme de concentration entre poitrines de porc. Le logarithme de la concentration z_i suit aussi une distribution normale :

$$\theta_{ij} \sim N(z_i, \lambda^2)$$

où μ est le logarithme de la concentration moyenne et σ^2 est la variance du logarithme de la concentration entre lots. Pour les distributions *a priori*, le paramètre μ suit une distribution normale et σ^2 et λ^2 suivent toutes deux une distribution inverse gamma.
Il n'y a pas d'effet unité dans le modèle B, donc $\lambda=0$. Inversement, il n'y a pas d'effet lot dans le modèle U, donc $\sigma=0$. Le modèle NS n'a aucun des deux effets mentionnés ci-dessus, donc $\lambda=\sigma=0$. Les modèles B, U et NS sont décrit dans au **Tableau 2**.

Tableau 2 : Description des modèles B, U et NS. Les indices *i, j, k* et *l* font référence respectivement à un lot, une poitrine de porc, une portion test et à la fraction analysée.

Model B	Model U	Model NS
$x_{ik} \sim Bin(1,1 - exp(-10^{z_i} S_k))$ $y_{ikl} \sim P(10^{z_i} S_k d_l)$ $z_i \sim N(\mu, \sigma^2)$	$x_{jk} \sim Bin(1,1 - exp(-10^{\theta_j} S_k))$ $y_{jkl} \sim P(10^{\theta_j} S_k d_l)$ $\theta_j \sim N(\mu, \lambda^2)$	$x_k \sim Bin(1,1 - exp(-10^{\mu} S_k))$ $y_{kl} \sim P(10^{\mu} S_k d_l)$

Pour déterminer les paramètres de la distribution *a priori*, des résultats d'auto-contrôles pratiqués par différentes entreprises du secteur ont été utilisés. Les distributions *a posteriori* des paramètres ont été estimées grâce au logiciel OpenBugs (Thomas *et al.* 2006). Dans le protocole utilisé pour les données décrites au Tableau 1, S_k=100 cm² et d_l=0,05. Les quantiles des distributions *a posteriori* pour les quatre modèles sont indiqués dans le **Tableau 3**.

Tableau 3 : Statistiques descriptives des distributions postérieures des modèles REF, B, U et NS.

Modèle	Paramètres	Statistiques descriptive des distributions postérieures				
		Moyenne	E.t.	2.5ème perc.	50ème perc.	97.5ème perc.
REF	μ	-3,09	0,53	-4,25	-3,05	-2,15
	σ	1,55	0,49	0,89	1,45	2,77
	λ	0,38	0,08	0,25	0,36	0,57
B	μ	-3,12	0,51	-4,21	-3,09	-2,18
	σ	1,72	0,47	1,06	1,63	2,86
U	μ	-3,51	0,15	-3,81	-3,51	-3,21
	λ	1,99	0,24	1,59	1,97	2,51
NS	μ	-0,94	0,.005	-0,95	-0,94	-0,93

"E.t." = écart-type, et "perc." = percentile.

Nous avons étudié la capacité des modèles à répliquer les données réelles à partir d'un critère visuel basé sur des données simulées. Les données de détection ont été simulées à partir des distributions *a posteriori* des paramètres (même nombre de données par lot et même nombre de lots que pour les données observées), puis les proportions de lots avec (1) uniquement des présences, (2) uniquement des absences ou (3) un mélange de présences et d'absences ont été décomptés. Ce processus a été répété *n* fois pour calculer la médiane et les intervalles de crédibilité à 50 % et 95 % des proportions de lots dans chacune des catégories. Un intervalle de crédibilité à *x* % signifie que *x* % des données simulées sont comprises dans l'intervalle. Le même procédé a été utilisé pour les données de dénombrement. Les résultats sont présentés à la **Figure 2**. Le modèle répliquant le mieux

Recherche

les données est le modèle B. Le modèle REF réplique un peu moins bien les données (données non montrées). Le modèle B est donc le meilleur des quatre modèles.

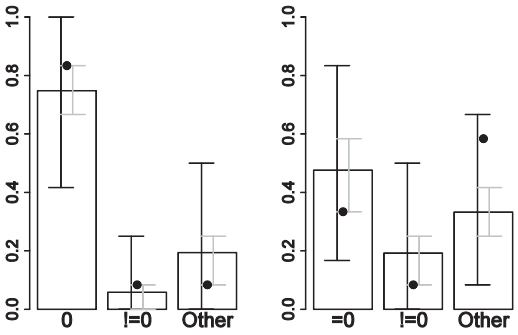

Figure 2a

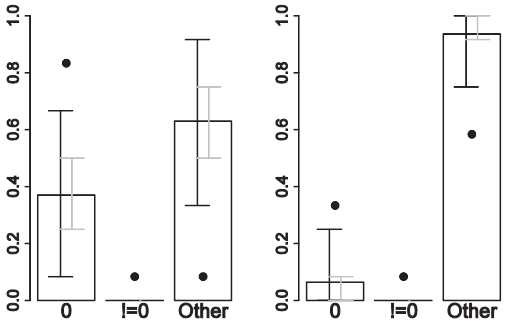

Figure 2b

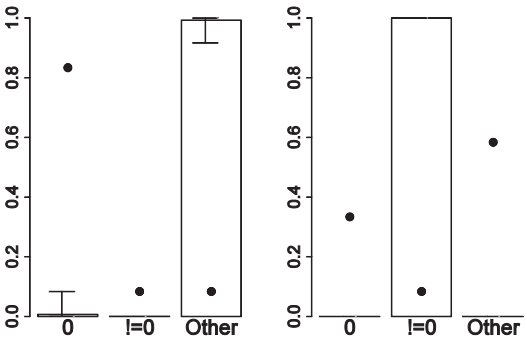

Figure 2c

Figure 2 : Données de contamination de poitrines de porc en sortie malaxage simulées et observées pour (a) le modèle B, (b) le modèle U et (c) le modèle NS. Chaque figure de gauche représente les données de détection et chaque figure de droite les données de dénombrement. Les histogrammes représentent la moyenne de chaque groupe ("=0" : proportion de lots ne comportant que des données nulles, "!= 0" : proportion de lots ne comportant que des données non nulles, "Other" : tous les autres lots). Les segments de droite représentent les intervalles de crédibilité à 50 % et les segments de gauche les intervalles de crédibilité à 95 %. Les points sont les données observées.)

Exemple illustrant la détermination de la taille optimal de l'échantillon pour la minimisation des coûts.

Connaître la distribution de la contamination permet de définir une stratégie pour les plans d'échantillonnage. Néanmoins, pour cette dernière, il faut également tenir compte des pratiques de l'industriel et des raisons de l'échantillonnage. Il existe plusieurs types de plans d'échantillonnage dans l'industrie agro-alimentaire, on se limitera ici au plan par attribut à deux classes. Le principe est le suivant : n produits sont prélevés et des analyses de recherche sont menées (généralement dans 25 g de produit). Si le nombre d'analyses positives y dépasse un certain nombre c, alors le lot est rejeté (détruit ou vendu dans une autre filière par exemple), sinon, le lot est livré. Ce type de plan suppose que la marchandise soit encore dans l'usine lorsque les résultats sont disponibles, ce qui n'est pas toujours le cas. Pour nous adapter à notre application, nous avons un peu modifié la définition du plan d'échantillonnage. Nous avons discuté avec un expert du secteur et nous sommes arrivés au plan suivant :

- l'échantillonnage ne porte pas sur un lot de production mais sur une période (1 semaine ou 1 mois par exemple) ;
- en fonction de la valeur de x, 3 décisions possibles sont prises par l'entreprise (ne rien faire, entreprendre une action corrective mineure car la prévalence en *L. monocytogenes* durant la période de production est jugée moyenne, entreprendre une action corrective majeure car la prévalence est jugée élevée).

Notre intention était de déterminer les valeurs optimales de la taille n de l'échantillon ainsi que des seuils c_1 et c_2, respectivement les valeurs de x au-delà desquelles l'action corrective mineure et l'action corrective majeure sont prises. Pour y parvenir, nous avons utilisé la théorie bayésienne de la décision. L'objectif de cette théorie est de déterminer la meilleure solution pour un opérateur en situation d'incertitude. Ceci est obtenu par différentes étapes :

- déterminer l'ensemble $\mathcal{D}$ de toutes les décisions possible (soit ici les 3 décisions décrites ci-dessus) ;
- déterminer toutes les valeurs de $\mathcal{S}$ des états de la nature (ici, la contamination par *L. monocytogenes* des poitrines de porc) et les distributions *a priori* ;
- déterminer la série de toutes les observations $\mathcal{O}$ (ici la détection et le dénombrement) et leurs distributions ;
- définir la fonction de perte L définie pour $\mathcal{D} \times \mathcal{S} \times \mathcal{O}$ in $\mathbb{R}^+$ (voir ci-dessous) ;
- déterminer la meilleur règle de décision (fonction qui à un ensemble d'observation o associe une décision d), obtenue en minimisant l'espérance de la fonction L sur les états de la nature et les observations.

Pour plus d'informations sur cette théorie, voir Berger, 1985 ; Parent, 2007 et Robert, 2006.

En fonction de la prévalence de la production sur une période, le client (ici un distributeur) peut appliquer une pénalité pour non-respect du cahier des charges et imposer des contrôles supplémentaires pendant une certaine période de temps. Le montant des pénalités dépend de la prévalence (plus la prévalence est forte et plus la pénalité est élevée) mais est modulé par l'éventuelle action corrective faite par l'entreprise (si l'entreprise a mis en place une action corrective, la pénalité diminue). Pour simplifier, la prévalence a été répartie en trois classes : faible, moyenne et forte. Nous avons demandé à notre expert d'estimer les coûts des pénalités et des actions correctives. Ces derniers sont résumés au **Tableau 4**.

Recherche

Tableau 4 : Coût supporté en euros par l'entreprise en fonction de la prévalence de la production et de la décision prise.

		Décision prise		
		Ne rien faire	a.c. mineure nécessaire	a.c. majeure nécessaire
Situation réelle	Prévalence faible	0	4250	14000
	Prévalence moyenne	6200	6110	14930
	Prévalence forte	92050	31900	27800

À tous ces coûts, il convient d'ajouter le coût d'échantillonnage, estimé à 20 euros par l'expert. Afin de mener le calcul à son terme, nous avons déterminé les seuils de prévalence : en dessous de 0,2, la prévalence est considérée comme faible et au-dessus de 0,6, elle est forte. Enfin, la distribution Beta de paramètres 2 et 3 a été utilisée pour décrire la prévalence (voir **figure 3**). Une distribution beta de paramètres α et β a une fonction de distribution de probabilité égale à $\frac{\Gamma(\alpha+\beta)}{\Gamma\alpha+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta}$, où $\Gamma t = \int_0^{\infty} z^{t-1}e^{-z}dz$. Avec toutes ces informations, nous pouvons calculer la fonction de perte qui associe un coût à un ensemble d'observation et une valeur de contamination. La décision ne dépend que de la valeur des observations ; leur connaissance déterminera automatiquement la décision à prendre.

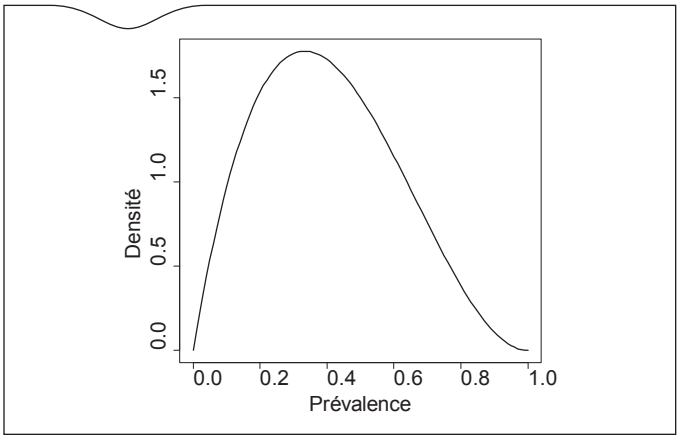

Figure 3 : Distribution de la prévalence entre les différentes périodes de production.

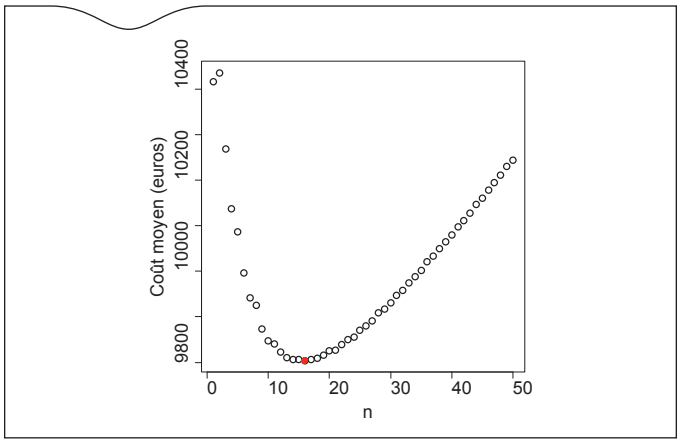

Figure 4 : Coût moyen en euros supporté par l'entreprise en fonction de la taille *n* de l'échantillon. Le coût minimal est atteint pour n=16, c_1=4 et c_2=11 (point rouge).

Munis de ces informations, nous avons calculé le coût moyen, selon la prévalence et les résultats des analyses, par période de production pour l'entreprise en fonction de la taille de l'échantillon. En faisant varier cette taille, nous pouvons déterminer quelle valeur minimise le coût moyen. Avec les valeurs numériques choisies, les différents coûts moyens sont présentés à la **figure 4**. Le minimum est atteint pour n=16, c_1=4 et c_2=11.

D'autres valeurs pour la distribution et les seuils de la prévalence ont été testé pour étudier leur impact sur le plan d'échantillonnage. Ainsi, lorsque la distribution beta a pour paramètres 2 et 20 et que les seuils de prévalence sont de 0,05 et 0,1, alors le coût moyen supporté par l'entreprise est atteint pour n=48, c_1=1 et c_2=6, ce qui est très différent du résultat précédent. Si les coûts varient, les résultats du plan d'échantillonnage varient également.

L'application de la théorie bayésienne de l'information permet d'apporter une aide au décideur en univers incertain. Elle nécessite de définir la population sur laquelle on travaille (définition du lot), de modéliser de la prévalence, de définir l'ensemble des décisions et des conséquences possibles, de déterminer les coûts et enfin d'effectuer des calculs probabilistes. Les valeurs finales sont très dépendantes du modèle employé et des coûts, ce qui signifie que toutes les entrées des modèles doivent être définies avec attention. Pour plus d'information sur ce travail, voir Commeau (2012).

Références

Berger J. 1985. Statistical decision theory and Bayesian analysis. *Springer Verlag, New York, second edition* : 617 pp.

Commeau N, (2012). Modélisation de la contamination par *Listeria monocytogenes* pour l'amélioration de la surveillance dans les industries agro-alimentaires. PhD thesis, AgroParisTech. http://tel.archives-ouvertes.fr/index.php?alsid=3e uus0iet62q4s4jjf39p0e935&view_this_doc=pastel-00770790&version=1

Gonzales-Barron U, Butler F. 2011. Characterisation of within-batch and between-batch variability in microbial counts in foods using Poisson-gamma and Poisson-lognormal regression models. *Food Control*, 22 :1268–1278.

International Commission on Microbiological Specifications for Foods (ICMSF). 2002. Microorganisms in Foods 7 : microbiological testing in food safety management. *New-York, Kluwer academic/Plenum Publisher :* 375 pp.

International Life Science Institute (ILSI). 2010. Impact of microbial distribution on food safety. 1-68.

Parent E, Bernier J. 2007. Le raisonnement bayésien. *Springer Verlag, France :* 327 pp.

Règlement (CE) N°2073/2005 de la Commission européenne concernant les critères microbiologiques applicables aux denrées alimentaires. *Journal officiel de l'Union européenne*, L338 :1–26.

Robert C. 2006. Le choix bayésien. *Springer-Verlag, France* : 639 pp.

Thomas, A., O'Hara, B., Ligges, U. and Sturtz, S. (2006). Making BUGS open. *R News*, 6 :12–17.