Nécessité d'évaluer l'exposition à *L. monocytogenes* et de caractériser les risques de listériose par consommation de salade de IV^{ème} gamme

Les éléments précédents montrent qu'un nombre important de travaux de recherche ont été réalisés dans lesquels la pathogénécité de *L. monocytogenes* ainsi que sa présence dans les végétaux ont été mis en évidence. *L. monocytogenes* représente donc un danger pour l'ensemble de la filière IV^{ème} gamme. Pourtant, aucune étude n'a été réalisée en France ou en Europe afin de déterminer les concentrations en *L. monocytogenes* à l'instant de la consommation (exposition) ainsi que le risque de tomber malade par ingestion de *Listeria monocytognes* suite à la consommation de produits de IV^{ème} gamme. Il est donc nécessaire de mettre en place une méthodologie d'évaluation de l'exposition et des risques pour le couple *L. monocytogenes*/salades de IV^{ème} gamme.

Dans le cadre formel de l'analyse de risque, l'évaluation de l'exposition se définit comme l'évaluation qualitative et/ou quantitative de l'ingestion d'agents biologiques, chimiques et physiques par le biais des aliments, ainsi que le cas échéant, par suite de l'exposition à d'autres sources. Cette étape consiste à évaluer les doses de contaminants ingérés à l'instant de la consommation. Les doses peuvent être évaluées soit par analyse directe des aliments consommés (études d'alimentation dupliquée) mais ces études sont rares (Leblanc et al., 2000, 2005), soit en croisant la contamination à l'instant de la consommation avec les quantités consommées en utilisant des outils statistiques plus ou moins complexes (Jaykus, 1996; Kroes et al., 2002). La caractérisation des risques s'effectue en intégrant les trois étapes précédentes (l'identification et la caractérisation du danger ainsi que l'évaluation de l'exposition) afin d'estimer qualitativement et/ou quantitativement, compte tenu des incertitudes inhérentes à l'évaluation, la probabilité de la fréquence et de la gravité des effets adverses connus ou potentiels sur la santé susceptibles de se produire dans une population donnée (Renwick et al., 2003). Ainsi, le risque est déterminé comme la probabilité de dépasser la dose tolérable ou comme la probabilité de tomber malade lors d'un acte de consommation dans le cas d'un

risque aigu.

L'évaluation de l'exposition et la caractérisation des risques pour le couple *L. monocytogenes*/salades de IV^{ème} gamme présentent certaines particularités qui nécessitent la construction d'outils spécifiques. Tout d'abord, les bactéries peuvent se développer lorsque les conditions environnementales leur sont favorables, contrairement à la plupart des contaminants chimiques présents en quantité constante au cours du temps dans les aliments. Par ailleurs, *Listeria monocytogenes* a la faculté de se multiplier même à de faibles températures. C'est pourquoi il est nécessaire dans les analyses de risques relatifs à des dangers microbiologiques d'évaluer les concentrations bactériennes tout au long de la chaîne alimentaire et d'intégrer alors le processus de croissance des bactéries. Ce type d'évaluation a un intérêt majeur pour les gestionnaires du risque (cf. le deuxième volet de l'analyse du risque) qui peuvent ainsi proposer des mesures de gestions à différent niveaux de la chaîne alimentaire et non pas seulement à l'instant de la consommation. Les bactéries et notamment les bactéries pathogènes (néfastes pour l'hôte) comme *Listeria monocytogenes* sont difficilement détectables et quantifiables. De plus, les méthodes analytiques de dénombrements des bactéries sont très coûteuses en terme de temps. Ainsi les données quantifiant le nombre de bactéries par unité d'analyse (g, ml, etc...) sont très peu nombreuses. Notons aussi qu'en comparaison avec certains dangers toxiques ou d'autres couples bactéries pathogènes/aliment, il existe peu de plans de surveillance nationaux de *L. monocytogenes* dans les végétaux et en particulier dans les végétaux frais non transformés. Ainsi les données disponibles sont pour la plupart des données d'origine bibliographique ou de centres techniques et leur recueil est une barrière supplémentaire à la connaissance des concentrations bactériennes. Il faut alors développer des outils permettant d'estimer les concentrations à l'aide des seules données disponibles. De plus, lorsque les concentrations en *L. monocytogenes* dans la salade ont pu être relevées, il s'avère que celles-ci sont très variables d'une analyse à l'autre. Le modèle d'évaluation de l'exposition et des risques proposé doit prendre en compte la variabilité propre aux concentrations microbiennes ainsi que l'incertitude des estimations due au manque d'information et de données. Le modèle doit aussi permettre de propager séparément l'incertitude totale (variabilité et incertitude) lors de la simulation de la croissance/décroissance des bactéries pendant les différentes étapes composant la chaîne alimentaire.

1.3 Enjeux scientifiques

1.3.1 Construire un modèle d'évaluation de l'exposition à *L. monocytogenes* et de caractérisation des risques de listériose par consommation de salade de IV^{ème} gamme

Nous proposons de construire un modèle d'évaluation de l'exposition et de caractérisation des risques pour le couple *L. monocytogenes*/salade de IV^{ème} gamme. Ce modèle est composée de plusieurs variables (ou paramètres) d'entrée et de sortie qui sont présentées dans la figure 1.2. L'objective est d'évaluer les variables de sortie qui sont les contaminations aux différentes étapes de la chaîne alimentaire, l'exposition des consommateurs, le risque de listériose et le nombre de cas. Pour cela, il est nécessaire d'introduire dans le modèle des variables d'entrée, comme la contamination de la matière première entrant dans l'usine, les paramètres de croissance, les températures et les durées de conservation du produit pour les différentes étapes constituant son circuit logistique (non détaillées dans la figure 1.2), la consommation de salade de IV^{ème} gamme, etc. Nous choisissons de réaliser l'évaluation de l'exposition et des risques en deux temps. Dans un premier temps, les distributions des variables d'entrée sont déterminées en utilisant des modèles développés afin de rendre compte de la complexité du comportement des micro-organismes (contamination, croissance, etc). Dans un second temps, les distributions des variables de sortie sont déterminées à l'aide de méthodes de simulation ayant en entrée les distributions des variables établies précédemment. Il s'agit alors de prendre en compte dans l'estimation des variables d'entrée la variabilité et l'incertitude et de les propager séparément au cours des simulations afin d'intégrer ces deux dimensions dans l'évaluation de l'exposition et des risques. Le modèle développé doit aussi permettre d'évaluer les conséquences en terme de concentrations bactériennes, d'exposition et de risque de scénarios comme le retrait du chlore du procédé de fabrication des salades de IV^{ème} gamme ou de scénarios futurs de gestion du risque pouvant être une meilleure gestion de la chaîne du froid. Ce modèle doit aussi être suffisamment générique pour être transposable aux différents produits de la filière IV^{ème} gamme et à d'autres filières alimentaires.

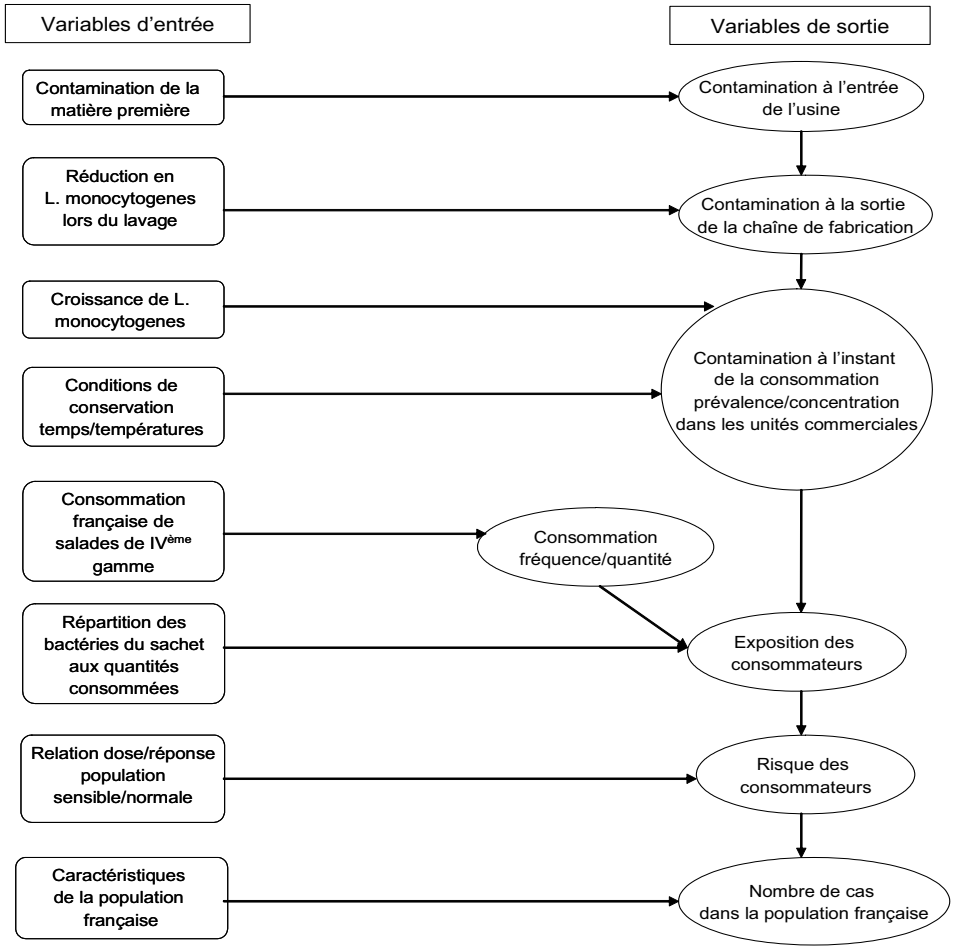

FIG. 1.2: Principaux composants du modèle d'évaluation de l'exposition et des risques lié à la présence de *L. monocytogenes* dans les salades de IV^{ème} gamme

1.3.2 Intégrer dans l'évaluation des risques alimentaires des méthodes statistiques permettant de prendre en compte séparément la variabilité et l'incertitude

La variabilité représente la diversité et l'hétérogénéité de la réponse dans la population étudiée. Elle est intrinsèque à cette population et ne peut être réduite par un complément d'information à la différence de l'incertitude qui représente l'ignorance partielle ou le manque de connaissances de l'évaluateur et qui peut être réduite par des données ou des études supplémentaires. L'incertitude totale est constituée de la variabilité et de l'incertitude. Ainsi tout phénomène biologique est intrinsèquement variable et la connaissance de ce phénomène ne peut être que partielle et cela même dans le cas d'information abondante. La statistique est la science adaptée à l'étude de la variabilité et de l'incertitude. Ainsi le Dr Daniel Schwartz 1994 déclare que "la variabilité est la raison d'être de la statistique" (Schwartz, 1994) ou encore Claudine Vergne dans son postulat pour les nouveaux programmes de statistique au lycée écrit "Prendre conscience de la variabilité d'un phénomène, ce n'est pas seulement constater que les résultats sont sujets à variation, c'est concevoir que, à notre échelle d'observation, les résultats sont nécessairement variables et imprévisibles, c'est accepter de prendre en compte les fluctuations, c'est faire le deuil de la certitude et s'engager dans le monde de l'incertain. Renonçant à des connaissances assurées, on peut alors, par des méthodes statistiques, en suivant des raisonnements de type inductif, accéder à une maîtrise relative de l'incertitude pour estimer, prévoir et prendre des décisions avec risque consenti". Il est alors indispensable lors de la réalisation d'analyse des risques alimentaires d'intégrer une mesure de l'incertitude totale. La commission du codex alimentarius recommande d'ailleurs de prendre en compte la variabilité du risque et d'évaluer les incertitudes autour des estimations du risque (Codex Alimentarius Commission, 2003).

Le moyen de prendre en compte la variabilité d'un phénomène est d'opter pour l'approche probabiliste (ou distributionnelle) qui est l'utilisation de distributions de probabilité en opposition à l'approche déterministe (ou ponctuelle) qui caractérise le phénomène étudié par une seule valeur qui est en général la moyenne. Bien que l'approche probabiliste soit de plus en plus employée lors d'évaluation des risques, peu de travaux prennent en compte tout en les distinguant les deux dimensions de l'incertitude totale (variabilité et incertitude). La prise en compte séparée de la variabilité et de l'incertitude dans l'évaluation des risques alimentaires est l'un des enjeux de cette thèse. Les méthodes statistiques permettant de mener à bien ce travail sont présentées dans cette partie.

L'approche bayésienne : une méthode statistique performante pour l'évaluation des risques et la prise en compte de la variabilité et de l'incertitude

Le paradigme bayésien Considérons un modèle statistique où la loi de probabilité $p(x|\theta)$ qui génère les observations x est donnée par un modèle paramétrique particulier qui dépend d'un paramètre inconnu de dimension $k : \theta \in \Theta \subset \mathbb{R}^k$. Dans l'approche classique de la statistique inférentielle, le paramètre θ, s'il est par nature inconnu, reste néanmoins fixé dans Θ. L'approche bayésienne considère toujours θ inconnu mais ce paramètre est, comme l'observation x, aléatoire (c'est-à-dire issu d'un phénomène aléatoire). Dans ce contexte, le paramètre θ est régi par une certaine loi de probabilité $p(\theta)$, dite loi *a priori* (supposée connue). Cette loi est subjective dans la mesure où elle représente la croyance de l'expérimentateur avant que l'expérience ne soit conduite. L'analyse bayésienne se fait en termes de lois régissant θ, l'inférence sur ce paramètre inconnu se réalisant au travers de ces lois. Plus spécifiquement, elle tire parti de l'observation x pour réactualiser la loi *a priori* $p(\theta)$: on construit sur la base de la loi *a priori* $p(\theta)$ et des observations x, une loi $p(\theta|x)$ qui régit θ, dite loi *a posteriori* (car déterminée après avoir observé x). Le principe est donc de corriger l'*a priori* que l'on suppose sur θ par l'information apportée par x. Une telle modélisation n'est pas universellement acceptée, le choix de la loi *a priori* n'obéissant pas nécessairement à des critères objectifs. Cependant, il est des cas où l'on dispose de résultats d'expériences permettant de conjecturer que le paramètre θ se trouve avec une forte probabilité dans une région déterminée de l'espace des paramètres. On peut ainsi rendre compte de cette information au travers d'une loi *a priori* se concentrant dans cette région. A l'inverse, lorsque aucune information n'est disponible sur le paramètre θ, cela peut être traduit par l'utilisation de lois *a priori* non informatives ou vagues (c'est-à-dire avec une variance élevée). Notons toutefois que le support de la loi *a posteriori* est inclue dans celui de la loi *a priori*. L'approche bayésienne est formalisée dans la figure 1.3.

D'un point de vue pratique, les calculs nécessaires pour déterminer la distribution *a posteriori* sont souvent complexes voire insolubles. En effet, la résolution analytique de l'intégrale (1.3) de la figure 1.3 est particulièrement difficile lorsque la loi $p(\theta)$ n'est pas conjuguée pour le modèle de vraisemblance $p(x|\theta)$, c'est-à-dire que les lois $p(\theta)$ et $p(\theta|x)$ n'appartiennent pas à la même famille de lois, ou lorsque le modèle présente une structure hiérarchique (Fig. 1.4), c'est-à-dire que les lois *a priori* sont découpées en plusieurs niveaux, les paramètres (appelés aussi hyperparamètres) des derniers niveaux possédant également des lois de probabilité. Le découpage hiérarchique donne donc lieu à une explosion de lois marginales et de lois *a posteriori*, suivant le niveau par rapport auquel on conditionne.

La distribution jointe de (x, θ) s'obtient simplement par

$$p(x, \theta) = p(x|\theta)p(\theta). \tag{1.1}$$

La formule de Bayes est basée sur la décomposition inverse de (1.1)

$$p(x, \theta) = p(\theta|x)p(x),$$

on obtient alors la distribution *a posteriori* de θ conditionnellement à l'observation de l'échantillon x

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}, \tag{1.2}$$

où le dénominateur $p(x)$ est la distribution marginale de x et ne dépend pas de θ. On a

$$p(x) = \int_{\Theta} p(x|\theta)p(\theta)d\theta. \tag{1.3}$$

La forme simplifiée de (1.2) s' écrit

$$p(\theta|x) \propto p(x|\theta)p(\theta). \tag{1.4}$$

où $\propto$ signifie "proportionnel à", la constante d'intégration étant donnée en (1.3)
Il est utile de pouvoir générer de futures valeurs de x, appelées aussi valeurs prédites et notées $\widetilde{x}$, étant donnée l'information récoltée sur θ grâce aux observations x. On définit alors la distribution prédictive *a posteriori* suivante

$$p(\widetilde{x}|x) = \int_{\Theta} p(\widetilde{x}|\theta)p(\theta|x)d\theta.$$

FIG. 1.3: Formalisme bayésien

La loi de probabilité $p(x|\theta)$ qui génère les observations x est donnée par un modèle paramétrique particulier qui dépend d'un paramètre inconnu de dimension k : $\theta \in \Theta \subset \mathbb{R}^k$. La loi de probabilité $p(\theta|\psi)$ du paramètre θ est elle aussi décrite par un modèle paramétrique qui dépend d'un paramètre (hyperparamètre) inconnu de dimension l : $\psi \in \Psi \subset \mathbb{R}^l$. L'hyperparamètre ψ est alors régi par la loi *a priori* $p(\psi)$. La distribution *a posteriori* de ψ conditionnellement à l'observation de l'échantillon x s'écrit alors :

$$p(\psi|x) \propto \int_{\Theta} p(x|\theta)p(\theta|\psi)p(\psi)d\theta. \tag{1.5}$$

On peut aussi obtenir la distribution *a posteriori* de θ

$$p(\theta|x) \propto \int_{\Psi} p(\theta|\psi)p(\psi|x)d\psi.$$

La distribution prédictive *a posteriori* dans ce cas est

$$p(\widetilde{x}|x) = \int_{\Theta}\int_{\Psi} p(\widetilde{x}|\theta,\psi)p(\theta,\psi|x)d\psi d\theta.$$

FIG. 1.4: Formalisme bayésien pour un modèle hiérarchique

Ainsi, dans le cas où les calculs sont complexes, le recours à des intégrations numériques est nécessaire pour approcher les distributions *a posteriori*. Pour cela, de nombreux algorithmes fondés sur les méthodes de Monte Carlo par Chaîne de Markov (MCMC) ont été développés et implémentés dans des logiciels, comme par exemple OpenBugs (variante de Winbugs), cf. Spiegelhalter et al. (2005). Le principe des algorithmes de MCMC est d'approcher les distributions *a posteriori* des paramètres (et/ou des hyperparamètres dans le cas d'un modèle hiérarchique), comme expliqué dans la figure 1.5. Les deux algorithmes de MCMC les plus utilisés sont l'algorithme de Hasting-Metropolis (Metropolis et al., 1953; Hastings, 1970) et l'échantillonnage de Gibbs (Geman & Geman, 1984; Gelfand & Smith, 1990). Le logiciel OpenBugs utilise principalement l'algorithme de Gibbs (ou une version hybride). Un exemple de cet algorithme est donné dans la figure 1.6. Pour un approfondissement des connaissances sur les méthodes bayésiennes et les méthodes MCMC, on peut par exemple se reporter aux ouvrages de Robert (2001); Gelman et al. (2004a) et Robert (1996).

Les méthodes MCMC permettent de simuler un échantillon $x_1, ..., x_n$ sans avoir recours à la distribution f (ici f peut-être la distribution *a posteriori* de l'équation (1.5)) et ainsi d'approcher l'intégrale $E_f[h(x)] = \int h(x)f(x)dx$. Le principe est d'utiliser une chaîne de Markov de loi stationnaire f : partant d'une valeur arbitraire $x^{(0)}$, une chaîne $x^{(t)}$ est générée à partir d'un noyau de transition k de loi stationnaire f, qui garantit la convergence en loi vers f. Pour T "assez grand", on peut considérer $x^{(T)}$ comme distribué suivant f et obtenir ainsi un échantillon $x^{(T)}, x^{(T+1)}, ...$, qui est effectivement distribué suivant f, même si les $(x^{(T+t)})_{t\geq 1}$ ne sont pas indépendants. Toute méthode produisant une chaîne de Markov $(x^{(t)})$ de loi stationnaire la distribution d'intérêt est appelée algorithme MCMC. L'utilisation d'une chaîne $(x^{(t)})$ produite par un algorithme MCMC est semblable à celle d'un échantillon indépendant et identiquement distribué suivant f au sens où le théorème ergodique (appellation de la loi des grands nombres dans ce cadre) garantit la convergence de la moyenne empirique vers son espérance théorique, c'est-à-dire

$$\frac{1}{B}\sum_{t=T}^{T+B} h(x^{(t)}) \underset{B\to\infty}{\to} E_f[h(x)].$$

La convergence vers la loi stationnaire est donc atteinte à partir d'un certain nombre T d'itérations. Ce nombre est à déterminer à l'aide des outils de contrôle de la convergence. La convergence peut être contrôlée graphiquement en identifiant le nombre d'itérations devant être réalisées afin que les valeurs produites, à partir de plusieurs chaînes ayant des valeurs initiales distinctes, se rejoignent. Une fois la convergence vers la loi stationnaire atteinte, il est possible de simuler B valeurs. En général, l'inférence bayésienne est fondée sur l'estimation des distributions marginales *a posteriori* des paramètres et des hyperparamètres. A partir des B dernières réalisations de la chaîne de Markov sont déduits, pour chaque paramètre et hyperparamètre, la moyenne, l'écart-type et certains percentiles ($2.5^{ème}$, $50^{ème}$, $97.5^{ème}$, etc). La loi jointe *a posteriori* des paramètres, formée par les B valeurs obtenues, peut également être utilisée pour simuler des valeurs de paramètres d'entrée d'un modèle de simulations de Monte Carlo.

FIG. 1.5: Principe des méthodes de Monte Carlo par Chaîne de Markov

L'algorithme de Gibbs pour approcher la loi jointe *a posteriori* $p(\theta, \psi|x)$ d'un modèle hiérarchique (Fig. 1.4) où x est le vecteur des observations, $\theta = (\theta_1, ..., \theta_i, ..., \theta_k)$ le vecteur des paramètres et $\psi = (\psi_1, ..., \psi_h, ..., \psi_l)$ le vecteur des hyperparamètres s'effectue comme suivant :

- Partir de valeurs initiales arbitraires $\psi^{(0)} = (\psi_1^{(0)}, ..., \psi_h^{(0)}, ..., \psi_l^{(0)})$ et générer des valeurs initiales $\theta^{(0)} = (\theta_1^{(0)}, ..\theta_i^{(0)}, ..\theta_k^{(0)})$ à partir de la distribution conditionnelle $p(\theta|\psi^{(0)}, x)$.
- Pour $t = 1, ..., T + B$, générer chaque $\theta_i^{(t)}$ et chaque $\psi_h^{(t)}$ en utilisant respectivement les distributions conditionnelles suivantes $p(\theta_i|\theta_{-i}^{(t-1)}, \psi^{(t-1)}, x)$ et $p(\psi_h|\theta^{(t)}, \psi_{-h}^{(t-1)}, x)$, où $\theta_{-i}^{(t-1)} = (\theta_1^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, ..., \theta_k^{(t-1)})$ et $\psi_{-h}^{(t-1)} = (\psi_1^{(t)}, ..., \psi_{h-1}^{(t)}, \psi_{h+1}^{(t-1)}, ..., \psi_l^{(t-1)})$ sont respectivement les vecteurs en θ sauf la $i^{ème}$ composante et en ψ sauf la $h^{ème}$ composante. Les valeurs de ces vecteurs correspondent aux valeurs prises à l'instant t ou $t-1$ selon si la mise à jour a été effectuée ou non.

FIG. 1.6: Algorithme de Gibbs

Modèles bayésiens hiérarchiques et prise en compte de la variabilité et de l'incertitude L'inférence bayésienne permet de quantifier l'incertitude sur les paramètres des modèles. Notons qu'en statistique classique, l'incertitude peut se quantifier en ayant recours au bootstrap (Efron & Tibshirani, 1993) cependant, ce type de problème est plus facile à appréhender par une approche bayésienne dans le sens où tous les paramètres (et hyperparamètres) sont traités comme des variables aléatoires. Dans le cas d'un modèle bayésien hiérarchique, il est possible d'interpréter la distribution de vraisemblance ainsi que la distribution des paramètres en terme de variabilité et la distribution des hyperparamètres en terme d'incertitude (Pouillot, 2006). De plus, l'effet de l'incertitude globale sur l'ensemble des hyperparamètres inconnus est prise en compte par la loi jointe *a posteriori* des hyperparamètres fournie par l'inférence bayésienne, contrairement aux méthodes de statistique classique où la relation de dépendance entre les paramètres n'est souvent pas considérée.

Méthodes bayésiennes et évaluation des risques microbiologiques Les méthodes bayésiennes sont de plus en plus employées dans l'évaluation des risques pour leur capacité à résoudre de nombreux problèmes propres à ce domaine. Les exemples sont nombreux pour les évaluations liées à des risques microbiologiques (Pouillot et al., 2007; Ranta et al., 2005; Barker et al., 2005). Certains auteurs comme Paulo et al. (2005) les ont aussi employées pour l'évaluation des risques liés aux pesticides. Enfin, ces méthodes sont utilisées dans d'autres domaines comme la science de l'eau (cf. par exemple la thèse de Pasanisi, 2004), l'épidémio-

logie avec par exemple les travaux de Richardson & Gilks (1993), ou encore l'écologie (cf. par exemple les travaux de Rivot et al., 2004). Nous relevons ci-dessous les différentes problématiques que pose l'évaluation des risques microbiologiques et pour lesquelles les méthodes bayésiennes peuvent apporter des solutions.

L'évaluation des risques microbiologiques nécessite l'intégration d'informations externes afin de compléter l'information apportée par les seules données disponibles. En effet, les données indispensables à la quantification de l'exposition et des risques microbiologiques sont souvent incomplètes, peu adaptées au problème ou peu représentatives du phénomène étudié. Il est alors nécessaire d'intégrer de l'information supplémentaire. Cette information peut être issue de l'opinion personnelle du statisticien, de techniques d'élicitation auprès d'experts (Kadane & Wolfson, 1998; O'Hagan & Oakley, 2007) ou d'études antérieures. L'approche bayésienne permet d'intégrer de l'information externe par l'utilisation de distributions de probabilité *a priori* sur le paramètre inconnu θ (ou ψ dans le cas d'un modèle hiérarchique). Notons que l'indépendance entre l'information *a priori* et les données étudiées est nécessaire afin que l'information ne soit pas redondante. L'utilisation de distribution *a priori* permet également de travailler sur des modèles éventuellement non identifiables. En effet, l'appui sur les distributions *a priori* permet de procéder à l'inférence même si les données utilisées ne permettent pas d'estimer certains paramètres du modèle. Dans ce cas, les distributions *a posteriori* sont calculées uniquement à partir des distributions *a priori*. Pour donner un exemple, dans cette thèse, les données rassemblées pour estimer les paramètres de croissance de *L. monocytogenes* dans la salade ne couvraient pas l'ensemble des températures auxquelles *L. monocytogenes* est susceptible de se développer. Ainsi, les températures optimales et maximales de croissance qui sont des paramètres de ce modèle n'étaient pas identifiables avec ces seules données. L'utilisation de distributions établies à partir d'une étude antérieure réalisée sur un autre aliment a permis de passer outre ce problème. Cela a toutefois pu être réalisé car les paramètres en question sont des paramètres propres à la bactérie étudiée et ne varient que très peu selon l'aliment. Il est donc conseillé de bien choisir les distributions *a priori* avant d'avoir recours à ce genre de pratique.

Les modèles de microbiologie prévisionnelle modélisant la croissance ou encore celui développé dans cette thèse afin d'évaluer la distribution de la contamination des végétaux par *L. monocytogenes* sont des modèles complexes c'est-à-dire comportant plusieurs niveaux hiérarchiques et un nombre important de paramètres. Le calcul analytique des distributions *a posteriori* est alors irréalisable et le recours à des intégrations numériques est obligatoire. L'inférence bayésienne a donné naissance aux algorithmes de MCMC (Monte Carlo par

Chaînes de Markov) très performants pour approcher les distributions *a posteriori* de ces modèles complexes.

Les méthodes de simulation de Monte-Carlo de second ordre : propagation séparée de la variabilité et de l'incertitude

Lorsque le modèle est complexe, l'utilisation de distributions de probabilité pour un grand nombre de variables d'entrée rend le calcul analytique des distributions des variables de sortie insoluble. C'est pourquoi les calculs sont menés numériquement par simulations de Monte-Carlo. Les méthodes de simulation de Monte-Carlo sont largement utilisées en évaluation des risques alimentaires (Nauta, 2000). En effet, ces méthodes permettent de travailler avec des modèles stochastiques (approche probabiliste) et de propager le hasard (la variabilité et l'incertitude) tout au long de la chaîne alimentaire. Les simulations de Monte-Carlo nécessitent un grand nombre d'itérations. A chaque itération, une valeur est tirée aléatoirement dans chaque distribution de probabilité des paramètres d'entrée du modèle. Chaque itération représente une situation réelle potentielle et une distribution de probabilité est obtenue pour chacune des variables de sortie.

Il est important et conseillé d'évaluer séparément la variabilité des sorties et l'incertitude des estimations produites. L'intérêt de séparer la variabilité de l'incertitude est de pouvoir déterminer laquelle des deux prédomine afin d'analyser les sorties en fonction de cette information. La méthode de Monte-Carlo à deux dimensions également appelée méthode de Monte-Carlo de second ordre (Nauta, 2000; Vose, 2000; Pouillot et al., 2007) permet de transférer séparément la variabilité et l'incertitude des paramètres dans le modèle. Le principe est de dissocier, lors de la simulation de Monte-Carlo, la simulation des paramètres incertains des paramètres variables, c'est-à-dire de réaliser deux simulations de Monte-Carlo imbriquées : une modélisation de la variabilité imbriquée dans une modélisation de l'incertitude (Fig. 1.7). Pour ce faire, il faut au préalable déterminer ce que représente les distributions utilisées en entrée du modèle : la variabilité du paramètre ou l'incertitude autour de l'estimation du paramètre. Il n'est pas toujours évident de qualifier de variabilité ou d'incertitude chaque distribution, cependant nous avons vu que l'utilisation de modèles bayésiens hiérarchiques facilite cette interprétation.

La simulation de Monte-Carlo à deux dimensions consiste à tirer aléatoirement un nombre n_u de valeurs dans chacune des distributions $p(\psi)$ des (hyper)paramètres, $\psi \in \mathbb{R}^l$, reflétant l'incertitude (voir le schéma illustratif Fig. 1.8). Notons que si la distribution de ψ a été établie par inférence bayésienne alors les n_u valeurs sont tirées dans la loi jointe *a posteriori* $p(\psi|x)$ de ψ. Puis, pour chaque valeur fixe des paramètres d'incertitude, un nombre n_v de valeurs est tiré aléatoirement dans chacune des distributions $p(\theta|\psi_u)$ des paramètres $\theta \in \mathbb{R}^k$ reflétant la variabilité (ou dans la distribution *a posteriori* $p(\theta|\psi_u, x)$ dans le cas où celle-ci a été établie par inférence bayésienne). Ces opérations sont effectuées pour chacune des variables d'entrée du modèle de simulation. Ensuite, par simulation de Monte-Carlo, pour chaque valeur fixée d'incertitude, une distribution empirique constituée des n_v valeurs est obtenue pour chaque variable de sortie (par exemple : concentrations, exposition, risque...). Cette distribution reflète la variabilité de chacune des variables de sortie et différentes statistiques de chaque variable sont calculées comme la moyenne, l'écart-type ou différents percentiles. Ainsi un nombre n_u de valeurs est obtenu pour chaque statistique de chaque variable et les percentiles de ces statistiques sont calculés. Le $50^{ème}$ percentile est proposé comme estimateur des différentes statistiques calculées, le $2.5^{ème}$ et le $97.5^{ème}$ percentiles sont proposés comme bornes de l'intervalle d'incertitude (UI, Uncertainty Interval) des estimations. La comparaison entre les intervalles d'incertitude des estimations et les percentiles représentant la variabilité du paramètre de sortie étudié permet de voir laquelle de la variabilité et de l'incertitude prédomine.

FIG. 1.7: Simulation de Monte-Carlo de second ordre

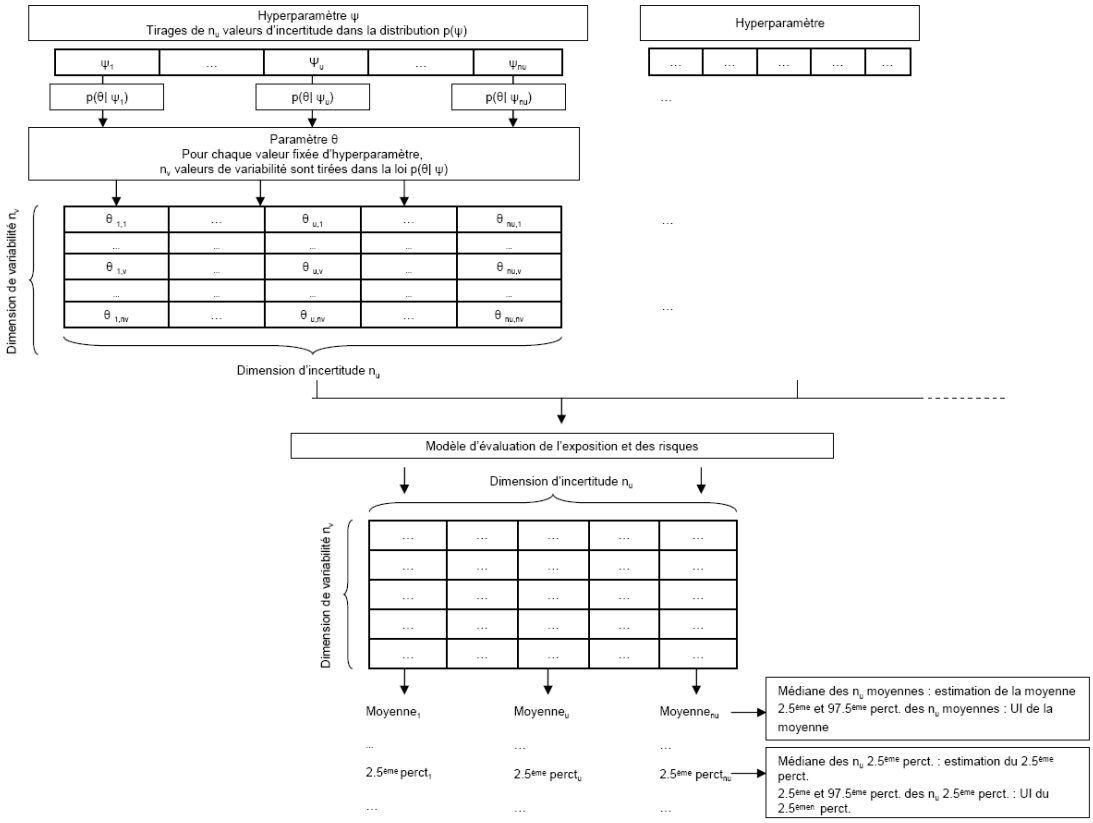

FIG. 1.8: Marche de la simulation de Monte-Carlo à deux dimensions

1.4 Organisation du manuscript

Cette thèse s'articule autour de six chapitres, le premier étant celui-ci et le dernier est une discussion générale sur les travaux présentés. Chacun des quatre chapitres intermédiaires est composé d'une introduction, d'une partie méthodologique, d'une description des données utilisées, d'une partie résultats et d'une discussion. Les chapitres 2 et 3 sont consacrés aux développements de modèles spécifiques afin de déterminer la distribution de la contamination microbiologique d'un aliment et les distributions des paramètres de croissance. Les deux chapitres suivants (chapitres 4 et 5) portent sur la construction du modèle de simulation des risques liés à la présence de *Listeria monocytogenes* dans les salades de IV^{ème} gamme.

Dans le chapitre 2, je me suis intéressée à un des problèmes majeurs de l'évaluation quantitative des risques microbiologiques qui est le manque de données de contamination adaptées. En effet, les données quantitatives de contamination microbiologique (les concentrations) sont peu nombreuses, et en particulier pour *L. monocytogenes* qui est généralement présente à de très faibles concentrations ce qui rend sa détection et sa quantification difficile. Ainsi, un modèle permettant d'estimer la distribution de contamination microbiologique d'aliments à l'aide des deux types de données disponibles (prévalence et concentration) a été développé. Ce modèle permet aussi de quantifier les concentrations inférieures à la limite de détection des méthodes analytiques qui sont rarement prises en compte dans les évaluations des risques alors que de faibles concentrations peuvent atteindre des valeurs élevées lorsque le produit est conservé assez longtemps dans des conditions favorables à la croissance. Ce chapitre, en proposant cette modélisation, présente un intérêt au niveau de l'évaluation des risques mais aussi au niveau statistique. En effet, on montre comment il est possible d'estimer les paramètres d'un mélange de deux lois normales lorsqu'une partie des données est absente mais que l'on dispose d'un estimateur de l'espérance conditionnelle. Les méthodes d'inférence bayésienne sont utilisées afin d'obtenir les distributions *a posteriori* des paramètres d'intérêt. Ce modèle est appliqué à *L. monocytogenes* dans les légumes toutes catégories confondues (transformés et non transformés) et de toutes origines géographiques afin d'avoir le plus de données possible.

Le troisième chapitre présente trois modèles permettant d'estimer les distributions des paramètres de croissance de *L. monocytogenes* dans les salades feuilles à partir de courbes de croissance obtenues par expérimentation. Ces modèles diffèrent légèrement au niveau de leur structure hiérarchique mais leur construction générale repose sur la méthodologie développée par Pouillot et al. (2003). Cette méthodologie consiste à estimer au sein d'un même

processus les paramètres des modèles de croissance primaire et secondaire de microbiologie prévisionnelle. L'estimation des paramètres s'effectue également par inférence bayésienne ce qui permet d'intégrer de l'information *a priori* issue de travaux antérieurs et indispensable pour estimer l'ensemble des paramètres. Une méthode permettant de valider les modèles proposés et de sélectionner le ou les plus performant(s) est développée.

Le quatrième chapitre est consacré à la modélisation de l'évolution des concentrations en *L. monocytogenes* sur l'ensemble de la chaîne alimentaire et tout au long de la vie du produit. La modélisation s'effectue en trois étapes :

1. l'évaluation, à l'aide du modèle développé dans le second chapitre, de la contamination de la matière première par *L. monocytogenes*.
2. l'évolution des concentrations pendant le processus de transformation de la matière première en produit de IV^{ème} gamme. La principale cause de variation des concentrations en *L. monocytogenes* étant le lavage des feuilles de salades à l'eau chlorée, cette étape est résumée par la modélisation de la réduction en micro-organismes pendant le lavage.
3. l'évolution des concentrations lors du stockage des produits de leur sortie de la chaîne de fabrication jusqu'à leur consommation. Cette étape intègre les résultats obtenus dans le chapitre précédent sur les distributions des paramètres de croissance. Les distributions des durées et des températures de stockage sont estimées à partir de diverses sources d'information comme des dires d'experts, des données bibliographiques ou des rapports d'étude.

Ainsi le pourcentage de sachets contenant des salades contaminées et la distribution de leur concentration à l'instant de leur consommation sont établis à l'aide de simulations de Monte Carlo intégrant les étapes précédentes. Différents scénarios fondés sur des hypothèses méthodologiques comme le choix du modèle de croissance ou sur des hypothèses liées au procédé de fabrication comme le retrait du chlore de l'eau de lavage seront testés afin d'évaluer les effets sur les concentrations bactériennes. Le traitement de l'eau de lavage par le chlore est effectué par les industriels afin d'améliorer la qualité microbiologique des aliments bien qu'aucune étude n'ait montré que le chlore ait un effet virucide. Son utilisation a une mauvaise image auprès des consommateurs même si celle-ci ne présente pas de risque sanitaire pour le consommateur (Afssa Saisine n°2003-SA-0015) dans le cas où les doses ajoutées à l'eau de lavage sont respectées[1]. Ainsi certains pays comme l'Allemagne l'ont déjà banni du procédé de

[1] **Guide des bonnes pratiques d'hygiène des produits végétaux crus prêts à l'emploi, 1996**

fabrication et son utilisation tolérée en France est remise en cause en vue d'une harmonisation des pratiques des pays européens. Enfin, les résultats obtenus sont comparés avec les données de contamination de salade de IV^{ème} gamme disponibles.

Dans le cinquième chapitre, l'exposition des consommateurs à *L. monocytogenes* est calculée en croisant la contamination à l'instant de la consommation avec les quantités consommées. L'enquête de consommation individuelle INCA et le panel des ménages Sécodip répertoriant les achats de ces derniers sur une année sont les deux enquêtes permettant d'évaluer la consommation des français. Le panel d'achat des ménages SECODIP est utilisé afin de déterminer la part de marché des salades en sachet par rapport à l'ensemble des salades achetées (transformées et non transformées), information non dissociée dans l'enquête INCA. Le nombre de bactéries présentes dans les quantités consommées est estimé à partir de la distribution de contamination des sachets et en utilisant un coefficient de groupement afin de prendre en compte la répartition hétérogène des bactéries. Une simulation de la croissance directement dans les quantités consommées est réalisée afin de comparer les deux méthodes. Puis, à l'aide d'une courbe dose-réponse de référence internationale (FAO/WHO, 2004) dont le paramètre diffère selon le type de population (sensible ou normale), le risque d'être atteint de la listériose invasive par consommation de salade en sachet est calculé. Le risque est calculé pour différentes sous-populations : enfants, adultes, personnes âgées et femmes enceintes et pour l'ensemble de la population française. Puisque certaines populations sont sous ou surreprésentées dans l'enquête INCA, le calcul du risque pour l'ensemble de la population française nécessite d'utiliser des poids de redressement afin de caler la répartition des populations de l'enquête INCA sur celle de la population française. Le nombre de cas de listériose invasive par an dans la population française est également simulé et comparé avec celui déclaré à l'institut de veille sanitaire (InVS) chargé de surveiller l'incidence de cette maladie.

Il est important de souligner que ces travaux ont été réalisés à partir d'un grand nombre de données. Le recueil de données a nécessité un travail considérable de prospection, d'harmonisation, de gestion et d'actualisation de base de données. Ces données ont été extraites de la bibliographie existante avec la collaboration de Frédéric Carlin ou m'ont été fournies par la DGCCRF et par les centres Actia chargés de mener en parallèle de mes travaux de modélisation des études de terrain en collaboration avec les entreprises afin que ce travail soit constamment alimenté, autant que possible, en données pertinentes et de qualité.