Titre: Title:	Estimation statistique de données manquantes en inventaire du cycle de vie
Auteur: Author:	Vincent Moreau
Date:	2012
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Moreau, V. (2012). Estimation statistique de données manquantes en inventaire du cycle de vie [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/869/

UNIVERSITÉ DE MONTRÉAL

ESTIMATION STATISTIQUE DE DONNÉES MANQUANTES EN INVENTAIRE DU CYCLE DE VIE

VINCENT MOREAU

DÉPARTEMENT DE GÉNIE CHIMIQUE

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION

DU DIPLÔME DE PHILOSOPHIAE DOCTOR

(GÉNIE CHIMIQUE)

MAI 2012

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée:

ESTIMATION STATISTIQUE DE DONNÉES MANQUANTES EN INVENTAIRE DU CYCLE DE VIE

présentée par: MOREAU Vincent

en vue de l'obtention du diplôme de : Philosophiae Doctor

a été dûment acceptée par le jury d'examen constitué de :

M. LEGROS Robert, Ph. D., président

M. SAMSON Réjean, Ph. D., membre et directeur de recherche

M. MARCOTTE Denis, Ph. D., membre et codirecteur de recherche

M. BAGE Gontran, Ph. D., membre et codirecteur de recherche

M. SAVADOGO Oumarou, D. d'état , membre

M. YOUNG Steven B., Ph. D., membre

DÉDICACE

A tous les objecteurs de croissance

REMERCIEMENTS

J'aimerais remercier en premier lieu mes trois directeurs de thèse, Réjean Samson, Denis Marcotte et Gontran Bage. Leurs réflexions préparaient déjà la terrain avant même que je choisisse ce projet de recherche, hors des sentiers battus. C'est la complémentarité de leurs contributions que j'ai le plus apprécié et qui m'a permis de développer un tel sujet. Je les remercie également de m'avoir questionné et soutenu pendant toute la durée de mon doctorat afin de présenter mes résultats de manière accessible. D'avoir travaillé sur un sujet aussi particulier, vos suggestions et commentaires m'ont continuellement encouragé dans l'approfondissement de mes recherches. Merci.

Un grand merci à Renée Michaud, Xavier Bengoa, Pablo Tirado et Steven Young pour la collecte de données. C'est grâce à leur collaboration que j'ai pu rassembler une partie des données essentielles à ce projet. Parmi les membres passés et présents du CIRAIG, je remercie particulièrement Pascal Lesage pour ses critiques précises et précieuses, et Ben Amor pour avoir partagé de nombreuses étapes du doctorat ensemble.

Mille mercis à toutes celles et ceux qui m'ont soutenu de près et de loin, en particulier mes amiEs Gabrielle Trepanier, Spencer Mann, Robert West, Cosmin Paduraru, Matthew Woods et surtout Stephanie Childs. Les critiques de Sybille van den Hove et Marc Le Menestrel sur les questions de fond ont également été très appréciées. Un immense merci à mes parents et mes frères pour leur présence malgré mon absence.

En dernier lieu, je remercie les partenaires industriels de la Chaire internationale en analyse du cycle de vie pour leur aide financière.

RÉSUMÉ

Un problème rencontré par de nombreuses méthodes d'évaluations d'impacts sur l'environnement est le manque de données fiables. L'analyse du cycle de vie (ACV), un outil d'évaluation des impacts pour produits, procédés et services sur l'ensemble de la chaîne de production, consommation et élimination, ne fait pas exception. Les raisons expliquant le manque de données sont multiples, la collecte peut être physiquement, économiquement ou légalement infaisable. Plus important encore sont l'absence de données et les incertitudes qui découlent de connaissances insuffisantes sur le fonctionnement ou l'état d'un système, d'une technologie émergente à l'évolution du climat. En ACV, l'inventaire du cycle de vie est une étape particulièrement gourmande en données puisqu'il s'agit de quantifier les flux d'énergie et de matériaux traversant les frontières d'un système, incluant les émissions. Différentes sources de données empiriques et génériques sont utilisées lorsque disponibles. Les bases de données d'inventaire et références bibliographiques, si elles ne sont pas nécessairement représentatives du contexte à l'étude, sont souvent ajustées en fonction de la situation géographique ou des technologies en vigueur. En parallèle, quelques approches existent pour estimer les données lorsque celles-ci ne sont pas disponibles, dont la contribution principale de cette thèse.

L'objectif de ce travail de recherche est d'améliorer la qualité des données d'inventaire en proposant une méthode d'estimation des données manquantes et des incertitudes correspondantes. Contrairement aux bilans de masse et d'énergie pour un procédé donné, l'approche développée ici consiste à modéliser les processus à l'aide d'estimateurs statistiques dont les propriétés sont adaptées aux échantillons de petites tailles et aux données très variables. L'hypothèse de recherche est la suivante: un estimateur nommé krigeage permet d'intégrer l'estimation de données manquantes et de leurs incertitudes de façon plus fiables qu'avec d'autres techniques linéaires. Emprunté à la géostatistique, il présente les avantages d'être flexible sur les modèles de relation entre variables dépendantes et indépendantes et d'être un estimateur exact, sans erreur statistique lorsque les données observées sont estimées, toute l'information est donc conservée. Une interprétation des paramètres du krigeage spécifique aux problèmes d'incertitude des

données, présente d'autres avantages. Qu'elles proviennent de sources diverses ou qu'il faille recourir à toutes les données disponibles, aussi peu nombreuses soient elles, leurs variations deviennent importantes. La résolution des équations du krigeage est ainsi modifiée pour permettre d'incorporer une incertitude spécifique à chaque observation. La procédure se base sur les caractéristiques techniques, variables indépendantes, des processus étudiés pour estimer leurs besoins en énergie et matériaux et leurs émissions sur l'ensemble du cycle de vie. En comparaison avec d'autres approches, peu de données additionnelles sont nécessaires.

La production d'électricité est un exemple phare du fait de son impact pour la plupart des procédés. L'hydroélectricité en particulier est mal représentée parmi les données existantes, sa production variant considérablement d'un site à l'autre. En d'autres termes il n'existe pas de centrale hydroélectrique générique. Contrairement aux flux d'inventaire de ces mêmes centrales, les spécifications techniques ou variables caractéristiques, telles que la capacité installée, la production effective, la surface d'un éventuel réservoir, etc. sont des données généralement accessibles. Les modèles de krigeage sont d'abord testés sur des données d'inventaire pour des éoliennes de puissances variées pour ensuite être appliqués d'un coté aux flux d'énergie et de matériaux nécessaires à la construction et l'exploitation de centrales et de l'autre aux émissions de gaz à effet de serre provenant des réservoirs hydroélectriques.

Les résultats montrent qu'il est possible d'améliorer l'estimation des données d'inventaire grâce au krigeage. En comparant plusieurs formes de krigeage et régression linéaire, les estimations par krigeage sont non seulement plus précises mais les écarts type couvrent aussi les données de manière plus exacte. Lorsque les données observées sont incomplètes, c'est à dire que les flux d'inventaire ne sont pas disponibles pour toutes les observations, les erreurs d'estimation sont plus faibles pour le krigeage que la régression linéaire. Plus précisément, le krigeage d'un flux d'inventaire sur base de deux variables caractéristiques présente des erreurs encore plus faibles que son cousin multivarié, le cokrigeage. Par validation croisée, les erreurs d'estimation sont en moyenne plus faible pour le krigeage que la régression linéaire, que les données observées soient complètes ou non. L'application de plusieurs variables caractéristiques améliore la qualité des estimations lorsqu'elles sont positivement corrélées. De plus, la modification du

système de krigeage pour intégrer une mesure d'incertitude propre à chaque observation, permet d'atteindre une réduction de la variation des données estimées. Autrement dit cette variabilité est incluse directement dans le modèle. Les estimations sont ainsi plus fiables lorsque proches des points d'échantillon dont les incertitudes sont faibles, et inversement. Pour chacun des échantillons, différents modèles reliant variables dépendantes et indépendantes sont testés, notamment les fonctions de covariance linéaire, exponentielle, sphérique et cubique, ainsi qu'une série de valeurs pour leurs paramètres.

Concernant l'analyse de systèmes de production d'électricité, ces résultats impliquent en particulier une estimation des données là où elles seraient difficiles à collecter et donc une simplification du processus de récolte. Pour des technologies aussi spécifiques aux conditions géologiques ou hydrologiques que l'hydroélectrique, l'estimation de flux d'inventaire par krigeage avec intégration de cette variabilité, se révèle plus représentative du contexte géographique ou technologique, d'où des données d'inventaire de meilleure qualité. Même si le krigeage présente de nombreux avantages – ses erreurs d'estimation sont en moyenne plus faibles que la régression linéaire – des limites existent quant à son application.

Les estimations se basent sur des variables indépendantes qui expliquent, avec des corrélations plus ou moins élevées, l'inventaire d'un processus. Dans la chaîne de production, il est donc possible de remonter jusqu'aux processus élémentaires dont les caractéristiques décrivent les différents flux d'énergie et de matériaux correspondants. En comparaison avec d'autres façon d'estimer ou de substituer des données manquantes, celle proposée ici ne repose sur aucun a priori ou jugement d'expert mais requiert certaines connaissances techniques pour identifier ces variables caractéristiques et valider le modèle de krigeage. Cette thèse apporte un ensemble d'évidences en faveur de l'hypothèse d'une amélioration de la modélisation des données d'inventaire du cycle de vie à l'aide du krigeage. L'approche se révèle particulièrement adaptée à l'estimation de données manquantes pour augmenter la fiabilité des inventaires et aux processus spécifiques dont les données existantes ne sont pas représentatives.

ABSTRACT

A problem frequently encountered in environmental impact assessment methods is the lack of reliable data. Life cycle assessment (LCA), a decision support tool which evaluates the impacts of products, processes and services from production to consumption and disposal, is no exception. Multiple reasons explain the lack of data, the sheer size of a product system or process can limit data collection, so do the economic costs as well as legal restrictions. More important is the absence of data and corresponding uncertainties due to a lack of knowledge regarding the state or performance of a system, from new and emerging technologies to climatic change. In LCA, life cycle inventory is a particularly data intensive and sensitive phase of the analysis, requiring the quantification of energy and material flows crossing a system's boundaries as well as emissions. When available, various sources of data, empirical and generic are compiled in an inventory. Although databases and literature references are not necessarily representative of the context, they are often adjusted according to geographic location and prevailing technologies. A few techniques to estimate missing inventory data also exist, to which this thesis is an important contribution.

Hence the main objective of the research work, to improve the quality of life cycle inventory data by developing a method to estimate missing data and corresponding uncertainties. Contrary to process-based models of mass and energy balance, this approach consists of statistical estimators which model processes from relatively small samples of usually high variability. The research hypothesis is as follows: the so called kriging estimator allows the combined estimation of missing data and their uncertainties in such ways that are more reliable than other linear estimators. Borrowed from spatial statistics, kriging is an estimator with several advantages, the flexibility associated with a choice of model function and the exact estimator property. In other words, kriging shows no statistical errors when estimating observed values, no data is averaged out. An interpretation of the kriging parameters specific to the problems of data uncertainty, offers more advantages. One parameter of the covariance function accounts for small scale variations of the data and taken as a proxy for uncertainty. Whether it be the variety of data

sources, the scarcity of the data itself or both, each and every source adds to data variability and uncertainty. The kriging system of equations is therefore modified such as to integrate a factor of uncertainty specific to each observations. Comparisons between the modified and conventional forms of kriging can be drawn. The procedure is based on the relationship between technical specifications, more readily available independent variables, and the dependent material and energy flows of the processes under consideration. Such material and energy requirements as well as emissions are estimated over the entire life cycle of products and processes. The needs for additional data are relatively low compared to other approaches, namely extended input output analysis.

For many products, processes and services, electricity generation and consumption account for a sizable share of the impacts. Hydroelectricity in particular is poorly represented within existing inventory data since production facilities vary considerably from one location to another. In other words, generic hydropower plants do not exist. Contrary to inventory flows, the technical specifications or characteristic variables of hydropower plants, such as the installed capacity, annual production or surface area of adjacent reservoirs, are usually publicly available. The kriging model is first tested on a data set which represents windmills of varying power capacity before it is applied to hydroelectricity. The experiment is divided according to data availability, on one hand the energy and materials required during construction, operation and maintenance of hydropower plants. On the other hand, the emissions of greenhouse gases from reservoirs are estimated.

The results show that estimation of inventory data can be improved thanks to kriging. When comparing different forms of kriging and linear regression, the kriging estimates are not only more precise but the standard deviations also cover the data more accurately. Where the observed data are incomplete, that is where inventory flows are missing for part of the observations, the estimation errors are lower for kriging than linear regression. Moreover, univariate kriging of inventory flows based on two characteristic variables, shows lower errors than its multivariate kin, cokriging. On average the statistical errors calculated from cross-validation are lower for kriging than they are for linear regression, whether the observed data are

complete or not. The application of several characteristic variables improves the quality of the estimates when they are positively correlated. In addition, the modified form of kriging which accounts for degrees of uncertainty specific to each observations, results in a reduction in the variations of the estimated inventory data. That is, data variability is incorporated directly in the model. Estimates closer to more reliable observations are shown to be less uncertain and vice versa. For each of the data sets, different relationships between dependent and independent variables are tested, for example the linear, exponential, spherical and cubic covariance functions as well as a range of parameter values.

For the analysis of electrical generation technologies, these results imply better estimates for data that are difficult to sample and therefore a simplified data collection process. In the case of site specific or variable processes such as hydroelectricity, the estimation of inventory data with kriging accounting for such data variability, proves more representative of the geographical or technological context. The quality of inventory data is consequently higher. Even if kriging has several advantages and its estimation errors are lower on average, some limitations to its application exist.

The estimation procedure is based on independent variables explaining, with different degrees of correlation, the inventory flows of a process. In the production and consumption chain, the inventory of elementary processes can be estimated as far as characteristic variables describe the corresponding material and energy flows or emissions. Compared to other techniques for estimating or substituting missing data, the proposed approach makes no particular assumption nor requires expert judgment but the technical knowledge to identify characteristic variables and validate the kriging model. This thesis brings evidence in favor of the hypothesis that the estimation of life cycle inventory data can be improved thanks to statistical estimators such as kriging. The methodological approach proves particularly suited to increasing both, the reliability of inventories by estimating missing data and the estimation of more representative inventory data for specific processes without existing quality data.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

ACV	Analyse du cycle de vie
EICV	Évaluation d'impact du cycle de vie
GES	Gaz à effet de serre
GHG	*Greenhouse gas*
ICV	Inventaire du cycle de vie
LCA	*Life cycle assessment*
LCI	*Life cycle inventory*
LCIA	*Life cycle impact assessment*
LOOCV	*Leave one out cross validation*
LOoOCV	*Leave one observation out cross validation*
LOvOCV	*Leave one variable out cross validation*
MAE	*Mean absolute error*
MAR	*Missing at random*
MCAR	*Missing completely at random*
MIET	*Missing inventory estimation tool*
MLE	*Maximum likelihood estimator*
MVUE	*Minimum variance unbiased estimator*
MWe	Mega Watt électrique
NMAR	*Non missing at random*
SETAC	*Society of environmental toxicology and chemistry*

LISTE DES ANNEXES

INTRODUCTION

L'analyse du cycle de vie est un outil d'évaluation quantitative de l'impact environnemental de produits et processus. Une approche globale y est privilégiée, à une échelle non seulement temporelle, le cycle de vie d'un produit ou service, mais aussi spatiale, les impacts pouvant découler d'un changement technologique ou d'une nouvelle réglementation externe au système à l'étude. Afin d'informer les processus décisionnels, d'ordres politique, économique ou écologique, la fiabilité d'une telle méthode est essentielle pour éviter une décision erronée. Lorsqu'elles sont quantifiables, les incertitudes peuvent infléchir ce processus d'un côté comme de l'autre et ne sont par conséquent pas négligeables. C'est leur évaluation qui est problématique, aussi bien pour la comparaison entre produits ou services que pour l'évaluation d'un processus en soi. En effet, les limites écologiques au développement technologique, soit l'extraction de matières premières et les rejets de polluants, ne sont pas relatives mais bien absolues.

La qualité d'une analyse du cycle de vie (ACV) dépend largement de la qualité des données récoltées pendant les différentes étapes, de l'inventaire à l'évaluation d'impact (Crettaz, Jolliet, and Saade 2005). En particulier la phase d'inventaire du cycle de vie (ICV) requiert un grand nombre de données. Par définition celle-ci nécessite de quantifier les flux d'énergie et de matériaux ainsi que les émissions qui traversent les frontières d'un système (ISO 2006). L'utilité et la qualité d'une ACV dépendent donc essentiellement de la qualité des données d'inventaire (Ayres 1995; Mongelli, Suh, and Huppes 2005). C'est à ce stade de l'analyse que se concentrent les travaux développés dans cette thèse avec l'intention d'améliorer la qualité des ICVs, plus spécifiquement via la prise en compte des incertitudes liées au manque de données.

L'incertitude associée aux données en ACV faisait déjà l'objet d'une rencontre spéciale de la *Society of Environmental Toxicology and Chemistry* (SETAC) en 1992 et les recherches n'ont cessé depuis (Fava et al. 1994). Un groupe de travail *Data availability and data quality* était ensuite mis sur pied et ses recherches culminèrent avec la publication des premiers standards de données en ACV (ISO 2002). La collecte de données et surtout le format sous lequel elles se présentent étaient considérablement améliorés. Par ailleurs la nécessité d'évaluer les incertitudes

en ACV était officiellement acquise. Néanmoins l'analyse des incertitudes reste peu fréquente, pour différentes raisons énumérées plus loin. La pratique courante se limite généralement à la propagation des erreurs à partir des données d'inventaire jusqu'à l'évaluation par catégorie d'impact, ou analyse de sensibilité (Lloyd and Ries 2007). Dans la plupart des cas, les analystes font appel aux simulations de type Monte Carlo, même si des techniques analytiques existent aussi. La majorité des études évaluent la sensibilité des résultats à un ou plusieurs paramètres plutôt que l'incertitude des données d'inventaire proprement dites.

Le manque de données d'inventaire est un facteur d'incertitude en soi qui mène généralement à l'usage de données génériques provenant d'études précédentes ou de base de données publiques et privées. La substitution de données manquantes par des données de sources secondaires présente cependant des inconvénients. Par exemple en évaluant l'impact de la production d'électricité, l'incertitude des résultats est plus élevée lorsque les données proviennent de sources secondaires que primaires, autrement dit empiriques (May and Brennan 2003). Des différences entre la qualité des inventaires sont également observables pour d'autres processus en comparant les sources de données (Peereboom et al. 1998; Miller and Theis 2006). Les limites à la collecte de données peuvent être tantôt imposées par la taille et la complexité des processus, tantôt par la confidentialité des données ou encore l'absence de connaissance sur un processus spécifique (Pehnt 2003). Par opposition à la modélisation de processus, l'approche développée ici consiste à estimer les données manquantes sur base d'échantillons grâce à des estimateurs statistiques.

L'exemple de la production d'électricité n'est pas anodin, quasi tous les processus nécessitent de l'énergie sous une forme ou une autre, et la plupart des procédés industriels requièrent de l'électricité (Curran, Mann, and Norris 2005). De plus les besoins en électricité génèrent souvent une grande part des impacts pour de nombreux procédés (Matsuno and Betz 2000; Weber et al. 2010). En effet les bases de données d'inventaire non représentatives d'un contexte technologique ou géographique seraient adaptées en priorité par modification du poids des différentes sources de production d'électricité dans l'approvisionnement local, régional ou sectoriel (Marriott, Matthews, and Hendrickson 2010). Développer des inventaires fiables pour la production d'électricité est une nécessité, non seulement parce que les usages de l'électricité sont vastes mais également de façon à réaliser des ACV plus exactes (de Haes and Heijungs 2007).

Autre particularité de l'électricité, la fin de vie qui contrairement à d'autres commodités n'est pas une décharge ou un incinérateur. D'où l'importance donnée à sa production en concentrant les cas d'applications sur les centrales hydroélectriques. L'approche méthodologique proposée ici n'en reste pas moins applicable à d'autres secteurs clés de l'économie dont les caractéristiques sont semblables.

Cette thèse se compose de trois articles, chacun présentant un aspect de la méthode à partir de données différentes. Le premier, relativement théorique, démontre les avantages d'un estimateur statistique en particulier et comment sélectionner un modèle adéquat. Les données proviennent de la base de données ecoinvent pour des éoliennes de puissance variable (Burger and Bauer 2007). Après une collecte de données plus complète sur les centrales hydroélectriques et leur réservoirs, le deuxième article montre comment évaluer les flux d'énergie et de matériaux nécessaires à la construction et au fonctionnement des installations par rapport à leurs multiples caractéristiques. Enfin un troisième article propose de modifier l'approche classique dans certains cas ou les données disponibles seraient très variables, afin de prendre en compte les incertitudes spécifiques à chaque observation. Là aussi les données proviennent de barrages hydroélectriques mais cette fois ce sont les émissions de gaz à effet de serre (GES) qui sont analysées. Une revue de la littérature et une présentation détaillée de la problématique précèdent les articles eux mêmes. Trois approches à la méthodologie, d'abord une description du modèle tel que modifié pour l'ACV, ensuite les explications spécifiques dans chacun des articles et finalement certains aspects mathématiques placés en annexe. Une discussion et des recommandations sur le cadre d'application du krigeage closent la thèse.

Chapitre 1 CADRE ET REVUE DE LITTÉRATURE

Le sujet abordé ci-après est propre à de nombreux outils quantitatifs d'évaluation d'impact sur l'environnement, et plus particulièrement à l'analyse du cycle de vie, le bilan carbone ou encore l'analyse de flux d'énergie et de matériaux. Bien que la plupart des références proviennent du domaine de l'ACV, il ne s'agit pas ici d'une analyse complète mais d'une contribution à l'inventaire du cycle de vie, une de ses étapes clés. Avec le développement de tels outils, les institutions que sont l'Organisation Internationale de Normalisation et Eurostat ont établi certaines règles notamment pour l'ACV et les bilans de matériaux et d'énergie, pour des systèmes très différents, d'une économie nationale à un produit ou service (EUROSTAT 2001; ISO 2006). Ces recommandations et normes stipulent qu'avant toute récolte de données, il importe de définir les objectifs d'une analyse, l'audience visée et les frontières du système étudié. Dans le cas d'une ACV, c'est la fonction du système qui compte, par exemple la production d'électricité, et l'unité fonctionnelle qui quantifie cette fonction, un kWh d'électricité injecté dans le réseau de distribution. Les flux d'inventaire désignent les quantités d'énergie ou d'émissions nécessaires pour exécuter cette fonction à hauteur de l'unité fonctionnelle. Pour un processus donné, ces flux d'énergie et de matériaux ainsi que les émissions sont généralement divisés en flux élémentaires et économiques (ISO 2006). Les premiers sont des matières premières, du pétrole, du charbon ou des déchets directement extraits ou rejetés dans l'environnement. Par exemple les minerais ou le phosphate proviennent de mines avant d'être transformés pour leur usage dans l'industrie ou l'agriculture. Toutes les émissions dans l'air, l'eau et les sols sont également considérées comme des flux élémentaires. Les seconds sont des matériaux, de l'énergie ou des services échangés entre processus, d'où leur nom de flux économiques. L'ensemble des flux élémentaires et économiques constitue l'inventaire du cycle de vie d'un processus selon l'unité fonctionnelle choisie. C'est à ce stade de l'analyse que ce projet de recherche entend apporter une contribution méthodologique afin de prendre en compte les données manquantes et les incertitudes correspondantes.

Une troisième étape d'évaluation d'impact du cycle de vie (EICV) n'est pas directement impliquée. L'approche méthodologique proposée ici, analyse pourtant les données d'inventaire de manière semblable à celle qui concerne actuellement les données EICV, issues d'expériences où les sujets sont plus nombreux qu'en ICV. Une revue des différentes approches présentée ci-dessous fait ainsi référence aux deux étapes du cycle de vie avec plus de détails pour l'étape d'inventaire. Bien que cette thèse concentre ses applications sur la production d'électricité, les résultats portent sur le système qui permet de fournir ce service. Or lui même, ses centrales, réservoirs (mines), ou lignes de transport, se subdivisent en éléments partageant les mêmes caractéristiques, turbines, transformateurs, etc. Tout système dont les spécifications techniques peuvent être isolées et ses éléments regroupés, est susceptible de bénéficier d'une approche semblable, de la fabrication de transistors aux processus de raffinage de produits pétroliers.

1.1 Méthodes d'analyse des incertitudes du cycle de vie

L'influence des données d'inventaire sur les résultats d'une analyse du cycle de vie, autrement dit la propagation des incertitudes à travers les étapes d'une analyse, a très vite fait partie intégrale de la recherche (Heijungs 1996; Peereboom et al. 1998). Nombreuses sont les techniques utilisées depuis, statistiques ou analytiques, pour évaluer l'incertitude des résultats. Trois catégories de méthodes se distinguent dont l'application séparée est recommandée pour conserver les qualités de chacune d'elles (May and Brennan 2003). La plus répandue simule à partir de distributions de probabilité les contributions des intrants sur les sortants, autrement dit une approche quantitative de l'analyse de sensibilité. Parmi elles, les simulations de Monte Carlo sont entrées dans les normes (p. ex. McCleese and LaPuma 2002; Sonnemann, Schuhmacher, and Castells 2003; Ciroth, Fleischer, and Steinbach 2004). Une autre façon de simuler l'incertitude des résultats consiste à appliquer différentes méthodes d'évaluation des impacts au même inventaire du cycle de vie (p. ex. Hung and Ma 2009).

Une deuxième catégorie procède de la même manière, excepté qu'il s'agit d'algorithmes plutôt que de simulations. Cette approche analytique fait notamment usage des séries de Taylor (Hong et al. 2010). Troisièmement, lorsque les données deviennent insuffisantes, des observations plus qualitatives peuvent améliorer l'analyse des incertitudes en s'intégrant dans une logique floue ou *fuzzy logic* (Ardente, Beccali, and Cellura 2004; Heijungs and Tan 2010). Ces

mesures d'évaluations sont multiples, de l'avis d'experts à la qualité des sources des données, et l'approche floue s'introduit parfois dans l'analyse multicritère (Benetto, Dujet, and Rousseaux 2008).

Les incertitudes provenant des inventaires du cycle de vie sont accompagnées d'autres sources qui influencent tout autant les résultats d'une ACV. Les frontières du système, hypothèses de rendement, etc. varient considérablement d'une étude à l'autre, pour des résultats surprenant dans certaines industries (Brandt 2012). Le développement de normes d'analyse du cycle de vie répond en partie aux problèmes causés par des études sur un même processus qui prétendraient à des conclusions très différentes. Une autre approche pour cerner l'incertitude due aux hypothèses et choix de modèles consiste à réaliser des scénarios, c'est à dire des inventaires plausibles avant d'en évaluer les impacts et de les comparer (Huijbregts et al. 2003). Tous les types d'incertitudes devraient être sujet à évaluation simultanée en ACV. De plus les techniques d'évaluation se chevauchent, et peuvent répondre à plusieurs interrogations sur les sources d'incertitude des résultats. Par contre ces méthodes translatent les incertitudes de la phase d'inventaire du cycle de vie à celle d'évaluation des impacts ou analyse de sensibilité. La raison d'être de cette thèse se situe en amont, à savoir quantifier l'incertitude des données d'inventaire proprement dites.

Évaluer l'incertitude des données d'inventaire se heurte à des écueils au moins aussi importants qu'en évaluation d'impact (Weidema 2009a). Les données manquantes par exemple sont souvent ignorées et implicitement égales à zéro (Ciroth, Fleischer, and Steinbach 2004). C'est là où l'approche cycle de vie atteint ses limites, à savoir quand un processus dont les données sont incomplètes aura un impact moindre qu'un équivalent dont l'inventaire est bien documenté. Lors de comparaisons entre deux produits ou processus, certaines données à fortiori inconnues, et apparaissant dans des proportions identiques peuvent être écartées. Le même raisonnement ne se justifie pas en termes absolus ou dans l'évaluation des conséquences dues à des changements de technologie ou de législation, aussi appelé ACV conséquencielle. Une des premières suggestions pour améliorer la fiabilité des données d'inventaire consiste en ajouts d'intervalles autour des valeurs observées (Chevalier and Le Teno 1996). Il s'agit là d'une approche quantitative que l'on distingue ici des qualitatives. L'expérience montre que ces deux approches ont plus de valeur séparément plutôt qu'intégrées dans un même cadre, qui pourrait se traduire par une perte d'information (May and Brennan 2003).

1.2 Qualification des données d'inventaire

Basé sur les travaux de Funtowicz et Ravetz (1990), l'évaluation qualitative ou semi quantitative des données d'inventaire est connue sous le nom d'approche pedigree (Weidema and Wesnaes 1996). A l'origine, cinq critères permettent d'estimer l'incertitude des données : leur représentativité géographique, technologique et temporelle, leur complétude et leur fiabilité. Ceux-ci sont évalués entre les scores 1, le meilleur, et 5, le pire. Weidema et sa collègue transforment ensuite ces scores en écarts type pour estimer l'incertitude des données d'inventaire (1996). Établir une distribution de probabilités pour représenter l'incertitude des données, pose par contre des problèmes lorsque la taille des échantillons est extrêmement réduite, comme indiqué dans la section suivante. Encore une fois, il est préférable de séparer les indicateurs de qualité des données d'une évaluation quantitative des incertitudes, par exemple en prenant compte de la disponibilité des données (Maurice et al. 2000). Cette double approche est retenue par une des principales bases de données d'inventaire (Frischknecht et al. 2005). Entre temps la méthode pedigree a été augmentée de 5 à 14 critères d'évaluation semi quantitatifs, appliqués à différents niveaux d'agrégation des données, flux individuel, processus et système de processus (Rousseaux et al. 2001). A noter que le choix du coefficient représentant l'incertitude des données y est calculé directement à partir de la variabilité des scores de chacun des critères, c'est à dire leur variance. L'incertitude des données peut généralement être réduite par des estimations plus exactes ou des mesures supplémentaires alors que leur variabilité, elle, est propre au phénomène observé. Les deux concepts sont facilement confondus (Björklund 2002). Néanmoins l'approche pedigree permet une substitution plus transparente des données manquantes par des données génériques. Cette pratique est largement répandue et sa validité, la justesse des données substituées, est laissée à la discrétion de chacun.

1.3 Quantification des données d'inventaire

Les expériences de quantification de données manquantes et de l'incertitude de données existantes en inventaire du cycle de vie sont relativement limitées et disparates. Aucun consensus au sein de la communauté ACV n'existe sur le type de modèles appropriés pour l'estimation de données manquantes en lieu et place de leur substitution (Sugiyama et al. 2005). Si le développement de méthodes quantitatives profite essentiellement à l'analyse de l'incertitude des

résultats d'une ACV, certaines expériences montrent qu'il est possible d'estimer les données manquantes afin de réduire l'incertitude autour des inventaires. Une hypothèse largement acquise en ACV veut que l'inventaire de *x* produits ou processus soit *x* fois supérieur ou inférieur à l'inventaire du produit ou processus en question. Une extrapolation ou interpolation linéaire est aussi appelée mise à l'échelle (scaling) déjà très utilisée en ingénierie des processus. Des règles ont été établies pour étendre des procédures semblables en ACV (Caduff-Kinkel et al. 2008). Celles-ci nécessitent pourtant un inventaire complet de données représentatives, en termes technologiques par exemple, à partir duquel l'inventaire d'un nouveau produits ou processus peut être calculé. Autre désavantage, ces estimations sont précisément insensibles à tout phénomène d'échelle ou comportement non linéaire (Shibasaki, Fischer, and Barthel 2007).

Deux grands axes de recherche se distinguent nettement; d'une part le développement de techniques faisant appel aux tables d'intrants et sortants pour un marché ou secteur économique donné, de l'autre l'application de méthodes statistiques. D'autres techniques, moins répandues existent aussi, par exemple la comparaison entre différentes sources de données d'inventaire, qui permet de valider ou d'invalider ces données (Jimenez-Gonzalez and Overcash 2000; Mongelli, Suh, and Huppes 2005). L'analyse des données d'inventaire à travers les lois de conservation de masse et d'énergie, s'assure que les bilans des inventaires soient cohérents (Geisler, Hofstetter, and Hungerbuhler 2004; Hau, Yi, and Bakshi 2007). En effet, avec le manque de données et des sources différentes, la probabilité d'incohérences – des processus qui créeraient de la matière – existe bel et bien. Une forme alternative de réconciliation des données d'inventaire avec la thermodynamique, déjà moins laborieuse, fait appel à la logique floue (Tan, Briones, and Culaba 2007). L'enseignement principal à retenir n'est pas tant les possibilités d'améliorer la qualité des données d'inventaire mais les dilemmes auxquels font face celles et ceux qui soupèsent les coûts et bénéfices de leurs applications. En pratique, l'analyse des incertitudes n'est pas qu'une question de crédibilité mais également de retour sur investissement.

1.3.1 Approches hybrides

L'analyse « input-output » élargie pour convertir des flux monétaires en flux de matière et d'énergie, est à la racine d'un outil nommé *Missing Inventory Estimation Tool* (MIET). MIET se base sur la valeur monétaire des échanges pour combler certaines données dont les quantités

physiques sont inexistantes ou inaccessibles (Suh and Huppes 2002). Le recours aux prix des ressources et capitaux présente là aussi des désavantages. En effet les prix des matières premières sont relativement volatiles et même si un inventaire s'en trouve plus complet, l'incertitude des données estimées peut s'avérer encore plus difficile à quantifier. De même, pour les biens de production qui ont généralement un coût économique élevé, leur faible impact par unité produite pourrait être surévalué. En outre, seuls les matériaux, l'énergie, les produits et services ou encore les déchets et émissions faisant partie du marché et dont les prix sont connus, peuvent être pris en compte. De plus les tables « input-output » ne couvrent qu'une fraction de l'économie de marché, les externalités sont par définition exclues. Une comparaison entre cette approche et une base de données d'inventaire classique a également permis d'isoler les problèmes liés aux différences d'agrégation entre les échanges économiques et les processus physico-chimiques (Mongelli, Suh, and Huppes 2005). L'analyse « input-output » s'oriente ainsi vers une approche hybride entre les données d'inventaire existantes et les données économiques (Williams, Weber, and Hawkins 2009). Les futures bases de données seront, elles aussi, hybrides ce qui ne manquera pas d'augmenter leur portée mais aussi leur complexité. Un avantage non négligeable sera la contre validation des données d'inventaire, une étape trop souvent ignorée en ACV (Ciroth 2006). L'approche « input-output » sert aussi d'autres objectifs que l'estimation de données manquantes, notamment une évaluation des échanges en analyse conséquencielle du cycle de vie, par exemple sur le marché de l'électricité.

1.3.2 Approches statistiques

La particularité de l'approche statistique, en analyse du cycle de vie et ailleurs, est d'incorporer les incertitudes avant de chercher à les réduire (Finnveden et al. 2009). Cela s'entend aussi bien pour l'estimation des données manquantes que pour l'analyse des incertitudes. Si par la même occasion ces dernières sont réduites, peu d'autres formes d'évaluation des incertitudes prétendent les incorporer et seraient donc comparables. Contrairement à l'analyse « input-output » environnementale, l'estimation statistique de données d'inventaire ne nécessite que peu d'information supplémentaire. Le niveau de complexité n'augmente pas comme avec l'intégration de données économiques, mais avec l'ampleur des systèmes étudiés. C'est précisément dans le contexte de systèmes aussi complexes que la production et distribution d'électricité et en

l'absence de données complète, qu'une approche statistique est intéressante. La majeure partie des contributions ayant fait avancer cette approche en inventaire du cycle de vie concerne toutefois la production de composés chimiques. Parmi toutes les substances produites, seule une fraction est suffisamment documentée pour faire l'objet d'analyses du cycle de vie (Wernet et al. 2008; Wernet et al. 2009). Si la confidentialité reste un obstacle à la récolte de données d'inventaire pour les produits chimiques, certains échantillons dépassent néanmoins la centaine d'observations (Capello et al. 2005; Wernet et al. 2008).

En statistiques, le pendant quantitatif à l'approche pedigree pour l'incertitude des données est de leur assigner une distribution de probabilités. Pour certains secteurs économiques ou les données sont relativement nombreuses, il est possible de tester si elles suivent ou non une loi de probabilités spécifique et d'en estimer les paramètres (Capello et al. 2005). Par manque de données d'échantillon – la plupart des processus élémentaires en ACV n'en comporte qu'une seule – il est fréquent d'assigner une distribution et un degré d'incertitude arbitraire ou proportionnel aux scores de l'approche pedigree. Les bases de données d'inventaire prennent des facteurs d'incertitude relativement uniformes qui consistent en écarts type, standards ou géométriques (Frischknecht et al. 2005). La distribution de choix est généralement log normale puisqu'il s'agit de quantités très physiques. D'autres distributions de probabilité sont aussi utilisées, par exemple triangulaire (Huijbregts 1998a). Le problème de l'estimation des paramètres d'une distribution reste entier, en particulier la moyenne et variance (Heijungs and Frischknecht 2005). Un moyen classique en statistiques pour l'estimation de ces paramètres est de recourir au maximum de vraisemblance (Maximum likelihood estimator ou MLE). Une fonction appelée vraisemblance est définie de sorte que les données, indépendantes et identiquement distribuées, soient à leur tour considérées comme paramètres et ces derniers comme variables. L'estimateur des paramètres recherchés est celui qui maximise le logarithme de la fonction de vraisemblance. Les logiciels statistiques possèdent cette fonction qui est aussi appliquée en ACV (Sugiyama et al. 2005).

Un fois dotées d'une distribution de probabilités et de ses paramètres, les données peuvent être considérées comme variables aléatoires et utilisées par exemple pour des simulations. Les données sont échantillonnées un nombre de fois suffisamment grand à partir de leur distribution. Comme cité plus haut les simulations de Monte Carlo sont bien intégrées en évaluation d'impact

du cycle de vie, notamment pour effectuer des analyses de sensibilité (Sonnemann, Schuhmacher, and Castells 2003). Un exemple particulier d'application des simulations de Monte Carlo est presque à l'image de l'analyse d'inventaire du cycle de vie : itératif. Au fur et à mesure des simulations, le but est de réévaluer la distribution de probabilités des données d'inventaire en fonction de la sensibilité des résultats (Lo, Ma, and Lo 2005). C'est l'approche bayésienne ou le choix d'une distribution et de ses paramètres est justifié ou non a posteriori. Enfin, d'autres données sont associées à des distributions de probabilités en EICV, en particulier les données épidémiologiques, toxicologiques et écotoxicologiques. Les résultats de tests en laboratoire sont généralement dérivés à partir de sujets assez nombreux pour qu'une structure se dessine. Un facteur de sécurité est ensuite appliqué pour extrapoler les effets toxiques sur d'autres milieux et espèces, le corps humain par exemple. La aussi ces distributions de probabilités permettent d'analyser le choix de modèle d'impact.

Le développement de modèles statistiques pour l'estimation de données d'inventaire est très récents. A l'exception des travaux de recherche de Wernet et ses collègues (2008; 2009), réalisés avec des échantillons relativement grands, le modèle proposé ici est unique. Il offre notamment plus de flexibilité avec moins de données. En général, un modèle statistique incorpore aussi bien des comportements systématiques qu'aléatoires (McCullagh and Nelder 1989). Autrement dit, un estimateur statistique offre une alternative potentiellement plus juste et simple pour estimer des données manquantes à partir d'échantillons restreints et très variables. De plus l'incertitude des données est souvent exprimée sous forme d'écarts type ou d'intervalles de confiance, qu'une approche statistique évalue aussi de manière assez simple.

Parmi les outils statistiques qui ont mené progressivement à l'application de modèles linéaires d'estimation des données d'inventaire, se trouve les coefficients de corrélation. Ceux-ci permettent d'évaluer les liens entre flux d'inventaire et de les étendre aux données manquantes, par exemple dans le cas de l'inventaire de composés chimiques comme le polyéthylène terephthalate (PET) (Sugiyama et al. 2005). Plus proche de la méthode proposée ici, les propriétés chimiques de certains produits peuvent être corrélées aux flux d'énergie et de matériaux ainsi qu'aux émissions inhérentes à leur production. En rassemblant des données pour un maximum de produits, cette approche permet d'estimer la quantité de ces flux qui serait ignorée dans l'analyse d'un produit donné (Dahllof and Steen 2007). Auparavant la stœchiométrie

de composés chimiques avait servi à calculer des valeurs de flux d'inventaire non disponibles (Hischier et al. 2005). Dans la même optique, et pour estimer aussi bien les besoins en énergie et matériaux et les émissions rejetées, que leurs incertitudes, un modèle statistique connue sous le nom de réseaux de neurones artificiels, utilise également les propriétés physico-chimiques des substances étudiées (Wernet et al. 2008). Le modèle est calibré et validé sur un échantillon d'une centaine de substances dont les différentes propriétés physico-chimiques sont documentées et accessibles. C'est précisément la démarche suivie ici en faisant appel aux caractéristiques techniques de produits ou processus sans pour autant connaître les détails de leur provenance ou de leur fonctionnement. Les résultats montrent qu'il est possible d'améliorer l'estimation des données manquantes et la prise en compte des incertitudes avec des estimateurs tels que les réseaux de neurones plutôt que la régression linéaire (Wernet et al. 2008).

L'importance de ce résultat rend possible l'estimation de données manquantes et de leurs incertitudes à partir de caractéristiques facilement observables, sans connaître a priori la physique ou la chimie du produit ou processus en question. Si les réseaux de neurones artificiels offrent des performances acceptables, un autre modèle appelé krigeage montre des résultats supérieurs dans d'autres domaines d'application (Matías et al. 2004). Ces deux modèles sont pourtant équivalents dans certains cas, sans compter la flexibilité du krigeage et son efficacité dans la reconstruction de données (Gunes, Sirisup, and Karniadakis 2006; Baccou and Liandrat 2011). Ce dernier est décrit en détails dans l'approche méthodologique.

Comme dans de nombreux domaines la régression linéaire est également appliquée en ACV, notamment pour estimer l'inventaire du cycle de vie de réseaux de production d'électricité (Hondo 2005). En statistiques toujours, une méthode d'estimation de données manquantes assez répandue est l'imputation, qui procède de plusieurs manières différentes (Tsiatis 2006). Celle-ci est également proposée pour estimer des données d'inventaire manquantes, par imputation multiple (Liu et al. 2009). Les données manquantes sont remplacées par un set de données plausibles dont la variation reflète en même temps leurs incertitudes. Les résultats montrent que l'imputation multiple améliore elle aussi la fiabilité et la qualité des inventaires. Avec les méthodes de vraisemblance et des estimateurs comme le krigeage ou les réseaux de neurones artificiels, l'imputation complète ce tableau des approches courantes pour estimer des données manquantes.

1.4 Conclusions de la revue de littérature

Quelque soit l'intention, de qualifier ou quantifier les incertitudes, l'approche statistique est nettement différente d'une analyse d'inventaire classique. De fait, tout son intérêt lorsque les données manquent est de faire abstraction de l'origine des produits ou services, ou des détails de chaque étape d'un processus, pour se concentrer sur les relations entre leurs spécifications techniques et leurs flux d'inventaire. Ce contraste avec les modèles de processus par bilan de matière et d'énergie ou taux de transfert, se reflète aussi entre les modes d'acquisition de données d'inventaires et de données (eco)toxicologiques en évaluation d'impact du cycle de vie. En ACV proprement dit, les approches statistiques restent limitées. Pourtant en pratique les analyses de sensibilité réalisées grâce aux simulations de Monte Carlo, dépendent largement des distributions de probabilités pour quantifier l'incertitude des données. C'est donc en amont que les approches statistiques sont susceptibles de contribuer de manière significative à la qualité des inventaires du cycle de vie. En aval, les approches hybrides ont les faveurs des bases de données d'inventaire plus que l'estimation statistique (Weidema 2009).

Chapitre 2 PROBLÉMATIQUE

Passion is inversely proportional to the amount of real information available.

Gregory Benford, Timescape, 1980

Un problème rencontré presque invariablement dans l'application de tous types d'évaluation quantitative des impacts environnementaux est l'absence de données. Celle-ci se manifeste plus fortement en analyse du cycle de vie par le fait que chacune des étapes, de la définition des objectifs à l'évaluation des impacts, nécessite des données exhaustives. En effet, l'utilité de l'ACV dépend de données valides et vérifiables (Ayres 1995). Le manque de données, représentatives ou non du contexte temporel, géographique et technologique pour un système donné, continue ainsi d'affecter la fiabilité des ACV et plus particulièrement sa phase intermédiaire d'inventaire du cycle de vie (Reap et al. 2008). Si un produit mal documenté se retrouvait mieux évalué qu'un autre pour lequel les données d'inventaire disponibles sont plus adéquates, l'analyse perd toute crédibilité. D'où l'importance d'évaluer les incertitudes afin de ne pas discréditer l'ACV (Weidema 2000; Williams, Weber, and Hawkins 2009). Le manque de données se décline différemment selon les domaines d'analyses, tant et si bien qu'il est utile de préciser sa signification en statistique et en ACV, deux domaines entre lesquels les allées et venues sont fréquentes dans les pages qui suivent. En revanche, une caractéristique du manque de données reste commune à de nombreux outils d'évaluation d'impact, l'incertitude des résultats s'en trouve généralement plus élevée.

2.1 Incertitudes des données

Le manque de données a son propre principe en statistique, tant ce problème est répandu. L'énoncé du *Missing Information Principle* est le suivant : un problème à priori simple requiert

une analyse plus complexe lorsque l'information manque (Orchard and Woodbury 1972). Les données manquantes vont de non réponses dans un sondage aux valeurs observables uniquement à des coûts excessifs ou confidentielles. Trois grandes catégories de données manquantes existent (Little and Rubin 2002). Tout d'abord la probabilité que des données manquent peut être totalement indépendante du reste de l'échantillon (*missing completely at random* ou MCAR). Un exemple est l'absence de données suite à la perte, pure et simple, de mesures. A noter que les cas où cette perte n'est pas la conséquence d'un événement imprévisible, peuvent refléter d'autres problèmes que l'absence de données elle-même. Ensuite, la probabilité que les données manquent peut dépendre uniquement des observations, tout en restant aléatoire (*missing at random* ou MAR). Par exemple, si deux mesures sont prises et qu'une troisième est requise lorsque les deux premières diffèrent largement, cette troisième mesure manque si les deux premières concordent. Finalement la catégorie la plus délicate concerne les données pour lesquelles les deux autres catégories ne s'appliquent pas. La probabilité qu'elles manquent peut alors dépendre de données non observables (*non missing at random* ou NMAR). En général, même si certaines données ne peuvent pas être observées, l'hypothèse MAR est peu restrictive, particulièrement dans les cas qui suivront ou plusieurs covariables ou variables secondaires sont, elles, observables (Tsiatis 2006).

Une autre source d'incertitude qui découle d'un manque de données est l'absence d'échantillon suffisant pour valider le choix d'un modèle ou de ses paramètres. Le statisticien John Tukey invente le concept de *uncomfortable science* pour décrire une situation où des sets distincts pour la calibration et validation d'un modèle sont exclus par manque d'observations (Hoaglin, Mosteller, and Tukey 1985). A priori, dans les cas où les données sont rares, les chances d'obtenir des échantillons importants sont relativement faibles. A noter qu'un large échantillon dont les incertitudes sont mal documentées ne compense pas nécessairement des données moins nombreuses mais plus fiables ou représentatives du contexte temporel ou technologique (Funtowicz and Ravetz, 1990). De plus, les données en inventaire du cycle de vie proviennent de sources diverses, à des coûts variables généralement plus élevés pour des données empiriques que pour des bases de données génériques ou des références bibliographiques.

2.2 Incertitudes des données d'inventaire

En ACV, et contrairement aux statistiques, le terme incertitude des paramètres correspond à l'incertitude des données qui se distingue d'autres formes d'incertitude, notamment celles qui proviennent des choix de modélisation (Huijbregts 1998b; Ross, Evans, and Webber 2002). Dans cette thèse, les deux terminologies sont parfois utilisées pour représenter le même problème, avec une nette préférence pour la dénomination statistique. De fait, les paramètres sont propres au modèle développé ici et non des données d'inventaire. Le terme de données, lui, est tantôt utilisé pour designer un ensemble de valeurs observées – les observations – tantôt pour une valeur spécifique, manquante ou estimée.

Le manque de données se subdivise encore en inventaire du cycle de vie. Une distinction importante réside entre d'un coté le manque de données représentatives d'un processus à l'étude et de l'autre l'absence pure et simple de données pour caractériser ce processus. Cette classification de l'incertitude des données en inventaire du cycle de vie est un des résultats du groupe de travail *Data availability and data quality* (Huijbregts et al. 2001). Son importance est essentiellement liée aux différentes approches qui s'appliquent dans un cas comme dans l'autre, par exemple simulation et estimation statistique. Si le manque de données est une des principales sources d'incertitude en inventaire, affectant leur qualité ainsi que les résultats d'une ACV, le part due à l'absence de données représentatives ou non est difficilement quantifiable (Björklund 2002). Certains experts affirment que 30 à 70 % des flux d'inventaire manquent sans pour autant préciser de quel type de données il s'agit (Weidema 2009). La disponibilité des données est souvent limitée par les facteurs décrits dans le tableau 1. En pratique toute estimation nécessite un minimum de données observées, qu'elles soient connexes ou non représentatives. S'il y a substitution de données manquantes il existe nécessairement au moins une observation servant de proxy. Dans une approche statistique la disponibilité d'un échantillon est un prérequis essentiel, d'où l'intérêt pour des méthodes qui proviennent de domaines ou les données observées sont aussi rares, indépendamment des causes décrites ci-dessous.

Tableau 1: Principales causes du manque de données en ACV (adapté de Pehnt 2003)

Type	Causes	Description	Exemple
I	Confidentialité	Données non disponibles pour des raisons de confidentialité	Marchés compétitifs, militaires
II	Complexité	Processus complexes ralentissant la collecte de données, relations causes à effets méconnues	Niveau d'intégration des processus élevé
III	Nouveauté	Processus et technologies en cours de développement	Nouveaux produits, nanotechnologies
IV	Contexte	Comportement des systèmes à l'étude imprévisible ou méconnu	Dérégulation du marché de l'électricité
V	Connaissances	Absence de connaissances sur les processus, le contexte ou les deux	

Que ce soient des bases de données ou des mesures sur le terrain, les causes les plus importantes quant au manque de données dans les applications qui suivent sont leur confidentialité et la complexité des processus étudiés. Les réseaux de production, transport et distribution d'électricité continuent de se développer à marche forcée, pour cause d'urbanisation croissante (Filion and Hebert 2004). Une des grandes tendances est la reconversion de barrages hydroélectriques en stations de pompage turbinage, une forme de batterie très lucrative qui complique l'évaluation des émissions par kWh d'hydroélectricité produit (Mijuk 2010; Deane, Gallachoir, and McKeogh 2010). Les connaissances continuent d'avancer sans toutefois fournir les réponses à tous les problèmes soulevés par la complexité des réseaux et les incertitudes demeurent élevées (Weber et al. 2010). En général, les données empiriques sont plus représentatives d'un contexte ou processus que les données génériques, bien que les premières fassent souvent partie des secondes après publication.

2.3 La problématique résumée

En résumé, le problème spécifique au travail de recherche présenté ci-après est le manque de données en inventaire du cycle de vie et comment y remédier. Étant donné les caractéristiques suivantes :

a) les contraintes sur la disponibilité des données d'inventaire, les besoins importants en données et leur manque chronique,

b) la complexité et l'ampleur de certains processus impliqués et imbriqués en analyse du cycle de vie,

c) l'importance des données représentatives du contexte temporel, géographique et technologique pour la qualité des inventaires,

les approches conventionnelles, par modélisation de processus ou encore par analyse « input-output », ont leurs limites que certains outils statistiques peuvent dépasser. A l'échelle de certains processus, les différences entre inventaires basés sur des données empiriques versus des données génériques, peuvent s'avérer significatives. L'idée principale est d'estimer les données manquantes à partir de doubles échantillons de données, primo les flux de matériaux et d'énergie et les émissions pour un processus spécifique et secundo les caractéristiques ou spécifications techniques de ce même processus. Le problème du manque de données n'est pas nouveau et les outils statistiques proposés non plus. Par contre les modifications qui y sont apportées, l'interprétation du problème et sa résolution, le sont.

Chapitre 3 OBJECTIFS

Cette thèse tente de répondre à deux objectifs principaux. Tout d'abord estimer les données manquantes pour améliorer la fiabilité des inventaires du cycle de vie. Ensuite, incorporer de manière simple l'incertitude des données existantes et évaluer celle des données estimées toujours dans l'optique d'augmenter l'exactitude des inventaires. D'où la proposition de modifier certains estimateurs statistiques pour lesquels la quantité de données additionnelles nécessaires reste limitée. Ceci mène directement aux questions suivantes :

Comment estimer des données d'inventaire alors que précisément elles sont absentes? Des outils statistiques sont ils mieux en mesure d'estimer des données manquantes que d'autres? Pourquoi a priori le krigeage est il mieux adapté que d'autres estimateurs linéaires?

Ces questions orientent le travail de recherche et si la revue méthodologique donne des réponses à la question de l'estimation statistique des données, elle n'en reste pas moins marginale en inventaire du cycle de vie. Il ressort donc que de combler ce vide méthodologique ajoute une valeur aussi bien théorique que pratique à l'analyse du cycle de vie.

L'hypothèse de recherche consiste à dire que oui, certains modèles statistiques, le krigeage en particulier, permettent d'estimer plus facilement et intégralement les données manquantes en inventaire.

Comme pour tous les modèles, quelques généralités s'imposent. Par essence un modèle est approximatif et ce sont les divergences entre les valeurs estimées et celles observées qu'il faut minimiser. Ces différences sont souvent nommées erreurs statistiques et confondues avec résidus. Parmi les propriétés d'un modèle, deux d'entre elles sont particulièrement souhaitables, la simplicité et l'étendue d'applications (McCullagh and Nelder 1989). A l'extrême, un modèle peut contenir autant de paramètres que d'observations et aucune erreur n'est commise. Un tel modèle serait inutilement lourd et contraint à un seul cas d'application. Un nombre raisonnable de paramètres est à mettre en opposition à une validation parfaite. De même un modèle doit couvrir

un champ d'application relativement large et avec exactitude. Afin de tester l'hypothèse de recherche et la validité du modèle, la moyenne des erreurs fournit suffisamment d'évidences pour ou contre celles-ci. Dans le détail, les objectifs de l'étude sont les suivants.

3.1 Objectifs spécifiques

L'objectif principal est de développer des outils statistiques, dont le krigeage, afin d'améliorer la qualité des inventaires du cycle de vie en estimant les données manquantes. A noter que les données d'inventaire, contrairement aux applications usuelles du krigeage, ne sont pas à priori distribuées dans l'espace. La relation qui importe ici repose d'un côté sur la disponibilité de données techniques, les spécifications qui caractérisent un objet, par exemple la puissance d'une turbine ou le voltage d'un transformateur. De l'autre les flux de matériaux et d'énergie ainsi que les émissions se rapportant au même objet et dont la disponibilité est fonction des contraintes énumérées au tableau 1. Ces derniers flux sont les variables dépendantes alors que les spécifications techniques sont les variables indépendantes du système. Celles-ci jouent le rôle de dimensions spatiales et sont généralement accessibles alors qu'arpenter l'ensemble des centrales d'un réseau de production, à la recherche de données pour les matériaux, l'énergie ou les émissions, semble au delà d'un horizon temporel raisonnable.

Les objectifs spécifiques sont repris ci dessous dans l'ordre dans lequel ils sont approfondis par la suite :

a) Établir des critères de comparaison pour différentes techniques statistiques, essentiellement des modèles linéaires dont le krigeage. Les comparer de manière à rassembler des arguments en faveur ou non de l'hypothèse de recherche. A posteriori, les données collectées pour les variables dépendantes sont d'ordres de grandeur très différents et donc normalisées pour une comparaison par erreurs moyennes absolues.

b) Premièrement, tester différents types de krigeage sur des échantillons restreints. Les résultats de ces tests sont contenus en grande partie dans les travaux précédents et inclus dans le premier article intitulé *Estimating material and energy flows in life cycle inventory with statistical models*. Deuxièmement, évaluer l'influence des paramètres du krigeage sur les résultats ainsi que les possibilités d'interprétation de ces paramètres en

inventaire du cycle de vie. Les conclusions des premier et deuxième articles sont les plus pertinentes à ce sujet, incluant des modifications à la résolution du système de krigeage.

c) Le troisième objectif visait à sélectionner un certain nombre de propriétés techniques, représentatives des objets à l'étude. Le niveau de résolution disponible pour la collecte de données d'inventaire est resté relativement bas si bien qu'à l'échelle d'une centrale électrique par exemple ou d'une éolienne, ces variables caractéristiques sont peu nombreuses. Elles n'en sont pas moins importantes pour l'ensemble du travail de recherche et résumées dans le deuxième article *Statistical estimation of missing data in life cycle inventory: an application to hydroelectric power plants.*

d) Appliquer la méthode à plus grande échelle, avec les données disponibles et estimer les flux de matériaux et d'énergie ainsi que les émissions pour des centrales encore en construction. L'évaluation d'impact du cycle de vie est au delà du cadre de ce travail, pourtant dans la mesure ou les applications sont principalement énergétiques, une évaluation des émissions de gaz à effets de serre apporte déjà un élément de comparaison. Les deuxième et troisième articles contiennent les données collectées pour les flux de matière et d'énergie et les émissions de GES, respectivement.

e) De deux choses; l'une définir un champ d'application des techniques d'estimation statistique des données d'inventaire. L'autre, adapter et montrer comment ces techniques prennent en compte l'incertitude des données observées et évaluent celle des données estimées. Cette dernière partie est essentiellement développée dans le troisième article qui s'intitule *Handling data variability and uncertainty with kriging: the case of hydroelectric reservoir emissions,* alors que les limites du krigeage sont discutées dans chacun des articles et dans la discussion plus large qui s'ensuit.

Malgré certains ajustements les objectifs ont été atteints et sont détaillés dans les articles mentionnés. Une description de l'approche méthodologique permet de compléter les bases sur lesquelles s'appuie ce travail de recherche. Certains détails sur le krigeage lui même se retrouvent en annexe, avec les informations additionnelles pour chaque article.

Chapitre 4 MÉTHODOLOGIE

Invention is the mother of necessity.

Melvin Kranzberg, Technology and History, 1986

La démarche servant à tester l'hypothèse de recherche et répondre aux objectifs du projet est présentée ci-dessous. Cette section décrit l'approche de façon générale et transversale, ne se substituant pas entièrement aux articles qui suivent. En effet, ces derniers considèrent chacun des objectifs de manière plus séquentielle. Il convient d'abord d'expliquer la recherche et récolte de données pour ensuite détailler la méthode elle même.

4.1 Sources des données

Les données proviennent de sources autant primaires que secondaires, en commençant par la société Hydro-Québec, qui a fourni des données en grande partie confidentielles, la base de données d'inventaire existante, ecoinvent, et un recueil de travaux sur le démantèlement des barrages : Clearinghouse for Dam Removal Information (ecoinvent Centre 2009; CDRI). Une large revue de la littérature publiée dans des journaux scientifiques et sous forme de rapports, par exemple par la Commission mondiale des barrages, la Commission international des grands barrages, ou le Comité suisse des barrages, complète ce tableau (World Commission on Dams 2000; ICOLD 2011). Une petite partie des données récoltées se trouvent dans le domaine public alors que d'autres sources de données, privées en l'occurrence, ont été écartées pour leurs coûts prohibitifs, notamment la base de données World Electric Power Plant Database, de la société Platts. Le choix d'analyser la production hydroélectrique se justifie d'abord par la collaboration avec Hydro-Québec et ensuite par la controverse autour du bilan carbone des barrages hydroélectrique ainsi que le manque de données d'inventaire pour la construction de grands barrages (Gleick 1992; Cullenward and Victor 2006; Ribeiro and Silva 2010).

4.1.1 Types de données

En analyse du cycle de vie, les données d'inventaire se présentent idéalement sous forme de processus élémentaires, par exemple l'irrigation d'un champ ou la construction d'un barrage hydroélectrique. Pour un nombre croissant de processus, les flux d'inventaire sont quantifiés par rapport à une unité de référence, un hectare irrigué ou un barrage construit. S'y retrouvent d'abord les flux élémentaires, intrants provenant de l'environnement et sortants émis dans l'air, les sols et l'eau. Ensuite, les flux économiques sont en général eux mêmes des processus, en amont ou en aval du processus en question. Ces flux possèdent chacun une quantité, unité, et, dans la mesure du possible, une variance et distribution de probabilités. Les données primaires, collectées directement auprès d'industriels sont souvent d'une autre nature et nécessitent d'être traitées avant utilisation. Pour les données secondaires, comme illustré dans le tableau 2, une documentation précise facilite la substitution de données manquantes en inventaire par ces données génériques. Plus que de substituer les données manquantes, dans la pratique la plupart des données pour les processus en amont, aval ou arrière plan, proviennent de bases de données. D'où l'importance d'améliorer la qualité des inventaires en estimant les données manquantes de processus aussi essentiels que la production d'électricité.

Tableau 2: Données d'inventaire pour la construction et l'élimination d'un barrage (adapté de ecoinvent Centre 2009)

	Quantité	Unité	Distribution	σ^2
Flux économiques				
Chromium steel 18/8, at plant	1.92E+08	kg	Log-normale	3
Diesel, burned in building machine	5.15E+09	MJ	Log-normale	1.7
Explosives, tovex, at plant	4.87E+07	kg	Log-normale	3
Gravel, round, at mine	7.80E+10	kg	Log-normale	3

Portland cement, strength class Z 42.5, at plant	8.97E+09	kg	Log-normale	3
Reinforcing steel, at plant	1.71E+08	kg	Log-normale	3
Steel, low-alloyed, at plant	7.68E+08	kg	Log-normale	3
Tap water, at user/RER U	4.95E+09	kg	Log-normale	3
Electricity, medium voltage, production UCTE, at grid	2.82E+09	kWh	Log-normale	2.1
Transport, lorry >16t, fleet average	4.10E+08	tkm	Log-normale	3
Transport, freight, rail	2.23E+09	tkm	Log-normale	3
Flux élémentaires rejetés dans l'air				
Heat, waste	1.02E+10	MJ	Log-normale	2.1
Particulates, < 2.5 μm	2.10E+06	kg	Log-normale	4.3
Particulates, > 10 μm	2.43E+07	kg	Log-normale	3.3
Particulates, > 2.5 μm, and < 10 μm	9.22E+06	kg	Log-normale	3
Élimination en fin de vie				
Disposal, building, reinforced concrete, to final disposal	9.21E+10	kg	Log-normale	2.1

Pour l'estimation de données manquantes grâce aux modèles statistiques expliqués ci-dessous, deux types de données ont été recueillis. D'une part les caractéristiques techniques d'un système, les cas d'éoliennes et de centrales hydroélectriques sont étudiés ici, d'autre part les flux d'énergie et de matériaux ainsi que les émissions qui s'y rapportent, à l'image de ceux présentés dans le tableau 2. Les caractéristiques techniques sont des variables, indépendantes, par exemple le design d'une turbine, sa capacité de production ou celle de toute une centrale, la production effective, etc. Ensemble, ces deux types de données décrivent des groupes fonctionnels qui font partie d'un même échantillon. Contrairement aux données d'inventaire, les spécifications

techniques sont en général aisément disponibles à des coûts minimes, soit dans la description même des installations de l'exploitant ou du fabricant. Pourtant la production effective annuelle d'une centrale peut être gardée confidentielle et largement inférieure à la capacité installée, ces deux variables sont donc complémentaires. Les producteurs dont les installations sont multiples et de types différents n'ont pas nécessairement intérêt à rendre ces informations publiques puisqu'elles peuvent indiquer comment ils répondent aux variations de la demande et ainsi dévoiler leurs coûts d'exploitation. En revanche obtenir les quantités de matériaux, d'énergies, de consommables et de rejets dans l'air, l'eau et les sols est beaucoup plus laborieux. Consulter les plans et autres factures pour chaque élément d'une centrale peut s'avérer une tache de longue durée, voire irréalisable. Pour cette raison, l'estimation à partir d'échantillons disponibles est importante.

Afin de rassembler un échantillon relativement homogène, la durée de vie des équipements et infrastructures est un facteur essentiel (Ribeiro and Silva 2010). Celle-ci est respectivement de 50 et 100 ans en moyenne et peut varier en fonction de nombreux paramètres, notamment le mode de fonctionnement, si l'installation couvre la charge de base ou de pointe. Comme pour des groupes multifonctionnels en ACV il arrive qu'une infrastructure tel qu'un barrage serve de réservoir à plusieurs centrales, nécessitant une forme d'allocation des ressources et des émissions à chacune des centrales concernées. La puissance d'une centrale hydroélectrique est fonction de deux facteurs principaux, la hauteur de chute et le flux d'eau dans les turbines, selon la relation suivante :

$$P = \eta \rho \, g \, Q \, H$$

ou P est la puissance (W), η un coefficient d'efficacité, ρ la densité de l'eau (~1000 kg/m3), g l'accélération du à la gravité (~9.8 m/s^2), Q le flux (m^3/s) et H la hauteur de chute (m). Une forme d'allocation peut donc être réalisée en pondérant par la hauteur de chute et le flux, autrement dit la puissance ou au contraire par la production moyenne annuelle et donc la quantité moyenne d'eau turbinée, la hauteur de chute ne variant que très peu.

En dehors des caractéristiques techniques, les flux d'énergie et de matériaux principaux en terme de masse, le béton et l'acier, ne sont pas nécessairement homogène. Par exemple, si le ciment est généralement transporté sur de longues distances, sa concentration dans le béton

mélangé sur place, varie en fonction de l'utilisation finale. Idem pour l'acier dont la qualité varie s'il s'agit de l'armature ou de l'équipement d'une centrale. Avec un échantillon suffisamment détaillé, il est ainsi préférable de conserver les catégories de matériaux séparément. A l'exception des émissions de deux des principaux gaz à effet de serre, la récolte de données prend essentiellement en compte les flux économiques et parmi eux les grands contributeurs. Comme pour de nombreux processus industriels, l'eau est la grande absente et son turbinage commence seulement à être caractérisé en ACV (Pfister, Saner, and Koehler 2011). Les raisons de ce choix sont d'abord la disponibilité des données et ensuite l'importance en terme de masse et d'impact dans la mesure où il ne s'agit pas d'éléments toxiques ou radioactifs.

4.1.2 Présentation des échantillons

Le but derrière la récolte de données a été de deux ordres, progresser rapidement dans l'évaluation des modèles statistiques et assembler un échantillon aussi large que possible sur les centrales hydroélectriques et leurs réservoirs. La disponibilité et le traitement des données s'échelonnant dans le temps, plusieurs échantillons ont servi à expérimenter et tester les différents aspects statistiques. D'abord la validité du krigeage pour des systèmes technologiques, le choix des modèles de covariance et l'évaluation de leurs paramètres. Ensuite l'application à plus grande échelle notamment multidimensionnel du krigeage, ses avantages et inconvénients et enfin, comment le krigeage permet de prendre en compte les incertitudes liées aux données d'inventaire. La séparation en deux échantillons distincts des flux d'inventaire propres aux centrales et aux réservoirs, facilite cette démonstration par le fait que les centrales hydro-électriques sont des groupes fonctionnels dont la phase de construction domine le cycle de vie alors que la majorité des émissions associées aux réservoirs commencent avec leur remplissage. De plus ces dernières sont encore plus variables d'un réservoir à l'autre, d'où l'intérêt d'évaluer leurs incertitudes.

Dans un premier temps les données sont donc extraites de la base de données d'inventaire ecoinvent : les processus se référant à cinq éoliennes de taille variant entre 30 kWe et 2 MWe. Les estimations se basent ainsi sur leur puissance comme variable caractéristique. Alors que les quatre premières sont construites à terre, la cinquième est typique d'une installation en mer. Pour chacune d'entre elles les données sont divisées en deux groupes, la tour statique et le rotor qu'elle

soutient. Un tableau contenant toutes ces données est disponible en annexe, elles servent aux premiers objectifs du projet et notamment à l'article intitulé *Estimating material and energy flows in life cycle inventory with statistical models*.

La recherche et le traitement des données proprement dites commencent dans un deuxième temps. L'objectif est de récolter des données d'inventaire ultérieures à une première étude de la société Hydro-Québec (Peisajovich 1997). Des données plus récentes proviennent directement des nouvelles installations de la société d'état, ainsi que d'une analyse détaillée des installations existantes. La principale difficulté consiste à attribuer les données fournies à chacune des centrales étant donné le déroulement de l'expansion des installations. En effet, la récolte de données primaires par Hydro-Québec s'effectue par grand projet, incluant diverses infrastructures et plusieurs centrales construites parfois simultanément. Une caractéristique importante est l'éloignement des centrales de la majorité des consommateurs, les grands projets permettent ainsi de réaliser des économies d'échelles sur le transport des matériaux autant que de l'électricité une fois produite et transmise par de larges corridors à très haute tension.

Aux centrales de la zone boréale s'ajoutent les études relativement complètes sur sept centrales alpines, très différentes de situation et de conception (Bauer et al. 2007). D'un coté des centrales dont les faibles hauteurs de chute sont compensées par un flux important et de l'autre des faibles débits pour les plus grandes hauteurs de chute qu'il soit. Le niveau de détail permet une bonne correspondance entre les données. De même, l'analyse du complexe Itaipu sur la rivière Parana, entre le Brésil et le Paraguay, peut aussi être intégrée (Ribeiro and Silva 2010). Sur les 27 centrales de l'échantillon final, 10 sont des centrales au fil de l'eau, dont le réservoir est négligeable par rapport au cours d'eau. Comme expliqué ci-dessous, l'approche par krigeage admet une relativement grande variabilité des données, particulièrement adaptée aux situations où les données observées sont limitées et les besoins d'estimation importants.

Si la variabilité des données n'est pas étrangère aux incertitudes pré-existantes, le cas des émissions de gaz à effet de serre provenant des réservoirs hydroélectriques est édifiant. La controverse autour de leur quantification, aussi bien du dioxyde de carbone que du méthane et même des oxydes d'azote, provient du fait que les mesures sont d'abord réalisées par les exploitants eux mêmes ou par des chercheurs à leur solde (Cullenward and Victor 2006). Un

consensus sur les flux à prendre en compte fait également défaut, certaines études prenant en compte la diffusion et/ou l'ébullition, rarement le dégazage de l'eau turbinée. Néanmoins, grâce à une série d'études récentes, un échantillon de 26 réservoirs hydroélectriques dont environ un tiers pour chacun des climats boréal, tempéré et tropical, est assemblé (Huttunen et al. 2002; Soumis et al. 2004; Soumis et al. 2005; Delmas et al. 2005; dos Santos et al. 2006; Demarty et al. 2009). Le dégazage reste une composante difficile à cerner et les mesures de diffusion et d'ébullition se rapportent à la surface des réservoirs, sans nécessairement prendre en compte leur profondeur. Un facteur soigneusement pris en compte par la plupart des études sur les réservoirs en milieu boréal est le nombre de jours sous couvert de glace durant lesquels les émissions seraient nulles. Pourtant des tests préliminaires montrent qu'au contraire, l'accumulation de gaz sous la glace, dégagée lors de la débâcle, n'est pas négligeable (Duchemin et al. 2006). En revanche il est encore plus difficile de mesurer les émissions causées par l'étiage, autrement dit la différence de surface entre les hauteurs maximales et minimales de l'eau dans les réservoirs. Certains réservoirs absorbent même du dioxyde de carbone, alors que des réservoirs de petites tailles en milieu tempéré peuvent émettre de grandes quantités de méthane (DelSontro et al. 2010). D'où la particularité de cet échantillon, la mesure des émissions étant nécessairement variable, un facteur d'incertitude est généralement rapporté sous forme d'écarts type pour chacune des études citées ci-dessus. L'intérêt est double, contrairement aux flux économiques d'une base de données d'inventaire qui proviendraient d'une seule observation, l'estimation des erreurs systématiques et aléatoires est plus fiable. De plus, cela permet d'illustrer la flexibilité du krigeage, les résultats sont présentés dans l'article intitulé *Handling data variability and uncertainty with kriging: the case of hydroelectric reservoir emissions*. Seul les émissions de dioxyde de carbone et de méthane sont prises en compte et uniformisées pour former un échantillon cohérent, en mg de gaz par m^2 de surface air-eau, par jour.

4.2 Modèle statistique

La méthodologie proposée est purement quantitative et se différencie des approches hybrides ne faisant appel à aucune donnée économique de type « input-output » afin de réduire la collecte de données, limiter les sources d'incertitude et permettre de comparer l'exactitude de plusieurs estimateurs. Emprunté à la géostatistique, le krigeage est, sous certaines conditions, un

estimateur sans biais, à variance minimum, pour l'interpolation de processus aléatoires dans l'espace (Ord 2006). Il se distingue d'autres estimateurs linéaires par la minimisation de la variance d'estimation (Isaaks and Srivastava 1989). Soit $Z(x_i)$ les valeurs observées aux points x_i et $Z(x_0)$ la valeur recherchée. Lorsque la moyenne des observations est inconnue, l'estimateur, dit de krigeage ordinaire, prend la forme (1):

$$\hat{Z}(x_0)=\sum_{i=1}^{N}\lambda_i Z(x_i) \quad (1) \qquad \sum_{i=1}^{N}\lambda_i=1 \quad (2)$$

ou λ_i sont les poids du krigeage et N le nombre d'observations. Cet estimateur est sans biais en imposant la contrainte (2). La variance d'estimation (4) est définie comme suite, à partir de l'erreur d'estimation (3) :

$$e=Z(x_0)-\hat{Z}(x_0) \quad (3) \qquad Var[e]=Var[Z(x_0)]+Var[\hat{Z}(x_0)]-2\,Cov[Z(x_0),\hat{Z}(x_0)] \quad (4)$$

ou *Var* et *Cov* sont les opérateurs variance et covariance, respectivement. En substituant l'expression de l'estimateur (1) dans (4), la variance d'estimation s'écrit de la façon suivante (5) :

$$\sigma_e^2=Var[Z(x_0)]+\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j Cov[Z(x_i),Z(x_j)]-2\sum_{i=1}^{N}\lambda_i Cov[Z(x_0),Z(x_i)] \quad (5)$$

Le problème revient donc à minimiser σ_e^2 sous la contrainte (2). En utilisant la méthode de Lagrange, c'est-à-dire en prenant les dérivées partielles du lagrangien L (6) par rapport à λ et un paramètre de Lagrange μ, un système de $N+1$ équations linéaires à $N+1$ inconnues est obtenu.

$$\begin{aligned} L(\lambda,\mu)&=\sigma_e^2+2\,\mu\left(\sum_{i=1}^{N}\lambda_i-1\right) \\ &=Var[Z(x_0)]+\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j Cov[Z(x_i),Z(x_j)]-2\sum_{i=1}^{N}\lambda_i Cov[Z(x_0),Z(x_i)]+2\mu\left(\sum_{i=1}^{N}\lambda_i-1\right) \end{aligned} \quad (6)$$

Les dérivées partielles de L égales à zéro donnent le système suivant (7) :

$$\begin{cases} \sum_{j=1}^{N}\lambda_j Cov[Z(x_i),Z(x_j)]+\mu=Cov[Z(x_0),Z(x_i)] \quad \forall \quad i=1,\ldots,N \\ \sum_{j=1}^{N}\lambda_j=1 \end{cases} \quad (7)$$

Le système d'équations linéaires en (7) est plus facilement représenté sous forme matricielle (8) :

$$\underbrace{\begin{pmatrix} \sigma^2 & Cov[Z(x_1),Z(x_2)] & \cdots & Cov[Z(x_1),Z(x_N)] & 1 \\ Cov[Z(x_2),Z(x_1)] & \sigma^2 & \cdots & Cov[Z(x_2),Z(x_N)] & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Cov[Z(x_N),Z(x_1)] & Cov[Z(x_N),Z(x_2)] & \cdots & \sigma^2 & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}}_{K_0} \underbrace{\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \mu \end{pmatrix}}_{\lambda_0} = \underbrace{\begin{pmatrix} Cov[Z(x_1),Z(x_0)] \\ Cov[Z(x_2),Z(x_0)] \\ \vdots \\ Cov[Z(x_N),Z(x_0)] \\ 1 \end{pmatrix}}_{k_0} \quad (8)$$

Les poids λ_i peuvent ainsi être insérés en (1). A l'optimum, la variance du krigeage ordinaire devient (9) :

$$\sigma^2 = Var[Z(x_0)] - \sum_{i=1}^{N} \lambda_i Cov[Z(x_0), Z(x_i)] - \mu \quad (9)$$

Deux hypothèses principales sont propres au krigeage. La première est la stationnarité, c'est à dire une moyenne des données *m* constante (10) :

$$E[Z(x)] = m \quad (10)$$

A cette stationnarité de premier ordre s'ajoute une autre hypothèse qui avance que la covariance ne dépend pas de la position des observations mais de la distance *h* qui les sépare (11) :

$$Cov[Z(x), Z(x+h)] = C(h) \quad (11)$$

ou *C*(*h*) est appelé fonction de covariance. Autrement dit cette hypothèse traduit de manière explicite le fait que deux observations proches l'une de l'autre se ressembleront davantage que deux autres éloignées, en moyenne. Une certaine continuité du phénomène observé est requise puisque la seconde hypothèse du krigeage est l'existence d'une fonction de covariance valide pour le calcul des matrices K_0 et k_0 en (8).

Plusieurs fonctions de covariance sont couramment utilisées, celles-ci comportent en général trois paramètres. Au delà d'une certaine distance *h*, appelée portée *a*, un seuil est franchi où deux observations n'ont plus d'influence l'une sur l'autre, leur covariance est donc nulle. Outre ce paramètre *a*, un effet de pépite C_0 intègre les variations à petite échelle. Celui-ci peut être interprété comme un paramètre de lissage, le cas d'un effet de pépite pur correspondant à la

régression ou lissage infini alors qu'un effet de pépite nul n'induit aucun effet de lissage (Marcotte and David 1988). Il peut également être considéré comme proxy pour le degré d'incertitude des données. Enfin, un troisième paramètre est la variance de la variable aléatoire $Var[Z(x)]$ ou palier. Plusieurs exemple de fonction de covariance sont décrits et testés dans les articles du chapitre suivant et dont les expressions sont également disponibles en annexe.

Pour améliorer les estimations par krigeage, il convient parfois d'utiliser des informations contenues dans d'autres variables que la principale, ou covariables. L'estimation simultanée de variables principale et secondaires est appelé cokrigeage (Wackernagel 2003). Le nombre de variables secondaires ne dépasse en général pas deux et le cokrigeage ne requiert pas nécessairement l'observation de la variable principale, leur ordre étant interchangeable. Soit $Y_j(x)$ des variables secondaires, l'estimateur de cokrigeage prend la forme suivante (12) :

$$\hat{Z}(x_0)=\sum_{i=1}^{N_Z}\lambda_i Z(x_i)+\sum_{i=1}^{N_{Y1}}\alpha_i^1 Y_1(x_i)+\ldots+\sum_{i=1}^{N_{Yp}}\alpha_i^p Y_p(x_i) \quad (12)$$

ou N_z, N_{yj} et p sont les nombres d'observations des variables Z et Y et le nombre de variables Y, respectivement. Les contraintes sur les poids du cokrigeage λ_i et α_i pour un estimateur sans biais sont les suivantes (13) :

$$\sum_{i=1}^{N_Z}\lambda_i=1 \qquad \sum_{i=1}^{N_{Yj}}\alpha_i^j=0 \quad \forall \quad j=1,\ldots,p \quad (13)$$

Le cokrigeage admet ainsi des poids négatifs. La même démarche que pour le krigeage ordinaire permet ensuite de calculer ces poids. Afin de répondre aux caractéristiques des processus industriels plutôt que géologiques et pour une meilleure prise en compte des incertitudes, les modifications apportées au krigeage et au cokrigeage sont présentées dans les deux sections ci-dessous.

4.2.1 Composantes déterministe et aléatoire

L'hypothèse de stationnarité des données est souvent adoptée pour représenter des phénomènes naturels. L'exemple des teneurs dans un gisement est à l'origine même du krigeage qui s'étend de plus en plus à des phénomènes tels que les niveaux de contamination dans un sol ou les concentrations de polluants dans l'air (p. ex. Soares 2010). En revanche, les propriétés des

processus industriels changent en fonction de leur conception et de leur production. Si leurs caractéristiques sont utilisées comme références pour estimer les quantités d'énergie et de matériaux qui traversent leurs frontières, les chances de respecter cette hypothèse s'amenuisent. Une adaptation du modèle de krigeage est possible et nécessaire dans ce cas et le résultat souvent dénommé krigeage universel. Le principe consiste à ajouter une dérive $m(x)$ sur la moyenne des observations (14) :

$$m(x)=\sum_{l=0}^{L} a_l x^l \quad L=0,1,2 \quad (14)$$

ou a_l sont des coefficients évalués en fonction des données, par exemple grâce à une régression, et x, la ou les « dimensions » caractéristiques d'un processus. Lorsque $L = 0$ dans l'équation (14), la moyenne des observations reste constante, lorsque $L = 1,2$, $m(x)$ est un polynôme de degré un ou deux. Ici, une dérive linéaire est intégrée dans le modèle de krigeage alors que d'autres possibilités existent pour représenter le comportement de la moyenne de observations, qu'elle soit connue ou non (Joseph 2006).

Le modèle de krigeage peut ainsi s'écrire de la manière suivante (15) :

$$Z(x)=m(x)+Y(x)=\sum_{l=0}^{L} a_l x^l+Y(x) \quad (15)$$

ou $Z(x)$ sont le ou les flux d'énergie et de matériaux que l'on cherche à estimer et $Y(x)$ la composante aléatoire du modèle. Cette combinaison déterministe et aléatoire apporte non seulement plus de flexibilité au krigeage mais permet aussi de représenter des processus dont les propriétés sont a priori très différentes des phénomènes naturels.

4.2.2 Effet de pépite variable

Parmi les objectifs de ce projet de recherche, la prise en compte des incertitudes est un élément essentiel pour la fiabilité des estimations par krigeage et leur comparaison avec d'autres méthodes, pourvu qu'elles aussi puissent fournir une évaluation des incertitudes associées à leurs résultats. En ACV, les données d'inventaire provenant de bases de données sont généralement distribuées selon une fonction de probabilités log-normale afin d'éviter toutes valeurs négatives puisqu'il s'agit de flux physiques (Frischknecht et al. 2005). L'erreur est uniforme sur l'ensemble

d'un échantillon et représentée par un écart type malgré le fait qu'une seule valeur soit souvent observée. De la même manière, l'effet de pépite est uniforme pour une fonction de covariance $C(h)$ donnée. Par exemple, pour une covariance exponentielle en fonction de la distance h (16) :

$$C(h)=\begin{cases} C\left[\exp(-3h/a)\right] & if \quad 0\le|h|\le a \\ 0 & if \quad |h|>a \end{cases} \quad (16)$$

ou la variance σ^2 est la somme de C et C_0 avec $C_0=1/2\,Var\left[Z(x)-Z(x+\epsilon)\right]$ l'effet de pépite. Comme précédemment, la portée effective, ou distance de corrélation, de la fonction de covariance est a.

Avec de multiples sources de données dont la variabilité est particulièrement importante, un degré d'incertitude uniforme est futile. Pour répondre aux objectifs concernant l'incertitude des données, une seconde modification est apportée au modèle de krigeage. En lieu et place d'un effet de pépite constant, un effet variable permet d'assigner une mesure d'incertitude propre à chaque observation. Soit C_0 l'incertitude des données observées, la variance σ^2 peut alors être associée à C (Cressie 1993). Le modèle de krigeage est modifiée comme suite (17) :

$$Z(x)=m(x)+Y(x)+e(x)=Y'(x)+e(x) \quad (17)$$

ou $e(x)$ est le terme d'erreur avec variance C_0. C'est $Y'(x_0)$ qui est estimé en tout point x_0, avec variance C, plutôt que $Z(x_0)$. Une valeur différente de C_0 est assignée à chacune des observations. La formulation en (16) de la fonction de covariance $C(h)$ est exempte d'effet de pépite, un effet variable peut donc être ajouté lors de la résolution du système d'équation du krigeage. Sous forme matricielle, soit K_0 la matrice des covariances $C(h)$ et C_0 le vecteur des erreurs ou facteurs d'incertitude, les équations du krigeage ordinaire se réécrivent de la façon suivante (18) :

$$\left[\underbrace{\begin{pmatrix} C & C(h_{12}) & \cdots & C(h_{1N}) & 1 \\ C(h_{21}) & C & \cdots & C(h_{2N}) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C(h_{N1}) & C(h_{N2}) & \cdots & C & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}}_{K_0} + \underbrace{\begin{pmatrix} C_{01} & 0 & \cdots & 0 & 0 \\ 0 & C_{02} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & C_{0N} & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}}_{diag(C_0)}\right] \underbrace{\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \mu \end{pmatrix}}_{\lambda} = \underbrace{\begin{pmatrix} C(h_{10}) \\ C(h_{20}) \\ \vdots \\ C(h_{N0}) \\ 1 \end{pmatrix}}_{k_o} \quad (18)$$

ou diag(C_0) est le vecteur C_0 sous forme diagonale, λ_i les poids du krigeage et k_0 les valeurs de la fonction de covariance $C(h)$ au point à estimer x_0. En ajoutant le vecteur C_0 à la diagonale de K_0

les propriétés de celle-ci, définie positive, restent inchangées. Le principe est le même lorsqu'une dérive est ajoutée à la moyenne des observations, comme dans l'équation (14). Cette modification est justifiée lorsqu'il existe des indices fiables de la variabilité des données, de leur incertitude, ou des deux simultanément. L'amélioration est visible en particulier dans le cas des données relatives aux émissions des barrages hydroélectriques, comme présenté dans l'article *Handling data variability and uncertainty with kriging: the case of hydroelectric reservoir emissions*.

4.2.3 Validation du modèle

Les échantillons décrits ci-dessus n'ont pas la taille disponible ailleurs, par exemple dans des travaux similaires mais avec d'autres estimateurs statistiques (Wernet et al. 2008). C'est une des raisons justifiant le choix de fonctions de covariance préétablies. C'est aussi une raison pour laquelle le modèle est testé avec une technique de validation croisée (Marcotte 1995). En général, peu de données sont disponibles pour établir un inventaire du cycle de vie et encore moins pour valider les résultats de manière fiable avec des échantillons distincts. La validation est donc absente de la plupart des ACV, malgré elle (Ciroth 2006). La validation croisée ne nécessite pas d'hypothèse particulière comme la normalité des données, et avec des échantillons de petites tailles, un cas spécial de *k-fold cross-validation*, soit *leave-one-out cross-validation* est applicable ici ($k = 1$). Chaque observation est à tour de rôle retirée puis estimée grâce à ses $N - 1$ voisines. Lorsque plusieurs variables sont estimées simultanément dans le cadre du cokrigeage, la validation peut se faire en retirant toutes les valeurs des variables d'une observation en même temps ou l'une après l'autre. Cela permet en outre de comparer les avantages et désavantages du cokrigeage.

La validation croisée permet également de calculer des écarts type pour les valeurs estimées par krigeage ou cokrigeage et leur comparaison avec des valeurs estimées notamment par régression linéaire. Les données étant d'échelles très différentes, il est préférable de normaliser chacune des variables caractéristiques et le cas échéant les flux élémentaires et économiques, par leur moyenne. Une fois les données normalisées, un autre critère de comparaison, les erreurs moyennes absolues (MAEs), peut aussi être calculé pour le krigeage ou cokrigeage en distinguant chaque fonction de covariance, comme dans les articles qui suivent. Par contre, l'évaluation des écarts type du krigeage nécessite un ajustement lorsque la fonction de

covariance n'est pas estimée à partir des données. En effet les modèles de covariance sont sélectionnés parmi une série de fonctions valides comme décrit ci-dessus. Ainsi, suite à la validation croisée, la moyenne des erreurs normalisées par les écarts type de krigeage, au carré, devrait être égale à un. En partant des erreurs de l'équation (3), le facteur *b* est calculé de la façon suivante (19):

$$E[\frac{(Z(x_i)-\hat{Z}(x_i))^2}{(C_{0i}+\sigma_{ki}^2)}]=b \quad i=1,\ldots,N \quad (19)$$

Ainsi $E[(Z(x_i)-\hat{Z}(x_i))^2/b(C_{0i}+\sigma_{ki}^2)]=1$. En d'autres termes, une fois divisées par le facteur *b*, les erreurs normalisées ont une variance de 1. Les mêmes estimées $\hat{Z}(x_i)$ sont obtenues mais les variances de krigeage sont multipliées par *b* et les écarts type ajustés aux données. Cela permet de comparer le krigeage avec d'autres estimateurs grâce à un critère représentant souvent l'incertitude et/ou la variabilité des valeurs estimées. Ces opérations sont principalement réalisées grâce au logiciel Matlab et font appel à la fonction COKRI, écrite par Denis Marcotte et augmentée depuis (1991). Elle est modifiée à des fins de comparaison avec d'autres scripts nécessaires notamment pour les fonctionnalités qui permettent d'intégrer les incertitudes sous forme d'effet de pépite variable.

Chapitre 5 PRÉSENTATION DES ARTICLES

Comme indiqué dans la description des objectifs, les articles suivants s'inscrivent dans une suite logique qui tente de vérifier l'hypothèse de recherche. Chacun d'entre eux expérimente avec un échantillon différent mais dont les propriétés sont identiques, à savoir une série de variables caractéristiques et les flux élémentaires et économiques correspondants. Certains flux ont jusqu'à 30% de données manquantes. La récolte de données sur les émissions des réservoirs a également permis d'inclure les écarts type pour toutes les valeurs observées. Seules les données du second article sur les installations hydroélectriques sont en partie confidentielles. Les autres échantillons sont repris intégralement en annexe. Voici un résumé des résultats obtenus.

5.1 Principaux résultats

5.1.1 Évaluation du krigeage et de ses paramètres en ICV

L'estimation de données d'inventaire en ACV par krigeage présente des avantages non négligeables. Tout d'abord il s'agit d'un estimateur exact, c'est à dire qu'aucune erreur n'est commise lorsque des valeurs observées sont estimées par le modèle. L'effet de lissage implique également que le krigeage est discontinu aux points observés. Ensuite les résultats montrent une meilleure adéquation des estimées par krigeage que par régression linéaire.

a) L'enveloppe des écarts type (68%) de krigeage contient les données de manière plus exacte que la régression. De la même façon, pour une estimation donnée, les écarts type de krigeage sont plus faibles que ceux d'une régression linéaire. Tel n'est pas nécessairement le cas pour le cokrigeage ou certains flux, covariables, sont mieux corrélés que d'autres. Lorsque des flux fortement corrélés sont cokrigés, leurs écarts type sont encore plus faibles que pour le krigeage.

b) L'évaluation des modèles de covariance et de leurs paramètres grâce à la moyenne des erreurs absolues (MAEs), présente dans ce cas de légères différences entre les modèles

cubique, sphérique et linéaire pour un effet de pépite faible. Lorsque la valeur de celui-ci augmente, les trajectoires de chacun des modèles sont sensiblement pareilles. Le choix des autres paramètres du krigeage peut être testé de la même manière et/ou s'avérer relativement arbitraire. La portée est définie de façon à ce que les covariances soient non nulles et toutes les données incluses. Ces résultats sont obtenus par validation croisée, ce qui permet également de comparer la moyenne des erreurs absolues pour le cokrigeage lorsqu'une seule valeur, ou toute une observation, est estimée à tour de rôle. Il ressort que le cokrigeage est plus précis lors de l'estimation d'une donnée manquante en utilisant des variables secondaires dont les données sont disponibles pour cette observation là.

5.1.2 Application du krigeage à l'inventaire de centrales hydroélectriques

Une différence importante ici est l'absence de données complètes pour la moitié des flux considérés, les gros contributeurs notamment durant la phase de construction des centrales hydroélectriques. Il est ainsi possible de comparer les résultats du krigeage avec des données manquantes. Un échantillon plus large permet également de tester un autre aspect de l'estimation par krigeage. S'il n'existe pas de points dont les coordonnées sont des dimensions spatiales en analyse d'inventaire, les variables caractéristiques n'en définissent pas moins un espace. La question est de savoir si plusieurs dimensions valent mieux qu'une, et si oui lesquelles. Trois dimensions sont étudiées, la production annuelle, la capacité installée et la hauteur de chute entre prise d'eau et turbine des centrales hydroélectriques. Le choix des modèles de covariance et l'estimation de leurs paramètres font appel aux résultats précédents et c'est une fonction de covariance linéaire qui est des plus appropriée selon les mêmes critères de moyennes d'erreurs absolues. Les principaux résultats sont les suivants :

a) Que ce soit dans le cas du krigeage ou du cokrigeage, les erreurs sont plus faibles lorsque les variables caractéristiques utilisées pour l'estimation des données manquantes sont positivement corrélées. Si elles le sont moins, comme c'est le cas de la production annuelle et de la hauteur de chute, les différences entre les données complètes et incomplètes sont moins marquées.

b) Les erreurs moyennes absolues du krigeage sont invariablement plus faibles que celles qui proviennent d'une estimation par régression linéaire. Le tableau est moins contrasté pour

le cokrigeage. En effet le cokrigeage n'apporte pratiquement aucun avantage sur le krigeage en présence de données manquantes. A l'inverse, lorsque les données sont complètes, et comme précédemment les flux élémentaires ou économiques sont très corrélés, les erreurs d'estimation du cokrigeage sont beaucoup plus faibles que pour les autres estimateurs, krigeage et régression linéaire.

c) Une comparaison des erreurs entre les versions unidimensionnelles et bidimensionnelles du krigeage et cokrigeage montre que deux variables caractéristiques sont meilleures qu'une seule lorsque celles-ci expliquent aussi bien l'une que l'autre le phénomène observé et qu'une partie des données manque. A nouveau, le cas du cokrigeage n'est pas univoque, mais quel que soit le nombre de variables caractéristiques, ses erreurs sont plus faibles que celles du krigeage pour l'estimation avec des données complètes. Par contre les erreurs d'estimation du krigeage avec des données complètes sont plus faibles dans le cas unidimensionnel que bidimensionnel.

Grâce à cette flexibilité, le krigeage, unidimensionnel notamment, se montre clairement plus adapté à l'estimation de données manquantes, en particulier lorsque les observations sont partielles et incomplètes. A l’évidence, ces résultats plaident en faveur de l’hypothèse de recherche.

5.1.3 Prise en compte des incertitudes observées et estimées

Le troisième article apporte des nouvelles modifications à l'approche classique par krigeage, celles-ci répondent directement à la prise en compte des incertitudes de données très variables. C'est le cas par exemple des émissions de gaz à effet de serre (GES) mesurées sur une série de réservoirs hydroélectriques qui ne présentent que peu, voire aucune, structure. En faisant varier un paramètre des fonctions de covariance selon le degré d'incertitude des données, il est non seulement possible de prendre en compte les incertitudes observées mais également de calculer les incertitudes correspondantes pour les valeurs estimées. Les résultats qui découlent de ces modifications se basent en partie sur ceux obtenus précédemment.

a) Contrairement aux estimations de flux économiques, les émissions de GES provenant des réservoirs hydroélectriques ne sont pas tant corrélées avec les variables caractéristiques telles que la puissance théorique d'une centrale et la surface de son réservoir. A priori,

inclure une composante déterministe, de type dérive linéaire sur la moyenne des observations, n'ajoute rien au modèle de krigeage. Les résultats montrent qu'effectivement une dérive linéaire n'améliore pas les estimations dans ce cas, au contraire.

b) La version modifiée du krigeage, qui inclut un effet de pépite ou facteur d'incertitude variable, donne des résultats cohérents non seulement avec les données observées mais également avec les prévisions quant aux émissions dues aux réservoirs sous les latitudes boréales ou tropicales. Si les écarts type du krigeage sont grands, il est néanmoins essentiel de constater que ceux-ci sont invariablement réduits par rapport au krigeage ordinaire avec effet de pépite constant, et ce pour les fonctions de covariance linéaire et cubique en particulier, indépendamment des gaz à effet de serre représentés. Grâce à l'introduction d'un effet de pépite variable, la variation des données qui serait autrement écartée peut être prise en compte de façon relativement simple.

c) A noter que les écarts type sont beaucoup plus faibles lors de l'estimation des émissions de GES pour un réservoir au nord du 50e parallèle qu'à l'équateur. Avec plus de données il y aurait donc possibilité de diviser l'échantillon de façon à prendre en compte les variations régionales ou selon le milieu.

Les trois articles présentés ci-dessous et leurs résultats respectifs sont interdépendants dans la mesure ou les expériences de l'un servent à développer le suivant. L'ordre des objectifs de recherche est respecté et permet de renforcer la complémentarité des résultats.

5.2 Estimating material and energy flows in life cycle inventory with statistical models

Moreau, V., G. Bage, D. Marcotte, R. Samson, submitted July 2010, *Journal of Industrial Ecology*, accepted August 2011 (Moreau et al. 2012)

5.2.1 Abstract

Data (un)availability and uncertainty are recurring problems in life cycle assessment and, particularly, inventory analysis. Advances in life cycle inventory have focused on the propagation and management of uncertainty, while this article addresses the question how to account for unavailable data and corresponding uncertainty. Large and complicated systems often lack complete data, due to confidential practices or the efforts required in the data collection process. Electricity production with multiple processes generating a single product is a classic example. Instead of the conventional process based models to estimate missing data, the approach developed in this article divides systems based on functionally equivalent objects. Each one of these objects is then described in terms of characteristic variables, such as power capacity. Kriging, a flexible statistical estimator, allows for the estimation of unknown material and energy flows based on the objects' characteristic variables. Both univariate and multivariate kriging are tested and compared to regression analysis. It is found that kriging performs better than linear regression, according to the mean absolute error criterion. Multivariate kriging provides an even more accurate joint estimation method to bridge data gaps scattered across inventories and when observable values of material and energy flows differ from one object to the next. Parameters of the underlying models are interpreted in terms of data uncertainty.

Keywords: Industrial ecology, missing data, statistical estimation, kriging, energy, methods

5.2.2 Introduction

Life cycle assessment (LCA) provides one comprehensive method of quantitative analysis for products, processes, and services. As with any decision making support method, its reliability is essential in order to set priorities. Should the uncertainty in the results be greater than the impact differences between products, services, or even policies, the wrong decisions could be

made. When assessing a system in itself, data uncertainty is equally problematic as the environment sets absolute, not relative, limits to economic and technological development.

The quality of an LCA largely depends on the quality of the data collected during the inventory phase of the assessment (Mongelli, Suh, and Huppes 2005). Life cycle inventory (LCI) is data intensive by definition, quantifying all of the material and energy flows crossing a given system's boundaries (ISO 2006). Thus, compiling an accurate inventory is a predeterminant of the overall assessment's quality. Data collection from primary sources is often problematic for reasons of confidentiality, product system size, or lack of knowledge (Pehnt 2003).

Almost all technical systems and processes involve some form of energy, and most industrial systems use electricity (Curran, Mann, and Norris 2005). In addition, electricity production and consumption generate a large share of the impacts for many processes, with relatively high uncertainty surrounding the embodied materials and energy of a given kWh (Matsuno and Betz 2000; Weber et al. 2010). The necessity to develop reliable inventories for electricity generation comes not only from the extensive use of power but also for the support of more accurate, future applications of LCA (de Haes and Heijungs 2007). For instance, assessing the impact of an electricity generation network represents a massive inventory that tends to yield high uncertainty if compiled from generic data (May and Brennan 2003). Differences in the quality of inventories have been highlighted for specific processes by, for example comparing data sources (Peereboom et al. 1998; Miller and Theis 2006). The inherent data gaps were identified as an important source of uncertainty by Huijbregts and colleagues (2001).

Hence, the purpose of this article is to show how one methodological approach can turn small samples from large systems into more complete and accurate inventories. In other words, this article aims to provide answers to the question of how to bridge data gaps scattered across inventories. Unlike the common process-based models, statistical estimators are experimented with beyond the usual linear interpolation and extrapolation techniques. In theory, one technique, known as kriging, has great potential in estimating life cycle inventory data. In practice, no consensus exists within the LCA community on the statistical estimation of inventory data (Sugiyama et al. 2005). So far, ignorance and substitution for missing data have been common methods for dealing with the issue (Ciroth, Fleischer, and Steinbach 2004). Most of the research

work in LCA involving statistical methods has focused on the propagation of uncertainties from inventory analysis to impact assessment (Lloyd and Ries 2007). A few studies have applied statistical techniques in cases of missing inventory data and associated uncertainty, for example with chemical compounds, where thousands have been produced but only a fraction analyzed (Dahllof and Steen 2007; Wernet et al. 2008).

In the methodological approach presented below, the key is to recognize that certain characteristic variables describing the objects of a system are more readily available than corresponding material and energy flows. Technical specifications are then used as independent variables to estimate the missing data. If the approach is aimed at inventory analysis from primary data sources, LCI databases also apply, especially where data availability is limited. Indeed, sample sizes are expected to remain small when estimating missing data. In this article, the data come from the ecoinvent database and consist of five records for wind turbines of different characteristics (ecoinvent Centre 2009). After a description of the methodology, results and limitations are presented and discussed. A conclusion, including several leads to expand this approach, ends this article.

5.2.3 Method

Data uncertainty continues to challenge life cycle assessment in many ways (Reap et al. 2008; Williams, Weber, and Hawkins 2009). In the analysis of systems involving either large populations or high levels of integration, or both, data scarcity becomes a critical issue. Even if all the data were accessible, collecting it would take time and energy beyond reasonable constraints. The lack of data is a problem faced by most practitioners of LCA as well as statisticians. As the missing information principle states: a simple problem at first requires more complex analysis when information is missing (Orchard and Woodbury 1972). In addition, with only small samples available, model validation becomes equally difficult to achieve. Statistician John Tukey coined the term 'uncomfortable science' for cases where not enough data can be collected for both training and testing data sets (Hoaglin, Mosteller, and Tukey 1985). This article introduces an estimator known as kriging, borrowed from spatial statistics, and uses cross validation to compare the different models.

Instead of spatial coordinates, characteristic variables replace one or more dimensions. That is, a process is decomposed into groups of objects that can be characterized by similar properties or technical specifications. For instance, power plants vary greatly, even when based on the same technology. Yet the embodied materials and energy of their building blocks, such as turbines, generators, and transformers, correlate positively with the power or current they are designed to generate or transfer. The power rating is one such characteristic or independent variable, according to which objects can be clustered. Material and energy flows, as well as emissions, are then dependent variables to be estimated. In LCA, materials, energy, and emissions are often referred to as economic and elementary flows, with the environment as a source and sink for elementary flows and the economic system as source and sink for economic flows. Analogies can be drawn between this approach, using functional groups of objects, and the functional unit in LCA, which uses reference flows as characteristic variables. Similarly, spatial statistics have the concept of regionalized variables, with not entirely random spatial patterns (Chiles and Delfiner 1999).

Under specific conditions, kriging is equivalent to neural networks previously used in LCI (Wernet et al. 2008); but kriging was developed as a method of interpolation for random processes in space, such as ore content in mines (Cressie 1990). The stakes are neither as high in LCA as in mining exploration, nor are the material and energy flows, or the emissions in LCI, spatially distributed. Nevertheless, in both cases data availability is very limited by such factors as accessibility and cost. The flexibility of kriging, however, should become obvious; it proves superior to other estimators precisely in cases of uncomfortable science and missing data (Gunes, Sirisup, and Karniadakis 2006). Kriging is often referred to as a minimum variance unbiased estimator (MVUE). And similar to least square estimators, kriging is a weighted linear combination of observations. With weights λ_i and N observations, kriging takes the following form:

$$\left\{estimated\ \ flow\,at\,characteristic\ variable_0\right\} = \sum_{i=1}^{N} \lambda_i \left\{flow\,observed\,at\,characteristic\,variable_i\right\}$$

A covariance function underlies every estimate by describing how the observations relate to one another. This function essentially translates into mathematical terms the notion that observations at distinct points close to one another should differ less than if they were far apart. Among the valid covariance functions, four common examples are presented in table 3. They are functions of the distance *h,* which is the difference between the values of the characteristic variables at each observation (power capacity in this case). Covariance functions usually have 3 parameters: a nugget effect (C_0) capturing variations on a small scale; a structured variance (C); and a range (a). The total variance is $C_0 + C$ and the range corresponds to the value of h for which the covariance is nil, when two observations are too far apart to be related.

Table 3: Types of covariance functions (adapted from Chiles and Delfiner 1999)

	Type	Expression
I	Nugget	$C(h)=\begin{cases} C_0 \;\; if \;\; h=0 \\ 0 \;\; if \;\; \lvert h\rvert>0 \end{cases}$
II	Linear	$C(h)=C[1-(h/a)]$
III	Spherical	$C(h)=\begin{cases} C\left[1-1.5(h/a)+0.5(h/a)^3\right] \;\; if \;\; 0\le\lvert h\rvert\le a \\ 0 \;\; if \;\; \lvert h\rvert>a \end{cases}$
IV	Cubic	$C(h)=\begin{cases} C\left[1-7(h/a)^2+8.75(h/a)^3-3.5(h/a)^5+0.75(h/a)^7\right] \;\; if \;\; 0\le\lvert h\rvert\le a \\ 0 \;\; if \;\; \lvert h\rvert>a \end{cases}$

When a nugget effect is present, the kriging estimator is discontinuous at data points, such that:

$$\lim_{\varepsilon\to 0}\hat{Z}(x_i+\varepsilon)\neq\hat{Z}(x_i) \quad (1)$$

where $\hat{Z}(x_i)$ is the estimate at x_i. The nugget effect can therefore be interpreted as an interpolating smoothing parameter (Marcotte and David 1988). The case of a pure nugget effect corresponds to regression or infinite smoothing, whereas zero nugget means no smoothing. Note that, in addition to its unbiased property, kriging is also an exact estimator, meaning estimates at data points equal observed values (2).

$$\hat{Z}(x_i) = Z(x_i) \quad \forall \quad i = 1, \ldots, N \quad (2)$$

where $Z(x_i)$ the observed value at x_i. In other words, no information is averaged out. One of the main distinctions between kriging and other estimators, such as regression, is the minimization of the error variance. Under the constraint of unbiasedness, which essentially states that the sum of the kriging weights (λ_i) should be equal to one, this optimization problem is solved by forming the Lagrangian (Isaaks and Srivastava 1989). See the supporting information in the annexes for details on how kriging is implemented in any mathematical software package.

Multivariate kriging, or so-called cokriging, estimates the values of a primary variable by using information from primary and secondary variables simultaneously (Wackernagel 2003). For example electricity generation from wind or water turbines rarely depends on installed capacity entirely. Other characteristic variables such as wind speed or water flow also influence production. Similarly, embodied material and energy flows behave in equivalent ways, but the availability (or unavailability) of observed values often changes from one flow to another. Thus, joint estimation derives more reliable estimates from incomplete data. For instance, the content of electrical conductors in a series of generators is well known, but only partial information is available on the amount of insulating material. Cokriging of conductor and insulator values together improves the reliability of estimates for both materials. This is extremely convenient, as material and energy flows often vary in equal proportions, and so cokriging further reduces uncertainties associated with missing data. Note that the role of primary and secondary variables can be reversed, and all estimations done jointly (Marcotte 1991). Details with respect to cokriging are available in the supporting information.

Life cycle inventory data are scarce. Even databases offer hardly more than one record for certain processes. Limited data sets are expected whenever estimating missing data in LCI. Thus, validation of a model and its parameter values should be done accordingly. Covariance functions and their parameters can be determined in several ways (e.g. maximum likelihood estimators) (Cressie 1993). To avoid assumptions, like normality of the data, and given small data sets, cross-validation as described below is preferred to other validation methods (Marcotte 1995).

Several types of cross-validation exist, all of which involve training and validation data sets. In leave-one-out cross-validation (LOOCV), successive observations are removed and

estimated from their neighbors exactly once. In other words, each observation forms the validation data once and part of the training data set ($N - 1$) times. Furthermore, with cokriging, the values of material and energy flows for a given observation can be part of the validation data at once or one after the other. These two possibilities are referred to as leave-one-observation-out cross-validation (LOoOCV) and leave-one-variable-out cross-validation (LOvOCV), respectively. Differences between the two validation procedures highlight, relatively simply, the usefulness of cokriging – that is, the contribution of values observed for one flow when estimating values for another. This is equivalent to having missing data scattered across an inventory of material and energy flows.

5.2.4 Results

The results were obtained by programming a mathematical software package (Matlab) for the different kriging estimators and cross validation procedures. Data, specifically the five records for wind turbines, came from the ecoinvent LCI database (Burger and Bauer 2007). Four of the turbines are located in Switzerland with 30 to 800 kW of capacity. The fifth turbine is a typical 2 MW offshore turbine in Denmark. The choice of data is motivated by the increasing number of wind turbines in operation and the limited number of documented cases in LCA. Data collection is a common problem of both large sources of electrical power (>30 MW) that require massive inventories and small and emerging technologies (<30 MW) that are growing rapidly. The economic flows of particular interest are materials, energy carriers, and fuels that have undergone some form of processing. These were selected for two functional groups: the static (tower) and the mobile (rotor) components of wind turbines. Although the methodology is intended for primary data sources, it applies equally well to LCI databases, and could provide more concise information for fitting models to their content. No information would be lost or averaged out, as kriging is an exact estimator.

All estimates are given with respect to one characteristic variable: power capacity. In other words, material and energy flows are kriged based on kilowatts. More importantly, they can be jointly estimated or cokriged. Because of different scales, and in order to compare estimates with mean absolute errors (MAEs), the data were normalized and the axes scaled relative to the mean. Figures 1 and 2 display cases of kriging for one flow from each of the functional groups

(concrete and copper) and varying power capacity. The results illustrate a spherical model of covariance and linear drift on the mean, as well as the regression line, which is the common form of extrapolation among LCA practitioners using limited data (see the annexes for parameter values). Also shown are the envelopes of standard deviations for kriging (68%), which shrinks towards data points because the estimator is exact, and for regression, which spreads much more.

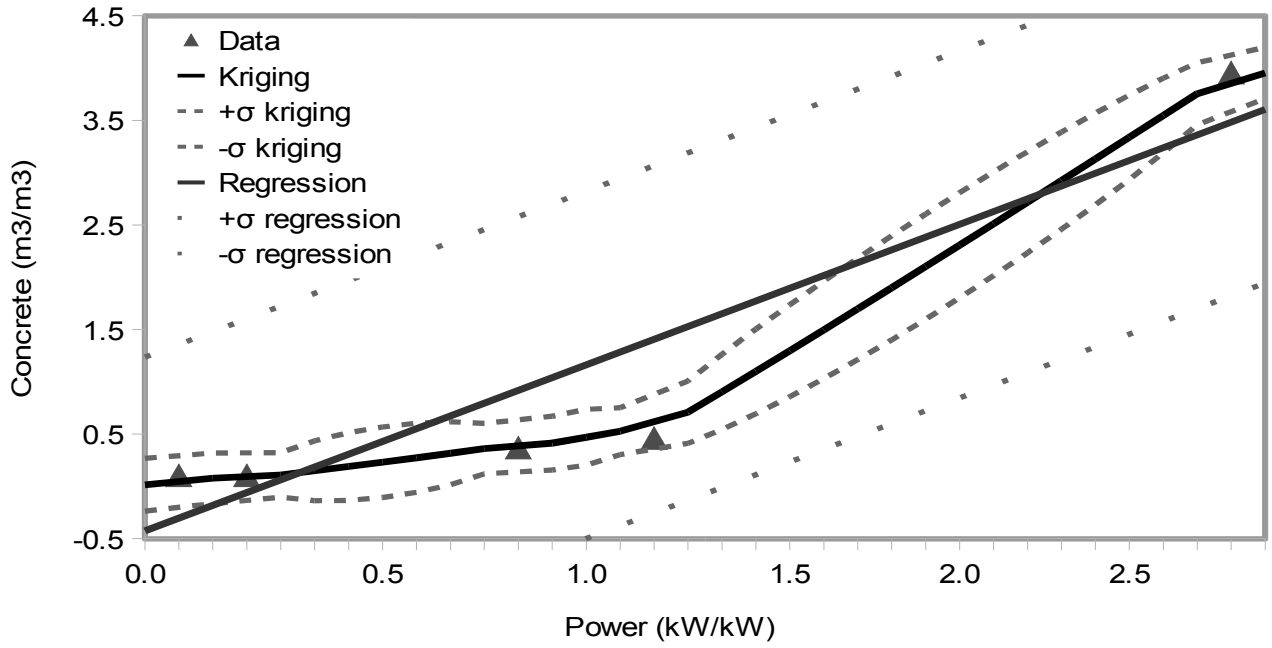


Figure 1: Kriging the volume of concrete necessary for the tower of wind turbines with respect to power capacity. For comparison purposes the linear regression is also provided and one standard deviation for both estimators.

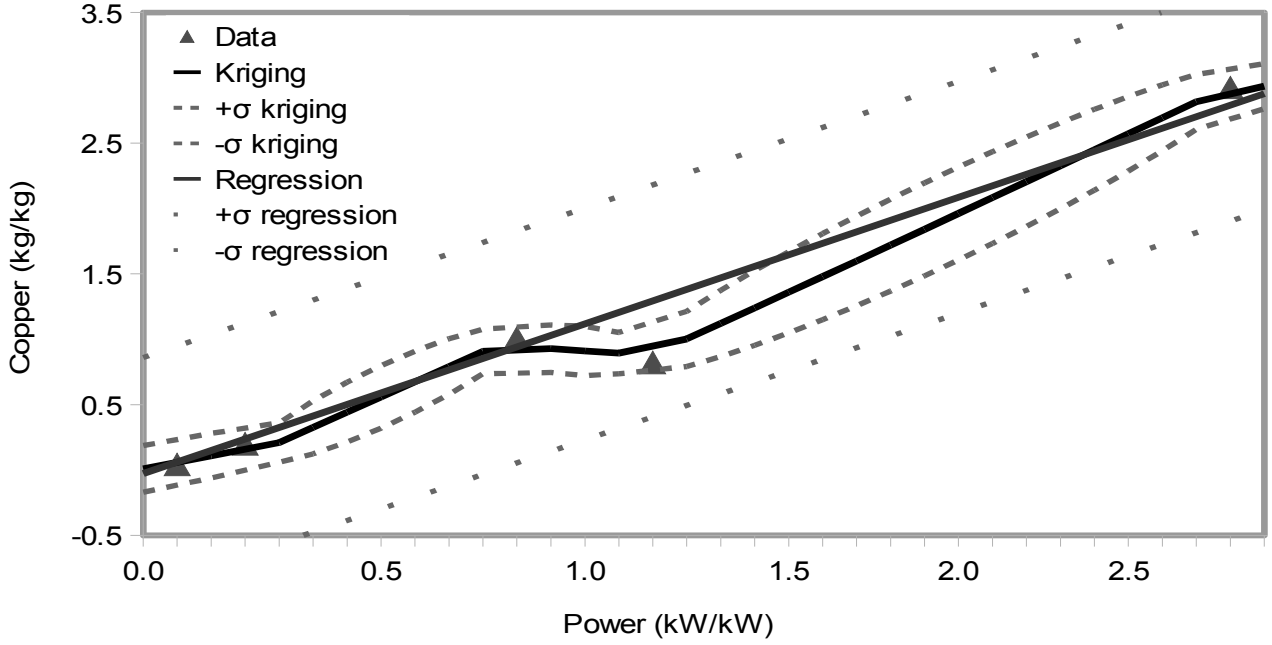


Figure 2: Kriging the amount of copper necessary for the rotor of wind turbines with respect to power capacity. For comparison purposes the linear regression is also provided and one standard deviation for both estimators.

Note that kriging is discontinuous at data points, yet provides a close fit. One benefit of kriging is the improvement of estimates even when very little data are available. In other words, the spherical model of covariance chosen in figures 1 and 2 is an educated guess; but it allows such estimates where a larger sample size, generally 30 or more observations, is needed to derive an ad hoc covariance function. A more formal selection procedure for covariance functions is described below. This approach can be expanded to include many more economic and elementary flows or characteristic variables for any type of object or system. Moreover, this approach applies to data for new or previously undocumented objects, as well as data missing at random from existing inventories. This is particularly the case with cokriging. To illustrate the former, consider a hypothetical 1.5 MW wind turbine. Table 4 shows the results of estimating material and energy flows with different estimators and their standard deviation. For potentially new and existing wind turbines, kriging thus provides data that are less variable than regression, for example.

Table 4: Example of kriging in the case of a hypothetical 1.5 MW wind turbine

Economic Flows	Kriging	Cokriging (pairwise estimation)		Regression
Concrete (m3)	547 ± 110	545 ± 58	557 ± 240	568 ± 366
Transport (tower) (10^3 tkm)	103 ± 21	103 ± 17		107 ± 65
Steel (tower) (tons)	152 ± 10		150 ± 46	150 ± 78
Welding (kg)	212 ± 29			207 ± 57
Epoxy (kg)	464 ± 125			440 ± 212
Electricity (MWh)	47 ± 5.5	47 ± 2.7	47 ± 18	48 ± 27
Aluminum (kg)	579 ± 73	578 ± 49		596 ± 333
Copper (tons)	1.5 ± 0.3		1.5 ± 0.5	1.5 ± 0.6
Chromium (tons)	36 ± 6.2			37 ± 20
Cast Iron (tons)	18 ± 2.7			19 ± 11
Fiberglass (tons)	28 ± 2.8			28 ± 17
Steel (rotor) (tons)	11 ± 1.1			11 ± 6.5

Columns 3 and 4 of table 4 are joint estimates of the volumes of concrete, with weight transported on one hand and the amount of steel necessary for the tower of the wind turbine on the other. Similarly, electricity requirements are cokriged with the amounts of aluminum and copper. Cokriging estimates vary less than regression, but only the results in column 3 are more accurate than univariate kriging. Thus the accuracy of the results depends on the relationship (correlation) between the material and energy flows estimated. The advantage of cokriging over kriging, if data are available for parts of an inventory and missing for others, is thus described in

the latter case above. For certain relationships between observed and unobserved variables, joint estimation improves the results.

In order to select covariance functions, the models in table 3 were compared based on mean absolute errors (MAEs). The errors were calculated from leave-one-variable-out cross validation (LOvOCV), and, for purposes of display, aggregated for three material and energy flows (electricity, aluminum, and copper). Figure 3 shows the results with respect to one parameter, the nugget effect, and on a scale relative to the mean of the observations. This parameter has no influence on regression, which is why the error remains constant there. The cubic model performs slightly better than the spherical one, with lower errors, especially for small values of the nugget effect. All three follow the same path of increasingly large errors with increasing values of nugget, which, as explained below, is a proxy data uncertainty. Clearly any choice of model and parameter value in figure 3 leads to more accurate results than linear regression.

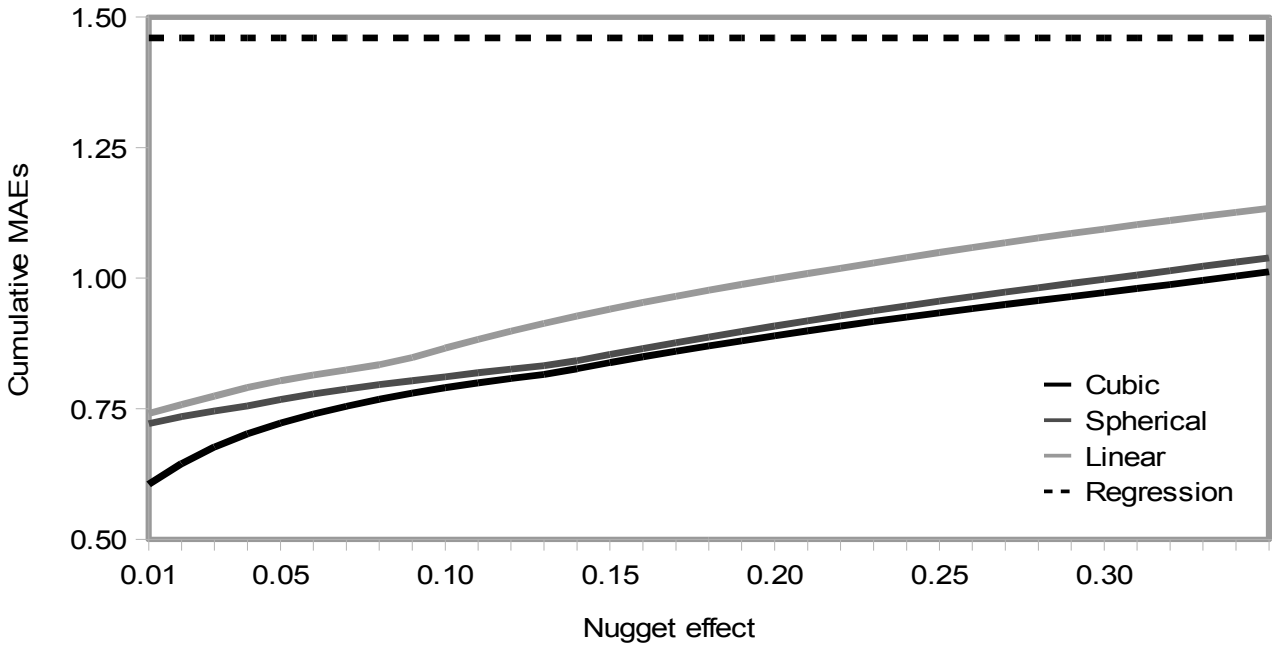


Figure 3: Mean absolute errors for cokriging as a function of the nugget parameter with different covariance models and for regression

The nugget effect measures variations on a small scale and can be interpreted as a factor of uncertainty for single observations or entire data sets. Various measures of uncertainty can be accounted for as long as they are consistent across observations. For example, standard deviation is a common measure of variability. Thus, comparison of covariance models with respect to changing nugget becomes relevant for two reasons. First, when dealing with missing data – which in itself is a source of uncertainty – the uncertainty assigned to the estimates should reflect that of the observed data. In other words, compounding effects or inconsistencies in the uncertainty of the estimated data should be avoided. Second, covariance functions have the convenient nugget parameter to account for data uncertainty in a relatively simple way. Analysis of data uncertainty is a highly recommended but seldom implemented practice in LCA (Ciroth, Fleischer, and Steinbach 2004). One of the main drawbacks remains the absence of appropriate accounting methods prior to propagating data uncertainty in the sequence of calculations from inventory analysis to impact assessment (Williams, Weber, and Hawkins 2009).

Assume the uncertainty of observed data translates into small values of nugget effects, say 0.01. With a cubic model of covariance, the mean absolute errors in the estimated data can be computed for cokriging in two ways. As explained in the methodological section, leave-one-out cross-validation with multivariate kriging takes out either single variables at a time or all of them at once for each observation, LOvOCV or LOoOCV respectively. The results in table 5 show the differences between these options as well as errors derived from linear regression. (See the supporting information in the annexes for parameter values other than nugget.)

Table 5: Comparison of mean absolute errors (N=5)

Economic flows	Cokriging LOvOCV	Cokriging LOoOCV	Regression (LOoOCV)
Electricity (kWh/kWh)	0.08	0.44	0.32
Aluminum (kg/kg)	0.09	0.47	0.42
Copper (kg/kg)	0.43	0.3	0.72

The comparison is straightforward in the first two rows of table 5. Estimating the value of a variable at a given point using the values of other variables from the same point clearly reduces mean absolute errors, as shown in the second column. This advantage of cokriging diminishes for variables like electricity and copper that have relationships without such high positive correlation. Nevertheless comparing the second and fourth columns of table 5 indicates a better estimation potential for cokriging over regression. Mean absolute errors are one among many criteria commonly used to evaluate discrepancies between estimates and actual values. MAEs are just as equally adequate as the other criteria, and more so in the case of normalized variables, where values are relative to the mean of the observed data. The level of confidence in the results increases more with a large number of estimates than it does by choice of criterion.

5.2.5 Discussion

The initial goal, to estimate data missing from or with limited availability in life cycle inventories using statistical estimators, led to a reliable approach. Indeed, kriging, as illustrated above, proves a superior method in cases of substantial or random data gaps, or both. Before some of the limitations are discussed, three main points should be emphasized. First, univariate and multivariate versions of kriging generally perform better than linear regression. Figures 1 and 2 as well as table 4 illustrate the better fit and more accurate estimates of kriging. Second, and most importantly, figure 3 and table 5 show how accuracy can be significantly improved with cokriging. MAEs derived from cross validation provide one indicator for selecting covariance functions and highlight the benefits of joint estimation for missing data. Third, the use of covariance functions and parameters, which have a direct interpretation as proxy to data uncertainty for example, imparts more flexibility to univariate and multivariate kriging even if it involves parameter evaluation.

A number of limitations exist for kriging, both independently of the type of data under consideration and with respect to the problem addressed in this article. In addition to minimum variance, one of the properties of kriging is such that its variance is generally less or equal to the variance of the observed data or underlying phenomenon being estimated. Mathematically this is expressed in equation (4) and means that small values are usually overestimated and large values underestimated.

$$Var(\hat{Z}(x_i)) \leq Var(Z(x_i)) \quad (4)$$

where *Var( )* is the variance and $\hat{Z}(x_i)$ and $Z(x_i)$ are the estimated and observed value at x_i respectively. This so called smoothing effect is found among other weighted average estimators and can be corrected for when necessary (Yamamoto 2005). Also, the choice of model and inclusion of a linear drift on the mean of the observations implies that, beyond the range of observed data, estimates are asymptotically linear. In other words, kriging converges towards linear regression as the estimated data get farther from the minimum and maximum of the observations.

Moreover, this approach may not apply to cases of missing data in every economic sector. It relies on the relationships between the characteristic variables of a given object and that object's embodied material and energy. For some objects, data unavailability may affect characteristic variables equally as much as material and energy flows. Electricity generation, however, is fertile ground for statistical developments, for two essential reasons. Unlike other products, once generated a kWh cannot be distinguished according to the processes involved. As a result, grid mixes are necessarily production averages, yet electricity sources vary constantly. The statistical approach presented in this article not only improves material and energy flow estimates, but it also saves on inventory data collection. With relatively small samples, kriging has the potential to cover large systems where cases of "uncomfortable science" may be the norm rather than the exception. Perhaps the two main hurdles to the application of kriging in LCI are the need for familiarity with the statistics, as well as the fact that some sources of uncertainty simply cannot be quantified.

5.2.6 Conclusions

Statistical estimation is one promising avenue for missing data in life cycle inventories and, if extension to economic and elementary flows is obvious, progress still needs to happen in characterizing functional groups. All estimates in this article were based on a single "dimension" or characteristic variable. A full procedure will most likely require the characterization of multidimensional objects, defined not only by their power capacity, but also, for example, their revolutions per minute. Even discrete dimensions, such as types of equipment or geographic location, may require characterization. Work also remains to be done in estimating the kriging

parameters, especially in the multivariate case and the description of how variables relate to one another. Numerous possibilities for improvement exist and statistical estimators like kriging could take a more prominent role in estimating LCI data and accounting for its uncertainty.

5.3 Statistical estimation of missing data in life cycle inventory: an application to hydroelectric power plants

Moreau, V., G. Bage, D. Marcotte, R. Samson, submitted August 2011, *The International Journal of Life Cycle Assessment*, rejected November 2011, revised and submitted February 2012, *Journal of Cleaner Production*. A shorter version was published as a book chapter and available in the annexes (Moreau et al. 2011).

5.3.1 Abstract

In evaluating the impacts of products and processes, life cycle assessment has taken a prominent role. Missing data, however, continue to affect the completeness and accuracy of such environmental assessment tools. Data are missing for many reasons, for specific processes, at random in existing inventories or worse, nonexistent. This article proposes a statistical approach to address the lack of data in life cycle inventories and applies it to hydroelectric power plants. Among large power production technologies, hydroelectricity varies considerably in scale and from one site to another. The procedure relies on relationships between the technical properties of a system or process, in this case hydropower plants, and the embodied material and energy flows in construction, operation and eventually dismantling. With highly flexible estimators known as kriging, predicting the value of material and energy flows becomes more accurate. From relatively small sample sizes, kriging allows better estimation without averaging out any of the observed data. Similarly, parameter estimation and model validation can be performed through cross validation which assumes very little on the data itself. Calculations of mean absolute errors for various forms of kriging and regression show that the former estimates the values of material and energy flows more accurately than the latter, more so in cases of incomplete data. Accounting for several technical characteristics as well as joint estimation of covariate material and energy flows provide different ways to further reduce the errors and improve the completeness of inventories where data are missing. The specificity of hydroelectric power makes existing inventory data largely unrepresentative and kriging offers new possibilities to increase the reliability of estimated, representative data.

Keywords: Life cycle inventory, missing data, statistical estimation, kriging, hydroelectricity

5.3.2 Introduction

Missing data in life cycle assessments (LCAs) are one of the main challenges to improving their quality and reliability (Reap et al. 2008; Weber et al. 2010). Life cycle inventory (LCI) is a particularly data intensive phase of LCA and was shown to be less accurate when compiled from generic data sources (May and Brennan 2003). Indeed, the quality of a life cycle inventory directly influences the overall quality of an LCA (Mongelli, Suh, and Huppes 2005). Several reasons explain the lack of data in LCI. Accessibility is limited because of confidential information on industrial processes or products (Wernet et al. 2009). The size and complexity of production systems and the novelty of a technology, largely prevent data collection (Pehnt 2003). The lack of knowledge regarding a product or process also translates into a lack of data, let alone the unknown unknowns. Consulting blueprints and bills of materials, if publicly available, goes beyond reasonable time and cost constraints for existing systems as large as electrical networks for example. Hence the goal of this paper, to show how statistical models improve the completeness of material and energy flow inventories by estimating missing data. A different terminology for the problem of missing data exist in LCA and in statistics. In the latter, a common assumption is that data sets are incomplete because of data missing at random (Tsiatis 2006). In LCI the lack of data is usually understood as either complete lack of data (data gaps) or lack of specific (representative) data (Huijbregts et al. 2001). The methodological approach introduces kriging, a class of minimum variance, unbiased estimators which can be formulated in different ways. Material and energy flows for processes described by similar characteristics are estimated from relatively small samples of inventory flows when data are unavailable.

Technological developments make assessments based on passed information relatively uncertain. Among the primary sources of energy, however, the best sites and highest quality resources have already been used up. Generating electricity from hydropower stations is a relevant example in many respects. Two main trends can be identified, first the continued construction of large hydro dams such as in the province of Quebec, Canada, at increasingly high costs. With two hydropower plants producing close to 1% of world hydroelectric output, Quebec is a quintessential example of large hydro. A second trend is the renewed development of pump back and storage facilities, particularly in Switzerland, Portugal and Austria (Deane, Gallachoir, and McKeogh 2010). Both strategies are problematic. In the first case flooded area per unit of

installed capacity increases and in the second, pump and storage plants are net consumers of electricity, turning “dirty” excess base load power (from nuclear or coal) into “clean” peak load hydroelectricity. In comparison to other large power generation facilities, hydropower plants vary considerably from one location to another and data cannot possibly be collected for every one of them (Egre and Milewski 2002). The performance of hydro dams and run-of-river plants depend on numerous technical conditions, the topography and hydrogeology, whereas the impacts are both social and environmental, challenging inventory analysis and impact assessments (Rosenberg, Bodaly, and Usher 1995). The model presented here extends the analysis to a wide range of power plants but remains essentially technical, leaving aside conflicts such as displacement of human populations and biodiversity loss. Generic hydropower plants hardly exist and current inventory databases are unrepresentative of large hydro, considering only a handful of alpine dams (Bauer et al. 2007). Thus, assessment of hydroelectric power mandates more accurate data than are available and the approach presented here attempts to estimate such missing data. In the methodological section below, details of the relevant properties of hydropower plants are provided, knowing that similar characteristics can be identified for other systems and processes. Results emphasize the flexibility of kriging and the potential to address both missing inventory data and lack of representative data. Finally the limitations of this approach and some recommendations are discussed.

5.3.3 Methodology

Life cycle inventory, as defined by international standards, requires the quantification of material and energy flows as well as emissions crossing a system's boundaries (ISO 2006). Among the current practices dealing with missing inventory data, ignorance remains widespread, setting values to zero (Ciroth, Fleischer, and Steinbach 2004). Substitution of missing data with non zero values is another common practice, assigning a degree of uncertainty which accounts for unrepresentativeness (Milà i Canals et al. 2011). Scaling data for a given process to the dimensions of another similar process is also possible provided a complete inventory exist (Caduff-Kinkel et al. 2008). Ignoring missing data has consequences on the reliability of the results and substitutions are limited by the availability and completeness of databases or literature sources. Several reasons lead us to hypothesize that statistical estimation can overcome these

limitations. Moreover the proposed estimator claims to be particularly suitable to data variability and site specificity for systems such as hydropower plants:

a) When ignored, missing data in LCI introduces a bias towards lower impacts and therefore favours the least documented processes or products. Also, the lack of (specific) data has been identified as an important source of uncertainty (Huijbregts et al. 2001).

b) Without generic production facilities, compiling the LCI of a kWh from given hydropower plants, or equivalently that of processes on similarly large scales, would require considerable data collection efforts. Even if data were technically available, access often runs into confidentiality issues.

c) LCI databases such as ecoinvent are not necessarily representative of major processes (ecoinvent Centre 2009). The data on hydroelectricity for example are scarce and refer to a few alpine dams, adjusting emissions for reservoir location, nothing on the scales of dams built in China, Brazil, Ethiopia or Canada. Moreover the size and scale of hydropower seldom translates into equivalently small or large impacts (Gleick 1992).

d) Although substitution of missing data for more representative values can be justified against ignoring emissions, statistical estimators have been relatively successful at evaluating data missing at random or complete lack of data (Wernet et al. 2008; Liu et al. 2009). As described below, the properties of kriging incorporate data into systems of equations, without averaging out observed values.

Regression and other linear methods have been applied to inventory analysis, including for electricity generation (Hondo 2005). In general, and the case of hydropower is no exception, predictor (independent) variables are necessary to estimate the dependent variables, missing material and energy flows for a given process. More than the scale of processes involved, their shared characteristics, such as technical specifications, become predictor variables or coordinates, and allow the formulation of statistical estimators. These characteristic variables can be several, discrete or continuous. For instance, within the process of hydroelectric power generation, the

methodology is applied to hydropower plants of varying capacity, some with reservoirs and others without. Despite widespread claims of hydroelectricity being a renewable source of power, less than a handful LCAs have been conducted on hydropower plants (Ribeiro and Silva 2010). Most studies have emphasized site specificity and scale as major issues as well as questionable comparisons with other large sources of power (e.g. Gleick 1992; Gagnon, Bélanger, and Uchiyama 2002).

5.3.3.1 System characteristics

The rated power of hydroelectric plants depends on two main factors, the water flow and the hydraulic head or height difference between water intake and turbine, according to the formula:

$$P = \eta \rho g Q H \quad (1)$$

where P is power (W), η is a coefficient of efficiency, ρ is the density of water (~1000 kg/m3), g is the acceleration due to gravity (~9.8 m/s^2), Q is the flow rate (m^3/s) and H is the head (m). In practice this is implemented with either low head and high flow (generally run-of-river plants), high head and low flow (generally with reservoirs) or combinations thereof. Again, this scale varies considerably with respect to site specifications, and both head and flow vary seasonally. The highest head in the world is 1883 m at Bieudron power station in Valais, Switzerland, and the largest flow, up to 950 m^3/s, can be found in a turbine of China's Three Gorges dam. The efficiency of hydropower plants is among the highest of large power sources, 82 to 88 % overall (Bauer et al. 2007). The difference between run-of-river and reservoir plants lies essentially with storage, which is one of the key advantages of hydroelectric dams. Electricity cannot be stored such that supply must match demand almost instantaneously and hydropower plants can respond and adjust their production within minutes.

Despite the specificity of every power plant and reservoir, a set of characteristics was chosen based primarily on data availability. Given the background description above, table 6 summarizes the characteristic variables of hydroelectric power plants referred to in this article. Clearly these are several among the many characteristics which define the inventory of hydroelectric power plants.

Table 6: System characteristics

	Characteristic variable	Related characteristics	Range
I	Type (0/1)	Dam(s)/dike(s)/spillway(s) River regime	Run-of-river/Reservoir
II	Total installed capacity (MW)	Size and number of turbines/alternators Dam(s)/dike(s)/spillway(s) River regime	20 – 14000
III	Annual production (GWh/year)	Size and number of turbines/alternators River regime	90 – 90000
IV	Hydraulic head (m)	Length of penstock Volume of turbined water	6 – 1400
V	Surface area of reservoir (km^2)	Flooded area Power per surface area	0 – 4200
VI	Volume of reservoir (km^3)	Electricity storage capacity	0 – 140

The construction of large hydroelectric power plants requires the temporary diversion and permanent impoundment of a river. Dams and dikes increase the storage capacity and raise the hydraulic head of a power plant. They may consist of concrete at various concentrations of cement as well as earth or rock filled dams with either clay or asphalt cores or concrete slab surfaces for imperviousness. Spillways are part of both run-of-river and reservoir plants and dimensioned according to river regime to evacuate flood waters. The line between large and small hydropower projects is somewhat arbitrary. Although nameplate capacity is often the defining criterion, no single variable is necessarily more important than another. The Manicouagan reservoir for example is among the world's largest but only about a quarter of its

volume can be used for electrical generation purposes. Installed capacity and annual production are useful to evaluate plants which continually operate at low capacity factors.

Technical characteristics can be identified for many different processes, including chemical manufacturing (Wernet et al. 2008). Decomposing processes into *N* objects described by one or more characteristic variables and applying the procedure *N* times leads to more consistent samples and eventually more complete inventories. Hydroelectric power plants may be divided according to their type and further into dams, spillways, power houses, turbines, alternators and transformers. In practice, data availability and accessibility dictate the level of aggregation which is often high as shown by a subset of the data provided in the annexes. Although material and energy requirements for structures and equipments are sometimes collected separately, hydroelectric power plants as a whole are considered here, independently of the type of dam or construction technique. Ribeiro and his colleague (2010) showed the inventory of a hydroelectric power plant to be sensitive to time horizons. Optimistically the useful lives of structures and equipments were adjusted to average 100 and 50 years, respectively. If they can technically last that long, chances are operations and economics dictate otherwise. Conversely, the age of certain reservoirs has no significant relationship with the level of emissions from decaying biomass (Rosa et al. 2004).

By and large the main material and energy flows involved in hydroelectric power construction and operation are cement and concrete, steel of various grades for structures and equipment, explosives, and fuels and lubricants for machinery and equipment. They are listed as inventory flows in the data available in the annexes. Their provision and the construction phase itself often account for an important share of hydropower's environmental impact (Peisajovich 1997; Bauer et al. 2007; Ribeiro and Silva 2010). Water is essential and most important by weight or volume, its quality is altered and course permanently changed. Often ignored, accounting for water consumption and withdrawal of power generation technologies is changing, however, out of necessity (Fthenakis and Kim 2010; Macknick et al. 2011). For hydroelectricity, quantities can be estimated from production data and equation (1) and the impacts of water use have recently been characterized in LCA (Pfister, Saner, and Koehler 2011). How these dependent variables, material and energy flows, are estimated based on the independent characteristic variables in table 6 is explained in the following section.

5.3.3.2 Kriging estimator

Borrowed from spatial statistics, kriging distinguishes itself on several accounts from other linear estimators. Clearly no cartesian points exist in LCIs such that the characteristic variables described in table 6 become coordinates and the material and energy flows or emissions to be estimated, dependent variables. Kriging is a weighted linear combination of the observations and in the single variate case, can be represented as follows for estimates at x_0:

$$\hat{Z}(x_0)=\sum_{i=1}^{N}\lambda_i Z(x_i) \quad (2) \qquad \sum_{i=1}^{N}\lambda_i=1 \quad (3)$$

with the x_i as characteristic variables and $Z(x_i)$ as inventory flows for a sample of size N with kriging weights λ_i. As shown in table 7, the estimator is unbiased provided equation (3) holds. The expected values of material and energy flows are unlikely to remain constant with respect to changing characteristic variables. A power plant generating 100 GWh annually will not have the same embodied material and energy or reservoir emissions on average as one generating ten times as much. Thus, assume a model combining a random function $Y(x)$ as well as a deterministic drift on the mean $m(x)$ or low order polynomial:

$$Z(x)=Y(x)+m(x)=Y(x)+\sum_{l=0}^{L}a_l x^l \quad L=0,1,2 \quad (4)$$

This particular combination reflects technological systems more appropriately than the estimator with constant mean as shown by other forms of variable means (Joseph 2006). More properties of kiging are listed in table 7:

Table 7: Properties of the kriging estimator (adapted from Isaaks and Srivastava 1989)

	Properties	Description	Expression
I	Unbiased estimator	Expected value of the estimator equals the expected value of the observations.	$E\left[\hat{Z}(x_i)\right]=E\left[Z(x_i)\right]$
II	Exact estimator	Estimating an observed value returns the observed value itself. No loss of information from averaging.	$\hat{Z}(x_i)=Z(x_i)$
III	Screening effect	Observations close to the estimated values receive higher weights.	
IV	Position and sample	Covariance functions *C(h)* account for relative positions of observations, comparatively smaller samples are sufficient.	$Cov[Z(x_i), Z(x_i+h)]=C(h)$
V	Smoothing effect	The kriging estimator usually varies less than the phenomena being estimated.	$Var(\hat{Z}(x_i))\leq Var(Z(x_i))$

where *E[ ]* and *Var[ ]* are the expectation and variance operators respectively. As above, the x_i are characteristic variables and the $Z(x_i)$ corresponding material and energy flows. Kriging minimizes the error variance, in other words the variance of the difference between observed and estimated values (see annexes). The main assumption claims that covariances are independent of position and thus equal to functions of the distance *h* between observations only, as shown in table 7. Such covariance functions $C(h)$ translate in formal terms the idea that distinct observations close to one another should agree more than if they were far apart. A number of valid covariance functions exist and can be found in the annexes (Chiles and Delfiner 1999). Each function has three parameters, the first of which is a so called nugget effect which captures variations on a small scale and can be understood as an interpolating smoothing parameter (Marcotte and David 1988). Note that since kriging is an exact estimator, it is discontinuous at every observations. The second parameter is a structured variance which equals the total variance when added to the nugget effect. Third, the range of a covariance function corresponds to the distance *h* at which the covariance is nil or observations are too far apart and unrelated.

If deterministic and random components as well as covariance functions impart great flexibility to kriging, another key advantage is the multivariate estimator, or cokriging. Multivariate kriging enables the estimation of primary variables using data from secondary variables simultaneously (Wackernagel 2003). All estimations can be performed jointly as the order between primary and secondary variables can be interchanged (Marcotte 1991). Moreover primary and secondary variables need not be observed at the same points. In other words when one variable cannot be observed, its missing values can be inferred based on the values of other variables or covariates. For example, if data on explosives are available for the construction of a series of power plants and dams whereas the volumes excavated are not, joint estimation might provide more accurate results for excavation. Since the main goal is to address the lack of data, the methodological approach can improve the completeness of inventories in spite of scarcity.

Small sample sizes (here $N = 27$) forfeit many attempts to validate a model and its parameters. In other words dividing the data into training and validation sets is problematic, a case sometimes referred to as uncomfortable science (Hoaglin, Mosteller, and Tukey 1985). Cross-validation however, has no prior assumption such as normality, and observations take part in both trial and validation data sets (Marcotte 1995). Leave-one-out cross-validation (LOOCV) successively removes observations and derives estimates from neighboring values exactly once. In other words each observation forms the validation data once and part of the training data set $N - 1$ times. This simple procedure allows for both the selection of parameter values and the calculation of estimation errors. Moreover, with cokriging, the values of secondary variables, or covariates, for a given observation can be part of the validation data all together or one after the other. These two options are referred to as Leave-one-observation-out cross-validation (LOoOCV) and Leave-one-variable-out cross-validation (LOvOCV), respectively. Comparing the results of the two validation procedures gives an indication of the advantages of cokriging, should the errors calculated from LOvOCV be lower than that of LOoOCV. The chosen criterion of comparison is mean absolute errors (MAEs) or the mean of the differences between observed and estimated values without regard to their sign. MAEs are a reliable measure of accuracy with scaled data, divided by their own means.

5.3.4 Results

The results draw upon the characteristics of hydropower plants, large systems with widely different properties and affected by data scarcity, but apply elsewhere. Besides secondary data sources, namely the ecoinvent database and Itaipu assessment, primary data were collected to achieve a reasonable sample size with characteristic variables in the ranges shown in table 6. Emphasis was put on material and energy flows in the construction and operation phase, data made accessible by public utilities and less controversial than reservoir greenhouse gas emissions (Cullenward and Victor 2006). Although more material and energy are needed for dismantling hydro dams and power plants, their end of life is usually ignored. No large dam has been dismantled to date but repository organizations track current works on small ones (see CDRI: Clearinghouse for Dam Removal Information).

Because of different scales the data were normalized, dividing variables by their mean values. Moreover, material flows were selected to compare the complete case – data on cement and steel are available for most observations – with data missing at random, excavated volumes and explosives have approximately 30 % of missing values. Figures 4 to 7 show mean absolute errors (MAEs) calculated from leave-one-observation-out cross-validation (LOoOCV) as well as leave-one-variable-out cross-validation (LOvOCV) in the case of cokriging. Estimations in figures 4 and 5 are based on two characteristic variables, annual production and installed capacity, whereas figures 6 and 7 use annual production and hydraulic head. They are identical for the different forms of kriging and regression. A linear covariance function was found to perform better than the other spherical or cubic types (see annexes).

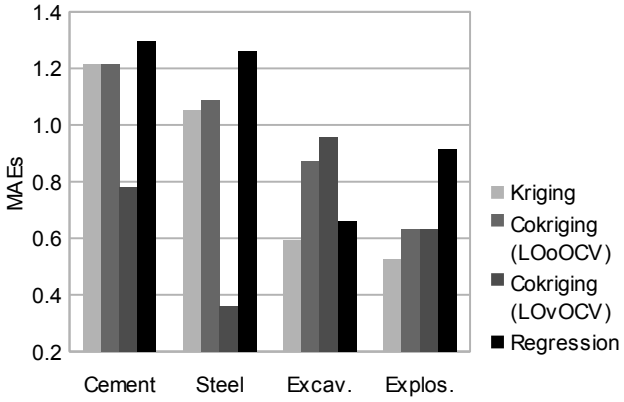


Figure 4: Comparison of MAEs based on production and capacity, pairwise cokriging of cement & steel, excavation & explosives

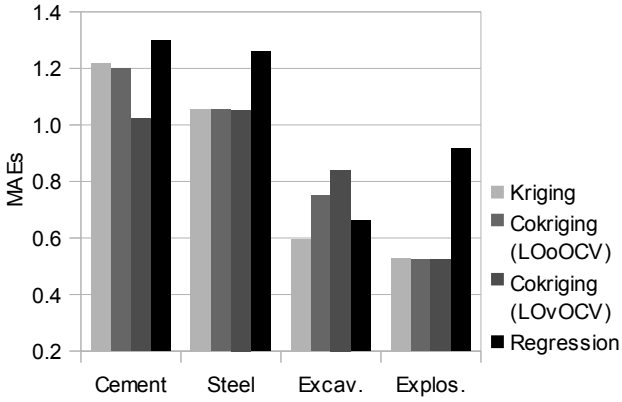


Figure 5: Comparison of MAEs based on production and capacity, pairwise cokriging of cement & excavation, steel & explosives

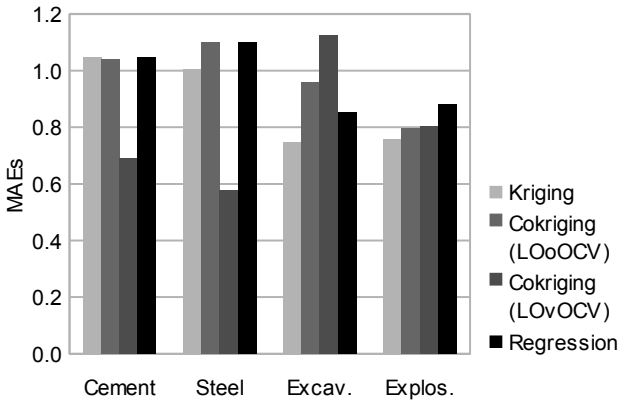


Figure 6: Comparison of MAEs based on production and head, pairwise cokriging of cement & steel, excavation & explosives

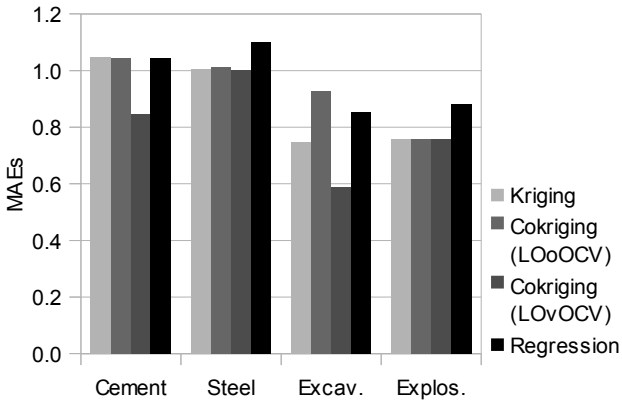


Figure 7: Comparison of MAEs based on production and head, pairwise cokriging of cement & excavation, steel & explosives

The important difference between figures 4 and 5, respectively 6 and 7, is how variables are paired in multivariate or cokriging estimation. First the well documented inventory flows of cement and steel are jointly estimated. Similarly for the flows with missing data, excavation volumes and explosives. Second, each of the well documented flows are co-estimated with one of the missing data flows. The most obvious result is also the most interesting: univariate kriging errors are lower than that of regression in all but two cases where kriging and regression fare

equally well. This shows not only that kriging performs better in the relative absence of data but also provides more accurate estimates. With the exception of excavated volumes, multivariate kriging also shows lower errors than regression, particularly when estimating one variable at a time (LOvOCV). An important observation which connects with the potential of estimating the complete lack of data is the following: multivariate kriging of material and energy flows for which data are relatively abundant and scarce, simultaneously, improves accuracy as shown by comparing the cokriging estimates of figure 4 with 5 and 6 with 7. Also, the mean absolute errors for univariate kriging are lower than that of the other estimators for inventory flows with missing data in all but one case. These observations favor the original hypothesis that estimators as versatile as kriging apply especially when data are missing.

Testing for the error reducing potential of accounting for several system characteristics with kriging, the one and two dimensional cases are compared below. Intuitively, the more variables from which to estimate missing values the better. Kriging and cokriging MAEs in one and two dimensions are presented in figure 8 for the same selected material flows.

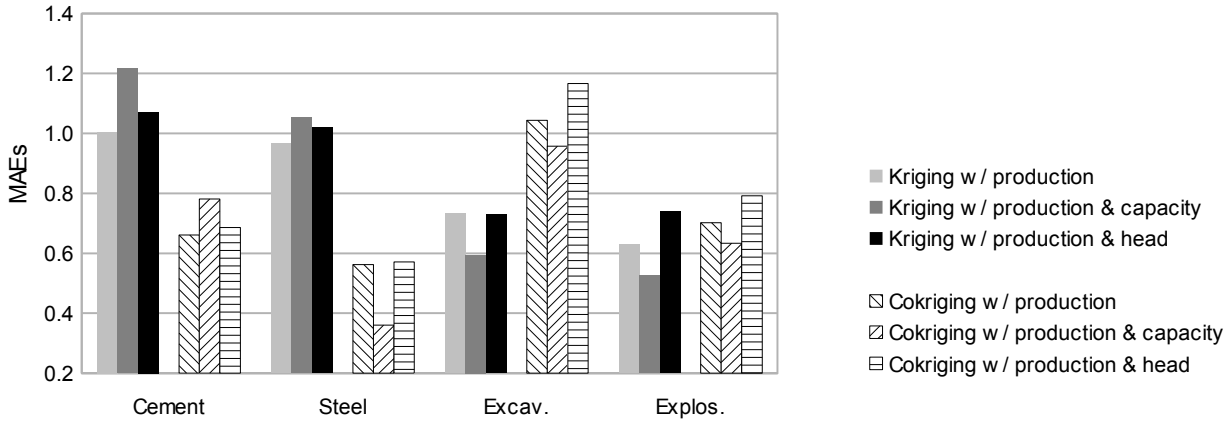


Figure 8: Comparison of MAEs for kriging and pairwise cokriging of cement & steel and excavation & explosives in 1 and 2 "dimensions"

The benefits of accounting for multiple characteristics are not straightforward. Annual production and installed capacity are positively and significantly correlated in this case whereas annual production and hydraulic head are not. From the right side of figure 8, adding capacity

improves the accuracy of the estimates but head does not. Indeed, hydraulic head even increases the errors observed compared to the one dimensional case. Although not split evenly the left side of figure 8 shows the opposite result. However, variables with incomplete data such as excavation and explosives above are precisely where two characteristic variables are better than one. Notice how the kriging errors decrease from left to right of figure 8 whereas cokriging errors alternate. Based on the results of figures 4 to 8, kriging consistently improves estimates especially when data are lacking. Even if the same conclusion holds for cokriging and the use of multiple characteristic variables, some exceptions occur.

5.3.4.1 Predicting material and energy requirements

A direct application of these results consist of estimating inventory flows required in the construction and operation of hydropower plants totaling 1550 MW and commissioned for the year 2020 by the provincial utility in Quebec, Canada. The project involves the construction of four dams and corresponding reservoirs covering 279.2 km^2 and replacing over 200 km of the Romaine river. Eight Francis turbines with average hydraulic head of 107 m are expected to produce 8000 GWh annually. Based on these characteristics, the predicted values of material flows are presented in table 8 for the kriging and cokriging models with linear covariance function, as well as regression. In the case of cokriging, cement and excavation volumes are jointly estimated and steel and explosives as well, to pair complete and incomplete inventory flows. While the observed data in this paper consist of very different hydropower plants covering a wide range of configurations as shown in table 6, the Romaine project includes four plants similar in design and location. The intervals shown in table 3 are one standard deviation (68%).

Table 8: Model estimates of selected material and energy flows for the Romaine project

Economic flows	Kriging	Cokriging	Regression
Cement (kt)	3.32E+02 ± 2.48E+02	3.00E+02 ± 2.43E+02	2.86E+02 ± 5.82E+02
Steel (t)	8.58E+04 ± 5.79E+04	8.35E+04 ± 5.79E+04	8.61E+04 ± 2.50E+05
Excavation (m^3)	2.83E+06 ± 1.14E+06	3.41E+06 ± 1.91E+06	3.33E+06 ± 1.26E+07
Explosives (t)	1.07E+03 ± 4.08E+02	1.07E+03 ± 4.08E+02	9.00E+02 ± 8.43E+02

Until the project is completed and operating at capacity, nothing can be said about the precision of either of these estimates. However, and despite differences of up to 20 % between estimators, the error calculations above show that kriging will be more accurate on average. The example in table 8 also shows how kriging provides better estimates compared to regression. Standard deviations are much lower for both kriging and cokriging than for regression, in some cases by an order of magnitude. As with the observed data from existing projects, these estimates could in theory be cross-validated with data from bills of material, engineering designs and other costs from contractors, provided they are accessible. Such validation process would require considerable data collection efforts and likely yield inconsistent results from one project to another, depending on data availability and level of aggregation. To the contrary, the statistical approach exemplified in this article shows how estimating missing inventory data can be conducted more systematically and accurately for both past and future projects.

5.3.5 Discussion and conclusions

The lack of data in life cycle inventory affects subsequent steps in the assessment and represents a significant source of uncertainty in the results. Data might be missing for specific processes or products, at random in existing inventories or nonexistent. In this article, the authors show how missing data on material and energy flows can be better estimated based on the design and operating characteristics of hydropower plants. Such characteristics can be identified for other systems and processes and shift the data collection efforts away from inaccessible or

unavailable inventory data. Indeed, a statistical approach differs substantially from the conventional process models but is complementary in many respects.

As elsewhere in statistical problems, the larger the sample size the greater the significance of the results but with kriging results can be conclusive from relatively small sample (Gunes, Sirisup, and Karniadakis 2006). Assuming a minimum sample size, kriging is particularly adaptive to low data availability and high requirements. In addition, kriging assumes a form of relationship between observations represented in mathematical terms by the covariance function. Typically thirty to fifty observations are necessary to establish a covariance function from the data, however, as shown above predefined functions and their parameters can be selected with cross-validation and from smaller sample sizes. With no prior assumptions on the data themselves, the selection of covariance functions and parameter estimation will differ according to the data set and influence the results. The combination of random and deterministic components in the model, allows the analysis of a greater diversity and variability of processes with distinct characteristics. In LCA, a sample size of more than one observation is arguably large already, challenging statistical analysis. However, considering the large differences in capacity for electricity generation purposes for example, there are cases where analyzing groups of processes becomes less cumbersome than each of them separately. This leads to more consistency in estimating missing data across inventories. Data collection remains a limiting factor in the statistical approach presented above but in a sense which suits the assessment of some processes better than others. Indeed, as long as characteristic variables exist and describe processes and their complete or incomplete inventories, estimates for the inventory flows of similar processes can be derived accurately.

The example of electricity generation is convenient because of the variety of processes yielding the exact same product among other reasons, there are limitations with respect to kriging which are more general. The combination of deterministic and random components in the model described above suits industrial processes but remains more appropriate for interpolation than extrapolation purposes. In other words, parametric models gradually converge towards the mean of the observations beyond the ranges of characteristics such as those shown in table 6. Given the first order drift on the mean, the kriging estimates above are asymptotically linear, eventually reaching that of regression. Thus, when extrapolating far beyond the range of characteristic

variables, the comparison between kriging estimators and regression becomes meaningless. Technological leaps and shifts make it necessarily difficult to estimate missing inventories for processes significantly different from existing ones. An important aspect of kriging is the smoothing effect as stated in table 7. On one hand this causes large values to be underestimated and small values to be overestimated. On the other hand the smoothing parameter can be adjusted to incorporate more or less variability in the observed data and therefore account for some of the uncertainties.

The versatility of kriging allows for better estimation especially in the absence of complete data and the complete absence of some data with the use of covariates. The results above show that when data are missing at random, univariate kriging has consistently lower mean absolute errors than regression with its multivariate equivalent being even more accurate in some cases. Similarly, when comparing actual estimates, the standard deviations are lower for kriging than they are for regression. Two potential improvements over process-based models or linear estimators, and future research objectives, are first to incorporate the uncertainties in the observations and translate them into that of the estimates. Second, to make statistical estimators for missing data more accessible or at least acceptable among practitioners. Careful analysis of the data as well as the additional steps kriging requires, are welcome but also obstacles to the automation of the procedure in life cycle inventory and in other material and energy flow analysis. To conclude, the approach presented in this article contributes not only to the estimate of missing data but also to much needed representative data which lack from existing inventory databases, particularly in the case of hydroelectricity.

5.4 Handling data variability and uncertainty with kriging: the case of hydroelectric reservoir emissions

Moreau, V., D. Marcotte, G. Bage, R. Samson, submitted April 2012, *Environmental and Ecological Statistics*

5.4.1 Abstract

Greenhouse gas emissions from hydroelectric reservoirs have been both challenging and controversial to measure and estimate over long periods of time. One type of environmental impact assessment particularly sensitive to such data is life cycle assessment for products, processes and services. Energy sources are often compared based on life cycle emissions and electricity generation accounts for a significant share of the impacts for most products and services. Statistical techniques have slowly become part of uncertainty analysis in life cycle assessment (LCA). This paper addresses the issue of data uncertainty in the inventory phase of LCA with a modified version of the kriging estimator. Measures of uncertainty for observed data are incorporated in the estimation of missing data thanks to nugget effects specific to each observations. In cases of low data availability, different data sources, and their respective degrees of uncertainty, can be accounted for relatively simply. Three different models of covariance functions are compared for both the ordinary and the modified form of kriging. Data on greenhouse gas emissions from hydroelectric reservoirs are particularly variable and suitable to the modifications of the kriging estimator implemented in this paper. It is found that the modified form of kriging provides estimates which are less variable and more consistent with expectations than ordinary kriging. In turn, this approach can be applied to various data sources in order to incorporate their specific uncertainties and calculate that of the estimates.

Keywords: kriging, data uncertainty, hydroelectric reservoir emissions, life cycle inventory

5.4.2 Introduction

Data uncertainty and variability continue to affect the assessment of environmental impacts and in particular the life cycle assessment of products, processes and services (Reap et al. 2008). The reliability of life cycle assessment (LCA) depends essentially on the quality of the

collected data (Ayres 1995; Mongelli, Suh, and Huppes 2005). In LCA, data quality has been characterized with five semi quantitative indicators, based on the NUSAP system of Funtowicz and Ravetz (1990). Data reliability and completeness are two indicators along side the geographical, temporal and technological representativeness of the data (Weidema and Wesnaes 1996). Poor scores for any of these criteria translate into high uncertainties. In other words uncertainty stems from inaccurate or inconsistent data, lack of representative data or lack of data altogether (Huijbregts et al. 2001).

Paradoxically, data uncertainty and quality are not mutual opposites. Data which uncertainty is high but well quantified or qualified, is often preferable to a precise value without any sense of (in)accuracy. The sources and types of uncertainties must therefore be distinguished (Funtowicz and Ravetz, 1990). Even if terminology describing different types of uncertainties changes across disciplines, the most common type of uncertainty is data uncertainty or parameter uncertainty, as it is referred to in LCA (Ross, Evans, and Webber 2002). It describes the uncertainty in the observed or collected data and other model inputs. Uncertainty introduced by modeling choices is beyond the scope of this paper. Moreover, uncertainty analysis of the data prior to life cycle inventories largely remains a missing link in LCA (Hong et al. 2010). A life cycle inventory (LCI) pulls together data on the material and energy flows as well as the emissions crossing a system's boundaries, for example the inputs and outputs of an electrical power plant over its life time per unit of kWh generated. Accounting for uncertainty in environmental assessment means risking the credibility of its conclusions. Few methods allow to integrate uncertainty in the assessment and the trade offs between their added value and the extra time and energy deters many practitioners (Björklund 2002).

Statistical approaches have the unique benefits of incorporating uncertainty prior to any attempt at reducing it (Finnveden et al. 2009). Comparative LCAs will reward less documented products or processes, for the lack of data lowers environmental impact. As a result, uncertainty in life cycle assessment and other environmental decision support tools tends to be underestimated (Lloyd and Ries 2007). The goal of this paper is to present a practical and simple approach to account for uncertainty in LCI. It shows how variable data sources, with and without measures of uncertainty, can be incorporated in estimating inventory data and corresponding

uncertainties. The proposed method calls upon the properties of kriging and the use of varying nugget effects as explained below.

5.4.3 Methodological approach

Inventory data without characterization of uncertainty does not translate into more accurate assessments, hence the emphasis on data collection and uncertainty (Ross, Evans, and Webber 2002; Williams, Weber, and Hawkins 2009). One growing field of research attempts to reduce uncertainty by extending economic input-output data to material and energy flow balances. However, there is no evidence or method to show that introducing notoriously volatile monetary flows actually reduces uncertainty (Williams, Weber, and Hawkins 2009). The problem consists of incorporating uncertainty. Empirical data are collected in various ways and so are generic data sources such as inventory databases, where uncertainty measures are not necessarily consistent from one source to another (Sugiyama et al. 2005).

Uncertainty can be quantified based on confidence intervals or parameters such as variance and standard deviation. The latter are often, but not always, associated with a probability distribution function, for example the lognormal when dealing with physical flows (Heijungs and Frischknecht 2005). Data variability is also an issue when assigning distribution functions. Indeed, to cover the requirements of most environmental assessment tools, a variety of data sources is needed. Lack of consistency in the data and associated uncertainties come from differences in measurements, processes as well as the absence or confidentiality of the data (Sugiyama et al. 2005). Moreover, data in LCI come primarily from secondary sources and databases where differences are even greater because entries range from point estimates to averages (Williams, Weber, and Hawkins 2009).

Thus, arguments for the proposed approach are threefold, first before attempting to reduce uncertainty it needs to be incorporated in the observed and estimated data necessary for environmental assessments. Second, procedures to account for uncertainty should support the use of different and variable data sources. In particular data are sometimes subject to confidentiality or unavailability and should be estimated. Contrary to variability, uncertainty can be reduced by providing accurate estimates in place of missing and unrepresentative data (Björklund 2002). Third, a variety of measures of uncertainty exist such as standard deviations and statistical errors

as well as more qualitative indicators (Funtowicz and Ravetz, 1990). Based on these considerations, a modified version of ordinary kriging offers a number of advantages. Clearly there are no spatial coordinates in life cycle inventories, however, material and energy flows depend to some extent on technical specifications such as capacity of a hydroelectric power plant or surface of a reservoir. These characteristics become dimensions and their availability remains high. Kriging distinguishes itself by minimizing the error variance. Assume the following estimator (1) subject to the unbiasedness constraint (2):

$$\hat{Z}(x_0)=\sum_{i=1}^{N}\lambda_i Z(x_i) \quad (1) \qquad \sum_{i=1}^{N}\lambda_i=1 \quad (2)$$

where the x_i are characteristic variables, the $Z(x_i)$ emissions or material and energy flows, λ_i the kriging weights and N the number of observations. The errors e and error variance are written as follows:

$$e=Z(x_0)-\hat{Z}(x_0) \quad (3) \qquad Var[e]=Var[Z(x_0)]+Var[\hat{Z}(x_0)]-2\,Cov[Z(x_0),\hat{Z}(x_0)] \quad (4)$$

Substituting (1) in (4), the error variance of estimation becomes:

$$\sigma_e^2=Var[Z(x_0)]+\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j Cov[Z(x_i),Z(x_j)]-2\sum_{i=1}^{N}\lambda_i Cov[Z(x_0),Z(x_i)] \quad (5)$$

The variance σ_e^2 in (5) is then minimized under the condition (2) by forming the Lagrangian and taking the derivatives with respect to λ and the Lagrange parameter μ equal to zero.

$$\begin{cases} \sum_{j=1}^{N}\lambda_j Cov[Z(x_i),Z(x_j)]+\mu=Cov[Z(x_0),Z(x_i)] \quad \forall \quad i=1,\ldots,N \\ \sum_{j=1}^{N}\lambda_j=1 \end{cases} \quad (6)$$

which is more easily seen in matrix form:

$$\underbrace{\begin{pmatrix} \sigma^2 & Cov[Z(x_1),Z(x_2)] & \cdots & Cov[Z(x_1),Z(x_N)] & 1 \\ Cov[Z(x_2),Z(x_1)] & \sigma^2 & \cdots & Cov[Z(x_2),Z(x_N)] & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Cov[Z(x_N),Z(x_1)] & Cov[Z(x_N),Z(x_2)] & \cdots & \sigma^2 & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}}_{K_0} \underbrace{\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \mu \end{pmatrix}}_{\lambda_0} = \underbrace{\begin{pmatrix} Cov[Z(x_1),Z(x_0)] \\ Cov[Z(x_2),Z(x_0)] \\ \vdots \\ Cov[Z(x_N),Z(x_0)] \\ 1 \end{pmatrix}}_{k_0} \quad (7)$$

The weights λ can then be inserted in (1). At the optimum, the kriging variance is given by:

$$\sigma^2 = Var[Z(x_0)] - \sum_{i=1}^{N} \lambda_i Cov[Z(x_0), Z(x_i)] - \mu \quad (8)$$

The main assumptions of kriging are first and second order stationarity:

$$E[Z(x)] = m \quad (9) \qquad Cov[Z(x), Z(x+h)] = C(h) \quad (10)$$

where *m* is a constant. This assumption can be relaxed and a drift on the mean of the observations added to the estimator. $C(h)$ is a covariance function which must be valid to calculate the covariance matrices K_0 and k_0 in (7). Parameters of the covariance function are the range *a* or distance *h* beyond which covariances are nil, a so called nugget effect C_0, capturing variations on a small scale, and a structured variance or sill *C*. Table 9 lists a series of valid functions.

Table 9: Types of covariance functions (adapted from Chiles and Delfiner 1999)

	Type	Expression
I	Nugget	$C(h) = \begin{cases} C_0 \ \ if \ \ h = 0 \\ 0 \ \ if \ \ \lvert h \rvert > 0 \end{cases}$
II	Linear	$C(h) = C[1 - (h/a)]$
III	Exponential	$C(h) = \begin{cases} C[\exp(-3h/a)] \ \ if \ \ 0 \le \lvert h \rvert \le a \\ 0 \ \ if \ \ \lvert h \rvert > a \end{cases}$
IV	Spherical	$C(h) = \begin{cases} C[1 - 1.5(h/a) + 0.5(h/a)^3] \ \ if \ \ 0 \le \lvert h \rvert \le a \\ 0 \ \ if \ \ \lvert h \rvert > a \end{cases}$

V	Cubic	$C(h)=\begin{cases} C\left[1-7(h/a)^2+8.75(h/a)^3-3.5(h/a)^5+0.75(h/a)^7\right] & if \quad 0\le\lvert h\rvert\le a \\ 0 \quad if \quad \lvert h\rvert>a \end{cases}$

The variance σ^2 is $C_0 + C$ and in general the nugget effect C_0 can be interpreted as a smoothing parameter (Marcotte and David 1988). A pure nugget effect or infinite smoothing corresponds to regression whereas zero nugget effect means no smoothing. Small values tend to be overestimated and large ones underestimated with smoothing. In other words kriging estimators vary equally or less than observed data, as expressed in (11):

$$Var[\hat{Z}(x_i)]\le Var[Z(x_i)] \quad (11)$$

When there are strong indications of variability in the observed data, smoothing or the magnitude of the nugget effect, need not be uniform across all observations. If C_0 is the uncertainty or error in the data, then the variance σ^2 can be evaluated as C (Cressie 1993). A modified version of the model takes the form (12):

$$Z(x)=Y(x)+e(x) \quad (12)$$

where $e(x)$ is the error term with variance C_0. At any given point x_0, $Y(x_0)$ is estimated instead of $Z(x_0)$, with variance C. A different value of C_0 can therefore be assigned to each observation as a measure of uncertainty. The linear system of equations in (7) is therefore modified as follows:

$$\left[\underbrace{\begin{pmatrix} C & C(h_{12}) & \cdots & C(h_{1N}) & 1 \\ C(h_{21}) & C & \cdots & C(h_{2N}) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C(h_{N1}) & C(h_{N2}) & \cdots & C & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}}_{K_0} + \underbrace{\begin{pmatrix} C_{01} & 0 & \cdots & 0 & 0 \\ 0 & C_{02} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & C_{0N} & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}}_{diag(C_0)}\right]\underbrace{\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \mu \end{pmatrix}}_{\lambda} = \underbrace{\begin{pmatrix} C(h_{10}) \\ C(h_{20}) \\ \vdots \\ C(h_{N0}) \\ 1 \end{pmatrix}}_{k_o} \quad (13)$$

Adding a positive diagonal of error or uncertainty factors C_0 to the covariance matrix K_0 maintains its positive definiteness.

To test the validity of the model adjusted for variable nugget effect and compare with ordinary kriging, standard deviations are one useful criterion. They are calculated through cross

validation, which is particularly appropriate with small samples (Marcotte 1995). A special case of *k-fold cross validation* with k =1, *leave one out cross validation* (LOOCV), successively removes each observation and estimates it with the $N-1$ neighbors. As the covariance functions in table 9 are not estimated from the data – a larger sample would be necessary to do so – an adjustment must be made to calculate the standard deviations. From the errors in equation (3), a value for b is computed as follows:

$$E[\frac{(Z(x_i)-\hat{Z}(x_i))^2}{(C_{0i}+\sigma_{ki}^2)}]=b \quad i=1,\ldots,N \quad (14)$$

Hence $E[(Z(x_i)-\hat{Z}(x_i))^2/b(C_{0i}+\sigma_{ki}^2)]=1$. Thus, the factor b ensures the normalized errors have unit variance. The same estimates $\hat{Z}(x_i)$ are obtained but the kriging variances are multiplied by b and the derived standard deviations adjusted to the data. The modified form of kriging is applied to the case of greenhouse gas emissions from hydroelectric reservoirs, which exhibit a wide range of ecological conditions and highly variable data sources and measurements.

5.4.4 Hydroelectric reservoir data

Hydroelectricity has often been considered a carbon free source of energy. However, extensive research over the greenhouse gas emissions of reservoirs has been conducted since publication of the first estimates (Rudd et al. 1993). As noted by Cullenward and Victor (2006), most of that research comes from scientists sponsored by large hydropower utilities. The uncertainties in their results have been exploited by utilities in much the same way climate change deniers have used uncertainties to steer the political debate. For comparison purposes with other sources of power such as coal and natural gas, impacts from reservoirs have focused on greenhouse gases (GHGs) and methyl mercury emissions (Rosenberg et al. 1997). GHG emissions from reservoirs are not accounted for in national inventories and could significantly alter that of Brazil and Canada in particular. Duchemin and his colleagues (2002) suggest that accounting for reservoir emissions would increase Canada's emissions by 3 % and the power utilities' by 17 %. With the exception of a spike during the first several years after flooding, there is no obvious pattern of decline in reservoir emissions as flooded trees decompose very slowly

and certain types of soils, such as peat, release carbon beyond the useful life of hydro reservoirs (St Louis et al. 2000; Huttunen et al. 2002; Soumis et al. 2005).

Three pathways dominate GHG emissions from reservoirs, diffusion and bubbling at the surface and degassing from turbines and spillways. Both, the prevailing pathway and the level of emissions depend on various factors, climate conditions, amount and type of vegetation flooded, water residence time, reservoir shape and volume (Demarty et al. 2009). Even old reservoirs behind run of river plants where residence times are short can be significant sources of GHG, more than twice the average bubbling methane emissions of tropical reservoirs according to a recent case study (DelSontro et al. 2010). Most ecosystems prior to flooding are carbon sinks, such that gross estimates necessarily underestimate GHG fluxes (St Louis et al. 2000). Few studies have attempted to measure net emissions (Bodaly et al. 2004; Demarty et al. 2009). Despite its importance in deeper reservoirs, degassing of turbined waters is problematic to measure (Fearnside 2004). In other words, on top of the variability between reservoirs, there is a lack of consistency across studies on what fluxes are measured, contributing to overall uncertainties. Without uncertainty analysis, the most recent study acknowledges that methods which capture the variability of reservoir emissions would further increase estimates (Barros et al. 2011). A statistical approach via resampling methods has already estimated historical methane emissions from large dams, including all three major pathways (Lima et al. 2008).

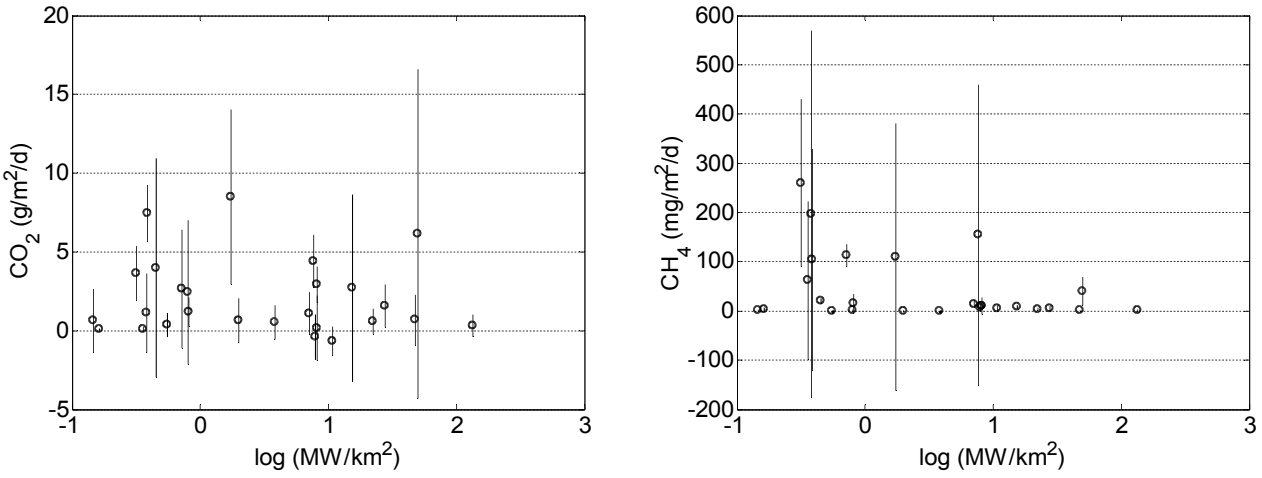


Figure 9: *a* Carbon dioxide (CO_2) and *b* methane (CH_4) emissions from selected reservoir case studies (Huttunen et al. 2002; Soumis et al. 2004; Soumis et al. 2005; Delmas et al. 2005; dos Santos et al. 2006; Demarty et al. 2009)

As shown in figure 9, the diversity of reservoirs is large and so is the variability (standard deviations). Among the reservoirs above, each of the tropical, temperate and boreal climates are approximately equally represented. The ratio of power capacity (MW) to surface area (km^2) often classifies hydroelectric reservoir stations in terms of emissions per kWh of output. According to the United Nations framework convention on climate change, hydroelectric dams with a ratio superior to 10 (dams with log $MW/km^2 > 1$ in figure 9) qualify for carbon credits. Some reservoirs are sinks for carbon dioxide which explains the negative values. No significant correlation can be observed between methane and carbon dioxide emissions. More than 8600 hydroelectric dams of 15 meters or higher exist (ICOLD 2011). Net emissions and long term studies would be prohibitively long and energy intensive for all of them. Hence the need to estimate emissions and quantify uncertainty even for the planned development of future large dams.

5.4.5 Results

The development of hydroelectric dams continues despite the increasing social, economic and environmental costs. China, Canada and Brazil are the world's largest producers of hydroelectric power and concentrate most of the large projects under construction (IEA 2010). Elsewhere, mainly in the Alps, dams are increasingly converted to pump and storage facilities, making estimation of the embodied emissions of kWh from hydropower more problematic (Deane, Gallachoir, and McKeogh 2010). Thus, estimates given below are based on two dimensions or characteristic variables, the installed capacity of power plants and surface area of reservoirs. While the former remains constant for a given power plant, the latter changes sometimes significantly with seasons, affecting the capacity factor.

A sample of 26 hydroelectric reservoirs of varying capacity and surface area, was assembled with the variances of methane and carbon dioxide emissions integrated as variable nugget effects in the model. Three projects were then selected and their emissions estimated, on the rivers Xingu (Belo Monte and Pimental dams, Brazil), Salween (TaSang dam, Burma) and Romaine (four dams numbered consecutively, Canada). They also illustrate three ecologically different sites, including two undammed rivers among which the longest in Southeast Asia. As

with many projects on similar scales, the development imperative skews the decision making process at the expenses of local populations, taxpayers, and environmental and energy conservation (Fearnside 2006). Tables 10 and 11 summarize the results for carbon dioxide and methane emissions respectively. For each model of covariance function in table 9, two values are given, first from the modified version of the model with varying nugget effect, second from conventional ordinary kriging.

Table 10: Comparison of estimates for carbon dioxide (CO_2) emissions with different covariance functions and variable versus constant errors

Dam	Capacity (MW)	Reservoir surface (km^2)	Linear ($g/m^2/d$)	Exponential ($g/m^2/d$)	Spherical ($g/m^2/d$)	Cubic ($g/m^2/d$)
Variable nugget effect C_0						
Belo Monte, Pimental	11181	441	0.34 ± 3.67	0.60 ± 4.62	0.26 ± 4.29	0.13 ± 4.31
Romaine 1,2,3,4	1550	279	1.10 ± 1.61	1.02 ± 2.21	1.10 ± 1.83	1.46 ± 1.16
TaSang	7110	870	0.77 ± 2.59	0.86 ± 3.70	0.80 ± 3.11	0.56 ± 2.44
Constant nugget effect C_0						
Belo Monte, Pimental	11181	441	0.57 ± 4.33	0.94 ± 4.87	0.06 ± 4.76	-0.87 ± 5.45
Romaine 1,2,3,4	1550	279	2.30 ± 2.53	2.00 ± 2.69	2.25 ± 2.57	2.98 ± 2.46
TaSang	7110	870	1.74 ± 3.50	1.54 ± 4.11	1.68 ± 3.82	1.21 ± 3.88

Table 11: Comparison of estimates for methane (CH_4) emissions with different covariance functions and variable versus constant errors

Dam	Capacity (MW)	Reservoir surface (km^2)	Linear ($mg/m^2/d$)	Exponential ($mg/m^2/d$)	Spherical ($mg/m^2/d$)	Cubic ($mg/m^2/d$)
Variable nugget effect C_0						
Belo Monte, Pimental	11181	441	7.5 ± 54.9	9.8 ± 61.3	5.7 ± 61.2	-1.5 ± 94.9
Romaine 1,2,3,4	1550	279	8.1 ± 21.2	8.5 ± 26.5	8.5 ± 23.0	2.5 ± 14.4
TaSang	7110	870	6.9 ± 40.2	8.2 ± 49.5	7.0 ± 45.3	-4.7 ± 53.4
Constant nugget effect C_0						
Belo Monte, Pimental	11181	441	5.7 ± 129.8	17.2 ± 153.2	11.0 ± 145.0	0.7 ± 157.4
Romaine 1,2,3,4	1550	279	22.1 ± 75.7	14.9 ± 84.6	19.4 ± 78.2	30.0 ± 71.1
TaSang	7110	870	17.7 ± 104.7	18.3 ± 129.1	17.7 ± 116.3	17.5 ± 112.0

An obvious result is the magnitude of standard deviations, making the dams potential sinks for methane, carbon dioxide or both. Nevertheless, the estimates and standard deviations are consistent with the observations, that is, equally variable. Results are also consistent across the four covariance functions, although the exponential and spherical models have mostly higher variations than the linear and cubic covariances. The more important result, however, is that standard deviations differ substantially between the modified and conventional forms of ordinary kriging, with more conservative estimates for the former. Such differences illustrate the fact that some measures are less uncertain than others which the modified form of kriging takes into account. In other words even if the results remain highly variable, incorporating errors specific to

each observation means kriging varies even less. Moreover estimated values of CO_2 and CH_4 emissions from the modified version of kriging disagree less with expectations for emissions from reservoirs in boreal versus tropical latitudes, contrary to conventional kriging. Figure 10 shows more clearly the differences in standard deviations for linear and cubic covariance functions with and without the variable nugget effects.

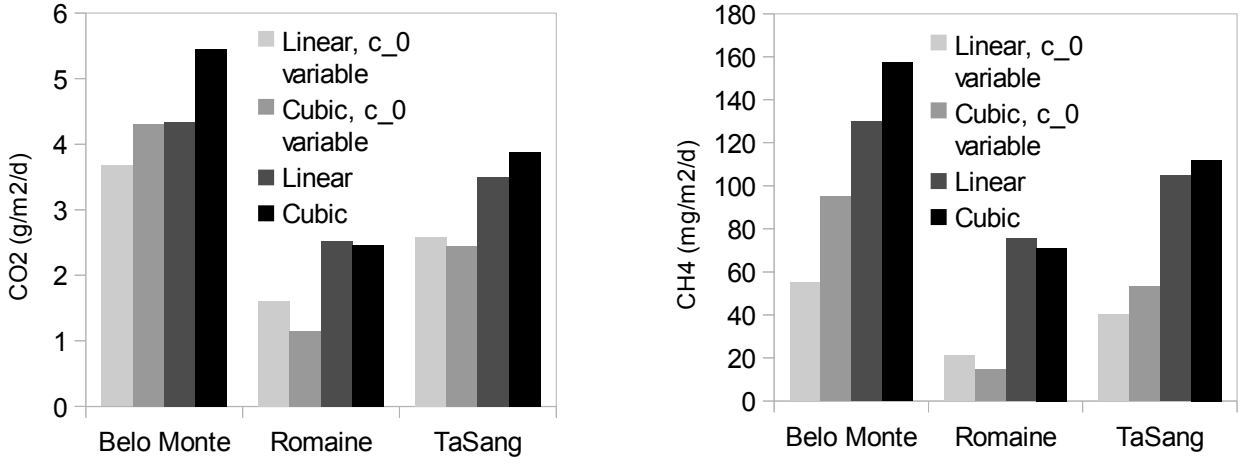


Figure 10: Comparing standard deviations from both forms of ordinary kriging with linear and cubic covariance and for CO_2 and CH_4 emissions

Standard deviations in figure 10 are invariably lower for the modified version of kriging. It therefore captures some of the subtleties inherent to highly variable data, and greenhouse gas emissions from reservoirs in particular, which would otherwise be averaged out. The greater variability of emissions from hydroelectric projects like Belo Monte makes them more uncertain whereas estimates for the emissions of northern reservoirs such as the Romaine are possibly more reliable. This illustrates the difficulty to predict greenhouse gas emissions from reservoirs, but as shown in tables 10 and 11, the estimates from kriging with variable nugget effect are also more conservative than conventional kriging. Thus, in the absence of data and for assessment purposes, the modified estimator provides a useful method of GHG evaluation.

Further tests were conducted by adding a linear drift on the mean of the observations. Instead of equation (9) above, the mean is allowed to vary according to:

$$E[Z(x)] = m(x) = \alpha x + \beta \quad (15)$$

where α and β are coefficients estimated from the data. While the estimates with non constant mean, also known as universal kriging, are not significantly different from the ones with *m* constant, the standard deviations are larger for both greenhouse gases. Moreover, the larger deviations with linear mean were found to be even larger in the case of constant nugget effect, the increase being negligible in the case of variable C_0. Therefore, with the absence of structure in the data, as shown in figure 9, including a linear tendency does not improve the analysis as much as accounting for the inherent variability of the data.

Cross-validation was also used to test for the value of *a*, the range of the covariance function, which yielded the lowest errors. A higher range fared better for both the linear and cubic covariances as well as the conventional and modified estimators, thus including positive covariances for almost all the observations in the sample. With such results in mind, the models of ordinary kriging with and without variable nugget effect are fitted to the data in figure 9, and illustrated in figure 11 for the cubic covariance function.

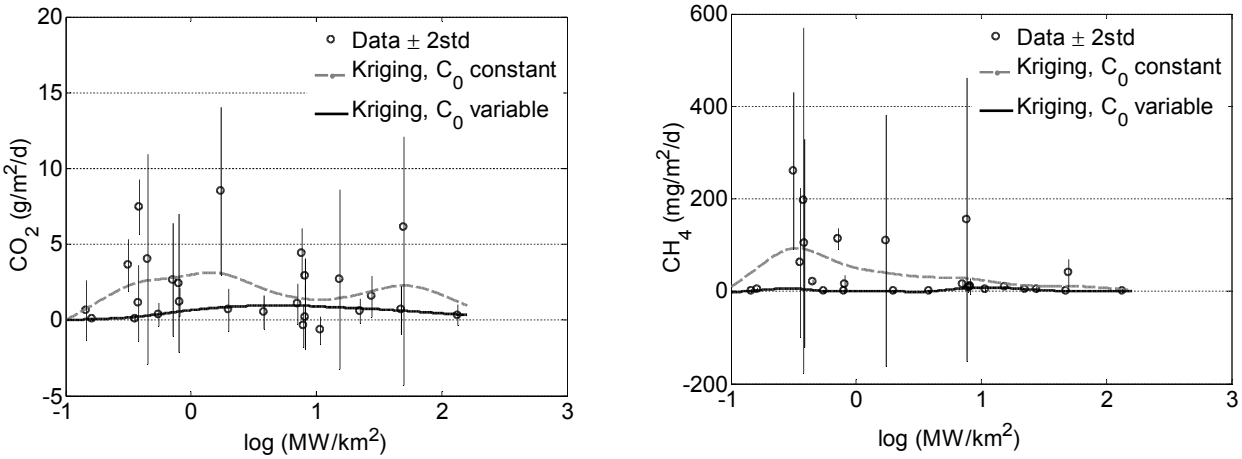


Figure 11: Kriging CO2 and CH4 emissions with and without variable nugget effect and a cubic covariance

As shown in figure 11, in this case the constant nugget effect in conventional ordinary kriging (dotted line) varies unnecessarily. To the contrary, the modified form of kriging (solid line) is less susceptible to data with large dispersion and provides more consistent estimates. They suggest that GHG emissions might be lower than that observed for certain reservoirs. However, the uncertainties remain high and even small daily values per square meter, add up to significant amounts of carbon per year. Thus, the approach developed here could serve as a basis for incorporating data variability and uncertainty in future environmental impact assessments and wherever data is both scarce and highly variable.

5.4.6 Discussion

The goal of this paper lies more with accounting for and less with reducing uncertainty in life cycle inventory and similar environmental assessment tools. In that sense the modified kriging model with varying nugget effect provides a relatively simple technique to incorporate different sources of data and corresponding uncertainties. Indeed when available, most databases and industry sources report data uncertainty differently (Sugiyama et al. 2005). Clearly variability of the data on reservoir emissions is high enough to justify such adjustment to the kriging estimator. Even then, the smoothing effect means kriging estimates tend to vary less than observed data. The use of relatively constant characteristic variables, power capacity and reservoir surface, means this approach emphasizes the permanent impact of reservoirs. Considerations of capacity factor or production in terms of kWh per year which constantly change depending on hydrological cycles and demand for power are less likely to reflect the same impact. Nevertheless, emission factors per kWh are frequently used even for hydroelectric dams and based on average production which remains confidential for many utilities.

The uncertainties in the results could be low enough for life cycle assessment practitioners to ignore them with the intention of simplifying the decision making process. However, given that uncertainties in LCA are generally underestimated, kriging with varying nugget effect possibly offers a more transparent picture precisely when data availability is low and uncertainty is high. For a complete interpretation of the results more data would be needed. One commonality between all three projects illustrated above is their design for export, either directly in the form of electricity (TaSang, Romaine) or indirectly as smelted alumina (Belo Monte, Romaine).

Expanding the system boundaries would require calculating the uncertainties of avoided emissions abroad.

In conclusion, the kriging estimator is extended here to more technological rather than spatial phenomena and the interpretation of its parameters differs from conventional applications. Cross validation is applied to evaluate a series of covariance functions and their parameters. An alternative use of the nugget effect allows error factors specific to each observation into the model. Modification of the kriging estimator in order to integrate widely different sources of environmental data, provides a relatively simple procedure to account for data uncertainties.

Chapitre 6 DISCUSSION GÉNÉRALE

Le but premier de cette thèse est d'améliorer la qualité des données d'inventaire en analyse du cycle de vie. Le manque de données est un problème récurrent en ACV et autant les coûts des données empiriques que le développement des bases de données, incitent les praticiens à le contourner, voire l'ignorer. C'est pourtant là une source d'incertitude qu'il est nécessaire d'analyser afin d'obtenir des résultats fiables et transparents. Un récapitulatif des points forts et faibles de l'approche proposée ici par rapport aux pratiques courantes permet d'évaluer l'utilité de celle-ci.

6.1 Avantages et inconvénients de l'approche statistique

6.1.1 Points forts par rapport à d'autres approches en ACV

Des techniques simples d'estimation des données manquantes, par exemple substitution avec des données existantes ou scaling, s'appliquent déjà en ACV même si l'absence de données est encore largement ignorée. L'hypothèse de linéarité des données y est acquise par défaut. Si la collecte de données concerne les processus élémentaires directement liés à la fonction étudiée, la plupart des données pour les processus en amont, en aval ou en arrière plan, provient de bases de données génériques. Ces dernières ne sont souvent ni représentatives du contexte technologique ou géographique ni suffisamment complètes (May and Brennan 2003). Le cas de hydroélectricité fait école mais des données manquantes se retrouvent dans de nombreux processus. L'approche par score pour évaluer une série de critères de représentativité des données fait appel à un jugement qui diffère d'une observation à l'autre. De plus, l'attribution d'une distribution de probabilités pour représenter l'incertitude d'une donnée s'avère problématique dans bien des cas (Heijungs and Frischknecht 2005). En revanche une approche quantitative grâce à des modèles statistiques qui nécessitent moins de données, est potentiellement plus appropriée en intégrant estimation des données manquantes et prise en compte des incertitudes existantes. Un grand avantage du krigeage est d'être, en moyenne, plus juste que d'autres estimateurs statistiques. Les

modifications apportées au krigeage ordinaire ne changent pas ses propriétés intrinsèques, au contraire, elles étendent sa flexibilité à des phénomènes plus technologiques que géologiques. Les processus étudiés ci-dessus requièrent pourtant une hypothèse, l'existence d'une relation entre les flux d'inventaire d'un processus donné et ses caractéristiques techniques. En somme il s'agit d'une forme de continuité de second ordre avec en lieu et place des variables régionalisées, des variables caractéristiques. Ces dernières sont généralement bien documentées et figurent dans le domaine public alors que les flux d'inventaires, lorsqu'ils sont disponibles, peuvent être confidentiels (Pehnt 2003; Wernet et al. 2008). Par comparaison avec d'autres méthodes d'estimation des données manquantes, les besoins en données additionnelles sont relativement faibles et, tout comme les flux d'inventaire, se résument à des données physiques et non économiques, telles que la puissance d'une turbine, la taille d'un réservoir, etc.

C'est là l'originalité du projet, d'abord de conserver une approche physique des données manquantes, c'est à dire de faire appel uniquement à un nombre limité de caractéristiques techniques permettant d'estimer les données d'inventaire. Cela répond aussi bien aux besoins de simplicité du modèle qu'à celui de portée, du fait que des données d'ordre économique par exemple manquent au même titre que les flux d'inventaire à estimer. Ensuite, l'approche méthodologique montre qu'avec certaines modifications, un outil statistique comme le krigeage fourni rapidement des résultats précis là où une modélisation des processus par bilans de masse et d'énergie se révélerait laborieuse. En outre, le modèle de krigeage élaboré ici permet de prendre en compte les incertitudes associées aux données observées et de calculer celles des données estimées. C'est tout l'avantage d'utiliser des fonctions de covariance dénuées d'effet de pépite, pour ajouter un effet spécifique à chacune des observations par après. D'autres approches, notamment hybrides, ne sont pas à priori développée pour une évaluation ou une prise en compte des incertitudes.

6.1.2 Limites de l'application du krigeage en ICV

Des parallèles peuvent être tracés entre les problèmes de données en géostatistique et en analyse d'inventaire du cycle de vie. Si les données manquent en inventaire pour de multiples raisons, en géologie leurs coûts d'acquisition peuvent être très élevés. Les besoins en données sont donc des phénomènes communs aux deux disciplines même si les enjeux sont très différents.

Les liens entre flux d'inventaire et variables caractéristiques peuvent être assimilés à ceux des variables régionalisées en géostatistique. Pourtant même après modification du krigeage pour contourner l'hypothèse de stationnarité, chaque étude demande une analyse des données et la sélection d'une fonction de covariance. Celle-ci peut se faire de manière empirique si l'échantillon est suffisamment large, sinon elle peut être sélectionnée parmi les fonctions préétablies. Le principal est d'identifier une relation entre flux d'inventaire et variables caractéristiques et d'éviter les échantillons où plusieurs processus, différant de par leurs données observées, correspondent à une même caractéristique. La matrice des covariances $C(h)$ pourrait devenir singulière et non inversible. Dans la pratique il faudrait que la plupart des observations d'un échantillon soient identiques ou de designs différents mais avec les mêmes spécifications techniques, un phénomène plutôt rare. Si tel est le cas, seule une des observations identiques ou leur moyenne peut être conservée.

Au delà des minima et maxima d'un échantillon, les valeurs estimées par krigeage se rapprochent asymptotiquement de la moyenne des observations. Si une dérive linéaire est ajoutée à celle-ci, les estimations par krigeage universel rejoignent progressivement celles d'une régression linéaire par exemple. Le krigeage est donc un meilleur interpolateur qu'extrapolateur. La validation croisée est appliquée ici comme alternative efficace à d'autres modes de sélection de fonctions de covariance et de leurs paramètres, pour laquelle la taille des échantillons peut être réduite (Marcotte 1995). L'application du krigeage nécessite évidemment une récolte de données suffisante, couvrant un maximum de processus dont les fonctions sont semblables, que les échelles de production et de consommation soient très variables ou non. Normaliser les données et sélectionner une fonction de covariance ne sont pas des étapes à négliger pour faciliter l'interprétation des résultats. Si une automatisation de cette approche à l'inventaire de processus choisis faciliterait son utilisation, les risques de voir des estimations invraisemblables par manque d'analyses préalables des données existent.

Finalement, la procédure d'estimation par krigeage nécessite au moins une variable caractéristique décrivant un processus. Dans la chaîne des processus élémentaires, il est ainsi possible de remonter jusqu'au point ou ces processus possèdent des spécifications observables. Par exemple la fabrication du ciment est un processus dont l'inventaire peut être estimer à partir de la conception d'une usine de fabrication. En revanche les caractéristiques de la production

agricole sont sujettes aux aléas du climat et les liens entre rendement et surface cultivées ou apports extérieurs ne sont pas aussi évidents. Le krigeage s'applique donc en priorité aux processus d'avant plan pour remonter progressivement la chaîne de production jusqu'aux activités industrielles qu'il est possible de mesurer et contrôler.

6.2 Implications des résultats pour l'ACV de la production d'électricité

La production d'électricité reste un facteur d'impact majeur dans l'analyse du cycle de vie de nombreux produits, processus et services. Une fois produit, il est impossible de distinguer la provenance d'un kWh, qu'il soit d'une centrale au charbon, à gaz, ou hydroélectrique par exemple. Avec l'intégration des réseaux de distribution, nationale et internationale, les producteurs répondent à la demande d'électricité en fonction des coûts et bénéfices escomptés. Ils peuvent la produire avec leurs propres installations ou acheter celle de leurs concurrents. Les échanges variant d'heure en heure, il est plus commode pour l'analyse du cycle de vie d'établir des moyennes de la provenance d'électricité sur un réseau donné. D'où l'adaptation des premières bases de données d'inventaire à d'autres réseaux en modifiant les proportions des différentes sources d'électricité. Pourtant, le choix d'un « grid mix » adéquat représente une grande source d'incertitude (Soimakallio, Kiviluoma, and Saikku 2011). Autre inconvénient, si ces moyennes changent d'un réseau à l'autre, l'inventaire des technologies de production, lui, reste sensiblement pareil. C'est le cas des centrales hydroélectriques par exemple, pour lesquelles l'inventaire de quelques barrages alpins fait référence. Or, peu de grandes centrales électriques varient autant que l'hydroélectrique. L'exception étant les centrales nucléaires qui fournissent moins d'électricité à l'échelle mondiale et pour lesquelles aucune analyse n'est en mesure de remplacer le cas par cas (Dones et al. 2005). Un autre problème en amont de la production d'électricité dans l'analyse du cycle de vie, est le traitement des minerais par exemple, qui requiert souvent de grandes quantités d'électricité. Non seulement l'approvisionnement de ces secteurs gourmands est généralement négocié hors du marché, mais aussi distribué via des lignes dédiées.

Ainsi l'analyse d'inventaires du cycle de vie spécifiques aux producteurs d'électricité de certains réseaux et même spécifiques à certaines industries reste importante (Finnveden et al. 2009; Marriott, Matthews, and Hendrickson 2010). C'est là en priorité que l'approche statistique présentée dans cette thèse trouve toute son utilité. En effet, à partir d'un échantillon de base, il est

possible d'estimer l'inventaire de centrales connectées à un réseau donné ou déconnectée du réseau pour desservir des activités industrielles particulières. Avec les spécifications techniques d'une série de technologies énergétiques, il est possible d'évaluer les flux économiques et élémentaires de sources d'énergie du même type. Les bases de données d'inventaire pourraient donc être conçues de façon légèrement différente, plutôt que la quantité nécessaire d'un processus, son type et ses caractéristiques serviraient ensuite à estimer l'inventaire requis pour répondre à une fonction ou fournir une quantité de produits ou services donnée. Avec une telle conception, un problème rencontré fréquemment dans la production d'électricité, à savoir quelle quantité de kWh supplémentaires nécessite une centrale additionnelle, ne se pose plus. Du moins le problème est ramené à la capacité de production elle-même, plutôt qu'aux échanges et surplus éventuels ailleurs. L'incertitude par rapport à quel producteur sera appelé à produire plus ou moins est très réel, notamment en ACV conséquencielle (Weidema 2009a).

Toutes les conséquences de la construction et de l'exploitation de ressources énergétiques ne sont pas quantifiables et encore moins caractérisées en ACV. Nombreux sont les problèmes d'ordre éthique pour lesquels d'autres approches existent. Néanmoins, l'approche statistique s'applique aussi à d'autres aspects importants de la production d'énergie, par exemple le retour sur investissement énergétique ou *energy payback,* soit le ratio entre l'énergie fournie et l'énergie dépensée pour le cycle de vie d'une centrale, de sa construction à son démantèlement. Celui-ci est un indicateur de plus en plus répandu pour comparer des sources de production d'électricité ou d'extraction conventionnelle et non conventionnelle de pétrole. Il peut être très élevé pour l'hydroélectrique, avec ou sans réservoir (Gagnon, Bélanger, and Uchiyama 2002). Comme le montre les exemples de centrales hydroélectriques présentés plus haut, la tendance vers une concentration de la production n'est pas prête de s'inverser. Par contre à l'avenir, une diversification des sources de production jouera en faveur d'une approche statistique, facilitant l'analyse d'inventaire d'une multitude de petites unités.

Chapitre 7 RECOMMANDATIONS ET CONCLUSIONS

Le débat sur l'estimation des incertitudes et des données d'inventaires s'enrichit doucement et se développe selon différents axes de recherche. Les principales recommandations et conclusions pouvant être tirées du projet de recherche effectué dans le cadre de cette thèse sont décrites ci-dessous.

7.1 Recommandations pour l'estimation de données manquantes en ICV

Parmi les différentes sources d'incertitude en ACV, l'incertitude des paramètres (données) reste une des plus étudiée suivie par les choix de modélisation. Pourtant dans la pratique, peu d'applications font état des incertitudes, par manque de ressources ou de méthodes. En effet, qui introduit une évaluation des incertitudes en ACV, court le risque d'obtenir des résultats peu concluants. Pour que l'approche cycle de vie soit réellement transparente, il est néanmoins important de qualifier et quantifier les incertitudes des données et rapporter celles-ci d'une façon comparable avec d'autres études. Les recommandations visent donc autant le développement méthodologique de l'analyse du cycle de vie que son application, en mettant l'accent sur la simplicité.

7.1.1 Propositions méthodologiques

Pour les données existantes, l'approche semi quantitative qui consiste à évaluer la qualité des données selon une série de critères est bien implantée en ACV et permet notamment de simuler l'influence de certains paramètres sur les résultats. Des progrès peuvent être accomplis, et cette thèse est une contribution dans ce sens, dans la caractérisation des données manquantes, dont la proportion dépasse celle des données disponibles selon certains experts. Il existe déjà plusieurs manières d'estimer ces données, par contre elles ne sont ni comparables entre elles ni en mesure de montrer les conséquences des données estimées sur l'incertitude des résultats. D’où l’intégration entre estimation des données manquantes et prise en compte des incertitudes établie

ici. Si le futur en inventaire du cycle de vie permet d'évaluer de la sorte chacune de ses propositions méthodologiques, un niveau supérieur de transparence serait atteint. Le cas des approches hybrides grâce à l'analyse « input-output » est particulièrement important tant elles se développent en ACV, sans pour autant apporter d'évidences pour ou contre une hypothétique évaluation des incertitudes.

Quelles que soient les approches privilégiées en ACV concernant l'absence de données d'inventaire, une distinction importante existe entre l'incertitude et la variabilité des données observées. Celle-ci ne se reflète pas nécessairement dans l'usage des écarts type, standards ou géométriques. D'autres mesures d'incertitude peuvent s'avérer utiles, notamment lorsque les données disponibles ne sont échantillonnées qu'une seule fois. En termes de processus élémentaire cela est fréquent lorsqu'un seul cas documenté sert de référence pour le même processus à des échelles très différentes. Si certaines possibilités d'intégration des méthodes statistiques en ICV sont déjà exploitées, il en existe d'autres comme le montre celles proposées ici qui à terme faciliteront sans doute l'analyse d'inventaire en soit, mais aussi augmenteront la cohérence entre les différentes approches.

7.1.2 Incertitudes des données d'inventaire en pratique

Une recommandation fréquente est d'évaluer les incertitudes dans toute analyse du cycle de vie. Les outils de simulations par exemple, contribuent aux analyses de sensibilité, mais ne concernent pas les données manquantes. Dans le cas présent une analyse d'incertitude s'applique avant tout aux données d'inventaire et puisque le secteur énergétique constitue généralement une part importante des impacts, il est naturel de commencer par ces données là. Toute proportion gardée, les incertitudes liées à la production d'électricité se retrouvent dans la plupart des cycles de vie des produits, processus ou services, presque indépendamment des frontières du système. Cette constante représenterait un premier pas vers l'évaluation des incertitudes de toutes les données d'inventaire. C'est aussi là que la contribution développée dans cette thèse serait des plus utile.

Outre la production d'électricité, une recommandation évidente est de favoriser l'usage de critères existants pour l'évaluation des incertitudes de façon à comparer les résultats d'une étude à l'autre et d'une méthode à l'autre. Puisque les simulations dépendent déjà de distribution de

probabilités, une approche statistique pour l'analyse des incertitudes se traduirait par des synergies notamment pour l'estimation des écarts type, standards ou géométriques.

7.1.3 Opérationalisation de l'approche statistique

Il ne s'agit pas de faire du krigeage une énième boite noire statistique qui réponde à tous les problèmes de données manquantes en inventaire du cycle de vie. De plus, comme expliqué ci-dessus, l'analyse ne bénéficierait pas d'une automatisation de toutes les étapes entre l'échantillon et les valeurs des flux élémentaires ou économiques recherchées. Deux options permettraient une meilleure intégration de l'approche. Tout d'abord l'adaptation des bases de données d'inventaire pour stocker en plus des données primaires, le système d'équation du krigeage correspondant, avec prise en compte des incertitudes et d'autres modifications réalisées ici le cas échéant. Ces bases de données pourraient ensuite être interrogées pour tout autre processus semblable et manquant dont les variables caractéristiques sont connues et constituent les intrants du modèle exécuté en arrière plan. Cela impliquerait une conservation des données d'échantillon dans les bases de données d'inventaire et poserait éventuellement des problèmes de confidentialité ou d'accès, même si ces données ne sont pas visibles pour l'utilisateur.

Ensuite, une conception autrement plus flexible et accessible consisterait en une structure autonome, reliée non pas par les données d'inventaire disponibles mais par les bases de données techniques. C'est à dire que les données échantillons restent indépendantes et la formulation des requêtes, semblable aux pratiques actuelles. En fonction de critères géographique, technologique, ou encore de capacité passée et future, la résolution ferait appel à la World Electric Power Plant Database (Platts) par exemple, qui regroupe quantité de variables caractéristiques pour des dizaines de milliers de centrales, toutes régions confondues. Bien que de telles bases de données soient accessibles, il s'agit d'informations exclusives pour lesquelles des droits doivent être acquittés.

Une application à grande échelle du krigeage en ACV ou en analyse de flux d'énergie et de matériaux justifierait une intégration de cette méthode avec les bases de données d'inventaire ou de variables caractéristiques. A l'inverse, si un consensus sur l'estimation statistique des données d'inventaire en ACV n'est pas une priorité, une approche par krigeage pourrait être réservée à l'analyse de processus dont les données font défauts, sous forme de programme

autonome comme il en existe en ACV, *Missing Inventory Estimation Tool* (MIET), CMLCA, openLCA, pour citer trois exemples.

7.2 Conclusions

L'absence de données pour l'évaluation de produits, processus ou services est un problème auquel les chercheurs sont continuellement confrontés. Il est paradoxal d'avoir besoin d'un échantillon raisonnablement grand pour pouvoir estimer de manière fiable l'impact d'une technologie particulière. Nombreux sont les exemples ou il n'est pas concevable d'observer suffisamment de données avant de devoir tout arrêter. Et inversement, certaines technologies se développent si rapidement qu'une fois atteint un seuil, la question de leur impact est malvenue. Il existe pourtant un milieu ou le recourt à des outils pour éclairer certaines décisions est nécessaire. Lorsqu'il s'agit de sources d'énergie, les enjeux sont considérables, que ce soit l'équité intra et intergénérationnelle ou la santé humaine et même celle de l'industrie. A mesure que la complexité des systèmes climatiques et technologiques se révèle, les prises de décisions en matière de santé et d'environnement sont distancées par des incertitudes grandissantes. C'est précisément là que les concepts de *uncomfortable science* et *post-normal science*, prennent tout leur sens. Plus encore qu'une évaluation systématique des incertitudes, l'analyse du cycle de vie, et ses outils apparentés, nécessite une participation étendue avant de voir les décisions qu'elle soutient contribuer à un réel changement. La question des incertitudes reste au cœur de cette thèse dont les points remarquables sont les suivants :

a) Les résultats des expériences menées dans le cadre de ce projet tendent à confirmer le potentiel du krigeage dans l'estimation de données manquantes en inventaire du cycle de vie. Des échantillons différents ont permis d'évaluer les propriétés du krigeage et d'illustrer les contributions importantes et originales.

b) Parmi les conclusions des articles ci-dessus, deux points ressortent clairement. D'abord en termes de réduction des erreurs d'estimation, le krigeage se distingue là où les données sont manquantes ou incomplètes, preuve qu'il s'agit d'un estimateur adapté à l'absence de données. Ensuite une interprétation alternative et une modification des paramètres du krigeage permettent une meilleure prise en compte de l'incertitude des sources de données, variables et variées.

c) Le krigeage se décline sous différentes formes, offrant une grande flexibilité. Par exemple l'estimation de données manquantes par cokrigeage réduit les erreurs moyennes absolues de manière significative lorsque ces flux d'inventaires sont corrélés. A l'inverse, en l'absence de données pour des flux d'inventaire non corrélés, le krigeage sur base de deux variables caractéristiques présente les erreurs les plus faibles.

En revanche l'utilisation d'estimateurs statistiques fait toujours débat au sein de la communauté ACV. Les résultats ci-dessus viennent non seulement nourrir ce débat mais montrent aussi qu'il est encore largement possible de développer ces applications en ICV. A mesure que les bases de données d'inventaire s'élargissent, notamment par l'intégration de facteurs économiques, des étapes de validation des données ou d'évaluation des incertitudes sont à prévoir. De plus, des études sectorielles ou régionales seront toujours nécessaires pour combler le manque de données représentatives et le krigeage montre qu'avec des échantillons de petite taille, nombreuses sont les données qui peuvent être estimées. Pour conclure avec un parallèle, l'analyse du cycle de vie s'insérerait naturellement dans une approche plus holistique, de type multicritère, alors qu'une approche statistique deviendrait un outil d'estimation parmi d'autres en analyse du cycle de vie.

LISTE DE RÉFÉRENCES

Ardente, F., M. Beccali, and M. Cellura. 2004. “FALCADE: a Fuzzy Software for the Energy and Environmental Balances of Products.” *Ecological Modelling* 176 (3-4) (September): 359–379.

Ayres, Robert U. 1995. “Life Cycle Analysis: A Critique.” *Resources, Conservation and Recycling* 14 (3–4): 199 – 223. doi:10.1016/0921-3449(95)00017-D.

Baccou, Jean, and Jacques Liandrat. 2011. “Kriging-based Subdivision Schemes: Application to the Reconstruction of Non-regular Environmental Data.” *Mathematics and Computers in Simulation* 81 (10): 2033 – 2050. doi:10.1016/j.matcom.2010.12.009.

Barros, Nathan, Jonathan J. Cole, Lars J. Tranvik, Yves T. Prairie, David Bastviken, Vera L. M. Huszar, Paul del Giorgio, and Fabio Roland. 2011. “Carbon Emission from Hydroelectric Reservoirs Linked to Reservoir Age and Latitude.” *Nature Geosci* advance online publication (July 31). doi:10.1038/ngeo1211. http://dx.doi.org/10.1038/ngeo1211.

Bauer, Christian, Rita Bollinger, Matthias Tuchschmid, and Mireille Faist-Emmenegger. 2007. “Wasserkraft.” In *Sachbilanzen Von Energiesystemen: Grundlagen Fur Den Okologischen Vergleich Von Energiesystemen Und Den Einbezug Von Energiesystemen in Okobilanzen Fur Die Schweiz*. Dones, R. et al. Final Report Ecoinvent No. 6-VIII. Paul Scherrer Institute Villingen, Swiss center for life cycle inventories Dubendorf.

Benetto, Enrico, Christiane Dujet, and Patrick Rousseaux. 2008. “Integrating Fuzzy Multicriteria Analysis and Uncertainty Evaluation in Life Cycle Assessment.” *Environmental Modelling & Software* 23 (12): 1461 – 1467. doi:10.1016/j.envsoft.2008.04.008.

Björklund, Anna. 2002. “Survey of Approaches to Improve Reliability in LCA.” *International Journal of Life Cycle Assessment* 7 (2): 64–72.

Bodaly, R. A, Kenneth G Beaty, Len H Hendzel, Andrew R Majewski, Michael J Paterson, Kristofer R Rolfhus, Alan F Penn, et al. 2004. “Experimenting with Hydroelectric Reservoirs.” *Environmental Science & Technology* 38 (18): 346A–352A. doi:10.1021/es040614u.

Brandt, Adam R. 2012. “Variability and Uncertainty in Life Cycle Assessment Models for Greenhouse Gas Emissions from Canadian Oil Sands Production.” *Environmental Science & Technology* 46 (2): 1253–1261. doi:10.1021/es202312p.

Burger, B., and Christian Bauer. 2007. “Windkraft.” In *Sachbilanzen Von Energiesystemen: Grundlagen Fur Den Okologischen Vergleich Von Energiesystemen Und Den Einbezug Von Energiesystemen in Okobilanzen Fur Die Schweiz*. Dones, R. et al. Final Report Ecoinvent No. 6-VIII. Paul Scherrer Institute Villingen, Swiss center for life cycle inventories Dubendorf.

Caduff-Kinkel, Marloes, Hans-Jorg Althaus, Stefanie Hellweg, Annette Koehler, and François Marechal. 2008. “Methodology for Establishing and Integrating Technical Scaling Laws in Life Cycle Assessment.” In Warsaw.

Capello, Christian, Stefanie Hellweg, Beat Badertscher, and Konrad Hungerbuhler. 2005. "Life-cycle Inventory of Waste Solvent Distillation: Statistical Analysis of Empirical Data." *Environmental Science and Technology* 39 (15) (August): 5885–5892.

CDRI. *Clearinghouse for Dam Removal Information*. http://library.ucr.edu/wrca/collections/cdri/.

ecoinvent Centre. 2009. *Ecoinvent Data V2.0*. Swiss Center for Life Cycle Inventories, St. Gallen.

Chevalier, J. L., and J. F. Le Teno. 1996. "Life Cycle Analysis with Ill-defined Data and Its Application to Building. Products." *International Journal of Life Cycle Assessment* 1 (2): 90–96.

Chiles, Jean-Paul, and Pierre Delfiner. 1999. *Geostatistics: Modeling Spatial Uncertainty*. Wiley Series in Applied Probability and Statistics. New York: Wiley Interscience.

Ciroth, A., G. Fleischer, and J. Steinbach. 2004. "Uncertainty Calculation in Life Cycle Assessments: A Combined Model of Simulation and Approximation." *International Journal of Life Cycle Assessment* 9 (4): 216–226.

Ciroth, Andreas. 2006. "Validation – The Missing Link in Life Cycle Assessment. Towards Pragmatic LCAs." *International Journal of Life Cycle Assessment* 11: 295–297.

Cressie, N. 1990. "The Origins of Kriging." *Mathematical Geology* 22 (3): 239–252.

———. 1993. *Statistics for Spatial Data*. Revised ed. Wiley Series in Probability and Statistics. New York, Toronto: John Wiley & Sons.

Crettaz, Pierre, Olivier Jolliet, and Myriam Saade. 2005. *Analyse Du Cycle De Vie, Comprendre Et Realiser Un Ecobilan*. Gerer L'environnement. Presses polytechniques et universitaire romandes.

Cullenward, Danny, and David Victor. 2006. "The Dam Debate and Its Discontents." *Climatic Change* 75: 81–86.

Curran, M. A., M. Mann, and G. Norris. 2005. "The International Workshop on Electricity Data for Life Cycle Inventories." *Journal of Cleaner Production* 13 (8): 853–862.

Dahllof, Lisbeth, and Bengt Steen. 2007. "A Statistical Approach for Estimation of Process Flow Data from Production of Chemicals of Fossil Origin." *International Journal of Life Cycle Assessment* 12 (2) (March): 103–108. doi:10.1065/lca2005.12.241.

Deane, J. P, B. P. O Gallachoir, and E. J McKeogh. 2010. "Techno-economic Review of Existing and New Pumped Hydro Energy Storage Plant." *Renewable & Sustainable Energy Reviews* 14 (4) (May): 1293–1302. doi:10.1016/j.rser.2009.11.015.

Delmas, Robert, Sandrine Richard, Frédéric Guérin, Gwénaël Abril, Corinne Galy-Lacaux, Claire Delon, and Alain Grégoire. 2005. "Long Term Greenhouse Gas Emissions from the Hydroelectric Reservoir of Petit Saut (French Guiana) and Potential Impacts." In *Greenhouse Gas Emissions - Fluxes and Processes*, ed. R. Allan, U. Förstner, W. Salomons, Alain Tremblay, Louis Varfalvy, Charlotte Roehm, and Michelle Garneau, 293–312. Environmental Science. Springer Berlin Heidelberg.

DelSontro, Tonya, Daniel F McGinnis, Sebastian Sobek, Ilia Ostrovsky, and Bernhard Wehrli. 2010. "Extreme Methane Emissions from a Swiss Hydropower Reservoir: Contribution from Bubbling Sediments." *Environmental Science & Technology* 44 (7): 2419–2425. doi:10.1021/es9031369.

Demarty, Maud, Julie Bastien, Alain Tremblay, Raymond H Hesslein, and Robert Gill. 2009. "Greenhouse Gas Emissions from Boreal Reservoirs in Manitoba and Québec, Canada, Measured with Automated Systems." *Environmental Science & Technology* 43 (23): 8908–8915. doi:10.1021/es8035658.

Dones, R., T. Heck, M. F. Emmenegger, and N. Jungbluth. 2005. "Life Cycle Inventories for the Nuclear and Natural Gas Energy Systems, and Examples of Uncertainty Analysis." *International Journal of Life Cycle Assessment* 10 (1): 10–23.

Duchemin, Éric, Marc Lucotte, René Canuel, and Nicolas Soumis. 2006. "First Assessment of Methane and Carbon Dioxide Emissions from Shallow and Deep Zones of Boreal Reservoirs Upon Ice Break-up." *Lakes & Reservoirs: Research & Management* 11 (1): 9–19.

Duchemin, Eric, Marc Lucotte, Vincent L. St Louis, and René Canuel. 2002. "Hydroelectric Reservoirs as an Anthropogenic Source of Greenhouse Gases." *World Resource Review* 14 (3): 334–353.

Egre, D., and J. C. Milewski. 2002. "The Diversity of Hydropower Projects." *Energy Policy* 30 (14): 1225–1230.

EUROSTAT. 2001. *Economy-wide Material Flow Accounts and Derived Indicators. A Methodological Guide*. Luxembourg: Statistical Office of the European Union.

Fava, J. A., A. A. Jensen, L. Lindfors, S. Pomper, B. de Smet, J. Warren, and B. Vigon. 1994. *Life Cycle Assessment Data Quality. A Conceptual Framework*. Ed. SETAC. Pensacola.

Fearnside, Philip M. 2004. "Greenhouse Gas Emissions from Hydroelectric Dams: Controversies Provide a Springboard for Rethinking a Supposedly 'Clean' Energy Source. An Editorial Comment." *Climatic Change* 66: 1–8.

———. 2006. "Dams in the Amazon: Belo Monte and Brazil's Hydroelectric Development of the Xingu River Basin." *Environmental Management* 38 (1): 16–27.

Filion, Pierre, and Christine Hebert. 2004. *L'Energie Au Quebec*. Secteur de l'energie et des changements climatiques Direction des politiques et des technologies de l'energie.

Finnveden, Göran, Michael Z. Hauschild, Tomas Ekvall, Jeroen Guinée, Reinout Heijungs, Stefanie Hellweg, Annette Koehler, David Pennington, and Sangwon Suh. 2009. "Recent Developments in Life Cycle Assessment." *Journal of Environmental Management* 91 (1): 1 – 21. doi:10.1016/j.jenvman.2009.06.018.

Frischknecht, R., N. Jungbluth, H. J. Althaus, G. Doka, R. Dones, T. Heck, S. Hellweg, et al. 2005. "The Ecoinvent Database: Overview and Methodological Framework." *International Journal of Life Cycle Assessment* 10 (1): 3–9.

Fthenakis, Vasilis, and Hyung Chul Kim. 2010. “Life-cycle Uses of Water in U.S. Electricity Generation.” *Renewable and Sustainable Energy Reviews* 14 (7): 2039 – 2048. doi:DOI: 10.1016/j.rser.2010.03.008.

Funtowicz, S. O., and J. R. Ravetz,. 1990. *Uncertainty and Quality in Science for Policy*. Vol. 15. Theory and Decision Library. Springer Berlin Heidelberg.

Gagnon, Luc, Camille Bélanger, and Yohji Uchiyama. 2002. “Life-cycle Assessment of Electricity Generation Options: The Status of Research in Year 2001.” *Energy Policy* 30 (14): 1267 – 1278. doi:DOI: 10.1016/S0301-4215(02)00088-5.

Gleick, Peter H. 1992. “Environmental Consequences of Hydroelectric Development: The Role of Facility Size and Type.” *Energy* 17 (8): 735 – 747. doi:DOI: 10.1016/0360-5442(92)90116-H.

Gunes, H., S. Sirisup, and G. E. Karniadakis. 2006. “Gappy Data: To Krig or Not to Krig?” *Journal of Computational Physics* 212 (1): 358–382.

de Haes, Helias A. Udo, and Reinout Heijungs. 2007. “Life-cycle Assessment for Energy Analysis and Management.” *Applied Energy* 84 (7-8): 817–827.

Heijungs, Reinout. 1996. “Identification of Key Issues for Further Investigation in Improving the Reliability of Life-cycle Assessments.” *Journal of Cleaner Production* 4 (3-4): 159–166.

Heijungs, Reinout, and Rolf Frischknecht. 2005. “Representing Statistical Distributions for Uncertain Parameters in LCA. Relationships Between Mathematical Forms, Their Representation in EcoSpold, and Their Representation in CMLCA.” *The International Journal of Life Cycle Assessment* 10 (4): 248–254.

Heijungs, Reinout, and Raymond Tan. 2010. “Rigorous Proof of Fuzzy Error Propagation with Matrix-based LCI.” *International Journal of Life Cycle Assessment* 15 (9): 1014–1019.

Hischier, R., S. Hellweg, C. Capello, and A. Primas. 2005. “Establishing Life Cycle Inventories of Chemicals Based on Differing Data Availability.” *International Journal of Life Cycle Assessment* 10 (1): 59–67.

Hoaglin, David Caster, Frederick Mosteller, and John Wilder Tukey. 1985. *Exploring Data Tables, Trends, and Shapes*. Wiley Series in Applied Probability and Statistics. New York, Toronto: John Wiley & Sons.

Hondo, Hiroki. 2005. “Life Cycle GHG Emission Analysis of Power Generation Systems: Japanese Case.” *Energy* 30 (11-12 SPEC. ISS.) (September): 2042–2056.

Hong, Jinglan, Shanna Shaked, Ralph Rosenbaum, and Olivier Jolliet. 2010. “Analytical Uncertainty Propagation in Life Cycle Inventory and Impact Assessment: Application to an Automobile Front Panel.” *International Journal of Life Cycle Assessment* 15 (5): 499–510.

Huijbregts, M. A. J. 1998a. “Part II: Dealing with Parameter Uncertainty and Uncertainty Due to Choices in Life Cycle Assessment.” *International Journal of Life Cycle Assessment* 3 (6): 343–351.

———. 1998b. “Application of Uncertainty and Variability in LCA.” *International Journal of Life Cycle Assessment* 3 (5): 273–280.

Huijbregts, M. A. J., W. Gilijamse, A. M. J. Ragas, and L. Reijnders. 2003. “Evaluating Uncertainty in Environmental Life-Cycle Assessment. A Case Study Comparing Two Insulation Options for a Dutch One-Family Dwelling.” *Environmental Science & Technology* 37 (11): 2600–2608.

Huijbregts, M. A. J., G. Norris, R. Bretz, A. Ciroth, B. Maurice, B. von Bahr, Bo P. Weidema, and A. S. H. de Beaufort. 2001. “Framework for Modelling Data Uncertainty in Life Cycle Inventories.” *International Journal of Life Cycle Assessment* 6 (3): 127–132.

Hung, Ming-Lung, and Hwong-wen Ma. 2009. “Quantifying System Uncertainty of Life Cycle Assessment Based on Monte Carlo Simulation.” *International Journal of Life Cycle Assessment* 14 (1) (January): 19–27. doi:10.1007/s11367-008-0034-8.

Huttunen, Jari T., Tero S. Vaisanen, Seppo K. Hellsten, Mirja Heikkinen, Hannu Nykanen, Hogne Jungner, Arto Niskanen, et al. 2002. “Fluxes of CH4, CO2, and N2O in Hydroelectric Reservoirs Lokka and Porttipahta in the Northern Boreal Zone in Finland.” *Global Biogeochem. Cycles* 16 (1) (January 19): 1003.

ICOLD. 2011. *World Register of Dams*. International Commission on Large Dams.

IEA. 2010. *Key World Energy Statistics*. International Energy Agency.

Isaaks, Edward H., and Mohan R. Srivastava. 1989. *Applied Geostatistics*. Oxford University Press.

ISO. 2002. *14048 Environmental Management - Life Cycle Assessment - Data Documentation Format*. International Organization for Standardization.

———. 2006. *14040 Environmental Management - Life Cycle Assessment - Principles and Framework*. International Organization for Standardization.

Joseph, V. Roshan. 2006. “Limit Kriging.” *Technometrics* 48 (4) (November): 458–466.

Lima, Ivan, Fernando Ramos, Luis Bambace, and Reinaldo Rosa. 2008. “Methane Emissions from Large Dams as Renewable Energy Resources: A Developing Nation Perspective.” *Mitigation and Adaptation Strategies for Global Change* 13 (2): 193–206.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. Second ed. Wiley Series in Probability and Statistics. New Jersey: John Wiley & Sons.

Liu, Yu, Xian Zheng Gong, Zhi Hong Wang, Wei Liu, and Zuoren Nie. 2009. “Multiple Imputation for Missing Data in Life Cycle Inventory.” *Materials Science Forum* Materials Research (610-613): 21–27.

Lloyd, S. M., and R. Ries. 2007. “Characterizing, Propagating, and Analyzing Uncertainty in Life-Cycle Assessment: A Survey of Quantitative Approaches.” *Journal of Industrial Ecology* 11 (1): 161–179.

Lo, Shih-Chi, Hwong-wen Ma, and Shang-Lien Lo. 2005. “Quantifying and Reducing Uncertainty in Life Cycle Assessment Using the Bayesian Monte Carlo Method.” *Science of The Total Environment* 340 (1–3): 23 – 33. doi:10.1016/j.scitotenv.2004.08.020.

Macknick, Jordan, Robin Newmark, Garvin Heath, and KC Hallett. 2011. *A Review of Operational Water Consumption and Withdrawal Factors for Electricity Generating Technologies*. Technical Report. NREL.

Marcotte, Denis. 1991. "Cokriging with Matlab." *Computers and Geosciences* 17 (9): 1265 – 1280. doi:DOI: 10.1016/0098-3004(91)90028-C.

———. 1995. "Generalized Cross-Validation for Covariance Model Selection." *Mathematical Geology* 27 (5): 659–672.

Marcotte, Denis, and M. David. 1988. "Trend Surface Analysis as a Special Case of IRF-k Kriging." *Mathematical Geology* 20 (7): 821–824. doi:10.1007/BF00890194.

Marriott, Joe, H. Scott Matthews, and Chris T. Hendrickson. 2010. "Impact of Power Generation Mix on Life Cycle Assessment and Carbon Footprint Greenhouse Gas Results." *Journal of Industrial Ecology* 14 (6): 919–928.

Matías, J., A. Vaamonde, J. Taboada, and W. González-Manteiga. 2004. "Comparison of Kriging and Neural Networks With Application to the Exploitation of a Slate Mine." *Mathematical Geology* 36 (4): 463–486.

Matsuno, Yasunari, and Michael Betz. 2000. "Development of Life Cycle Inventories for Electricity Grid Mixes in Japan." *International Journal of Life Cycle Assessment* 5 (5): 295–305.

May, J. R., and D. J. Brennan. 2003. "Application of Data Quality Assessment Methods to an LCA of Electricity Generation." *International Journal of Life Cycle Assessment* 8 (4): 215–225.

McCleese, D. L., and P. T. LaPuma. 2002. "Using Monte Carlo Simulation in Life Cycle Assessment for Electric and Internal Combustion Vehicles." *International Journal of Life Cycle Assessment* 7 (4): 230–236.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Monographs on Statistics and Applied Probability. Chapman and Hall.

Mijuk, Gordana. 2010. "Batterie Europas." *NZZ Am Sonntag*. http://www.nzz.ch/nachrichten/startseite/bat_terie_europas_1.8742799.html.

Milà i Canals, Llorenç, Adisa Azapagic, Gabor Doka, Donna Jefferies, Henry King, Christopher Mutel, Thomas Nemecek, et al. 2011. "Approaches for Addressing Life Cycle Assessment Data Gaps for Bio-based Products." *Journal of Industrial Ecology* 15 (5): 707–725. doi:10.1111/j.1530-9290.2011.00369.x.

Miller, Shelie A., and Thomas L. Theis. 2006. "Comparison of Life-Cycle Inventory Databases: A Case Study Using Soybean Production." *Journal of Industrial Ecology* 10 (1-2): 133–147.

Mongelli, I., S. Suh, and G. Huppes. 2005. "A Structure Comparison of Two Approaches to LCA Inventory Data, Based on the MIET and ETH Databases." *International Journal of Life Cycle Assessment* 10 (5) (September): 317–324.

Moreau, Vincent, Gontran Bage, Denis Marcotte, and Réjean Samson. 2011. “Modelling the Inventory of Hydropower Plants.” In *Towards Life Cycle Sustainability Management*, ed. Matthias Finkbeiner, 443–450. Springer Netherlands. http://dx.doi.org/10.1007/978-94-007-1899-9_43.

———. 2012. “Estimating Material and Energy Flows in Life Cycle Inventory with Statistical Models.” *Journal of Industrial Ecology* 16 (3): 399–406. doi:10.1111/j.1530-9290.2012.00459.x.

Orchard, Terence, and Max A Woodbury. 1972. “A Missing Information Principle: Theory and Applications.” In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of Statistics*, 697 – 715. Berkeley, Calif.: Univ. California Press.

Ord, J. K. 2006. “Kriging.” *Encyclopedia of Statistical Sciences*. John Wiley & Sons.

Peereboom, Eric Copius, Rene Kleijn, Saul Lemkowitz, and Sven Lundie. 1998. “Influence of Inventory Data Sets on Life-Cycle Assessment Results: A Case Study on PVC.” *Journal of Industrial Ecology* 2 (3): 109–130.

Pehnt, M. 2003. “Assessing Future Energy and Transport Systems: The Case of Fuel Cells Part I: Methodological Aspects.” *International Journal of Life Cycle Assessment* 8 (5): 283–289.

Peisajovich, A. 1997. *Etude Du Cycle De Vie De L’electricite Produite Et Transportee Au Quebec*. Direction principale Communication et Environnement, Hydro-Quebec.

Pfister, Stephan, Dominik Saner, and Annette Koehler. 2011. “The Environmental Relevance of Freshwater Consumption in Global Power Production.” *The International Journal of Life Cycle Assessment* 16 (6): 580–591.

Reap, John, Felipe Roman, Scott Duncan, and Bert Bras. 2008. “A Survey of Unresolved Problems in Life Cycle Assessment.” *International Journal of Life Cycle Assessment* 13 (4) (June): 290–300. doi:10.1007/s11367-008-0008-x.

Ribeiro, Flavio de Miranda, and Gil Anderi da Silva. 2010. “Life-cycle Inventory for Hydroelectric Generation: a Brazilian Case Study.” *Journal of Cleaner Production* 18 (1): 44–54. doi:10.1016/j.jclepro.2009.09.006.

Rosa, Luiz Pinguelli, Marco Aurelio dos Santos, Bohdan Matvienko, Ednaldo Oliveira dos Santos, and Elizabeth Sikar. 2004. “Greenhouse Gas Emissions from Hydroelectric Reservoirs in Tropical Regions.” *Climatic Change* 66: 9–21.

Rosenberg, D. M., F. Berkes, R. A Bodaly, R. E. Hecky, Carol A Kelly, and John W. M Rudd. 1997. “Large-scale Impacts of Hydroelectric Development.” *Environmental Reviews* 5 (1): 27–54. doi:doi:10.1139/er-5-1-27.

Rosenberg, D. M., R. A. Bodaly, and P. J. Usher. 1995. “Environmental and Social Impacts of Large Scale Hydroelectric Development: Who Is Listening?” *Global Environmental Change* 5 (2): 127 – 148. doi:DOI: 10.1016/0959-3780(95)00018-J.

Ross, Stuart, David Evans, and Michael Webber. 2002. “How LCA Studies Deal with Uncertainty.” *International Journal of Life Cycle Assessment* 7 (1): 47–52.

Rudd, John W. M, R. Harris, Carol A Kelly, and R. E. Hecky. 1993. "Are Hydroelectric Reservoirs Significant Sources of Greenhouse Gases?" *Ambio* (22): 246–248.

dos Santos, Marco Aurelio, Luiz Pinguelli Rosa, Bohdan Sikar, Elizabeth Sikar, and Ednaldo Oliveira dos Santos. 2006. "Gross Greenhouse Gas Fluxes from Hydro-power Reservoir Compared to Thermo-power Plants." *Energy Policy* 34 (4): 481 – 488. doi:DOI: 10.1016/j.enpol.2004.06.015.

Shibasaki, Maiya, Matthias Fischer, and Leif Barthel. 2007. "Effects on Life Cycle Assessment — Scale Up of Processes." In *Advances in Life Cycle Engineering for Sustainable Manufacturing Businesses*, ed. Shozo Takata and Yasushi Umeda, 377–381. Springer London. http://dx.doi.org/10.1007/978-1-84628-935-4_65.

Soares, Amílcar. 2010. "Geostatistical Methods for Polluted Sites Characterization." In *geoENV VII – Geostatistics for Environmental Applications*, ed. P. M. M. Atkinson and C. D. D. Lloyd, 16:187–198. Quantitative Geology and Geostatistics. Springer Netherlands. http://dx.doi.org/10.1007/978-90-481-2322-3_17.

Soimakallio, Sampo, Juha Kiviluoma, and Laura Saikku. 2011. "The Complexity and Challenges of Determining GHG (greenhouse Gas) Emissions from Grid Electricity Consumption and Conservation in LCA (life Cycle Assessment) – A Methodological Review." *Energy* 36 (12): 6705 – 6713. doi:10.1016/j.energy.2011.10.028.

Sonnemann, Guido W., Marta Schuhmacher, and Francesc Castells. 2003. "Uncertainty Assessment by a Monte Carlo Simulation in a Life Cycle Inventory of Electricity Produced by a Waste Incinerator." *Journal of Cleaner Production* 11 (3) (May): 279–292.

Soumis, Nicolas, Eric Duchemin, Rene Canuel, and Marc Lucotte. 2004. "Greenhouse Gas Emissions from Reservoirs of the Western United States." *Global Biogeochem. Cycles* 18 (3): GB3022.

Soumis, Nicolas, Marc Lucotte, René Canuel, Sebastian Weissenberger, Stéphane Houel, Catherine Larose, and Éric Duchemin. 2005. "Hydroelectric Reservoirs as Anthropogenic Sources of Greenhouse Gases." In *Water Encyclopedia*, 203–210. John Wiley & Sons, Inc. http://dx.doi.org/10.1002/047147844X.sw791.

St Louis, Vincent L., Carol A Kelly, Eric Duchemin, John W. M Rudd, and D. M. Rosenberg. 2000. "Reservoir Surfaces as Sources of Greenhouse Gases to the Atmosphere: A Global Estimate." *Bioscience* 50 (9): 766–775.

Sugiyama, H., Y. Fukushima, M. Hirao, S. Hellweg, and K. Hungerbuehler. 2005. "Using Standard Statistics to Consider Uncertainty in Industry-based Life Cycle Inventory Databases." *International Journal of Life Cycle Assessment* 10 (6) (November): 399–405.

Suh, S., and G. Huppes. 2002. "Missing Inventory Estimation Tool Using Extended Input-Output Analysis." *International Journal of Life Cycle Assessment* 7 (3): 134–140.

Tan, Raymond R., Lee Michael A. Briones, and Alvin B. Culaba. 2007. "Fuzzy Data Reconciliation in Reacting and Non-reacting Process Data for Life Cycle Inventory Analysis." *Journal of Cleaner Production* 15 (10): 944–949.

Tsiatis, Anastasios. 2006. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Berlin: Springer.

Wackernagel, Hans. 2003. *Multivariate Geostatistics : an Introduction with Applications*. 3rd ed. Berlin: Springer.

Weber, Christopher L, Paulina Jaramillo, Joe Marriott, and Constantine Samaras. 2010. "Life Cycle Assessment and Grid Electricity: What Do We Know and What Can We Know?" *Environmental Science and Technology* 44 (6): 1895 – 1901.

Weidema, Bo P. 2000. "Increasing Credibility of LCA." *The International Journal of Life Cycle Assessment* 5 (2): 63–64.

———. 2009a. "Avoiding or Ignoring Uncertainty." *Journal of Industrial Ecology* 13 (3): 354–356.

———. 2009b. "Consequential and Hybrid LCA Presentation" Ecole Polytechnique de Montreal.

Weidema, Bo P., and Marianne Suhr Wesnaes. 1996. "Data Quality Management for Life Cycle Inventories–an Example of Using Data Quality Indicators." *Journal of Cleaner Production* 4 (3-4): 167–174.

Wernet, Gregor, Stefanie Hellweg, Ulrich Fischer, Stavros Papadokonstantakis, and Konrad Hungerbuehler. 2008. "Molecular-Structure-Based Models of Chemical Inventories Using Neural Networks." *Environmental Science and Technology* 42 (17): 6717–6722.

Wernet, Gregor, Stavros Papadokonstantakis, Stefanie Hellweg, and Konrad Hungerbuehler. 2009. "Bridging Data Gaps in Environmental Assessments: Modeling Impacts of Fine and Basic Chemical Production." *Green Chemistry* 11 (11): 1826–1831. doi:10.1039/b905558d.

Williams, Eric D., Christopher L. Weber, and Troy R. Hawkins. 2009. "Hybrid Framework for Managing Uncertainty in Life Cycle Inventories." *Journal of Industrial Ecology* 13 (6): 928–944.

World Commission on Dams. 2000. *Dams and Development: A New Framework for Decision Making*. London, Sterling VA: Earthscan Publications Ltd.

Yamamoto, Jorge. 2005. "Correcting the Smoothing Effect of Ordinary Kriging Estimates." *Mathematical Geology* 37 (1): 69–94.

ANNEXES

A.1 Compléments au premier article *Estimating material and energy flows* [...]

Supplementary material:

The following provides mathematical details for implementing kriging. The results were obtained by programming and running scripts on Matlab. Assume the following model for kriging (1):

$$Z(x)=m(x)+Y(x)=\sum_{l=0}^{L} a_l x^l+Y(x) \quad (1) \quad m(x)=\sum_{l=0}^{L} a_l x^l \quad L=0,1,2 \quad (2)$$

where the x are independent or characteristic variables while the $Z(x)$ are dependent ones, that is energy and material flows. $m(x)$ is a deterministic drift represented by a low order polynomial (2) and $Y(x)$ is the random function of the model. The combination of random and deterministic components in the model accommodates the fact that expected values of material and energy flows are unlikely to remain constant as the characteristic variables of a given object increase or decrease. Indeed, turbines in the tens of kilowatts have not the same embodied materials and energy as turbines in the tens of megawatts, on average. This is known as universal kriging.

The estimator takes the form (3) and is unbiased provided the set of unbiasedness constraints (4) are fulfilled. In other words the sum of kriging weights λ_i should be equal to one.

$$\hat{Z}(x_0)=\sum_{i=1}^{N} \lambda_i Z(x_i) \quad (3) \qquad \sum_{i=1}^{N} \lambda_i=1 \quad (4)$$

where x_0 are the points to be estimated. The minimization of the error variance σ^2 under this constraint is achieved by forming the Lagrangian $L(\lambda, \mu)$ with parameter μ:

$$\begin{aligned} L(\lambda,\mu) &= \sigma^2+2\mu\left(\sum_{i=1}^{N} \lambda_i-1\right) \quad (5) \\ &= Var[Z(x_0)]+\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j Cov[Z(x_i),Z(x_j)]-2\sum_{i=1}^{N} \lambda_i Cov[Z(x_0),Z(x_i)]+2\mu\left(\sum_{i=1}^{N} \lambda_i-1\right) \end{aligned}$$

where *Var*[ ] and *Cov*[ ] are the variance and covariance operators respectively. Solving equation (5), by setting partial derivatives to zero, gives the kriging weights λ_i.

Similarly the cokriging estimator has the following form (6) with unbiasedness constraints on the primary variable (7) and on the secondary variables (8):

$$\hat{Z}(x_0)=\sum_{i=1}^{N_Z}\lambda_i Z(x_i)+\sum_{i=1}^{N_{Y1}}\alpha_i^1 Y_1(x_i)+\ldots+\sum_{i=1}^{N_{Yp}}\alpha_i^p Y_p(x_i) \quad (6)$$

$$\sum_{i=1}^{N_Z}\lambda_i x_i^l=x_0^l \quad \forall \quad l=0,\ldots,L \quad (7) \qquad \sum_{i=1}^{N_{Yj}}\alpha_i^j x_i^l=0 \quad \forall \quad l=0,\ldots,L\,; \quad j=1,\ldots,p \quad (8)$$

where $Z(x)$ is the primary variable and $Y(x)$ the p secondary variables (usually $p = 1$ or 2). The $Y(x)$ additional variables need not be observed at the same points as the primary one. In the case of cokriging, cross-covariances are required to describe the interrelationships between variables. A common simple model is the linear coregionalisation model (LCM) of the form (9):

$$C(h)=B_0\,Nugget(C_0=1)+B_1\,Model(C=1) \quad (9)$$

where B_0 and B_1 are (p+1) squared and symmetric positive definite matrices of coefficients and Model is any valid univariate covariance model (see table 2).

The following parameter values were used used in calculations, note that the range *a* is a constant 3.00:

Kriging and cokriging parameters for figures 1, 2 and table 3

Parameters	Concrete	Lorry	Steel	Welding	Epoxy	kWh	Al	Cu	Cr	Fe	Fiberglass	Steel
C_0	0.14	0.10	0.06	0.01	0.04	0.08	0.08	0.02	0.07	0.08	0.09	0.10
$C_0 + C$	2.62	1.84	1.13	0.13	0.76	1.50	1.46	0.34	1.26	1.52	1.76	1.95
B_0	0.14 0.00 0.00 0.10					0.08 0.00 0.00 0.08						
B_1	2.62 2.19 2.19 1.84					1.50 1.48 1.48 1.46						
B_0	0.14 0.00 0.00 0.06					0.08 0.00 0.00 0.02						
B_1	2.62 1.64 1.64 1.13					1.50 0.54 0.54 0.34						

Cokriging parameters in table 4

Parameter	kWh & Al	kWh & Cu
B_0	0.016 0.000 0.000 0.015	0.016 0.000 0.000 0.004
B_1	1.56 1.54 1.54 1.53	1.56 0.56 0.56 0.36

Ecoinvent data for wind turbines (shaded is tower) :

Rated power	kW	30	150	600	800	2000
Electricity, medium voltage, at grid/CH U	kWh		2	5	10	
Concrete, normal, at plant/CH U	m3	23.2	23.2	81.7	102	872
Diesel, burned in building machine/GLO U	MJ	5070	5070	25800	27900	876
Epoxy resin, liquid, at plant/RER U	kg	37	95	144	360	547
Reinforcing steel, at plant/RER U	kg	2370	2370	11200	14000	80000
Steel, low-alloyed, at plant/RER U	kg	5300	15000	38900	69400	122000
Sheet rolling, steel/RER U	kg	5300	15000	38900	69400	122000
Welding, arc, steel/RER U	m	84	115	152	190	228
Zinc coating, pieces/RER U	m2	74	190			
Zinc coating, pieces, adjustment per um/RER U	m2	1480	3800			
Transport, lorry 20-28t, fleet average/CH U	tkm	10800	21000	16300	22300	162000
Transport, freight, rail/CH U	tkm	4640	10500	70300	117000	43700
Electricity, medium voltage, at grid/CH U	kWh	2	2	10	10	
Electricity, medium voltage, production UCTE, at grid/UCTE U	kWh	575	3980	17500	17500	67500
Aluminium, primary, at plant/RER U	kg	14.7	70.6	204	207	845
Cast iron, at plant/RER U	kg	268	1290	7270	6480	26400
Chromium steel 18/8, at plant/RER U	kg	736	3530	17800	14500	51600
Copper, at regional storage/RER U	kg	240	484	1300	1460	1590
Glass fibre reinforced plastic, polyamide, injection moulding, at plant/RER U	kg	206	2790	7140	9660	41000
Lead, at regional storage/RER U	kg	0.5	0.5	0.5	0.5	0.5
Lubricating oil, at plant/RER U	kg	10	57	50.4	58.8	1150
Polyethylene, HDPE, granulate, at plant/RER U	kg	246	465	603	621	27
Polypropylene, granulate, at plant/RER U	kg	20	20	20	20	
Polyvinylchloride, bulk polymerised, at plant/RER U	kg	164	322	421	434	6
Steel, low-alloyed, at plant/RER U	kg	112	298	2710	3750	15900
Synthetic rubber, at plant/RER U	kg	3.2	15	100	100	100
Tin, at regional storage/RER U	kg	0.5	0.5	0.5	0.5	0.5
Section bar rolling, steel/RER U	kg	380	1680	9920	10200	42200
Sheet rolling, aluminium/RER U	kg	14.7	70.6	204	207	845
Sheet rolling, chromium steel/RER U	kg	736	3530	17800	14400	51600
Wire drawing, copper/RER U	kg	20	481	1300	1460	1590
Transport, lorry 20-28t, fleet average/CH U	tkm	3730	9850	5720	5710	28200
Transport, freight, rail/CH U	tkm	951	4130	49200	47800	28200

A.2 *Statistical estimation of missing data in life cycle inventory: an application to hydroelectric power plants* (version courte)

Cette version de l'article a été publiée sous le titre *Modelling the Inventory of Hydropower Plants* dans *Towards Life Cycle Sustainability Management*, ed. Matthias Finkbeiner, pages 443-450. Springer Netherlands. L'original est disponible sous www.springerlink.com.

Abstract

Life cycle inventories are data intensive by definition and missing data continues to hinder more complete and accurate assessments. This article proposes a statistical approach to address data gaps in life cycle inventories applied to large scale hydroelectric power. The procedure relies on relationships between the technical characteristics of hydropower plants and the material and energy flows necessary throughout the life cycle of such systems. With highly flexible estimators known as kriging, predicting the value of material and energy flows suddenly becomes more accurate. From relatively small sample sizes, kriging allows better estimation without averaging out any of the original data. Similarly, parameter estimation and model validation can be performed through cross validation which assumes very little on the data itself. Mean absolute errors for various forms of kriging and regression show that the former performs better than the latter, more so in cases of incomplete data.

Introduction

Hydroelectric power generation accounts for approximately 16% of world electricity output, with China, Canada and Brazil being the leading producers (IEA 2010). Despite widespread claims of hydroelectricity being a renewable source of power, very few studies on its potential impact exist. Indeed, less than a handful of life cycle assessments (LCAs) have been conducted on hydro power plants (Ribeiro and Silva 2010). Hydroelectric projects depend on numerous local conditions, topography, hydrology, geology as well as factors affecting human populations and the environment. This specificity of hydroelectric power accounts for much of the lack of necessary data, making inventory analysis and impact assessments notoriously difficult. Moreover, new hydropower projects became larger over time, multiplying data

collection efforts and challenging the development of environmental assessment tools (Rosenberg et al. 1997). Generic hydropower plants hardly exist.

Quantifying and qualifying the gains and loses to society and the environment from hydropower projects is a monumental task the World Commission on Dams managed to undertake (World Commission on Dams 2000). Most studies have emphasized the site specific issues which translate into questionable comparisons with other large sources of power. Hence the goal of this article, taking a statistical approach to enable better estimation of life cycle inventory data on hydropower plants and more generally, showing how data gaps in life cycle inventories (LCIs) and associated databases can be alleviated with appropriate statistical estimators. Besides primary sources of data, this article draws upon existing inventories, namely the ecoinvent database and Itaipu assessment (Bauer et al. 2007; Ribeiro and Silva 2010). A description of the hydropower systems follows before the methodological approach is explained. Results emphasize the importance of the construction phase of hydropower and a conclusion ends this article.

System characteristics

The power of hydroelectric plants usually depends on two factors, the water flow and the hydraulic head or height difference between water intake and turbine according to the relation:

$$P = \eta \rho \, g \, Q \, H$$

where P is power (W), η is a coefficient of efficiency, ρ is the density of water (~1000 kg/m3), g is the acceleration due to gravity (~9.8 m/s2), Q is the flow rate (m3/s) and H is the head (m). In practice this is implemented with either low head, high flow (generally run-of-river plants), high head, low flow (generally with reservoirs) or combinations in between. Again this scale varies considerably with respect to site specifications.

Overall, hydroelectric power plants are 82 to 88 % efficient (Bauer et al. 2007). The difference between run-of-river and reservoir plants lies essentially with storage which is one of the key advantages of hydroelectric dams (Egre and Milewski 2002). Electricity cannot be stored such that supply must match demand almost instantaneously. Technically, hydropower plants can

respond and adjust their production within minutes and are therefore well suited for both base and peak load production, particularly with a reservoir.

Despite the specificity of every power plant and reservoir, a set of characteristics was chosen based primarily on data availability. The table below summarizes the characteristic variables of hydroelectric power plants referred to in this article.

System characteristics

Characteristic variables	Ranges	Notes
Type	0/1	Run-of-river/Reservoir
Total installed capacity (MW)	20 – 14000	-
Annual production (GWh/year)	90 – 90000	Average over ≥ 7 years
Hydraulic head (m)	6 – 1400	-
Surface of reservoir (km²)	0 – 4200	-
Volume of reservoir (km3)	0 – 140	-

Methodology

Life cycle inventory, as defined by international standards, requires the quantification of material and energy flows as well as emissions crossing a system's boundaries (ISO 2006). The quality of a LCI directly influences the overall quality of an LCA (Mongelli, Suh, and Huppes 2005). A number of reasons lead us to hypothesize that statistical estimation can overcome both the specificity of hydropower with respect to construction sites and operations and the limitations of current practices dealing with missing data in LCI. Analyzing the inventory of electricity generation tends to yield higher uncertainties when compiled from generic data (May and Brennan 2003).

Moreover, statistical models have been relatively successful in such cases as with the assessment of chemical manufacturing (Wernet et al. 2008). Regression and other linear methods have also been applied to inventory analysis of electricity generation (Hondo 2005). A more

flexible estimator is presented here. Borrowed from spatial statistics, kriging distinguishes itself on several accounts. Clearly no cartesian points exist in LCIs such that the characteristic variables described above become coordinates and the material and energy flows or emissions to be estimated, dependent variables. Kriging is a weighted linear combination of the observations. We assume a model with both a random function and a deterministic drift or low order polynomial. Provided a set of unbiasedness constraints are met, the properties in the table below hold:

Properties of the kriging estimator (adapted from Isaaks and Srivastava 1989)

Properties	Description	Expression
Unbiased estimator	Expected values of estimates and observations are equal	$E\left[\hat{Z}(x_i)\right]=E\left[Z(x_i)\right]$
Exact estimator	No loss of information	$\hat{Z}(x_i)=Z(x_i)$
Screening effect	Weights vary according to the distance from estimates	-
Size and position	Covariance functions C(h) to model observations in space.	$h=\left\|x_i-x_j\right\|$
Smoothing effect	Kriging varies less than the estimated phenomena.	$Var(\hat{Z}(x_i))\leq Var(Z(x_i))$

where E[ ] and Var( ) are the expectation and variance operators respectively. The xi's are characteristic variables and the Z's corresponding material and energy flows. Covariance functions C(h) translate in formal terms the idea that distinct observations close to one another should agree more than if they were far apart. A number of valid covariance functions exist, linear, exponential, spherical, etc. Each function has 3 parameters, the first of which is a nugget effect which captures variations on a small scale and can be understood as an interpolating smoothing parameter (Marcotte and David 1988). Note that since kriging is an exact estimator, it is discontinuous at every observations. The second parameter is a structured variance which equals the total variance when added to the nugget effect. Third, the range of a covariance

function corresponds to the distance h at which the covariance is nil or observations are unrelated.

If deterministic and random components as well as covariance functions impart great flexibility to the kriging estimator, another advantage is multivariate kriging, or cokriging. Multivariate kriging enables the estimation of primary variables using data from secondary variables simultaneously (Wackernagel 2003). All estimations can be performed jointly as the order between primary and secondary variables can be interchanged (Marcotte 1991). For example, if data on explosives is available for the construction of most power plants and dams whereas the excavated volumes are not, joint estimation might provide more accurate results for excavation. In general, primary and secondary variables need not be observed at the same points.

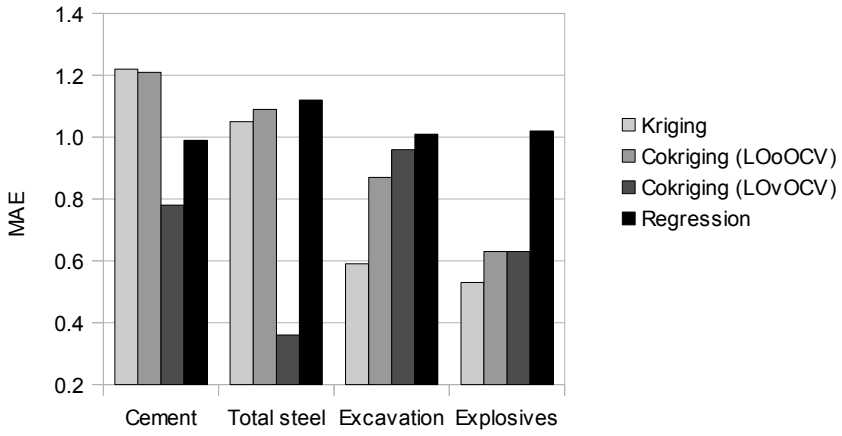
The relatively small sample sizes (here $N = 27$) available for power generation systems would forfeit many attempts to validate a model and its parameters. Cross-validation however, has no prior assumption such as normality and observations take part in both trial and validation sets (Marcotte 1995). Leave-one-out cross-validation (LOOCV) successively removes observations and derives estimates from neighboring values exactly once. Moreover, with cokriging, the values of secondary variables for a given observation can be part of the validation data all together or one after the other. These two options are referred to as Leave-one-observation-out cross-validation (LOoOCV) and Leave-one-variable-out cross-validation (LOvOCV), respectively. Comparing the results of the two validation approaches indicate some of the advantages of cokriging, should the errors calculated from LOvOCV be lower than that of LOoOCV.

Results

Data on the construction, operation and removal of hydropower plants is scarce. A sample from different sources and for widely different power plants was carefully assembled. Emphasis was put on the construction phase, less controversial than reservoir flooding and responsible for most of the overall impact in many cases (Peisajovich 1997; Bauer et al. 2007; Ribeiro and Silva 2010). Time is clearly an important factor to be considered with construction. In particular, Ribeiro and his colleague (2010) have shown the LCI of hydroelectric power plants to be sensitive to time horizons. Optimistically the useful lives of infrastructures and equipments were

adjusted to 100 and 50 years, respectively. If dams and equipment can technically last that long, chances are operations and economics dictate otherwise.

By and large the main material and energy flows by weight or volume involved in hydroelectric power construction and operation are water, cement and concrete, steel of various grades for structures and equipment, explosives, and fuels and lubricants for machinery, transportation and operation. Their provision and the construction phase itself account for an important share of hydropower's total impact, although water use is often ignored. Because of different scales, the data was normalized, dividing variables by their mean values. Mean absolute errors (MAEs) can then be derived from performing leave-one-observation-out cross-validation (LOoOCV). As shown in the figure below comparisons between the different estimators are not straightforward.



Comparison of MAEs for different estimators

The most evident result is the most interesting: the kriging and cokriging errors are lower than that of regression for the two rightmost flows of the chart. This shows not only that kriging performs better in the presence of data gaps but also provides more accurate estimates. Estimating one variable at a time reduces somewhat the MAEs, as the comparison of cokriging errors indicates. Both observations support the original hypothesis that estimators as versatile as kriging are particularly well suited in situations where data is missing and needed.

Does accounting for several variables simultaneously further reduce kriging errors? Besides providing more information from which to estimate missing values, the benefits are not obvious either. In the case of kriging, univariate or multivariate, both characteristics, installed capacity and annual production, enter in the calculation. Regression uses installed capacity only. It appears that the advantage of multivariate over univariate kriging (and regression) are particularly important when data is relatively complete, as shown on the left of the figure above. To the contrary, univariate kriging shows lower errors when data is missing. One explanation is weaker relationships between the different material flows themselves than with the characteristic variables, which is a reason why they are characteristic. Hence the preliminary work necessary to identify appropriate variables describing the properties of a system.

Conclusion

The lack of data affects inventory analysis of many hydropower plants and other systems. In this article, the authors show how material and energy flows involved in the initial phases of hydropower plants can be better estimated based on their design and operating characteristics. The constraints and drawbacks with respect to data collection and data gaps can therefore be loosened, thanks to appropriate and accurate statistical estimators. The versatility of kriging allows for better estimation especially in the absence of complete data. Limitations to this procedure exist, kriging is better suited for interpolation purposes. A minimum sample size (typically $N = 30 – 50$) is necessary to establish covariance functions, such that this experiment used predefined functions and cross-validation to estimate the better model and its parameters. While no prior assumptions on the data are needed, the results of this parameter selection might differ from one data set to the next. Nevertheless, the approach presented above contributes not only to the estimate of missing data but also to much needed representative data which lacks from existing inventory databases.

A.3 Compléments au deuxième article *Statistical estimation of missing data* [...]

Online resources:

Kriging minimizes the estimation variance (2) defined from the estimation errors (1):

$$e = Z(x_0) - \hat{Z}(x_0) \quad (1) \qquad Var[e] = Var[Z(x_0)] + Var[\hat{Z}(x_0)] - 2\,Cov[Z(x_0), \hat{Z}(x_0)] \quad (2)$$

where Var and Cov are the variance and covariance operators respectively. Substituting the expression for kriging in (2), the estimation variance can be rewritten as follows (3) :

$$\sigma_e^2 = Var[Z(x_0)] + \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j Cov[Z(x_i), Z(x_j)] - 2\sum_{i=1}^{N} \lambda_i Cov[Z(x_0), Z(x_i)] \quad (3)$$

To minimize σ_e^2 under the constraint of unbiasedness, the Lagrange method takes the partial derivatives of the Lagrangian L (4) with respect to λ and Lagrange parameter μ.

$$\begin{aligned} L(\lambda, \mu) &= \sigma_e^2 + 2\mu\left(\sum_{i=1}^{N} \lambda_i - 1\right) \\ &= Var[Z(x_0)] + \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j Cov[Z(x_i), Z(x_j)] - 2\sum_{i=1}^{N} \lambda_i Cov[Z(x_0), Z(x_i)] + 2\mu\left(\sum_{i=1}^{N} \lambda_i - 1\right) \end{aligned} \quad (4)$$

With the partial derivatives of L equal to zero, a system of N +1 linear equations with N +1 unknowns is obtained (5) :

$$\begin{cases} \sum_{j=1}^{N} \lambda_j Cov[Z(x_i), Z(x_j)] + \mu = Cov[Z(x_0), Z(x_i)] \quad \forall \quad i = 1, \ldots, N \\ \sum_{j=1}^{N} \lambda_j = 1 \end{cases} \quad (5)$$

The cokriging estimator has the following form (6) with unbiasedness constraints on the primary variable (7) and on the secondary variables (8):

$$\hat{Z}(x_0) = \sum_{i=1}^{N_Z} \lambda_i Z(x_i) + \sum_{i=1}^{N_{Y1}} \alpha_i^1 Y_1(x_i) + \ldots + \sum_{i=1}^{N_{Yp}} \alpha_i^p Y_p(x_i) \quad (6)$$

$$\sum_{i=1}^{N_z} \lambda_i x_i^l = x_0^l \quad \forall \quad l=0,\ldots,L \quad (7) \qquad \sum_{i=1}^{N_{yj}} \alpha_i^j x_i^l = 0 \quad \forall \quad l=0,\ldots,L; \quad j=1,\ldots,p \quad (8)$$

where Z is the primary variable and Y the p secondary variables (usually p= 1 or 2). The Y additional variables need not be observed at the same points as the primary one. In the case of cokriging, cross-covariances are required to describe the interrelationships between variables. A simple model is the linear coregionalisation model (LCM) of the form (9):

$$C(h) = B_0 \, Nugget(C_0 = 1) + B_1 \, Model(C = 1) \quad (9)$$

where B_0 and B_1 are (p+1) squared and symmetric positive definite matrices of coefficients and Model is any valid univariate covariance model such as the ones presented in the table below.

Types of covariance functions

	Type	Expression
1	Nugget	$C(h) = \begin{cases} C_0 \;\; if \;\; h=0 \\ 0 \;\; if \;\; \lvert h \rvert > 0 \end{cases}$
2	Linear	$C(h) = C[1-(h/a)]$
3	Spherical	$C(h) = \begin{cases} C\left[1-1.5(h/a)+0.5(h/a)^3\right] \;\; if \;\; 0 \le \lvert h \rvert \le a \\ 0 \;\; if \;\; \lvert h \rvert > a \end{cases}$
4	Cubic	$C(h) = \begin{cases} C\left[1-7(h/a)^2+8.75(h/a)^3-3.5(h/a)^5+0.75(h/a)^7\right] \;\; if \;\; 0 \le \lvert h \rvert \le a \\ 0 \;\; if \;\; \lvert h \rvert > a \end{cases}$

The parameters of covariance functions *C(h)* are a nugget effect C_0 capturing variations on a small scale, a structured variance C and a range a. The total variance is C_0 + C, the range corresponds to the value of h for which the covariance is nil, when two observations are too far apart to be related.

Power plant	Type	Capacity (MW)	Head (m)	Reservoir surface (km2)	Reservoir volume (10E6m3)	Cement (10E6kg)	Total Steel (t)	Excavation (m3)	Explosives (t)	Sources
Itaipu	1	14000	118	1350	29000	2.48E+03	1.27E+06	5.65E+07		Ribeiro and Silva 2010 and Itaipu Binacional
Robert Bourassa (LG-2)	1	5328	137.16	2835	61715	1.71E+02	5.79E+04	5.38E+06		Own calculations and Peisajovich 1997
LG-3	1	2304	79	2420	60020	2.22E+02	4.13E+04	4.75E+06	1.45E+03	
LG-4	1	2650	116.7	765	19530	1.29E+02	3.64E+04	2.41E+06		
LG-1	0	1368	27.5	70	1228	1.33E+02	5.10E+04	3.10E+06	1.36E+03	
LG-2A	1	1998	138.5	2835	61715	2.63E+01	2.03E+04	3.30E+06		
La Forge-1	1	840	57.3	1288	7777	3.16E+01	2.12E+04	5.40E+06		
La Forge-2	0	310	27.4	260	1840	3.50E+01	1.51E+04	2.70E+06	1.19E+03	
Brisay	1	446	37.5	4275	53790	2.10E+01	1.95E+04	0.00E+00	0.00E+00	
Eastmain-1	1	480	63	603	6940	1.86E+01	6.40E+03	1.64E+06	7.22E+02	
Jean Lesage (Manic-2)	0	1145	70.11	124	35000	2.13E+02	1.69E+04	4.80E+05	1.78E+02	
Manic-5	1	1596	141.8	1950	142000	5.24E+02	1.21E+05	3.31E+05	1.22E+02	
Manic-1	0	184	36.58			8.85E+00	1.20E+04			
Rene Levesque (Manic-3)	0	1244	94.19	236		1.95E+03	4.25E+05			
Outardes-4	1	785	120.55	650	24300	2.13E+00	5.52E+03		2.05E+02	
Outardes-3	0	1026	143.57			8.50E+01	1.19E+04	1.50E+05	5.55E+01	
Outardes-2	0	523	82.3			8.85E+00	5.71E+03		0.00E+00	
Bersimis-1	1	1178	266.7	798	13900	6.73E+02	1.56E+05	1.69E+06	6.24E+02	
Bersimis-2	0	869	115.83			1.95E+02	4.84E+04	1.58E+05	5.83E+01	
Beauharnois	0	1657	24.39			1.46E+02	6.63E+04	5.71E+06	2.11E+03	
Malvaglia-Biasca	1	315.6	644	0.19	4.6	3.80E+01	0.00E+00	2.60E+06	2.60E+03	Own calculations and Bauer et al. 2007
Punt dal Gall-Ova Spin	1	51.6	120	4.71	164.6	1.84E+02	5.71E+03			
Sufers-Baerenburg	1	230.4	306	0.94	17.5	6.00E+00	2.67E+03	1.20E+06		
Limmern-Linthal	1	261	963	1.36	93	2.40E+02	8.86E+03		1.50E+03	
Salanfe-Mieville	1	60	1332	1.85	40	5.90E+01	3.00E+03			
Curnera	1	150	539	0.81	41.1	1.32E+02	4.01E+04			
Ruppoldingen	0	21.1	6.5			1.70E+01	8.31E+03	7.30E+05		

A.4 Compléments au troisième article *Handling data variability and uncertainty* [...]

Hydroelectric reservoir data

Reservoir	Capacity (MW)	Area (Km2)	CH4 (mg/m2/d)	Stdev	CO2 (mg/m2/d)	Stdev	Sources
Tucurui	4240	2430	109.40	135.76	8474.50	2769.74	(dos Santos et al. 2006)
Samuel	216	559	104.00	112.57	7447.00	905.10	
Xingo	3000	60	40.10	14.29	6138.50	5230.47	
Serra da Mesa	1275	1784	113.00	11.31	2644.00	1878.08	
Tres Marias	396	1040	196.25	186.61	1113.50	1255.50	
Miranda	390	50.6	154.15	153.09	4388.00	837.21	
Barra Bonita	140.76	312	20.90	2.40	3985.50	3462.70	
Itaipu	12600	1549	10.70	3.11	170.50	1034.50	
Segredo	1260	82	8.75	1.63	2695.00	2961.36	
Manic 5	1596	1950	15.00	10.00	1170.00	470.00	(Soumis et al. 2005)
Grand Rapids	481	3277	0.58	1.17	624.00	1007.00	(Demarty et al. 2009)
Jenpeg	133.8	1	1.11	1.11	316.00	347.00	
Kettle	1225.2	319	0.01	0.07	514.00	548.00	
McArthur	55.3	100	0.04	0.19	367.00	388.00	
Eastmain 1	480	603	0.77	0.99	2426.00	2287.00	
Riviere des Prairies	48	1	0.49	0.47	665.00	794.00	
La Grande 2	5616	2815	0.14	0.27	661.00	708.00	
Petit Saut	115	365	260.09	84.82	3631.08	863.20	(Delmas et al. 2005)
Lokka	150	418	62.15	80.40	77.00	79.20	(Huttunen et al. 2002)
Porttipahta	35	214	3.65	1.63	65.50	41.72	
FD Roosevelt	6809	306	3.77	1.62	598.04	405.23	(Soumis et al. 2004)
Dworshak	400	37	4.05	3.19	-675.68	459.46	
Wallula	1120	157	14.19	3.92	1076.43	668.79	
Shasta	629	77	10.38	8.60	2909.09	571.43	
Oroville	939	34	4.29	2.71	1558.82	676.47	
New Melones	300	38	7.21	4.29	-394.74	710.53	

A.5 Liste des contributions écrites et orales réalisées dans le cadre de cette thèse

Le tableau ci-dessous reprend les présentations qui découlent directement des travaux reliés à cette thèse. Sans atteindre la même diffusion, le travail exploratoire et la rédaction d'une demande de subventions retenue par le Conseil de recherches en sciences naturelles et en génie du Canada, a permis de démarrer sur des bases solides.

Type	Titre, auteurs	Date
Conférence	The potential of kriging for modeling life cycle inventory data. Moreau V., G. Bage, R. Samson. The Eigth International Conference on EcoBalance, Tokyo. pp.335-336	Décembre 2008
Conférence	Estimating economic flows in life cycle inventory with statistical models. Moreau V., G. Bage, D. Marcotte, R. Samson. Cycle2010, Montreal	Mai 2010
Article	Estimating material and energy flows in life cycle inventory with statistical models. Moreau V., G. Bage, D. Marcotte, R. Samson. *Journal of Industrial Ecology* 16(3) pp. 399-406	Soumis juillet 2010 Accepté août 2011
Conférence	Modeling life-cycle inventories of hydroelectric power plants. Moreau V., G. Bage, D. Marcotte, R. Samson. Sixth International Conference on Industrial Ecology, Berkeley, California	Juin 2011
Article	Statistical estimation of missing data in life cycle inventory: an application to hydroelectric power plants. Moreau V., G. Bage, D. Marcotte, R. Samson. *The International Journal of Life Cycle Assessment*	Soumis août 2011 Refusé novembre 2011 Révisé et soumis février 2012, *Journal of Cleaner Production*
Affiche	Modeling the life-cycle inventory of hydroelectric power plants. Moreau V., G. Bage, D. Marcotte, R. Samson. Life Cycle Management, Berlin, Germany	Août 2011

Chapitre	Modelling the inventory of hydropower plants. Moreau V., G. Bage, D. Marcotte, R. Samson. In *Towards Life Cycle Sustainability Management*, ed. M. Finkbeiner, 443-450. Springer Netherlands.	Août 2011
Article	Handling data variability and uncertainty with kriging : the case of hydroelectric reservoir. Emisions. Moreau V., D. Marcotte, G. Bage, R. Samson. *Environmental and ecological statistics*	Soumis avril 2012