

— Proposition de thèse —
— mathématiques appliquées, statistique, ML, optimisation —

Optimisation bayésienne pour l'appréciation quantitative des risques en microbiologie

Mots-clés : optimisation bayésienne, processus gaussiens, statistique & ML, simulation numérique, planification d'expériences numériques, sécurité alimentaire, appréciation quantitative des risques, microbiologie

Contexte

Cette thèse en mathématiques appliquées (optimisation, statistique et ML) s'inscrit dans le cadre du projet européen ArtiSaneFood, dont l'un des objectifs pour la France est de permettre la poursuite de la production de fromages au lait cru, qui est aujourd'hui très surveillée et potentiellement en danger en raison de futures réglementations européennes. Elle fait l'objet d'un partenariat entre l'ANSES (agence nationale de sécurité sanitaire), CentraleSupélec, le CNIEL (Centre National Interprofessionnel de l'Economie Laitière) et d'autres acteurs importants de l'économie du lait.

Dans le domaine de la microbiologie pour l'agroalimentaire, l'*appréciation quantitative des risques* (AQR) est une discipline scientifique qui analyse les risques microbiologiques. Il s'agit d'analyser les liens entre les pratiques des producteurs et des consommateurs et les risques de contamination alimentaire.

La qualité microbiologique d'un aliment est le résultat de nombreuses opérations qui interviennent le long du processus de fabrication et de conservation. Les différents ingrédients sont progressivement transformés dans des équipements avec contrôle des paramètres des procédés (températures, durées, débits, pH, etc.).

Les modèles AQR sont des modèles de type équations différentielles ordinaires en présence de conditions initiales aléatoires et/ou équations différentielles stochastiques, parfois de grande dimension, qui simulent le développement microbiologique lors des étapes de fabrication en fonction des facteurs contrôlant les procédés. A partir de ces modèles (stochastiques) et des simulations qui en résultent, des analyses statistiques sont effectuées pour mesurer l'impact des mesures de maîtrise sur le risque (analyse statistique par scénarios). Dans un contexte de prise de décision, les estimations des quantités d'intérêt doivent s'accompagner de mesures d'incertitude.

Dans le domaine de l'AQR, l'état de l'art consiste à effectuer des analyses de sensibilité aux différents facteurs contrôlant les procédés afin d'identifier parmi les *mesures de maîtrise des risques*, celles ayant le plus d'impact (Lamboni et al., 2014). Les méthodes de type *Approximate Bayesian Computation* (Duret et al., 2014) peuvent aider à déterminer des valeurs de la mesure de maîtrise pour répondre à un objectif de sécurité fixé par le gestionnaire du risque.

Objectifs

Dans ce contexte, il apparaît aujourd'hui nécessaire de franchir une étape supplémentaire pour non plus seulement analyser les risques mais également les contrôler et les optimiser et ainsi garantir avec grande probabilité la sûreté des produits. Toutefois, les modèles AQR sont coûteux en temps de calcul et l'optimisation des risques à partir de ces modèles est difficile (Rigaux et al., 2013). L'optimisation jointe de plusieurs mesures de maîtrise et ou pour plusieurs sorties du modèle AQR se heurte donc à un problème méthodologique (Guillier et al., 2016).

Les temps de calcul des simulations de risques sanitaires étant potentiellement longs, l'optimisation des stratégies de minimisation de risque implique d'utiliser des algorithmes adaptés à l'optimisation de fonctions coûteuses.

Depuis quelques années, l'utilisation dans le domaine de la simulation numérique d'algorithmes bayésiens issus de la statistique et du machine learning, et en particulier l'apparition d'algorithmes d'optimisation bayésiens (Jones et al. 1998, Villemonteix et al. 2008, Emmerich et al. 2010, Hernandez-Herbato et al. 2016...), avec notamment des travaux à CentraleSupélec (Villemonteix et al. 2008, Bect et al. 2010, Feliot et al., 2017; Stroh et al., 2017...), laissent envisager une réponse au problème de l'optimisation des risques sanitaires liés aux contaminations microbiennes. Ces algorithmes ont montré des performances intéressantes sur ce cas simple mais ne semblent toutefois pas entièrement adaptés aux futurs besoins en matière d'optimisation. En particulier :

- Pour une stratégie et un jeu de facteurs de contrôle des procédés donnés, les variations des mesures de quantification des risques autour de leur espérance peuvent être non négligeables et doivent être prises en compte. L'espérance des mesures de risque n'est d'ailleurs probablement pas le meilleur critère pour mesurer l'efficacité et la robustesse d'une stratégie. Des indicateurs statistiques plus pertinents mais aussi plus difficiles à estimer (quantiles, super-quantiles...) doivent donc être envisagés.
- L'optimisation d'un risque unique n'est pas satisfaisante si l'on considère la complexité du problème tel qu'il se pose en pratique. Les contaminations microbiennes peuvent être de nature variée et il y a donc plusieurs mesures de risque sanitaire. Le coût économique des stratégies de contrôle des risques doit également être pris en compte. Pour ces différentes raisons, des formulations multi-objectifs et/ou sous contraintes doivent être étudiées.
- Un producteur ne souhaite pas nécessairement connaître avec précision les paramètres optimaux d'une stratégie. Il sera davantage intéressé par les plages de valeurs que les paramètres peuvent prendre pour satisfaire une tolérance donnée sur les objectifs. Cette façon d'envisager l'optimisation, consistant à rechercher des régions « quasi optimales » de l'espace des solutions plutôt que la solution exacte, ne semble pas avoir été considérée dans la littérature de l'optimisation bayésienne.

Pour résumer, les objectifs de cette thèse sont triples :

- Se familiariser et comprendre les modèles AQR multi-dangers, en proposer de nouveaux si nécessaire.
- Construire des algorithmes d'optimisation multi-objectif, en présence d'évaluations stochastiques (puisque les simulateurs sont non-déterministes), et avec des limites fortes sur le nombre d'évaluations qu'il est possible de conduire.
- Établir des recommandations méthodologiques à destination de l'industrie agroalimentaires.

Profil du candidat recherché & contacts

- Titulaire d'un M2R ou d'un diplôme d'ingénieur.
- Formation en mathématiques appliquées : probabilités/statistiques, machine learning, optimisation...
- Maîtrise d'au moins un langage de programmation « scientifique » de haut niveau (type Python, Matlab/Octave, R).
- Bon niveau en anglais écrit et parlé.
- Aptitudes personnelles souhaitées : autonomie, ouverture d'esprit, sens du travail en équipe, écoute, synthèse...
- Contacts :
 - emmanuel.vazquez@centralesupelec.fr
 - julien.bect@centralesupelec.fr

Références

Bect, J., Ginsbourger D., Li L., Picheny V., Vazquez E., Sequential design of computer experiments for the estimation of a probability of failure, *Statistics and Computing* 22 (3), 773-793

Duret, S., Guillier, L., Hoang, H.-M., Flick, D., Laguerre, O., 2014. Identification of the significant factors in food safety using global sensitivity analysis and the accept-and-reject algorithm: application to the cold chain of ham. *International journal of food microbiology* 180, 39-48.

Wagner, T., Emmerich, M., Deutz, A., & Ponweiser, W. (2010, September). On expected-improvement criteria for model-based multi-objective optimization. In *International Conference on Parallel Problem Solving from Nature* (pp. 718-727). Springer, Berlin, Heidelberg.

Feliot, P., Bect, J., Vazquez, E., 2017. A Bayesian approach to constrained single-and multi-objective optimization. *Journal of Global Optimization* 67(1-2), 97-133.

Guillier, L., Duret, S., Hoang, H.-M., Flick, D., Nguyen-Thé, C., Laguerre, O., 2016. Linking food waste prevention, energy consumption and microbial food safety: the next challenge of food policy? *Current Opinion in Food Science* 12, 30-35.

Hernández-Lobato, D., Hernandez-Lobato, J., Shah, A., & Adams, R. (2016, June). Predictive entropy search for multi-objective bayesian optimization. In *International Conference on Machine Learning* (pp. 1492-1501).

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4), 455-492

Lamboni, M., Sanaa, M., Tenenhaus-Aziza, F., 2014. Sensitivity analysis for critical control points determination and uncertainty analysis to link FSO and process criteria: application to *Listeria monocytogenes* in soft cheese made from pasteurized milk. *Risk analysis* 34(4), 751-764.

Rigaux, C., Ancelet, S., Carlin, F., Nguyen-thé, C., Albert, I., 2013. Inferring an augmented Bayesian network to confront a complex quantitative microbial risk assessment model with durability studies: application to *Bacillus cereus* on a courgette purée production chain. *Risk Analysis* 33(5), 877-892.

Stroh, R., Demeyer, S., Fischer, N., Bect, J., Vazquez, E., 2017. Sequential design of experiments to estimate a probability of exceeding a threshold in a multi-fidelity stochastic simulator. arXiv preprint arXiv:1707.08384.

Villemonaix, J., Vazquez, E. Walter E., 2008, An informational approach to the global optimization of expensive-to-evaluate functions, *J. Global Optimization*, 44 (4), 509