STIC Environnement / 11-13 mai 2011 / Saint Etienne

Session Apprentissage statistique et classification

Un modèle de régression pour données censurées de retombées atmosphériques

Pascaud Aude[1,3], Roustant Olivier[2], Sauvage Stéphane[1,3], Coddeville Patrice[1,3]
[1]Ecole des Mines de Douai, Dépt. Chimie et Environnement, F-59500 Douai, France
[2]Ecole des Mines de Saint-Etienne, LSTI/CROCUS, F-42023 St-Etienne, France
[3]Université Lille Nord de France, F-59000 Lille, France

aude.pascaud@mines-douai.fr

Thèse de doctorat - PROJET SESAME PRIMEQUAL2/PREDIT
Directeur de thèse : Coddeville Patrice
Encadrants : Roustant Olivier - Sauvage Stéphane

Plan de l'exposé

I. Introduction
II. Méthodologie
III. Résultats
IV. Conclusions
V. Perspectives

Cadre de l'étude

Introduction

Transformations dans les phases gazeuse, particulaire et dans les gouttelettes nuageuses

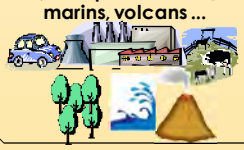
H_2SO_4, HNO_3, NH_4HSO_4 ...

Transport atmosphérique

SO_2, NO_x, NH_3, COV

Emissions

Chauffage, Production d'électricité, Industries, Transports routiers, aérosols marins, volcans ...



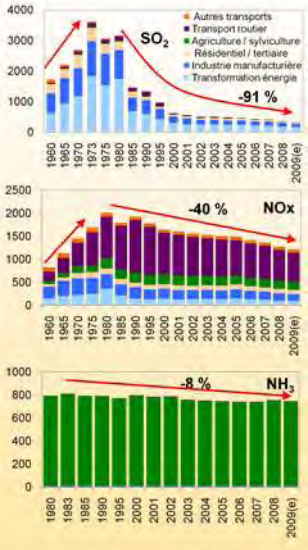
Retombées humides
Retombées sèches

H^+, SO_4^{2-}, NO_3^- NH_4^+

Ecosystèmes terrestre et aquatique

Effets sur la faune et la flore

Constat : Émissions Françaises



Réduction très importante SO_2 depuis 1980

Réduction moyenne NOx depuis 1990

Réduction très faible NH_3 depuis 1980

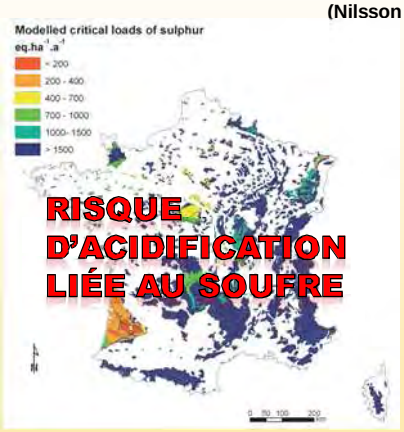
France	SO_2	NOx	NH_3
Estimations 2009	324	1215	746
Göteborg objectifs 2010 (Kt)	400	860	780
Directive 2001/81CE objectifs 2010 (Kt)	375	810	780

(CITEPA, 2010)

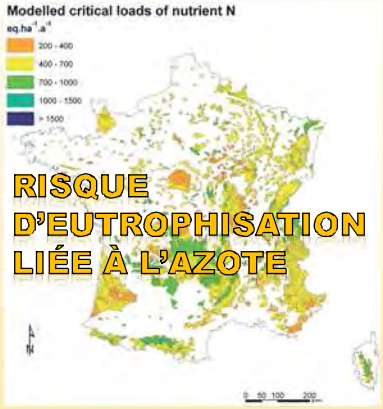
Constat : des écosystèmes sensibles

charges critiques

« valeur d'exposition à un ou plusieurs contaminants en dessous de laquelle des effets significatifs sur des composantes sensibles de l'écosystème n'apparaissent pas, en l'état actuel des connaissances »
(Nilsson et Grennfelt 1988)



Zones sensibles — Zones peu sensibles

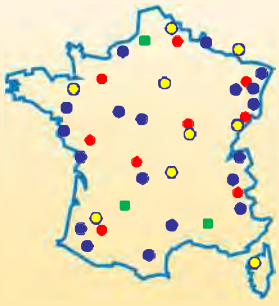


Zones sensibles — Zones peu sensibles

(Probst et Leguédois, 2008)

Observatoires de retombées atmosphériques

Dispositif MERA | Sous-réseaux CATAENAT | Réseau VAG



	Nombres de sites	Période d'observation
MERA	8	13 ans
CATAENAT	27	16 ans
BAPMON	3	31 ans

Trois observatoires ≠ modes de fonctionnement

Fréquences de prélèvements / Types de préleveurs / Méthodes d'analyses

Assurance qualité données

Suivi fonctionnement / Modes opératoires / Traçabilité échantillons / Intercomparaison

PROJET SESAME – PRIMEQUAL2 /PREDIT

Objectifs

BD : 20 ans chimie pluies 3 observatoires {MERA, CATAENAT, VAG}

- ***Analyse préalable*** *: permettre la comparaison des bases*
- ***Tendances – Prévisions*** *: relier les stratégies de réduction d'émissions et les évolutions dans les retombées atmosphériques*
- ***Approche source-récepteur*** *: déterminer des profils chimiques et des contributions des sources d'émissions et les localiser*
- ***Approche géostatistique*** *: spatialiser les concentrations et les dépôts associés aux retombées atmosphériques, possibilité de relier avec l'analyse temporelle.*
- ***Développement méthodologique des charges critiques*** *: améliorer l'estimation et définir des cartes de dépassements selon les scénarios et les résultats obtenus par les étapes précédentes.*

Points +

Base de données consistante
Approches multidisciplinaires et multipartenaires (EMD, EMSE, ECOLAB, ONF, METEO FRANCE)

Analyse préalable : comparabilité

OBJECTIFS

fort niveau de confiance à la donnée
-> données suspectes
limite de détection de la technique analytique
-> données censurées

Pouvoir détecter les valeurs suspectes

Définir des valeurs de remplacement des données censurées pour construire des moyennes mensuelles

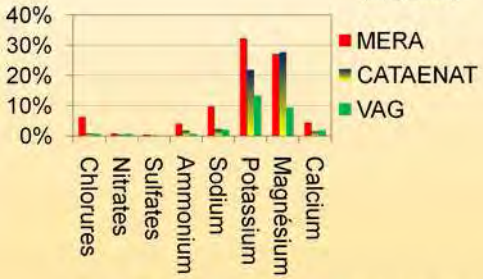
Méthodes de Remplacement

METHODOLOGIE

	Effectif faible	Effectif important	Effectif très important
Utilisation exploratoire	LD/2 (si qq échantillons < LD)	LD/2 (si <15% des échantillons < LD)	Méthode de Cohen (distribution normale) Méthode Kaplan Meier (autres distributions)
Utilisation publication	Kaplan Meier	Kaplan Meier	Méthode de Cohen (distribution normale) Méthode Kaplan Meier (autres distributions)
Utilisation réglementaire	Utilisation publication	Kaplan Meier	Kaplan Meier

MERA → LD/2
CATAENAT → 0
VAG → LD

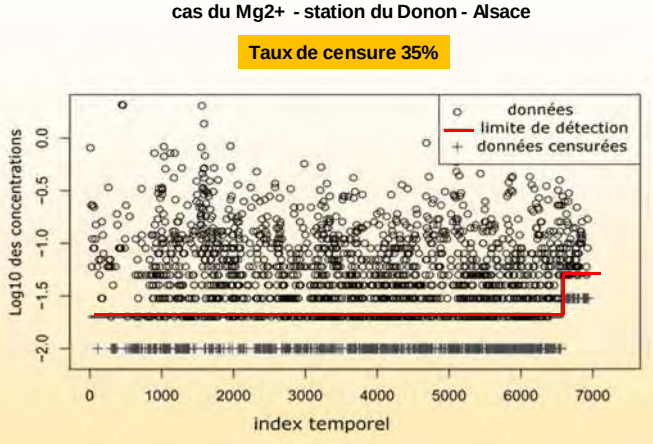
(EPA 2009)



Limites de détection

cas du Mg2+ - station du Donon - Alsace

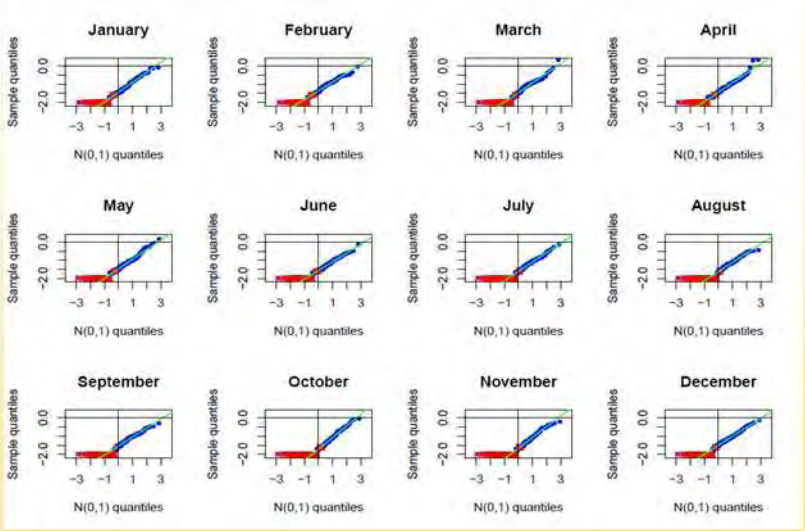
Taux de censure 35%



EVOLUTION TECHNIQUE DES MÉTHODES ANALYTIQUES

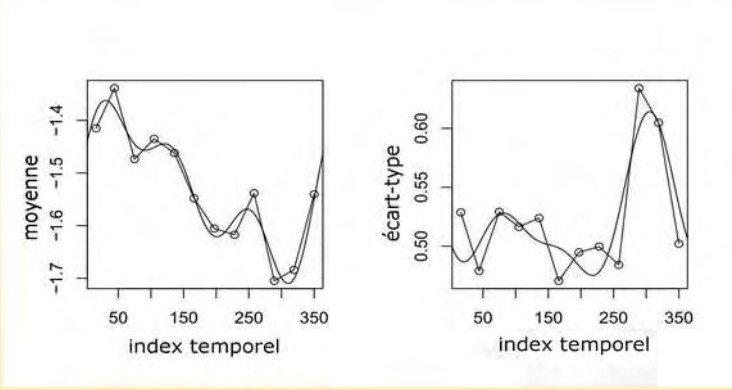
Log-normalité

cas du Mg2+ - station de Donon - le Bas-Rhin (Alsace)



Saisonnalités

cas du Mg2+ - station de Donon - le Bas-Rhin (Alsace)



μ et σ sont ici estimés par les ordonnées à l'origine et les pentes des diagrammes précédents

Conclusion et modélisation

Soient $y_1 = \log_{10}(C_1), \ldots, y_n = \log_{10}(C_n)$

- Si $y(t) > L(t)$, $y(t)$ provient d'une loi $N(\mu(t), \sigma(t)^2)$

Dans la suite on fait l'hypothèse que ce résultat est encore valable lorsque $y(t) < L(t)$

modèle de régression linéaire tronqué

- généralisation du modèle de Tobit, 1958,
- et de celui de Liu et al, 1997, utilisé en chimie environnementale

Modèle de régression tronqué

Mathématiquement, on a :

$$y(t) = y^*(t) \quad si \quad y^*(t) >= L(t)$$
$$y(t) = L(t) \quad si \quad y^*(t) < L(t)$$

avec :

$$y^*(t) = \mu(t) + \sigma(t)e(t), \quad t=1,\ldots,n$$

et $\mu(t) = x(t)'\beta$, $\sigma(t)=z(t)'\gamma$ (>0)

e(1), …, e(n) sont des v.a. indépendantes de loi N(0,1)
β et γ sont des paramètres à estimer
x(t) and z(t) sont des variables explicatives

Choix des variables explicatives

- Endogènes :

– Tendance: monotone, affine par morceaux, …
– Saisonnalités: sous la forme de premiers termes d'une série de Fourier

par exemple : $\mu(t) = b_0 + b_1 t + b_2 \cos(wt) + b_3 \sin(wt)$

- Exogènes:

– Emissions des polluants, variables météorologiques …

Valeurs suspectes et imputation

- **Repérage de valeurs suspectes réalisé avec un calcul d'un quantile d'ordre élevé**

  Seuil utilisé : $\mu(t) + \Phi^{-1}(0.998)*\sigma(t)$

- **Remplacement d'une valeur censurée par une valeur (imputation simple), obtenue par prévision, en généralisant la formule de Liu [Liu et al., 1997] au cas hétéroscédastique.**

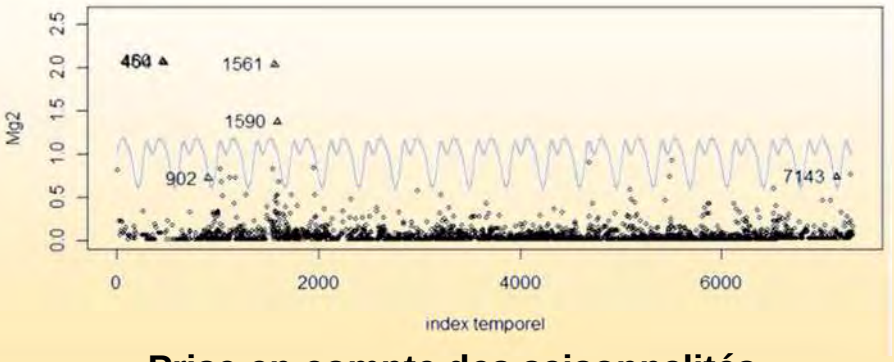
- **Possibilité de reconstituer toute la distribution des données censurées par simulation**

CAS D'UNE SÉRIE TRÈS CENSURÉE

APPLICATIONS

Repérage de valeurs suspectes

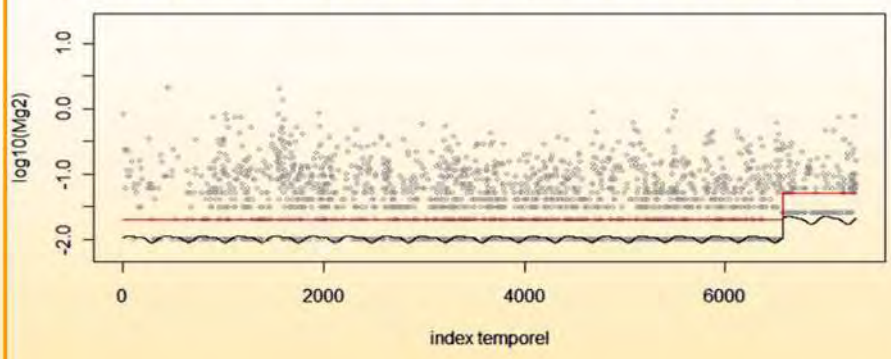
cas du Mg2+ - station de Donon - le Bas-Rhin (Alsace) n=2851 taux censure =35%



Prise en compte des saisonnalités
-> repérage des données 902 et 7143

CAS D'UNE SÉRIE TRÈS CENSURÉE

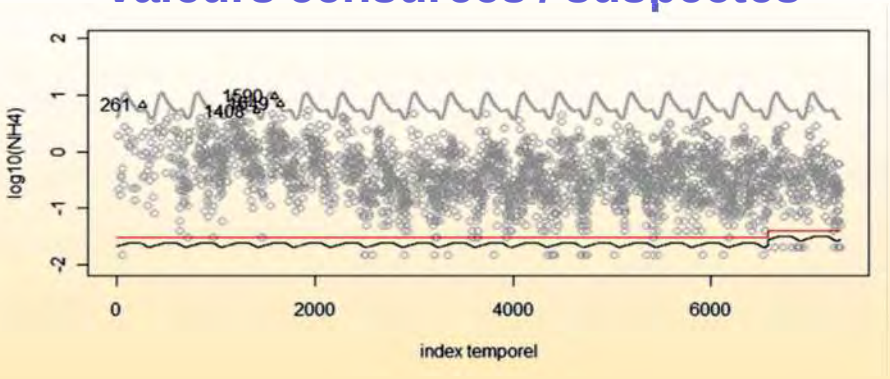
Remplacement de valeurs censurées



Le résultat de l'imputation simple est ici voisin de LD/2

CAS D'UNE SÉRIE PEU CENSURÉE

valeurs censurées / suspectes



Le résultat de l'imputation simple est ici plus proche de LD

CONCLUSION

- **détection saisonnière et automatique des valeurs suspectes**
- **valeur de remplacement saisonnière et moins arbitraire que LD/2**

  -> méthodologie appliquée au calcul de moyennes mensuelles
- **modèle adapté à l'incorporation de tendances (endogènes ou exogènes), prise en compte de la saisonnalité, de l'hétéroscédasticité des données**

PERSPECTIVES

- calcul d'indicateurs statistiques
  -> nécessite souvent de reconstituer la distribution des données censurées
- définition de tendances et choix de variables exogènes
  -> exemple : la hauteur de pluie
- comparaison avec d'autres méthodes

REMERCIEMENTS

- Mireille Batton-Hubert et Djamel Mimoun, ainsi que les participants au séminaire interne EMSE du 20/01/2011 pour leur remarques constructives
- Tous les opérateurs des dispositifs de surveillance
- Ministère chargé de l'environnement et l'ADEME pour le financement du projet SESAME dans le cadre de PRIMEQUAL/PREDIT

MERCI DE VOTRE ATTENTION !!

REFERENCES

CITEPA, "Les émissions dans l'air en France métropole – Substances relatives à l'acidification, l'eutrophisation et à la pollution photochimique." *Centre Interprofessionnel d'Etudes de la pollution Atmosphérique*, 17pp, 2009.

Liu S., Lu J.C., Kolpin D., and Meeker W., Analysis of environmental data with censored observations. Environmental Science & Technology, 31:3358–3362, 1997.

Probst A., Leguédois S. French National Focal Center report In: J-P. Hettelingh, M. Posch, J. Slootweg (eds.) Critical Load, Dynamic Modelling and Impact Assessment in Europe, *CCE Status Report 2008*, Coordination Center for Effects, RIVM, Bilthoven, pp. 134−140, 2008.

Tobin J., Estimation of relationships for limited dependent variables, *Econometrica*, vol. 26, p. 24-36, 1958.

U.S. Environmental Protection Agency, *Air toxics data analysis workbook*, disponible sur http://www.epa.gov/ttnamti1/files/ambient/airtox/workbook/AirToxicsWorkbook6-09.pdf, 2009. dernière consultation 11/05/2011