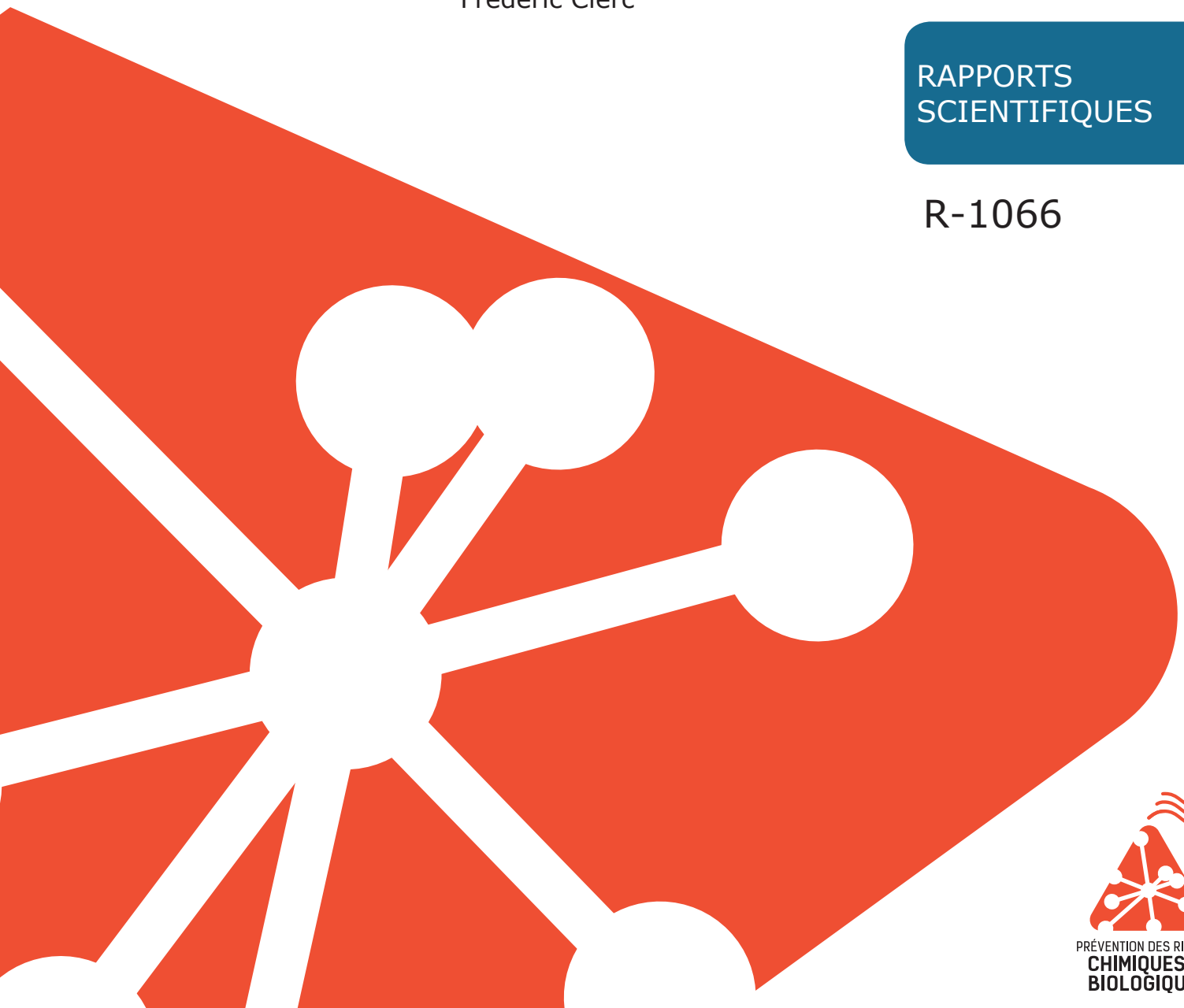
Institut de recherche
Robert-Sauvé en santé
et en sécurité du travail

WebExpo
Vers une meilleure interprétation des mesures d'exposition professionnelle aux substances chimiques sur les lieux de travail

Jérôme Lavoué
Lawrence Joseph
Tracy L. Kirkham
France Labrèche
Gautier Mater
Frédéric Clerc

RAPPORTS SCIENTIFIQUES

R-1066



PRÉVENTION DES RISQUES
CHIMIQUES ET
BIOLOGIQUES

NOS RECHERCHES travaillent pour vous !

Solidement implanté au Québec depuis 1980, l'Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST) est un organisme de recherche scientifique reconnu internationalement pour la qualité de ses travaux.

Mission

Contribuer, par la recherche, à la prévention des accidents du travail et des maladies professionnelles ainsi qu'à la réadaptation des travailleurs qui en sont victimes;

Assurer la diffusion des connaissances et jouer un rôle de référence scientifique et d'expertise;

Offrir les services de laboratoires et l'expertise nécessaires à l'action du réseau public de prévention en santé et en sécurité du travail.

Doté d'un conseil d'administration paritaire où siègent en nombre égal des représentants des employeurs et des travailleurs, l'IRSST est financé par la Commission des normes, de l'équité, de la santé et de la sécurité du travail.

Pour en savoir plus

Visitez notre site Web ! Vous y trouverez une information complète et à jour.
De plus, toutes les publications éditées par l'IRSST peuvent être téléchargées gratuitement.
www.irsst.qc.ca

Pour connaître l'actualité de la recherche menée ou financée par l'IRSST, abonnez-vous gratuitement :

- au magazine *Prévention au travail*, publié conjointement par l'Institut et la CNESST (preventionautravail.com)
- au bulletin électronique InfoIRSST

Dépôt légal
Bibliothèque et Archives nationales du Québec
2020
ISBN : 978-2-89797-084-0
ISSN : 0820-8395

IRSST - Direction des communications
et de la valorisation de la recherche
505, boul. De Maisonneuve Ouest
Montréal (Québec)
H3A 3C2
Téléphone : 514 288-1551
publications@irsst.qc.ca
www.irsst.qc.ca

Janvier 2020

WebExpo
Vers une meilleure interprétation des mesures d'exposition professionnelle aux substances chimiques sur les lieux de travail

Jérôme Lavoué[1,2], Lawrence Joseph[3], Tracy L. Kirkham[4], France Labrèche[5], Gautier Mater[6], Frédéric Clerc[6]

[1] Département de santé environnementale et santé au travail, École de santé publique, Université de Montréal
[2] Centre de recherche du CHUM
[3] Division de l'Épidémiologie clinique, Centre universitaire de santé McGill
[4] Dalla Lana School of Public Health, Université de Toronto
[5] IRSST
[6] Métrologie des polluants, INRS

RAPPORTS SCIENTIFIQUES

R-1066

Cette étude a été financée par l'IRSST. Les conclusions et recommandations sont celles des auteurs.
Cette publication a été traduite, seule la version originale (R-1065) fait foi.

ÉVALUATION PAR DES PAIRS
Conformément aux politiques de l'IRSST, les résultats des travaux de recherche publiés dans ce document ont fait l'objet d'une évaluation par des pairs.

REMERCIEMENTS

Nous tenons à souligner la contribution inestimable des trois programmeurs qui ont pris part à ce projet : Patrick Bélisle, François Lemay et Daniel Margulius. Nous aimerions également remercier les membres des comités d'experts et d'intervenants pour leur disponibilité ainsi que leurs remarques et leurs suggestions éclairées.

SOMMAIRE

Une grande partie de l'activité des hygiénistes du travail consiste à mesurer les niveaux d'exposition professionnelle des travailleurs. La plupart des rapports d'évaluation de l'exposition font état d'une importante variabilité spatiale et temporelle quant à l'intensité de l'exposition, laquelle fluctue souvent du simple au décuple en dépit de conditions apparemment similaires. Il en résulte depuis toujours un défi de taille en ce qui concerne l'interprétation des niveaux mesurés par rapport aux valeurs limites d'exposition professionnelle (VLEP). Il existe désormais un cadre consensuel – issu d'une élaboration progressive au cours des deux dernières décennies – concernant l'analyse des niveaux d'exposition par rapport aux limites d'exposition. Ce cadre repose sur la présomption que les niveaux d'exposition présentent une distribution lognormale, à tout le moins approximativement. Plusieurs paramètres de la distribution sous-jacente tenus pour être indicatifs d'un risque pour la santé sont estimés à partir d'un certain nombre de mesures et comparés à la VLEP.

Bien que ces méthodes permettent une meilleure évaluation du risque que les approches traditionnelles, elles n'ont pas été largement adoptées par les intervenants en hygiène du travail, et elles font appel à des notions statistiques généralement non abordées dans les programmes de formation habituels. Elles requièrent en outre des calculs difficilement réalisables avec des outils courants tels que calculatrices et chiffriers. Bien que des outils aient été développés en ce sens au fil des années – généralement à l'initiative de bénévoles – la plupart restent déficients à plusieurs égards, que ce soit en termes d'accessibilité, de fonctionnalité, de convivialité ou de complexité. Par ailleurs, l'incertitude relative aux estimations a surtout été prise en compte par voie de tests d'hypothèses formels ou de calculs d'intervalles de confiance, dont les résultats ne sont pas faciles à présenter aux décideurs, ce qui entrave la capacité des intervenants à communiquer efficacement les risques. Enfin, les outils disponibles sont des logiciels autonomes difficiles à intégrer à un système de gestion de données préexistant.

Le projet WebExpo visait à améliorer les pratiques actuelles en matière d'interprétation des niveaux d'exposition professionnelle grâce à la création d'une bibliothèque de solutions algorithmiques aux questions les plus fréquentes concernant l'évaluation du risque en hygiène du travail. La plupart de ces questions nécessitent l'estimation des paramètres d'une ou plusieurs distributions statistiques, et WebExpo a eu recours aux statistiques bayésiennes pour effectuer les tâches requises à cette fin. Les méthodes bayésiennes ont été retenues du fait qu'elles présentent deux grands avantages. Premièrement, elles fournissent directement des inférences probabilistes (p. ex., Quelles sont les chances que... ?), ce qui facilite la communication des risques. Deuxièmement, elles permettent d'aborder des questions méthodologiques rarement prises en compte, notamment en ce qui a trait aux valeurs signalées comme étant non détectées (un sujet de préoccupation fréquent). Les trois objectifs spécifiques de WebExpo étaient les suivants : 1) évaluer les besoins actuels en matière de calculs, de documentation et de communication des risques en lien avec l'interprétation des données d'exposition professionnelle, 2) créer une bibliothèque de codes de programmation informatique s'appuyant sur les statistiques bayésiennes pour répondre à une série de questions formulées dans le cadre du premier objectif spécifique, et 3) créer, à partir des codes de programmation élaborés dans le cadre du deuxième objectif spécifique, des prototypes d'outils répondant aux besoins définis dans le cadre du premier objectif spécifique.

Le premier objectif spécifique a été atteint en réalisant une revue des lignes directrices en vigueur à l'échelle internationale ainsi que des publications pertinentes les plus récentes, complétées par des rencontres avec des comités d'experts et d'intervenants. Le deuxième objectif spécifique a été atteint en développant des solutions bayésiennes applicables à la liste de calculs établie à la première étape, en convertissant ces algorithmes en code statistique et en traduisant ce code en langage de programmation. Finalement, les algorithmes de programmation ont été utilisés pour créer des prototypes d'analyse de données fonctionnels illustrant les calculs possibles et pouvant servir de point de départ à la création d'outils complets d'analyse de données.

La liste de calculs pertinents issue du premier objectif spécifique et ayant par la suite servi de base à l'élaboration de formules mathématiques, d'algorithmes et de prototypes comportait deux grands volets. Le premier concernait l'estimation des paramètres d'une seule distribution, soit l'approche conventionnelle axée sur l'évaluation d'un groupe dit « d'exposition similaire ». Les mesures sont alors présumées provenir d'une distribution d'expositions partagées par un groupe de travailleurs accomplissant des tâches similaires. En guise d'illustration, ce modèle permet de répondre à la question : « Quelle est la probabilité que les expositions non mesurées au sein de ce groupe dépassent la VLEP plus de 5 % du temps ? » Le second volet étendait la portée du premier modèle de manière à pouvoir estimer la mesure dans laquelle un groupe de travailleurs partage ou non des expositions similaires. La variabilité globale de l'exposition est alors divisée en variabilité intra-travailleur et en variabilité inter-travailleur. Cela permet non seulement d'évaluer le risque de groupe, mais aussi de déterminer si certains travailleurs individuels sont plus à risque que le groupe. En guise d'illustration, ce modèle permet de répondre à la question : « Bien que l'exposition de groupe semble acceptable, quelle est la probabilité qu'un travailleur au hasard subisse une exposition supérieure à la VLEP plus de 5 % du temps ? » Tous les modèles incluent le traitement des valeurs non détectées et tiennent compte des erreurs de mesure associées au prélèvement et à l'analyse.

Les algorithmes issus de l'exercice sont disponibles en R pour les chercheurs, en C# pour les applications autonomes hors ligne ou sur serveur, et en JavaScript pour les applications Web. Ils couvrent la saisie de données, l'estimation bayésienne, des modules d'interprétation numérique et une interface utilisateur limitée aux prototypes en C# et en JavaScript. Le code est publiquement accessible aux termes de la licence libre Apache 2.0 pour permettre aux utilisateurs de créer leurs propres applications.

Le projet WebExpo devrait générer une trousse d'outils complète à la disposition de la communauté de l'hygiène du travail aux fins d'interprétation des niveaux d'exposition professionnelle, tout en offrant aux utilisateurs la souplesse de créer ou d'adapter leur propre logiciel plutôt que d'en utiliser un nouveau.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ACRONYMES ET ABRÉVIATIONS

AIHA : American Industrial Hygiene Association

BOHS : British Occupational Hygiene Society

C95 : 95^{e} centile

CEN : Comité européen de normalisation

CV, CV_e : coefficient de variation

ÉTG écart-type géométrique

GES : groupe d'exposition similaire

HT : hygiène du travail / hygiéniste du travail

IDC : intervalle de crédibilité

INRS : Institut national de recherche et de sécurité

LIMS : *Laboratory Information Management System*

LQ : limite de quantification

MA : moyenne arithmétique

MCCM : méthode de Monte-Carlo par chaînes de Markov

MG : moyenne géométrique

MP : moyenne pondérée

NIOSH : National Institute of Occupational Safety and Health

NVvA : Société néerlandaise d'hygiène du travail

OSHA : Administration américaine de la santé et de la sécurité au travail

RSPSAT : Réseau de santé publique en santé au travail

VLEP : valeur limite d’exposition professionnelle

VLEP-PLT : VLEP moyenne pondérée à long terme

1. INTRODUCTION

1.1 Maîtrise et gestion de l'exposition aux substances chimiques dans l'air des lieux de travail

Le but premier de l'hygiène du travail consiste à identifier les dangers et à évaluer, maîtriser et gérer les risques en milieu de travail. Une part importante de ces activités repose sur l'acquisition de connaissances relatives au niveau d'exposition des travailleurs à des substances chimiques dans l'air qu'ils respirent. L'évaluation de l'exposition peut être requise à plusieurs fins, mais consiste souvent à comparer le niveau d'exposition des travailleurs à des valeurs limites d'exposition professionnelle (VLEP) ou à des valeurs définies dans les lignes directrices adoptées par divers gouvernements et organismes. L'évaluation de l'exposition peut aussi servir à comprendre les facteurs qui déterminent l'intensité de l'exposition, de manière à mieux cibler les interventions.

Alors que certains des besoins en matière d'évaluation de l'exposition peuvent être satisfaits par des méthodes indirectes, comme la gestion graduée des risques ou l'utilisation de modèles mathématiques, de nombreuses situations exigent des mesures directes de l'exposition par échantillonnage et analyse de l'air respiré par les travailleurs.

1.2 Implications de la variabilité environnementale sur l'évaluation de l'exposition : le profil d'exposition

Lorsqu'on mesure l'exposition sur les lieux de travail, on cherche généralement non seulement à acquérir des connaissances sur la période faisant l'objet de la mesure, mais aussi à en déduire les conditions habituelles d'exposition dans des circonstances comparables. D'où l'objectif d'obtenir un portrait représentatif de l'exposition correspondant à un ensemble de conditions. Par exemple, lorsqu'on évalue le niveau d'exposition d'un travailleur au cours d'un quart de travail complet, on aimerait pouvoir utiliser cette information pour en tirer des connaissances sur tous les jours où l'on ne procède à aucune mesure. C'est en effet l'ensemble des expositions journalières subies par le travailleur, son « profil d'exposition », qui reflète le risque.

Il a très tôt été reconnu que les niveaux d'exposition en milieu de travail varient considérablement dans le temps et dans l'espace, de même qu'entre les travailleurs. Même les mesures intégrées sur un quart de travail complet et prises à répétition dans des conditions de travail similaires peuvent souvent présenter des variations allant du simple au décuple d'une journée à l'autre (N. Esmen, 1979; Kumagai et Matsunaga, 1995; Oldham, 1953; Rappaport, 2000). Par conséquent, l'exposition correspondant à une situation particulière (p. ex. peindre des pièces métalliques dans un atelier de carrosserie) ne peut pas être décrite par une seule valeur de concentration type. Elle est plutôt caractérisée par un ensemble de niveaux d'exposition différents en raison d'infimes variations propres à de nombreux facteurs déterminants (p. ex. la surface à peindre, le fait que les portes soient ouvertes ou fermées, les déplacements d'air, l'expérience des travailleurs). L'estimation de cette variabilité est essentielle pour tracer un portrait exact du profil d'exposition et pour évaluer le risque de façon fiable.

1.3 Statistiques en hygiène du travail : la distribution lognormale

Des méthodes statistiques conçues pour relever le défi posé par la variabilité environnementale ont commencé à voir le jour dans la littérature scientifique au cours des années 1960 (Breslin, Ong, Glauberman, George et Leclare, 1967; Kerr, 1962; Roach, 1966), et elles ont peu à peu convergé vers l'établissement de lignes directrices publiées par diverses instances œuvrant à guider la pratique de l'hygiène du travail dans différents pays. Ainsi l'American Industrial Hygiene Association (AIHA) (Hawkins, Norwood et Rock, 1991), le National Institute for Occupational Safety and Health (NIOSH) (N. A. Leidel, Busch et Lynch, 1977), les sociétés britannique et néerlandaise d'hygiène du travail (BOHS et NVvA, respectivement) (BOHS-NVvA, 2011; BOHS Technology Committee Working Group, 1993) et l'Institut national de recherche et de sécurité (INRS) en France (INRS, 2018) ont-ils tous publié des lignes directrices sur la comparaison des niveaux d'exposition à des valeurs limites d'exposition. L'Union européenne a par ailleurs récemment mis à jour des recommandations (CEN, 2018) initialement publiées en 1995 (CEN, 1995).

Les méthodes préconisées supposent que la variabilité environnementale est adéquatement modélisée par le modèle de distribution lognormal. Selon cette approche, on présume que, pour un groupe d'exposition donné (p. ex. les soudeurs d'acier inoxydable dans une usine de fabrication de pièces), l'ensemble des niveaux d'exposition des travailleurs du groupe au cours d'une période où les conditions de travail sont relativement stables (p. ex. une année) – ensemble ci-après appelé « distribution des expositions » – suit un modèle lognormal. Par suite, lorsqu'on procède à une série de mesures, on présume qu'elles constituent un échantillon aléatoire de cette distribution d'expositions. Il devient dès lors possible de tirer des conclusions sur la distribution globale des expositions à partir de cet échantillon, inclusion faite des journées faisant l'objet de mesures ou non. Nombre d'études probantes publiées à ce jour suggèrent que le modèle lognormal constitue une hypothèse par défaut raisonnable pour la plupart des situations d'exposition impliquant des vapeurs et des aérosols (N. A. Esmen et Hammad, 1977; Kumagai et Matsunaga, 1995; Oldham, 1953; Roach, 1977).

1.4 Meilleures pratiques actuelles en matière d'interprétation des données de mesure en hygiène du travail

Les approches recommandées pour comparer les niveaux d'exposition mesurés à une VLEP ont considérablement évolué au cours des trente dernières années. La première ligne directrice en matière d'interprétation de données fondée sur un cadre statistique a été proposée par le NIOSH en 1977. L'organisme recommandait alors que les expositions soient maîtrisées de telle sorte que moins de 5 % des niveaux d'exposition subis par un travailleur dépassent la VLEP (N. A. Leidel *et al.*, 1977) (autrement dit, la « fraction de dépassement » devait être inférieure à 5 %, un critère qu'on retrouve aussi dans la plus récente norme européenne [CEN, 2018]). À l'époque, le NIOSH suggérait de vérifier le respect de cette exigence en comparant une valeur d'exposition unique à un seuil d'intervention fixé à la moitié de la VLEP. Bien que cette proposition ait été fondée sur des bases statistiques et qu'elle ait fourni un moyen pratique d'évaluer le risque, il a par la suite été établi que la comparaison d'une mesure unique à la moitié de la VLEP ne permettait pas de s'assurer que 95 % des expositions non mesurées seraient inférieures à la VLEP (Buringh et Lanting, 1991; Lyles et Kupper, 1996; Rappaport, 1984; Tornero-Velez, Symanski, Kromhout, Yu et Rappaport, 1997). Des développements méthodologiques survenus au cours des décennies suivantes ont permis de cerner plusieurs

indices de risque fondés sur la distribution lognormale (voir ci-dessous), lesquels ont été retenus lors d'un atelier tenu en 2008 sur la mise à jour des directives du NIOSH (Ramachandran, 2008). Dans tous les cas, la distribution des expositions correspond à l'ensemble des concentrations auxquelles est exposé un groupe de travailleurs tenu pour partager des conditions d'exposition similaires (ce qu'on appelle un « groupe d'exposition similaire » ou GES).

1.4.1 Proportion des expositions supérieures à la VLEP (fraction de dépassement)

Cet indice est directement lié à la proposition initiale du NIOSH selon laquelle moins de 5 % des expositions devaient dépasser la VLEP (N. A. Leidel *et al.*, 1977; N. Leidel, Busch et Crouse, 1975). En ce qui concerne les expositions sur l'ensemble d'un quart de travail, la distribution d'intérêt engloberait toutes les expositions moyennes pondérées survenues au cours d'une période de conditions stables (généralement une année). Il s'agirait alors de prélever un échantillon aléatoire de cette distribution et d'estimer la fraction de dépassement, c'est-à-dire la proportion de jours où l'on s'attend à ce que l'exposition soit supérieure à la VLEP. Le calcul de la fraction de dépassement est recommandé par l'INRS en France, par les sociétés britannique et néerlandaise d'hygiène du travail (BOHS/NVvA) ainsi que par le Comité européen de normalisation (CEN), et il est au fondement même de la réglementation française en vigueur (BOHS-NVvA, 2011; CEN, 2018; INRS, 2018; République française, 2009). Étant donné que l'estimation de la fraction de dépassement s'effectue à partir d'un échantillon de la distribution des expositions, l'incertitude doit être prise en compte dans le calcul des limites de confiance autour de l'estimation. Dans la recommandation ci-dessus, le respect de la VLEP repose sur la démonstration que la limite de confiance supérieure à 70 % de la fraction de dépassement est inférieure à 5 %. En termes plus simples, il faut pouvoir démontrer avec un degré de certitude d'au moins 70 % que moins de 5 % des expositions dépassent la VLEP. La comparaison de la fraction de dépassement à 5 % est numériquement équivalente à la comparaison du 95^{e} centile estimé de la distribution sous-jacente à la VLEP (Clerc et Vincent, 2014). Ce dernier calcul est recommandé dans les lignes directrices actuelles de l'AIHA (Jahn, Bullock et Ignacio, 2015), assorti de l'établissement d'une limite de confiance supérieure à 95 % (plutôt qu'à 70 %, comme ci-dessus).

1.4.2 Moyenne arithmétique à long terme de la distribution des expositions

Les modèles toxicocinétiques ont démontré que la moyenne arithmétique (MA) de la distribution à long terme des niveaux d'exposition constitue un indice de risque plus pertinent en ce qui a trait à l'évaluation des dommages cumulatifs résultant de l'exposition à la plupart des substances qui présentent une toxicité chronique que les indices axés sur la borne supérieure de la distribution (comme la fraction de dépassement) (Rappaport, 1991). Dans ce contexte, il s'agirait d'effectuer un certain nombre de mesures, d'estimer la moyenne arithmétique de la distribution des expositions sous-jacente ainsi que les limites de confiance autour de l'estimation ponctuelle, et de les comparer à la VLEP. L'utilisation de cet indice a soulevé certains débats du fait qu'il est moins prudent que les indices de dépassement (p. ex., dans le cas d'une distribution lognormale type, une MA à hauteur de la VLEP correspondrait à une fraction de dépassement d'environ 30 %) (P Hewett, 1997; Lyles et Kupper, 1996; Tornero-Velez *et al.*, 1997). Les lignes directrices actuelles de l'AIHA recommandent cette approche dans les cas où la limite d'exposition a été explicitement définie comme un indice de dose

cumulative à long terme (« VLEP moyenne pondérée à long terme » [VLEP-PLT]) (Jahn *et al.*, 2015).

1.4.3 Probabilité de surexposition individuelle

Dans la foulée des travaux charnières de Kromhout, Rappaport et Symanski (Kromhout, Symanski et Rappaport, 1993; Rappaport, Kromhout et Symanski, 1993), il a été reconnu au cours des années 1990 que la pratique habituelle consistant à réunir sous l'appellation de « groupe d'exposition homogène » les travailleurs effectuant des tâches similaires dans un même environnement risquait d'entraîner une sous-estimation du risque pour certains membres du groupe. En dépit d'une distribution acceptable des expositions de groupe, il se pourrait en effet qu'en présence d'une forte variabilité de l'exposition entre les travailleurs, certains d'entre eux présentent une distribution d'expositions individuelles inacceptable. Cette considération trouve clairement écho dans les lignes directrices de l'AIHA, où « groupe d'exposition homogène » a été remplacé par « groupe d'exposition similaire » (GES) dans les plus récentes éditions. L'AIHA recommande également d'avoir recours à des méthodes d'analyse de la variance lorsqu'on dispose de suffisamment de données pour déterminer de manière empirique si l'exposition du groupe est effectivement « homogène » (Hawkins *et al.*, 1991; Ignacio et Bullock, 2008; Mulhausen et Diamano, 1998). Cette considération fait partie intégrante de la plus récente ligne directrice de la BOHS et de la NVvA, intitulée *Testing Compliance with Occupational Exposure Limits for Airborne Substances* [Vérification du respect des limites d'exposition professionnelle aux substances aéroportées] (BOHS-NVvA, 2011). Cette ligne directrice comporte deux volets. La distribution des expositions du groupe est d'abord évaluée afin de déterminer si moins de 5 % des expositions sont supérieures à la VLEP (ce qui rejoint la recommandation européenne mentionnée ci-dessus). Si le risque de groupe est acceptable, la ligne directrice préconise des tests visant à déterminer s'il existe une variabilité d'exposition significative entre les travailleurs du groupe afin d'estimer la probabilité que la distribution des expositions d'un travailleur au hasard corresponde à une fraction de dépassement supérieure à 5 %. Si on estime cette probabilité à plus de 20 %, le diagnostic de la ligne directrice en est un de « non-respect ». Dans leurs premières recommandations, Rappaport *et al.* et Lyles *et al.* suggéraient de déterminer la probabilité que la moyenne arithmétique des expositions d'un travailleur au hasard soit supérieure à la VLEP, et de la comparer à un seuil de 10 % (Lyles, Kupper et Rappaport, 1997b, 1997a; Rappaport, Lyles et Kupper, 1995). Selon cette approche, le degré de variation inter-travailleur peut être établi en calculant le coefficient de corrélation rho intra-travailleur, ou ce qu'on appelle le rapport R. Rho est proche de 1 quand les différences inter-travailleur sont importantes : plus les travailleurs sont différents, plus les mesures propres à un même travailleur sont proches les unes des autres par rapport à celles d'autres travailleurs. Le rapport R, initialement proposé par Rappaport *et al.*, correspond plus ou moins au rapport de la moyenne géométrique (MG) des expositions du travailleur le plus exposé à la MG des expositions du travailleur le moins exposé (Rappaport *et al.*, 1993).

En résumé, les lignes directrices actuelles en matière d'interprétation des données d'hygiène du travail recommandent principalement quatre indices comme étant les plus pertinents à l'évaluation du risque : la fraction de dépassement, le 95^{e} centile de la distribution des expositions, la moyenne arithmétique lorsqu'on dispose de VLEP moyennes pondérées à long terme, et la probabilité de surexposition individuelle (la surexposition étant définie comme une moyenne arithmétique individuelle supérieure à la VLEP ou un 95^{e} centile individuel supérieur à la VLEP). Ces indices peuvent également être utilisés pour d'autres analyses que la

comparaison avec la VLEP, y compris l'évaluation de l'effet des déterminants de l'exposition (p. ex. l'effet d'une intervention).

1.5 Méthodes d'interprétation bayésiennes des données d'exposition professionnelle

Les statistiques bayésiennes offrent une alternative à l'approche conventionnelle, dite « fréquentiste », lorsqu'il s'agit de tirer des conclusions au sujet d'une population à partir d'un échantillon. La méthode d'inférence bayésienne repose sur l'établissement de présomptions a priori à propos d'un ensemble de paramètres inconnus sous la forme de distributions de probabilités. Le théorème de Bayes est ensuite utilisé pour raffiner ces présomptions à la lumière d'observations empiriques, ce qui permet d'obtenir des distributions de probabilités « a posteriori » pour les paramètres d'intérêt. Bien que la théorie date du 18^{e} siècle, les techniques bayésiennes n'ont gagné en popularité qu'assez récemment, grâce à la puissance de calcul grandissante des ordinateurs. On en est venu à suggérer que les statistiques bayésiennes soient appliquées à l'hygiène du travail du fait qu'elles permettent d'intégrer un jugement expert (sous forme de présomptions a priori) aux données de mesure (S. Banerjee, Ramachandran, Vadali et Sahmel, 2014; Paul Hewett, Logan, Mulhausen, Ramachandran et Banerjee, 2006; Ramachandran et Vincent, 1999; Sottas *et al.*, 2009).

1.5.1 Principe de l'analyse de données bayésienne

L'analyse de données bayésienne débute par l'établissement de distributions de probabilités a priori (les « a priori ») relatives aux paramètres d'un modèle – ces a priori correspondant aux connaissances disponibles sur les paramètres en question – avant de considérer le jeu de données actuel. Ces distributions a priori sont ensuite mises à jour à la lumière de l'information tirée du jeu de données actuel par le truchement de la fonction de vraisemblance, ce qui permet d'obtenir la distribution a posteriori des paramètres (Gelman, 2013; McElreath, 2016). La distribution a posteriori correspond à l'ensemble des connaissances désormais disponibles, les renseignements passés ayant été jumelés à ceux tirés du jeu de données actuel. Toutes les inférences découlent par conséquent de cette distribution a posteriori.

Pour effectuer cette mise à jour, on a recours au théorème de Bayes, qui peut simplement s'énoncer comme suit :

$$a\ posteriori = \frac{a\ priori * vraisemblance}{constante\ de\ normalisation}$$

La constante de normalisation ne sert qu'à faire en sorte que la distribution a posteriori s'intègre à l'unité.

La distribution a priori correspond à l'information disponible avant l'analyse du jeu de données actuel – il est possible de définir une distribution a priori de manière à ce que sa teneur en information soit très élevée (on la qualifie alors d'« informative ») ou très peu élevée (on la qualifie alors de « faiblement informative » ou « non informative »). Dans le contexte de l'HT, les a priori informatifs sont attrayants parce que l'information ajoutée aux mesures effectives compense, dans une certaine mesure, la taille réduite des échantillons couramment utilisés dans les évaluations d'HT (Sudipto Banerjee, Ramachandran, Vadali et Sahmel, 2014; Paul Hewett *et al.*, 2006; Ramachandran et Vincent, 1999; Sottas *et al.*, 2009).

1.5.2 L'analyse de données bayésienne en hygiène du travail

Parmi les premiers à introduire les méthodes bayésiennes dans notre domaine, Ramachandran et Vincent proposaient de recourir au jugement expert pour guider la reconstitution des expositions passées (Ramachandran et Vincent, 1999). Hewett *et al.* ont proposé un outil pour évaluer la probabilité que le 95^{e} centile de la distribution des expositions se trouve dans chacune des catégories de gestion de l'exposition de l'AIHA (Paul Hewett *et al.*, 2006). Sottas *et al.* ont proposé un outil combinant les mesures aux informations a priori provenant d'un jugement expert, d'une base de données d'exposition existante et d'un modèle mécaniste (Sottas *et al.*, 2009). Plus récemment, Banerjee *et al.* ainsi que McNally *et al.* ont proposé des cadres d'analyse bayésiens pour comparer les données d'exposition aux VLEP (Sudipto Banerjee *et al.*, 2014; McNally *et al.*, 2014). Jones et Burstyn, de même que Quick *et al.* ont proposé des distributions a priori particulières à utiliser lors de l'interprétation des données de mesure au moyen des statistiques bayésiennes, tandis que Huynh *et al.* ont comparé statistiques traditionnelles et statistiques bayésiennes quant au traitement des valeurs non détectées (Huynh *et al.*, 2016; Jones et Burstyn, 2017; Quick, Huynh et Ramachandran, 2017). Plus récemment, Remy-Martin *et al.* ainsi que Groth *et al.* ont proposé des solutions bayésiennes pour traiter les données censurées bivariées aux fins de régression linéaire (Groth *et al.*, 2017; Martin Remy et Wild, 2017).

Le premier outil d'analyse de données bayésien en HT reposait sur l'établissement d'informations a priori sous la forme de probabilités a priori que le 95^{e} centile de la distribution des expositions se trouve dans chacune des catégories de gestion du risque de l'AIHA, les probabilités en question étant mises à jour en fonction des données d'observation par voie d'analyse bayésienne (Paul Hewett *et al.*, 2006). Quelques années plus tard, un modèle bayésien a été élaboré pour l'outil Advanced REACH (ART[1]), dans lequel une approche mécaniste jumelée à une base de données de mesures correspondant à divers scénarios d'exposition renseigne les distributions a priori (McNally *et al.*, 2014). Depuis, plusieurs autres options ont été proposées pour produire des a priori informatifs, même si, à notre connaissance, aucune d'elles n'a donné lieu au développement d'outils pratiques. Parmi ces options, retenons l'utilisation de modèles mécanistes (Zhang, Banerjee, Yang, Lungu et Ramachandran, 2009), d'études existantes pertinentes (Quick, Huynh et Ramachandran, 2017), d'un simple logiciel comme Excel pour estimer les distributions a posteriori (Jones et Burstyn, 2017), et d'a priori provenant de bases de données d'expositions antérieures (Sottas *et al.*, 2009). L'approche bayésienne traditionnelle recommande d'évaluer la robustesse d'une analyse à travers un éventail d'a priori différents afin d'élargir la portée de l'interprétation (Gelman, 2013). Ainsi, elle sera applicable à une plus grande variété d'interprétations des informations antérieures. Compte tenu des tailles d'échantillons réalistes dans notre domaine (5 à 10 observations), des a priori informatifs comme ceux décrits ci-dessus ont généralement un effet non négligeable sur les estimations finales de l'exposition (Jones et Burstyn, 2017).

Il existe d'autres avantages importants à utiliser les statistiques bayésiennes pour interpréter les données d'hygiène du travail. L'inférence bayésienne est de nature probabiliste. Ainsi, plutôt que de renvoyer un test d'hypothèse ou un intervalle de confiance, dont la juste interprétation est parfois difficile à communiquer au profane, l'analyse bayésienne fournit directement des réponses aux questions du type « Quelle est la probabilité que... ? » (p. ex., « Quelle est la

[1] https://www.advancedreachtool.com/

probabilité que ce groupe soit surexposé plus de 5 % des jours ? » ou « Quelle est la probabilité que cette intervention réduise les niveaux d'exposition d'au moins 50 % ? »). Cela facilite la communication des risques liés à des notions complexes, aux dirigeants comme aux travailleurs. En outre, deux défis techniques que les approches traditionnelles n'abordent pas adéquatement à l'heure actuelle, à savoir le traitement des données non détectées et la prise en compte des erreurs de mesure dans les évaluations, sont faciles à intégrer dans une approche bayésienne (Espino-Hernandez, Gustafson et Burstyn, 2011; Mcbride, Williams et Creason, 2007; McNally *et al.*, 2014; Morton, Cotton, Cocker et Warren, 2010; Pilote *et al.*, 2000; Wild, Hordan, Leplay et Vincent, 1996).

Le cadre bayésien semble donc très prometteur en ce qui concerne l'amélioration de l'analyse et de l'interprétation des données en hygiène du travail. Malheureusement, sa mise en œuvre est actuellement hors de portée de la plupart des intervenants, dans la mesure où les calculs bayésiens nécessitent des logiciels de pointe et des connaissances techniques avancées généralement réservés à des spécialistes universitaires.

1.6 Le traitement des valeurs non détectées dans l'interprétation des données de mesure en hygiène du travail

En 1990, Hornung et Reed écrivaient que la réduction des niveaux d'exposition depuis les années 1970 – lesquels n'avaient été que partiellement compensés par une amélioration progressive des méthodes d'analyse – avait entraîné une augmentation de la proportion des données d'exposition déclarées comme étant non détectées (Hornung et Reed, 1990). Plus récemment, Lavoué *et al.* ont fait état d'un taux de données non détectées de 60 % sur 1,4 million de mesures consignées dans la base de données du laboratoire de Salt Lake City de l'Occupational Safety and Health Administration (OSHA) des États-Unis, une majorité des échantillons d'analyse ayant été prélevés par des inspecteurs du travail de l'OSHA depuis 1979 (Lavoué, Friesen et Burstyn, 2013). Sarazin *et al.* ont quant à eux signalé 40 % de valeurs non détectées parmi 0,5 million de mesures contenues dans la base de données du système de gestion de l'information du laboratoire de l'IRSST (*Laboratory Information Management System* [LIMS]), qui renferme les résultats d'analyse d'échantillons recueillis au Québec depuis 1985 par des hygiénistes du travail gouvernementaux (Sarazin, Labrèche, Lesage et Lavoué, 2018).

Il a été amplement démontré que l'élimination des valeurs non détectées ou leur remplacement par une quelconque valeur fixe fausse l'estimation de la plupart des paramètres d'intérêt (D. R. Helsel, 2012; D. Helsel, 2005). L'ampleur de l'erreur augmente au gré de la proportion des valeurs non détectées, et elle devient particulièrement importante au moment de procéder à des tests statistiques ou de calculer des intervalles de confiance. Malgré l'omniprésence des valeurs non détectées et leur impact potentiel sur l'interprétation des données, peu de méthodes relatives à leur traitement ont été proposées dans notre domaine, et ce, malgré les éditoriaux parus dans *Annals of Occupational Hygiene* et réclamant des avancées à cet égard (Dennis Helsel, 2010; T. L. Ogden, 2010). D'importants progrès ont toutefois récemment été signalés (Flynn, 2010; Ganser et Hewett, 2010; Groth *et al.*, 2017; Krishnamoorthy, Mallick et Mathew, 2009; Martin Remy et Wild, 2017), plusieurs études de simulation ayant été menées pour comparer différentes approches (Paul Hewett et Ganser, 2007; Huynh *et al.*, 2014, 2016).

Il s'avère que les méthodes bayésiennes sont tout indiquées pour relever ce défi, car elles permettent de multiples points de censure et estiment avec précision l'incertitude inhérente lorsque les valeurs de données ne sont connues qu'à hauteur d'un certain intervalle (Huynh

et al., 2016). Ces récentes percées n'ont malheureusement pas encore été intégrées aux outils d'analyse de données pratiques.

1.7 L'erreur de mesure dans l'interprétation des données d'hygiène du travail

En plus de la variabilité des niveaux d'exposition comme tels, chaque valeur d'un ensemble de mesures d'exposition comporte un facteur d'erreur dû à l'échantillonnage et à l'analyse. Ce facteur, généralement exprimé sous forme de coefficient de variation (CV), est traditionnellement pris en compte au moment d'interpréter une mesure d'exposition unique afin de déterminer si la concentration réelle de l'exposition sous-jacente était supérieure ou inférieure à la VLEP. À titre d'illustration, Leidel et Bush proposent différentes formules pour estimer les intervalles de confiance correspondant à une valeur moyenne pondérée sur la base de l'erreur d'échantillonnage et d'analyse (N. A. Leidel et Busch, 2000). Cependant, l'erreur de mesure n'a traditionnellement pas été considérée en lien avec l'interprétation d'un ensemble de mesures d'exposition en vue d'estimer les paramètres de la distribution des niveaux d'exposition. Le défi que soulève ce type d'analyse tient au fait que la variabilité environnementale est généralement modélisée par une distribution lognormale, alors que l'erreur de mesure due à l'échantillonnage et à l'analyse est plutôt modélisée par une distribution normale (Ashley et Bartley, 2004; Bartley, 2001; Bartley et Lidén, 2008). Une analyse combinée de cet ordre est en effet impensable au moyen des statistiques conventionnelles.

La pratique actuelle consistant à ignorer l'erreur de mesure dans l'interprétation des ensembles de données d'HT est corroborée par deux études où la distribution normale liée à l'erreur de mesure était approximée par une distribution de probabilités lognormale (Grzebyk et Sandino, 2005; Nicas, Simmons et Spear, 1991). Nicas *et al.* (Nicas *et al.*, 1991) ont estimé la part de l'erreur de mesure dans la variabilité totale observée, tandis que Grzebyk et Sandino (Grzebyk et Sandino, 2005) ont formulé des équations permettant d'évaluer le biais encouru quant à l'estimation de la moyenne géométrique (MG) et de l'écart-type géométrique (ÉTG). Les deux études concluaient que l'erreur de mesure joue un rôle négligeable lorsque le CV correspondant est inférieur à 30 % et que la variabilité environnementale est élevée (ÉTG > 2). Cependant, la variabilité peut être faible dans certains lieux de travail, et certaines méthodes d'échantillonnage comportent une importante erreur de mesure. Par ailleurs, aucune approche n'a encore été proposée pour estimer l'erreur liée à une valeur d'exposition sur un quart de travail complet calculée à partir d'une série d'échantillons partiels ne couvrant pas la totalité du quart, laquelle pourrait être plus importante que l'erreur d'échantillonnage et d'analyse type. Enfin, ni Grzebyk et Sandino ni Nicas *et al.* n'ont estimé l'impact de l'erreur de mesure sur les indices décisionnels décrits ci-dessus. Par conséquent, bien que le biais d'estimation de l'ÉTG puisse sembler faible (le fait d'ignorer l'erreur de mesure entraînerait généralement une surestimation de l'ÉTG vrai), son impact réel sur la limite de confiance supérieure du 95^{e} centile de la distribution des expositions (souvent utilisé dans la prise de décision) pourrait être important.

Comme dans le cas des valeurs non détectées, les statistiques bayésiennes représentent une alternative prometteuse aux statistiques fréquentistes, car elles peuvent tenir compte de l'erreur de mesure de manière flexible (Espino-Hernandez *et al.*, 2011; Morton *et al.*, 2010; Pilote *et al.*, 2000).

1.8 Défis liés à l'interprétation des données et à la communication des risques

De récentes études sur le jugement expert ont révélé que les hygiénistes du travail évaluaient mieux les niveaux d'exposition lorsqu'ils avaient été formés aux statistiques lognormales (P. Logan, Ramachandran, Mulhausen et Hewett, 2009; P. W. Logan, Ramachandran, Mulhausen, Banerjee et Hewett, 2011). Au Québec, les approches modernes à l'interprétation des données ont été examinées et résumées dans un rapport de l'IRSST (Drolet *et al.*, 2013). Les auteurs y précisent que ces approches nécessitent des notions statistiques et des outils de calcul qui ne sont pas très répandus dans le domaine.

La communication des risques est un autre défi à relever. Toute amélioration à cet égard serait bienvenue, car les notions statistiques connexes restent souvent obscures pour les décideurs et les travailleurs. À titre d'exemple, il se peut, dans un contexte d'exposition donné, que les valeurs mesurées soient toutes inférieures à la VLEP, mais que la proportion estimée des expositions qu'on s'attend à voir dépasser la VLEP les jours ne faisant l'objet d'aucune mesure soit nettement plus élevée que les 5 % généralement tenus pour acceptables. Cette évaluation particulière semblerait sans doute contre-intuitive à un public non averti, mais elle devient intuitivement raisonnable lorsqu'on sait que la présence de plusieurs observations tout juste au-dessous de la valeur seuil dans une distribution présentant une longue queue (comme la lognormale) peut indiquer une forte probabilité qu'une valeur se trouve dans cette queue, et par conséquent au-dessus de la VLEP. La difficulté de communiquer efficacement des résultats statistiques de manière convaincante à des non-spécialistes et le manque d'outils pour ce faire peuvent également expliquer la lente mise en œuvre des lignes directrices modernes par les intervenants actifs dans le domaine.

1.9 Besoins en matière d'analyse numérique et statistique dans l'interprétation des données d'exposition professionnelle

Les procédures statistiques applicables aux paramètres lognormaux et à l'incertitude qui les entoure ne sont pas décrites dans les manuels de statistique standards, pour la plupart axés sur la distribution normale. Elles ont progressivement été élaborées depuis les années 1960, et elles demeurent en constante évolution. Bien que ces avancées aient peu à peu transpiré des rapports de recherche, jusqu'à se voir intégrées, avec le temps, dans les lignes directrices d'associations d'hygiène du travail, leur mise en œuvre peut s'avérer complexe, et donc difficile pour les intervenants qui ne disposent pas des connaissances et des outils statistiques nécessaires pour effectuer les calculs requis. Au Québec, le *Guide d'échantillonnage des contaminants de l'air en milieu de travail* (Drolet et Beauchamp, 2013) est cité dans la réglementation comme une référence en ce qui concerne le niveau de précision requis pour évaluer la conformité réglementaire des expositions aux VLEP. Ce guide fournit des instructions détaillées sur la façon de comparer une mesure à la VLEP afin de déterminer si l'exposition était conforme le jour où la mesure a été effectuée, ce qui est essentiel pour les spécialistes de la conformité réglementaire. Néanmoins, il ne renferme pas de documentation détaillée sur la distribution lognormale et les indices de risque connexes. Nous n'avons relevé que cinq outils d'évaluation pratiques disponibles mettant l'accent sur l'estimation des statistiques nécessaires à l'évaluation du risque inhérent aux substances chimiques dans l'air des milieux de travail (c.-

à-d. « statistiques d'hygiène du travail ») : IHSTAT[2] (chiffrier Excel gratuit), Altrex Chimie[3] (logiciel autonome téléchargeable et gratuit), IHData Analyst[4] (logiciel gratuit), BW_Stat[5] (chiffrier Excel gratuit) et HYGINIST[6] (logiciel autonome téléchargeable et gratuit). Il convient aussi de mentionner ProUCL[7], mis à disposition par l'Environmental Protection Agency des États-Unis, à savoir un ensemble d'outils générique axé sur les contaminants environnementaux et applicable à des ensembles de données d'exposition professionnelle. De plus, l'outil ART mentionné à la section 1.5.2, bien que plus précisément axé sur le cadre d'évaluation du risque défini par le règlement REACH, permet également d'estimer les centiles d'une distribution d'expositions et l'incertitude connexe (McNally *et al.*, 2014). Les outils propres à l'HT mentionnés ci-dessus présentent des similitudes, et ils permettent tous d'évaluer le risque de surexposition à partir d'un ou plusieurs indices. À ce titre, ils marquent un important pas en avant pour ce qui est de rendre les statistiques d'hygiène au travail plus accessibles. Cependant, aucun d'eux n'offre une solution unique intégrée et complète à l'interprétation des données lognormales. Leurs limites les plus notables concernent l'absence de traitement adéquat des valeurs non détectées, l'interprétation de données au-delà de l'évaluation de la surexposition (p. ex. l'effet d'une intervention) et le soutien à la communication probabiliste des risques.

En outre, nombre d'établissements et d'entreprises privées qui procèdent à des mesures d'exposition routinières possèdent leur propre base de données d'exposition ; cependant, aucun des outils présentés ci-dessus ne peut être facilement intégré à un système de gestion de données existant. Par conséquent, afin de pouvoir effectuer les calculs pertinents, il s'avère nécessaire d'exporter leurs données vers les outils existants pour ensuite réimporter les résultats dans le système interne. D'où le besoin d'une trousse d'outils statistiques en hygiène du travail à même de faciliter la programmation des calculs nécessaires dans les systèmes existants.

1.10 Résumé des lacunes et des besoins au chapitre des connaissances

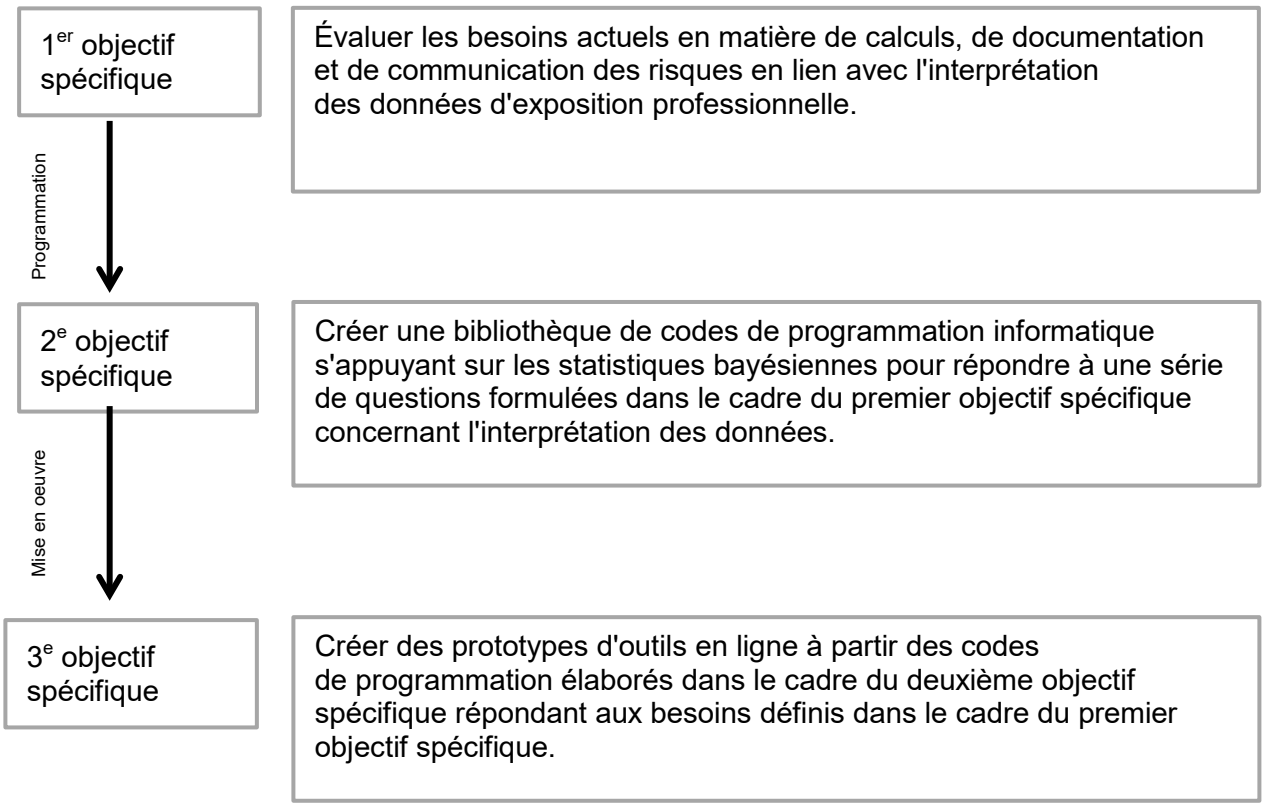
L'importante variabilité spatiale et temporelle observée dans les niveaux d'exposition a toujours constitué un défi de taille quant à leur interprétation. Un cadre d'analyse consensuel fondé sur la distribution lognormale existe désormais, mais bien que les avancées qui en découlent permettent une meilleure évaluation du risque que les approches traditionnelles, elles n'ont pas été largement adoptées par les intervenants en hygiène du travail. Elles font en effet appel à des notions statistiques généralement non abordées dans les programmes de formation habituels, et nécessitent des calculs difficilement réalisables avec de simples outils courants (p. ex. calculatrices ou chiffriers). Les quelques outils pertinents actuellement disponibles marquent un important pas en avant, mais ils ne répondent pas encore entièrement aux besoins des intervenants. De surcroît, les outils disponibles sont indépendants les uns des autres, et ils ne s'intègrent pas facilement à une structure de gestion de données existante. Enfin, bien que les méthodes bayésiennes offrent une approche très prometteuse en ce qui concerne

[2] https://www.aiha.org/get-involved/VolunteerGroups/Pages/Exposure-Assessment-Strategies-Committee.aspx
[3] http://www.inrs.fr/accueil/produits/mediatheque/doc/outils.html?refINRS=outil13
[4] https://www.easinc.co/ihda-software/
[5] https://www.bsoh.be/?q=en/node/89
[6] http://www.tsac.nl/hyginist.html
[7] https://www.epa.gov/land-research/proucl-software

l'interprétation des données d'hygiène au travail, elles ne sont actuellement pas accessibles à la plupart des intervenants. En conclusion, pour favoriser l'adoption de lignes directrices modernes quant à l'interprétation des données d'hygiène du travail et pour améliorer les pratiques d'évaluation des risques chimiques, d'importants besoins doivent être comblés, notamment en matière de transfert des connaissances et en outils d'analyse de données accessibles et complets.

2. OBJECTIFS DE RECHERCHE

Le projet WebExpo visait à améliorer le transfert des meilleures pratiques actuelles en ce qui concerne l'interprétation des niveaux d'exposition professionnelle aux fins d'évaluation des risques dans le domaine de l'hygiène du travail.



3. MÉTHODES

3.1 1er objectif spécifique : Évaluation des besoins actuels en matière de calculs, de documentation et de communication des risques

Sur la base de la revue présentée à la section 1, il est possible de dresser une liste indicative des caractéristiques essentielles qui devraient à tout le moins figurer dans un outil d'interprétation de données complet :

Indices pertinents à l'évaluation de groupe de la surexposition :

- Fraction de dépassement
- 95e centile
- Moyenne arithmétique

Indices pertinents à l'évaluation de la surexposition individuelle :

- Composantes de la variabilité inter-travailleur et intra-travailleur
- Corrélation intra-travailleur (rho)
- Rapport R
- Probabilité que la fraction de dépassement individuelle d'un travailleur soit trop élevée
- Probabilité que le 95e centile individuel d'un travailleur soit trop élevé
- Probabilité que la moyenne arithmétique individuelle d'un travailleur soit trop élevée

Gestion de l'incertitude :

- Calcul des intervalles de confiance propres à tous les indices ci-dessus
- Traitement des valeurs non détectées
- Traitement de l'erreur de mesure

Nous avons validé cette liste et envisagé de l'allonger en sollicitant la rétroaction d'intervenants réunis en deux comités.

Nous avons tout d'abord formé un comité d'intervenants constitué de spécialistes en hygiène du travail du Québec. Ce comité comptait neuf membres : un hygiéniste du travail (HT) d'une société d'experts-conseils, deux HT d'entreprises privées, un HT de l'IRSST, un HT du Réseau de santé publique en santé au travail (RSPSAT), un technicien en hygiène du travail et un médecin du travail du RSPSAT. Ce comité d'intervenants s'est réuni deux fois à raison d'une demi-journée chaque fois, soit au début et à la fin du projet. L'objectif principal de ce comité était de formuler des commentaires et des suggestions du point de vue des intervenants du Québec concernant leurs besoins et les obstacles à l'application des lignes directrices actuelles en matière d'interprétation des données.

En second lieu, nous avons créé un comité d'experts composé de spécialistes canadiens et étrangers dans le domaine des statistiques d'hygiène du travail, actifs aussi bien en milieu universitaire qu'en entreprise (tableau A1 de l'annexe A). Ce comité d'experts s'est réuni une fois pendant deux jours au début du projet. L'objectif principal de ce comité était de formuler des commentaires et des suggestions sur les choix méthodologiques relatifs aux calculs et aux caractéristiques à inclure dans les algorithmes.

Les deux comités ont contribué à l'établissement de la liste définitive des fonctionnalités et des calculs intégrés dans WebExpo.

3.2 2e objectif spécifique : Création d'une bibliothèque de codes de programmation informatique

Cette tâche peut être divisée en deux parties. La première consistait à définir des solutions bayésiennes théoriques aux problèmes d'estimation de la liste établie en 3.1. Comme l'explique en détail la section 3.2.1, cette étape consistait essentiellement à transposer dans le domaine de l'HT des techniques statistiques déjà utilisées dans d'autres domaines. La deuxième partie portait sur l'expression des calculs nécessaires sous forme d'algorithmes afin de faciliter leur utilisation par un large public, au-delà des utilisateurs de progiciels statistiques spécialisés. Comme l'explique la section 3.2.2, il a d'abord fallu appliquer les solutions retenues dans le progiciel statistique R, pour ensuite traduire le code de R dans des langages utilisés pour programmer des applications Web ou autonomes.

3.2.1 Établissement de modèles bayésiens applicables aux problèmes d'estimation définis en 3.1

Les modèles bayésiens du projet WebExpo ont été élaborés sur la base de la littérature disponible en HT, des techniques en usage dans d'autres domaines, ainsi que de l'expertise des membres de l'équipe de biostatistique de l'Université McGill (Lawrence Joseph et Patrick Bélisle), qui ont été invités à préciser les fondements mathématiques sous-jacents. Nous avons utilisé les modèles présentés dans Banerjee *et al.* (S. Banerjee *et al.*, 2014) et dans McNally *et al.* (McNally *et al.*, 2014), respectivement, comme point de départ pour les modèles axés sur les GES et la variabilité entre travailleurs, que nous avons par la suite élargis de manière à inclure la censure, l'erreur de mesure et différents types d'a priori.

On s'attendait à ce que les modèles du projet WebExpo soient trop complexes et à ce que les distributions a posteriori qui en découlent ne puissent être facilement exprimées en forme analytique fermée, c'est-à-dire pouvant être décrites sous forme d'équations. En conséquence, la simulation Monte-Carlo par chaîne de Markov (MCCM) serait nécessaire pour obtenir des échantillons des distributions a posteriori à partir desquels des inférences pourraient être faites. À titre d'exemple, considérons l'estimation de la moyenne μ d'une distribution normale ayant un écart-type σ. Après avoir défini des a priori pour les paramètres inconnus et recueilli un échantillon d'observations, l'extrant type d'une MCCM consisterait, par exemple, en 10 000 valeurs aléatoires de μ issues de la distribution a posteriori. L'estimation ponctuelle de μ correspondrait à la médiane de ces 10 000 valeurs, et le 2,5e et le 97,5e centiles des 10 000 valeurs constitueraient un intervalle de crédibilité symétrique à 95 % (IDC de 95 %). Les intervalles de crédibilité bayésiens sont interprétés comme des énoncés probabilistes directs : la probabilité que μ se trouve dans l'intervalle est de 95 %, compte tenu des observations, de la distribution a priori et de la fonction de vraisemblance utilisées.

Une caractéristique de l'analyse de données bayésienne particulièrement utile dans notre projet tient à ce que dès lors qu'on dispose d'échantillons de la distribution a posteriori en lien avec les paramètres du modèle, on dispose automatiquement d'échantillons de la distribution a posteriori pour toute fonction qui en dérive. Dans l'exemple ci-dessus, supposons que nous ayons obtenu 10 000 valeurs de μ et 10 000 valeurs de σ à partir de leur distribution a posteriori conjointe. Il

est alors facile d'obtenir 10 000 valeurs pour le coefficient de variation de la distribution, car il suffit de calculer CV = σ/μ pour chaque paire d'échantillons de μ et σ. Les extrants des modèles bayésiens étant des estimations de paramètres distributionnels de base (p. ex. moyenne géométrique et écart-type), nous avons également formulé des équations pour transformer les chaînes MCCM de ces estimations en indices pertinents, notamment ceux retenus en 3.1 (p. ex. le 95^{e} centile). Cette procédure reposait principalement sur les lignes directrices et les publications existantes en hygiène du travail (voir la section 1 pour les références). La figure 1 illustre le processus d'estimation bayésien comme il a été mis en œuvre dans WebExpo.

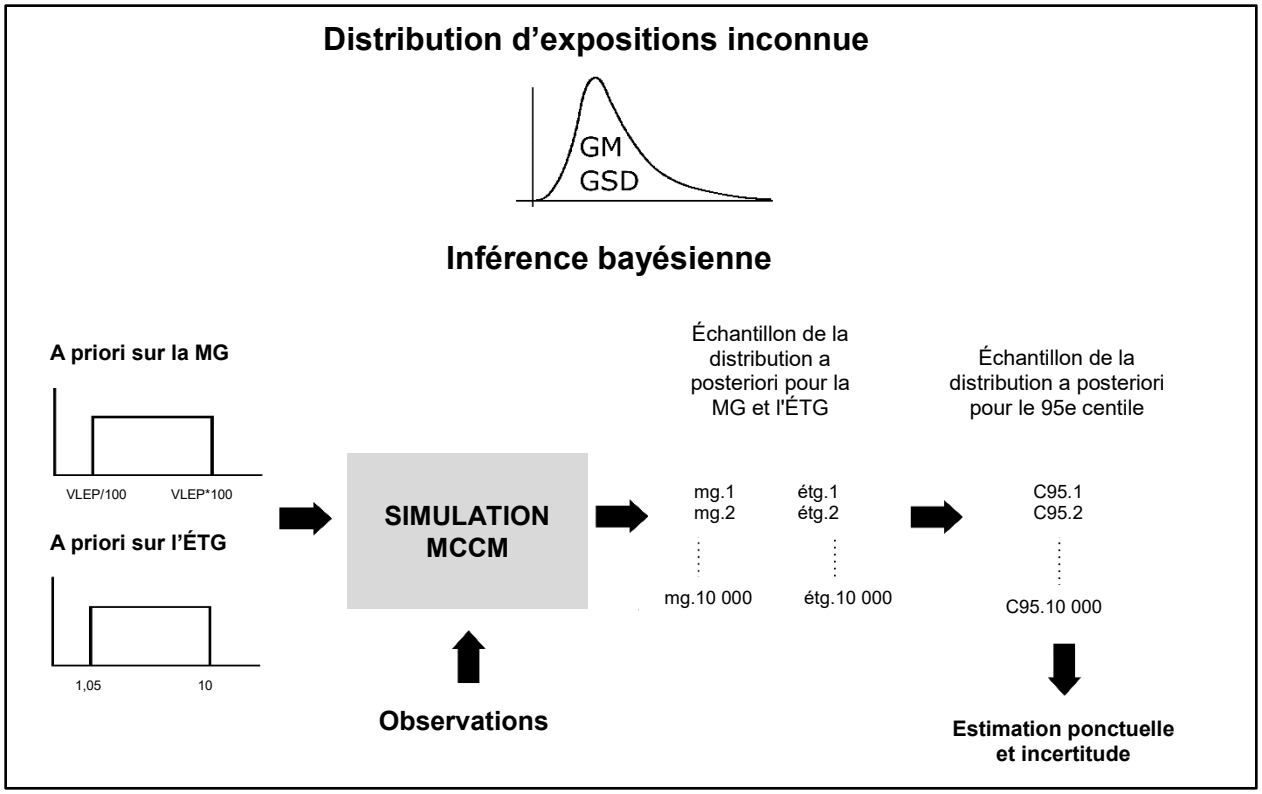


Figure 1. Illustration du cadre d'analyse bayésien dans WebExpo.

3.2.2 Création d'une bibliothèque de codes de programmation informatique

3.2.2.1 Approche générale

Les méthodes de Monte-Carlo par chaîne de Markov sont gourmandes en ressources informatiques, et les calculs bayésiens afférents sont généralement effectués au moyen de logiciels spécialisés comme Openbugs[8], Winbugs[9], JAGS[10] ou STAN[11]. Le code sur lequel reposent de telles applications est généralement reproduit dans des rapports de recherche comme celui de Banerjee *et al.* (2014). Cependant, ces programmes sont trop complexes pour une utilisation au jour le jour par les intervenants en HT. Par conséquent, la mise en application

[8] http://www.openbugs.net/w/FrontPage
[9] https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/
[10] http://mcmc-jags.sourceforge.net/
[11] http://mc-stan.org/

des algorithmes de WebExpo au moyen des logiciels mentionnés ci-dessus ne répondrait pas vraiment aux besoins des intervenants en matière de calcul.

Nous avons donc résolu d'implémenter dans un premier temps les modèles bayésiens de WebExpo au moyen d'un moteur MCCM sur mesure écrit dans le langage statistique libre de R (R Core Team, 2014) en utilisant des fonctions de calcul de base, une telle application pouvant par la suite être traduite dans les langages de programmation habituellement utilisés pour créer des outils pratiques ne nécessitant aucune licence particulière.

La première étape de mise en œuvre des algorithmes théoriques consistait à créer une bibliothèque de codes en R, inclusion faite du formatage initial des données, des calculs bayésiens et de la conversion des extrants des fonctions bayésiennes en indices d'exposition pertinents. Il en résulterait une bibliothèque de codes en R permettant d'effectuer tous les calculs inhérents au projet WebExpo. Toutefois, l'utilisation de cette bibliothèque de codes nécessiterait tout de même une certaine expertise en R, ainsi qu'une version locale de R sur l'ordinateur de l'utilisateur.

L'étape suivante a donc consisté à traduire cette bibliothèque de codes en deux langages de programmation informatique, de manière à faciliter l'exécution des routines correspondantes dans d'autres environnements informatiques sans avoir à utiliser des progiciels d'analyse statistique, le code ainsi traduit pouvant en outre être utilisé par des programmeurs pour créer diverses applications. La première traduction a été faite dans le langage de programmation Web JavaScript[12]. Cette version permet d'effectuer directement des calculs bayésiens dans un environnement Web standard (c.-à-d. dans le navigateur même de l'utilisateur). La deuxième traduction a été faite en C# (*C sharp* en anglais)[13], un langage de programmation répandu utilisé pour créer des applications autonomes. Il permet notamment de créer des logiciels téléchargeables pouvant exécuter toutes les routines produites dans le cadre du projet, ou d'intégrer ces routines à des applications de gestion de données existantes.

Enfin, les calculs ont aussi été codés dans R à l'aide d'appels à l'application JAGS – un moteur bayésien tiers permettant une simulation MCCM rapide pour un large éventail de modèles – par le biais du progiciel RJAGS[14]. Cet ensemble de fonctions en R+JAGS a été créé pour permettre aux utilisateurs de R d'effectuer les calculs décrits dans ce rapport avec une efficacité optimale, les modèles en code pur de R étant nettement plus lents que leurs homologues en R+JAGS, tout particulièrement en ce qui concerne l'erreur de mesure.

3.2.2.2 Contrôle de la qualité

La théorie sous-jacente aux modèles bayésiens dans le projet WebExpo repose sur des études publiées et concerne de surcroît des modèles relativement simples (modèles fondés sur l'estimation d'une seule distribution ou sur l'estimation des composantes de la variance). En conséquence, nous n'avons pas tenté d'effectuer des simulations, par exemple pour vérifier que notre procédure fournissait des estimations justes de la moyenne géométrique, ou que la couverture des intervalles de crédibilité était exacte. Les résultats du traitement bayésien des données censurées ont de même déjà été évalués par d'autres. Les modèles théoriques et les

[12] https://en.wikipedia.org/wiki/JavaScript
[13] https://en.wikipedia.org/wiki/C_Sharp_(programming_language)
[14] https://cran.r-project.org/web/packages/rjags/index.html

fonctions en R ont été écrits par deux statisticiens bayésiens d'expérience, Lawrence Joseph et Patrick Bélisle.

Notre principale préoccupation dans ce projet était que l'implémentation des algorithmes MCCM – qui reposent sur la génération de nombres aléatoires – permette d'obtenir des résultats comparables sur différentes plateformes de calcul (R, R+JAGS, C# et JavaScript). Même au sein d'une seule plateforme, le caractère aléatoire des simulations MCCM implique que l'exécution répétée d'une même analyse donnera des résultats légèrement variables. Lorsque le code pur de R est traduit en C# et en JavaScript, les résultats pourraient différer d'une plateforme à l'autre, soit parce que les fonctions mathématiques de base s'écrivent différemment dans ces langages, soit parce que les mécanismes de génération de nombres aléatoires ou les procédures d'arrondissement y diffèrent. Quant aux différences entre R et R+RJAGS, bien que les scripts en code pur de R et JAGS découlent du même modèle théorique, les algorithmes MCCM comme tels sont différents.

Lors de la traduction du code pur de R vers C# et JavaScript, nous avons cherché à faire en sorte que les différences entre les plateformes soient aussi minimes que possible, et qu'à cette fin, les codeurs en C# et en JavaScript communiquent régulièrement entre eux, en particulier lorsque des écarts notables étaient observés entre les plateformes. Concrètement, des échantillons standards (de taille, distribution, degré de censure et variabilité variables) ont été analysés sur toutes les plateformes, et les résultats ont été comparés.

Il était possible d'utiliser le même générateur de nombres aléatoires en C# et en R, ce qui a permis de comparer les chaînes MCCM pour chaque échantillon standard à chaque itération entre R et C# ; des statistiques descriptives des différences entre les itérations ont ensuite été utilisées pour mesurer l'accord entre les deux plateformes.

Dans le cas du JavaScript, bien que la même procédure ait en principe pu être appliquée, nous avons plutôt comparé des quantiles d'échantillons de la distribution a posteriori, à savoir les 1er, 2,5^{e}, 5^{e}, 25^{e}, 50^{e}, 75^{e}, 95^{e}, 97,5^{e} et 99^{e} centiles des chaînes MCCM pour tous les paramètres inconnus. Cette procédure était en effet plus simple à mettre en œuvre et moins chronophage que l'approche itération par itération.

La procédure a aussi été simplifiée au moment de comparer R à R+JAGS, puisqu'il y avait moins de risques d'écarts entre les plateformes (les calculs sont implémentés en R dans les deux cas). Pour chaque procédure d'estimation, un échantillon standard a été soumis 50 fois aux fonctions de R et de R+RJAGS, après quoi nous avons calculé, sur l'ensemble des répétitions, les plages d'estimations ponctuelles et de limites de crédibilité relatives aux paramètres inconnus. Les plages en question ont ensuite été comparées entre R et RJAGS pour s'assurer qu'elles étaient comparables, compte tenu de la variabilité observée à l'intérieur de chaque approche.

3.3 3^{e} objectif spécifique : Création de prototypes d'outils

Les bibliothèques de codes en JavaScript et C# décrites en 3.2 ont été utilisées pour créer des prototypes d'outils d'interprétation de données dans les deux langages. Ces prototypes, également d'exploitation libre, visent à mettre en valeur les calculs que les algorithmes permettent d'effectuer ; ils présentent une interface utilisateur minimale, mais n'en affichent pas moins les résultats numériques essentiels. Les deux prototypes offrent une interface de saisie de données dans laquelle des valeurs doivent être inscrites pour tous les paramètres

nécessaires aux fonctions parallèlement au jeu de données à analyser. Les extrants comprennent les chaînes MCCM comme telles de même que les indices d'exposition du tableau 2 et les intervalles de crédibilité. Aucune illustration graphique ni interprétation des résultats n'est fournie. Les deux prototypes serviront de point de départ à la création ultérieure d'un outil d'interprétation de données pratique complet exclusif à l'IRSST.

4. RÉSULTATS

4.1 1er objectif spécifique : Évaluation des besoins actuels en matière de calculs, de documentation et de communication des risques

La revue de la littérature présentée à la section 1 a fait ressortir deux approches principales correspondant à deux modèles statistiques différents en ce qui concerne l'interprétation des données d'HT. La première, ci-après appelée « analyse de GES », fait référence à des situations où l'on dispose d'un ensemble de mesures pour un groupe de travailleurs dont l'exposition est similaire (c.-à-d. de travailleurs présumés partager la même distribution d'expositions) ou pour un seul travailleur. Le cas échéant, l'analyse consiste à estimer les paramètres d'une distribution d'expositions unique à partir desquels les indices d'exposition, comme la fraction de dépassement, peuvent être dérivés. La deuxième approche porte sur l'analyse dite des « différences inter-travailleur » et s'applique lorsqu'on dispose de mesures répétées pour certains travailleurs au sein d'un groupe. Elle permet de séparer la variabilité totale en composantes inter-travailleur et intra-travailleur, d'évaluer l'homogénéité de l'exposition au sein du groupe, et de déterminer si certains travailleurs individuels peuvent être à risque malgré une exposition de groupe acceptable. Comme cette dichotomie se reflète dans la section des résultats, nous avons jugé nécessaire de l'introduire ici.

4.1.1 Rétroaction du comité d'intervenants du Québec

La majorité des commentaires formulés par ce comité avaient trait à des recommandations relatives à la conception d'outils de calcul pratiques en HT plutôt qu'aux procédures d'estimation numériques comme telles ou aux indices d'exposition pertinents. Ces commentaires seront très utiles dans la prochaine phase de ce projet pour créer un outil à partir des prototypes développés dans la présente phase du projet. Ils ne sont toutefois pas pertinents à la sélection et à l'implémentation algorithmique des routines de calcul, de sorte que nous n'avons pas reproduit ici l'entièreté des notes de réunion. Nous devons néanmoins ajouter que le comité d'intervenants – tout comme le comité d'experts, d'ailleurs (voir ci-dessous) – a souligné l'importance de faciliter la communication des risques, que ce soit par l'entremise de résultats numériques faciles à comprendre, d'outils graphiques ou d'une documentation complète accessible aux non-spécialistes.

4.1.2 Rétroaction du comité d'experts international

Les notes finales de la réunion de deux jours tenue par le comité d'experts figurent à l'annexe A. Au cours de cette réunion, après une introduction générale, la liste proposée des indices de base décrits à la section 3.1 a été présentée aux participants aux fins de discussion. D'autres points de discussion à l'ordre du jour comprenaient le traitement des valeurs non détectées, l'erreur de mesure, l'utilisation d'a priori bayésiens informatifs dans les calculs, et la communication des risques. Les participants étaient libres d'ajouter tout autre point jugé pertinent.

Tant pour l'analyse de GES que pour l'analyse des différences inter-travailleur, le comité a confirmé l'intérêt d'estimer tous les indices dans la proposition initiale.

Le traitement des données censurées était jugé essentiel, mais pourrait se limiter aux données censurées à gauche (étant donné que la censure à droite et la censure par intervalle s'appliquent plus rarement aux données de mesure en HT).

L'inclusion d'une certaine forme d'erreur de mesure dans les calculs était jugée d'intérêt, mais plutôt sous forme d'option, car on estimait son impact négligeable dans la plupart des situations.

Le comité a exprimé peu d'intérêt pour les tests d'hypothèses formels visant à évaluer l'adéquation de la distribution lognormale. Il semble en effet qu'en deçà de 30 à 50 données de mesure – ce qui serait le cas dans la majorité des situations d'analyse de données en HT –, un test d'hypothèse, ou même une évaluation graphique de type Q-Q, ne fournit pas d'information utile sur la forme de la distribution.

Enfin, le comité a souligné l'importance de produire des extrants numériques accessibles aux non-spécialistes, et tout particulièrement susceptibles de refléter adéquatement l'incertitude liée aux analyses.

4.1.3 Élaboration d'un cadre d'interprétation de données probabiliste

La mention par les deux comités de l'importance de faciliter la communication de l'incertitude nous a conduits à concevoir une alternative à la seule utilisation des intervalles de confiance. Le cadre d'interprétation qui en résulte, décrit ci-dessous, vise essentiellement à fournir une réponse à la question suivante : « Quelle est la probabilité que cette situation donne lieu à une surexposition ? ».

L'évaluation de l'incertitude a traditionnellement fait appel à des intervalles de confiance et à des tests d'hypothèses. Par exemple, pour répondre à la question : « Même si l'estimation ponctuelle du 95^{e} centile (C95) pour un groupe de travailleurs est inférieure à la VLEP, comment pouvons-nous être sûrs que la valeur vraie est effectivement inférieure à la VLEP ? » Un test statistique commun pour répondre à cette question consisterait à formuler une hypothèse nulle telle que « le vrai 95^{e} centile est supérieur à la VLEP ». On pourrait alors effectuer le test et espérer pouvoir rejeter l'hypothèse nulle avec une faible erreur de type I. En guise d'alternative, on pourrait calculer une limite de confiance supérieure et espérer qu'elle soit inférieure à la VLEP. Une caractéristique commune à ces procédures est que leur résultat, selon un degré de confiance prédéfini, est binaire. Par exemple, lorsqu'il s'agit de calculer une limite de confiance supérieure à 90 % à l'égard du 95^{e} centile, soit cette valeur limite est inférieure à la VLEP, auquel cas nous pouvons être sûrs à 90 % que la valeur vraie est inférieure à la VLEP, soit elle est supérieure à la VLEP, auquel cas la conclusion est : « Nous ne pouvons pas démontrer avec un degré de confiance de 90 % que le vrai 95^{e} centile est inférieur à la VLEP. »

L'incertitude pourrait être exprimée sous une autre forme, plus directe. Par exemple, en calculant la probabilité que le vrai 95^{e} centile soit inférieur à la VLEP, laquelle devrait être élevée (> 90 % dans l'exemple ci-dessus), ou, au contraire, en calculant la probabilité que le vrai 95^{e} centile soit supérieur à la VLEP, laquelle devrait être faible (< 10 % dans l'exemple ci-dessus). Les énoncés qui en découlent sont informatifs dans les deux cas, et faciles à communiquer aux travailleurs ou aux employeurs, car ils fournissent une réponse directe à la question « Quelles sont les chances que l'exposition soit trop élevée ? »

L'analyse bayésienne permet naturellement de tels énoncés directs sur le degré d'incertitude inhérent aux conclusions qu'on peut tirer des données. Le cadre d'interprétation probabiliste implémenté dans le projet WebExpo comporte deux étapes menant à une estimation de la probabilité que l'exposition ne soit pas adéquatement maîtrisée, que nous avons choisi d'appeler « probabilité de surexposition » ou « risque de surexposition ».

Étape 1 – Définition d'une surexposition : quelle caractéristique de la distribution des expositions correspond à une situation inacceptable ?

À titre d'exemple, pour ce qui est de l'analyse de GES, la revue de la littérature présentée à la section 1 a fait ressortir trois définitions possibles d'une surexposition :

Fraction de dépassement ≥ 5 %
95^{e} centile ≥ VLEP
Moyenne arithmétique ≥ VLEP

Étape 2 – Analyse des données d'observation à l'aide des modèles bayésiens.

Au-delà des estimations ponctuelles de paramètres assorties d'intervalles de crédibilité, la probabilité que le critère de surexposition soit satisfait est estimée à partir des échantillons de la distribution a posteriori (p. ex. la probabilité que le vrai 95^{e} centile soit supérieur ou égal à la VLEP compte tenu des données). Cette quantité – le risque de surexposition – peut directement être utilisée comme intrant pour la gestion de l'exposition : le risque de surexposition est-il assez faible pour qu'on puisse considérer l'exposition bien maîtrisée, ou est-il assez élevé pour justifier une forme d'intervention (p. ex., envisager la mise en œuvre de mesures visant à maîtriser l'exposition) ?

Bien que le risque de surexposition fournisse toute l'information voulue sur l'incertitude, les gestionnaires du risque préfèrent souvent recevoir des résultats sous forme de recommandation – « Cette situation exige-t-elle une intervention, ou non ? ». La formulation d'une telle recommandation nécessite l'établissement d'un seuil de risque de surexposition ; la situation peut alors être dite maîtrisée de manière adéquate si le risque de surexposition est inférieur à la valeur définie, faute de quoi elle sera tenue pour mal maîtrisée. La valeur couramment attribuée au seuil de risque de surexposition est de 5 %, ce qui revient à dire que le risque de surexposition doit être inférieur à 5 % pour déclarer une situation acceptable (Jahn *et al.*, 2015).

Afin d'illustrer la correspondance entre cette forme d'expression de l'incertitude et d'autres approches plus traditionnelles, nous utiliserons l'exemple de C95 ≥ VLEP comme critère de surexposition. Un risque de surexposition inférieur à 5 % correspond à : « les chances que le vrai 95^{e} centile soit supérieur à la VLEP sont de moins de 5 % ». Cela signifie que nous sommes sûrs à au moins 95 % que le 95^{e} centile est inférieur à la VLEP. Ce qui est équivalent à l'énoncé suivant : « la limite de confiance supérieure à 95 % du 95^{e} centile est inférieure à la VLEP » (une formule plus traditionnelle). Les lignes directrices britannico-néerlandaises et européennes en vigueur de même que la réglementation française recommandent de comparer la limite de confiance supérieure à 70% de la fraction de dépassement à 5 %, ce qui correspond à un seuil de risque de surexposition de 30 % selon le critère de surexposition « fraction de dépassement ≥ 5 % » (BOHS-NVvA, 2011; CEN, 2018; T. Ogden et Lavoué, 2012; République française, 2009). Bien que les algorithmes de WebExpo ne comparent pas comme tel le risque de surexposition à un seuil déterminé, une telle comparaison est facile à réaliser en utilisant la

valeur du risque de surexposition fournie. La figure 2 ci-dessous illustre la correspondance entre l'utilisation traditionnelle de limites de confiance ou de crédibilité et le risque de surexposition.

L'analyse de l'incertitude faisant appel au calcul des variabilités inter-travailleur et intra-travailleur comporte une couche de complexité supplémentaire en ce qu'elle fournit des estimations de probabilité de surexposition individuelle ; par exemple, la probabilité qu'un travailleur au hasard présente une distribution d'expositions individuelles inacceptable. Un seuil a été proposé, selon lequel la probabilité de surexposition individuelle doit être inférieure à 20 % (BOHS-NVvA, 2011). Cependant, cette probabilité étant estimée, elle demeure incertaine. Par conséquent, plutôt que de simplement comparer l'estimation ponctuelle de la probabilité de surexposition individuelle à 20 %, on peut évaluer les chances que la valeur vraie soit supérieure ou égale à 20 %, soit les chances qu'une intervention soit requise. Un exemple type d'extrant serait « Compte tenu des données, on estime que la probabilité de surexposition individuelle (la surexposition étant définie comme C95 > VLEP) est de 12 % (IDC à 90 %; 6-50) ; les chances que la valeur vraie soit supérieure au seuil de 20 % sont de 25 %.

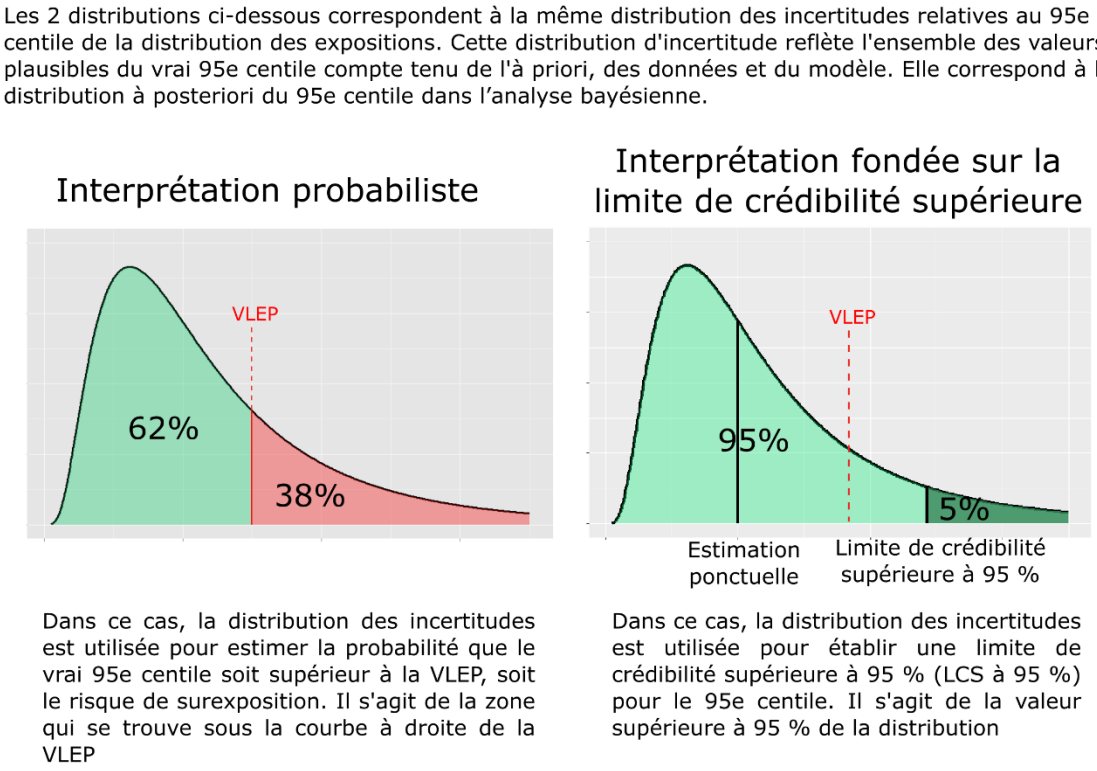


Figure 2. Illustration de la correspondance entre le risque de surexposition et les limites de crédibilité du 95^{e} centile.

4.1.4 Liste de calculs finale du projet WebExpo

Le tableau 1 présente un glossaire des termes et des indices utilisés dans le projet WebExpo. Le tableau 2 présente la liste des indices sélectionnés aux fins d'inclusion sur la base de la revue documentaire initiale et de la rétroaction des comités.

Le lecteur constatera que le tableau 2 n'affiche pas la probabilité de surexposition individuelle exprimée en fonction de la définition de la surexposition où la fraction de dépassement est supérieure au seuil de dépassement. Cela est dû au fait que cette quantité est égale à Prob.ind.overexpo.perc, pour autant que le centile retenu corresponde au seuil de dépassement (dans les valeurs par défaut : 95^{e} centile / seuil de dépassement de 5 %).

Tableau 1. Glossaire de termes

Fraction de dépassement	Proportion des niveaux d'exposition dans la population d'intérêt qui sont supérieurs à la limite d'exposition. Ou encore la probabilité qu'une seule valeur d'exposition au hasard soit supérieure à la VLEP.
95^{e} centile	Le 95^{e} centile d'une distribution est défini comme la valeur sous laquelle se trouve 95 % de la distribution.
Surexposition	Caractéristique d'une distribution d'expositions qui est inacceptable, c.-à-d. qui déclencherait une action préventive.
Seuil de dépassement	Proportion des niveaux d'exposition supérieurs à la VLEP utilisée comme seuil pour définir la surexposition (traditionnellement 5 %).
Centile critique	Centile de la distribution des expositions qui sera comparé à la VLEP pour évaluer la surexposition (traditionnellement le 95^{e} centile).
Risque de surexposition	Probabilité que le critère utilisé pour définir la surexposition soit satisfait (p. ex. 95^{e} centile ≥ VLEP). Dans la pratique : probabilité qu'une situation d'exposition soit inacceptable.
Seuil de risque de surexposition	Risque de surexposition maximal admissible. Cette valeur, définie a priori par l'utilisateur, est utilisée pour créer une dichotomie entre « exposition adéquatement maîtrisée » et « exposition mal maîtrisée » en fonction du risque de surexposition. Une valeur traditionnelle utilisée dans le domaine des statistiques serait de 5 %. La définition du respect de la VLEP dans les lignes directrices européennes correspond à un seuil de risque de surexposition de 30 %.
Probabilité de surexposition individuelle	Probabilité qu'un travailleur au hasard dans un groupe présente une distribution d'expositions individuelles correspondant à une surexposition (p. ex. probabilité que le 95^{e} centile de la distribution d'expositions individuelle d'un travailleur au hasard soit supérieur à la VLEP). Peut également s'énoncer comme suit : proportion des travailleurs dont la distribution individuelle d'expositions correspond à une surexposition.
Intervalle de crédibilité	Bien que l'équivalence ne soit pas tout à fait exacte, les intervalles de crédibilité bayésiens sont généralement interprétés d'une manière comparable aux intervalles de confiance traditionnels.
Rapport R	Le rapport R a été défini par Rappaport *et al.* comme le rapport du 97,5^{e} centile de la distribution des moyennes arithmétiques individuelles des travailleurs sur le 2,5^{e} centile de la même distribution.
Différence R	Définie en adaptant le rapport R à la distribution normale. Différence entre le 97,5^{e} centile de la distribution des moyennes arithmétiques individuelles des travailleurs et le 2,5^{e} centile de la même distribution, exprimée en pourcentage de la moyenne arithmétique de groupe.

Tableau 2. Indices d'exposition calculés pour la distribution lognormale dans le projet WebExpo

Analyse de GES
Estimation des paramètres distributionnels (estimation ponctuelle et intervalles de crédibilité)
Moyenne géométrique
Écart-type géométrique
Fraction de dépassement de la VLEP
Centile de la distribution des expositions (c.-à-d. centile critique, par défaut le 95^{e})
Moyenne arithmétique de la distribution des expositions
Décision relative à l'acceptabilité de l'exposition (risque de surexposition)
Probabilité que la fraction de dépassement soit supérieure ou égale au seuil de dépassement (par défaut 5 %)
Probabilité que le centile critique (par défaut le 95^{e}) soit supérieur ou égal à la VLEP
Probabilité que la moyenne arithmétique soit supérieure ou égale à la VLEP
Différences inter-travailleur*
Estimation des paramètres distributionnels (estimation ponctuelle et intervalles de crédibilité)
Moyenne géométrique de groupe
Écart-type géométrique intra-travailleur
Écart-type géométrique inter-travailleur
Coefficient de corrélation intra-travailleur (rho)
Probabilité que rho soit supérieur à un seuil prédéfini (Prob.rho.overX)
Rapport R (R.ratio)
Probabilité que R soit supérieur à 2 (Prob.R.over2, seuil servant à définir des groupes hétérogènes dans Kromhout *et al.*, 1993)
Probabilité que R soit supérieur à 10 (Prob.R.over10, seuil servant à définir des groupes très hétérogènes dans Kromhout *et al.*, 1993)
Paramètres permettant de quantifier la possibilité que certains travailleurs soient surexposés (probabilité de surexposition individuelle)
Proportion des travailleurs individuels dont le centile critique est supérieur à la VLEP (Prob.ind.overexpo.perc)
Proportion des travailleurs individuels dont la moyenne arithmétique est supérieure à la VLEP (Prob.ind.overexpo.am)
Probabilité que la valeur vraie de Prob.ind.overexpo.perc soit supérieure à un seuil donné (Prob.ind.overexpo.perc.overX, par défaut 20 %)
Probabilité que la valeur vraie de Prob.ind.overexpo.am soit supérieure à un seuil donné (Prob.ind.overexpo.am.overX, par défaut 20 %)
Paramètres personnalisables
Probabilité des intervalles de crédibilité (par défaut 90 %)
Seuil de dépassement (5 %)
Centile critique (par défaut le 95^{e})
Seuil du coefficient de corrélation intra-travailleur (par défaut 0,2)
Couverture de la population pour le rapport R (par défaut 80 %)
Seuil de probabilité de surexposition individuelle (par défaut 20 %)

* *De plus, pour tout travailleur individuel : tous les indices de l'analyse de GES*

Enfin, bien que l'estimation de la distribution de probabilités lognormale soit au cœur de l'interprétation des données d'hygiène du travail, la distribution normale est également utilisée dans certains cas (p. ex., bien que l'exposition à des substances chimiques – le centre d'intérêt du présent rapport – suit le plus souvent le modèle lognormal, les niveaux d'exposition au bruit exprimés en décibels présentent généralement une distribution normale). En outre, les deux formes de distribution sont étroitement liées, car si X suit une distribution normale avec une MG et un ÉTG, Y = ln(X) suit une distribution normale avec une moyenne ln(MG) et un écart-type ln(ÉTG). De ce fait, tous les modèles bayésiens ont été faciles à adapter au modèle normal, et offrent donc la possibilité d'analyser les données selon une distribution lognormale (l'option par défaut) ou normale. Dans les sections qui suivent, l'accent porte sur le modèle lognormal, mais une sous-section décrit brièvement l'option normale et ses particularités. Les résultats précisément liés au modèle normal sont présentés à l'annexe C. Le tableau C1 de cette annexe résume les indices calculés dans WebExpo en lien avec le modèle normal.

4.2 2e objectif spécifique : Création d'une bibliothèque de codes de programmation informatique

Une présentation mathématique détaillée des modèles et des algorithmes MCCM élaborés par l'équipe de McGill se trouve à l'annexe B.

4.2.1 Modèles bayésiens créés dans WebExpo – Analyse de GES (« SEG » [similar exposure group] *dans les noms de modèles)*

L'hypothèse principale qui sous-tend ce modèle est que le schéma d'exposition étudié est bien représenté par une distribution lognormale, et qu'un échantillon représentatif de cette distribution a été obtenu.

Soit X la variable aléatoire représentant les niveaux d'exposition.

Soit Y défini comme Y = ln(X). Y correspond donc aux niveaux d'exposition soumis à une transformation logarithmique.

Étant donné que X suit une distribution lognormale, Y suit une distribution normale, ce qui peut être exprimé sous la forme $Y \sim N(\mu,\sigma)$.

La moyenne géométrique de la distribution des expositions est définie par $MG = \exp(\mu)$.

L'écart-type géométrique est défini par $ÉTG = \exp(\sigma)$.

μ et σ représentent les paramètres inconnus du modèle.

4.2.1.1 Définition des distributions a priori (les « a priori »)

Comme μ et σ sont les paramètres d'intérêt dans ce modèle, il nous fallait définir des distributions a priori pour chacun d'eux.

Pour notre modèle de base [SEG.informedvar, section 3 de l'annexe B], nous avons, dans le cas de μ, opté pour une distribution a priori faiblement informative, sous la forme d'une distribution uniforme bornée comme elle est décrite, entre autres, dans Huynh *et al.* ou Banerjee *et al.* (S. Banerjee *et al.*, 2014; Huynh *et al.*, 2016). Dans le cas de σ, nous nous

sommes inspirés du modèle décrit par McNally *et al.* (McNally *et al.*, 2014), fondé sur la population des valeurs observées dans un jeu de données présenté par Kromhout *et al.* (Kromhout *et al.*, 1993) et Rappaport *et al.* (Rappaport *et al.*, 1993). Les auteurs ont dégagé des estimations de variabilité pour près de 200 groupes d'exposition. À partir du tableau A1 de Kromhout *et al.*, nous avons tabulé 165 valeurs de σ. L'évaluation graphique suggérait que la distribution de ces valeurs présentait une forme lognormale. L'ajustement des données à une distribution lognormale donne une MG de 0,84 pour σ (ce qui correspond à un ÉTG de 2,32 pour les niveaux d'exposition), et un ÉTG de 1,87 (cette quantité exprimant la variabilité des valeurs de sigma, et non des niveaux d'exposition). Cette distribution correspond à 95 % des valeurs d'ÉTG des niveaux d'exposition entre 1,3 et 17,6 – 70 % de la distribution se trouve entre 1,5 et 4,5. L'a priori de la variabilité dans le modèle [SEG.informedvar] est donc exprimé sous forme de distribution lognormale des ÉTG logtransformés de la distribution des expositions. Les valeurs par défaut sont celles ci-dessus, mais elles peuvent être modifiées par l'utilisateur.

Ce choix d'a priori maintient le niveau d'information a priori au plus bas dans le cas de la moyenne géométrique, tout en le rendant quelque peu informatif dans le cas de la variabilité sur la base des données historiques. Par souci de flexibilité, nous avons ajouté deux choix supplémentaires d'a priori pour ce modèle :

1- [SEG.uninformative, section 2 de l'annexe B] : Ce modèle utilise une distribution a priori uniforme pour σ comme pour μ, les plages étant définies par l'utilisateur. En définissant de larges plages, le modèle devient non informatif dans la pratique.

2- [SEG.riskband, section 5 de l'annexe B] : Ce modèle élargit l'approche de Hewett *et al.*, où les renseignements a priori reposent sur l'attribution de limites supérieure et inférieure à σ comme à μ, en assignant aussi des probabilités au fait que le 95^{e} centile se trouve dans les différentes bandes de gestion de l'exposition définies par l'AIHA (Jahn *et al.*, 2015). Ces bandes sont définies comme suit : < 0,01*VLEP; [0,01 * VLEP – 0,1 * VLEP]; [0,1 * VLEP – 0,5 * VLEP]; [0,5 * VLEP – VLEP]; et ≥ VLEP. La dernière bande correspond à une exposition inacceptable. Compte tenu de ces définitions, l'utilisateur doit attribuer une probabilité à chaque bande de sorte que la somme de ces probabilités totalise 1. L'attribution d'une probabilité de 0,2 aux cinq bandes engendre un a priori non informatif (dans la mesure où les plages de μ et σ sont raisonnablement larges), tandis que l'attribution d'une probabilité élevée à une ou plusieurs bandes rendra l'a priori d'autant plus informatif. L'attribution des probabilités peut s'appuyer sur un jugement expert, sur des modèles d'émission mathématiques ou sur d'autres jeux de données (Arnold, Stenzel, Drolet et Ramachandran, 2016; Jayjock, Chaisson, Franklin, Arnold et Price, 2009; P. Logan *et al.*, 2009; P. W. Logan *et al.*, 2011). Par souci de flexibilité, nous avons étendu la proposition de Hewett *et al.* à un nombre personnalisable de bandes et de limites de bandes.

Nous voulions également permettre aux utilisateurs d'enrichir les calculs en utilisant des distributions a priori basées sur d'autres données pertinentes, présentées sous forme de paramètres de synthèse (c.-à-d. moyenne, écart-type et taille d'échantillon). Quick *et al.* ont décrit de telles formes de distributions a priori dans *Annals of Occupational Hygiene* après que le projet WebExpo ait été lancé, de sorte que leur proposition n'aurait pu être intégrée sans ressources supplémentaires (Quick *et al.*, 2017). Nous avons cependant proposé une

modification au modèle [SEG.informedvar] à même de produire des résultats similaires. Les mathématiques sous-jacentes à la proposition sont décrites à l'annexe B. Les calculs effectués sont essentiellement équivalents à l'entrée par l'utilisateur d'un jeu de données dans le modèle [SEG.informedvar] comprenant toutes les données, les observations actuelles et le jeu de donnée représenté par les paramètres de synthèse. Les utilisateurs qui sélectionnent cette option doivent fournir la moyenne (dans la même échelle que μ), l'écart-type (dans la même échelle que σ) et la taille de l'échantillon. Le modèle correspondant est appelé [SEG.past.data].

4.2.1.2 Analyse des données censurées

Une façon de modéliser les données censurées dans l'analyse bayésienne consiste à les traiter comme des données manquantes devant se trouver dans la portion censurée de la distribution (Gelman, 2013). Une donnée censurée à gauche sous la limite de quantification (< LQ) est ainsi traitée comme une observation manquante dans la portion de la distribution qui se trouve sous la LQ. À chaque itération du processus MCCM, les valeurs manquantes sont imputées sous ces conditions. Cette contrainte influe sur la distribution a posteriori de la moyenne et de l'écart-type estimés par le modèle. Lorsque les a priori sont peu informatifs, la procédure s'apparente étroitement à la méthode de vraisemblance maximale fréquentiste. Nous avons appliqué cette approche aux données censurées à gauche, censurées par intervalle et censurées à droite (les deux derniers cas, bien que non prioritaires selon le groupe d'experts, sont de simples extensions du premier). Les points de censure peuvent en outre être spécifiques à chaque observation (c.-à-d. que plusieurs valeurs de LQ sont permises).

4.2.1.3 L'erreur de mesure

4.2.1.3.1 L'erreur de mesure exprimée sous forme d'écart-type

Le modèle d'erreur de mesure classique pour une quantité mesurée est généralement exprimé comme suit :

X_observé = X_vrai + erreur

En supposant qu'il n'y a aucun biais – que des fluctuations aléatoires autour de la valeur vraie – , le modèle traditionnel de la quantité « erreur » est une distribution normale avec une moyenne nulle et une erreur type σ_e fixe potentiellement inconnue. Cette structure d'erreur de mesure a été ajoutée comme option dans les modèles d'analyse de GES.

Le modèle bayésien serait donc défini comme suit :

$$\mathbf{X\_observé \sim N(X\_vrai, \sigma_e)} \quad \mathbf{(1)}$$

avec Y_vrai = ln(X_vrai)

$$\mathbf{Y\_vrai \sim N(\mu,\sigma)} \quad \mathbf{(2)}$$

σ_e, est considéré comme inconnu, avec une distribution a priori uniforme bornée. Si σ_e est présumé connu, l'utilisateur peut définir les limites inférieure et supérieure comme étant égales.

Dans la pratique, l'énoncé mathématique effectif du modèle présenté à l'annexe B est légèrement différent de celui ci-dessus. Comme les valeurs vraies sont censées suivre un

modèle lognormal, elles sont strictement positives. Cependant, l'équation 1 peut produire des valeurs négatives si l'écart-type est élevé par rapport à la valeur vraie. En conséquence, la distribution normale qui définit la valeur observée en fonction de la valeur vraie est tronquée de telle sorte que seules des valeurs observées positives puissent être générées (voir la section 6.1.1 de l'annexe B).

4.2.1.3.2 L'erreur de mesure exprimée sous forme de coefficient de variation

Comme mentionné précédemment, il est courant, en hygiène du travail, d'exprimer l'erreur de mesure en termes de CV, soit en considérant l'erreur proportionnelle au niveau d'exposition. Les CV vont généralement de quelques points de pourcentage, dans le cas d'une valeur moyenne pondérée sur huit heures issue de l'analyse chimique d'un tube adsorbant, à environ 30 % dans le cas des tubes colorimétriques de détection instantanée. Un CV constant sur un ensemble de mesures implique un écart-type différent pour chaque mesure. Un tel modèle a également été créé pour WebExpo, où il offre une seconde option de traitement de l'erreur de mesure.

$$\mathbf{X\_observé \sim N(X\_vrai, CV_e * X\_vrai)} \quad \textbf{(3)}$$

avec CV_e comme coefficient de variation exprimant l'erreur de mesure,

et avec Y_vrai = ln(X_vrai)

$$\mathbf{Y\_vrai \sim N(\mu,\sigma)} \quad \textbf{(4)}$$

Dans les modèles de WebExpo, tout comme pour σ_e, CV_e est traité comme une inconnue avec une distribution a priori uniforme bornée. Si CV_e est présumé connu sans la moindre incertitude, l'utilisateur peut définir les limites inférieure et supérieure comme étant égales.

4.2.1.4 Modification des modèles pour la distribution normale

Lorsque l'option de distribution normale est sélectionnée, les observations ne sont pas soumises à une transformation logarithmique préalable. Par conséquent, les paramètres μ et σ correspondent directement à la moyenne et à l'écart-type de la distribution sous-jacente.

Bien que le modèle normal accepte des valeurs négatives, le modèle de WebExpo pour la distribution normale est limité aux valeurs positives. Comme l'erreur de mesure peut être exprimée sous forme de CV, des valeurs négatives de X entraîneraient des valeurs négatives d'écart-type du terme d'erreur. Par conséquent, les utilisateurs désireux d'intégrer des données comportant des valeurs négatives ou presque nulles doivent transformer leurs données par l'ajout d'une constante positive avant l'analyse.

4.2.1.5 Interprétation des extrants du modèle bayésien

Comme mentionné à la section 4.1, l'extrant type de l'analyse bayésienne estimée par MCCM est un large échantillon de la distribution combinée a posteriori des paramètres inconnus, duquel découlent toutes les inférences. Ainsi, dans le cas de l'analyse de GES, l'extrant brut des algorithmes dont disposent les utilisateurs est, par exemple, 25 000 couples μ/σ. Nous

avons sur cette base appliqué un certain nombre d'équations afin d'estimer les indices décrits à la section 4.1.4.

La figure 3 illustre le flux de traitement des données dans l'analyse de la distribution lognormale. Les intrants comprennent les observations réelles, la valeur limite d'exposition professionnelle, les paramètres propres au modèle bayésien (choix et spécification de l'a priori, paramètres MCCM, choix et spécification de l'erreur de mesure) ainsi que les paramètres utilisés pour interpréter les échantillons de la distribution a posteriori. Aux fins du modèle lognormal, nous avons choisi de diviser les observations par la VLEP avant de les soumettre aux routines de traitement bayésiennes. Cette étape de normalisation fait en sorte que les quantités traitées dans le cadre des analyses bayésiennes – quels que soient l'état ou l'unité de départ – se trouveront à l'intérieur d'une plage plus ou moins centrée sur 1. Cette uniformité permet de proposer des limites inférieures et supérieures de μ qui conviennent à la plupart des situations.

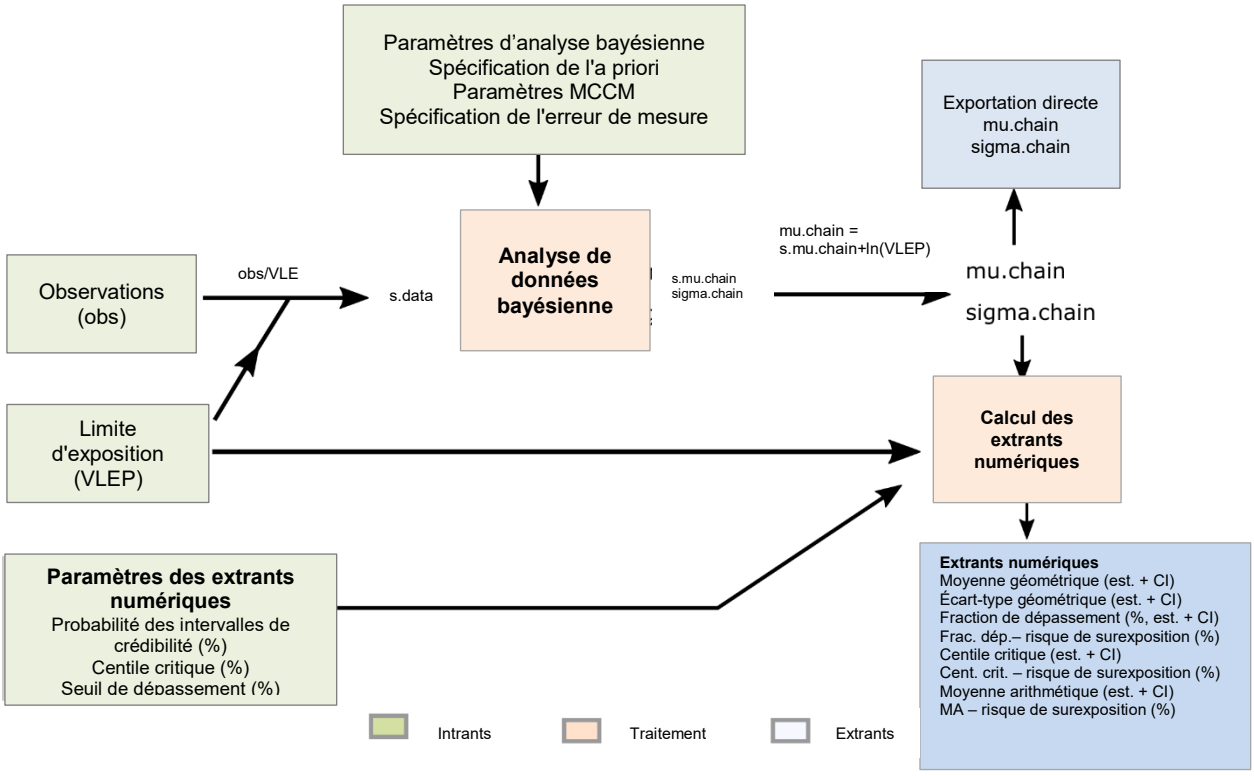


Figure 3. Flux de traitement des données pour les analyses de GES – Distribution lognormale.

Les indices de sortie comprennent la moyenne géométrique, l'écart-type géométrique, la fraction de dépassement de la VLEP, tout centile de la distribution (par défaut le 95^{e}) et la moyenne arithmétique, obtenue à partir des équations ci-dessous.

Moyenne géométrique de la distribution des expositions :

$$\boldsymbol{MG = \exp(\mu)} \quad \textbf{(5)}$$

Écart-type géométrique de la distribution des expositions :

$$\text{É}\boldsymbol{TG = \exp(\sigma)} \quad \textbf{(6)}$$

X^e centile de la distribution des expositions :

$$\boldsymbol{CX = \exp\{\mu + \Phi^{-1}(X) * \sigma\}} \quad \textbf{(7)}$$

où Φ^{-1} est la fonction de distribution cumulative inverse de la distribution normale standard.

Fraction de dépassement de la VLEP :

$$\boldsymbol{F(\%) = 100 * \left\{1 - \Phi\left(\frac{\ln(\text{VLEP}) - \mu}{\sigma}\right)\right\}} \quad \textbf{(8)}$$

où Φ est la fonction de distribution cumulative de la distribution normale standard.

Moyenne arithmétique de la distribution des expositions

$$\boldsymbol{MA = \exp\{\mu + 0.5 * \sigma^2\}} \quad \textbf{(9)}$$

L'incertitude relative aux indices précédents est caractérisée en les calculant pour toutes les valeurs combinées de μ et σ dans l'échantillon de la distribution a posteriori. Par exemple, l'équation 8, appliquée à l'échantillon de la distribution combinée a posteriori des μ et des σ, fournira 25 000 valeurs de fraction de dépassement, ce qui correspond à notre connaissance de ce paramètre compte tenu du modèle, de la distribution a priori et des observations. Ces valeurs définissent l'incertitude entourant le processus d'estimation. L'estimation ponctuelle de la fraction de dépassement correspondra à la médiane des 25 000 valeurs, et, par exemple, leur 5^e et leur 95^e centiles formeront un intervalle de crédibilité symétrique à 90 %. L'incertitude peut aussi être exprimée sous la forme de ce que nous avons défini comme étant le risque de surexposition à la section 4.1.3, auquel cas la proportion des 25 000 valeurs a posteriori qui dépasse le seuil de 5 % correspond à la probabilité que la fraction de dépassement vraie soit supérieure ou égale à 5 %.

4.2.1.6 Exemples

Illustrons les calculs d'analyse de GES à l'aide d'un jeu de données hypothétique provenant d'une distribution connue avec une moyenne géométrique MG_vraie = 30 et un écart-type géométrique ÉTG_vrai = 2. Le 95^{e} centile vrai de cette distribution est de 84, et sa moyenne arithmétique vraie est de 38. Avec une VLEP arbitraire fixée à 100, le schéma d'exposition réelle serait acceptable selon les actuelles définitions consensuelles de la surexposition.

Nous utiliserons un échantillon aléatoire de taille neuf (comme le recommande la récente ligne directrice de la communauté européenne) de cette distribution vraie pour appliquer les modèles bayésiens créés pour WebExpo. Ces chiffres peuvent représenter, à titre d'exemple, neuf concentrations moyennes pondérées de toluène mesurées pour un GES.

24,7 / 64,1 / 13,8 / 43,7 / 19,9 / 133 / 32,1 / 15 / 53,7

[échantillon 1 de l'annexe E]

Pour cet exemple, nous allons d'abord présumer l'absence d'erreur de mesure et effectuer les calculs avec le modèle [SEG.informedvar]. L'extrant brut des calculs bayésiens consiste en un échantillon de 25 000 valeurs de la distribution combinée a posteriori de μ et σ, tel que μ = ln(MG_vraie) et σ = ln(ÉTG_vrai). Dans la pratique, le modèle a été appliqué en utilisant les algorithmes de R+JAGS (voir la section 4.3) et les paramètres par défaut (voir l'annexe D).

La figure 4 présente les histogrammes des échantillons de la distribution a posteriori de μ et de σ. Ces histogrammes reflètent les connaissances acquises sur μ et σ compte tenu des données, des a priori et du modèle. Ils représentent notre estimation de l'incertitude entourant ces paramètres. Dans cet exemple, les valeurs les plus plausibles de μ se situent probablement entre 3 et 4 (bien que, comme le révèle l'histogramme, des valeurs plus extrêmes soient possibles). La médiane des 25 000 valeurs de μ de l'échantillon de la distribution a posteriori dans l'histogramme correspond à l'estimation ponctuelle de μ : 3,53. Des valeurs plausibles de σ se trouveraient entre 0,5 et 1,5, avec une estimation ponctuelle de 0,78.

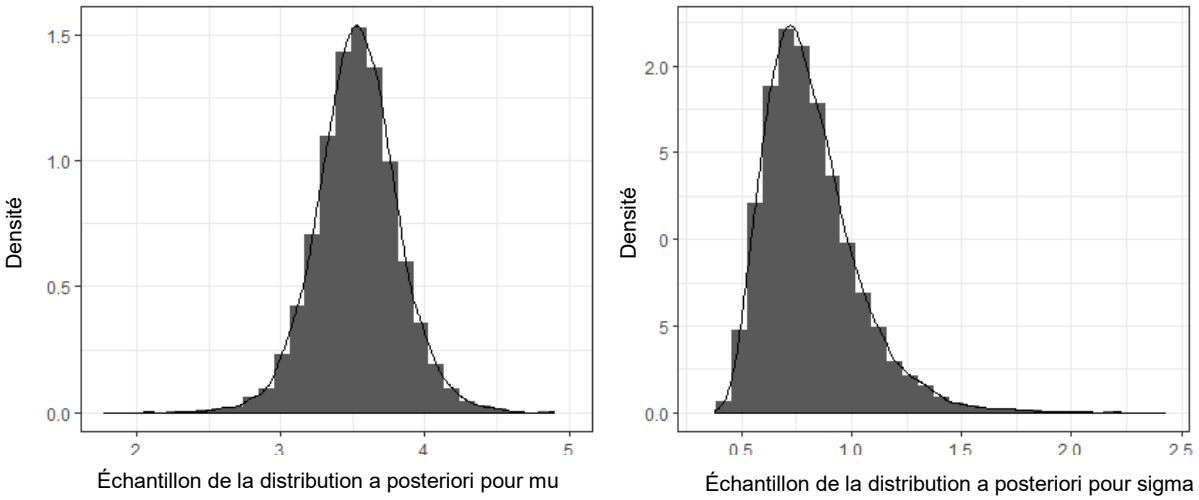


Figure 4. Échantillons de la distribution a posteriori pour la moyenne géométrique et l'écart-type logtransformés fournis par le modèle SEG.informedvar (modèle lognormal).

À partir des valeurs présentées dans la figure 4, il est facile d'appliquer les équations 5 à 9 afin d'obtenir des échantillons de la distribution a posteriori pour les différents paramètres d'intérêt. La figure 5 ci-dessous porte sur des échantillons de la distribution a posteriori pour le 95^{e} centile (équation 7) et la moyenne arithmétique (équation 9).

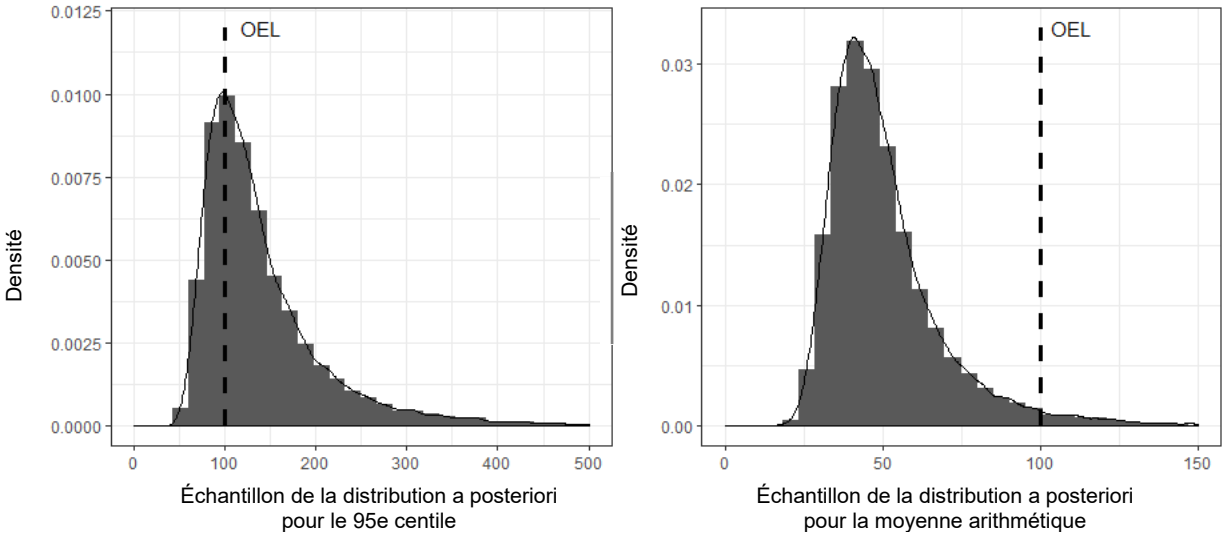


Figure 5. Échantillons de la distribution a posteriori pour le 95^{e} centile et la moyenne arithmétique calculés à partir des extrants du modèle SEG.informedvar (modèle lognormal).

Les histogrammes de la figure 5 illustrent l'incertitude entourant l'estimation de ces indices à partir d'un échantillon de taille neuf, étant donné que les valeurs des histogrammes couvrent une large plage. La notion de risque de surexposition est bien illustrée dans la figure 5 : en prenant pour critère de surexposition un 95^{e} centile supérieur ou égal à la VLEP (p. ex. VLEP = 100 µg/m^{3}), la probabilité (ou le risque) de surexposition est représentée par la portion de la surface de l'histogramme qui se trouve à droite du trait vertical correspondant à la VLEP, soit la proportion des valeurs de l'échantillon de la distribution a posteriori pour le 95^{e} centile qui dépassent la VLEP.

Le tableau 3 résume les résultats de l'interprétation des échantillons de la distribution a posteriori, y compris les estimations ponctuelles et les intervalles de crédibilité à 90 %, de même que le risque de surexposition (en fonction du 95^{e} centile et de la MA) et les probabilités relatives aux bandes de risque de l'AIHA, c'est-à-dire les probabilités que le 95^{e} centile vrai (ou la MA vraie) soit < 0,01 * VLEP, [0,01 * VLEP – 0,1 * VLEP], [0,1 * VLEP – 0,5 * VLEP], [0,5 * VLEP – VLEP] et ≥ VLEP, respectivement.

Pour illustrer les notions d'estimation ponctuelle et d'intervalle de crédibilité, le tableau 3 indique que la valeur la plus plausible en ce qui concerne la MG est de 34,2, avec une probabilité de 90 % que la valeur vraie se situe entre 21,6 et 54,5. Le tableau 3 indique également que la valeur la plus plausible en ce qui concerne la fraction de dépassement est de 8,29%, avec une probabilité de 90 % que la valeur vraie se situe entre 1,46 % et 26,5 %, et une probabilité de 71 % qu'elle soit supérieure à 5 % (risque de surexposition). Au regard des bandes de risque de

l'AIHA appliquées au 95^{e} centile, bien que l'estimation ponctuelle soit de 122 par rapport à une VLEP de 100, le tableau 3 indique que la probabilité que le 95^{e} centile vrai soit supérieur à la VLEP (100 µg/m^{3}) est de 71 %, que les chances qu'il se situe entre 0,5 * VLEP (50 µg/m^{3}) et la VLEP (100 µg/m^{3}) sont de 29 %, et que la probabilité qu'il se trouve dans une des autres catégories est inférieure à 1 %. Ainsi, en dépit d'un 95^{e} centile vrai inférieur à la VLEP, l'inférence découlant de l'échantillon disponible suggère qu'il existe une forte probabilité (71%) que la valeur vraie (que nous savons inférieure à la VLEP) soit supérieure à la VLEP. Cela illustre la difficulté à faire une inférence concluante à partir d'un échantillon de faible taille lorsque la situation réelle est acceptable, quoique de façon marginale. Le tableau 3 illustre également la différence entre la surexposition définie en fonction du 95^{e} centile (risque de surexposition de 71 %) par rapport à la moyenne arithmétique (risque de surexposition de 3,7 %).

Tableau 3. Estimations ponctuelles et intervalles de crédibilité des indices d'exposition dans un exemple de calcul bayésien selon le modèle lognormal

Paramètre	**Estimation ponctuelle et intervalle de crédibilité à 90 %**
MG	34,2 [21,6 - 54,5]
ÉTG	2,18 [1,72 - 3,37]
Fraction de dépassement (%)	8,29 [1,46 - 26,5] Risque de surexposition : 71 %
95^{e} centile	122 [72,1 - 303] Risque de surexposition : 71 %
Probabilités des bandes de l'AIHA en % (95^{e} centile)	0 / 0 / 0,048 / 29 / 71
Moyenne arithmétique	46,7 [30,4 - 91,6] Risque de surexposition : 3,7 %
Probabilités des bandes de l'AIHA en % (MA)	0 / 0 / 59 / 37 / 3,7

Afin d'illustrer l'influence du choix de la distribution a priori, nous avons analysé le même échantillon en utilisant les autres options de WebExpo en ce qui concerne les renseignements a priori. Nous avons inclus le modèle [*uninformative]*, le modèle [*riskband]* (semblable à celui proposé par Banerjee *et al.* et Hewett *et al.*), ainsi que le modèle *[past.data]*. Dans le cas du modèle informatif [*riskband]*, nous avons défini la connaissance a priori comme l'évaluation préalable d'un expert hypothétique jugeant la situation vraisemblablement acceptable, mais par une faible marge, d'où les choix de probabilités a priori suivants pour les bandes de l'AIHA : < 0,01 * VLEP (10 %); [0,01 * VLEP – 0,1 * VLEP] (20 %); [0,1 * VLEP – 0,5 * VLEP] (50 %); [0,5 * VLEP – VLEP] (10 %); et ≥ VLEP (10 %). Dans le cas du modèle *[past.data]*, nous présumerons l'existence préalable d'un ensemble de données de cinq mesures avec une moyenne géométrique de 5 et un écart-type géométrique de 2,4, jugés pertinents aux fins des présentes analyses. Pour ces différentes analyses, les paramètres autres que ceux mentionnés ci-dessus étaient les paramètres par défaut décrits à l'annexe D. Le tableau 4 en présente les résultats.

Tableau 4. Estimations ponctuelles et intervalles de crédibilité des indices d'exposition pour 4 choix de distribution a priori

Paramètre	*Informedvar*	*Uninformative*	*Past.data*	*Riskband*
MG (IDC à 90 %)	34,2 [21,7 - 54,1]	34,3 [20,9 - 56,8]	17,2 [9,91 - 29,7]	29,8 [19,1 - 46,1]
ÉTG (IDC à 90 %)	2,18 [1,73 - 3,38]	2,3 [1,75 - 4,15]	3,33 [2,49 - 5,45]	2 [1,66 - 3,19]
Fraction de dépassement (%) (IDC à 90 %)	8,30 [1,51 - 26,3]	9,77 [1,76 - 30,3]	7,16 [1,81 - 19,7]	3,71 [0,872 - 21,2]
95^{e} centile (IDC à 90 %)	122 [72,8 - 302]	134 [74,9 - 418]	124 [64 - 342]	90,8 [65,6 - 247]
Risque de surexposition (%, C95) (IDC à 90 %)	71 %	76 %	69 %	26 %
MA (IDC à 90 %)	46,6 [30,7 - 91,3]	49,1 [31,2 - 118]	35,9 [20,2 - 90,4]	37,6 [27,3 - 76,6]
Risque de surexposition (%, MA)	3,7 %	7,5 %	3,7 %	2,4 %

Le tableau 4 illustre l'influence de différents choix de renseignements a priori lorsque la taille de l'échantillon est relativement faible. Bien que les deux a priori peu informatifs (*informedvar* et *uninformative*) produisent des résultats semblables (mais non égaux), les deux a priori informatifs (*riskband* et *past.data*) ont une influence marquée. Ainsi l'application du modèle *past.data*, fondé sur un jeu de données comportant des niveaux d'exposition inférieurs à ceux de l'échantillon, a-t-elle eu pour effet de réduire l'estimation de la MG alors que celle de l'ÉTG s'en est trouvée augmentée (sans doute en raison des écarts de niveaux entre les deux ensembles de données). L'effet combiné de la réduction de la MG et de l'augmentation de l'ÉTG a relativement peu modifié la fraction de dépassement et le 95^{e} centile, mais a entraîné une réduction de la moyenne arithmétique. La distribution a priori de type *riskband*, qui impliquait des niveaux d'exposition plus faibles, a eu pour effet d'abaisser les valeurs estimées, et la réduction de la MG a entraîné une réduction du 95^{e} centile, de la fraction de dépassement, de la moyenne arithmétique et des valeurs de risque de surexposition connexes.

Afin d'illustrer l'impact de l'erreur de mesure, nous présentons ici l'analyse d'un échantillon provenant d'une distribution connue. Nous avons d'abord généré un échantillon de taille 100 à partir d'une distribution lognormale ayant une MG de 60 et un ÉTG de 1,5 [échantillon 2 de l'annexe E]. L'erreur de mesure a été ajoutée pour chaque point sous forme d'écart aléatoire par rapport à la valeur d'exposition vraie sous-jacente, selon un coefficient de variation de 30 %. Nous avons ensuite analysé cet échantillon suivant trois approches : une analyse présumant l'absence d'erreur de mesure, une analyse présumant une erreur de mesure connue comme étant de 30 %, et une analyse présumant une erreur de mesure inconnue, mais située entre 15 % et 45 %. Le tableau 5 montre les résultats de cette analyse.

Tableau 5. Estimations ponctuelles et intervalles de crédibilité des indices d'exposition au regard de l'erreur de mesure

Paramètre	Aucune erreur de mesure (A)	CV connu (30 %)	CV inconnu (15-45 %)
MG (IDC à 90 %)	56,9 [52,2 - 61,9]	59,8 [54,8 - 65,1]	58,9 [53,7 - 64,3]
ÉTG (IDC à 90 %)	1,68 [1,59 - 1,79]	1,49 [1,39 - 1,62]	1,55 [1,4 - 1,7]
Fraction de dépassement (%) (IDC à 90 %)	13,7 [9,62 - 18,8]	9,8 [5,27 - 15,8]	11,2 [5,82 - 17,1]
95^{e} centile (IDC à 90 %)	133 [118 - 153]	115 [101 - 135]	121 [103 - 142]
MA (IDC à 90 %)	65 [59,6 - 71,5]	64,8 [59,4 - 70,9]	64,8 [59,3 - 71]

(A) : Les données présentent en réalité des erreurs de mesure ; l'en-tête de colonne indique le type d'analyse appliqué aux données.

Le tableau 5 indique que le fait de ne pas tenir compte de l'erreur de mesure n'a eu que peu d'effet sur la MG, alors qu'il a entraîné une surestimation de l'ÉTG. Cette surestimation a eu un impact sur l'estimation de la partie supérieure de la distribution, entraînant une surestimation de la fraction de dépassement et du 95^{e} centile. L'impact s'est avéré moindre en ce qui concerne la moyenne arithmétique. Il convient de noter que l'intervalle de crédibilité de l'ÉTG dans le cas de l'analyse naïve ne contenait pas la valeur vraie, contrairement aux deux analyses intégrant l'erreur de mesure. Par rapport à l'analyse présumant un CV connu, la présomption d'un CV situé à l'intérieur d'une plage donnée n'a eu que peu d'effet dans cet exemple.

4.2.2 Modèles bayésiens créés dans WebExpo – Analyse des différences inter-travailleur

La principale hypothèse sur laquelle repose ce modèle est que le schéma d'exposition au sein d'un groupe est adéquatement représenté par une structure hiérarchique dans laquelle la distribution des expositions individuelles des travailleurs au sein du groupe est adéquatement représentée par une distribution lognormale. Les distributions des travailleurs diffèrent par leur tendance centrale (MG), mais possèdent la même variabilité. L'ensemble des MG propres aux travailleurs eux-mêmes suit une distribution lognormale.

Soit X une variable aléatoire représentant les niveaux d'exposition.

Soit Y = ln(X), où Y correspond dès lors aux niveaux d'exposition logtransformés.

Soit y_{ij} la valeur correspondant à la mesure prise le j^{e} jour pour la i^{e} personne.

Le modèle hiérarchique à effets aléatoires à un niveau s'énonce comme suit :

$$y_{ij} = \mu_y + b_i + e_{ij}$$

pour i = 1, 2, ..., k travailleurs le j = 1, 2, ..., n_i jours

μ_y est la moyenne du groupe, b_i est l'effet aléatoire pour le travailleur i, et e_{ij} est l'écart aléatoire le j^{e} jour par rapport à la moyenne $\mu_y + b_i$ du i^{e} travailleur.

Dans ce modèle à effets aléatoires, b_i et e_{ij} sont mutuellement indépendants et présentent des distributions normales dont les moyennes sont nulles. L'écart-type inter-travailleur (de b_i) est σ_b, et l'écart-type intra-travailleur (d'e_{ij}) est σ_w.

La moyenne géométrique de groupe est définie par $MG = \exp(\mu_y)$.

L'écart-type géométrique inter-travailleur est défini par $ÉTG_B = \exp(\sigma_B)$.

L'écart-type géométrique intra-travailleur est défini par $ÉTG_W = \exp(\sigma_w)$.

La distribution des expositions individuelles de tout travailleur est définie par :

$MG_i = \exp(\mu_y + b_i)$

et

$ÉTG_i = \exp(\sigma_w)$.

4.2.2.1 Définition des distributions a priori (les « a priori »)

Avec ce modèle (décrit à la section 4 de l'annexe B), nous devions définir une distribution a priori pour les trois paramètres μ_y, σ_b et σ_w.

Pour notre première analyse, nous avons défini des a priori similaires à ceux du modèle [SEG.informedvar] précédemment décrit. Dans le cas de μ_y, l'a priori était le même que pour le modèle [SEG.informedvar], soit une distribution uniforme dont les bornes étaient -20 et 20.

En ce qui concerne les paramètres de variabilité, nous avons utilisé les mêmes données publiées que pour le modèle [SEG.informedvar], mais avec les composantes inter-travailleur et intra-travailleur de la variance présentées dans l'article de Kromhout *et al.* plutôt que la variabilité totale. L'évaluation graphique suggérait également une forme lognormale pour la distribution des écarts-types inter-travailleur (σ_b) et intra-travailleur (σ_w). L'ajustement des données à des distributions lognormales a fourni les paramètres suivants :

Pour la variabilité inter-travailleur :

MG de la distribution lognormale pour σ_b : 0,415
ÉTG de la distribution lognormale pour σ_b : 2,18
Cette distribution correspond à 95 % des valeurs d'$ÉTG_B$ entre 1,1 et 6,7 ; 70 % de la distribution se trouve entre 1,2 et 2,5.

Pour la variabilité intra-travailleur :

MG de la distribution lognormale pour σ_w : 0,844
ÉTG de la distribution lognormale pour σ_w : 1,88
Cette distribution correspond à 95 % des valeurs d'$ÉTG_W$ entre 1,3 et 18,3 ; 70 % de la distribution se trouve entre 1,6 et 5,1.

Ce choix d'a priori maintient le niveau d'information a priori au plus bas dans le cas de la moyenne géométrique, tout en le rendant quelque peu informatif dans le cas de la variabilité sur

la base des données historiques. Ces distributions sont très semblables à celles décrites par McNally *et al.* (McNally *et al.*, 2014). Comme dans le cas de l'analyse de GES, les valeurs des paramètres distributionnels ci-dessus sont les valeurs proposées par défaut, mais elles peuvent être modifiées par l'utilisateur.

Nous avons ajouté un choix supplémentaire d'a priori pour ce modèle, selon lequel les distributions a priori pour σ_b et σ_w sont uniformes, leurs bornes étant définies par l'utilisateur. De larges plages correspondent à une distribution a priori non informative.

4.2.2.2 Analyse des données censurées

Dans le modèle des différences inter-travailleur, les données censurées sont traitées selon la même approche que celle décrite en 4.2.1.2. Plusieurs points de censure sont permis, et les données peuvent être censurées à gauche, à droite ou par intervalle. Les mathématiques sous-jacentes sont détaillées à l'annexe B.

4.2.2.3 Erreur de mesure

L'erreur de mesure pour ce modèle est traitée de la même façon que celle décrite en 4.2.1.3 et détaillée à l'annexe B.

4.2.2.4 Modification des modèles pour la distribution normale

Les modifications apportées aux modèles bayésiens pour l'analyse des différences inter-travailleur selon le modèle normal plutôt que lognormal sont les mêmes que celles décrites en 4.2.1.4.

4.2.2.5 Interprétation des extrants des modèles bayésiens

Dans le cas de l'analyse des différences inter-travailleur, l'extrant brut des algorithmes présenté aux utilisateurs a la forme de l'exemple suivant : 25 000 valeurs combinées de μ_y / σ_b / σ_w / b_i (i = 1 à k travailleurs). Nous avons ensuite appliqué au résultat obtenu plusieurs équations afin d'estimer les indices décrits à la section 4.1.4 (voir ci-dessous).

La figure 6 illustre le flux de traitement des données dans l'analyse de la distribution lognormale. Les intrants comprennent les observations réelles assorties d'un identifiant de travailleur, la valeur limite d'exposition professionnelle, les paramètres propres au modèle bayésien (choix et spécification de l'a priori, paramètres MCCM, choix et spécification de l'erreur de mesure) ainsi que les paramètres utilisés pour interpréter les échantillons de la distribution a posteriori. Comme pour l'analyse de GES du modèle lognormal, nous avons choisi de diviser les observations par la VLEP avant de les soumettre aux routines de traitement bayésiennes.

Les équations suivantes décrivent la façon dont les différents indices présentés à la section 4.1.4 sont calculés à partir de l'extrant MCCM.

Moyenne géométrique de groupe :

$$\boldsymbol{MG_{groupe} = \exp(\mu_Y)} \quad \textbf{(10)}$$

Écart-type géométrique inter-travailleur :

$$\boldsymbol{\acute{E}TG_b = \exp(\sigma_b)} \quad \textbf{(11)}$$

Écart-type géométrique intra-travailleur :

$$\boldsymbol{\acute{E}TG_w = \exp(\sigma_w)} \quad \textbf{(12)}$$

Coefficient de corrélation intra-travailleur :

$$\boldsymbol{\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}} \quad \textbf{(13)}$$

Rapport $R_{X\%}$: plage de modification renfermant le X % central de la distribution des moyennes géométriques, des moyennes arithmétiques ou des centiles propres à des travailleurs donnés. Au départ, X a été fixé à 95 par les premiers partisans de son utilisation (Kromhout et al., 1993; Rappaport et al., 1993). Nous proposons plutôt une valeur par défaut de 80 %.

$$\boldsymbol{R_{X\%} = \exp\left(2 * \Phi^{-1}\left(\frac{1+X}{2}\right) * \sigma_b\right)} \quad \textbf{(14)}$$

Probabilité qu'un seul travailleur au hasard affiche une moyenne arithmétique supérieure à la VLEP :

$$\boldsymbol{P_{ind}^{MA}(\%) = 100 * \left\{1 - \Phi\left(\frac{\ln(\mathrm{VLEP}) - \left(\mu_Y + 0,5 * \sigma_w^2\right)}{\sigma_b}\right)\right\}} \quad \textbf{(15)}$$

Probabilité qu'un seul travailleur au hasard affiche un X^e centile supérieur à la VLEP (ce qui équivaut à la probabilité qu'un seul travailleur au hasard présente un degré de dépassement de la VLEP supérieur à (100-X) % :

$$\boldsymbol{P_{ind}^{PX}(\%) = 100 * \left\{1 - \Phi\left(\frac{\ln(\mathrm{VLEP}) - \left(\mu_Y + \Phi^{-1}(X) * \sigma_w\right)}{\sigma_b}\right)\right\}} \quad \textbf{(16)}$$

En plus de ce qui précède, il est possible d'obtenir des indices spécifiques à toute distribution d'expositions individuelle. En conséquence, par définition, la distribution des expositions pour un travailleur *i* est définie par :

Moyenne géométrique de la distribution des expositions :

$$\boldsymbol{MG = \exp(\mu_Y + b_i)} \quad \textbf{(17)}$$

Écart-type géométrique de la distribution des expositions :

$$\boldsymbol{\acute{E}TG = \exp(\sigma_w)} \quad \textbf{(18)}$$

X^e centile de la distribution des expositions :

$$\boldsymbol{PX = \exp\{\mu_Y + b_i + \Phi^{-1}(X) * \sigma_w\}} \quad \textbf{(19)}$$

où Φ^{-1} est la fonction de distribution cumulative inverse de la distribution normale standard.

Fraction de dépassement de la VLEP :

$$\boldsymbol{F(\%) = 100 * \left\{1 - \Phi\left(\frac{\ln(VLEP) - \mu_Y - b_i}{\sigma_w}\right)\right\}} \quad \textbf{(20)}$$

où Φ est la fonction de distribution cumulative de la distribution normale standard.

Moyenne arithmétique de la distribution des expositions :

$$\boldsymbol{MA = \exp\{\mu_Y + b_i + 0,5 * \sigma_w^2\}} \quad \textbf{(21)}$$

Il convient de noter que les indices propres aux travailleurs ci-dessus, bien qu'applicables à un seul travailleur, sont estimés par ajustement du modèle bayésien à l'ensemble des données, et non seulement aux données du travailleur *i*.

Comme pour l'analyse de GES, l'incertitude relative aux indices précédents est caractérisée en les calculant pour toutes les valeurs combinées de μ, σ_B, σ_w et b_i dans l'échantillon de la distribution a posteriori.

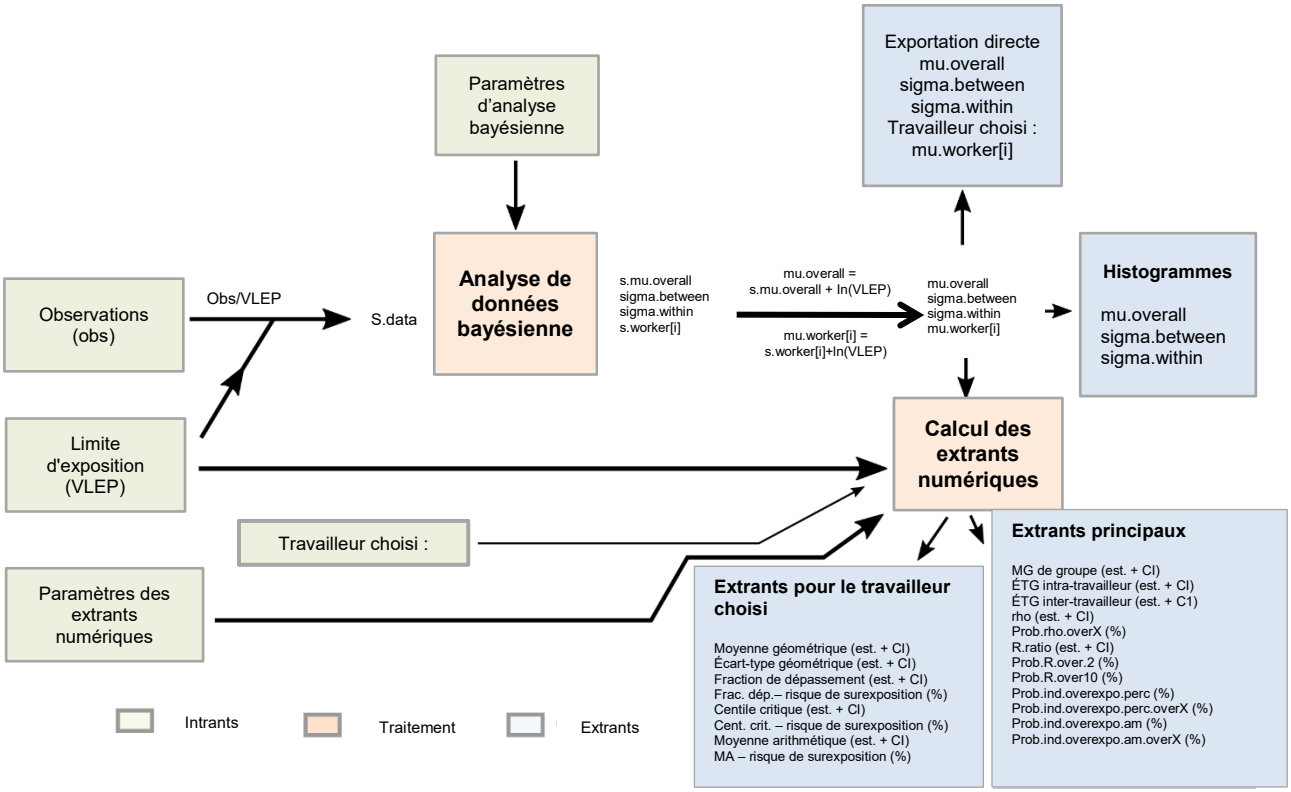


Figure 6. Flux de traitement des données pour les analyses de différences inter-travailleur – Distribution lognormale.

4.2.2.6 Exemples

Nous allons illustrer l'analyse des différences inter-travailleur à partir de deux échantillons fictifs tirés de distributions connues. La première distribution présente une MG de 30, un ÉTG global de 2,5 et une faible corrélation intra-travailleur (rho = 0,06 – cette valeur correspondant au 25^{e} centile de la distribution des valeurs de rho dans le jeu de données de Kromhout *et al.* décrit en 4.2.2.1 et 4.2.1.1). La seconde distribution affiche la même MG et le même ÉTG, mais une forte corrélation intra-travailleur (rho = 0,66 – cette valeur correspondant au 75^{e} centile de la distribution des valeurs de rho dans le jeu de données de Kromhout *et al.*)

Nous avons tiré de chaque distribution un échantillon de 100 observations, soit 10 observations par travailleur provenant de 10 travailleurs différents. L'analyse repose sur une VLEP de 150, légèrement plus élevée que le 95^{e} centile théorique du groupe (C95 vrai du groupe = 135). Les deux échantillons sont présentés à l'annexe E [échantillon 3 et échantillon 4, respectivement].

Pour cet exemple, nous avons présumé l'absence d'erreur de mesure et effectué les calculs selon le modèle [between-worker differences.informedvar] implémenté en R + RJAGS (voir 4.3).

L'extrant brut des calculs bayésiens consiste en un échantillon de 50 000 valeurs de la distribution combinée a posteriori des μ_y / σ_b / σ_w / b_i (i = 1 à k travailleurs). L'interprétation de ces valeurs est semblable à celle décrite en 4.2.1.6. Le tableau 6 résume les résultats de

l'interprétation des échantillons de la distribution a posteriori, y compris les estimations ponctuelles, les intervalles de crédibilité à 90 % et les indices de risque de surexposition.

Tableau 6. Estimations ponctuelles et intervalles de crédibilité des indices d'exposition dans un exemple de calcul bayésien selon le modèle lognormal (analyses de différences inter-travailleur)

Paramètre	Faible corrélation intra-travailleur (rho = 0,06)	Forte corrélation intra-travailleur (rho = 0,66)
MG de groupe (IDC à 90 %)	28,6 [23,6 - 34,7]	28,3 [18,7 - 43,4]
ÉTG inter-travailleur (IDC à 90 %)	1,24 [1,09 - 1,54]	2,15 [1,72 - 3,18]
ÉTG intra-travailleur (IDC à 90 %)	2,34 [2,14 - 2,62]	1,74 [1,64 - 1,88]
Corrélation intra-travailleur (rho) (IDC à 90 %)	0,06 [0,00908 - 0,206]	0,654 [0,47 - 0,818]
Probabilité que rho soit supérieur à 0,2	5,5 %	100 %
R.ratio (IDC à 90 %)	1,74 [1,24 - 3,01]	7,09 [3,99 - 19,4]
Probabilité que R soit supérieur à 2	30 %	100 %
Probabilité que R soit supérieur à 10	0 %	25 %
Probabilité de surexposition individuelle (95^{e} centile) en % (IDC à 90 %)	12,1 [0,0248 - 52,4]	16,3 [4,94 - 36,5]
Chances que la probabilité ci-dessus soit supérieure à 20 %	34 %	36 %
Probabilité de surexposition individuelle (moyenne arithmétique) en % (IDC à 90 %)	9,5e-08 [0 - 0,177]	2,38 [0,164 - 13,1]
Chances que la probabilité ci-dessus soit supérieure à 20 %	0 %	1,4 %

Le tableau 6 indique que, pour les deux échantillons – de taille relativement élevée et de variabilité respectivement faible et élevée –, l'analyse fournit des résultats très proches de la distribution théorique en termes de MG de groupe et de valeurs de rho. Les ÉTG de groupe (non inclus dans le tableau 6) sont également proches de la valeur théorique de 2,5, les estimations ponctuelles étant respectivement de 2,4 et 2,6 pour les échantillons à faible et forte corrélation intra-travailleur.

Dans le cas de l'échantillon à faible corrélation intra-travailleur, la faible corrélation avec un ÉTG de groupe de 2,5 donne un faible ÉTG inter-travailleur (1,24), comme en témoigne le faible rapport R (1,7), avec 70 % de chances que la valeur vraie soit inférieure à 2, un seuil initialement proposé par Kromhout *et al.* (1993) pour définir « homogène ». La variabilité inter-travailleur étant faible, la majeure partie de la variabilité survient intra-travailleur (ÉTG = 2,3). Le phénomène inverse se produit dans le second échantillon, où l'ÉTG inter-travailleur est de 2,15, ce qui correspond à un rapport R de 7, avec 100 % de chances que la valeur vraie soit supérieure au seuil de 2, et 25 % de chances qu'elle soit supérieure à 10. Dans ce cas, la majeure partie de la variabilité totale survient entre les travailleurs, avec un faible ÉTG intra-travailleur correspondant (1,7).

Pour illustrer la notion de surexposition individuelle, prenons le cas de l'échantillon à faible corrélation intra-travailleur, pour lequel le tableau 6 indique que la probabilité que le 95^{e} centile

d'un travailleur au hasard soit supérieur à la VLEP est estimée à 12,1 % (IDC à 90 %; 0,02 % - 52,4 %). Les chances que la valeur vraie de cette probabilité soit supérieure à 20 % – le critère utilisé par la NVvA et la BOHS – sont de 34 %. De même, la probabilité que la moyenne arithmétique d'un travailleur au hasard soit supérieure à la VLEP est estimée à ~0 % (IDC à 90 %; 0,0 % - 0,2 %). Les chances que la valeur vraie de cette probabilité soit supérieure à 20 % – le critère utilisé par la NVvA et la BOHS – sont de 0 %.

Le fait que la probabilité de surexposition individuelle (tant pour le 95^{e} centile [près de 15 %] que pour la moyenne arithmétique [près de 0 %]) soit similaire dans les deux échantillons en dépit d'importants écarts de variabilité inter-travailleur est digne de mention. Il s'agit là d'une conséquence de la variabilité globale partagée par les deux groupes. Dans le cas du premier échantillon, malgré le peu de différences entre les travailleurs en termes de MG (selon la mesure du rapport R), la forte variabilité (intra-travailleur) au jour le jour implique la possibilité de valeurs relativement élevées du 95^{e} centile ou de la MA (qui dépendent tous deux de σ_w), qui seront similaire entre les travailleurs. Dans le cas du deuxième échantillon, malgré d'importantes différences entre les travailleurs en termes de MG (selon la mesure du rapport R), la faible variabilité (intra-travailleur) au jour le jour implique des valeurs relativement faibles du 95^{e} centile ou de la MA, mais qui pourront grandement varier d'un travailleur à l'autre, avec quelques valeurs élevées. Pour illustrer ce point, nous présentons dans le tableau 7 les indices d'exposition propres aux travailleurs les moins exposés et les plus exposés (en termes de MG) des deux échantillons.

Le tableau 7 montre clairement les importantes différences qui existent entre les deux échantillons sous l'angle des MG calculées pour les travailleurs les moins exposés et les plus exposés : MG_{moins} = 26 et MG_{plus} = 33 pour l'échantillon à faible corrélation intra-travailleur, et MG_{moins} = 7 et MG_{plus} = 130 pour l'échantillon à forte corrélation intra-travailleur. Les estimations ponctuelles du 95^{e} centile indiquent que tous les travailleurs du premier échantillon ont des 95^{e} centiles comparables avec, dans tous les cas (toujours sans tenir compte de l'incertitude), une distribution d'expositions acceptable, quoique plutôt marginalement (95^{e} centile autour de 100-140 pour une VLEP de 150). Dans le cas de l'autre échantillon, les estimations du 95^{e} centile propre aux différents travailleurs varient entre 18 (ce qui est très faible par rapport à la VLEP, une situation manifestement acceptable) et 325 (soit plus du double de la VLEP, une situation manifestement inacceptable). Ces contrastes appuient la proposition de Kromhout *et al.* et Rappaport *et al.* selon laquelle l'utilisation de ce type de modèle peut être utile pour axer les mesures de prévention sur le plan collectif plutôt qu'individuel (Kromhout *et al.*, 1993; Pesch *et al.*, 2015).

Tableau 7. Estimations ponctuelles et intervalles de crédibilité des indices d'exposition propres aux travailleurs les moins exposés et les plus exposés dans deux échantillons respectivement à faible et à forte corrélation intra-travailleur

Paramètre	Faible corrélation intra-travailleur (rho = 0,06)		Forte corrélation intra-travailleur (rho = 0,66)	
	Travailleur le moins exposé (MG)	Travailleur le plus exposé (MG)	Travailleur le moins exposé (MG)	Travailleur le plus exposé (MG)
MG (IDC à 90 %)	26,2 [18,7 - 34,7]	33,4 [25,1 - 48,4]	7,13 [5,32 - 9,53]	130 [97,3 - 174]
ÉTG (IDC à 90 %)	2,34 [2,14 - 2,62]	2,34 [2,14 - 2,62]	1,74 [1,64 - 1,88]	1,74 [1,64 - 1,88]
Fraction de dépassement (%) (IDC à 90 %)	2,02 [0,57 - 5,06]	3,97 [1,57 - 9,53]	0 [0 - 0]	40,1 [22,1 - 60,6]
95^{e} centile (IDC à 90 %)	106 [73,6 - 151]	137 [99,5 - 202]	17,8 [13,1 - 24,7]	325 [243 - 445]
MA (IDC à 90 %)	37,8 [26,7 - 51,2]	48,3 [36,1 - 70,1]	8,32 [6,21 - 11,2]	152 [114 - 204]

Le tableau 8 illustre l'analyse d'un échantillon présentant la même MG et le même ÉTG de groupe que ci-dessus, mais une corrélation intra-travailleur moyenne (rho = 0,22, ce qui correspond à la valeur médiane dans la base de données de Kromhout *et al.*) estimée à partir d'un échantillon de taille réaliste selon la ligne directrice de la BOHS et de la NVvA (n = 12, avec quatre mesures répétées sur trois travailleurs) [échantillon 5 de l'annexe E].

Tableau 8. Estimations ponctuelles et intervalles de crédibilité des indices d'exposition dans un exemple de calcul bayésien selon le modèle lognormal (analyses de différences inter-travailleur) avec un échantillon de taille réaliste

Paramètre	Estimation ponctuelle et intervalle de crédibilité à 90 %
MG de groupe (IDC à 90 %)	30,9 [16,9 - 56,4]
ÉTG inter-travailleur (IDC à 90 %)	1,4 [1,11 - 2,56]
ÉTG intra-travailleur (IDC à 90 %)	2,31 [1,84 - 3,42]
Corrélation intra-travailleur (rho) (IDC à 90 %)	0,135 [0,0121 - 0,577]
Probabilité que rho soit supérieur à 0,2	37 %
R.ratio (IDC à 90 %)	2,35 [1,29 - 11,1]
Probabilité que R soit supérieur à 2	62 %
Probabilité que R soit supérieur à 10	5,8 %
Probabilité de surexposition individuelle (95^{e} centile) en % (IDC à 90 %)	29,7 [0,0114 - 99,2]
Chances que la probabilité ci-dessus soit supérieure à 20 %	59 %
Probabilité de surexposition individuelle (moyenne arithmétique) en % (IDC à 90 %)	0,0204 [0 - 21,3]
Chances que la probabilité ci-dessus soit supérieure à 20 %	5,4 %

Les résultats du tableau 8 révèlent l'importante incertitude liée aux paramètres estimés à partir du modèle des différences inter-travailleur avec les tailles d'échantillons actuellement proposées. De fait, l'intervalle de crédibilité à 90 % de R renferme aussi bien des valeurs considérées comme indicatives d'une exposition homogène (< 2) que des valeurs considérées comme indicatives d'une exposition très hétérogène (> 10). Cela se reflète également dans l'incertitude entourant la probabilité de surexposition individuelle (selon le critère du 95^{e} centile supérieur à la VLEP), estimée à 29,7 % pour cet échantillon, avec un intervalle de crédibilité à 90% allant de 0,01 à 99,2 %.

4.3 Les algorithmes de WebExpo

Le principal livrable du projet WebExpo tient à la mise à disposition publique des algorithmes qui permettent l'exécution des analyses bayésiennes et numériques décrites dans le présent rapport. Les algorithmes sont tous accessibles à https://github.com/webexpo/ et sous licence libre Appache 2.0[15]. Ce rapport tient lieu de documentation d'accompagnement.

Comme l'indique la section 3.3.2, les modèles et l'interprétation numérique des échantillons de la distribution a posteriori ont d'abord été écrits dans le langage de R. Les modèles comme tels ont été codés de deux façons. Ils ont d'abord été écrits en code pur de R, sans appel à un progiciel externe, pour permettre leur traduction en C# et en JavaScript. Comme mentionné en 3.2.2.1, ils ont également été codés à l'aide de l'application JAGS pour permettre aux utilisateurs de R d'effectuer les calculs avec une efficacité optimale.

Dans un deuxième temps, le code pur de R a été traduit en C# et en JavaScript. Dès les premiers efforts de traduction, il s'est avéré que les ressources nécessaires pour traduire la composante de l'erreur de mesure du code pur de R seraient très coûteuses, et que son implémentation intégrale se ferait au détriment d'autres composantes du projet (p. ex. moins d'options de distributions a priori). Nous avons alors choisi de définir les priorités en matière de traduction selon les objectifs suivants :

1. faire en sorte que les analyses jugées essentielles pour les modèles de GES et de différences inter-travailleur soient implémentées dans tous les langages pour les modèles normal et lognormal ;

2. faire en sorte que l'erreur de mesure soit implémentée dans au moins un langage autre que celui de R.

[15] https://www.apache.org/licenses/LICENSE-2.0

En conséquence, l'erreur de mesure n'a été implémentée qu'en langage C#, où elle n'est exprimée que sous forme de CV (tenu pour le plus utile). Le traitement des données censurées à droite, à gauche et par intervalle a été inclus dans tous les langages. Le tableau 9 présente en détail les diverses composantes implémentées dans chacun des quatre langages.

Tableau 9. Diverses composantes implémentées dans chacun des quatre langages

	Pas d'erreur de mesure				**Erreur de mesure sous forme de CV**				**Erreur de mesure sous forme d'ÉT**			
	R	R+JAGS	C#	Java Script	R	R+JAGS	C#	Java Script	R	R+JAGS	C#	Java Script
Analyse de GES												
Informedvar	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-
Uninformative	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-
Past.data	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-
Riskband	✓	✓	-	-	✓	✓	-	-	✓	✓	-	-
Différences inter-travailleur												
Informedvar	✓	✓	✓	✓	✓	✓	-	-	✓	✓	-	-
Uninformative	✓	✓	✓	✓	✓	✓	-	-	✓	✓	-	-

4.3.1 Organisation des scripts

Les scripts en R de WebExpo (pur R et R+JAGS) sont regroupés en sous-sections, comme suit :

Scripts de formatage de données – Ces scripts servent à formater les observations avant de les soumettre aux fonctions de calcul bayésiennes.

Scripts de modèles bayésiens – C'est ici que se trouve la plus grande partie de la bibliothèque créée. Les algorithmes correspondants servent à l'échantillonnage MCCM, et leurs extrants comprennent les échantillons des distributions a posteriori.

Scripts d'interprétation des extrants bayésiens – Les scripts de cette catégorie appliquent les équations 5 à 21 aux échantillons des distributions a posteriori de manière à obtenir les indices d'exposition décrits dans les sections correspondantes.

Les scripts en R comprennent aussi des fonctions de génération de données servant à produire des échantillons aléatoires sur mesure correspondant aux différents modèles implémentés dans WebExpo. Enfin, la bibliothèque de codes en R comprend également des scripts servant à reproduire tous les résultats numériques présentés dans le présent rapport.

Tout en utilisant la même architecture globale, les bibliothèques de codes en C# et en JavaScript sont organisées quelque peu différemment en raison de la nature de ces langages de programmation. À l'instar des exemples de la bibliothèque en R, les bibliothèques en C# et en JavaScript comprennent des algorithmes explicatifs servant à guider les utilisateurs à travers les différents modèles bayésiens. Enfin, les bibliothèques en C# et en JavaScript comprennent également le code correspondant aux prototypes, lequel offre une interface de saisie de données et une interface de présentation de résultats numériques.

4.3.2 Paramètres de calcul

Les algorithmes de WebExpo ont besoin d'un ensemble d'intrants pour effectuer les calculs MCCM et fournir des indices d'exposition interprétés assortis de l'incertitude connexe. Ces intrants peuvent être séparés en trois grandes catégories : observations, paramètres bayésiens et paramètres d'interprétation numériques.

4.3.2.1 Observations

Cette catégorie regroupe les valeurs réelles saisies par l'utilisateur. En théorie, il n'y a pas de limite inférieure au nombre d'observations fournies par les utilisateurs, puisque l'exécution des calculs bayésiens en l'absence de données fera simplement en sorte que les échantillons des distributions a posteriori reproduisent les distributions a priori. Nous recommandons cependant un seuil pragmatique d'au moins trois observations non censurées (p. ex. six mesures de concentration, dont trois valeurs non détectées). Il n'y a pas non plus de limite supérieure théorique au nombre d'observations soumises pour analyse, bien que l'utilisation de la mémoire risque à certains points de dépasser la capacité de l'ordinateur. À titre d'exemple, nous avons analysé un ensemble de données de 100 000 observations à l'aide du script en R+RJAGS selon le modèle lognormal d'analyse de GES sur un ordinateur de bureau ordinaire en un peu plus de 120 minutes. En ce qui concerne les valeurs d'échantillonnage comme telles, les seules restrictions sont qu'elles soient strictement positives et dans les limites définies pour les distributions a priori des modèles bayésiens. Pour ce qui est du modèle lognormal, nous avons effectué une division par la VLEP avant de lancer les calculs bayésiens, ce qui a permis de limiter la plage de valeurs effectivement analysées, tout en nous permettant de proposer des valeurs « universelles » par défaut pour les distributions lognormales a priori (p. ex. : limites de la distribution uniforme pour la moyenne). Nous avons également inclus dans cette première catégorie d'intrants la valeur limite d'exposition professionnelle, qui doit être exprimée selon la même échelle que les observations.

4.3.2.2 Paramètres bayésiens

Cette série de paramètres peut se diviser en trois parties. L'utilisateur doit tout d'abord définir un modèle d'analyse. Il doit à cette fin choisir entre une distribution normale ou lognormale, de même qu'entre une analyse de GES et une analyse des différences inter-travailleur, et déterminer si l'erreur de mesure doit être prise en compte ou non.

Le deuxième jeu d'intrants bayésiens porte sur les paramètres des distributions a priori sélectionnées. Le tableau 10 ci-dessous présente une liste sommaire de ces paramètres, tandis que l'annexe D en présente la liste complète et détaillée, assortie de valeurs par défaut et des plages recommandées.

Le troisième ensemble de paramètres consiste en intrants relatifs aux procédures d'échantillonnage MCCM. Ils comprennent le nombre d'itérations de rodage (itérations rejetées au terme de l'échantillonnage, utilisées pour permettre à la procédure MCCM de dégager une distribution stationnaire), le nombre d'itérations retenues (c.-à d. la taille de l'échantillon de distribution a posteriori), ainsi que les valeurs initiales à attribuer aux paramètres à estimer. Nous recommandons d'utiliser 25 000 itérations pour 2500 itérations de rodage dans le cas des modèles d'analyse de GES, et 50 000 itérations pour 5000 itérations de rodage dans le cas des modèles d'analyse des différences inter-travailleur. Les valeurs de départ doivent être

plausibles pour les paramètres d'intérêt, c'est-à-dire dans les limites définies pour les distributions a priori.

Tableau 10. Paramètres servant à définir les distributions a priori dans les modèles de WebExpo

Modèle	Paramètres a priori
SEG.informedvar	Bornes de la distribution uniforme pour μ
	Paramètres distributionnels de la distribution lognormale pour σ
SEG.past.data	Bornes de la distribution uniforme pour μ
	Paramètres distributionnels de la distribution lognormale pour σ
	Moyenne, écart-type et taille d'échantillon du jeu de données externe
SEG.uninformative	Bornes de la distribution uniforme pour μ
	Bornes de la distribution uniforme pour σ
SEG.riskband	Bornes de la distribution uniforme pour μ
	Bornes de la distribution uniforme pour σ
	Nombre de bandes de risque et limites connexes
	Probabilité a priori associée à chaque bande
Between-worker.informedvar	Bornes de la distribution uniforme pour μ
	Paramètres distributionnels de la distribution lognormale pour σ_w et σ_b
Between-worker.uninformative	Bornes de la distribution uniforme pour μ
	Bornes de la distribution uniforme pour σ_w et σ_b

Remarque : pour la plupart des paramètres, l'échelle dépend du choix de distribution (normale/lognormale) ; par exemple, dans le cas de SEG.past.data, la moyenne fournie est la moyenne arithmétique des observations pour le modèle normal, mais la moyenne géométrique logtransformée et normalisée en fonction de la VLEP pour le modèle lognormal.

4.3.2.3 Paramètres d'interprétation numériques

Les paramètres d'interprétation numériques n'ont aucune influence sur les calculs bayésiens en soi ; ils sont utilisés pour transformer l'information contenue dans les échantillons de distribution a posteriori en indices pertinents à l'évaluation des risques et pour exprimer l'incertitude entourant leur estimation. Ils comprennent notamment :

- la probabilité des intervalles de crédibilité (par défaut 90 %) ;
- le seuil de dépassement (par défaut 5 %) – seuil définissant la proportion acceptable de niveaux d'exposition supérieurs à la VLEP ;
- le centile critique (par défaut le 95^{e}) – centile d'intérêt dans la distribution des expositions ;
- les paramètres spécifiques à l'analyse des différences inter-travailleur :
    - le seuil du coefficient de corrélation intra-travailleur (par défaut 0,2) – la ligne directrice de la BOHS recommande de procéder à une évaluation détaillée des différences inter-travailleur lorsque l'estimation ponctuelle de ce coefficient est supérieure à 0,2 ;

- la couverture de la population pour le rapport R (par défaut 80 %) – la proposition initiale de Kromhout *et al.* était de 95 % (Kromhout *et al.*, 1993). Cela correspondrait plus ou moins à la comparaison des travailleurs les plus exposés et les moins exposés dans une population de 100. Notre proposition moins rigoureuse correspondrait plutôt à la comparaison des travailleurs les plus exposés et les moins exposés dans une population de 10, un chiffre plus représentatif des groupes d'exposition dans la pratique quotidienne en HT ;

- le seuil de probabilité d'une surexposition individuelle (par défaut 20 %) – la ligne directrice de la BOHS et de la NVvA recommande de juger une situation d'exposition non conforme lorsque la probabilité d'une surexposition individuelle (calculée en fonction du 95^{e} centile) est supérieure à 20 %. Rappaport *et al.* ont suggéré d'utiliser un seuil de 10 % et de considérer la probabilité d'une surexposition individuelle définie en fonction de la moyenne arithmétique (Rappaport *et al.*, 1995).

4.3.3 Performance

4.3.3.1 Précision numérique

Comme mentionné à la section 3.2.2.2, nous n'avons pas effectué de simulations pour évaluer la précision des estimations produites par nos modèles ni celle de nos procédures d'estimation, car elles se fondent sur des études solidement documentées. Nous avons toutefois mis l'accent sur la vérification de la reproductibilité des résultats entre les différentes plateformes (R, R+RJAGS, JavaScript, C#).

C#

Afin de tester les modèles d'analyse de GES en C#, nous avons utilisé un échantillon standard pour chacun des 24 scénarios définis par des combinaisons des caractéristiques suivantes :

- taille d'échantillon de 10 ou 100 ;
- faible variabilité (ÉTG = 1,5 pour une distribution lognormale, ÉT = 2 pour une distribution normale) ou variabilité élevée (ÉTG = 3,5 pour une distribution lognormale, ÉT = 15 pour une distribution normale) ;
- aucune censure, censure faible (15 % pour n = 100, 30 % pour n = 10) ou censure élevée (50 %) ;
- modèle normal ou lognormal.

Les échantillons ont été ajustés à l'aide des algorithmes en C# et en R pour chacune des routines suivantes : SEG.informedvar, SEG.past.data et SEG.uninformative.

Pour l'analyse des différences inter-travailleur, une approche similaire a été utilisée, suivant laquelle C# et R ont été comparés par rapport à 24 scénarios :

- 5 travailleurs à raison de 3 mesures par travailleur, 10 travailleurs à raison de 5 mesures par travailleur, et 20 travailleurs à raison de 20 mesures par travailleur ;
- faible corrélation intra-travailleur (rho = 0,2) ou forte corrélation intra-travailleur (rho = 0,8) ;
- aucune censure ou censure à 50 % ;
- modèle lognormal ou normal.

Enfin, pour le module d'erreur de mesure des fonctions SEG.informedvar et SEG.uninformative, 12 scénarios ont été testés :

- taille d'échantillon de 5, 10 ou 100 ;
- aucune censure ou censure à 60 % ;
- modèle lognormal ou normal.

Dans le cas de la distribution lognormale, l'erreur de mesure a été définie comme étant inconnue entre 20 % et 30 %. Dans le cas de la distribution normale, elle a été définie comme étant inconnue entre 0,1 % et 1,2 %.

À travers tous ces essais, les écarts entre C# et R, calculés à chacune des 25 000 itérations pour les paramètres inconnus de μ et σ, étaient en moyenne de l'ordre d'environ 10^{-15}, avec des valeurs maximales de l'ordre d'environ 10^{-10}. Les écarts entre C# et R étaient dans tous les cas moins importants de plusieurs ordres de grandeur que les écarts observés à l'intérieur de chaque plateforme (p. ex. en répétant une analyse avec R ou C# à partir d'une autre valeur de départ aléatoire).

JavaScript

Lors de la traduction du code pur de R en JavaScript, nous avons comparé les quantiles des échantillons de distributions a posteriori en ce qui a trait aux paramètres inconnus. Dans le cas des fonctions SEG.informedvar, SEG.past.data et SEG.uninformative, les 4 scénarios suivants ont été testés avec un échantillon standard de taille 100 – aucune censure ou censure à 60 % ; distribution lognormale ou normale.

Dans le cas des fonctions between worker.informedvar et between worker.uninformative, les 8 échantillons testés se présentaient comme suit : 10 travailleurs à raison de 5 mesures par travailleur et 20 travailleurs à raison de 20 mesures par travailleur ; aucune censure ou censure à 60 % ; distribution lognormale ou normale.

Au cours de ces essais, les écarts entre JavaScript et R, calculés pour tous les paramètres inconnus à chacun des 9 centiles des chaînes MCCM, étaient en moyenne de l'ordre d'environ 10^{-15}, avec des valeurs maximales de l'ordre d'environ 10^{-10}. Une fois de plus, les écarts entre JavaScript et R étaient dans tous les cas moins importants de plusieurs ordres de grandeur que les écarts observés à l'intérieur de chaque plateforme (p. ex. en répétant une analyse avec R ou JavaScript).

R+JAGS

Quant aux écarts entre R et RJAGS, les fonctions suivantes ont été testées en utilisant le protocole décrit à la section 3.2.2, pour les modèles aussi bien lognormal que normal : SEG.informedvar avec erreur de mesure (exprimée sous la forme d'une valeur de CV connue), SEG.past.data, SEG.uninformative, SEG.riskband, between worker.informedvar et between worker.uninformative (avec et sans erreur de mesure). Ces essais ont démontré que R et R+RJAGS produisaient des résultats raisonnablement comparables. À titre d'illustration, le tableau 11 présente les résultats de l'analyse d'un échantillon lognormal de taille 9 (< 25,7 / 17,1 / 168 / 85,3 / 66,4 / < 49,8 / 33,2 / < 24,4 / 38,3 [échantillon 6 de l'annexe E]) avec 30 % de données censurées et une forte variabilité (ÉTG vrai = 2,5). L'analyse a été réalisée 50 fois avec R et R+RJAGS, et le tableau affiche les valeurs minimales et maximales observées pour 10 paramètres. Y figurent aussi les résultats d'un essai avec C# et d'un essai avec JavaScript.

Le tableau 11 donne un aperçu de la variabilité attendue lors de telles analyses, et confirme que les variations sont peu significatives par rapport à l'incertitude entourant les estimations ponctuelles. On s'attend à ce que les variations les plus importantes touchent les limites de crédibilité, qui correspondent aux valeurs extrêmes des échantillons de distributions a posteriori, ainsi que les quantités assorties d'une grande incertitude (comme la fraction de dépassement et le 95^{e} centile, liés à la queue de la distribution des expositions).

Tableau 11. Comparabilité des résultats entre les plateformes

Paramètre	R+JAGS (min) (a)	R+JAGS (max) (b)	R (min) (a)	R (max) (a)	C#	JavaScript
Estimation ponctuelle de la MG	34,1	34,7	34,0	34,6	34,3	34,4
LCI à 95 % de la MG	16,3	17,2	16,5	17,0	16,8	16,6
LCS à 95 % de la MG	60,1	61,9	60,2	61,6	60,8	61,0
Estimation ponctuelle de l'ÉTG	2,63	2,69	2,63	2,68	2,66	2,66
LCI à 95 % de l'ÉTG	1,88	1,91	1,89	1,90	1,90	1,90
LCS à 95 % de l'ÉTG	5,21	5,56	5,21	5,47	5,32	5,34
Estimation ponctuelle de la fraction de dépassement (%)	13,2	13,5	13,2	13,5	13,3	13,5
LCI à 95 % de la fraction de dépassement (%)	3,33	3,53	3,34	3,51	3,41	3,37
LCS à 95 % de la fraction de dépassement (%)	32,6	33,9	33,0	34,0	33,5	33,5
Fraction de dépassement relative au risque de surexposition (%)	88,8	89,9	88,9	89,6	89,1	89,1

(a) : valeurs minimales dans les 50 analyses ; (b) : valeurs maximales dans les 50 analyses ; LCI : limite de confiance inférieure ; LCS : limite de confiance supérieure.

4.3.3.2 Vitesse de calcul

La vitesse de calcul des algorithmes de ce projet dépend de plusieurs facteurs. En ce qui concerne l'analyse bayésienne de type MCCM, une taille d'échantillon et un nombre d'itérations élevés augmenteront le temps de calcul. La prise en compte de l'erreur de mesure et la censure de données augmenteront également le temps de calcul, étant donné qu'elles suscitent la génération d'un plus grand nombre de valeurs aléatoires. Au-delà de ces considérations,

différentes plateformes pourraient être mieux adaptées ou optimisées pour le traitement des algorithmes, et la puissance de l'ordinateur demeure un facteur déterminant. Nous n'avons pas mené d'enquête pour évaluer le temps de calcul dans un large éventail de situations. Cependant, compte tenu de ce que nous estimons couvrir la plupart des situations ($n < 50$ et aucune erreur de mesure), toutes les plateformes utilisées dans WebExpo devraient effectuer les calculs (avec 25 000 ou 50 000 itérations, selon le modèle) instantanément ou en quelques secondes. La prise en compte de l'erreur de mesure est le facteur le plus susceptible d'influer sur la vitesse de calcul de nos algorithmes : la plateforme la plus rapide, R+RJAGS, affiche un temps de calcul de moins d'une minute, suivie de C# (jusqu'à 10 minutes) et de R (de 30 minutes pour de petits échantillons à plusieurs heures pour le modèle d'analyse plus complexe des différences inter-travailleur et des tailles d'échantillons plus élevées).

4.4 3e objectif spécifique : Les prototypes de WebExpo

Les prototypes en C# et en JavaScript sont disponibles à https://github.com/webexpo/. Ces deux prototypes intègrent les modèles de WebExpo décrits dans le tableau 9. En conséquence, contrairement aux plateformes R et R+JAGS, qui intègrent toutes les possibilités décrites dans le présent rapport, le prototype en C# n'accepte qu'un type d'erreur de mesure (exprimée en CV) pour les modèles d'analyse de GES, et il ne comprend pas le modèle des bandes de risque. Quant au prototype en JavaScript, il ne comprend ni le traitement de l'erreur de mesure ni le modèle des bandes de risque. Les deux prototypes produisent les mêmes extrants numériques, inclusion faite de tous les indices énumérés dans le tableau 2, ainsi que les échantillons des distributions a posteriori pour les utilisateurs qui souhaitent effectuer divers calculs post-MCCM. Dans le cas de l'analyse des différences inter-travailleur, les deux prototypes permettent également de sélectionner un travailleur et d'obtenir des indices d'exposition propres à ce travailleur, ainsi que l'échantillon de la distribution a posteriori des moyennes propres au travailleur en question. Enfin, les deux prototypes sont disponibles en français et en anglais, et comportent une infrastructure multilingue facilitant la traduction dans d'autres langues. Les figures 7 et 8 présentent les interfaces utilisateur des prototypes en C# et en JavaScript, respectivement.

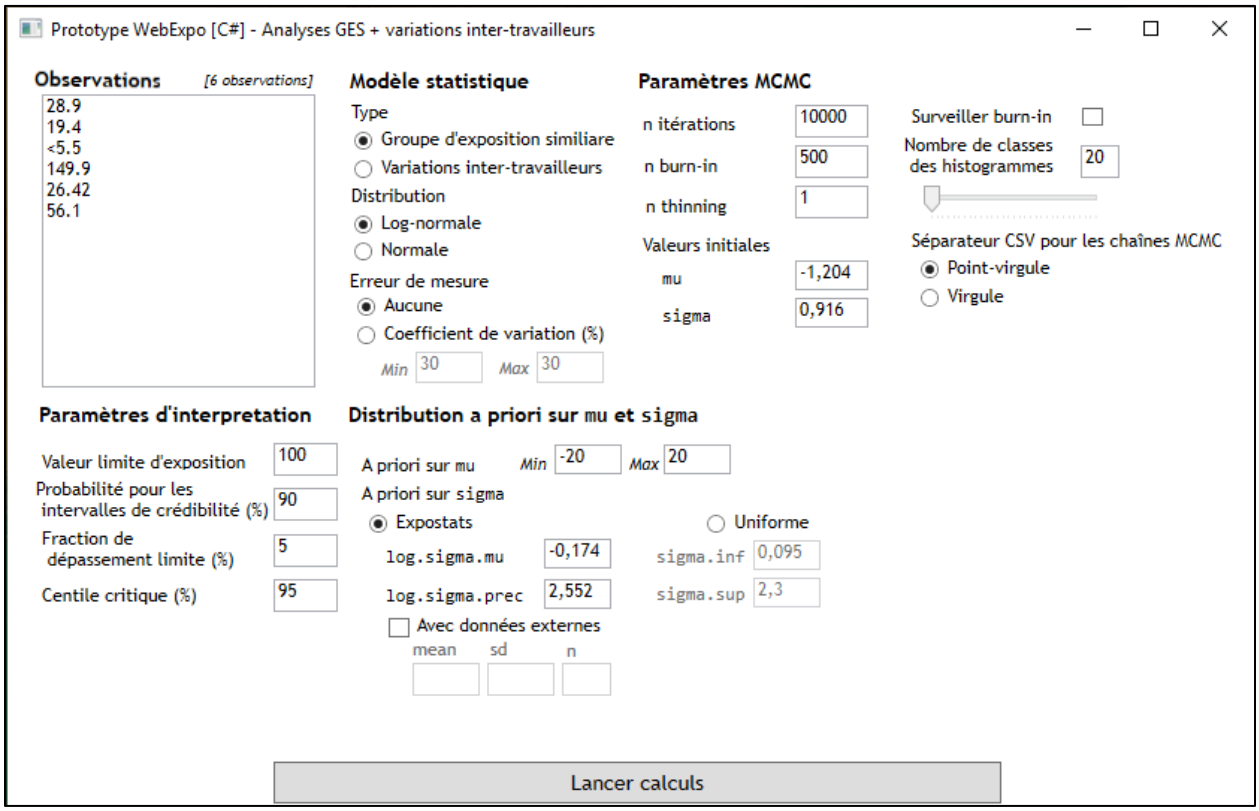


Figure 7. Interface utilisateur du prototype de WebExpo en C#.

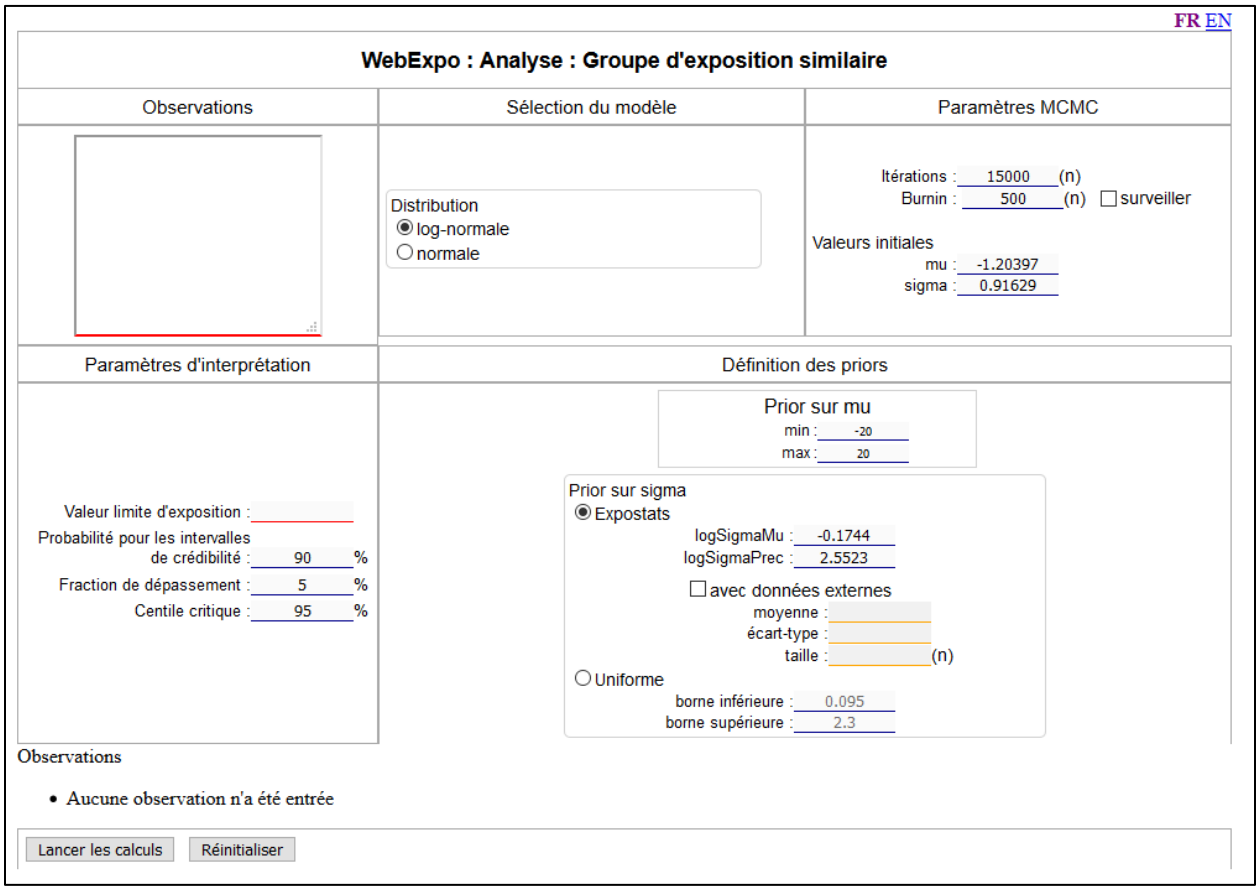


Figure 8. Interface utilisateur du prototype de WebExpo en JavaScript.

5. DISCUSSION

5.1 Aperçu général

Malgré l'existence d'un cadre consensuel concernant l'analyse des données de mesure en hygiène du travail – cadre ayant évolué jusqu'à la dernière décennie et reflété dans les récentes directives européennes (CEN, 2018) –, les intervenants manquent d'outils pour effectuer les calculs connexes, plutôt complexes, surtout en ce qui a trait aux données comportant des valeurs de concentration non détectées, comme il s'en présente fréquemment dans notre domaine.

Le projet WebExpo visait à créer une liste de calculs à même de fournir des réponses détaillées aux questions d'interprétation des données d'HT à partir d'un ensemble de lignes directrices actuelles ; à produire des algorithmes permettant d'implémenter ces calculs dans un cadre méthodologique unifié ; et à partager ces algorithmes dans différents langages pour faciliter leur utilisation dans la création d'outils pratiques. Nous avons atteint ces objectifs en dressant une liste de calculs fondée sur l'examen des lignes directrices en vigueur et des publications récentes, suivi d'une consultation auprès d'un groupe d'experts en évaluation de l'exposition à l'échelle internationale ; en implémentant les calculs à l'aide de statistiques bayésiennes ; et en créant des algorithmes dans des langages statistiques et de programmation dont les capacités sont illustrées par deux prototypes d'outil.

Les algorithmes de WebExpo permettent d'analyser des données lognormales ou normales pour estimer aussi bien une distribution unique que les composantes de la variance intra-travailleur et inter-travailleur lorsque des mesures répétées sont disponibles sur certains individus. Dans chacun des cas, plusieurs types de renseignements a priori peuvent être utilisés, y compris de l'information provenant de données externes pertinentes et de jugements experts. Le traitement des données censurées (à gauche, à droite ou par intervalle) est parfaitement intégré dans tous les calculs, et l'erreur de mesure peut être prise en compte dans l'analyse. Enfin, outre les traditionnels intervalles de confiance, la nature des statistiques bayésiennes permet d'exprimer l'incertitude sous forme d'énoncés probabilistes.

5.2 Choix d'a priori bayésiens

Les modèles de WebExpo offrent plusieurs choix de distributions a priori selon le type d'analyse à effectuer. Ces modèles peuvent essentiellement se résumer comme suit : les modèles *informedvar*, où il y a peu d'information a priori sur la moyenne, mais où la variabilité est minimalement fondée sur des données historiques ; les modèles *uninformative*, où les distributions a priori sont uniformes et assorties de larges plages (personnalisables) ; et les modèles *riskband* et *past.data*, qui peuvent être très informatifs. L'approche bayésienne traditionnelle recommande d'évaluer la robustesse à travers un éventail d'a priori différents afin d'élargir l'interprétation d'une analyse (Gelman, 2013), puisqu'elle s'appliquera alors à une plus grande variété d'interprétations. Compte tenu des tailles d'échantillons réalistes dans notre domaine (5 à 10 observations), des a priori informatifs ont généralement un effet non négligeable sur les estimations finales de l'exposition (Jones et Burstyn, 2017) par rapport aux a priori non informatifs. Les intervenants qui analysent des données d'hygiène du travail à partir de différentes distributions a priori (informatives ou non informatives) risquent donc de se retrouver aux prises avec des résultats très différents selon les a priori utilisés, ce qui ne manquerait pas de justifier des efforts d'évaluation plus poussés. Soulignons toutefois qu'il s'agit

là d'une force plutôt que d'une faiblesse de l'approche bayésienne, et qu'il en ressort clairement que la taille des échantillons recueillis pour évaluer l'exposition en milieu de travail est souvent insuffisante. Des échantillons de taille appropriée permettraient de tirer de solides conclusions, quels que soient les a priori utilisés, ce qui est rarement le cas dans ce domaine.

Nous recommandons l'utilisation des a priori *informedvar* comme distributions par défaut dans la bibliothèque de WebExpo, car ils offrent un compromis raisonnable, avec un a priori non informatif pour la moyenne géométrique, mais un a priori modérément informatif quant à la variabilité, fondé sur une population de valeurs disponibles en matière de variabilité sur les lieux de travail, comme celui utilisé par McNally *et al.* et récemment préconisé par Jones et Burstyn (Jones et Burstyn, 2017; McNally *et al.*, 2014).

5.3 Atouts

À notre connaissance, les algorithmes de WebExpo couvrent la liste la plus complète de calculs jugés pertinents pour l'interprétation des données d'hygiène du travail selon les meilleures pratiques actuelles. Bien que tous les calculs proposés dans nos algorithmes ne soient pas nécessairement d'intérêt pour tous, le fait que WebExpo soit d'exploitation libre devrait permettre la création d'applications adaptées à des besoins plus ciblés. En outre, la contribution majeure de ce projet étant son moteur bayésien et la création d'échantillons de distributions a posteriori par simulation MCCM, tout traitement ultérieur de ces données en dehors des calculs des tableaux 2 et 3 est simple et direct.

En dépit du fait que le traitement des valeurs non détectées présente depuis longtemps un défi de taille dans l'interprétation des données d'HT, les percées récentes à cet égard (p. ex. Huynh *et al.* et Krishnamoorty et Mathews [Huynh *et al.*, 2016; Krishnamoorthy *et al.*, 2009]) n'ont généralement pas été intégrées à des outils pratiques. IHSTAT, sans doute l'outil d'analyse de données le plus utilisé en HT, n'admet pas de données censurées, quoique Lavoué ait récemment proposé un outil permettant d'implémenter la régression sur les statistiques d'ordre dans IHSTAT (J. Lavoué, 2013). HYGINIST limite la régression sur les statistiques d'ordre à un seul point de censure, tout comme BWSTAT. IHData analyst et ProUCL intègrent plusieurs procédures existantes, mais aucune qui repose sur l'imputation multiple. WebExpo utilise la même approche bayésienne que celle décrite dans Huynh *et al.* (2014) et implémentée dans ART (McNally *et al.*, 2014), appliquée à tous les modèles et élargie au-delà des seules données censurées à gauche (sans doute le cas le plus fréquent en hygiène du travail) pour inclure les données censurées par intervalle et les données censurées à droite.

La gestion de l'incertitude est un élément essentiel de l'évaluation des risques (Waters *et al.*, 2015). Dans WebExpo, nous avons tiré parti de la nature probabiliste des statistiques bayésiennes pour proposer, parallèlement aux calculs plus traditionnels, un cadre de travail permettant d'estimer la probabilité que les critères de surexposition soient respectés. Inspirés par Hewett *et al.* (Paul Hewett *et al.*, 2006), dont nous avons étendu la proposition à d'autres indices et à d'autres types d'analyses, nous sommes d'avis que la présentation de l'interprétation des données d'exposition sous la forme « Il y a XX % de chances que notre critère de surexposition soit respecté » est plus efficace que les traditionnels rapports statistiques complexes au moment de communiquer avec les gestionnaires et les travailleurs.

Les algorithmes et les prototypes de WebExpo offrent un potentiel d'application plus large que l'analyse des données de mesure en HT. Ainsi avons-nous, pour l'essentiel, créé des moteurs bayésiens permettant d'estimer les paramètres lognormaux et normaux de toute quantité pour laquelle ces modèles peuvent être considérés comme pertinents. De plus, le modèle d'analyse des différences inter-travailleur peut être utilisé avec des unités de regroupement autres que les travailleurs, notamment l'établissement, la profession ou le site contaminé en matière de données de pollution environnementale.

Nos algorithmes sont les premiers à permettre la prise en compte de l'erreur de mesure dans l'analyse des données d'HT. Comme mentionné dans l'introduction, seulement deux publications ont évalué si la variabilité analytique pouvait avoir un impact sur l'évaluation de la variabilité environnementale. Nous ne croyons pas que l'erreur de mesure doive toujours être prise en compte, vu la puissance de calcul requise, mais elle peut néanmoins être rigoureusement traitée dans les situations où l'on juge important de le faire. Plus important encore, peut-être, la possibilité d'inclure l'erreur de mesure dans l'interprétation des données d'HT sans recourir à une quelconque forme de simplification devrait permettre de revoir les travaux fondamentaux de Nicas *et al.* (Nicas *et al.*, 1991) et de Grzebyk et Sandino (Grzebyk et Sandino, 2005) afin de dégager une image plus précise de l'impact de l'erreur de mesure sur la prise de décision.

Enfin, les algorithmes de WebExpo sont ouvertement disponibles sous licence libre Apache 2.0. Ils peuvent donc être utilisés sans restriction par quiconque désire créer des outils ou les inclure dans un système de gestion de données préexistant (sous réserve d'une juste reconnaissance des chercheurs et des bailleurs de fonds). Nous espérons que cela favorisera leur utilisation par les établissements et les entreprises actifs dans des domaines liés à l'hygiène du travail. Bien qu'elle ne vise pas directement les intervenants, cette disponibilité devrait ultimement faciliter une interprétation robuste des données dans la pratique.

5.4 Limites

Tous les modèles créés par l'équipe de statisticiens n'ont pas été implémentés sur toutes les plateformes. Par conséquent, le traitement de l'erreur de mesure n'est possible qu'en R et en C#, et le modèle de distribution a priori des bandes de risque n'est disponible qu'en R. Ces restrictions se sont avérées nécessaires en raison du défi imprévu posé par la traduction du code de R avec prise en compte de l'erreur de mesure. À titre d'illustration, le cœur du code MCCM est passé de ~500 à ~3500 lignes en ajoutant le traitement de l'erreur de mesure. Comme la pratique courante en matière d'analyse de données d'HT ne tient pas compte des erreurs de mesure, nous avons estimé que la plupart des besoins actuels seraient couverts par les modèles implémentés sur toutes les plateformes de notre projet. Les utilisateurs qui souhaitent effectuer des analyses plus poussées peuvent utiliser le code en C# ou le code en R. Comme le code pur de R donne accès à toutes les fonctionnalités, l'élargissement des possibilités actuellement offertes en C# ou en JavaScript est également envisageable avec des ressources appropriées.

WebExpo n'inclut ni statistiques non paramétriques ni fonctions de vérification de la forme de distribution des échantillons. Ces deux particularités ont été discutées dans le cadre de la rencontre du comité d'experts. Compte tenu de la faible puissance statistique associée aux approches non paramétriques, il a été jugé que leur ajout ne serait pas utile dans le contexte des pratiques courantes vu la taille des échantillons (typiquement < 10 pour évaluer une situation particulière). De même, l'utilité pratique des tests d'hypothèses formels pour évaluer la

normalité ou la lognormalité (des versions de ces tests populaires en HT comprennent ceux de Shapiro-Wilk et de Shapiro-Francia [Shapiro et Francia, 1972a, 1972b]) a été jugée limitée pour ce qui est de décider s'il convient ou non d'utiliser l'approche lognormale dans l'évaluation des risques. En bref, le fait de pouvoir répondre à la question « Est-ce que cet échantillon provient d'une distribution lognormale ? », soit la question du test d'hypothèse formel, n'est pas directement pertinent à la question réelle, à savoir « Est-ce que le modèle lognormal permet de tirer des conclusions utiles ? ». De plus, la puissance de ces tests est très limitée compte tenu de la taille actuelle des échantillons en HT. Par conséquent, en ce qui concerne l'interprétation des données d'exposition sur les lieux de travail, nous recommandons de présumer la lognormalité – une hypothèse par défaut raisonnable compte tenu de la littérature existante sur le sujet –, mais aussi d'utiliser des outils graphiques de type graphe quantile-quantile pour détecter tout écart important. Nonobstant l'avis de l'équipe de WebExpo sur les tests statistiques formels, les utilisateurs sont libres d'utiliser ou de définir toute procédure visant à évaluer des données avant de les soumettre aux algorithmes de calcul lognormaux ou normaux de WebExpo.

Nous avons initialement songé à inclure certains outils d'analyse des « déterminants de l'exposition » dans WebExpo. Nous avons entre autres examiné la possibilité d'analyser l'effet d'une variable nominale (analyse de variance) ou continue (régression linéaire ou lissée), dans un modèle pouvant inclure jusqu'à deux variables. Le fait est que de telles analyses constituent désormais l'essentiel des analyses de jeux de données sur l'exposition publiées dans les revues d'HT, et qu'elles sont de plus en plus utiles grâce à l'accumulation de données d'exposition sous forme numérique par les entreprises et les établissements. Elles ont cependant été exclues du champ d'application de WebExpo, les experts ayant jugé que le niveau d'expertise statistique nécessaire pour effectuer adéquatement de telles analyses exigerait de maîtriser un progiciel statistique, rendant du coup inutile la création d'un outil à cette fin. Nous tenons néanmoins à préciser que l'exécution d'une analyse du modèle GES distincte pour chaque catégorie d'une variable nominale (p. ex. données de quart de jour ou de nuit) en combinant les échantillons MCCM permettrait la réalisation d'analyses apparentées à l'analyse de variance traditionnelle (p. ex. estimation de l'écart de la fraction de dépassement, + intervalle de crédibilité, entre deux catégories quelconques). Apparentées, mais pas équivalentes pour autant, car l'analyse de variance utilise l'ensemble des données plutôt que des jeux de données séparés par catégorie. En outre, l'analyse de variance présume une variance commune aux différentes catégories, par opposition à une variance propre à chaque catégorie.

Enfin, bien que nous croyions que les algorithmes développés dans le cadre de ce projet seraient en définitive avantageux pour les intervenants en HT, ils ne constituent pas en soi une boîte à outils d'interprétation de données facilement utilisable d'entrée de jeu. Par conséquent, les deux comités ont jugé qu'une introduction accessible aux statistiques lognormales, des graphiques explicatifs et des extrants d'une complexité adaptée au niveau d'expertise des différents utilisateurs s'avéraient essentiels à la composition d'une boîte à outils réellement utile pour l'interprétation des données d'HT. Nous sommes tout à fait d'accord avec cette évaluation, et nous tenons à ajouter que les prototypes en C# et en JavaScript ne doivent pas être considérés comme des outils pratiques en HT, car ils ne sont assortis d'aucun de ces éléments. Nous estimons néanmoins que nous avons créé une base de calcul solide et complète qui devrait servir de point de départ à la création d'outils pouvant être adaptés aux besoins particuliers des différents intervenants.

5.5 Rapport entre WebExpo et la boîte à outils d'interprétation de données en ligne Expostats

Vers la même époque où une demande de financement a été soumise à l'IRSST concernant le présent projet (en 2014), notre équipe à l'Université de Montréal a lancé la première version de la boîte à outils Expostats[16], une suite logicielle en ligne gratuite axée sur l'interprétation des données d'hygiène du travail. Au départ très limitée quant à son interface utilisateur et ses capacités, cette suite a évolué jusqu'à fournir un jeu d'outils complet pouvant désormais être aussi utilisé hors ligne. Lavoué *et al.* en ont récemment fait la description dans *Annals of work exposures and health* (Jérôme Lavoué *et al.*, 2018). Les outils d'Expostats permettent des calculs similaires à ceux décrits dans le présent rapport, mais ils n'autorisent qu'un seul type de distribution a priori (*informedvar*) et ne permettent pas de prendre en compte l'erreur de mesure. L'exécution des modèles bayésiens d'Expostats repose sur des scripts en JAGS et en R, de même que sur l'application SHINY[17], qui sert d'interface entre R et les utilisateurs en ligne. La boîte à outils Expostats est installée sur des serveurs de capacité limitée, ce qui restreint le nombre d'utilisateurs simultanés, et pour des raisons de licence, le moteur de calcul ne peut pas facilement être utilisé par d'autres afin de créer leurs propres outils. En gros, Expostats visait à fournir aux intervenants des outils de calcul de pointe à court terme, et il est fonctionnel depuis maintenant plusieurs années. En revanche, Webexpo visait à créer une base algorithmique d'exploitation libre pour le même ensemble de calculs, afin de permettre aux établissements ou aux entreprises de créer des solutions adaptées à leurs propres besoins et, à plus long terme, de favoriser l'adoption à plus grande échelle de pratiques de pointe en matière d'interprétation de données dans notre domaine.

[16] http://www.expostats.ca/site/index.html
[17] https://shiny.rstudio.com/

6. CONCLUSION

L'interprétation des données quantitatives d'HT n'est qu'une partie de l'analyse des risques sur les lieux de travail ; dans de nombreux cas, des décisions peuvent même être prises sans qu'aucune mesure n'ait à être prélevée. Cependant, lorsque des données d'exposition quantitatives sont disponibles, elles exigent une interprétation adéquate. Cette partie de l'équation continue de poser un défi dans l'évaluation des risques sur les lieux de travail, notamment en raison d'une forte variabilité environnementale, d'un cadre d'analyse statistique plutôt complexe, et d'une étonnante pénurie d'outils pratiques. Le projet WebExpo représente une importante initiative d'application pratique des récentes avancées informatiques et théoriques. Bien qu'à ce stade, les algorithmes et les prototypes proposés ne puissent être directement utilisés par les intervenants, ils peuvent être librement utilisés par tout établissement, individu ou entreprise comme une solide et rigoureuse base de développement d'outils d'interprétation de données d'HT de nouvelle génération. Plus précisément, les deux prototypes serviront de point de départ à la création ultérieure d'un outil d'interprétation de données pratique et complet exclusif à l'IRSST. Il demeure important de garder à l'esprit que la robustesse des conclusions dépendra de la pertinence du modèle lognormal ou normal dans chaque situation, de la représentativité et de la qualité des échantillons recueillis, ainsi que de la pertinence des distributions a priori.

BIBLIOGRAPHIE

Arnold, S. F., Stenzel, M., Drolet, D. et Ramachandran, G. (2016). Using checklists and algorithms to improve qualitative exposure judgment accuracy. *Journal of Occupational and Environmental Hygiene*, *13*(3), 159-168.

Ashley, K. et Bartley, D. L. (2004). Analytical performance criteria. *Journal of Occupational and Environmental Hygiene*, *1*(4), D37-D41.

Banerjee, S., Ramachandran, G., Vadali, M. et Sahmel, J. (2014). Bayesian hierarchical framework for occupational hygiene decision making. *Annals of Occupational Hygiene*, *58*(9), 1079-1093.

Bartley, D. L. (2001, July). Definition and assessment of sampling and analytical accuracy. *Annals of Occupational Hygiene, 45*(5), 357-364.

Bartley, D. et Lidén, G. (2008). Measurement uncertainty. *The Annals of Occupational Hygiene*, *52*(6), 413-417.

BOHS-NVvA. (2011). *Testing compliance with occupational exposure limits for airborne substances*. Tiré de http://www.bohs.org/library/technicalpublications

BOHS Technology Committee Working Group. (1993). *British Occupational Hygiene Society Technical Guide No. 11 : Sampling strategies for airborne contaminants in the Workplace*. Leeds, Canberra, Australie: Libraries Australia

Breslin, A. J., Ong, L., Glauberman, H., George, A. C. et Leclare, P. (1967). The accuracy of dust exposure estimates obtained from conventional air sampling. *American Industrial Hygiene Association Journal*, *28*(1), 56-61.

Buringh, E. et Lanting, R. (1991). Exposure variability in the workplace: Its implications for the assessment of compliance. *American Industrial Hygiene Association Journal*, *52*(1), 6-13.

CEN. (1995). *Workplace atmospheres: Guidance for the assessment of exposure by inhalation to chemical agents for comparison with limit values and measurement strategy*. Standard, EN 689:2018. Brussels, Belgium: CEN.

CEN. (2018). *Workplace exposure: Measurement of exposure by inhalation to chemical agents: Strategy for testing compliance with occupational exposure limit values*. Standard, EN 689:2018. Brussels, Belgium: CEN.

Clerc, F. et Vincent, R. (2014). Assessment of occupational exposure to chemicals by air sampling for comparison with limit values: The Influence of Sampling Strategy. *The Annals of Occupational Hygiene*, *58*(4), 437-449.

Comité européen de normalisation. (1995). *Atmosphères des lieux de travail : conseils pour l'évaluation de l'exposition aux agents chimiques aux fins de comparaison avec des valeurs limites et stratégie de mesurage* Norme NF EN 689. Bruxelles, Belgique: Communauté européenne.

Drolet, D. et Beauchamp, G. (2013). *Sampling guide for air contaminants in the workplace* (8e éd.). (Rapport no T-15). Montréal, QC: IRSST.

Drolet, D., Goyer, N., Roberge, B., Lavoué, J., Coulombe, M. et Dufresne, A. (2013). *Stratégies de diagnostic de l'exposition des travailleurs aux substances chimiques (Rapport no 665)*. Montréal, QC: IRSST.

Esmen, N. (1979). Retrospective industrial hygiene surveys. *American Industrial Hygiene Association Journal*, *40*(1), 58-65.

Esmen, N. A. et Hammad, Y. H. (1977). Log-normality of environmental sampling data. *Journal of Environmental Science and Health*, *A12*, 29-41.

Espino-Hernandez, G., Gustafson, P. et Burstyn, I. (2011). Bayesian adjustment for

measurement error in continuous exposures in an individually matched case-control study. *BMC Medical Research Methodology*, *11*(1), 67.

Flynn, M. R. (2010). Analysis of censored exposure data by constrained maximization of the Shapiro-Wilk W statistic. *The Annals of Occupational Hygiene*, *54*(3), 263-271.

Ganser, G. H. et Hewett, P. (2010). An accurate substitution method for analyzing censored data. *Journal of Occupational and Environmental Hygiene*, *7*(4), 233-244.

Gelman, A. (2013). *Bayesian data analysis (3e éd.)*. Boca Raton, FL: CRC Press.

Groth, C., Banerjee, S., Ramachandran, G., Stenzel, M. R., Sandler, D. P., Blair, A., . . . Stewart, P. A. (2017). Bivariate left-censored bayesian model for predicting exposure: Preliminary analysis of worker exposure during the deepwater horizon oil spill. *Annals of Work Exposures and Health*, *61*(1), 76-86.

Grzebyk, M.,et Sandino, J. P. (2005). Aspects statistiques et rôle de l'incertitude de mesurage dans l'évaluation de l'exposition professionnelle aux agents chimiques. *Hygiène et Sécurité du Travail*, *200*, 9-22.

Hawkins, N. C., Norwood, S. K. et Rock, J. C. (1991). *A strategy for occupationnal exposure assessment*. Fairfax, VA: American Industrial Hygiene Association.

Helsel, D. (2005). *Non detects and data analysis :Statistics for censored environmental data*. Hoboken, NJ: John Wiley & Sons.

Helsel, D. (2010). Much ado about next to nothing: Incorporating nondetects in science. *The Annals of Occupational Hygiene*, *54*(3), 257-262.

Helsel, D. R. (2012). *Statistics for censored environmental data using minitab and R (2e. éd.)*. Hoboken, NJ: John Wiley & Sons.

Hewett, P. (1997). Mean testing: I. Advantages and disadvantages. *Applied Occupational and Environmental Hygiene*, *12*(5), 339-346.

Hewett, P. et Ganser, G. H. (2007). A comparison of several methods for analyzing censored data. *The Annals of Occupational Hygiene*, *51*(7), 611-32.

Hewett, P., Logan, P., Mulhausen, J., Ramachandran, G. et Banerjee, S. (2006). Rating exposure control using Bayesian decision analysis. *Journal of Occupational and Environmental Hygiene*, *3*(10), 568-581.

Hornung, R. et Reed, L. D. (1990). Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene*, *5*(1), 46-51.

Huynh, T., Quick, H., Ramachandran, G., Banerjee, S., Stenzel, M., Sandler, D. P., . . . Stewart, P. A. (2016). A comparison of the β-substitution method and a Bayesian method for Analyzing Left-Censored Data. *The Annals of Occupational Hygiene*, *60*(1), 56-73.

Huynh, T., Ramachandran, G., Banerjee, S., Monteiro, J., Stenzel, M., Sandler, D. P., . . . Stewart, P. A. (2014). Comparison of methods for analyzing left-censored occupational exposure data. *The Annals of Occupational Hygiene*, *58*(9), 1126-1142.

Ignacio, J. S. et Bullock, W. H. (2008). *A strategy for assessing and managing occupational exposures* (3e éd.). Fairfax, VA: AIHA Press.

INRS. (2018). *Interprétation statistique des résultats de mesure*. Paris, France: INRS.

Jahn, S. D., Bullock, C. et Ignacio, J. S. (2015). *A strategy for assessing and managing occupational exposures* (4e éd.). Fairfax, VA: AIHA Press.

Jayjock, M. A, Chaisson, C. F., Franklin, C. A, Arnold, S. et Price, P. S. (2009). Using publicly available information to create exposure and risk-based ranking of chemicals used in the workplace and consumer products. *Journal of Exposure Science & Environmental Epidemiology*, *19*(5), 515-524.

Jones, R. M. et Burstyn, I. (2017). Bayesian analysis of occupational exposure data with conjugate priors. *Annals of Work Exposures and Health*, *61*(5), 504-514.

Kerr, G. W. (1962). Use of statistical methodology in environmental monitoring. *American Industrial Hygiene Association Journal*, *23*(1), 75-82.

Krishnamoorthy, K., Mallick, A. et Mathew, T. (2009). Model-based imputation approach for data analysis in the presence of non-detects. *The Annals of Occupational Hygiene*, *53*(3), 249-263.

Kromhout, H., Symanski, E. et Rappaport, S. M. (1993). A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *The Annals of Occupational Hygiene*, *37*(3), 253-270.

Kumagai, S. et Matsunaga, I. (1995). Changes in the distribution of short-term exposure concentration with different averaging times. *American Industrial Hygiene Association Journal*, *56*(1), 24-31.

Lavoué, J. (2013). Dealing with non-detects in occupational hygiene datasets. *Exposure*, December 2013:13-16.

Lavoue, J., Friesen, M. C et Burstyn, I. (2013). Workplace measurements by the US occupational safety and health administration since 1979: Descriptive analysis and potential uses for exposure assessment. *Annals of Occupational Hygiene*, *57*(1):77-97.

Lavoué, J., Joseph, L., Knott, P., Davies, H., Labrèche, F., Clerc, F., . . . Kirkham, T. (2018). Expostats: A Bayesian toolkit to aid the interpretation of occupational exposure measurements. *Annals of Work Exposures and Health*. Tiré de http://doi.org/10.1093/annweh/wxy100.

Leidel, N. A. et Busch, K. A. (2000). Statistical design and data analyis. In R. L. Harris (Ed.), *Patty's industrial hygiene* (5e éd., p. 2387-2514). New York, NY: John Wiley & Sons.

Leidel, N. A., Busch, K. A. et Lynch, C. F. (1977). *NIOSH Occupational exposure sampling strategy manual*. Cincinnati, OH: US Department of Health, Education, and Welfare.

Leidel, N., Busch, K. et Crouse, W. E. (1975). NIOSH *Technical information : Exposure measurement action level and occupational environmental variability* (NIOSH 76-131). Cincinnati, OH: NIOSH. Tiré de https;//www.cdc.gov/niosh/docs/76-131/pdfs/76-131.pdf.

Logan, P., Ramachandran, G., Mulhausen, J. et Hewett, P. (2009). Occupational exposure decisions: Can limited data interpretation training help improve accuracy? *The Annals of Occupational Hygiene*, *53*(4), 311-324.

Logan, P. W., Ramachandran, G., Mulhausen, J. R., Banerjee, S. et Hewett, P. (2011). Desktop study of occupational exposure judgments: Do education and experience influence accuracy? *Journal of Occupational and Environmental Hygiene*, *8*(12), 746-758.

Lyles, R. H. et Kupper, L. L. (1996). On strategies for comparing occupational exposure data to limits. *American Industrial Hygiene Association Journal*, *57*(1), 6-15.

Lyles, R. H., Kupper, L. L. et Rappaport, S. M. (1997a). A lognormal distribution-based exposure assessment method for unbalanced data. *Annals of Occupational Hygiene*, *41*(1), 63-76.

Lyles, R. H., Kupper, L. L. et Rappaport, S. M. (1997b). Assessing regulatory compliance of occupational exposures via the balanced one-way random effects ANOVA model. *Journal of Agricultural, Biological, and Environmental Statistics*, *2*(1), 64-86.

Martin Remy, A. et Wild, P. (2017). Bivariate left-censored measurements in biomonitoring: A Bayesian model for the determination of biological limit values based on occupational exposure limits. *Annals of Work Exposures and Health*, *61*(5), 515-527.

Mcbride, S., Williams, R. et Creason, J. (2007). Bayesian hierarchical modeling of personal exposure to particulate matter. *Atmospheric Environment*, *41*(29), 6143-6155.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

McNally, K., Warren, N., Fransman, W., Entink, R. K., Schinkel, J., van Tongeren, M., . . . Tielemans, E. (2014). Advanced REACH Tool: A Bayesian model for occupational exposure assessment. *The Annals of Occupational Hygiene*, *58*(5), 551-565.

Morton, J., Cotton, R., Cocker, J. et Warren, N. D. (2010). Trends in blood lead levels in UK workers, 1995-2007. *Occupational and Environmental Medicine*, *67*(9), 590–595.

Mulhausen, J. R. et Diamano, J. (1998). *A strategy for assessing and managing occupational exposures* (2e éd. Fairfax, VA: AIHA Press.

Nicas, M., Simmons, B. P. et Spear, R. C. (1991). Environmental versus analytical variability in exposure measurements. *American Industrial Hygiene Association Journal*, *52*(12), 553-577.

Ogden, T. L. (2010). Handling results below the level of detection. *The Annals of Occupational Hygiene*, *54*(3), 255-256.

Ogden, T. et Lavoué, J. (2012). Testing compliance with occupational exposure limits: Development of the British-Dutch guidance. *Journal of Occupational and Environmental Hygiene*, *9*(4), D63-70.

Oldham, P. D. (1953). The nature of the variability of dust concentrations at the coal face. *British Journal of Industrial Medicine*, *10*(4), 227-234.

Pesch, B., Kendzia, B., Hauptmann, K., Van Gelder, R., Stamm, R., Hahn, J.-U., . . . Brüning, T. (2015). Airborne exposure to inhalable hexavalent chromium in welders and other occupations: Estimates from the German MEGA database. *International Journal of Hygiene and Environmental Health*, *218*(5), 500-506.

Pilote, L., Joseph, L., Bélisle, P., Robinson, K., Van Lente, F. et Tager, I. B. (2000). Iron stores and coronary artery disease: A clinical application of a method to incorporate measurement error of the exposure in a logistic regression model. *Journal of Clinical Epidemiology*, *53*(8), 809-816.

Quick, H., Huynh, T.et Ramachandran, G. (2017). A method for constructing informative priors for Bayesian modeling of occupational hygiene data. *Annals of Work Exposures and Health*, *61*(1), 67-75.

R Core Team. (2014). R: A language and environment for statistical computing. Vienne, Autriche: R Foundation for Statistical Computing. Tiré de http://www.r-project.org/

Ramachandran, G. (2008). Toward better exposure assessment strategies: The new NIOSH initiative. *The Annals of Occupational Hygiene*, *52*(5), 297-301.

Ramachandran, G. et Vincent, J. H. (1999). A Bayesian approach to retrospective exposure assessment. *Applied Occupational and Environmental Hygiene*, *14*(8), 547–557.

Rappaport, S. M. (1984). The rules of the game: An analysis of OSHA's enforcement strategy. *American Journal of Industrial Medicine*, *6*(4), 291-303.

Rappaport, S. M. (1991). Assessment of long-term exposures to toxic substances in air. *The Annals of Occupational Hygiene*, *35*(1), 61-121. Tiré de http://doi.org/10.1093/annhyg/35.6.674.

Rappaport, S. M. (2000). Interpreting levels of exposures to chemical agents. Dans R. L. Harris (Ed.), *Patty's industrial hygiene* (5e éd. p. 679-745). New York, NY: John Wiley & Sons.

Rappaport, S. M., Kromhout, H. et Symanski, E. (1993). Variation of exposure between workers in homogeneous exposure groups. *American Industrial Hygiene Association Journal*, *54*(11), 654-662.

Rappaport, S. M., Lyles, R. H. et Kupper, L. L. (1995). An exposure-assessment strategy accounting for within- and between-worker sources of variability. *Annals of Occupational Hygiene*, *39*(4), 469-495.

République française. (2009). Arrêté du 15 décembre 2009 relatif aux contrôles techniques des valeurs limites d'exposition professionnelle sur les lieux de travail et aux conditions d'accréditation des organismes chargés des contrôles. *Journal officiel de la République française*, Texte 35 sur 156.

Roach, S. A. (1966). A more rational basis for air sampling progammes. *American Industrial Hygiene Association Journal*, *27*(1), 1-12.

Roach, S. A. (1977). A most rational basis for air sampling programmes. *The Annals of Occupational Hygiene*, *20*(1), 65-84.

Sarazin, P., Labrèche, F., Lesage, J. et Lavoué, J. (2018). *Étude comparative des banques de données de mesures d'exposition IMIS (OSHA) et LIMS* (Rapport n^{o} R-1032). Montréal, QC:IRSST.

Shapiro, S. S. et Francia, R. (1972a). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, *67*(337), 215-216.

Shapiro, S. S. et Francia, R. S. (1972b). An approximate analysis of variance test for normality. Journal of American Statistical Association, 67(32), 215-216

Sottas, P.-E., Lavoué, J., Bruzzi, R., Vernez, D., Charrière, N., & Droz, P.-O. (2009). An empirical hierarchical Bayesian unification of occupational exposure assessment methods. *Statistics in Medicine*, *28*(1), 75-93.

Tornero-Velez, R., Symanski, E., Kromhout, H., Yu, R. C. et Rappaport, S. M. (1997). Compliance versus risk in assessing occupational exposures. *Risk Analysis* , *17*(3), 279-292.

Waters, M., McKernan, L., Maier, A., Jayjock, M., Schaeffer, V. et Brosseau, L. (2015). Exposure estimation and interpretation of occupational risk: Enhanced information for the occupational risk manager. *Journal of Occupational and Environmental Hygiene*, *12*(S1), S99-S111.

Wild, P., Hordan, R., Leplay, A. et Vincent, R. (1996). Confidence intervals for probabilities of exceeding threshold limits with censored log-normal data. *Environmetrics*, *7*(3), 247-259.

Zhang, Y. F., Banerjee, S., Yang, R., Lungu, C. et Ramachandran, G. (2009). Bayesian modeling of exposure and airflow using two-zone models. *The Annals of Occupational Hygiene*, *53*(4), 409-424.

ANNEXE A : NOTES DE RENCONTRE DE LA RÉUNION INTERNATIONALE D'EXPERTS

Réunion WebExpo
18-19 février 2016
Montréal
Compte-rendu final de la réunion

TABLE DES MATIÈRES

1. INTRODUCTION

Ce document résume les discussions menées au cours de la réunion WebExpo, tenue les 18 et 19 février 2016 à Montréal. Il a été rédigé par Jérôme Lavoué et commenté/révisé par les participants à la réunion. Très peu de renseignements contextuels y sont fournis étant donné qu'ils sont présentés dans le protocole scientifique et dans les documents préparatoires à la réunion. Ce résumé est écrit de façon à dégager les impressions et les idées générales recueillies au cours de la réunion plutôt que les arguments précis de personnes en particulier.

Tableau A1 : liste des participants à la réunion d'experts

Nom	Affiliation	Domaine d'expertise	Titre
Jérôme Lavoué	Université de Montréal	Science de l'exposition	PI
Lawrence Joseph	Université McGill	Statistiques	Co-PI
Simon Aubin	IRSST	Métrologie en HT	Collaborateur
France Labrèche	IRSST	Épidémiologie	Collaboratrice
Tracy Kirkham	Université de Toronto	Hygiène du travail	Collaboratrice
Gautier Mater	INRS	Hygiène du travail	Collaborateur
Frédéric Clerc	INRS	Statistiques	Collaborateur
Patrick Bélisle	Université McGill	Programmation statistique	Documentaliste
Dunia Ouedraogo	Université de Montréal	Science de l'exposition	Doctorant
Martie van Tongeren	IOM	Science de l'exposition	Expert
Martine Chouvet	ITGA	Hygiène du travail	Experte
Paul Bozek	Université de Toronto	Hygiène du travail	Expert
Hugh Davies	Université de la	Hygiène du travail	Expert

	Colombie-Britannique		
Michel Gérin	Université de Montréal	Hygiène du travail	Expert

2. LISTE DES CALCULS DE BASE – FONCTIONNALITÉS INCLUSES DANS WebExpo

2.1 Évaluation de groupe (p. ex. données provenant d'un groupe d'exposition similaire)

Il y avait un large consensus concernant le fait que les 3 indices de risque proposés – la fraction de dépassement (nécessite de préciser la VLEP), le 95^{e} centile et la moyenne arithmétique – devraient être inclus dans WebExpo.

2.2 Modèle de variabilité intra-travailleur et inter-travailleur

Le groupe a convenu que les indices suivants étaient pertinents pour l'analyse de la variabilité intra-travailleur et inter-travailleur :

- probabilité qu'un travailleur au hasard présente un 95^{e} centile supérieur à la VLEP ;
- probabilité qu'un travailleur au hasard présente une moyenne arithmétique supérieure à la VLEP ;
- variabilité inter-travailleur ;
- variabilité intra-travailleur (une valeur moyenne pour l'ensemble des travailleurs) ;
- coefficient de corrélation intra-travailleur ;
- rapport de Rappaport.

D'autres propositions visant à refléter les différences entre les travailleurs comprenaient la mention d'un intervalle ou d'un écart-type relatif aux indices de risque des travailleurs (ex. : la fraction de dépassement moyenne était de 15 % avec un écart-type de 5 % sur l'ensemble des travailleurs). Un ajustement du rapport de Rappaport a aussi été proposé, en utilisant des centiles moins extrêmes de la distribution inter-travailleur (p. ex. 10 % et 90 % au lieu de 2,5 % et 97,5 %) afin d'en faciliter l'interprétation pour les petits groupes de travailleurs (qu'est-ce que le 2,5^{e} centile d'une population de 10 travailleurs ?).

Pour ce qui est d'évaluer si une exposition de groupe est homogène ou non, l'utilité du test d'analyse de la variance a été fortement critiquée pour les raisons suivantes : la plupart des gens n'en comprennent pas la signification réelle, et il ne fournit pas de réponse à la question d'intérêt fondamental : « Les différences entre les travailleurs sont-elles suffisantes pour avoir un impact sur le diagnostic final ? ». D'une part, un échantillon de petite taille ne permet pas de détecter d'importants écarts ; d'autre part, il se peut que l'hypothèse nulle soit rejetée même si les différences sont très faibles. Il y avait consensus au sein du groupe concernant le fait qu'il vaut mieux fournir des estimations de l'amplitude des différences inter-travailleur.

Le groupe a également convenu qu'il était utile de fournir des estimations d'exposition individuelle pour chaque travailleur échantillonné, même si ces estimations sont très approximatives. Des considérations d'ordre éthique concernant la tendance à « pointer du doigt » doivent néanmoins être prises en compte lors de la présentation ou de la communication des résultats.

Il a été souligné que, bien qu'il puisse être intéressant d'estimer la variabilité intra-travailleur propre à chaque travailleur, il n'est pas réaliste de le faire avec les tailles d'échantillons habituelles. Il y avait un intérêt pour les projets de simulation explorant l'impact d'une absence de modélisation de ces différences dans le diagnostic.

2.3. Autres fonctionnalités

Le groupe a convenu qu'il serait intéressant d'inclure la possibilité de modéliser l'influence d'une variable nominale. Cette fonctionnalité devrait être souple et permettre la saisie de données par le biais de fichiers Excel pour plusieurs candidats pouvant être analysés séparément par simple sélection de la part de l'utilisateur.

Le groupe a manifesté de l'intérêt pour l'éventuel ajout d'une variable continue (p. ex. durée de l'échantillon, tendance temporelle), mais a reconnu les difficultés techniques supplémentaires que présentait une telle option (p. ex. adéquation d'une simple pente, interprétation de la pente). L'intérêt le plus marqué concernait les tendances temporelles.

Le groupe ne s'est pas montré intéressé outre mesure par la possibilité d'analyser les données censurées par intervalle (plutôt que censurées à gauche seulement) ni par l'évaluation de la corrélation sérielle.

Certains établissements recommandent de signaler les situations présentant un ÉTG élevé afin de détecter des conditions anormales. Il n'y avait pas de consensus au sein du groupe concernant l'intérêt de cette procédure ni quant aux valeurs d'ÉTG « acceptables ». Un ÉTG élevé peut simplement refléter des conditions d'exposition très variables (p. ex. lutte contre les incendies).

La possibilité d'inclure l'erreur d'échantillonnage a été soulevée, car elle est facile à modéliser dans le cadre d'analyse bayésien, et l'incertitude entourant une valeur portant sur un quart de travail complet pourrait être importante lorsque seulement une partie du quart a fait l'objet de mesure. Comme les études disponibles (deux articles) montraient que l'erreur de mesure en HT n'importait que dans de rares situations par rapport à la variabilité environnementale, le groupe a convenu qu'il n'était pas nécessairement utile de la modéliser globalement, et qu'il valait mieux offrir la possibilité de l'inclure dans certaines conditions restreintes.

3. DISCUSSION SUR LES DISTRIBUTIONS A PRIORI BAYÉSIENNES

Il a été mentionné que la plupart des utilisateurs n'auraient pas les connaissances requises pour évaluer la pertinence ou la valeur des différentes distributions a priori proposées. Des a priori génériques pourraient être applicables dans plusieurs situations (p. ex. ÉTG fondé sur la base de données de Kromhout *et al.*), mais ne fourniraient que très peu d'information. D'autre part, des a priori informatifs ciblés (p. ex. de type BDA) peuvent fournir plus d'informations, mais il pourrait être très coûteux de s'assurer qu'ils sont exacts. D'ailleurs, il ne serait pas vraiment possible de savoir s'ils sont vraiment exacts, étant donné que les données réelles seraient vraisemblablement insuffisantes pour s'en assurer.

Nous nous trouvons dans une situation (échantillons de faible taille) qui ne permet pas vraiment d'évaluer la pertinence des différentes distributions a priori, alors que les résultats finaux

dépendront largement des a priori. Par conséquent, si les a priori ne concordent pas, la seule conclusion valable sera que les données ne sont pas assez fiables et que plus d'échantillons sont nécessaires.

En ce qui concerne les a priori qui ressemblent à un autre ensemble de données (p. ex. appelant à fournir une MG, un ÉTG et une taille d'échantillon virtuel), il a été souligné que les utilisateurs seraient généralement confus quant à la détermination de la taille de l'échantillon virtuel. En guise d'alternative, il serait possible de prévoir une barre de défilement permettant de visualiser l'impact de différentes valeurs des paramètres en question.

Il a également été soulevé qu'au lieu d'un paramètre reflétant directement la taille de l'échantillon virtuel, il serait préférable d'utiliser un paramètre qui reflète la « proximité » des valeurs a priori par rapport à une situation donnée (p. ex. une barre de défilement allant de « correspond à ma situation » à « peu pertinent »).

Les statisticiens du groupe ont souligné que les données généralement recueillies en HT ne semblent pas suffisantes pour caractériser adéquatement l'exposition dans de nombreuses situations. Il y avait consensus au sein du groupe quant au fait que les outils devraient par conséquent faire clairement ressortir l'importante incertitude entourant les estimations de l'exposition. Une option possible consisterait à détourner l'accent mis sur les estimations ponctuelles de manière à mieux refléter la gamme des valeurs probables.

Le groupe a convenu que l'inclusion de plusieurs distributions a priori différentes dans l'outil présentait un intérêt, mais qu'il était difficile de les présenter et de fournir un soutien à l'interprétation des résultats (des a priori inappropriés peuvent fausser les résultats). Il a été suggéré que des études soient menées sur la façon dont les résultats de différents a priori se comparent, dans l'espoir qu'elles puissent fournir une orientation de base.

En ce qui concerne les bandes a priori de la BDA et de l'AIHA, le groupe les a jugées difficiles à utiliser pour susciter le choix d'une distribution a priori de la part des utilisateurs (« larges catégories inégales », en fonction de paramètres inconnus, C95). Il a été jugé que les utilisateurs savaient mieux estimer les tendances centrales. Cette approche pourrait être utile pour présenter les résultats, mais elle semble moins intéressante pour ce qui est de susciter le choix de bonnes distributions a priori.

4. DISCUSSION SUR LES PETITS ENSEMBLES DE DONNÉES

Les lignes directrices de l'INRS recommandent une taille d'échantillon minimale absolue de 3. Dans le cas des échantillons d'une taille inférieure à 6, il n'y a pas de calcul de paramètres distributionnels ; on compare plutôt la valeur maximale de la série à une fraction de la VLEP. Cette approche repose sur une valeur a priori d'ÉTG, et non estimée à partir des données.

Il a été soulevé que dans le cas des échantillons de petite taille, les intervalles de confiance fondés sur un calcul fréquentiste pourraient ne pas être fiables, ce qui appuie l'idée d'une approche alternative en pareil cas. Il a toutefois été précisé que les intervalles de crédibilité bayésiens ne sont pas touchés par ce problème, car l'incertitude entourant la moyenne et l'écart-type est pleinement prise en compte dans les a priori. Une approche fondée sur un cadre bayésien n'obligerait donc pas à traiter différemment les très petits ensembles de données.

5. DISCUSSION SUR LES APPROCHES NON PARAMÉTRIQUES – PRÉSOMPTIONS RELATIVES AU MODÈLE LOGNORMAL

Le consensus autour de cette question était le suivant :

- les tests d'hypothèses comme le test de Shapiro-Wilk ne sont pas utiles pour évaluer si la distribution sous-jacente est lognormale :
  - ils n'indiquent pas la mesure dans laquelle nous sommes éloignés de la lognormale, ni si l'éloignement en question a une quelconque incidence sur l'interprétation ;
  - la forme de la distribution ne peut être évaluée à partir des tailles d'échantillon « courantes » (5-10) ;
- dans le cas des échantillons dont la taille est inférieure à 20-30, un graphique Q-Q n'est pas très utile non plus ;
- les bandes d'incertitude sur un graphique Q-Q sont en effet influencées par des points extrêmes, à tel point qu'un écart très extrême serait requis pour qu'un point tombe à l'extérieur.

En conclusion, les tests d'hypothèses formels ne seront probablement pas inclus dans la boîte à outils de WebExpo. Selon les statisticiens du groupe, il ne serait pas non plus utile de chercher à déterminer la forme d'une distribution en utilisant un graphique Q-Q comptant moins de 30 points. En dessous de ce seuil, des graphiques plus simples peuvent aider à repérer les valeurs aberrantes. Le choix du modèle de distribution reposerait donc sur une décision a priori (p. ex. lognormale pour les expositions à des agents chimiques aéroportés, normale pour le bruit). L'option de graphique Q-Q sera probablement incluse, assortie d'un avertissement quant au fait qu'elle risque de ne pas être très utile lorsque la valeur de n est faible.

Les approches non paramétriques ont été brièvement discutées. Il a été mentionné qu'elles ne seraient vraisemblablement pas utiles compte tenu de la taille actuelle des échantillons, vu l'importante perte de puissance. Un simple graphique séquentiel où divers symboles correspondraient à différents travailleurs fournirait un bon résumé et aiderait à repérer les valeurs extrêmes.

6. DISCUSSION SUR LA COMMUNICATION DES RISQUES

Le groupe était fortement d'accord avec la proposition d'envoyer différents messages à différents publics. Les collègues de l'INRS ont partagé leur expérience de la création d'un jeu-questionnaire pour déterminer la complexité du message en fonction des réponses. Avec le recul, ils ne recommanderaient pas la même approche pour WebExpo. Le groupe a convenu qu'une approche du genre « cliquez ici pour une interprétation plus détaillée » fonctionnerait bien. Il faudrait toutefois faire en sorte que les utilisateurs chevronnés n'aient pas à faire ce choix de façon répétée. Il y avait aussi consensus quant au fait que l'outil devrait encourager l'utilisateur à prendre connaissance de la réponse plus détaillée. La notion de jeu-questionnaire était populaire non pas en ce qui concerne la sélection automatique de la complexité des résultats d'interprétation des données, mais plutôt comme élément du matériel pédagogique.

Pour ce qui est des graphiques servant à illustrer les résultats de l'interprétation des données, la représentation de la fraction de dépassement au moyen de calendriers où certains jours

seraient grisés était populaire. D'autres suggestions comprenaient la courbe de densité de la distribution estimée, voire un nuage de courbes de densité lognormales visant à refléter l'incertitude. Le type de graphique devrait dépendre du niveau de compréhension des utilisateurs (p. ex., une courbe de densité n'est pas à la portée de beaucoup de gens). Le groupe a convenu qu'une composante « simulateur lognormal » serait souhaitable à des fins pédagogiques.

En ce qui concerne la diffusion des algorithmes développés, il a été souligné que, bien qu'une page wiki puisse bien fonctionner pour les documents (p. ex. exemples, lignes directrices, matériel pédagogique), il est peu probable qu'elle convienne pour les codes de programmation informatique. En outre, la maintenance et la tenue à jour d'un système wiki requièrent une source de financement constante (difficile à assurer ici). Les divers algorithmes n'utiliseront que des bibliothèques à libre accès, de sorte que leur distribution et leur utilisation ne soulèveront aucune question. Des versions Java des algorithmes seront produites, en partie parce que des collègues de l'INRS projettent de les utiliser pour la prochaine itération de leur propre outil d'interprétation de données, ALTREX2.

7. RÉFÉRENCES

1. Kromhout H, Symanski E, Rappaport SM. A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. Ann Occup Hyg 1993; 37:253–70. http://www.ncbi.nlm.nih.gov/pubmed/8346874 (accessed Oct. 9, 2014).

2. Banerjee S, Ramachandran G, Vadali M, *et al.* Bayesian Hierarchical Framework for Occupational Hygiene Decision Making. Ann Occup Hyg 2014; 58:1079-1093. doi:10.1093/annhyg/meu060.

3. McNally K, Warren N, Fransman W, *et al.* Advanced REACH Tool: a Bayesian model for occupational exposure assessment. Ann Occup Hyg 2014; 58:551–65. doi:10.1093/annhyg/meu017.

4. Arnold SF, Stenzel M, Drolet D, *et al.* Using checklists and algorithms to improve qualitative exposure judgment accuracy. J Occup Environ Hyg 2016; 13:159–68. doi:10.1080/15459624.2015.1053892.

5. Logan P, Ramachandran G, Mulhausen J, *et al.* Occupational exposure decisions: can limited data interpretation training help improve accuracy? Ann Occup Hyg 2009; 53:311–24. doi:10.1093/annhyg/mep011.

6. Vadali M, Ramachandran G, Mulhausen JR, *et al.* Effect of training on exposure judgment accuracy of industrial hygienists. J Occup Environ Hyg 2012; 9:242–56. doi:10.1080/15459624.2012.666470.

ANNEXE B :
DOCUMENTATION TECHNIQUE DES MODÈLES BAYÉSIENS

WebExpo: The R algorithms

August 2, 2018

1 Introduction

This document addresses sampling from the posterior distributions from several models in Industrial Hygiene. In each model, the prior distribution $f(\theta)$ for the hyperparameters $\theta = (\mu, \sigma)$ is relatively simple.

1.1 Generating a sample from posterior distribution via Markov Chain Monte Carlo (MCMC)

From Bayes theorem, the posterior distribution $f(\theta|x)$ for θ given data x is proportional to

$$f(\theta|x) \propto f(\theta) \times f(x|\theta) \ ,$$

where $f(\theta)$ is the prior distribution for θ and $f(x|\theta)$ is the likelihood function.

In most situations encountered in this work, the posterior for θ does not have an analytic solution but we can use Markov Chain Monte Carlo simulation to draw a sample from it. When θ consists of a series of parameters, say $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$, if the full conditional posterior distribution for θ_i can be written as a function of other components, that is, if we can write $f(\theta_i|\theta_{-i}, x)$, for $i = 1, 2, \ldots, q$ — where $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_q)$ — then the MCMC algorithm is as follows: sample θ_i from the above distribution for $i = 1, 2, \ldots, q$, collect the sampled values and repeat a large number of times; in the long run, the sample collected along these lines converges to a sample from the posterior distribution $f(\theta|x)$.

Sections 2–5 present all the models in absence of measurement error while Section 6 presents the modifications to bring to each of them in the presence of measurement error.

2 Uninformative model [SEG.uninformative]

The joint prior distribution for μ and σ in the uninformative model can be constructed from

$$\begin{aligned} \mu &\sim U(\mu_0, \mu_1) \\ \tau = \frac{1}{\sigma^2} &\sim \text{Gamma}(\alpha, \beta) \end{aligned} \tag{1}$$

with $\mu_0 = -1000, \mu_1 = 1000$ and $\alpha = \beta = 0.001$. The likelihood is

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

for $i = 1, 2, \ldots, N$, where the Y_i are independenty (and identically) distributed.

The joint posterior distribution for (μ, σ) is thus

$$\begin{aligned} f(\mu, \sigma|y) &\propto \frac{1}{\sigma^N} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu)^2\right\} \frac{1}{\sigma^{2\alpha+1}} \exp -\frac{\beta}{\sigma^2} I_\mu(\mu_0, \mu_1) \\ &= \frac{1}{\sigma^{N+2\alpha+1}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu)^2\right\} \exp -\frac{\beta}{\sigma^2} I_\mu(\mu_0, \mu_1) \end{aligned}$$

where $I_\theta(a,b)$ is the indicator function, that is,

$$I_\theta(a,b) = \begin{cases} 1 & \text{when } a \le \theta \le b \text{ and} \\ 0 & \text{elsewhere.} \end{cases}$$

From the above, we easily derive the full conditional posterior density for μ to get

$$f(\mu|\sigma, y) \sim N(\bar{y}, \sigma^2/N) \tag{2}$$

truncated to (μ_0, μ_1), while the full conditional posterior density for σ is

$$f(\sigma|\mu, y) \quad \propto \quad \frac{1}{\sigma^{N+2\alpha+1}} \exp -\frac{1}{\sigma^2}\left\{\beta + \frac{1}{2}\sum (y_i - \mu)^2\right\},$$

that is, $\tau = \sigma^{-2} \sim \text{Gamma}(\alpha + N/2, \beta + \frac{1}{2}\sum (y_i - \mu)^2)$ from (B.2).

NOTE: After discussion, the gamma prior distribution for τ in (1) was dropped from the package in favor of the uniform prior on σ introduced in next section.

2.1 Alternative posterior when the prior distribution for σ is (improper) uniform

An alternative to the above model is to use a uniform prior distribution on σ rather than the Gamma prior used in (1), that is, to use

$$\sigma \sim U(\sigma_0, \sigma_1)$$

where the range may be left unspecified, that is, with $\sigma_0 = 0$ and $\sigma_1 = \infty$, in which case σ's prior is said to be *improper*, that is, its density does not integrate to 1; this is not a problem in theory but in practice it may leave a high probability for large values for σ — especially when sample size is small — which do not make sense in practice. Hence we encourage the use of a finite upper bound $\sigma_1 < \infty$ based on the scale of the measurements taken.

The conditional posterior density for σ is then

$$f(\sigma|\mu, y) \propto \frac{1}{\sigma^N} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu)^2\right\}, \tag{3}$$

that is,

$$\begin{aligned} \tau = \sigma^{-2} \quad &\sim \quad \text{Gamma}((N-1)/2, b) \text{ from } (B.2) \\ &\quad \text{where } \; b = \frac{1}{2}\sum_{i=1}^{N}(y_i - \mu)^2 \;\text{ when }\; N > 1. \end{aligned}$$

When $N \le 1$, σ values can be sampled from the inverse cumultative density function (icdf) method presented in Appendix A, with $a = N = 1$ and b defined as above.

3 Kromhout model [SEG.informedvar]

The joint prior distribution for the Kromhout model is given by

$$\begin{aligned} \mu \quad &\sim \quad U(\mu_0, \mu_1) \\ \log(\sigma) \quad &\sim \quad N(\mu^*, \sigma^{*2}) \end{aligned} \tag{4}$$

and the likelihood is

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

where the Y_i's, $i = 1, 2, \ldots, N$ are independently distributed and can be left-, right- or interval-censored. The hyperparameter values are $\mu_0 = -101.38161, \mu_1 = 98.61839$, $\mu^* = -0.1744$ and $\sigma^{*-2} = 2.5523$.

The joint posterior for (μ, σ) is hence given by

$$f(\mu, \sigma|y) \propto \frac{1}{\sigma^N} \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right\}\frac{1}{\sigma}\exp\left\{-\frac{(\log(\sigma)-\mu^*)^2}{2\sigma^{*2}}\right\}I_\mu(\mu_0,\mu_1) \tag{5}$$

The full conditional posterior density for μ is thus given by

$$\begin{aligned} f(\mu|\sigma, y) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum y_i^2 - 2\mu\sum y_i + N\mu^2\right)\right\}I_\mu(\mu_0,\mu_1) \\ &\propto \exp\left\{-\frac{N}{2\sigma^2}\left(\mu^2 - 2\mu\bar{y}\right)\right\}I_\mu(\mu_0,\mu_1)\ , \end{aligned} \tag{6}$$

that is, $\mu \sim N(\bar{y}, \sigma^2/N)$ truncated to the interval (μ_0, μ_1) and the full conditional posterior density for σ is proportional to

$$f(\sigma|\mu, y) \propto \frac{1}{\sigma^{N+1}} \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right\}\exp\left\{-\frac{(\log(\sigma)-\mu^*)^2}{2\sigma^{*2}}\right\} \tag{7}$$

Generating MCMC values for μ from its full conditional posterior density (6) is straightforward, while σ values will be sampled from its full conditional posterior density (7) through the inverse cumulative density function sketched in Appendix A, with

$$\begin{aligned} a &= N, \\ b &= \frac{1}{2}\left(\sum y_i^2 - 2\mu\sum y_i + N\mu^2\right), \\ \tilde{\mu} &= \mu^* \text{ and} \\ \tilde{\sigma}^2 &= \sigma^{*2}\ . \end{aligned}$$

If there are any right-censored values y_i, that is, values specified as $y_i < z_i$ for some z_i's, then at each loop in the MCMC process, corresponding y_i values are sampled from $N(\mu, \sigma^2)$ on the interval $)-\infty, z_i)$. Similar sampling is also performed for left- and interval-censored y_i values.

3.1 Two-Level Kromhout model

The Two-Level Kromhout model is the same as Kromhout model discussed above but applied to two groups, that is, it is a model with the following prior distributions

$$\begin{aligned} \mu_j &\sim U(\mu_0, \mu_1) \\ \log(\sigma_j) &\sim N(\mu^*, \sigma^{*2}) \end{aligned}$$

independently for groups $j = 1, 2$ and the likelihood is given by

$$Y_{ji} \sim N(\mu_j, \sigma_j^2)$$

for $i = 1, 2, \ldots, N_j, j = 1, 2$. The hyperparameter values are $\mu^* = -0.1744$ and $\sigma^{*-2} = 2.5523$ with limits for μ slightly different from the one-group model, $\mu_0 = -100$ and $\mu_1 = 100$.

3.2 Use of past data

One might want to include past data — available through sample size n, observed mean $\bar{p}$ and standard deviation s_p — in the analysis. The likelihood of past data $\mathbf{p}$ — measured without error

— is given by

$$\begin{aligned} f(p|\mu,\sigma) &= \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(p_i-\mu)^2\right\} \\ &= \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(p_i-\bar{p}+\bar{p}-\mu)^2\right\} \\ &= \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(p_i-\bar{p})^2+2(p_i-\bar{p})(\bar{p}-\mu)+(\bar{p}-\mu)^2]\right\} \\ &= \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(p_i-\bar{p})^2\right\}\exp\left\{-\frac{n}{2\sigma^2}(\bar{p}-\mu)^2\right\} \\ &= \frac{1}{\sigma^n}\exp\left\{-\frac{(n-1)s_p^2}{2\sigma^2}\right\}\exp\left\{-\frac{n}{2\sigma^2}(\bar{p}-\mu)^2\right\} \quad . \end{aligned} \tag{8}$$

The joint posterior for (μ,σ) is hence given by the product of (5) and the above likelihood of past data, that is,

$$\begin{aligned} f(\mu,\sigma|y,p) &\propto \frac{1}{\sigma^{N+n+1}}\exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right\} \\ &\times \exp\left\{-\frac{(n-1)s_p^2}{2\sigma^2}\right\}\exp\left\{-\frac{n}{2\sigma^2}(\bar{p}-\mu)^2\right\} \\ &\times \exp\left\{-\frac{(\log(\sigma)-\mu^*)^2}{2\sigma^{*2}}\right\}I_\mu(\mu_0,\mu_1) \quad . \end{aligned}$$

The full conditional posterior density for μ is thus given by

$$\begin{aligned} f(\mu|\sigma,y,p) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum(y_i-\mu)^2+n(\bar{p}-\mu)^2\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(N\mu^2-2\mu\sum y_i+n\mu^2-2\mu n\bar{p}\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\mu^2(N+n)-2\mu(N\bar{y}+n\bar{p})\right)\right\}I_\mu(\mu_0,\mu_1) \\ \Longrightarrow \mu|\sigma,y,p &\sim \mathrm{N}\left(\frac{N\bar{y}+n\bar{p}}{N+n},\frac{\sigma^2}{N+n}\right)I_\mu(\mu_0,\mu_1) \end{aligned}$$

while the full conditional posterior distribution for σ is

$$f(\sigma|\mu,y,p) \propto \frac{1}{\sigma^{N+n+1}}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum(y_i-\mu)^2+(n-1)s_p^2+n(\bar{p}-\mu)^2\right)\right\}\exp\left\{-\frac{(\log(\sigma)-\mu^*)^2}{2\sigma^{*2}}\right\} .$$

Values for σ can be sampled from its full conditional posterior density through the inverse cumulative density function sketched in Appendix A, with

$$\begin{aligned} a &= N+n, \\ b &= \frac{1}{2}\left(\sum y_i^2-2\mu\sum y_i+N\mu^2+(n-1)s_p^2+n(\bar{p}-\mu)^2\right), \\ \tilde{\mu} &= \mu^* \text{ and} \\ \tilde{\sigma}^2 &= \sigma^{*2} \quad . \end{aligned}$$

Originally, the algorithm including past data was based on a uniform and improper distribution for σ which is not used anymore; the rationale for this deprecated version of the algorithm is relegated to Appendix C.5.

3.2.1 Limitations / warnings

If it is thought that the past data were measured with error, they sould NOT be used (indeed, the above section assumed that the past data was measuread without measurement error).

If the measurement error (in past data) was proportional to true (unmeasured) values — that is, measurement error would be modeled through a coefficient of variation — they sould DEFINITELY not be used (the assumptions on which the algorithm is based seem to be violated in a unfixable fashion).

If measurement error (in past data, again) was constant and relatively small when compared to σ, they could still be used, but with some caution. Indeed, the above calculations intrinsically assume that $(n-1)s_p^2/\sigma^2 \sim \chi^2_{n-1}$, which is NOT the case when past values are measured with error. If the measurement error is small, then we may not be very far from that distribution and the algorithm and past data still provide useful results.

The only way that past data obtained with measurement error could be used is if we have access to the complete list of observed values $p_1, p_2, \ldots, p_n$ rather than the usual summary statistics $(\bar{p}, s_p^2)$. In this case, if we additionally assume that the measurement error in past data was of same nature and size as in our actual data $(y_1, y_2, \ldots, y_N)$, then one could simply include past data as (additional) new data. The case when the measurement error is of different nature and/or size than the measurement error in the current data is beyond the scope of this program.

If the outcome of interest follows a log-normal distribution (rather than a normal distribution), then the mean and standard deviation of past data must have been calculated on the log values as well in order to be usable.

4 McNally model [Between.worker]

McNally's model is a hierarchical model with overall mean μ, having prior distribution

$$\mu \sim U(\mu_0^*, \mu_1^*) \, .$$

The model includes random worker effects $\mu_j, j = 1, \ldots, M$ with independent and identical prior distributions

$$\mu_j \sim N(0, \sigma_B^2) \, .$$

The parameter σ_B^2 is the between-worker variance, with prior distribution

$$\log(\sigma_B) \sim N(\mu_B^*, \sigma_B^{*2}) \tag{9}$$

where $\mu_B^* = -0.8786$ and $\sigma_B^{-2*} = 1.634$. The likelihood is given by

$$Y_i \sim N(\mu + \mu_{w_i}, \sigma_W^2)$$

where w_i is the worker index of observation $i, i = 1, \ldots, N$ and σ_W^2 is the within-subject variance, with prior distribution

$$\log(\sigma_W) \sim N(\mu^*, \sigma^{2*}) \tag{10}$$

where $\mu^* = -0.4106$ and $\sigma^{-2*} = 1.9002$.

The joint posterior density is proportional to

$$\begin{aligned} f(\mu, \mu_1, \ldots, \mu_M, \sigma_W^2, \sigma_B^2 | y) \quad \propto \quad & \frac{1}{\sigma_W^N} \exp\left\{ -\frac{1}{2\sigma_W^2} \sum_{i=1}^{N} (y_i - \mu - \mu_{w_i})^2 \right\} \\ \cdot \quad & \frac{1}{\sigma_B^M} \exp\left\{ -\frac{1}{2\sigma_B^2} \sum_{j=1}^{M} \mu_j^2 \right\} \\ \cdot \quad & f(\sigma_W^2) f(\sigma_B^2) I_\mu(\mu_0, \mu_1) \, . \end{aligned}$$

It follows that the conditional posterior density for worker k's mean is propotional to

$$\begin{aligned} f(\mu_k|y,\mu,\sigma_W,\sigma_B) &\propto \exp\left\{-\frac{1}{2\sigma_W^2}\sum_{i:w_i=k}(y_i-\mu-\mu_k)^2\right\}\exp\left\{-\frac{1}{2\sigma_B^2}\mu_k^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma_W^2}\left(\sum_{i:w_i=k}(y_i-\mu)^2-2\sum_{i:w_i=k}(y_i-\mu)\mu_k+\mu_k^2 n_k\right)\right\}\exp\left\{-\frac{1}{2\sigma_B^2}\mu_k^2\right\}, \end{aligned}$$

where n_k is the number of observations in worker k. Thus

$$f(\mu_k|y,\mu,\sigma_W,\sigma_B) \propto \exp\left\{-\frac{n_k}{2\sigma_W^2}\left(\mu_k^2-2(\bar{y_k}-\mu)\mu_k\right)\right\}\exp\left\{-\frac{1}{2\sigma_B^2}\mu_k^2\right\},$$

where $\bar{y_k}$ is the average of observations for worker k.

Simple algebra reduces the above expression to

$$\mu_k|y,\mu,\sigma_W,\sigma_B \sim N\left(\frac{(\bar{y_k}-\mu)\sigma_B^2}{\sigma_W^2/n_k+\sigma_B^2}, \frac{\sigma_W^2\sigma_B^2/n_k}{\sigma_W^2/n_k+\sigma_B^2}\right). \tag{11}$$

The full conditional posterior density for μ is proportional to

$$\begin{aligned} f(\mu|y,\mu_1,\ldots,\mu_M,\sigma_W^2) &\propto \exp -\frac{1}{2\sigma_W^2}\left\{\sum_{i=1}^{N}(y_i-\mu-\mu_{w_i})^2\right\}I_\mu(\mu_0,\mu_1) \\ &= \exp -\frac{1}{2\sigma_W^2}\left\{\sum_i (y_i-\mu_{w_i})^2-2\mu\sum_i(y_i-\mu_{w_i})+N\mu^2\right\}I_\mu(\mu_0,\mu_1) \\ \Rightarrow \mu|y,\mu_1,\ldots,\mu_M,\sigma_W^2 &\sim N\left(\frac{\sum_i(y_i-\mu_{w_i})}{N},\frac{\sigma_W^2}{N}\right) \quad \text{truncated on } (\mu_0,\mu_1) \\ &\sim N\left(\bar{y}-\frac{\sum n_k\mu_k}{N},\frac{\sigma_W^2}{N}\right) \quad \text{truncated on } (\mu_0,\mu_1) \end{aligned} \tag{12}$$

The full conditional posterior density for σ_W is proportional to

$$\begin{aligned} f(\sigma_W|y,\mu,\mu_1,\ldots,\mu_M) &\propto \frac{1}{\sigma_W^N}\exp\left\{-\frac{1}{2\sigma_W^2}\sum_{i=1}^{N}(y_i-\mu-\mu_{w_i})^2\right\}\frac{1}{\sigma_W}\exp\left\{-\frac{1}{2\sigma^{*2}}\left(\log(\sigma_W)-\mu^*\right)^2\right\} \\ &= \frac{1}{\sigma_W^{N+1}}\exp\left\{-\frac{1}{2\sigma_W^2}\sum_{i=1}^{N}(y_i-\mu-\mu_{w_i})^2\right\}\exp\left\{-\frac{1}{2\sigma^{*2}}\left(\log(\sigma_W)-\mu^*\right)^2\right\} \end{aligned} \tag{13}$$

It follows that σ_W values can be sampled from the inverse cumulative density function (icdf) method presented in Appendix A, with

$$\begin{aligned} a &= N+1\,, \\ b &= \frac{1}{2}\sum_{i=1}^{N}(y_i-\mu-\mu_{w_i})^2\,, \\ \tilde{\mu} &= \mu^* \text{ and} \\ \tilde{\sigma}^2 &= \tilde{\sigma}^{*2}\,. \end{aligned}$$

Finally, the full conditional posterior density for σ_B^2 is proportional to

$$f(\sigma_B|y,\mu_1,\ldots,\mu_M) \propto \frac{1}{\sigma_B^M}\exp\left\{-\frac{1}{2\sigma_B^2}\sum_{j=1}^{M}\mu_j^2\right\}\frac{1}{\sigma_B}\exp\left\{-\frac{1}{2\sigma_B^{*2}}\left(\log(\sigma_B)-\mu_B^*\right)^2\right\}$$

Thus, σ_B can be sampled from its conditional posterior density through icdf method with

$$\begin{aligned} a &= M, \\ b &= \frac{1}{2}\sum_{j=1}^{M}\mu_j^2, \\ \tilde{\mu} &= \mu_B^* \text{ and} \\ \tilde{\sigma}^2 &= \sigma_B^{*2}\ . \end{aligned}$$

4.1 Alternative posterior when σ_W and σ_B prior distributions are uniform

An alternative to the above model is to use a uniform prior distribution on both σ_W and σ_B rather than the log-normal prior distributions specified in (9) and (10); the two *sigma* variables could possibly be defined on a specified range only, rather than on $\Re^+$.

When using uniform prior distribution on σ_B and σ_W, their respective full conditional posterior distributions are

$$\begin{aligned} f(\sigma_B|y,\mu_1,\ldots,\mu_M) &\propto \frac{1}{\sigma_B^M}\exp\left\{-\frac{1}{2\sigma_B^2}\sum_{j=1}^{M}\mu_j^2\right\} \\ \text{and } f(\sigma_W|y,\mu_1,\ldots,\mu_M) &\propto \frac{1}{\sigma_W^N}\exp\left\{-\frac{1}{2\sigma_W^2}\sum_{i=1}^{N}\left(y_i-\mu-\mu_{w_i}\right)^2\right\} \end{aligned} \tag{14}$$

from which we can easily sample using the algorithm described in Appendix A if M(or N) ≤ 1, or from an Inverted-Gamma distribution otherwise.

5 Banerjee model [SEG.riskband]

In the Banerjee model, the outcome follows either a normal distribution

$$Y_i \sim N(\mu,\sigma^2), i = 1,2,\ldots,N$$

or a lognormal distribution, given parameters (μ,σ).

Prior knowledge is expressed around the 95th percentile of the data distribution, that is, $Y_{0.95} = \mu + 1.645\sigma$; a series of cut-offs points $A_1, A_2, \ldots, A_{R-1}$ is given along with prior probabilities that (μ,σ) falls in either of the regions

$$\begin{aligned} \mathcal{R}_1 &= \{(\mu,\sigma) : \mu + z\sigma \leq A_1\} \\ \mathcal{R}_2 &= \{(\mu,\sigma) : A_1 < \mu + z\sigma \leq A_2\} \\ &\vdots \\ \mathcal{R}_{R-1} &= \{(\mu,\sigma) : A_{R-2} < \mu + z\sigma \leq A_{R-1}\} \\ \mathcal{R}_R &= \{(\mu,\sigma) : \mu + z\sigma > A_{R-1}\} \end{aligned}$$

with probabilities

$$P\{(\mu,\sigma) \in \mathcal{R}_j\} = p_j, \quad j = 1,2,\ldots,R \tag{15}$$

where z is the desired quantile of the normal distribution, in general the 95th, that is, $z = 1.645$.

We assume a piecewise-uniform joint prior distribution for (μ,σ) on its rectangle domain $(\mu_0,\mu_1)\times(\sigma_0,\sigma_1)$, where (μ_0,μ_1) and (σ_0,σ_1) are the lower and upper limits for μ and σ respectively, that is

$$f(\mu,\sigma) = \begin{cases} f_1 \text{ for } (\mu,\sigma) \in \mathcal{R}_1 \\ f_2 \text{ for } (\mu,\sigma) \in \mathcal{R}_2 \\ \vdots \\ f_R \text{ for } (\mu,\sigma) \in \mathcal{R}_R\ . \end{cases}$$

Given the clinical cut-off points $A_1, A_2, \ldots, A_{R-1}$, the domain of (μ, σ) is split into R regions. The figure below illustrates a typical case of the domain partitioning done that way.

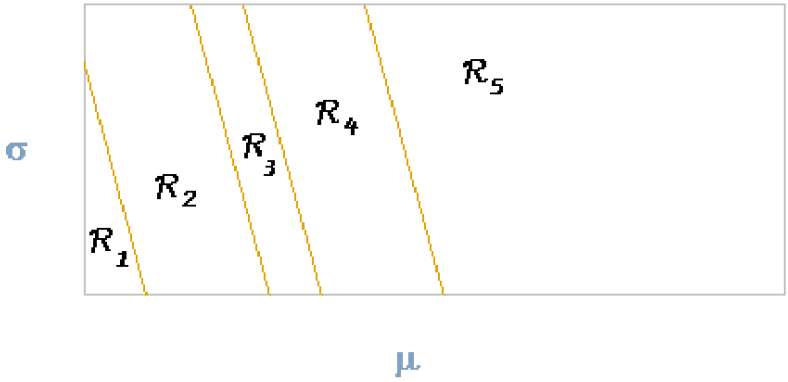


Figure 1: Regions defined by a series of clinical cut-off points.

If we let $S_j = \text{area}(\mathcal{R}_j), j = 1, 2, \ldots, R$, then the constant densities $f_1, f_2, \ldots, f_R$ for all points in each of the regions $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_R$ can be chosen such that

$$\begin{aligned} f_j &\geq 0, \quad j = 1, 2, \ldots, R, \\ \sum f_j &= 1 \\ \text{and } f_j \times S_j &\propto p_j, \quad j = 1, 2, \ldots, R. \end{aligned}$$

That is easily done by setting $f'_1 = 1$ and f'_j such that

$$\frac{f'_j S_j}{f'_1 S_1} = \frac{p_j}{p_1}, j \geq 2$$

and then $f_j = f'_j / \sum_k f'_k S_k, j = 1, 2, \ldots, R$.

The posterior distribution of (μ, σ) is thus

$$f(\mu, \sigma | y) \propto \frac{1}{\sigma^N} \exp\left\{-\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2\right\} f(\mu, \sigma) \; .$$

The full conditional posterior distribution for μ is proportional to

$$\begin{aligned} f(\mu | y, \sigma) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(N\mu^2 - 2\mu \sum y_i + \sum y_i^2\right)\right\} f(\mu | \sigma) \\ \Rightarrow \mu | y, \sigma &\sim N\left(\bar{y}, \frac{\sigma^2}{N}\right) f(\mu | \sigma) \; . \end{aligned}$$

Therefore, the full conditional posterior distribution for μ is a Normal density with sections defined by the μ-partition brought by $f(\mu|\sigma)$ weighed with the corresponding weights $f(\mu|\sigma)$.

In an attempt to clarify the result above, consider the typical figure below, suppose that $\sigma = \sigma^*$ and that we want to sample from $f(\mu | y, \sigma^*)$.

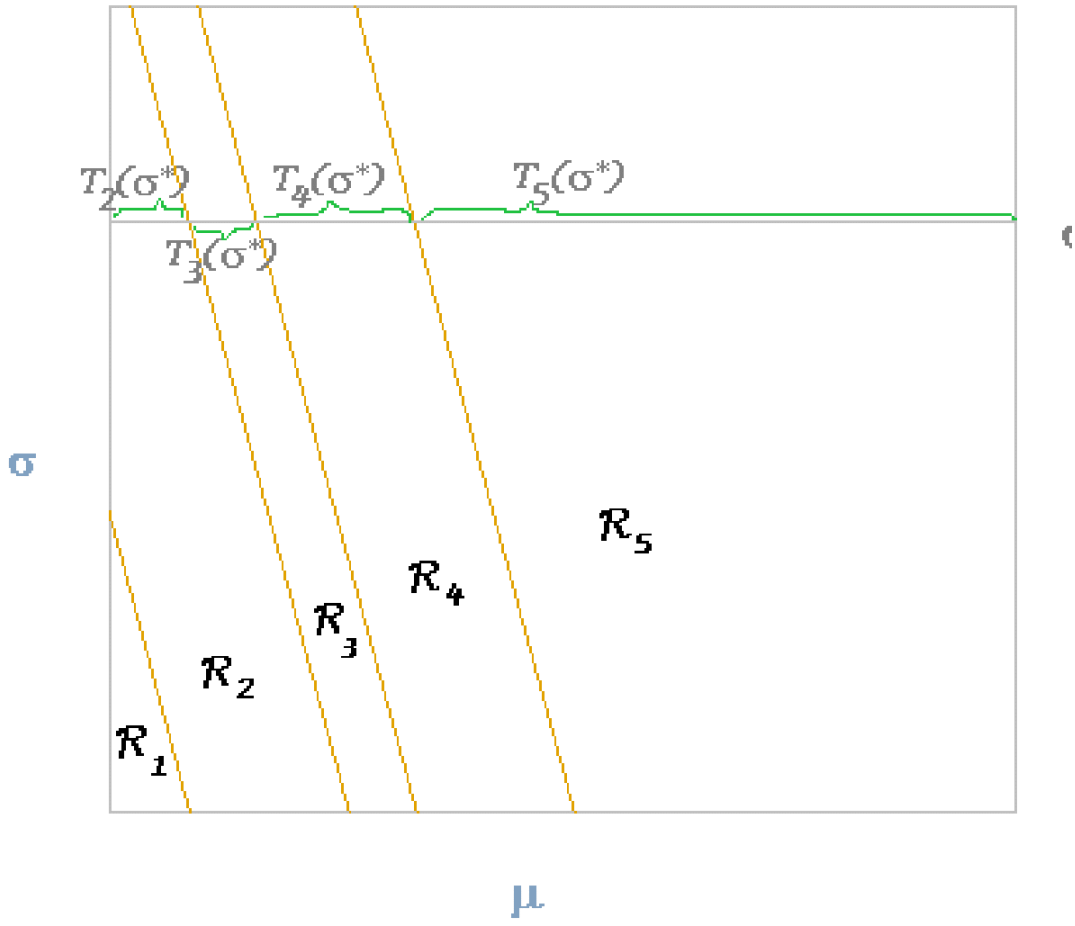


Figure 2: Partitioning μ-domain for a value $\sigma = \sigma^*$.

For the values $\mu \in (\mu_0, \mu_1)$, the couple (μ, σ^*) falls on either of the segments $T_2(\sigma^*)$, $T_3(\sigma^*)$, $T_4(\sigma^*)$ or $T_5(\sigma^*)$ and the density for the points in each segment $T_j(\sigma^*)$ is f_j, $j = 2, \ldots, 5$. Hence

$$f(\mu|\sigma^*) = f_j \text{ for } \mu \in T_j(\sigma^*),\ j = 2, \ldots, R \tag{16}$$

and the sampling of a value from μ conditional posterior distribution $f(\mu|y, \sigma^*)$ is trivial.

The full conditional posterior distribution for σ is given by

$$f(\sigma|y, \mu) \propto \frac{1}{\sigma^N} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right\} f(\sigma|\mu)\ ,$$

that is, we can sample σ values from its conditional posterior distribution using the algorithm presented in Appendix A with

$$\begin{aligned} a &= N \text{ and} \\ b &= \frac{1}{2}\sum_i (y_i - \mu)^2, \end{aligned}$$

with sections defined through $f(\sigma|\mu)$, weighed by the corresponding f-weights. The partition emerging from $f(\sigma|\mu)$ can be derived the same way as μ-partition was derived from $f(\mu|\sigma)$ above.

5.1 Problems with the algorithm suggested by Banerjee

The model and the algorithm described in previous section follow the idea described in Banerjee's paper, but NOT the algortihm he suggested.

Erdogan Gunel raised the fact that the algorithm suggested by Banerjee does not provide a sample from the different regions $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_R$ with probabilities $p_1, p_2, \ldots, p_R$ as expected. Even though his demonstration — based on simulation — is incorrect (we reproduced his results by assuming an incorrect understanding of Banerjee's algorithm) and exacerbates the shift between observed proportions of pairs (μ, σ) in each region $\mathcal{R}_1$ to $\mathcal{R}_R$ and the corresponding proportions p_1 to p_R, he is right: the algorithm suggested by Banerjee does not provide a prior sample from each region with the expected probabilities p_1 to p_R.

Indeed, if the joint prior distribution for (μ, σ) is constant within each region R_j, then the marginal prior distribution for σ is **not** uniform — unless the region $\mathcal{R}_j$ is paralleliped-shaped, which is not the case for all regions (see regions $\mathcal{R}_1, \mathcal{R}_2$ and $\mathcal{R}_5$ in Figure 1); it is particularly easy to realize when considering the triangle-shaped region $\mathcal{R}_1$ from that Figure: consider $\sigma_1 > \sigma_0$ both falling in that triangle. The set of values $(\mu, \sigma_1) \in \mathcal{R}_1$ is clearly smaller than the set $(\mu, \sigma_0) \in \mathcal{R}_1$, and hence $f(\sigma_1) < f(\sigma_0)$, which shows that the marginal distribution for σ is NOT uniform; however, the algorithm suggested by Banerjee uses a uniform distribution as the marginal distribution for σ, which is clearly incorrect.

Writing a correct model in either WinBUGS or RJags for Banerjee's model is made difficult by the need to write a piecewise-uniform prior distribution for (μ, σ). Writing a Gibbs model (in R) based on repetitive sampling from the full conditional distributions $f(\mu|y, \sigma)$ and $f(\sigma|y, \mu)$ removes that difficulty as we have seen in the previous section.

The algorithm in the previous section also addresses the incorrect piecewise-density function suggested by Banerjee. Indeed, if $f(\mu, \sigma) = p_j$ for $(\mu, \sigma) \in \mathcal{R}_j$, as Banerjee suggests, then the prior probabilities for each region is $P\{(\mu, \sigma) \in \mathcal{R}_j\} \propto f_j \times S_j$, which can be far off the expected probabilites $p_1, p_2, \ldots, p_R$ when the areas $S_1, S_2, \ldots, S_R$ are dissimilar.

Finally, a word on the model suggested by Gunel: the author suggests a normal prior distribution for μ and an inverted-gamma distribution for σ, which can differ substantially from Banerjee's model. More importantly, the model does not include any prior knowledge about the probabilities of (μ, σ) in each region, even though the author seemed to consider the lack of control of the latter as the main caveat of Banerjee's model; hence it can hardly be presented as an alternative to Banerjee's model. Finally, the inverted-gamma prior distribution on σ decreases the model's value in practice.

5.2 Implementation in RJags

In an RJags version of Banerjee's algorithm, we need to sample (μ, σ) from their joint prior distribution $f(\mu, \sigma)$, presented in the above sections: that will be done through first sampling σ from its marginal prior distribution $f(\sigma)$ and then sampling μ from its conditional prior distribution $f(\mu|\sigma)$.

Suppose that the ranges for μ and σ and the cut-off values $A_1, A_2, \ldots, A_{R-1}$ are such that the full space for (μ, σ) corresponds to Figure 3.

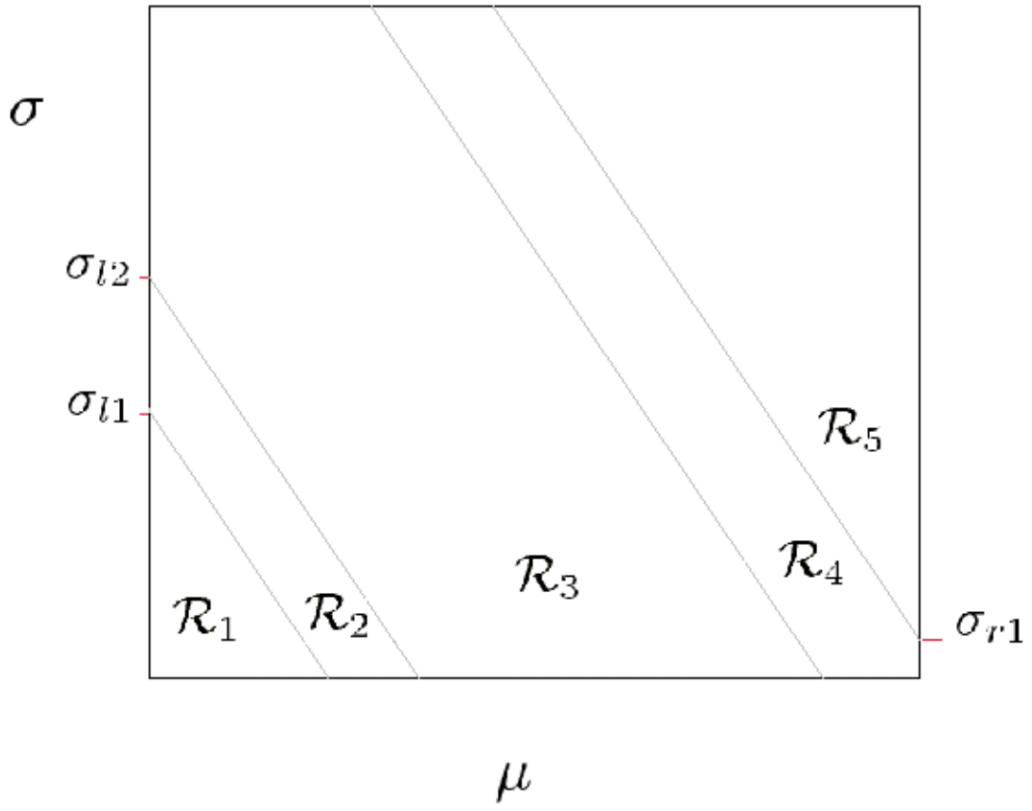


Figure 3: Regions and σ intercepts on left- and right-hand sides

It is easy to find the intercepts of the diagonal lines $\mu + z\sigma = A_i, i = 1, 2, \ldots, R$ with both the left-hand and the right-hand side borders of (μ, σ)'s domain — in the above figure, the values $(\sigma_{l1}, \sigma_{l2})$ and σ_{r1}, respectively. These values, once sorted, give the limits of J ($J = 4$ in this example) distinct intervals for σ, $\Sigma_1, \Sigma_2, \ldots, \Sigma_J$ such that $\cup\Sigma_j$ is equal to the whole domain for σ, as sketched in Figure 4.

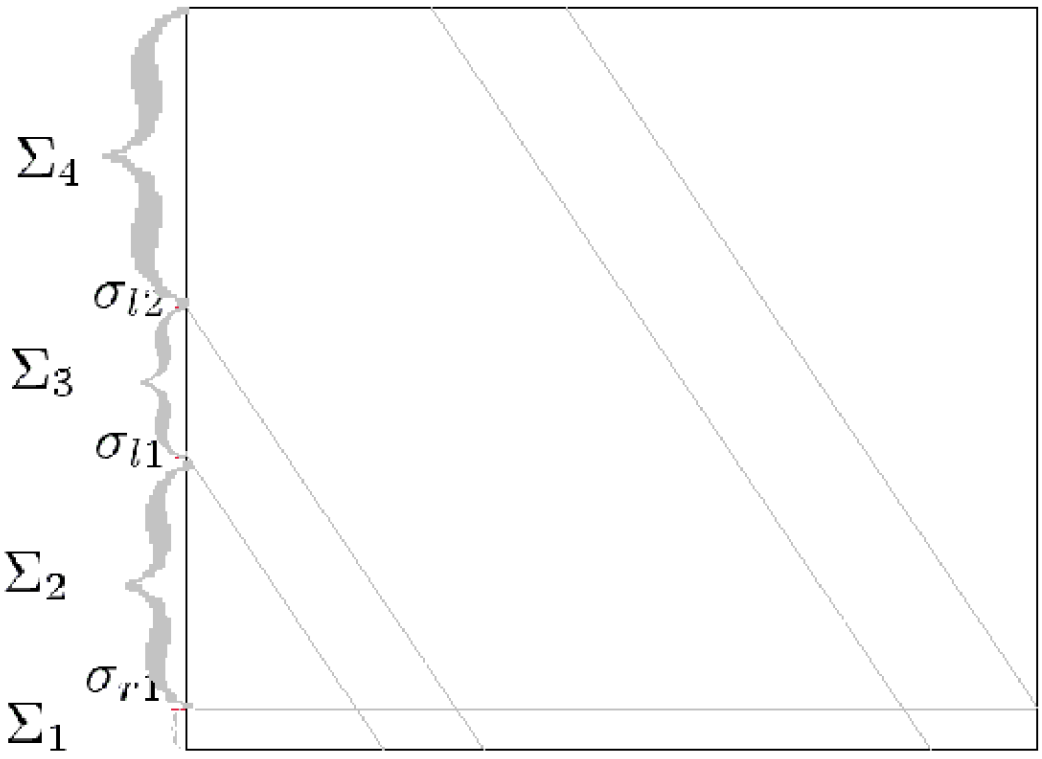


Figure 4: Segmenting σ's domain

The marginal density $f(\sigma)$ is easily calculated as

$$f(\sigma) = \sum_{j=1}^{R} l(T_j) f_j \tag{17}$$

where $l(T_j)$ is the length of the interval $T_j(\sigma)$, introduced in Figure 2. It is easy to see that $f(\sigma)$ is linear on each segment $\Sigma_1, \Sigma_2, \ldots, \Sigma_J$ and hence that the marginal prior distribution $f(\sigma)$ is piecewise-linear, as presented in Figure 5 below.

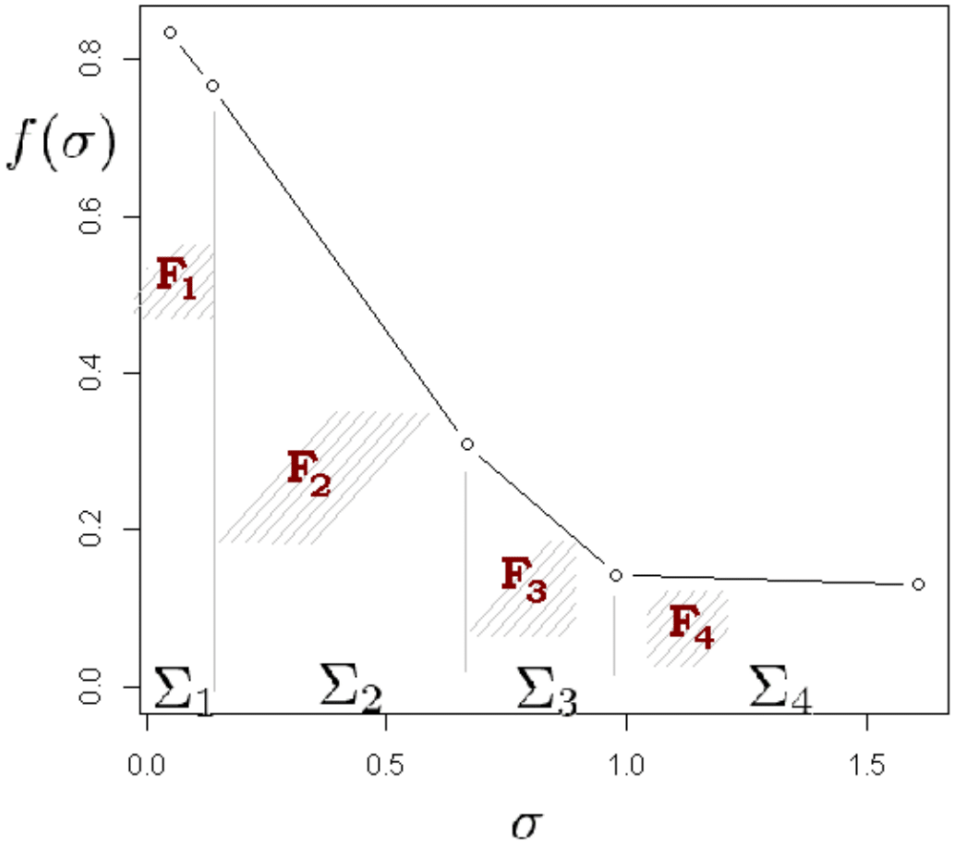


Figure 5: Piecewise-linear marginal prior density function $f(\sigma)$

Hence the marginal cumulative density function for σ, defined as $F(\sigma) = \int_{\sigma_0}^{\sigma} f(\sigma')d\sigma'$ is piecewise-quadratic. Let label Q_j the quadratic form of $F(\sigma)$ over the interval $\Sigma_j, j = 1, 2, \ldots, J$. Then one can sample σ through inverse cumulative density function by first sampling

$$U_\sigma \sim \text{Uniform}(0, 1)$$

and then solving

$$\sigma_j = Q_j^{-1}(U_\sigma)$$

for $j = 1, 2, \ldots, J$. The value σ_j is then accepted only if it is in the interval Σ_j: only one of $\{\sigma_1, \sigma_2, \ldots, \sigma_J\}$ is thus accepted, and the sample value for σ for this iteration is assigned the value of that unique solution.

Given a value sampled for σ, the conditional prior density for $f(\mu|\sigma)$ is trivially defined as

$$f(\mu|\sigma) = f_j \quad \text{if } \mu \in T_j(\sigma)$$

and sampling from it is also straightforward, as the conditional marginal cumulative density function for μ given σ is piecewise-linear. If we label L_j the linear equation describing $F(\mu|\sigma)$ over $T_j(\sigma), j = 1, 2, \ldots, R$, we can the sample a value μ by first sampling

$$U_\mu \sim \text{Uniform}(0, 1)$$

and then solving

$$\mu_j = L_j^{-1}(U_\mu)$$

for $j = 1, 2, \ldots, R$. The value μ_j is then accepted only if it is in the interval $T_j(\sigma)$: only one of $\{\mu_1, \mu_2, \ldots, \mu_R\}$ is thus accepted, and the sample value for μ for this iteration is assigned the value of that unique solution.

6 Modification of the algorithms in the presence of measurement error

The previous sections address the different models in the absence of measurement error. When measurement error exists, however, all need little adjustments. In the next two sections, we introduce two measurement error models. We first introduce the classical measurement error, assuming that the sd of the measurement error is the same for each observation — a model that appears to be of little interest in the context of Industrial Hygiene. In Section 6.2, we will thus introduce a second measurement error model where the scale of the measurement error is specified through a coefficient of variation ν, that is, where the sd of the measurement error is assumed to be some percentage of the (unobserved) true value.

In both measurement error models, the true values $T_i, i = 1, \ldots, N$ are latent variables. Depending on the model, T_i is assumed a Normal or a log-Normal distribution, that is, the likelihood is either

$$f(T_i|\mu,\sigma) = \begin{cases} f_1(T_i,\mu,\sigma) = \frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(T_i-\mu)^2\right\} & \text{when } T_i \sim N(\mu,\sigma^2) \\ f_2(T_i,\mu,\sigma) = \frac{1}{T_i\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\log(T_i)-\mu)^2\right\} & \text{when } T_i \sim \text{logN}(\mu,\sigma^2) \end{cases} \tag{18}$$

6.1 Classical measurement error model

In the classical measurement error model, we assume that the observed values Y_i are normally distributed around the corresponding true values T_i with a standard deviation ξ independent of true value T_i and for which we assume a uniform prior distribution

$$\xi \sim \text{U}(\xi_0, \xi_1)\ .$$

Given the parameters (μ, σ) from the density of T_i, this measurement error model assumes that

$$Y_i|T_i \sim N(T_i, \xi^2),$$

that is,

$$f(Y_i|T_i,\xi) = g_1(Y_i,T_i,\xi) \propto \frac{1}{\xi}\exp\left\{-\frac{1}{2\xi^2}(Y_i-T_i)^2\right\}. \tag{19}$$

The full conditional posterior distribution for ξ is

$$f(\xi|Y,T) \propto \frac{1}{\xi^N}\exp\left\{-\frac{1}{\xi^2}\underbrace{\frac{1}{2}\sum_i (Y_i-T_i)^2}_{\beta}\right\} I_\xi(\xi_0,\xi_1)\ .$$

.

Hence we can easily sample from ξ's posterior distribution through a (truncated) Inverted-Gamma distribution if $N > 1$, or the icdf algorithm presented in Appendix A when $N \leq 1$.

6.1.1 Modification when the outcome is log-normally distributed

When the true values T_i are assumed to be log-normally distributed, their values will be positive. Consequently, assuming that the observed values Y_i are normally distributed around the latent true values seems unnatural, since it leads to non-zero probabilities of getting negative measured values; hence we propose the use of a normal distribution (around true values) *restricted to the positive domain* when the outcome is log-normally distributed.

In this context, the data distribution (19) being restricted to values $Y_i > 0$ must be divided by the standardizing constant $\kappa = \Pr(Y_i > 0) = \Phi(T_i/\xi)$ to integrate to 1; that is, we have

$$f(Y_i|T_i,\xi) = g_1^*(Y_i,T_i,\xi) \propto \frac{1}{\xi}\exp\left\{-\frac{1}{2\xi^2}(Y_i-T_i)^2\right\}\cdot\frac{1}{\Phi\left(\frac{T_i}{\xi}\right)} \tag{20}$$

and the full conditional posterior distribution for ξ must be modified accordingly, leading to

$$f(\xi|Y,T) \propto \frac{1}{\xi^N\prod_i\Phi\left(\frac{T_i}{\xi}\right)}\exp\left\{-\frac{1}{\xi^2}\underbrace{\frac{1}{2}\sum_i(Y_i-T_i)^2}_{\beta}\right\} I_\xi(\xi_0,\xi_1)\ . \tag{21}$$

Sampling values from the above distribution requires the use of the inverse cumulative density function method; the computation of the first two derivatives of $\log(f)$ necessary to do so is relegated to Section 8.

6.2 Measurement error specified through a coefficient of variation

In this measurement error model, we assume that the values Y_i are normally distributed around the corresponding true values T_i with a standard deviation equal to some percentage of the true value; that is, we assume a coefficient of variation ν known within some relatively close lower and upper limits ν_0 and ν_1, respectively, assume a uniform prior distribution

$$\nu \sim \mathrm{U}(\nu_0,\nu_1)$$

— e.g. $\nu \sim \mathrm{U}(15\%, 17\%)$ — and assume that the measurements Y_i are normally distributed around the true values T_i with standard deviation equal to $\nu \times T_i$, that is

$$Y_i|T_i \sim N(T_i, \mathrm{sd} = \nu T_i). \tag{22}$$

Before continuing on the elicitation of the full posterior distrbutions for parameter ν, we must emphasize on the latter making sense if and only if the values of $T_i, i = 1,2,\ldots,N$ are positive.

Warning: a restriction when the outcome is normally distributed. The data conditional (on the true values) distribution (22) makes sense if the outcome is log-normally distributed; however, if the data are normally distrbuted, there is a non-null probability that $T_i < 0$ and hence a non-null probability that $\mathrm{sd}(Y_i|T_i) = \nu T_i < 0$, which is obviously problematic! A work-around is to assume that the true values are **not** normally distributed, but rather normally distributed with values restricted to its positive domain, that is,

$$\begin{aligned} f(t_i|\mu,\sigma) &\propto f_{N(\mu,\sigma^2)}(t_i)I(t_i>0) \\ &= \frac{1}{\kappa}\cdot\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{t_i-\mu}{\sigma}\right)^2\right\}I(t_i>0) \end{aligned} \tag{23}$$

where κ is a standardizing constant. In order to have f integrate to 1, we must set

$$\kappa = \Pr(t_i>0) = \Phi\left(\frac{\mu}{\sigma}\right)\ ;$$

that being a function of μ and σ, the posterior distributions developed earlier for these two parameters clearly need to be adjusted: that work is relegated to Section 7.

Since the range for true values T_i is restricted to positive values, it seems natural to confine the measured values Y_i to the positive domain as well, and hence the conditional distribution of data Y_i given true values T_i, including the appropriate standardizing constant $\kappa = \Pr(Y_i > 0) = \Phi(T_i/T_i\nu) = \Phi(1/\nu)$, is given by

$$f(Y_i|T_i,\nu) = g_2(Y_i,T_i,\nu) \propto \frac{1}{T_i\nu}\exp\left\{-\frac{1}{2T_i^2\nu^2}(Y_i-T_i)^2\right\}\cdot\frac{1}{\Phi(1/\nu)}. \tag{24}$$

The full conditional distribution for the coefficient of variation ν is proportional to

$$f(\nu|Y,T) \propto \frac{1}{\nu^N\Phi^N(1/\nu)}\exp\left\{-\frac{1}{\nu^2}\underbrace{\frac{1}{2}\sum_i\left(\frac{Y_i}{T_i}-1\right)^2}_{\beta}\right\} I_\nu(\nu_0,\nu_1)\ . \tag{25}$$

6.3 Conditional posterior distribution for T_i

In presence of measurement error, the full conditional posterior distribution for true value T_i is given by

$$f(T_i|Y_i,\mu,\sigma,\xi \text{ or } \nu) \propto f(Y_i|T_i,\xi \text{ or } \nu)f(T_i|\mu,\sigma^2) \tag{26}$$

which is equal to either $f_1g_1, f_1g_2, f_2g_1^*$ or f_2g_2 — described in (18), (19), (20) and (24) — depending on whether T_i is assumed to have a normal or a log-normal distribution and on the measurement error model.

The easiest case is when T_i is assumed to have a normal distribution and a classical measurement error model: in that scenario, the full posterior distribution for T_i is given by

$$\begin{aligned}
f(Y_i|T_i,\xi)f(T_i|\mu,\sigma^2) &\propto \exp\left\{-\frac{1}{2\xi^2}(Y_i-T_i)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}(T_i-\mu)^2\right\}\\
&= \exp\left\{-\frac{1}{2\xi^2}(Y_i^2-2Y_iT_i+T_i^2)\right\}\exp\left\{-\frac{1}{2\sigma^2}(T_i^2-2T_i\mu+\mu^2)\right\}\\
&\propto \exp-\frac{1}{2}\left\{T_i^2\underbrace{\left(\frac{1}{\xi^2}+\frac{1}{\sigma^2}\right)}_{\tau^*} - 2T_i\underbrace{\left(\frac{Y_i}{\xi^2}+\frac{\mu}{\sigma^2}\right)}_{\mu_i^*}\right\}\\
&= \exp-\frac{\tau^*}{2}\left\{T_i^2-2T_i\frac{\mu_i^*}{\tau^*}\right\}\\
\Longrightarrow T_i|Y_i &\sim N\left(\frac{\mu_i^*}{\tau^*}, \text{precision}=\tau^*\right)\ .
\end{aligned} \tag{27}$$

6.3.1 When the measurement error is specified through a Coefficient of Variation and the outcome is log-normally distributed

The other three scenarios lead to conditional posterior distributions for T_i that are similar to each other but more complex than the above. For example, in the context of an outcome that is log-normally distributed and with a measurement error specified through a coefficient of variation (see Section 6.2), the full conditional posterior distribution for T_i is given by

$$\begin{aligned}
f(T_i|Y_i,\nu,\mu,\sigma) \propto f_2(T_i,\mu,\sigma,\nu,Y_i) &\propto \frac{1}{\nu T_i}\exp\left\{-\frac{1}{2\nu^2T_i^2}(Y_i-T_i)^2\right\}\cdot\frac{1}{T_i}\exp\left\{-\frac{1}{2\sigma^2}(\log(T_i)-\mu)^2\right\}\\
&\propto \frac{1}{T_i^2}\exp\left\{-\frac{1}{2\nu^2T_i^2}(Y_i-T_i)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}(\log(T_i)-\mu)^2\right\}\ .
\end{aligned}$$

As one can see from the above, the full conditional distribution $f(T_i|Y_i,\nu,\mu,\sigma)$ is quite complex and there does not seem to be an easy way to sample from it; however, given values Y_i,μ,σ and ν, one can numerically evaluate the corresponding cumulative density and thus use the inverse cumulative density function algorithm to sample a random value from it. Sampling from T_i's posterior distribution in the other two scenarios is performed along the same lines.

Initial values for Y_i's are generated with a Normal or log-Normal distribution with parameters μ and σ^2 and limited to their corresponding censoring range.

At each following iteration, the values $T_i, i = 1,\ldots,N$ are generated from an icdf algorithm as mentioned above and then the censored values in $\{Y_i\}_{i=1}^N$ — if any — are generated with distributions (19) or (24) and limited to their corresponding censoring range.

6.3.2 When the measurement error is specified through a Coefficient of Variation and the outcome is normally distributed

When the outcome is normally distributed and the measurement error is specified through a Coefficient of Variation, the full conditional posterior distribution for T_i is given by

$$\begin{aligned} f(T_i|Y_i,\nu,\mu,\sigma) &\propto \frac{1}{T_i\nu\Phi(1/\nu)}\exp\left\{-\frac{1}{2T_i^2\nu^2}(Y_i-T_i)^2\right\}\exp\left\{-\frac{1}{2}\left(\frac{T_i-\mu}{\sigma}\right)^2\right\} \\ &\propto \frac{1}{T_i}\exp\left\{-\frac{1}{2T_i^2\nu^2}(Y_i-T_i)^2\right\}\exp\left\{-\frac{1}{2}\left(\frac{T_i-\mu}{\sigma}\right)^2\right\} \end{aligned}$$

for $T_i > 0$, from (23) and (24).

6.3.3 When the measurement error is classical and the outcome is log-normally distributed

Under the classical measurement error model and when the outcome is log-normally distributed, the full conditional posterior distribution for T_i is given by

$$\begin{aligned} f(T_i|Y_i,\xi,\mu,\sigma) &\propto \frac{1}{\xi}\exp\left\{-\frac{1}{2\xi^2}(Y_i-T_i)^2\right\}\frac{1}{\Phi\left(\frac{T_i}{\xi}\right)}\cdot\frac{1}{T_i\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\log(T_i)-\mu)^2\right\} \\ &\propto \frac{1}{T_i\Phi\left(\frac{T_i}{\xi}\right)}\exp\left\{-\frac{1}{2\xi^2}(Y_i-T_i)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}(\log(T_i)-\mu)^2\right\} \end{aligned} \tag{28}$$

from (18) and (20).

7 Revisiting posterior distributions in the presence of Measurement Error specified through a Coefficient of Variation when the outcome is normally distributed

As we have discussed in Section 6.2, the full conditional posterior distributions for μ,σ (and μ_k,σ_W in McNally's model) need to be revisited in presence of measurement error specified through a coefficient of variation when the outcome is normally distributed. The sections below will address these issues one model at a time, taking advantage of results developed in Appendix C.

Each of the full conditional posterior distributions developed below involve a term $\Phi\left(\frac{\mu}{\sigma}\right)$ in the denominator, making the sampling of random values from them impossible in a direct manner: we will rather use an inverse cumulative density function. For that algorithm to be efficient, we need to find the domain on which the conditional posterior density function is not negligeable, which will be done by first finding its mode and then remote left and right values (such that the density at these two endpoints is very small when compared to density at the mode). Hence we need to compute the first two derivatives of the log of each full conditional posterior distribution.

7.1 Uninformative model

The full conditional posterior distribution for μ (2) becomes

$$\begin{aligned} f &\propto \frac{e^{-\frac{N}{2\sigma^2}(\mu-\bar{y})^2}}{\Phi^N\left(\frac{\mu}{\sigma}\right)} \qquad (29)\\ \text{therefore } \log(f) &= C - \frac{N}{2\sigma^2}(\mu-\bar{y})^2 - N\log\Phi\left(\frac{\mu}{\sigma}\right) \end{aligned}$$

where the constant C can be ignored (we will omit it in the $\log(f)$ expressions throughout the remainder of this section). The two terms in $\log f$ above correspond to the functions h_4 and h_3, respectively, described in (C.5), and we thus easily find its first two derivatives

$$\begin{aligned} \frac{\partial}{\partial\mu}\log f &= -\frac{N}{\sigma^2}(\mu-\bar{y}) - \frac{N\phi}{\Phi\sigma} \qquad (30)\\ \frac{\partial^2}{\partial\mu^2}\log f &= -\frac{N}{\sigma^2} + \frac{N\phi}{\sigma^2\Phi^2}\left(\frac{\mu\Phi}{\sigma}+\phi\right) \qquad (31)\\ \text{where } \phi &= \phi\left(\frac{\mu}{\sigma}\right)\\ \text{and } \Phi &= \Phi\left(\frac{\mu}{\sigma}\right)\ . \end{aligned}$$

If we let

$$\begin{aligned} z &= \frac{\mu}{\sigma}\\ \text{and } \varphi &= \frac{\phi}{\Phi} \end{aligned}$$

then results (30) and (31) simplify to

$$\begin{aligned} \frac{\partial}{\partial\mu}\log f &= -\frac{N}{\sigma^2}(\mu-\bar{y}) - \frac{N\varphi}{\sigma} \qquad (32)\\ \frac{\partial^2}{\partial\mu^2}\log f &= -\frac{N}{\sigma^2} + \frac{N}{\sigma^2}\left(\frac{z\varphi}{\sigma^2}+\varphi^2\right)\\ &= \frac{N}{\sigma^2}\left(z\varphi+\varphi^2-1\right)\ . \qquad (33) \end{aligned}$$

The full conditional posterior distribution for σ (3) becomes

$$\begin{aligned} f &\propto \frac{1}{\sigma^N}\exp\left\{-\frac{b}{\sigma^2}\right\}\frac{1}{\Phi^N\left(\frac{\mu}{\sigma}\right)},\\ \text{therefore } \log f &= -N\log\sigma - \frac{b}{\sigma^2} - N\log\Phi\left(\frac{\mu}{\sigma}\right) \end{aligned}$$

whose terms correspond to functions h_1, h_2 and h_3 in (C.5), respectively. We thus easily find

$$\begin{aligned} \frac{\partial}{\partial\sigma}\log f &= -\frac{N}{\sigma} + \frac{2b}{\sigma^3} + \frac{N\mu\phi}{\Phi\sigma^2}\\ \text{and } \frac{\partial^2}{\partial\sigma^2}\log f &= \frac{N}{\sigma^2} - \frac{6b}{\sigma^4} + \frac{N\mu\phi}{\Phi^2\sigma^4}\left\{\frac{\mu^2}{\sigma}\Phi + \mu\phi - 2\sigma\Phi\right\}\ ; \end{aligned}$$

in terms of z and φ, the above two reduce to

$$\begin{aligned} \frac{\partial}{\partial\sigma}\log f &= -\frac{N}{\sigma} + \frac{2b}{\sigma^3} + \frac{Nz\varphi}{\sigma}\\ \text{and } \frac{\partial^2}{\partial\sigma^2}\log f &= \frac{N}{\sigma^2} - \frac{6b}{\sigma^4} + \frac{N\mu\phi}{\Phi^2\sigma^4}\left\{\frac{\mu^2}{\sigma}\Phi + \mu\phi - 2\sigma\Phi\right\}\\ &= \frac{N}{\sigma^2} - \frac{6b}{\sigma^4} + \frac{Nz}{\sigma^2}\left((z^2-2)\varphi + z\varphi^2\right). \end{aligned}$$

7.2 Kromhout model

The full conditional posterior distribution for μ (6) rewrites as (29) and therefore its first two derivatives are given by (32) and (33).

The full conditional posterior density for σ (7) rewrites

$$\begin{aligned} f &\propto \frac{1}{\sigma^{N+1}} \exp\left\{-\frac{b}{\sigma^2}\right\} \exp\left\{-\frac{(\log(\sigma)-\mu^*)^2}{2\sigma^{*2}}\right\} \frac{1}{\Phi^N\left(\frac{\mu}{\sigma}\right)} \\ \text{and } \log f &= -(N+1)\log\sigma - \frac{b}{\sigma^2} - \frac{(\log(\sigma)-\mu^*)^2}{2\sigma^{*2}} - N\log\Phi\left(\frac{\mu}{\sigma}\right) \end{aligned}$$

which terms correspond to functions h_1, h_2, h_5 and h_3 in (C.5), respectively.

Hence we easily get its first two derivatives, given by

$$\begin{aligned} \frac{\partial}{\partial\sigma}\log f &= -\frac{N+1}{\sigma} + \frac{2b}{\sigma^3} - \frac{1}{\sigma\sigma^{*2}}(\log(\sigma)-\mu^*) + \frac{N\mu\phi}{\Phi\sigma^2} \\ \text{and } \frac{\partial^2}{\partial\sigma^2}\log f &= \frac{N+1}{\sigma^2} - \frac{6b}{\sigma^4} + \frac{\log(\sigma)-\mu^*-1}{\sigma^{*2}\sigma^2} + \frac{N\mu\phi}{\Phi^2\sigma^4}\left\{\frac{\mu^2}{\sigma}\Phi + \mu\phi - 2\sigma\Phi\right\}\ ; \end{aligned}$$

in terms of z and φ, the above two reduce to

$$\begin{aligned} \frac{\partial}{\partial\sigma}\log f &= -\frac{N+1}{\sigma} + \frac{2b}{\sigma^3} - \frac{1}{\sigma\sigma^{*2}}(\log(\sigma)-\mu^*) + \frac{Nz\varphi}{\sigma} \\ \text{and } \frac{\partial^2}{\partial\sigma^2}\log f &= \frac{N+1}{\sigma^2} - \frac{6b}{\sigma^4} + \frac{\log(\sigma)-\mu^*-1}{\sigma^{*2}\sigma^2} + \frac{Nz}{\sigma^2}\left((z^2-2)\varphi + z\varphi^2\right). \end{aligned}$$

7.3 McNally model

The full conditional posterior density for μ (12) becomes

$$\begin{aligned} f &\propto \frac{\exp\left\{-\frac{N}{2\sigma_W^2}(\mu-\theta)^2\right\}}{\prod_k \Phi^{n_k}\left(\frac{\mu+\mu_k}{\sigma_W}\right)} \quad \text{where } \theta = \bar{y} - \frac{\sum n_k\mu_k}{N} \\ \text{therefore } \log f &= -\frac{N}{2\sigma_W^2}(\mu-\theta)^2 - \sum_k n_k \log\Phi\left(\frac{\mu+\mu_k}{\sigma_W}\right) \end{aligned}$$

which terms correspond to h_4 and h_3 in (C.5), respectively. Its first two derivatives are easily obtained as

$$\begin{aligned} \frac{\partial}{\partial\mu}\log f &= -\frac{N}{\sigma_W^2}(\mu-\theta) - \frac{1}{\sigma_W}\sum_k \frac{n_k\phi_k}{\Phi_k} \\ \text{and } \frac{\partial^2}{\partial\mu^2}\log f &= -\frac{N}{\sigma_W^2} + \sum_k \frac{n_k\phi_k}{\sigma_W^2\Phi_k^2}\left(\frac{(\mu+\mu_k)\Phi_k}{\sigma_W} + \phi_k\right) \\ \text{where } \phi_k &= \phi\left(\frac{\mu+\mu_k}{\sigma_W}\right) \\ \text{and } \Phi_k &= \Phi\left(\frac{\mu+\mu_k}{\sigma_W}\right). \end{aligned}$$

If we let

$$\begin{aligned} z_k &= \frac{\mu+\mu_k}{\sigma_W} \\ \text{and } \varphi_k &= \frac{\phi_k}{\Phi_k} \end{aligned}$$

then the above two derivatives reduce to

$$\begin{aligned}\frac{\partial}{\partial\mu}\log f &= -\frac{N}{\sigma_W^2}(\mu-\theta)-\frac{1}{\sigma_W}\sum_k n_k\varphi_k \\ \text{and } \frac{\partial^2}{\partial\mu^2}\log f &= -\frac{N}{\sigma_W^2}+\frac{1}{\sigma_W^2}\sum_k n_k(z_k\varphi_k+\varphi_k^2)\ .\end{aligned}$$

The full conditional posterior density for μ_k (11) becomes

$$\begin{aligned} f &\propto \frac{\exp\left\{-\frac{1}{2\sigma_k^{*2}}(\mu_k-\mu_k^*)^2\right\}}{\Phi^{n_k}\left(\frac{\mu+\mu_k}{\sigma_W}\right)} \\ \text{therefore } \log f &= -\frac{1}{2\sigma_k^{*2}}(\mu_k-\mu_k^*)^2-n_k\log\Phi\left(\frac{\mu+\mu_k}{\sigma_W}\right)\end{aligned}$$

where μ_k^* and σ_k^{*2} are the mean and variance of the distribution in (11). The components of $\log f$ above correspond to h_4 and h_3 in (C.5), respectively. Its first two derivatives are easily obtained as

$$\begin{aligned}\frac{\partial}{\partial\mu_k}\log f &= -\frac{1}{\sigma_k^{*2}}(\mu_k-\mu_k^*)-\frac{n_k\phi_k}{\Phi_k\sigma_W} \\ \text{and } \frac{\partial^2}{\partial\mu_k^2}\log f &= -\frac{1}{\sigma_k^{*2}}+\frac{n_k\phi_k}{\sigma_W^2\Phi_k^2}\left(\frac{(\mu+\mu_k)\Phi_k}{\sigma_W}+\phi_k\right)\ ;\end{aligned}$$

in terms of z_k and φ_k, the above two reduce to

$$\begin{aligned}\frac{\partial}{\partial\mu_k}\log f &= -\frac{1}{\sigma_k^{*2}}(\mu_k-\mu_k^*)-\frac{n_k\varphi_k}{\sigma_W} \\ \text{and } \frac{\partial^2}{\partial\mu_k^2}\log f &= -\frac{1}{\sigma_k^{*2}}+\frac{n_k}{\sigma_W^2}(z_k\varphi_k+\varphi_k^2)\ .\end{aligned}$$

When the prior distribution for σ_W is log-normal, its posterior distribution (13) becomes

$$\begin{aligned} f &\propto \frac{1}{\sigma_W^{N+1}}\exp\left\{-\frac{b}{\sigma_W^2}\right\}\exp\left\{-\frac{1}{2\sigma^{*2}}\left(\log(\sigma_W)-\mu^*\right)^2\right\}\frac{1}{\prod_k\Phi^{n_k}\left(\frac{\mu+\mu_k}{\sigma_W}\right)} \\ \text{therefore } \log f &= -(N+1)\log\sigma_W-\frac{b}{\sigma_W^2}-\frac{1}{2\sigma^{*2}}\left(\log(\sigma_W)-\mu^*\right)^2-\sum_k n_k\log\Phi\left(\frac{\mu+\mu_k}{\sigma_W}\right)\end{aligned}$$

whose terms correspond to h_1, h_2, h_5 and h_3 in (C.5), respectively. Its first two derivatives are easily obtained as

$$\begin{aligned}\frac{\partial}{\partial\sigma_W}\log f &= -\frac{N+1}{\sigma_W}+\frac{2b}{\sigma_W^3}-\frac{1}{\sigma_W\sigma^{*2}}(\log(\sigma_W)-\mu^*)+\frac{1}{\sigma_W^2}\sum_k\frac{n_k(\mu+\mu_k)\phi_k}{\Phi_k} \\ \text{and } \frac{\partial^2}{\partial\sigma_W^2}\log f &= \frac{N+1}{\sigma_W^2}-\frac{6b}{\sigma_W^4}+\frac{\log(\sigma_W)-\mu^*-1}{\sigma^{*2}\sigma_W^2} \\ &\quad +\sum_k\frac{n_k(\mu+\mu_k)\phi_k}{\Phi_k^2\sigma_W^4}\left\{\frac{(\mu+\mu_k)^2}{\sigma_W}\Phi_k+(\mu+\mu_k)\phi_k-2\sigma_W\Phi_k\right\}\ ;\end{aligned}$$

in terms of z_k and φ_K, the above two reduce to

$$\begin{aligned}\frac{\partial}{\partial\sigma_W}\log f &= -\frac{N+1}{\sigma_W}+\frac{2b}{\sigma_W^3}-\frac{1}{\sigma_W\sigma^{*2}}(\log(\sigma_W)-\mu^*)+\frac{1}{\sigma_W}\sum_k n_k z_k\varphi_k \\ \text{and } \frac{\partial^2}{\partial\sigma_W^2}\log f &= \frac{N+1}{\sigma_W^2}-\frac{6b}{\sigma_W^4}+\frac{\log(\sigma_W)-\mu^*-1}{\sigma^{*2}\sigma_W^2}+\frac{1}{\sigma_W^2}\sum_k n_k z_k((z_k^2-2)\varphi_k+z_k\varphi_k^2)\ .\end{aligned}$$

When the prior distribution for σ_W is uniform, its posterior distribution (14) rewrites

$$\begin{aligned} f &\propto \frac{1}{\sigma_W^N} \exp\left\{-\frac{b}{\sigma_W^2}\right\} \frac{1}{\prod_k \Phi^{n_k}\left(\frac{\mu+\mu_k}{\sigma_W}\right)} \\ \text{therefore } \log f &= -N\log\sigma_W - \frac{b}{\sigma_W^2} - \sum_k n_k \log\Phi\left(\frac{\mu+\mu_k}{\sigma_W}\right) \end{aligned}$$

which terms correspond to h_1, h_2 and h_3 in (C.5), respectively. Its first two derivatives are easily obtained as

$$\begin{aligned} \frac{\partial}{\partial\sigma_W}\log f &= -\frac{N}{\sigma_W} + \frac{2b}{\sigma_W^3} + \frac{1}{\sigma_W^2}\sum_k \frac{n_k(\mu+\mu_k)\phi_k}{\Phi_k} \\ \text{and } \frac{\partial^2}{\partial\sigma_W^2}\log f &= \frac{N}{\sigma_W^2} - \frac{6b}{\sigma_W^4} + \sum_k \frac{n_k(\mu+\mu_k)\phi_k}{\Phi_k^2\sigma_W^4}\left\{\frac{(\mu+\mu_k)^2}{\sigma_W}\Phi_k + (\mu+\mu_k)\phi_k - 2\sigma_W\Phi_k\right\} ; \end{aligned}$$

in terms of z_k and φ_K, the above two reduce to

$$\begin{aligned} \frac{\partial}{\partial\sigma_W}\log f &= -\frac{N}{\sigma_W} + \frac{2b}{\sigma_W^3} + \frac{1}{\sigma_W}\sum_k n_k z_k \varphi_k \\ \text{and } \frac{\partial^2}{\partial\sigma_W^2}\log f &= \frac{N}{\sigma_W^2} - \frac{6b}{\sigma_W^4} + \frac{1}{\sigma_W^2}\sum_k n_k z_k\left\{(z_k^2-2)\varphi_k + z_k\varphi_k^2\right\} . \end{aligned}$$

7.4 Banerjee model

Since the posterior distributions for μ and σ under the Banerjee model are piecewise, the adjustments needed in the presence of measurement error specified through a coefficient of variation when the outcome is normally distributed are the same as for uninformative model (see Section 7.1).

7.5 Posterior distribution for ν

Whether the outcome is normally or log-normally distributed, the posterior distribution for the coefficient of variation ν (25) becomes

$$\begin{aligned} f &\propto \frac{1}{\nu^N \Phi^N\left(\frac{1}{\nu}\right)} \exp\left\{-\frac{\beta}{\nu^2}\right\} \\ \text{therefore } \log f &= -N\log\nu - \frac{\beta}{\nu^2} - N\log\Phi\left(\frac{1}{\nu}\right) \end{aligned}$$

whose terms correspond to functions h_1, h_2 and h_3 in (C.5), respectively. Its first two derivatives can thus be easily written as

$$\begin{aligned} \frac{\partial}{\partial\nu}\log f &= -\frac{N}{\nu} + \frac{2\beta}{\nu^3} + \frac{N\phi}{\Phi\nu^2} \\ \text{and } \frac{\partial^2}{\partial\nu^2}\log f &= \frac{N}{\nu^2} - \frac{6\beta}{\nu^4} + \frac{N}{\Phi^2\nu^4}\left\{\frac{\phi}{\nu^3}\Phi\nu^2 - \phi\left[-\frac{\phi}{\nu^2}\nu^2 + 2\Phi\nu\right]\right\} \\ &= \frac{N}{\nu^2} - \frac{6\beta}{\nu^4} + \frac{N\phi}{\Phi^2\nu^4}\left\{\frac{\Phi}{\nu} + \phi - 2\Phi\nu\right\} \\ \text{where } \phi &= \phi\left(\frac{1}{\nu}\right) \\ \text{and } \Phi &= \Phi\left(\frac{1}{\nu}\right) ; \end{aligned}$$

if we let $\varphi = \phi/\Phi$, then the above two results reduce to

$$\begin{aligned} \frac{\partial}{\partial\nu}\log f &= -\frac{N}{\nu} + \frac{2\beta}{\nu^3} + \frac{N\varphi}{\nu^2} \\ \text{and } \frac{\partial^2}{\partial\nu^2}\log f &= \frac{N}{\nu^2} - \frac{6\beta}{\nu^4} + \frac{N}{\nu^4}\left\{\varphi\left(\frac{1}{\nu} - 2\nu\right) + \varphi^2\right\} . \end{aligned}$$

8 Posterior distribution for classical measurement error parameter (ξ) when the outcome is log-normally distributed

Sampling values for ξ from its posterior distribution when the outcome is log-normally distributed (see equation 21) also requires the computation of the first two derivatives of $\log(f)$.

Since

$$\log(f) = -N\log(\xi) - \sum_i \log\Phi\left(\frac{T_i}{\xi}\right) - \frac{\beta}{\xi^2}$$

where the second term corresponds to h_3 in (C.5) — replacing $\mu+\theta$ by T_i and σ by ξ — we easily get

$$\frac{\partial}{\partial\xi}\log(f) = -\frac{N}{\xi} + \sum_i \frac{1}{\Phi\left(\frac{T_i}{\xi}\right)}\cdot\phi\left(\frac{T_i}{\xi}\right)\frac{T_i}{\xi^2} + \frac{2\beta}{\xi^3}$$

from derivation chain rule and

$$\begin{aligned} \frac{\partial^2}{\partial\xi^2}\log(f) &= \frac{N}{\xi^2} + \sum_i \frac{T_i\phi_i}{\Phi_i^2\xi^4}\left\{\frac{\Phi_i T_i^2}{\xi} + \phi_i - 2\xi\Phi_i\right\} - \frac{6\beta}{\xi^4} \\ &\quad \text{where } \phi_i = \phi\left(\frac{T_i}{\xi}\right) \text{ and } \Phi_i = \Phi\left(\frac{T_i}{\xi}\right) \\ &= \frac{N}{\xi^2} + \sum_i \frac{T_i\varphi_i}{\xi^5}\left\{T_i^2 + \xi\varphi_i T_i - 2\xi^2\right\} - \frac{6\beta}{\xi^4} \\ &\quad \text{where } \varphi_i = \frac{\phi_i}{\Phi_i}. \end{aligned}$$

A Generating values for σ from its inverse cumulative density function

If U is a random variable with a Uniform(0,1) density, then the variable $X = F^{-1}(U)$ has the cumulative density function F.

This method will be used in the context of WebExpo for σ when its conditional posterior distribution is given by either

$$f(\sigma|y, \text{other parameters}) \propto \frac{1}{\sigma^{a+1}} \exp\left\{-b/\sigma^2\right\} \exp\left\{-\frac{(\log(\sigma) - \tilde{\mu})^2}{2\tilde{\sigma}^2}\right\} ,$$

as is the case in the Kromhout, Two-Level Kromhout and McNally models, or

$$f(\sigma|y, \text{other parameters}) \propto \frac{1}{\sigma^{a}} \exp -b/\sigma^2 ,$$

as is the case in the Banerjee model and the uninformative model when the prior on σ is uniform.

In either case, the cumulative density function $F(\sigma) = \int_{-\infty}^{\sigma} f(\sigma')d\sigma'$ does not have an analytic solution but can be estimated numerically in R with the `integrate()` function for any value σ.

Hence, one can sample a value U from a uniform $U(0, 1)$ distribution and use a Newton-Raphson algorithm to find the value for σ such that

$$\frac{F(\sigma) - F(\sigma_0)}{F(\sigma_1) - F(\sigma_0)} = U$$

where (σ_0, σ_1) are the boundaries of the σ-domain; the resulting value σ is thus sampled from its corresponding f posterior density.

B Distribution of σ when precision τ is Gamma distributed

When the precision $\tau = 1/\sigma^2$ is Gamma-distributed, that is, when

$$\begin{aligned} \tau &\sim \text{Gamma}(\alpha, \beta) \\ \text{or } f(\tau) &\propto \tau^{\alpha-1} e^{-\beta\tau} \end{aligned}$$

then the distribution of σ is given by

$$f(\sigma) \propto \frac{1}{\sigma^{2(\alpha-1)}} \exp\left\{-\frac{\beta}{\sigma^2}\right\} \cdot \sigma^{-3}$$

since

$$\begin{aligned} \tau &= \sigma^{-2} \\ \text{and } \frac{d\tau}{d\sigma} &= -2\sigma^{-3}. \end{aligned}$$

Hence

$$\begin{aligned} f(\sigma) &\propto \frac{1}{\sigma^{2\alpha-2+3}} \exp\left\{-\frac{\beta}{\sigma^2}\right\} \\ &= \frac{1}{\sigma^{2\alpha+1}} \exp\left\{-\frac{\beta}{\sigma^2}\right\} \end{aligned} \tag{B.1}$$

Conversely, if

$$\begin{aligned} f(\sigma) &\propto \frac{1}{\sigma^a} \exp\left\{-\frac{\beta}{\sigma^2}\right\} \\ \text{then } \tau &\sim \text{Gamma}\left(\frac{a-1}{2}, \beta\right). \end{aligned} \tag{B.2}$$

C Derivatives

C.1 First derivative of ϕ and Φ

Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the normal density and cumulative density functions, respectively, that is

$$\begin{aligned} \phi(z) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \\ \text{and } \Phi(z) &= \int_{-\infty}^{z} \phi(x)dx \, . \end{aligned}$$

If $u = u(x)$, then the first derivatives of $\phi = \phi(u(x))$ and $\Phi = \Phi(u(x))$ are

$$\begin{aligned} \phi' &= \frac{d}{dx}\phi(u(x)) \\ &= \frac{1}{\sqrt{2\pi}} \frac{d}{dx} e^{-\frac{u^2(x)}{2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2(x)}{2}} \cdot -\frac{2u(x)}{2} \cdot u' \\ &= -\phi u u' \end{aligned} \tag{C.1}$$

and

$$\begin{aligned} \Phi' &= \frac{d}{dx}\Phi(u(x)) \\ &= \phi(u(x)) \cdot \frac{d}{dx} u(x) \\ &= \phi u' \, . \end{aligned} \tag{C.2}$$

C.2 Derivatives of $\log(\Phi(g(z)))$

The first two derivatives of $\log(\Phi(g(z)))$ are given by

$$\begin{aligned} \frac{\partial}{\partial z} \log(\Phi(g(z))) &= \frac{1}{\Phi(g(z))} \cdot \phi(g(z))g'(z) \\ &\quad \text{from derivation chain rule} \\ &= \varphi(g(z))g'(z) \text{ where } \varphi(z) = \frac{\phi(z)}{\Phi(z)} \end{aligned} \tag{C.3}$$

and

$$\begin{aligned} \frac{\partial^2}{\partial z^2} \log(\Phi(g(z))) &= \frac{\partial}{\partial z}\left[\frac{\phi \cdot g'}{\Phi}\right] \\ &= \frac{[\phi g']'\Phi - \phi g'\Phi'}{\Phi^2} \\ &= \frac{(\phi' g' + \phi g'')\Phi - \phi g'\phi g'}{\Phi^2} \text{ from (C.2)} \\ &= \frac{\phi' g' + \phi g''}{\Phi} - \left[\frac{\phi g'}{\Phi}\right]^2 \\ &= \frac{-\phi g g'^2 + \phi g''}{\Phi} - \left[\frac{\phi g'}{\Phi}\right]^2 \text{ from (C.1)} \\ &= \frac{\phi}{\Phi}\left[-g g'^2 + g''\right] - \left[\frac{\phi g'}{\Phi}\right]^2 \\ &= \varphi\left[-g g'^2 + g''\right] - \left[\frac{\phi g'}{\Phi}\right]^2 . \end{aligned} \tag{C.4}$$

C.3 Derivatives of functions involved in full conditional posterior distributions

The posterior distributions elicited in Section 7 are the product of density functions (from likelihood and prior distributions) which can be regrouped in 5 types of functions, which we label $h1, \ldots, h_5$, defined as

$$h_i = \begin{cases} \frac{1}{\sigma^a} & \text{when } i = 1 \\ \exp -\frac{\beta}{\sigma^2} & \text{when } i = 2 \\ \frac{1}{\Phi^N\left(\frac{\mu+\theta}{\sigma}\right)} & \text{when } i = 3 \\ \exp -\frac{N}{2\sigma^2}(\mu-\theta)^2 & \text{when } i = 4 \\ \exp -\frac{1}{2\sigma^{*2}}(\log(\sigma)-\mu^*)^2 & \text{when } i = 5 \end{cases} . \tag{C.5}$$

We take their log

$$\log h_i = \begin{cases} -a\log\sigma & \text{when } i = 1 \\ -\frac{\beta}{\sigma^2} & \text{when } i = 2 \\ -N\log\Phi\left(\frac{\mu+\theta}{\sigma}\right) & \text{when } i = 3 \\ -\frac{N}{2\sigma^2}(\mu-\theta)^2 & \text{when } i = 4 \\ -\frac{1}{2\sigma^{*2}}(\log(\sigma)-\mu^*)^2 & \text{when } i = 5 \end{cases} \tag{C.6}$$

and compute the first two derivatives (relatively to μ and σ) for each of them to get

$$\frac{\partial}{\partial\sigma}(\log h_i) = \begin{cases} -\frac{a}{\sigma} & \text{when } i = 1 \\ 2\frac{\beta}{\sigma^3} & \text{when } i = 2 \\ N\varphi\frac{(\mu+\theta)}{\sigma^2} & \text{when } i = 3 \text{ (from (C.3))} \\ -\frac{1}{\sigma\sigma^{*2}}(\log(\sigma)-\mu^*) & \text{when } i = 5 \end{cases} \tag{C.7}$$

$$\frac{\partial^2}{\partial\sigma^2}(\log h_i) = \begin{cases} \frac{a}{\sigma^2} & \text{when } i = 1 \\ -6\frac{\beta}{\sigma^4} & \text{when } i = 2 \\ -N\left\{\varphi\left[-\frac{(\mu+\theta)}{\sigma}\frac{(\mu+\theta)^2}{\sigma^4}+\frac{2(\mu+\theta)}{\sigma^3}\right]-\varphi^2\frac{(\mu+\theta)^2}{\sigma^4}\right\} & \text{when } i = 3 \text{ from (C.4)} \\ = \frac{N\varphi\cdot(\mu+\theta)}{\sigma^3}\left\{\frac{(\mu+\theta)^2}{\sigma^2}+\varphi\cdot\frac{(\mu+\theta)}{\sigma}-2\right\} & \\ -\frac{1}{\sigma^{*2}}\frac{\left(\frac{1}{\sigma}\sigma-(\log(\sigma)-\mu^*)\right)}{\sigma^2} & \text{when } i = 5 \\ = \frac{\log(\sigma)-\mu^*-1}{\sigma^{*2}\sigma^2} & \end{cases} \tag{C.8}$$

$$\frac{\partial}{\partial\mu}(\log h_i) = \begin{cases} -N\frac{\phi}{\Phi\sigma} = -\frac{N\varphi}{\sigma} & \text{when } i = 3 \text{ (from (C.3))} \\ -\frac{N}{\sigma^2}(\mu-\theta) & \text{when } i = 4 \end{cases} \tag{C.9}$$

$$\frac{\partial^2}{\partial\mu^2}(\log h_i) = \begin{cases} -N\left\{\varphi\left[-\left(\frac{\mu+\theta}{\sigma}\right)\cdot\frac{1}{\sigma^2}+0\right]-\left[\frac{\phi}{\Phi\sigma}\right]^2\right\} & \text{when } i = 3 \text{ from (C.4)} \\ = \frac{N\varphi}{\sigma^2}\left[\frac{\mu+\theta}{\sigma}+\varphi\right] & \\ -\frac{N}{\sigma^2} & \text{when } i = 4 \end{cases} \tag{C.10}$$

C.4 Approximation of $\log(\Phi(g(z)))$ by Taylor series development

Let

$$f(z) = \log(\Phi(g(z))) \; . \tag{C.11}$$

A good approximation of $f(z)$ is obtained from the first three terms of its Taylor series development, that is

$$f(z) \approx f(0) + f'(0)z + \frac{f''(0)}{2}z^2 \; .$$

From (C.3), with $g(z) = z$, we easily obtain the first two derivatives of $f(z)$, respectively given by

$$f'(z) = \frac{\phi(z)}{\Phi(z)} \tag{C.12}$$

and

$$f''(z) = \frac{\phi'(z)\Phi(z) - \phi(z)\phi(z)}{\Phi^2(z)} \; ; \tag{C.13}$$

From (C.1), we obtain that $\phi'(0) = 0$; since $\phi(0) = \frac{1}{\sqrt{2\pi}}$ and $\Phi(0) = 1/2$ we easily get that

$$\begin{aligned} f'(0) &= \frac{1/\sqrt{2\pi}}{1/2} = \sqrt{\frac{2}{\pi}} \\ \text{and } f''(0) &= -\frac{\phi^2(0)}{\Phi^2(0)} = -[f'(0)]^2 = -\frac{2}{\pi} \; . \end{aligned}$$

Since $f(0) = \log(\Phi(0)) = \log(1/2) = -\log(2)$, we finally get the following second-degree approximation

$$f(z) \approx -\log(2) + \sqrt{\frac{2}{\pi}}z - \frac{1}{\pi}z^2 \; . \tag{C.14}$$

In some occasions, we may want to use a third-degree Taylor series development approximation. From (C.13), we have

$$f''(z) = \frac{\phi'(z)}{\Phi(z)} - \left[\frac{\phi(z)}{\Phi(z)}\right]^2$$

from which we easily get f's third derivative

$$f'''(z) = \frac{\phi''(z)\Phi(z) - \phi'(z)\phi(z)}{\Phi^2(z)} - 2\frac{\phi(z)}{\Phi(z)}\left\{\frac{\phi'(z)\Phi(z) - \phi^2(z)}{\Phi^2(z)}\right\} \; ;$$

since

$$\begin{aligned} \phi'(z) &= -z\phi(z) \text{from (C.1)} && \text{(C.15)} \\ \text{we obtain } \phi''(z) &= -\phi(z) - z\phi'(z) \; . && \text{(C.16)} \end{aligned}$$

The Taylor series z^3-coefficient can now be easily calculated, as

$$\begin{aligned} f(0) &= \log(1/2) = -\log(2) \\ \phi(0) &= \frac{1}{\sqrt{2\pi}} \\ \Phi(0) &= \frac{1}{2} \\ \phi'(0) &= 0 \text{ (from (C.15))} \\ \text{and } \phi''(0) &= -\phi(0) = -\frac{1}{\sqrt{2\pi}} \\ \text{and finally } f'''(0) &= \frac{\phi''(0)}{\Phi(0)} + 2\left[\frac{\phi(0)}{\Phi(0)}\right]^3 = -\sqrt{\frac{2}{\pi}} + 2\sqrt{\frac{8}{\pi^3}} = \left(\frac{4}{\pi} - 1\right)\sqrt{\frac{2}{\pi}} \end{aligned}$$

from which we get the following third-degree approximation

$$\log(\Phi(z)) \approx -\log(2) + \sqrt{\frac{2}{\pi}}z - \frac{1}{\pi}z^2 + \frac{1}{6}\left(\frac{4}{\pi} - 1\right)\sqrt{\frac{2}{\pi}}z^3 \quad . \tag{C.17}$$

The second-degree approximation of $\log(\Phi(z))$ is good for $z \leq 1.5$ while its third-degree approximation is good for $z \leq 5.5$; figure below shows the very good fit for both versions in their respective domains.

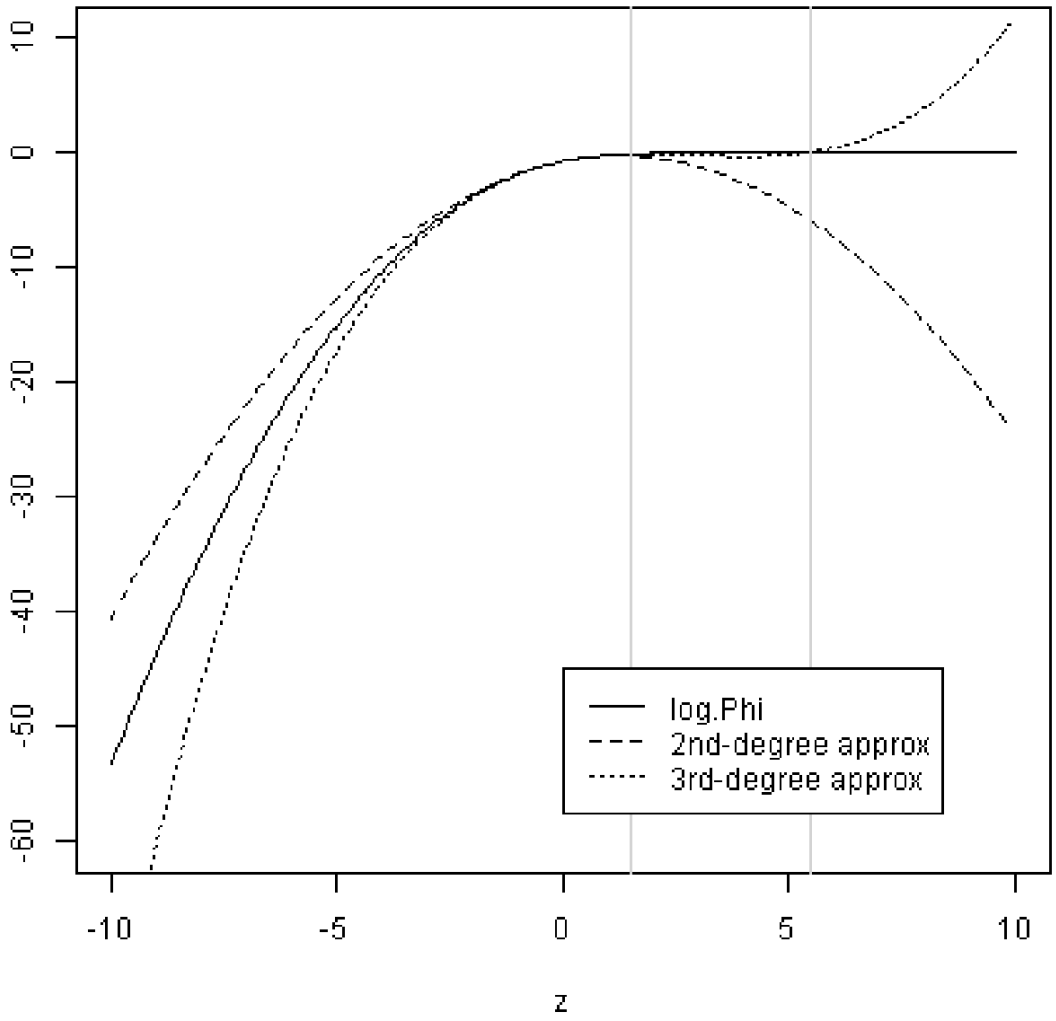


Figure 6: Second- and third-degree Taylor series approximation of $\log(\Phi(z))$.

C.5 Incorrect use of past data

The algorithm `SEG.informedvar` (section 3) including past data in the analysis was based on an improper prior (flat on $(0, \infty)$; while being technically correct, it was of limited interest, as no other algorithm was based on a uniform prior on σ. The current appendix documents that deprecated algorithm for historical reasons.

The likelihood of past data $\mathbf{p}$ (see equation 8)

$$l(\mathbf{p}|\mu,\sigma) = \underbrace{\frac{1}{\sigma^n}\exp\left\{-\frac{(n-1)s_p^2}{2\sigma^2}\right\}}_{g(\sigma^2)}\underbrace{\exp\left\{-\frac{n}{2\sigma^2}(\bar{p}-\mu)^2\right\}}_{l(\bar{p}|\mu,\sigma)}$$

is the product of the likelihood of $\bar{p}$ and a function $g(\sigma^2)$, which which we will approximate by a log-normal distribution with parameters (μ_p^*, σ_p^*) to be able to use the results of section 3. We approximate $g(\sigma^2)$ by a log-normal distribution for σ with parameters such that the means and 95% prior interval ranges match.

If we let $t = \sigma^2$, then the distribution function for t is easily obtained as

$$\begin{aligned} f(t) &\propto t^{-1/2}\frac{1}{(t^{1/2})^n}\exp\left\{-\frac{(n-1)s_p^2}{2t}\right\} \\ &= \frac{1}{t^{\frac{n+1}{2}}}\exp\left\{-\frac{(n-1)s_p^2}{2t}\right\} \end{aligned}$$

and hence $t = \sigma^2 \sim \text{InverseGamma}(\alpha = \frac{n-1}{2}, \beta = \frac{(n-1)s_p^2}{2})$.

If a random variable D has an Inverse-gamma distribution with parameters (α, β), then its mean is $E(D) = \log(\beta) - \psi(\alpha)$ where $\psi(\cdot)$ is the digamma function (see the inverse gamma distribution Wikipedia page).

Since σ^2 has an inverse gamma distribution, as shown above, we get

$$E(\log(\sigma^2)) = \log(\beta) - \psi(\alpha) \ .$$

The mean of its log-normal approximation μ_p^* is thus set to

$$\mu_p^* = \frac{\log(\beta) - \psi(\alpha)}{2} \ .$$

Let denote $\gamma_{\alpha,\beta;0.025}$ and $\gamma_{\alpha,\beta;0.975}$ the lower and upper 2.5 percentiles of a gamma distribution with parameters (α, β): then the lower and upper corresponding percentiles for $\log(\sigma^2)$ are given by $\log(\gamma^{-1}_{\alpha,\beta;0.975})$ and $\log(\gamma^{-1}_{\alpha,\beta;0.025})$ respectively, and hence the parameter σ_p^* is defined as

$$\sigma_p^* = \frac{1}{4\Phi^{-1}(0.975)}(\log(\gamma^{-1}_{\alpha,\beta;0.025}) - \log(\gamma^{-1}_{\alpha,\beta;0.975})) \ .$$

ANNEXE C : RÉSULTATS PROPRES AU MODÈLE NORMAL

A. Liste des indices d'exposition calculés pour la distribution normale dans le projet WebExpo

Tableau C1 : Indices d'exposition calculés pour la distribution normale dans le projet WebExpo

Analyse de GES
Estimation des paramètres distributionnels (estimation ponctuelle et intervalles de crédibilité)
Moyenne arithmétique
Écart-type arithmétique
Fraction de dépassement de la VLEP
Centile de la distribution des expositions (c.-à-d. centile critique, par défaut le 95^{e})
Décision relative à l'acceptabilité de l'exposition (risque de surexposition)
Probabilité que la fraction de dépassement soit supérieure ou égale au seuil de dépassement (par défaut 5 %)
Probabilité que le centile critique (par défaut le 95^{e}) soit supérieur ou égal à la VLEP
Probabilité que la moyenne arithmétique soit supérieure ou égale à la VLEP
Différences inter-travailleur
Estimation des paramètres distributionnels (estimation ponctuelle et intervalles de crédibilité)
Moyenne arithmétique de groupe
Écart-type arithmétique intra-travailleur
Écart-type arithmétique inter-travailleur
Coefficient de corrélation intra-travailleur (rho)
Probabilité que rho soit supérieur au seuil (Prob.rho.overX)
Différence R (R.diff)
Paramètres permettant de quantifier la possibilité que certains travailleurs soient surexposés (probabilité de surexposition individuelle)
Proportion des travailleurs individuels dont le centile critique est supérieur à la VLEP (Prob.ind.overexpo.perc)
Proportion des travailleurs individuels dont la moyenne arithmétique est supérieure à la VLEP (Prob.ind.overexpo.am)
Probabilité que la valeur vraie de Prob.ind.overexpo.perc soit supérieure à un seuil donné (Prob.ind.overexpo.perc.overX, par défaut 20 %)
Probabilité que la valeur vraie de Prob.ind.overexpo.am soit supérieure à un seuil donné (Prob.ind.overexpo.am.overX, par défaut 20 %)
Pour tout travailleur individuel : tous les indices de l'analyse de GES
Paramètres personnalisables
Probabilité des intervalles de crédibilité (par défaut 90 %)
Seuil de dépassement (5 %)
Centile critique (par défaut le 95^{e})
Seuil du coefficient de corrélation intra-travailleur (par défaut 0,2)
Couverture de la population pour la différence R (par défaut 80 %)
Seuil de probabilité de surexposition individuelle (par défaut 20 %)

B. Interprétation des extrants des modèles bayésiens – analyse de GES

La figure C1 illustre le flux de traitement des données dans l'analyse de la distribution normale. Dans le cas du modèle normal, les observations ne sont pas soumises à un prétraitement avant l'application des routines bayésiennes. Par conséquent, les utilisateurs doivent spécifier les a priori et l'erreur de mesure compte tenu de l'échelle de leur quantité d'intérêt. Les valeurs par défaut de WebExpo ont été déterminées en fonction de niveaux d'exposition au bruit exprimés en décibels (annexe D).

Les indices de sortie comprennent la moyenne arithmétique, l'écart-type arithmétique, la fraction de dépassement de la VLEP et tout centile de la distribution (par défaut le 95^{e}), obtenus à partir des équations ci-dessous.

Moyenne arithmétique de la distribution des expositions :

$$\boldsymbol{MA = \mu} \quad \textbf{(1)}$$

Écart-type arithmétique de la distribution des expositions :

$$\text{É}\boldsymbol{TA = \sigma} \quad \textbf{(2)}$$

X^{e} centile de la distribution des expositions :

$$\boldsymbol{PX = \mu + \Phi^{-1}\,(X) * \sigma} \quad \textbf{(3)}$$

où Φ^{-1} est la fonction de distribution cumulative inverse de la distribution normale standard.

Fraction de dépassement de la VLEP :

$$\boldsymbol{F(\%) = 100 * \left\{1 - \Phi\left(\frac{\text{VLEP} - \mu}{\sigma}\right)\right\}} \quad \textbf{(4)}$$

où Φ est la fonction de distribution cumulative de la distribution normale standard.

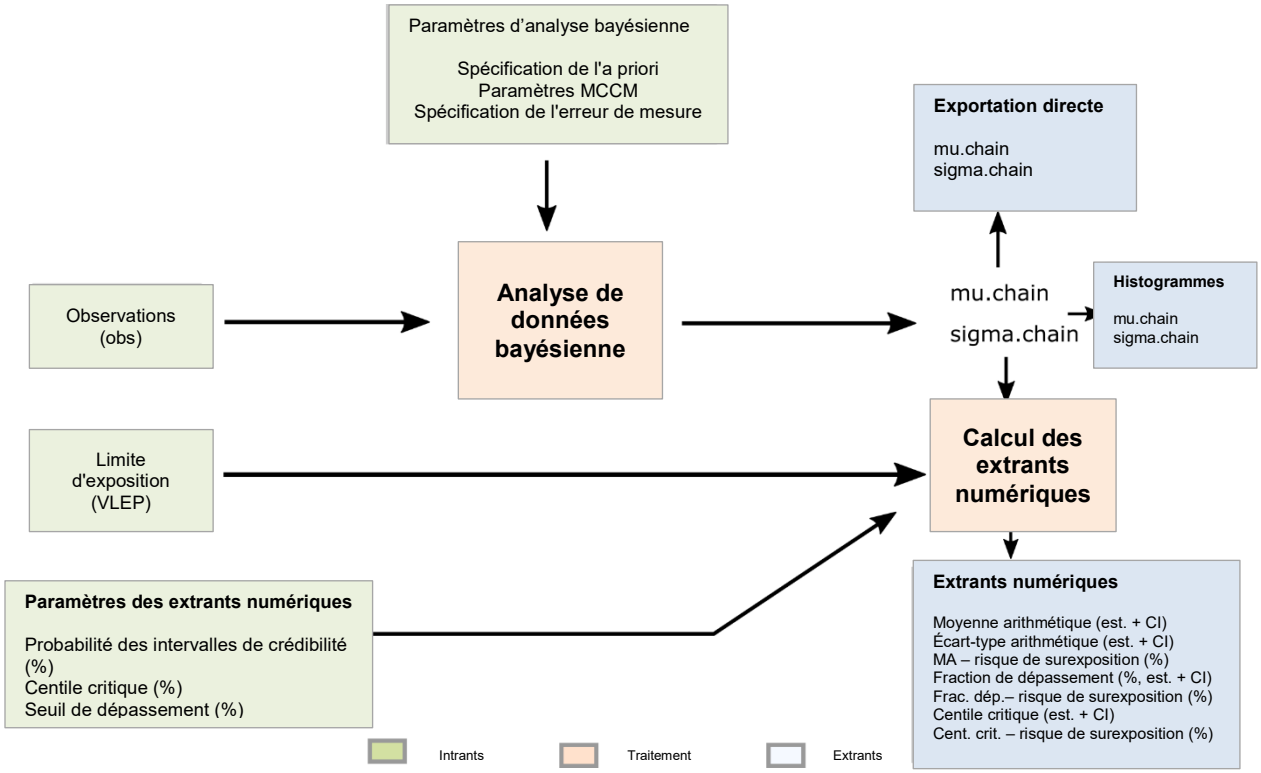


Figure C1. Flux de traitement des données pour les analyses de GES – Distribution normale.

Exemple

Nous présentons ci-dessous un exemple d'analyse d'un échantillon de taille 9 provenant d'une distribution normale où MA vraie = 80 et ÉTA vrai = 5 (compte tenu d'une VLEP de 85) [échantillon C1 de l'annexe E]. Le tableau C2 présente les résultats de cette analyse selon le modèle bayésien pour la distribution normale et une distribution a priori non informative en utilisant les paramètres par défaut (voir l'annexe D). L'étroitesse marquée des intervalles de crédibilité présentés dans le tableau C2 illustre la variabilité beaucoup plus faible de la distribution normale utilisée dans cet exemple, ce qui correspondrait à une distribution plausible d'expositions au bruit, comparativement à la distribution lognormale utilisée pour modéliser la variabilité environnementale des expositions à des substances chimiques.

Tableau C2. Estimations ponctuelles et intervalles de crédibilité des indices d'exposition dans un exemple de calcul bayésien selon le modèle normal

Paramètre	**Estimation ponctuelle et intervalle de crédibilité à 90 %**
Moyenne arithmétique	78,6 [76,9 - 80,4]
Écart-type arithmétique	3,01 [2,03 - 5,13]
Fraction de dépassement (%)	1,72 [0,0517 - 13,7]
95^{e} centile	83,6 [81,4 - 87,6]

C. Interprétation des extrants des modèles bayésiens – analyse des différences inter-travailleur

La figure C2 illustre le flux de traitement des données dans l'analyse de la distribution normale. Dans le cas du modèle normal (comme pour le modèle normal d'analyse de GES), les observations ne sont pas soumises à un prétraitement avant l'application des routines bayésiennes. Par conséquent, les utilisateurs doivent spécifier les a priori et l'erreur de mesure compte tenu de l'échelle de leur quantité d'intérêt. Les valeurs par défaut de WebExpo ont été déterminées en fonction de niveaux d'exposition au bruit exprimés en décibels (voir l'annexe D).

Les équations suivantes relatives aux indices d'exposition normaux ont été obtenues en adaptant les équations lognormales ci-dessus.

Moyenne arithmétique de groupe :

$$\boldsymbol{MA_{groupe} = \mu_Y} \qquad \textbf{(5)}$$

Écart-type arithmétique inter-travailleur :

$$\text{É}\boldsymbol{TA_b = \sigma_b} \qquad \textbf{(6)}$$

Écart-type arithmétique intra-travailleur :

$$\text{É}\boldsymbol{TA_w = \sigma_w} \qquad \textbf{(7)}$$

Coefficient de corrélation intra-travailleur :

$$\boldsymbol{\rho = \frac{\sigma_b^2}{\sigma_b^2+\sigma_w^2}} \qquad \textbf{(8)}$$

Différence $R_{X\%}$: Distance relative, exprimée en % de la moyenne arithmétique de groupe, intégrant le X % central de la distribution des moyennes arithmétiques propres aux travailleurs individuels, ou d'un quelconque centile. Cette proposition constitue une tentative d'expression de l'hétérogénéité d'une façon comparable au rapport R dans le cas du modèle lognormal, adaptée à l'échelle normale. Nous proposons une valeur par défaut de X de 80 %.

$$\boldsymbol{Rdiff_{X\%} = \frac{100 * \left(2 * \Phi^{-1}\left(\frac{1+X}{2}\right) * \sigma_b\right)}{MA_{groupe}}} \qquad \textbf{(9)}$$

Probabilité qu'un seul travailleur au hasard affiche une moyenne arithmétique supérieure à la VLEP :

$$\boldsymbol{P_{ind}^{MA}(\%) = 100 * \left\{1 - \Phi\left(\frac{\text{VLEP} - \mu_Y}{\sigma_b}\right)\right\}} \quad \textbf{(10)}$$

Probabilité qu'un seul travailleur au hasard affiche un X^e centile supérieur à la VLEP (ce qui équivaut à la probabilité qu'un seul travailleur au hasard présente un degré de dépassement de la VLEP supérieur à (100-X) % :

$$\boldsymbol{P_{ind}^{PX}(\%) = 100 * \left\{1 - \Phi\left(\frac{\text{VLEP} - \left(\mu_Y + \Phi^{-1}(X) * \sigma_w\right)}{\sigma_b}\right)\right\}} \qquad \textbf{(11)}$$

En plus de ce qui précède, il est possible d'obtenir des indices spécifiques à toute distribution d'expositions individuelles. En conséquence, par définition, la distribution des expositions pour un travailleur *i* est définie par :

Moyenne arithmétique de la distribution des expositions :

$$\boldsymbol{MA = \mu_Y + b_i} \quad \textbf{(12)}$$

Écart-type arithmétique de la distribution des expositions :

$$\text{É}\boldsymbol{TA = \sigma_w} \qquad \textbf{(13)}$$

X^e centile de la distribution des expositions :

$$\boldsymbol{PX = \mu_Y + b_i + \Phi^{-1}(X) * \sigma_w} \quad \textbf{(14)}$$

où Φ^{-1} est la fonction de distribution cumulative inverse de la distribution normale standard.

Fraction de dépassement de la VLEP :

$$\boldsymbol{F(\%) = 100 * \left\{1 - \Phi\left(\frac{VLEP - \mu_Y - b_i}{\sigma_w}\right)\right\}} \qquad \textbf{(15)}$$

où Φ est la fonction de distribution cumulative de la distribution normale standard.

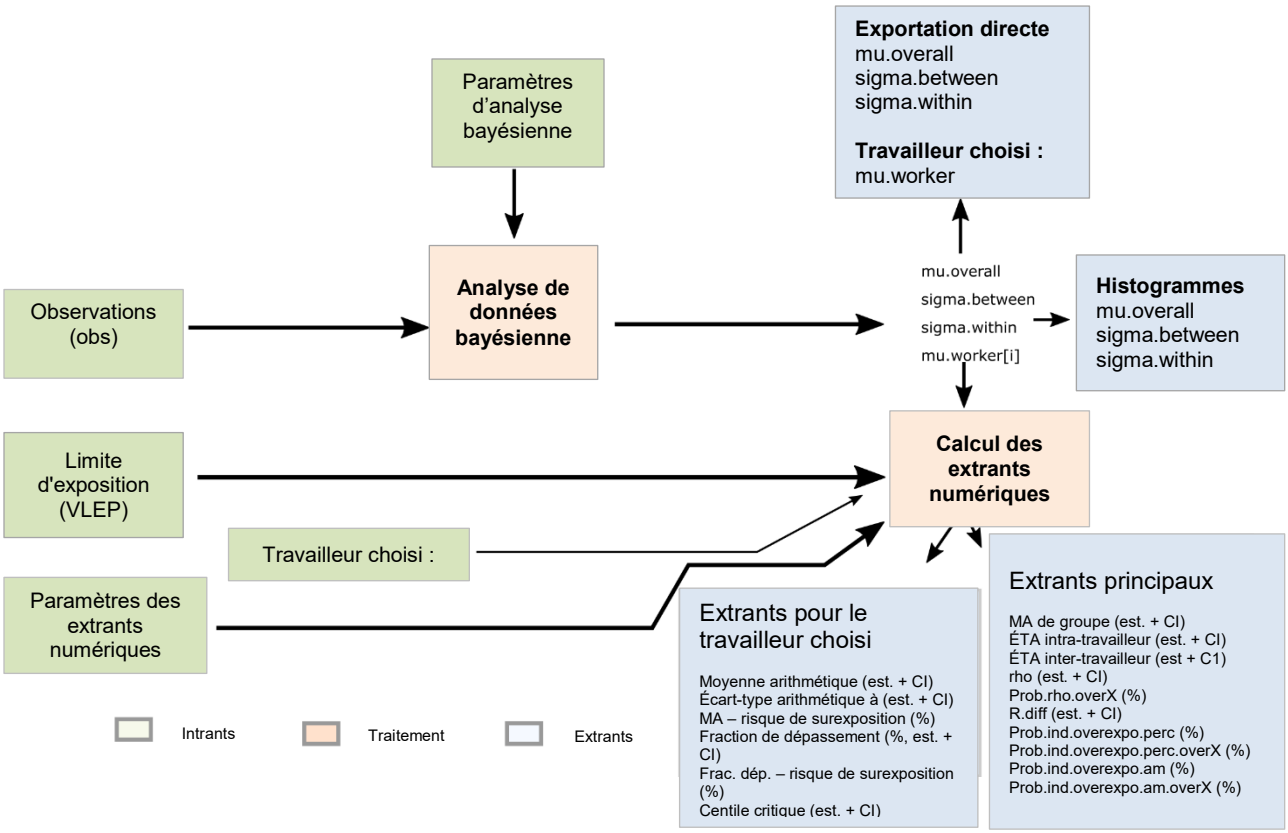


Figure C2. Flux de traitement des données pour les analyses de différences inter-travailleur – Distribution normale.

Exemple

Nous présentons ci-dessous un exemple d'analyse d'un échantillon de taille 100 (10 répétitions avec 10 travailleurs) provenant d'une distribution normale où MA vraie = 80 et ÉTA vrai = 5 [échantillon C2 de l'annexe E]. En l'absence de données empiriques sur des valeurs réalistes pour la corrélation intra-travailleur, nous avons utilisé la valeur médiane trouvée pour les substances chimiques dans la base de données de Kromhout *et al.* : rho = 0,22. Le tableau 11 présente les résultats de cette analyse. Comme pour l'analyse de GES en 4.2.1.6, les intervalles de crédibilité présentés dans le tableau C3 sont plus étroits que dans le cas lognormal. Pour cet exemple, nous avons présumé l'absence d'erreur de mesure et effectué les calculs selon le modèle [between-worker differences.informedvar] implémenté dans R + RJAGS (voir 4.3).

Tableau C3. Estimations ponctuelles et intervalles de crédibilité des indices d'exposition dans un exemple de calcul bayésien selon le modèle normal (analyse des différences inter-travailleur)

Paramètre	Faible corrélation intra-travailleur
Moyenne arithmétique (IDC à 90 %)	80,8 [78,8 - 82,7]
Écart-type arithmétique inter-travailleur (IDC à 90 %)	3,17 [2,02 - 5,16]
Écart-type arithmétique intra-travailleur (IDC à 90 %)	4,37 [3,88 - 4,96]
Corrélation intra-travailleur (rho) (IDC à 90 %)	0,345 [0,166 - 0,591]
Probabilité que rho soit supérieur à 0,2	90 %
R.difference (IDC à 90 %)	10 [6,4 - 16,4]
Probabilité de surexposition individuelle (95^{e} centile) en % (IDC à 90 %)	82,7 [59,1 - 96,8]
Chances que la probabilité ci-dessus soit supérieure à 20 %	100 %
Probabilité de surexposition individuelle (95^{e} centile) en % (IDC à 90 %)	8,81 [1,01 - 29]
Chances que la probabilité ci-dessus soit supérieure à 20 %	15 %

ANNEXE D : PARAMÈTRES D'ENTRÉE POUR LES MODÈLES BAYÉSIENS DANS WEBEXPO

Instructions générales en matière d'entrée de données pour tous les modèles

Bien qu'en théorie les modèles bayésiens puissent accepter comme intrant minimal jusqu'à zéro observation – auquel cas les distributions a posteriori vont simplement reproduire les distributions a priori –, nous proposons un intrant minimal « raisonnable » de 3 résultats non censurés, les résultats non censurés devant représenter au moins 30 % de l'échantillon total.

Dans le cas des modèles lognormaux, les données doivent être strictement positives et se trouver entre VLEP/1000 et 1000*VLEP. Bien qu'il s'agisse là de limites techniques informelles, ces bornes sont compatibles avec les paramètres par défaut proposés et présentés ci-dessous pour les fonctions bayésiennes.

Dans le cas des modèles normaux, les données doivent être positives et éloignées de zéro en raison du modèle de traitement de l'erreur de mesure. Sur la base des données de bruit exprimées en décibels, nous proposons que les valeurs soient comprises entre 40 et 140.

Les observations non censurées doivent être entrées sous forme numérique, les observations censurées à gauche sous la forme < X, les observations censurées à droite sous la forme > X, et les observations censurées par intervalle sous la forme [X1-X2].

A. SEG.informedvar (incluant le modèle past.data)

Paramètre	Recommandation par défaut	Fourchette raisonnable
n.chain[(A)]	1	1-5
n.iter	25 000	10000-100000
n.burnin	5000	100-5000
n.thin[(B)]	1	1
mu.lower	Lognormal : -20 Normal : 40	Lognormal : [-100; -0,5] et < min(observations) Normal (dB) : [20-85] et < min (observations)
mu.upper	Lognormal : 20 Normal (dB) : 125	Lognormal : [0,5; 100] et > max(observations) Normal (dB) : [85-140] et > max (observations)
log.sigma.mu	Lognormal : -0,1744 (MG = 0,84) Normal : 1,098612 (MG = 3)	Lognormal : La MG pour la distribution lognormale des valeurs de log(ÉTG) est comprise entre 0,405 et 1,609 ; correspond à log.sigma.mu entre -0,90 et 0,48 Normal : La MG pour la distribution lognormale de sigma est comprise entre 0,5 et 10 ; correspond à log.sigma.mu entre -0,69 et 2,30

Paramètre	Recommandation par défaut	Fourchette raisonnable
log.sigma.prec	Lognormal : 2,5523 Normal : 1,191059	Lognormal : L'ÉTG pour la distribution lognormale des valeurs de log(ÉTG) est comprise entre 1,5 et 5 ; correspond à log.sigma.prec entre 0,40 et 6,0 Normal : L'ÉTG pour la distribution lognormale de sigma est comprise entre 1,5 et 5 ; correspond à log.sigma.prec entre 0,40 et 6,0
init.mu	Lognormal : log(0,3) Normal : 85	Lognormal : Entre VLEP/1000 et 100*VLEP (init.mu entre -6,908 et 4,605) Normal : Entre 50 et 120dB (init.mu entre 50 et 120)
init.sigma	Lognormal : log(2,5) Normal : 3	Lognormal : ÉTG entre 1,5 et 10 (init.sigma entre 0,405 et 2,303) Normal : Sigma entre 0,5 et 10 (init.sigma entre 0,5 et 10)
past.data (moyenne)	s. o.	Lognormal : Entre VLEP/1000 et 100*VLEP (moyenne entre -6,908 et 4,605) Normal : Entre 50 et 120dB (moyenne entre 50 et 120)
past.data (é.-t.)	s. o.	Lognormal : ÉTG entre 1,5 et 10 (é.-t. entre 0,405 et 2,303) Normal : Sigma entre 0,5 et 10 (é.-t. entre 0,5 et 10)
past.data (n)	s. o.	1-1000
me.sd.range	s. o.	Lognormal : 0,001-1 (ce qui signifie entre VLEP/1000 et VLEP) Normal : 0,1-10
cv.range	s. o.	0,01-2 (ce qui signifie entre 1 et 200 %)

(A) Bien que le nombre de chaînes MCCM soit un paramètre des fonctions en code R créées par l'équipe de McGill, la relative simplicité des modèles ne justifie pas de chaînes multiples. Par conséquent, dans toutes les autres fonctions et dans tous les autres algorithmes, ce paramètre est fixé à 1.

(B) Parallèlement à la note (A), le facteur de dilution est fixé à 1 dans tous les algorithmes, sauf pour les fonctions en R.

B. SEG.uninformative (limité aux paramètres non décrits en A)

Paramètre	Recommandation par défaut	Fourchette raisonnable
sd.range	Lognormal : ÉTG entre 1,1 et 10 sd.range = [0,095-2,3] Normal : ÉTA entre 0,1 et 20 dB sd.range = [0,1-20]	Lognormal : ÉTG entre 1,01 et 20 sd.range inclus dans [0,01-3] Normal : ÉTA entre 0,1 et 100 dB sd.range inclus dans [0,1-100]
Init.sd	Lognormal : log(2,5) Normal : 3	Lognormal : ÉTG entre 1,5 et 10 (init.sd entre 0,405 et 2,303) Normal : Sigma entre 0,5 et 10 (init.sd entre 0,5 et 10)

C. SEG.riskband (limité aux paramètres non décrits en A)

Paramètre	Recommandation par défaut	Fourchette raisonnable
A **(points de rupture définissant les bandes)**	Lognormal (bandes de l'AIHA) 0,01/0,1/0,5/1 Normal : 70/80/85/90	Valeurs ordonnées de la plus petite à la plus grande entre 0.001 et 100.
target_perc **(centile de la distribution utilisé pour définir les probabilités a priori sur les bandes)**	95 (95^{e} centile)	1 - 99
region.prior.prob **(probabilités a priori pour chaque bande)**	0,2/0,2/0,2/0,2/0,2	Toutes valeurs comprises entre 0 et 1, à condition qu'elles totalisent 1
mu.lower.riskb	Voir mu.lower section A.	Voir mu.lower section A.
mu.upper.riskb	Voir mu.upper section A.	Voir mu.upper section A.
sigma.lower	Voir sd.range section B.	Voir sd.range section B.
sigma.upper	Voir sd.range section B.	Voir sd.range section B.

D. Between.worker differences

Les contraintes relatives aux observations sont semblables à celles des fonctions d'analyse de GES. Nous recommandons en outre l'entrée d'au moins 3 travailleurs avec au moins 2 mesures chacun. Tous les travailleurs n'ont pas à avoir des mesures répétées. Notez que l'incertitude entourant les estimations d'expositions propres aux travailleurs individuels sera directement liée au nombre de mesures prélevées pour chaque travailleur d'intérêt.

Paramètre	Recommandation par défaut	Fourchette raisonnable
n.iter	50 000	15000-200000
n.burnin	5000	500-10000
mu.overall.lower	Lognormal : -20 Normal : 40	Lognormal : [-100; -0,5] et < min(observations) Normal (dB) : [20-85] et < min(observations)
mu.overall.upper	Lognormal : 20 Normal (dB) : 125	Lognormal : [0,5; 100] et > max(observations) Normal (dB) : [85-140] et > max(observations)
log.sigma.between.mu	Lognormal : -0,8786 (MG = 0,415) Normal : 1,098612 (MG = 3)	Lognormal : La MG pour la distribution lognormale des valeurs de log(ÉTG) est comprise entre 0,105 et 1,609 ; correspond à log.sigma.between.mu entre -2,25 et 0,48 Normal : La MG pour la distribution lognormale de sigma est comprise entre 0,1 et 10 ; correspond à log.sigma.mu entre -2,30 et 2,30
log.sigma. between.prec	Lognormal : 1,634 Normal : 1,191059	Lognormal : L'ÉTG pour la distribution lognormale des valeurs de log(ÉTG) est comprise entre 1,5 et 5 ; correspond à log.sigma.between.prec entre 0,40 et 6,0 Normal : L'ÉTG pour la distribution lognormale de sigma est comprise entre 1,5 et 5 ; correspond à log.sigma.between.prec entre 0,40 et 6,0
log.sigma.within.mu	Lognormal : -0,4106 (MG = 0,415) Normal : 1,098612 (MG = 3)	Lognormal : La MG pour la distribution lognormale des valeurs de log(ÉTG) est comprise entre 0,405 et 1,609 ; correspond à log.sigma.mu entre -0,90 et 0,48 Normal : La MG pour la distribution lognormale de sigma est comprise entre 0,5 et 10 ; correspond à log.sigma.mu entre -0,69 et 2,30

Paramètre	Recommandation par défaut	Fourchette raisonnable
log.sigma. within.prec	Lognormal : 1,9002 Normal : 1,191059	Lognormal : L'ÉTG pour la distribution lognormale des valeurs de log(ÉTG) est comprise entre 1,5 et 5 ; correspond à log.sigma.within.prec entre 0,40 et 6,0 Normal : L'ÉTG pour la distribution lognormale de sigma est comprise entre 1,5 et 5 ; correspond à log.sigma.within.prec entre 0,40 et 6,0
init.mu.overall	Lognormal : log(0,3) Normal : 85	Lognormal : Entre VLEP/1000 et 100*VLEP (init.mu.overall entre -6,908 et 4,605) Normal : Entre 50 et 120dB (init.mu.overall entre 50 et 120)
init.sigma.between	Lognormal : log(2,5) = 0,916 Normal : 3	Lognormal : ÉTG entre 1,1 et 10 (init.sigma.between entre 0,095 et 2,303) Normal : Sigma entre 0,5 et 10 (init.sigma.between entre 0,5 et 10)
init.sigma.within	Lognormal : log(2,5) = 0,916 Normal : 3	Lognormal : ÉTG entre 1,1 et 10 (init.sigma.within entre 0,095 et 2,303) Normal : Sigma entre 0,5 et 10 (init.sigma.within entre 0,5 et 10)
sigma.between.range	Lognormal : ÉTG entre 1,00 et 10 Sigma.between.range = [0-2,3] Normal : ÉTA entre 0 et 20 dB Sigma.between.range = [0-20]	Lognormal : ÉTG entre 1,0 et 20 Sigma.between.range inclus dans [0,00-3] Normal : ÉTA entre 0 et 100 dB Sigma.between.range inclus dans [0-100]
Sigma.within.range	Lognormal : ÉTG entre 1,1 et 10 Sigma.within.range = [0,095-2,3] Normal : ÉTA entre 0,1 et 20 dB Sigma.within.range = [0,1-20]	Lognormal : ÉTG entre 1,01 et 20 Sigma.within.range inclus dans [0,01-3] Normal : ÉTA entre 0,1 et 100 dB Sigma.within.range inclus dans [0,1-100]
me.sd.range	s. o.	Normal : 0,1-10 Lognormal : 0,001-1 (ce qui signifie entre VLEP/1000 et VLEP)
cv.range	s. o.	0,01-2 (ce qui signifie entre 1 % et 200 %)

ANNEXE E:
ÉCHANTILLONS UTILISÉS POUR LES EXEMPLES NUMÉRIQUES

Échantillon 1 : Analyse de GES – exemple principal

24,7	64,1	13,8	43,7	19,9	133	32,1	15	53,7

Échantillon 2 : Analyse de GES – exemple d'erreur de mesure

96,6	38,3	80,8	15,1	34	73,4	14,5	64,8	27,4	48,7
43,3	43,4	57,8	94,9	44,1	44,3	62,9	117	51,6	64,7
50,1	74,7	221	46,8	84,3	93,4	126	46,9	29,5	73,8
66,9	61,3	30,2	101	22,6	191	29,3	68	114	33,7
52,5	118	49,7	60,4	36,6	55,9	31,9	84,3	75,8	39,5
28,3	56,5	44,2	48	36,6	70	37	72	48	66,1
72,4	80,9	69,1	162	67,3	75,2	40,5	25,6	44	120
56,3	42,9	6,63	24,9	40,9	81	97,2	74,7	79,6	48,8
75,3	54,8	66,5	71,3	28,7	87,5	51,9	19,6	60,8	45,9
46,9	84,8	120	103	36,7	92,7	32,8	73,8	214	65,3

Échantillon 3 : Différences inter-travailleur – faible corrélation intra-travailleur

Travailleur									
1	2	3	4	5	6	7	8	9	10
185	4,79	8,85	16,4	14,7	37,9	22	69,9	28,1	113
34,8	23	31,7	6,91	59,6	96,9	44,8	30,5	7,49	7,68
16,7	7,54	15,8	87,4	15	40,8	37,5	33,4	16	85,6
12,4	62,3	89,6	20	21,8	106	16,6	53	23	196
18,6	8,55	164	16,8	20,6	21,7	30,7	70,7	99,9	35
47,4	9,28	40,5	7,12	96,1	25,8	7,07	78,3	12	17,6
52,6	43,6	47,6	6,99	16,8	51,3	7,18	18	11,8	60,7
15,3	94,2	75,5	16,4	15,8	23	80,9	45,2	57,4	15,5
27,6	44,6	10,7	12,6	8,02	18,9	44,5	51,4	8,79	34,3
26,3	66,6	62,3	63,9	26,7	20,2	135	33,7	24	12,1

Échantillon 4 : Différences inter-travailleur – forte corrélation intra-travailleur

Travailleur									
1	2	3	4	5	6	7	8	9	10
66,8	14,2	186	23,5	43,8	41	6,56	9,21	19,6	78,7
46	53,9	84,6	16,2	31,1	11,4	9,5	9,42	14,3	28,2
61,1	21,8	94,4	40,2	13,1	4,44	6,97	28,7	22,8	41,3
54,6	47,8	218	130	24,1	12,9	5,92	72,9	35,1	14,4
31,7	48,8	189	42,2	27,7	22,7	2,42	35,6	28,9	72,9
74,3	76,5	130	25,7	23,9	20,5	14	17,2	36,9	10,2
60,9	41,3	107	35,4	40,2	12,6	12,3	20,2	13	16,2
53,4	20,4	80,6	40,8	60,3	8,35	3,07	13,4	13,3	15,8
38,9	31,9	288	109	29,8	13,6	7,01	10,5	13,6	42,2
27,5	31,1	173	40,9	37,2	28,1	6,49	26,3	37	61

Échantillon 5 : Différences inter-travailleur – taille d'échantillon réaliste

Travailleur-1	Travailleur-1	Travailleur-1	Travailleur-1	Travailleur-2	Travailleur-2
31	60,1	133	27,1	61,1	5,27
Travailleur-2	Travailleur-2	Travailleur-3	Travailleur-3	Travailleur-3	Travailleur-3
30,4	31,7	20,5	16,5	15,5	71,5

Échantillon 6 : Variabilité des résultats entre les plateformes de calcul

< 25,7	17,1	168	85,3	66,4	< 49,8	33,2	< 24,4	38,3

Exemple C1 : Analyse de GES – échantillon normal

81	79,5	80,7	78,1	80,1	74,8	74,8	79,8	79,8

Exemple C2 : Différences inter-travailleur – échantillon normal

Travailleur									
1	2	3	4	5	6	7	8	9	10
76,2	70,6	79,2	79,1	85,3	77,8	79,1	80	80	89,1
82,3	78,7	77,7	77,6	92,2	89	80,7	76,6	81,2	85,4
81,7	77,6	73,5	81,2	75,8	81,9	85,8	84,6	73,8	81,8
73,7	76,9	78,9	82,6	84,1	80,4	84,8	77,1	80,7	88,1
79,4	79,5	81,6	81,6	76,1	88,5	88,5	81,5	76,9	86,4
79,1	84,8	83,1	82,4	84,6	87	82,6	77,4	77,5	81,6
80,2	77,6	85,1	76,9	78,9	85	78,6	73,5	74,6	86,8
71	65,5	84,2	87,6	75,8	88,1	90,1	82,2	70,6	81,4
86,9	74,1	79,8	80,4	89	81,3	82,9	74,4	82,3	86,7
75,6	69,9	84,1	79,7	87,1	90,6	83	77,6	66,4	83,6