Une approche paramétrique de la régression linéaire floue - Formalisation par intervalles

Amory Bisserier

THESE

présentée à

l'UNIVERSITE DE SAVOIE

pour obtenir

le grade de DOCTEUR

discipline : STIC - Traitement de l'Information

par

M. Amory BISSERIER

Une approche paramétrique de la régression linéaire floue - Formalisation par intervalles

---

Soutenue le 9 juillet 2010

Préparée au LISTIC et encadrée par : Pr. Sylvie Galichet
Dr. Reda Boukezzoula

COMPOSITION DU JURY

Mme. Mylène MASSON	Rapporteur
Mr. Didier DUBOIS	Rapporteur
Mr. Luc JAULIN	Président
Mme. Sylvie GALICHET	Directeur de thèse
Mr. Reda BOUKEZZOULA	Codirecteur de thèse

THESE

présentée à

l'UNIVERSITE DE SAVOIE

pour obtenir

le grade de DOCTEUR

discipline : STIC - Traitement de l'Information

par

M. Amory BISSERIER

Une approche paramétrique de la régression linéaire floue - Formalisation par intervalles

---

Soutenue le 9 juillet 2010

Préparée au LISTIC et encadrée par : Pr. Sylvie Galichet
Dr. Reda Boukezzoula

COMPOSITION DU JURY

Mme. Mylène MASSON	Rapporteur
Mr. Didier DUBOIS	Rapporteur
Mr. Luc JAULIN	Président
Mme. Sylvie GALICHET	Directeur de thèse
Mr. Reda BOUKEZZOULA	Codirecteur de thèse

Remerciements

Cette thèse est l'aboutissement d'un travail réalisé au sein du Laboratoire d'Informatique, Systèmes et Traitement de l'Information et de la Connaissance (LISTIC) de l'Université de Savoie. Je remercie donc tout naturellement ces deux institutions m'ayant accueilli dans leurs locaux. Je remercie tout particulièrement P. Bolon, directeur du LISTIC durant ces années, pour son écoute et sa disponibilité envers les doctorants, ainsi que le personnel administratif, Joelle et Samia, dont le travail de l'ombre est essentiel.

Je tiens à exprimer mes remerciements à Mme Masson, Mr Dubois pour avoir accepter de rapporter mes travaux, ainsi que Mr Jaulin pour avoir présidé le jury. Leurs remarques, suggestions et discussions lors de la soutenance ont contribué à faire de cette ultime étape un moment extrêmement intéressant.

Je tiens également à exprimer ma profonde gratitude et mes remerciements les plus sincères à mes encadrants, Sylvie Galichet et Reda Boukezzoula. Leurs connaissances, leurs compétences ainsi que leur grande rigueur scientifique m'ont permis de bénéficier d'apports inestimables durant la thèse. Cela reste un sacré challenge que de les satisfaire ! Que Sylvie sache par ailleurs que les discussions variées que nous avons pu avoir en diverses occasions ont toujours été d'agréables instants, toujours motivants.

Je remercie bien entendu mes collègues du laboratoire, et tout particulièrement Lionel Valet, toujours prompt à partager son expérience et son vécu. Comment ne pas ici mettre en avant messieurs Olivier Passalacqua, Grégory Païs, Nabil Fakhfakh et Florent Martin, avec qui je partageais le fameux bureau B206. Que les deux premiers sachent que si leurs goûts musicaux restent discutables, voire scandaleux, le fait de les partager avec eux a toujours été l'occasion de moments très divertissants !

D'un point de vue plus personnel, je voudrais ici remercier chaleureusement mes parents, sans qui je n'aurais pu faire toutes ces études. Je leur serai éternellement reconnaissant de leurs efforts. Je n'oublie évidemment pas mes amis, qui m'ont toujours soutenu et encouragé : Hélène et Benoît et leur fameux cake, Aïnoa, Aurélien, Marie, Olivier, Dwarffy, Zoreil, Romain, Ben, Solange, Olive, Ad, Yann, Lolotte, Sébastien, et j'en oublie...

Mes derniers remerciements, mais non les moindres, iront à Marion. Elle en connait les raisons !

Table des matières

Introduction

De tous temps, l'Homme a cherché à comprendre le monde qui l'entoure. Sans doute est-ce même un des principes fondamentaux de l'humanité lui ayant permis d'évoluer, sinon de progresser. Une illustration de cette volonté d'appréhender ce fonctionnement global se trouve dans l'orientation actuelle de la physique théorique, et plus particulièrement la physique des particules. En effet, les tentatives se multiplient pour développer une "théorie du Tout" visant à unifier les interactions fondamentales présentes dans l'Univers. L'objectif est simple : mettre en équations pour comprendre.

Ainsi, l'objectif de cette théorie en particulier, comme de toute autre en général, est de lier un certain nombre de variables, de constantes, au sein d'un modèle mathématique. La validité de ce modèle doit par la suite être éprouvée par des expériences, pour le conforter, l'enrichir, ou le réfuter le cas échéant. Le comportement de tout système ainsi modélisé, dénommé "boîte blanche" en théorie des systèmes, est donc parfaitement connu.

Cependant, dans de nombreux domaines, sciences de l'ingénieur, économie, informatique, sciences humaines, ..., il peut être difficile de connaitre un modèle physique explicitant les mécanismes de fonctionnement d'un système. Ce dernier, alors dénommé "boîte noire", ne peut donc être appréhendé qu'au travers de ses réactions (communément appelées sorties) à des sollicitations extérieures (dénommées entrées).

Si l'ensemble des observations des interactions entre les entrées et les sorties d'un système permet de caractériser son comportement, il n'en reste pas moins que cette approche n'offre pas la possibilité d'une exploitation aisée. Une solution à cela est de chercher à identifier un modèle mathématique à partir de ces observations. La finalité de ce modèle est donc de représenter les propriétés ou le comportement du système de manière synthétique à partir d'un ensemble de mesures.

Dans cette optique, de très nombreuses techniques concernant l'identification des systèmes [41] ont été développées : machine learning [47], data mining [60], techniques régressives... Ces dernières, sur laquelle nous focalisons notre attention, sont elles-mêmes composées d'un certain nombre de grandes familles distinctes, selon qu'une approche paramétrique ou non, linéaire

ou non, ... , soit considérée. L'approche paramétrique linéaire est particulièrement appréciée car elle permet l'obtention de modèles facilement interprétables, du fait du nombre restreint de paramètres les définissant. Cette approche reste néanmoins performante, du fait du grand nombre de modèles (polynomiaux, multilinéaires, dynamiques, ...) qu'il est possible d'identifier. Il est donc tout à fait compréhensible que cette approche ait suscité un intérêt important. Ainsi, de nombreuses techniques (moindres carrés ordinaires ou généralisés, régression quantile, maximum de vraisemblance, ...) ont été développées dans ce cadre de régression paramétrique linéaire.

Toutes ces techniques ont un point commun : elles reposent sur le postulat que les observations issues du système à modéliser sont précises, quoique éventuellement soumises à variabilité, et qu'il en est de même du modèle obtenu. Or, dans de nombreuses situations, les observations peuvent être imprécises, si elles résultent de mesures ou de protocoles expérimentaux manquant de fiabilité. Le système en lui-même, et plus précisément ses paramètres, peuvent être également imprécis, reflétant ainsi des phénomènes liés par exemple aux conditions limites ou aux tolérances.

Dans cette thèse, nous nous plaçons dans une démarche de construction inductive d'un modèle générique à partir des informations observées imprécises. Autrement dit, on s'intéresse plus particulièrement à une problématique de modélisation. Notre approche peut être assimilée à une méthodologie d'apprentissage, souvent exploitée dans le domaine de l'intelligence artificielle, ou encore à une stratégie d'inférence paramétrique, très populaire dans le domaine des statistiques.

Sachant que décrire précisément le comportement d'un système imprécis est illusoire, une approche commune dans ce cadre est de considérer son modèle comme l'association d'un modèle nominal précis auquel on attache des incertitudes. L'analyse du modèle est alors abordée de façon à étudier sa performance et sa robustesse vis-à-vis de ces imprécisions. Une autre approche, objet de ces travaux, vise à intégrer les imprécisions dans le modèle. Ces incertitudes sont alors considérées comme une caractéristique intrinsèque du modèle.

Dans ce contexte, deux grandes problématiques peuvent être soulevées :

- Comment est-il possible de représenter des imprécisions dans le modèle ?
- Comment peut-on gérer ces imprécisions au cours du processus d'identification ?

Concernant la représentation des imprécisions, nous nous intéressons aux ensembles flous introduits par Zadeh [74], particulièrement bien adaptés pour formaliser des informations imprécises [58], et les gérer à l'aide d'opérateurs dédiés. Afin de s'affranchir des inconvénients liés à ces opérateurs basés sur le principe d'extension et dans l'optique de répondre à la seconde problématique, nous faisons le choix de manipuler les ensembles flous au travers de l'arithmétique

des intervalles [44]. Dans ce contexte, la problématique essentielle de ces travaux est de chercher à identifier un modèle imprécis sur des données imprécises, en gérant, au travers d'une formalisation par intervalles, les ensembles flous modélisant ces imprécisions tout au long du processus d'identification.

Plan du mémoire Les travaux présentés dans ce mémoire tentent de répondre aux objectifs précédemment exposés. Pour ce faire, trois chapitres ont été introduits.

Le chapitre I nous permettra dans un premier temps de justifier le choix de modèles régressifs paramétriques linéaires. Dans ce contexte, en introduisant les concepts liés aux nombres flous et à leur manipulation au travers de l'arithmétique des intervalles, les modèles linéaires flous seront alors formalisés. Une fois les techniques fondamentales d'identification de tels modèles présentées, elles seront discutées au travers d'une analyse critique mettant en évidence leurs limites. Ces dernières concerneront essentiellement deux points fondamentaux, à savoir l'imprécision des modèles flous identifiés, et la relation d'inclusion entre observations et sorties produites par ces modèles.

Dans le chapitre II, les propositions faites en vue de remédier à ces limites seront introduites. Ces améliorations concerneront dans un premier temps la structure des modèles linéaires flous considérés, au travers de la nature de leurs paramètres mais aussi de la forme de leur modèle mathématique. Ensuite, la technique d'identification en elle-même sera étudiée, en vue d'en proposer une version revisitée. Chacun de ces points seront présentés et illustrés dans un premier temps dans le cas de modèles linéaires simples. Par conséquent, une généralisation, ainsi que diverses extensions, en seront proposées en fin de chapitre.

Le chapitre III aura pour objectif d'illustrer les champs d'applications possibles de la méthode régressive proposée. Ainsi, les identifications de modèles polynomiaux, multilinéaires et dynamiques seront successivement abordées. Il sera ainsi possible d'étudier quelques problématiques, que ce soit au niveau du choix de la forme analytique des modèles ou encore de l'influence des données d'identification sur la qualité des modèles obtenus. Par ailleurs, l'application illustrant l'identification de modèles dynamiques permettra de mettre en évidence la nécessité de considérer des modèles imprécis plus généraux, ainsi qu'une présentation critique des travaux préliminaires menés à ce sujet.

Chapitre 1

La régression dans un environnement imprécis : positionnement

Sommaire

1.1 Introduction

De manière générale, la régression est présentée comme étant une méthode statistique utilisée pour analyser la relation d'une variable par rapport à une ou plusieurs autres. Cette relation est définie au travers d'un modèle régressif, dont les paramètres sont identifiés lors d'un processus d'ajustement mathématique du modèle par rapport aux données observées.

Cette définition ne présente la régression que comme un outil statistique, hors de tout contexte. Il peut cependant être utile d'en avoir une vue plus contextuelle (figure 1.1). Considérons un expert d'un domaine scientifique quelconque, dont l'objectif est d'analyser des

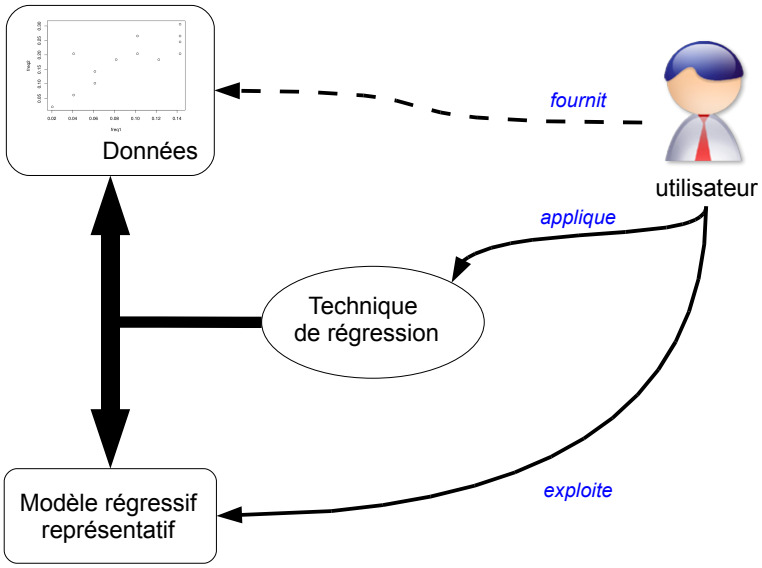


Fig. 1.1: Une vue systémique de la régression

données collectées mises à sa disposition, suite à des mesures ou à des observations. Cette analyse est potentiellement difficile à mener en l'état, notamment dans le cas où une représentation synthétique des données, par exemple graphique, n'est pas disponible. Ainsi, pour initier le processus d'analyse, cet expert est souvent amené à utiliser une technique de régression lui permettant d'identifier un modèle représentatif des données. L'interprétation de ce modèle, son exploitation, devra donc lui permettre de poursuivre plus efficacement l'analyse des données, et à terme de tirer des conclusions constructives.

Il est important de remarquer ici que le terme utilisateur employé dans la figure 1.1 met en avant le fait que l'expert d'un domaine scientifique quelconque devient un simple utilisateur de l'analyse régressive. L'utilisation de la technique de régression mise à disposition ne doit donc pas nécessiter une expertise particulière dans le domaine des techniques régressives. Ce point est

particulièrement important, et impose un certain nombre de contraintes dans la définition que nous retiendrons d'une technique régressive.

D'un point de vue utilisateur, les actions à réaliser doivent être limitées à la fourniture des données à analyser, à l'application d'une technique régressive sur celles-ci, et enfin à l'exploitation du modèle identifié en résultant. Il n'est donc pas souhaitable d'impliquer l'utilisateur dans le processus de développement ou de mise au point de l'outil de régression.

La technique de régression est donc vue comme un outil mis à disposition de l'utilisateur. Par conséquent, sa mise en oeuvre doit être la plus simple possible, et son fonctionnement transparent, toujours dans une optique de simplification du problème d'analyse des données.

Le dernier point concerne le modèle identifié à l'aide de la technique de régression mise à disposition de l'utilisateur. Son exploitation doit elle aussi être relativement aisée et l'interprétation des caractéristiques du modèle identifié doit faciliter une analyse pertinente des données initiales.

La définition contextuelle d'un outil régressif, à l'origine des différentes contraintes à prendre en compte dans la suite de cette étude, a été présentée dans le cas (que nous qualifierons de conventionnel dans la suite) où aucune imprécision des données observées n'est explicitement prise en compte. Cependant, il est possible d'étendre cette vision de la régression à un cadre dans lequel les imprécisions ne sont plus négligées (figure 1.2). Celles-ci interviennent potentiellement à plusieurs niveaux. Bien évidemment, les données mesurées à analyser sont les plus sujettes aux imprécisions, suite à des mesures ou encore des protocoles expérimentaux manquant de fiabilité. Dans ce cas, le modèle représentatif des données doit prendre en considération ces imprécisions, afin que l'utilisateur puisse en tenir compte dans son interprétation et son exploitation. Par conséquent, une méthode régressive utilisée dans un environnement imprécis doit être adaptée à cet environnement afin de fournir un modèle régressif pertinent en adéquation avec les attentes de l'utilisateur.

Cette vision contextuelle de la régression dans un environnement imprécis met en lumière le fait que les contraintes définies dans un cadre conventionnel doivent encore être respectées. Si l'on considère que les données fournies par l'utilisateur sont entachées d'imprécision, la technique d'identification doit être adaptée à l'identification d'un modèle régressif représentatif de ces données, et donc de l'imprécision qui leur est associée. Dans cette optique, nous retiendrons le formalisme des ensembles flous introduits par Zadeh [74] afin de formaliser les imprécisions à modéliser, aussi bien au niveau des données que des modèles représentatifs. En effet, selon [58], cette théorie est particulièrement adaptée pour représenter des informations dont la valeur est imprécise, sans hypothèse forte sur la nature des imprécisions.

A la vue de toutes ces considérations préliminaires, il s'agira dans ce chapitre de déterminer dans un premier temps quelle approche de technique régressive semble être la mieux à même de satisfaire les contraintes que nous nous imposons. Cette étude se fera aussi bien dans un cadre

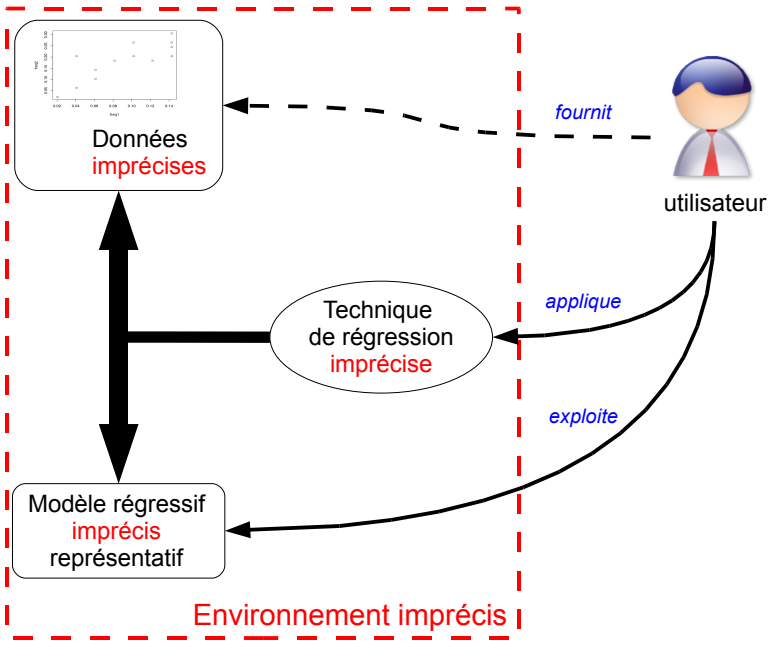


FIG. 1.2: Une vue systémique de la régression dans un environnement imprécis

conventionnel que dans un environnement imprécis. Une fois cet axe d'étude dégagé, une analyse plus fine et critique des outils régressifs existants gérant les imprécisions sera menée.

1.2 Les différentes approches de régression du point de vue de l'utilisateur

1.2.1 Régression non paramétrique ou paramétrique ?

Une des premières distinctions à faire dans les différentes techniques de régression concerne l'aspect paramétrique ou non de celles-ci [59]. Cette distinction est particulièrement importante dans le cadre de la vision contextuelle que nous avons adoptée d'une technique régressive (section 1.1).

Dans le cas d'une approche non paramétrique, la technique de régression que l'utilisateur va appliquer sur le jeu de données à analyser va lui retourner la structure du modèle (que l'on peut assimiler à la forme générale du modèle), ainsi que les paramètres correspondant permettant une représentation optimale des données (figure 1.3). Dans le cas d'une approche paramétrique, l'utilisateur va appliquer une technique de régression sur les données, en ayant déterminé a priori la structure du modèle lui semblant le plus pertinent pour représenter les données. Ainsi,

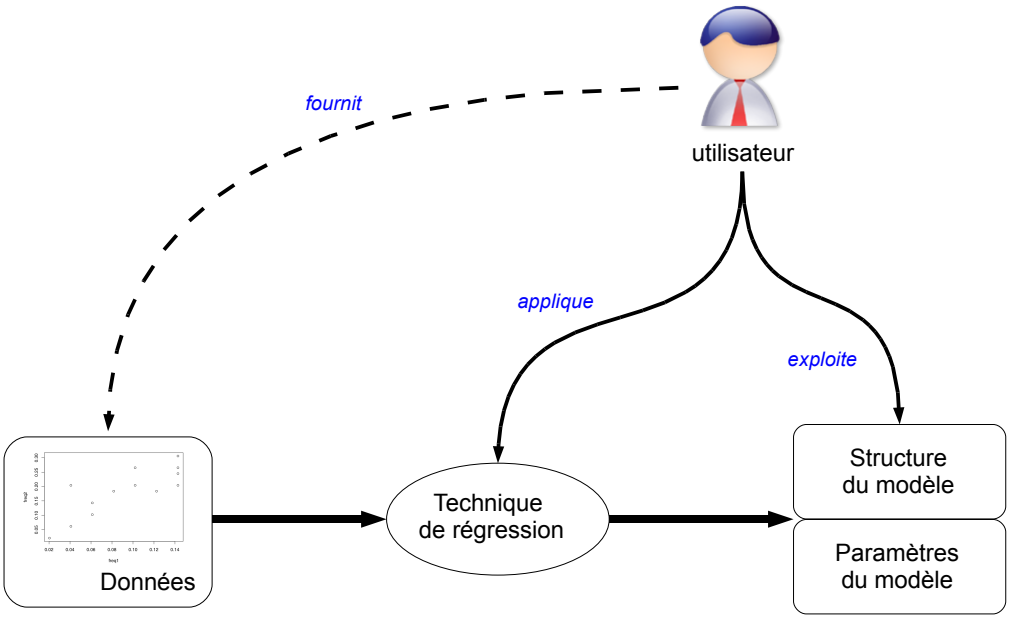


FIG. 1.3: Une vision systémique de la régression non paramétrique

l'outil régressif va retourner la valeur des paramètres présents dans la définition mathématique du modèle imposée par l'utilisateur (figure 1.4).

La différence fondamentale entre ces deux approches se situe donc au niveau du rôle de l'utilisateur lors de l'application de la technique régressive sur les données à étudier, ainsi que sur l'exploitation éventuelle des modèles obtenus (figure 1.5).

Dans une approche paramétrique, l'utilisateur doit définir la forme mathématique du modèle à identifier. Ainsi, cela lui permet d'avoir une action de contrôle sur le nombre de paramètres de ce modèle, paramètres qui seront donc les inconnues à déterminer à l'aide de la technique régressive. Il lui sera donc possible de chercher à réduire le nombre de ces paramètres, en vue de simplifier le modèle, et donc de faciliter son exploitation une fois la régression réalisée (Takezawa [59]). Cette action de contrôle de la complexité du modèle, donc du nombre des paramètres n'est pas possible dans une approche non paramétrique, puisque c'est la meilleure représentation des données qui va être recherchée, quelle que soit la complexité du modèle ainsi obtenu. Ainsi, les approches non paramétriques fournissent un modèle qui ne peut pas être décrit par un petit nombre de paramètres. La seule interprétation possible de ce modèle par l'utilisateur ne pourra donc se faire qu'au travers d'une représentation graphique.

Bien évidemment, le fait que l'utilisateur doive déterminer la forme mathématique du modèle à identifier dans une approche paramétrique peut aussi être vu comme un inconvénient. En effet, le modèle spécifié par l'utilisateur peut être inapproprié pour représenter au mieux les

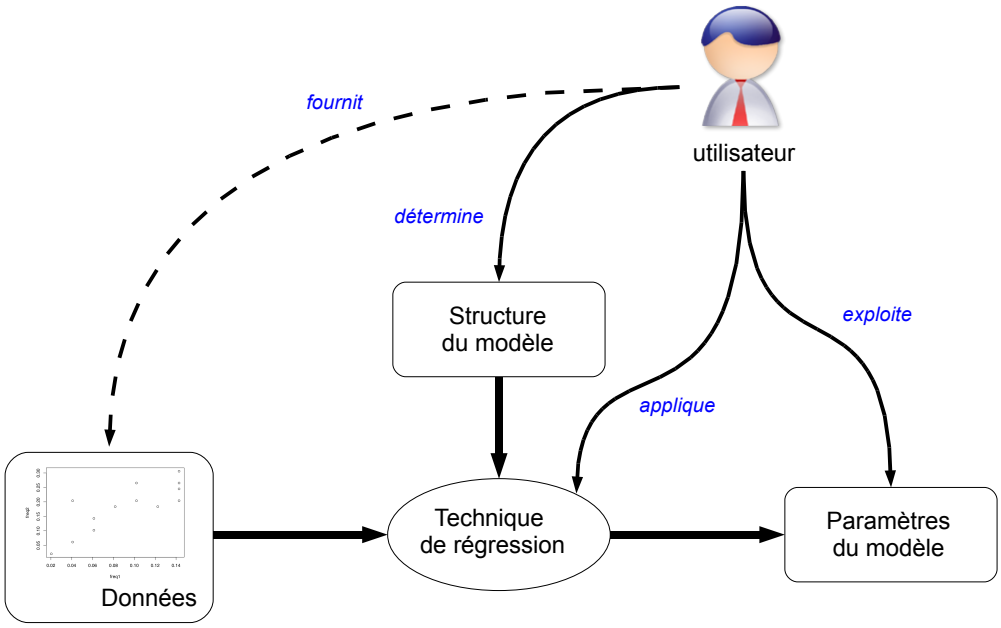


FIG. 1.4: Une vision systémique de la régression paramétrique

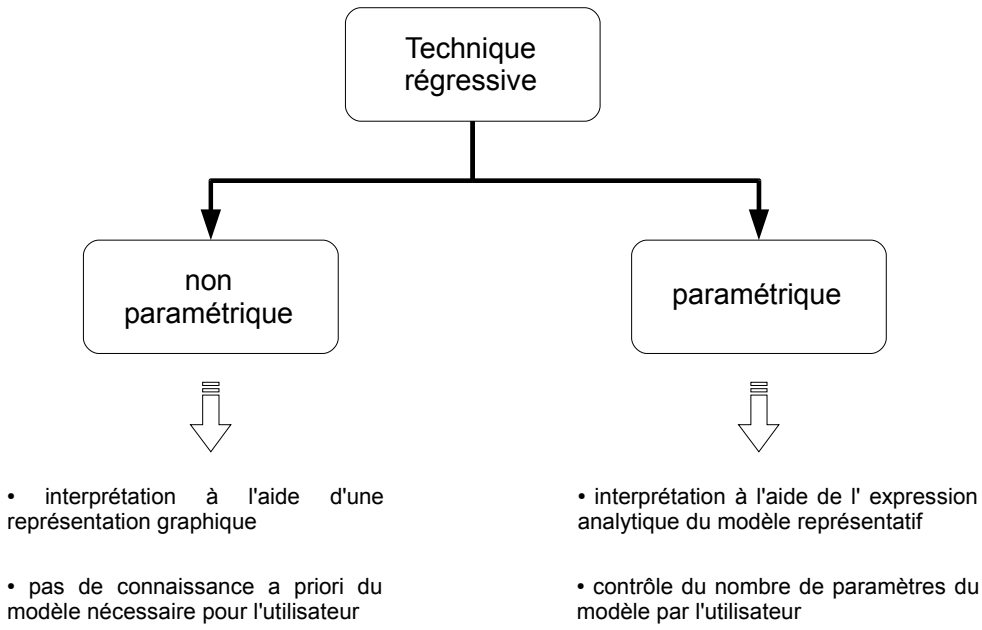


FIG. 1.5: Comparatif des techniques régressives non paramétrique et paramétrique

données, quels que soient les paramètres optimaux fournis par la technique régressive, aussi performante soit-elle (Eubank [22]). Cela peut notamment être le cas si les connaissances a priori de l'utilisateur sur les données à analyser sont insuffisantes pour proposer une structure de modèle adéquate. On remarquera ici cependant que le choix d'une approche non paramétrique ne doit se faire que si les données à analyser sont suffisamment riches en informations pour permettre de déterminer et la structure du modèle, et les nombreux paramètres correspondant, lors de la phase de régression.

Ces différents avantages et inconvénients entre approches paramétrique et non paramétrique ont été soulignés dans des études menées dans le cadre conventionnel. Dans le contexte d'imprécision qui nous intéresse, plusieurs approches non paramétriques peuvent être distinguées. Certains travaux concernent l'utilisation des *Support Vector Machines* dans un cadre de régression non paramétrique. Citons par exemple les travaux de Hao et al. [27], Hwang et al. [34], Hong et al. [32]. Il s'agit dans ces approches de déterminer un modèle imprécis, soit sur un jeu de données précises [34], soit sur des données elles-mêmes imprécises [32], ce qui correspond donc à notre définition contextuelle de la régression imprécise.

Un des avantages de ces méthodes, mis en avant dans les travaux cités précédemment, concerne là encore la non-nécessité d'avoir une connaissance a priori de la structure du modèle à identifier. Malheureusement, on retrouve dans ces approches non paramétriques imprécises le fait que les modèles obtenus ne sont interprétables qu'au travers de la représentation graphique de l'évolution des données. Ce dernier point est particulièrement vrai dans le cas de méthodes régressives robustes aux éventuels points aberrants ([32]), ou encore dans des approches appliquées sur des données ayant été classifiées, afin d'en dégager différentes tendances distinctes, lors d'une phase de clustering initiale [15].

En plus du manque d'interprétabilité des modèles produits, ces différentes approches non paramétriques présentent des inconvénients importants vis-à-vis des contraintes imposées par la définition que nous retenons d'une technique régressive, et ce à plusieurs niveaux.

Tout d'abord, les méthodes de régression proposées dans ce cadre non paramétrique présentent une certaine complexité dans leur mise en oeuvre, avec notamment, la nécessité pour l'utilisateur de définir au préalable des paramètres intervenant dans la technique régressive. La qualité du modèle identifié est fortement impactée par ce choix, comme souligné par Hong et al [32]. Ainsi, si l'utilisateur semble être affranchi de la détermination de la structure mathématique du modèle, son intervention est reportée sur le choix essentiel de ces paramètres. Cela va donc à l'encontre de la contrainte de simplicité de mise en oeuvre que nous nous imposons.

Ces méthodes non paramétriques dans un cadre d'imprécision ne permettent pas à l'utilisateur de disposer au final d'une expression analytique du modèle identifié, de manière similaire aux méthodes non paramétriques développées dans un cadre conventionnel. Ainsi, si elles per-

mettent d'obtenir une bonne représentation des données, elles ne fournissent pas un modèle simple d'exploitation permettant une analyse fine par l'utilisateur, ou encore une potentielle action de contrôle de ce modèle.

Pour finir, d'un point de vue conceptuel, ces approches peuvent sembler en contradiction avec la notion même d'environnement imprécis dans lequel nous nous positionnons. En effet, dans ce cadre d'étude, les données collectées par l'utilisateur sont potentiellement imprécises (figure 1.2). Or, dans les approches non paramétriques, les données sont vues comme suffisamment riches pour permettre l'identification simultanée de la structure et des paramètres du modèle adéquat. Ce paradoxe entre l'imprécision des données et leur richesse supposée suffisante pour l'utilisation des approches non paramétriques peut conduire à s'interroger sur le bien-fondé de l'utilisation de ces dernières.

Ainsi, afin de s'affranchir de toute interrogation concernant la richesse de l'information apportée par les données imprécises nécessaires à l'identification, et afin de pouvoir rester dans un cadre où les modèles obtenus sont exprimés analytiquement, et ainsi potentiellement contrôlables, voire inversibles, nous nous focaliserons dans la suite de ce document sur les approches de régression paramétrique.

1.2.2 Approche linéaire ou non linéaire ?

Les techniques régressives paramétriques ([24], [12], [26]) permettent de modéliser par une expression analytique, une relation existant entre des données numériques. Celles-ci, collectées par un utilisateur en vue de leur analyse, sont composées de manière générale :

- d'une *variable dépendante*, également nommée réponse, sortie, ou bien encore mesure, et notée y
- d'une ou plusieurs *variables indépendantes*, également nommées entrées, et notées $\mathbf{x} = [x_i, i = 1, ..., N]$, où $\mathbf{x}$ est le vecteur d'entrées de N composantes.

L'objectif d'une quelconque méthode d'analyse régressive paramétrique est de déterminer une relation f exprimant l'évolution de la sortie en fonction des entrées, c'est-à-dire :

$$y = f(\mathbf{x}, \mathbf{a}) \tag{1.1}$$

où $\mathbf{a} = [a_i, i = 1, ..., k]$ est le vecteur de dimension k des paramètres de la fonction f. Les k composantes de ce vecteur sont donc les inconnues à déterminer à l'aide de la méthode régressive utilisée, sachant que la forme de la fonction f en elle-même doit être connue et imposée *a priori* par l'utilisateur.

Deux cas sont traditionnellement distingués concernant la structure de cette fonction f. Elle peut ainsi être linéaire, ou non linéaire (figure 1.6), conduisant ainsi à l'utilisation d'approches régressives adaptées à chacun de ces cas. Cette linéarité doit s'entendre au sens des paramètres,

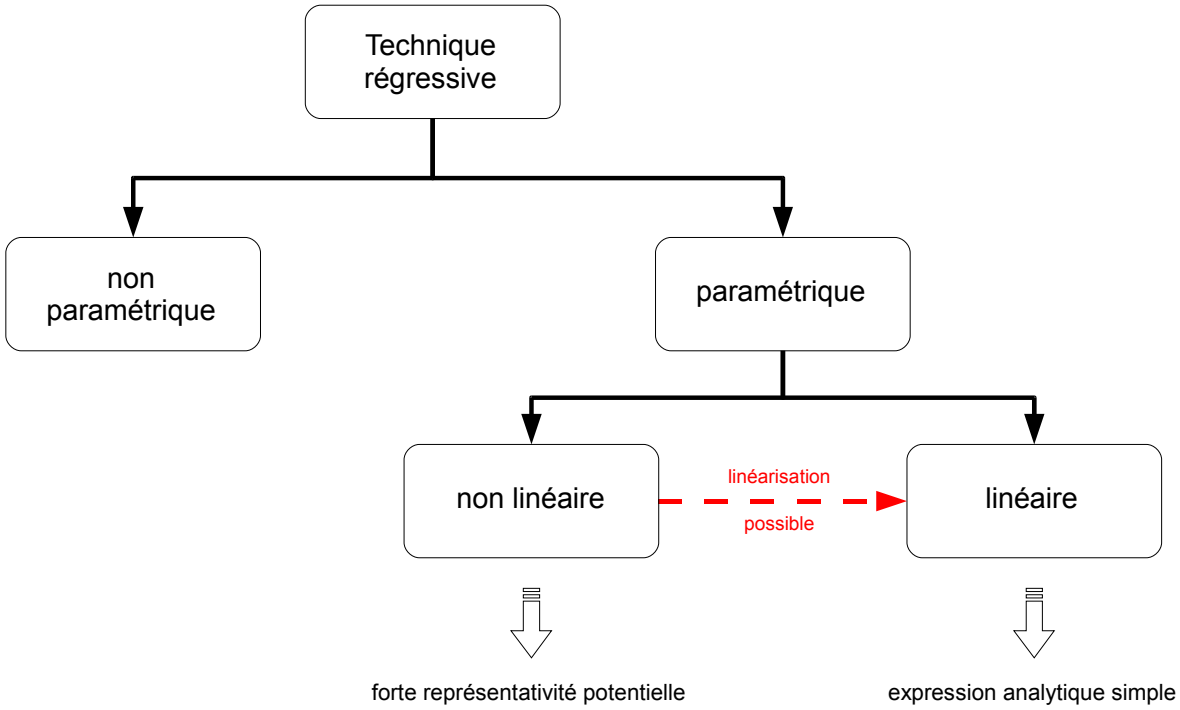


FIG. 1.6: Comparatif des techniques régressives paramétriques non linéaire et linéaire

et non des variables indépendantes.

Un modèle non linéaire pouvant avoir une forme mathématique quelconque, il est évident qu'il présente un meilleur potentiel de représentativité des données observées qu'un modèle linéaire en ses paramètres, pour lequel la contrainte de linéarité limite le choix de sa structure mathématique. Cependant, cette linéarité en les paramètres imposée dans le choix de la forme du modèle dans un cadre de régression paramétrique linéaire permet de garantir l'obtention d'un modèle représentatif de forme analytique simple, et donc facilement exploitable par l'utilisateur.

Comme introduit dans [30], il est également possible de linéariser un modèle non linéaire, afin de faciliter l'identification de ses paramètres (l'exemple le plus simple concerne les modèles exponentiels, qui peuvent être linéarisés en combinant la fonction les définissant avec un logarithme népérien). Ainsi, il est envisageable dans certains cas de pouvoir appliquer une technique régressive paramétrique linéaire bien que les données collectées initialement imposeraient le choix d'un modèle non linéaire.

Dans le contexte imprécis qui nous intéresse particulièrement ici, remarquons les travaux de Ishibuchi et al [35], basés sur l'utilisation de réseaux de neurones comme représentations de relations non linéaires liant les variables dépendantes et indépendante d'un jeu de données. Les

poids associés à ces réseaux de neurones, assimilables aux coefficients du modèle représentatif identifié, sont des intervalles flous, permettant ainsi de représenter l'imprécision intrinsèque au modèle. Là encore, la représentativité du modèle obtenu est très bonne, mais aucune forme analytique simple n'est mise à disposition de l'utilisateur en vue de l'exploitation de ce modèle.

Afin de respecter les contraintes imposées par la définition retenue de l'analyse régressive, on privilégiera l'obtention d'un modèle de forme analytique simple et donc facilement exploitable et inversible le cas échéant par l'utilisateur. Ainsi, dans la suite, nous nous focaliserons sur les techniques régressives paramétriques linéaires. Afin de prendre en compte le contexte d'imprécision dans lequel nous plaçons nos travaux, nous étudierons donc les techniques de régression paramétriques linéaires adaptées à ce cadre.

1.3 La régression paramétrique linéaire en environnement imprécis

Selon la vision contextuelle que nous retenons de la régression paramétrique en environnement imprécis l'utilisateur doit déterminer la structure du modèle linéaire le mieux à même de représenter ses données. Au préalable, il est nécessaire que soient précisés plusieurs points relatifs au choix d'une représentation floue de l'information. Ainsi, dans un premier temps, les différentes formes mathématiques de modèle imprécis possibles dans un contexte de régression floue seront présentées selon la nature des données.

Le deuxième point détaillé ensuite concerne la formalisation mathématique des sous-ensembles flous (nombres ou intervalles flous) intervenant dans la définition du problème régressif. Ces derniers permettront de représenter toutes les imprécisions manipulées que ce soit au niveau des données collectées, des modèles identifiés, ou encore des techniques régressives proposées.

1.3.1 Régression linéaire floue : données, modèles

Dans le cadre retenu de régression paramétrique linéaire, l'utilisateur doit fixer la structure du modèle à identifier. L'objectif ici est de déterminer les différentes formes mathématiques possibles en fonction des données disponibles.

On admet que l'utilisateur dispose d'un jeu de données observées composé de M échantillons. Chacun de ces échantillons est composé d'un vecteur d'entrées à N composantes, et de la sortie mesurée qui lui est associée.

Dans le contexte d'imprécision dans lequel nous nous positionnons, la nature de ces données

est un point qu'il est indispensable d'étudier. En effet, les entrées aussi bien que les sorties peuvent être précises, ou non. Cette présence ou non d'imprécision est cruciale, car elle impacte directement la nature même des modèles flous à identifier. Bien entendu, leur structure linéaire en les paramètres reste inchangée, seule la nature précise ou non des paramètres est influencée. Par souci de clarté, nous adopterons par la suite les conventions de notation suivantes :

- une grandeur précise, c'est-à-dire un réel, sera notée en minuscule : x
- une grandeur imprécise, c'est-à-dire un intervalle flou, sera notée en majuscule : X

Dans [65], Tanaka, qui a initialement introduit le concept de régression floue, se place dans le cadre où seules les sorties observées sont imprécises, les entrées étant précisément connues. Dans ce cas, l'ensemble des données observées est défini par :

$$([x_{1j}, x_{2j}, ..., x_{Nj}], Y_j), j = 1, ..., M \tag{1.2}$$

où :

- les N composantes x_{kj}, $k = 1, ..., N$ du j^{eme} échantillon, $j = 1, ..., M$, du vecteur d'entrée sont des réels
- la sortie associée Y_j, $j = 1, ..., M$ est un intervalle flou

Dans ce cas [65], l'imprécision de la sortie du modèle identifié ne peut avoir pour origine que l'imprécision de ses paramètres. Le modèle que l'utilisateur est amené à identifier est donc de la forme :

$$Y = A_0 \oplus A_1 \odot x_1 \oplus A_2 \odot x_2 ... \oplus A_N \odot x_N \tag{1.3}$$

où :

- les coefficients $A_i, i = 0, ..., N$ sont des intervalles flous
- le symbole $\oplus$ est l'opérateur d'addition entre deux grandeurs floues
- le symbole $\odot$ est l'opérateur de multiplication entre une grandeur floue et un réel.

Il est important de remarquer ici que le cas où toutes les données collectées sont précises est un cas particulier de cette approche. En effet, l'utilisateur peut chercher à identifier un modèle imprécis, en fait un modèle à paramètres imprécis, à partir de données précises. Dans ce cas, l'expression mathématique (1.3) reste valable et aussi bien les entrées que les sorties collectées sont des réels, seuls les paramètres du modèle sont des intervalles flous. L'imprécision du modèle identifié reflète alors soit la variabilité des entrées/sorties précises, soit la non-validité de l'hypothèse faite sur la structure du modèle ou encore sur sa linéarité [56].

Cependant, les imprécisions sur les données peuvent concerner aussi bien les entrées que les sorties observées. Ce cas peut donc être vu comme une généralisation du précédent. Dans ce cas, l'ensemble des données observées est défini par :

$$([X_{1j}, X_{2j}, ..., X_{Nj}], Y_j), j = 1, ..., M \tag{1.4}$$

où :

- les N composantes X_{kj}, $k = 1, ..., N$ du vecteur d'entrée du j^{eme} échantillon, $j = 1, ..., M$, sont des intervalles flous
- la sortie associée Y_j, $j = 1, ..., M$ est un intervalle flou

Plusieurs types de modèles peuvent alors être identifiés dans ce cas. La première approche sur ce type de données [18] vise à identifier un modèle pour lequel l'incertitude de sa sortie n'a pour origine unique que celle de ses entrées. Le modèle à identifier est alors de la forme :

$$Y = a_0 \oplus a_1 \odot X_1 \oplus a_2 \odot X_2 ... \oplus a_N \odot X_N \tag{1.5}$$

où les coefficients $a_i, i = 0, ..., N$ sont des réels.

En fait, cette approche ne correspond pas tout à fait à l'objectif fixé de la régression linéaire floue. En effet, étant donné que la structure du modèle est précise, ce dernier ne représente qu'un transfert précis entre entrées et sorties floues. Dans ce cas, l'imprécision résultant du modèle lui-même est inexistante et le modèle est en fait précis.

Il est possible [17] d'introduire une part d'imprécision dans le modèle (1.5) en cherchant à identifier un modèle de la forme :

$$Y = A_0 \oplus a_1 \odot X_1 \oplus a_2 \odot X_2 ... \oplus a_N \odot X_N \tag{1.6}$$

où les coefficients $a_i, i = 1, ..., N$ sont toujours des réels, tandis que le coefficient A_0 est maintenant un intervalle flou. Cette approche a pour avantage d'introduire une certaine imprécision dans la structure du modèle même, c'est-à-dire au niveau du paramètre A_0. Cependant, cette approche n'en reste pas moins biaisée. En effet, il n'y a pas de connaissance *a priori* sur le fait que seul le paramètre A_0 soit imprécis, et pas les autres. Ainsi, les contraintes sous-jacentes sur la nature des paramètres lors de l'identification sont trop restrictives pour garantir l'identification d'un modèle de forme suffisamment générale.

Une approche plus générale est proposée dans les travaux de Sakawa et Yano [54]. Dans le cadre où les entrées et les sorties observées sont imprécises, le modèle identifié est de la forme :

$$Y = A_0 \oplus A_1 \otimes X_1 \oplus A_2 \otimes X_2 ... \oplus A_N \otimes X_N \tag{1.7}$$

où :

- les coefficients $A_i, i = 1, ..., N$ sont des intervalles flous.
- le symbole $\otimes$ est l'opérateur de multiplication entre deux grandeurs floues.

Dans ce cas, le modèle permet de représenter aussi bien l'incertitude due aux données que celle inhérente à sa structure imprécise, tous ses paramètres étant des intervalles flous. De plus, un réel pouvant être vu comme le cas particulier d'un intervalle flou d'imprécision nulle, le modèle (1.7) généralise tous les précédents, que ce soit le modèle (1.3) qui considère des entrées précises ou encore les modèles (1.5) et (1.6) à paramètres précis.

Le tableau 1.1 présente un récapitulatif des différents modèles à identifier selon la nature des données observées et le choix d'un modèle précis ou imprécis. Dans tous les cas, ce sont les paramètres du modèle qui le rendent précis ou imprécis. Quant à la sortie du modèle, elle est systématiquement imprécise dans le cadre de cette thèse, le cas précis n'étant pas traité. Selon le contexte de représentation choisi, l'imprécision de la sortie du modèle régressif provient alors soit de celle des entrées, soit de celle du modèle, soit des deux. On remarquera l'impossibilité du cas où l'utilisateur chercherait à identifier un modèle précis sur des données collectées dont les sorties sont imprécises alors que les entrées sont précises. En effet, dans ce cas, l'imprécision des sorties mesurées serait en contradiction avec la précision des sorties estimées.

Dans toute la suite de ce chapitre, nous nous focalisons sur des données de type (1.2) et des modèles (1.3). Le fait de se limiter au cas des entrées précises peut sembler restrictif, celles-ci pouvant également présenter une forme d'imprécision qu'il est préjudiciable de négliger. On remarquera cependant que dans le cadre d'un problème d'identification, les entrées appliquées au système peuvent être considérées comme parfaitement connues et définies. Dans ce cas, seules les sorties mesurées sont potentiellement imprécises, conséquence d'une part des éventuelles erreurs de mesure, et d'autre part de l'imprécision du système lui-même. L'introduction de cette dernière dans le modèle est l'objectif essentiel des méthodes régressives linéaires floues présentées dans la suite. Ce contexte d'étude est par ailleurs le plus couramment considéré dans la littérature concernant la régression linéaire floue.

		DONNEES		
		Entrées précises Sorties précises	Entrées précises Sorties imprécises données (1.2)	Entrées imprécises Sorties imprécises données (1.4)
MODELE	Paramètres précis	Non traité ici	*Cas impossible*	modèle (1.5)
	Paramètres imprécis	modèle (1.3)	modèle (1.3)	modèle (1.7)

TAB. 1.1: Nature des imprécisions prises en compte selon la nature des données et celle des paramètres des modèles

Synthèse 1 : Dans la suite de ce chapitre, il sera supposé que le jeu de données disponible est de la forme :

$$([x_{1j}, x_{2j}, ..., x_{Nj}], Y_j), j = 1, ..., M \tag{1.8}$$

et que l'on cherche à représenter les sorties observées imprécises par un modèle linéaire imprécis de la forme :

$$Y = A_0 \oplus A_1 \odot x_1 \oplus A_2 \odot x_2 ... \oplus A_N \odot x_N \tag{1.9}$$

1.3.2 Intervalles, Intervalles flous et arithmétique associée

Comme présenté dans la section précédente, dans une approche régressive paramétrique linéaire en environnement imprécis, les imprécisions peuvent intervenir aussi bien au niveau des sorties observées que de la structure du modèle déterminée par l'utilisateur. Il est donc important de pouvoir représenter ces imprécisions, afin de pouvoir les exhiber, les quantifier, les manipuler et à terme les exploiter.

Ainsi, nous proposons de présenter le formalisme des ensembles flous introduits par Zadeh [74], ainsi que les opérations associées intervenant dans les modèles imprécis. Cela permettra par la suite de discuter des techniques régressives imprécises en ayant à disposition un formalisme unifié et rigoureux. Les sous-ensembles flous, plus précisément dans notre cas des nombres flous ou intervalles flous, peuvent être vus comme des familles d'intervalles emboîtés. Dans un premier temps, nous introduisons donc les notations associées à la représentation des intervalles conventionnels.

Un intervalle a est défini comme l'ensemble d'éléments de $\Re$ compris entre une borne inférieure a^- et une borne supérieure a^+, c'est-à-dire :

$$a = \{x | a^- \leq x \leq a^+, x \in \Re\} \tag{1.10}$$

Il est possible de caractériser cet intervalle a dans un espace permettant d'exhiber de manière immédiate son incertitude en choisissant l'espace de représentation Midpoint/Radius. Dans ce cas, le Midpoint représente le point Milieu de l'intervalle, et est donné par :

$$M_a = (a^- + a^+)/2 \tag{1.11}$$

Quant au Radius, ou demi-largeur, il représente l'incertitude de l'intervalle a, qui devient donc directement accessible. Le Radius est donné par :

$$R_a = (a^+ - a^-)/2 \tag{1.12}$$

Ainsi, il est possible de définir l'intervalle a de deux manières distinctes, soit dans l'espace des bornes :

$$a = [a^-, a^+] \tag{1.13}$$

soit dans l'espace Midpoint / Radius :

$$a = (M_a, R_a) \tag{1.14}$$

Le passage d'une représentation à l'autre est réalisé grâce aux équations (1.11) et (1.12) ou encore selon les équivalences suivantes :

$$a^- = M_a - R_a \tag{1.15}$$

$$a^+ = M_a + R_a \tag{1.16}$$

En théorie des ensembles, on associe à l'intervalle a de $\Re$ sa fonction caractéristique, généralement notée χ_a ou encore $\mathbb{1}_a$. Cette dernière explicite l'appartenance de tout élément de $\Re$ à l'intervalle a. Formellement, χ_a est la fonction définie par :

$$\begin{array}{rcl} \chi_a : \Re & \longrightarrow & \{0,1\} \\ x & \longmapsto & \chi_a(x) = \begin{cases} 1 \text{ si } x \in a \\ 0 \text{ sinon} \end{cases} \end{array} \tag{1.17}$$

Graphiquement, la fonction χ_a est la fonction rectangulaire illustrée à la figure 1.7.

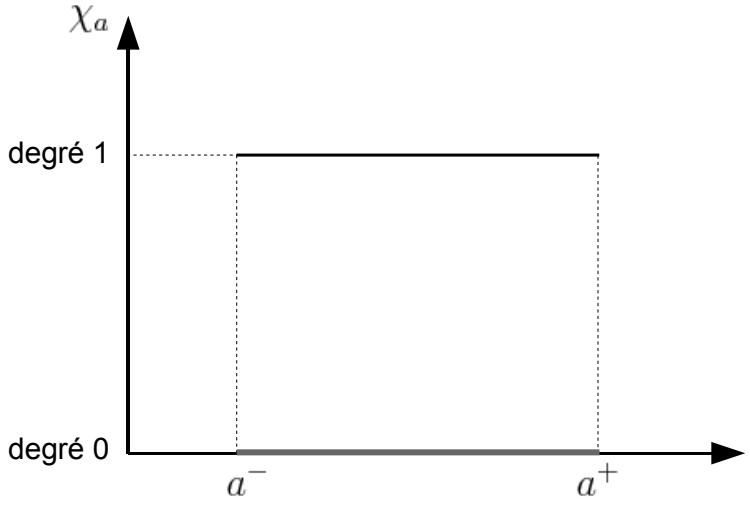


FIG. 1.7: Représentation de la fonction caractéristique d'un intervalle

Dans le cas plus général d'un intervalle flou A, une distribution de possibilité définie sur $\Re$ et modélisée par une fonction d'appartenance notée μ_A, est associée à A. Formellement, cette

fonction μ_A est définie par :

$$\begin{array}{rcl} \mu_A : \Re & \longrightarrow & [0,1] \\ x & \longmapsto & \mu_A(x) \end{array} \tag{1.18}$$

La figure 1.8 en donne une représentation graphique, et met en évidence que deux types d'informations sont à considérer pour définir un intervalle flou, correspondant à deux dimensions distinctes :

- la dimension horizontale, commune aux intervalles conventionnels, qui est l'axe des réels $\Re$
- la dimension verticale, permettant la représentation des degrés d'appartenance, qui est donc l'intervalle $[0,1]$

La fonction d'appartenance constitue une extension de la fonction caractéristique, dans la mesure où elle prend ses valeurs dans l'intervalle $[0,1]$ et non plus uniquement dans l'ensemble $\{0,1\}$.

Il est possible d'associer à un intervalle flou deux intervalles particuliers correspondant aux deux valeurs extrêmes du degré d'appartenance. Ainsi, le support de A est l'intervalle constitué des éléments appartenant au moins un peu à A. Il est donc défini au degré 0 par $Support(A) = \{x \in \Re | \mu_A(x) > 0\}$. Le noyau de A est quant à lui l'ensemble des éléments appartenant totalement à A. C'est donc l'intervalle défini au degré 1 par $Noyau(A) = \{x \in \Re | \mu_A(x) = 1\}$.

Lorsque la fonction d'appartenance est linéaire, l'intervalle flou A est trapézoïdal (figure 1.8). Il est alors complètement défini par son support et son noyau. Dans tous la suite, on notera :

$$S_A = Support(A) = [S_A^-, S_A^+] \tag{1.19}$$

$$K_A = Noyau(A) = [K_A^-, K_A^+] \tag{1.20}$$

Quant à l'intervalle flou trapézoïdal A, il sera noté $A = (K_A, S_A)$. En utilisant une représentation des intervalles par les bornes conformément à (1.13), on obtient alors :

$$A = ([K_A^-, K_A^+], [S_A^-, S_A^+]) \tag{1.21}$$

Il est également possible de noter l'intervalle flou trapézoïdal A en exploitant une représentation dans l'espace Midpoint / Radius (cf. équation (1.14)) de chacun de ses intervalles significatifs, c'est-à-dire en exhibant de manière directe l'imprécision du support (imprécision maximale envisageable) et celle du noyau (imprécision attachée aux éléments d'appartenance maximale) :

$$A = ((M_{K_A}, R_{K_A}), (M_{S_A}, R_{S_A})) \tag{1.22}$$

Lorsque le noyau d'un intervalle flou est parfaitement défini, c'est-à-dire d'imprécision nulle ($R_{K_A} = 0$), l'intervalle flou est unimodal. Dans le cas d'une fonction d'appartenance linéaire,

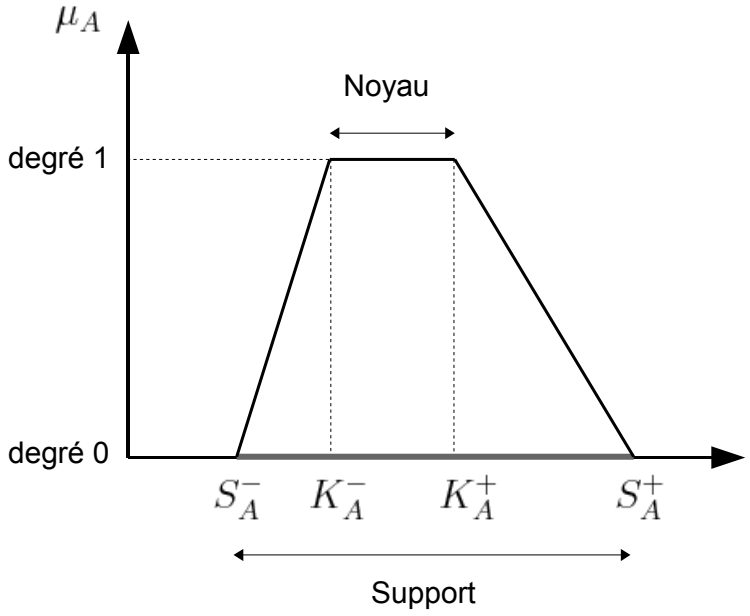


FIG. 1.8: Représentation d'un intervalle flou trapézoïdal

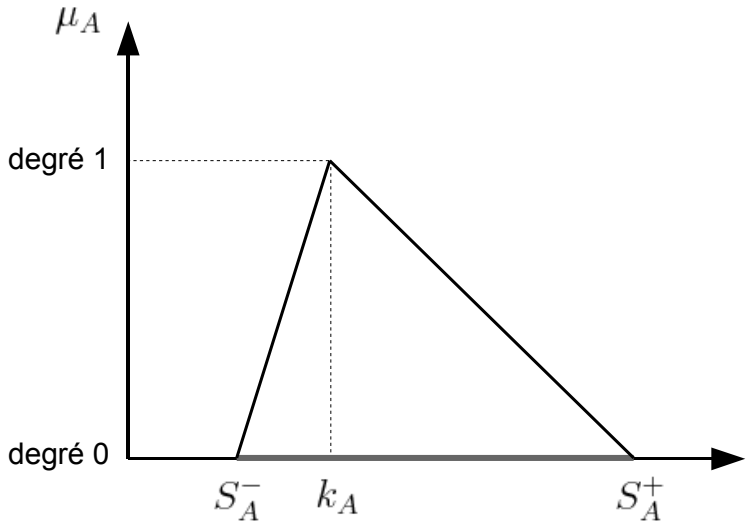


FIG. 1.9: Représentation d'un intervalle flou triangulaire

l'intervalle flou A devient triangulaire (figure 1.9). Il est alors complètement défini par son intervalle support $[S_A^-, S_A^+]$ et son noyau précis K_A restreint à un unique élément, c'est-à-dire $K_A = \{k_A\}$, $R_{K_A} = 0$ et $M_{K_A} = k_A$. Dans ce cas, l'intervalle flou triangulaire A sera noté :

$$A = (k_A, [S_A^-, S_A^+]) \tag{1.23}$$

Bien entendu, il est là aussi possible de le représenter dans l'espace Midpoint / Radius,

l'imprécision du support étant alors directement accessible :

$$A = (k_A, (M_{S_A}, R_{S_A})) \tag{1.24}$$

Un autre cas particulier concerne les intervalles flous triangulaires symétriques (figure 1.10). Dans ce cas, l'unique élément du noyau est défini comme étant le milieu de l'intervalle support,

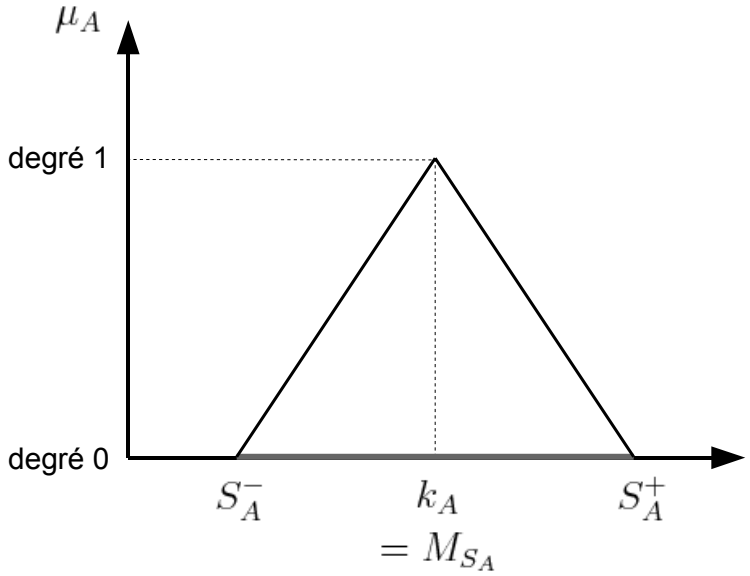


FIG. 1.10: Représentation d'un intervalle flou triangulaire symétrique

c'est-à-dire :

$$k_A = M_{S_A} \tag{1.25}$$

L'intervalle flou triangulaire symétrique A est alors complètement défini par son intervalle support, puisqu'en substituant l'égalité (1.25) dans (1.24), on obtient :

$$A = (M_{S_A}, (M_{S_A}, R_{S_A})) \tag{1.26}$$

Sachant qu'un unique intervalle définit alors complètement l'intervalle flou A (figure 1.11), pour une raison de simplicité, il sera noté :

$$A = (M_A, R_A) \tag{1.27}$$

ou dans l'espace des bornes :

$$A = [S_A^-, S_A^+] \tag{1.28}$$

Reste à souligner que les représentations (1.27) et (1.28) sont similaires à (1.13) et (1.14). Seul le contexte d'utilisation (a intervalle ou A intervalle flou triangulaire symétrique) permet de les distinguer.

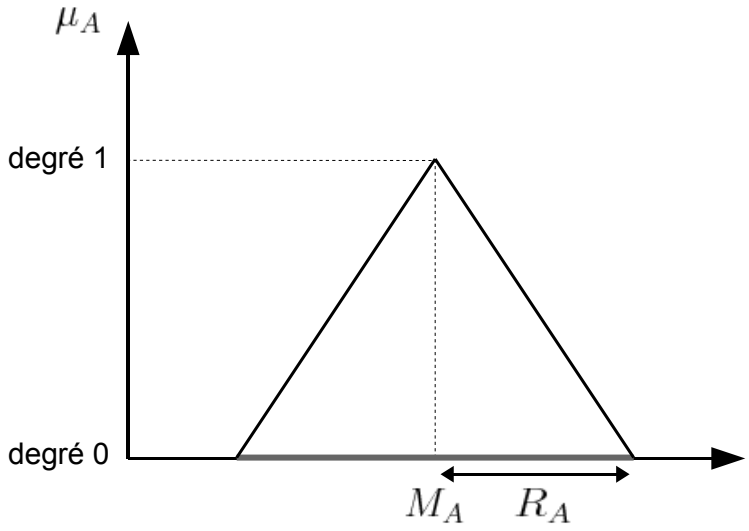


FIG. 1.11: Représentation d'un intervalle flou triangulaire symétrique

Quel que soit le type d'intervalle flou considéré, il est nécessaire que les intervalles support et noyau soient bien définis. Cela se traduit de manière générale par :

- dans l'espace des bornes : celles-ci doivent être bien ordonnées, c'est-à-dire $S_A^- \leq S_A^+$ et $K_A^- \leq K_A^+$;
- dans l'espace Midpoint / Radius : les Radius doivent être positifs, c'est-à-dire $R_{S_A} \geq 0$ et $R_{K_A} \geq 0$.

Les intervalles flous pouvant être vus comme une généralisation des intervalles conventionnels par l'ajout d'une dimension verticale supplémentaire correspondant au degré d'appartenance de chaque élément, il est possible de les manipuler comme une collection d'intervalles conventionnels. Cela se fait en utilisant le principe des α-coupes (figure 1.12). Une α-coupe $[A]_\alpha$ d'un intervalle flou A est définie par l'intervalle composé des éléments dont le degré d'appartenance à A est supérieur ou égal à la valeur α considérée :

$$[A]_\alpha = \{x \in \Re | \mu_A(x) \geq \alpha\} \tag{1.29}$$

Une fois le principe des α-coupes retenu, pour un degré α fixé a priori, tout calcul se ramène à manipuler des intervalles conventionnels, et ce, au travers de l'arithmétique qui leur est associée. Il devient donc possible d'exploiter les opérations de base définies sur des intervalles conventionnels, c'est-à-dire la somme de deux intervalles ainsi que la multiplication par un scalaire pour évaluer la sortie d'un modèle (1.9), dont les paramètres sont imprécis, et les entrées précises.

Soient deux intervalles $a = [a^-, a^+] = (M_a, R_a)$ et $b = [b^-, b^+] = (M_b, R_b)$, leur somme est définie dans les deux espaces de représentation (bornes et Midpoint / Radius) par :

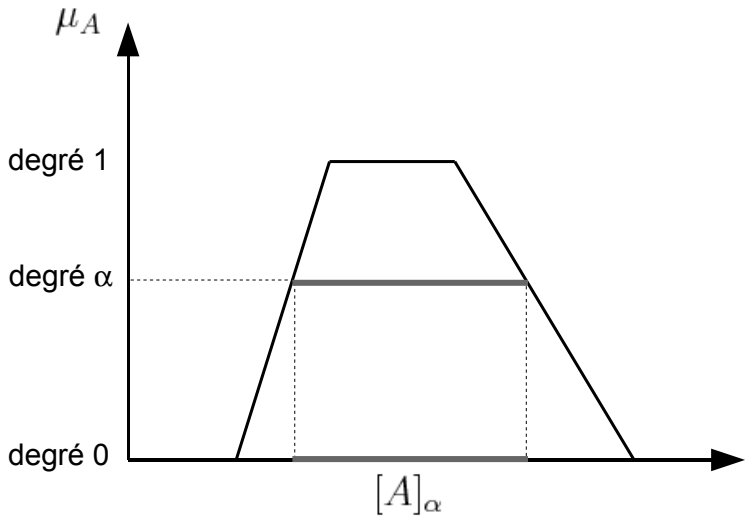


FIG. 1.12: Visualisation d'une α-coupe d'un intervalle flou

- dans l'espace des bornes :

$$a \oplus b = [a^- + b^-, a^+ + b^+] \tag{1.30}$$

- dans l'espace Midpoint / Radius :

$$a \oplus b = (M_{a\oplus b}, R_{a\oplus b}) \tag{1.31}$$

avec :

$$\begin{cases} M_{a\oplus b} = M_a + M_b \\ R_{a\oplus b} = R_a + R_b \end{cases} \tag{1.32}$$

Le produit de l'intervalle a par le scalaire $\omega \in \Re$ est défini par :

- dans l'espace des bornes :

$$\omega \odot a = \begin{cases} [\omega \cdot a^-, \omega \cdot a^+] \text{ , si } \omega \geq 0 \\ [\omega \cdot a^+, \omega \cdot a^-] \text{ , si } \omega < 0 \end{cases} \tag{1.33}$$

- dans l'espace Midpoint / Radius :

$$\omega \odot a = (M_{\omega\odot a}, R_{\omega\odot a}) \tag{1.34}$$

avec :

$$\begin{cases} M_{\omega\odot a} = \omega \cdot M_a \\ R_{\omega\odot a} = |\omega| \cdot R_a \end{cases} \tag{1.35}$$

Pour une formulation complète du problème de régression, il sera par la suite nécessaire de disposer d'opérateurs ensemblistes, notamment d'une relation d'inclusion entre deux intervalles. En conservant le même formalisme que pour les opérateurs arithmétiques, l'inclusion de a dans b, c'est-à-dire $a \subseteq b$ est définie comme suit :

- dans l'espace des bornes :

$$a \subseteq b \Leftrightarrow \begin{cases} b^- \leq a^- \\ a^+ \leq b^+ \end{cases} \tag{1.36}$$

- dans l'espace Midpoint / Radius [10]

$$a \subseteq b \Leftrightarrow |M_b - M_a| \leq R_b - R_a \tag{1.37}$$

Il est également utile d'introduire la notion d'intersection non vide entre a et b, c'est-à-dire $a \cap b \neq \emptyset$, définie comme suit :

- dans l'espace des bornes :

$$a \cap b \neq \emptyset \Leftrightarrow \begin{cases} a^- \leq b^+ \\ b^- \leq a^+ \end{cases} \tag{1.38}$$

- dans l'espace Midpoint / Radius

$$a \cap b \neq \emptyset \Leftrightarrow |M_b - M_a| \leq R_a + R_b \tag{1.39}$$

Tous ces concepts et opérations basiques permettent maintenant de quantifier et manipuler des imprécisions dans un contexte de régression paramétrique linéaire, aussi bien au niveau des sorties observées que des paramètres des modèles à identifier.

1.4 Les approches fondamentales de la régression linéaire floue

Dans les paragraphes précédents, les différents types de modèles que l'utilisateur peut être amené à identifier sur un jeu de données, en environnement imprécis, ont été présentés, avant que n'ait été fait le choix du contexte particulier de la synthèse 1 (équations (1.8) et (1.9)). Puis, le concept d'intervalles flous retenu pour la formalisation des imprécisions a été introduit. En exploitant le principe des α-coupes, le calcul flou nécessaire à l'évaluation de la sortie du modèle (1.9) a ensuite été abordé au travers de l'arithmétique des intervalles conventionnels. Tout cela permet maintenant de se focaliser sur les techniques régressives linéaires floues qu'un utilisateur aura à sa disposition pour mener à bien la phase d'identification du modèle dont il souhaite déterminer les paramètres.

L'objectif ici est donc de présenter les deux principales catégories de techniques se dégageant en régression linéaire floue. Le but de la technique de régression est alors d'identifier les paramètres $A_0, A_1, ..., A_N$ du modèle (1.9) à l'aide des M échantillons de données disponibles (1.8).

Une première classification fondamentale des techniques de régression linéaire floue est proposée par Tanaka et Diamond dans [18]. Deux approches peuvent ainsi être distinguées :

- l'approche dite possibiliste, introduite par Tanaka [65], et dénommée ainsi car basée sur la manipulation des distributions de possibilités constituant les fonctions d'appartenance des intervalles flous considérés pour modéliser l'imprécision [18].

- l'approche dite des moindres carrés flous, introduite par Diamond [17], vue comme une extension de l'approche classique de régression linéaire basée sur l'idée de la recherche de la meilleure adéquation du modèle aux données.

Dans la suite, nous nous attachons à revisiter ces différentes approches au travers du formalisme présenté auparavant, basé sur le principe des α-coupes, permettant d'utiliser les opérateurs arithmétiques sur les intervalles.

1.4.1 Approche possibiliste

1.4.1.1 Formulation du problème

Les premiers travaux sur la régression linéaire floue ont été menés par Tanaka et al. [65], dans un cadre de régression dite possibiliste. Cette approche repose sur le principe global de l'optimisation de l'imprécision du modèle sous un ensemble de contraintes exprimant la relation entre la sortie du modèle et les données d'identification. Plusieurs points fondamentaux sont à souligner dans cette approche.

Dans les travaux [65], les auteurs considèrent des données d'identification dont les sorties imprécises sont des intervalles flous triangulaires symétriques. En adoptant la notation de l'équation (1.27) pour les sorties observées Y_j, les données d'identification (1.8) deviennent :

$$([x_{1j}, x_{2j}, ..., x_{Nj}], (M_{Y_j}, R_{Y_j})), j = 1, ..., M \tag{1.40}$$

Ce choix de forme des données est dicté par des raisons de simplicité, deux paramètres, c'est-à-dire le Midpoint et le Radius, étant suffisant pour parfaitement définir les sorties imprécises observées.

Les sorties observées étant des intervalles flous triangulaires symétriques, la structure du modèle est choisie de façon à produire des sorties de même nature. Les entrées considérées étant précises, et donc modélisées sous forme de réels, seule la nature des paramètres du modèle influence celle de sa sortie. Ainsi, tout comme les sorties observées, les paramètres du modèle (1.9) sont des intervalles flous triangulaires de la forme :

$$A_i = (M_{A_i}, R_{A_i}), i = 0, ..., N \tag{1.41}$$

Les sorties triangulaires symétriques du modèle, ou prédictions, sont alors notées :

$$\hat{Y}_j = (M_{\hat{Y}_j}, R_{\hat{Y}_j}), j = 1, ..., M \tag{1.42}$$

Ainsi, en exploitant les équations (1.32) et (1.35) le Midpoint de la j^{ieme} sortie du modèle est donné par :

$$M_{\hat{Y}_j} = M_{A_0} + \sum_{i=1}^{N} M_{A_i} x_{ij}, j = 1, ..., M \tag{1.43}$$

Le Radius est quant à lui donné par :

$$R_{\hat{Y}_j} = R_{A_0} + \sum_{i=1}^{N} R_{A_i}|x_{ij}|, j = 1, ..., M \tag{1.44}$$

Les variables à identifier sont donc le Midpoint et le Radius des $N + 1$ paramètres du modèle. Comme mentionné précédemment, l'identification de ces paramètres se fait en optimisant l'imprécision du modèle sous un ensemble de contraintes modélisant la relation entre les sorties observées et celles du modèle. Deux points distincts sont donc à étudier, le critère d'optimisation d'une part qui découle d'un choix de représentation de l'imprécision du modèle, et les contraintes d'autre part.

1.4.1.2 Représentation de l'imprécision du modèle

Dans le cadre de la régression linéaire floue dite possibiliste, les paramètres du modèle sont identifiés de manière à optimiser un critère linéaire, sachant que les sorties du modèle doivent respecter des contraintes vis-à-vis des sorties observées. Le critère doit alors permettre de quantifier l'imprécision du modèle, une représentation de celle-ci devant donc être définie.

Dans un premier temps, Tanaka et al. [65] considèrent que l'imprécision du modèle peut être vue comme étant la somme de celle de ses paramètres $A_i, i = 0, ..., N$. Ceux-ci étant des intervalles flous triangulaires symétriques, leur imprécision est donnée par leur demi-largeur, soit leur Radius. Ainsi, le critère J_1 à optimiser est la somme des Radius des paramètres du modèle, soit :

$$J_1(\mathbf{R_A}) = \sum_{i=0}^{N} R_{A_i} \tag{1.45}$$

où $\mathbf{R_A}$ représente le vecteur regroupant les Radius des différents paramètres.

Cependant, ce critère présente plusieurs inconvénients, constatés par les auteurs eux-mêmes [62] :

- L'optimisation du critère J_1 sous contraintes de Radius positifs tend à fournir un modèle dont la plupart des paramètres sont des nombres précis (Radius nul). En fait, il est montré dans [36] que lorsqu'une entrée domine toutes les autres en valeur absolue, il est possible d'atteindre la valeur minimale de J_1 en augmentant suffisamment le Radius du paramètre associé à la variable dominante pour pouvoir annuler le Radius de tous les autres paramètres. Ce comportement apparaît clairement dans les résultats fournis dans [28].
- Dans J_1, l'incertitude du modèle n'est définie qu'à partir de celle de ses paramètres. Cette formulation de l'imprécision du modèle présente l'inconvénient de négliger la contribution des entrées. Pourtant, si l'on se réfère à l'équation (1.44), celles-ci pondèrent la somme des Radius des paramètres dans l'évaluation de l'imprécision des sorties prédites $R_{\hat{Y}_j}, \forall j = 1, ..., M$ et devraient donc contribuer à l'imprécision du modèle.

Pour remédier à ces inconvénients, une autre formulation linéaire de l'imprécision d'un modèle linéaire flou, et donc du critère à optimiser dans le cadre de son identification est proposée dans [62]. Dans ce cas, l'imprécision d'un modèle linéaire est considéré comme étant celle de sa sortie. Ainsi, le critère à optimiser proposé correspond à l'imprécision totale du modèle pour l'ensemble des entrées considérées lors de l'identification. Il s'agit donc de sommer le Radius des sorties prédites $R_{\hat{Y_j}}, \forall j = 1, ..., M$, le Radius étant défini par l'équation (1.44). Le critère J_2 obtenu est alors :

$$J_2(\mathbf{R_A}) = M.R_{A_0} + \sum_{i=1}^{N}\sum_{j=1}^{M} R_{A_i}|x_{ij}| \tag{1.46}$$

Quelques remarques importantes peuvent être faites sur ce critère.

Le critère est linéaire en les variables d'optimisation, que sont les Radius des paramètres du modèle. En effet, les entrées $x_{ij}, i = 1, ..., N, j = 1, ..., M$ sont des valeurs numériques précises et connues, l'introduction de leur valeur absolue dans le critère n'induit donc pas sur sa linéarité.

L'imprécision du modèle à optimiser est ici définie comme celle de sa sortie aux points observés utilisés pour l'identification. Ainsi, l'imprécision (Radius) des paramètres est prise en compte, tout comme les entrées observées qui pondèrent la contribution des différents paramètres.

Pour les deux critères définis, l'optimisation se fait selon le Radius des paramètres, leur Midpoint n'apparaissant ni dans l'expression de J_1 (1.45) ni dans celle de J_2 (1.46). Ainsi, l'optimisation de ces critères linéaires en l'état, hors contraintes, conduit à l'obtention de paramètres tous nuls. Ce résultat trivial (le modèle le moins imprécis possible est un modèle précis) n'a évidemment aucun intérêt mais met en évidence la nécessité d'une optimisation sous contraintes. Celles-ci, présentées dans la suite, devront donc faire apparaitre aussi bien les Radius que les Midpoints des paramètres du modèle comme variables d'optimisation.

1.4.1.3 Les différentes contraintes envisageables

Différentes relations entre sorties observées et prédites pouvant être considérées ont été proposées par Tanaka et al. dans [61], [66] et [62], comme une extension aux travaux initiaux [65].

Afin de s'affranchir de tout calcul flou et par conséquent de se ramener à la manipulation d'intervalles conventionnels, les auteurs utilisent le principe des α-coupes et proposent de résoudre le problème de régression à un α fixé.

Trois cas de figure sont distingués, selon qu'il est imposé, pour toute donnée d'identification, que :

- la sortie du modèle couvre la sortie observée (**problème Minimal**),
- la sortie du modèle soit couverte par la sortie observée (**problème Maximal**),

- l'intersection entre la sortie du modèle et celle observée ne soit pas vide (**problème Conjonctif**).

On remarquera ici qu'indépendamment du problème traité, il est impératif de garantir par contraintes que les intervalles flous identifiés comme paramètres du modèle soient bien des intervalles flous. Ainsi, leur Radius doit être positif. Par conséquent, dans tous les cas, il est nécessaire d'introduire les contraintes linéaires communes suivantes :

$$R_{A_i} \geq 0, i = 0, ..., N \tag{1.47}$$

Il est ensuite possible d'introduire les contraintes spécifiques à chacun des cas cités auparavant.

Le problème Minimal : Il s'agit dans ce cas de garantir par contraintes l'inclusion des sorties observées Y_j dans celles prédites $\hat{Y}_j^{min}$, c'est-à-dire, pour le degré α considéré :

$$[Y_j]_\alpha \subseteq [\hat{Y}_j^{min}]_\alpha, j = 1, ..., M \tag{1.48}$$

Dans ce cas, le modèle, également nommé modèle de possibilité dans [18], est identifié en

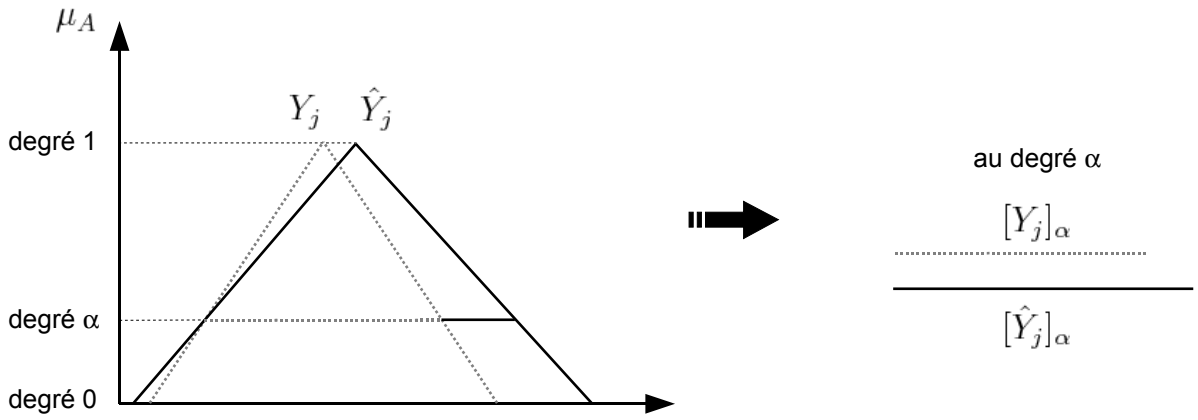


FIG. 1.13: Le problème Minimal

minimisant le critère choisi sous les contraintes (1.48) et (1.47). Ainsi, il s'agit de déterminer les paramètres du modèle tels que sa sortie soit la moins imprécise, tout en englobant les sorties observées au degré α fixé.

Le problème Maximal : Il s'agit dans ce cas de garantir par contraintes l'inclusion des sorties prédites $\hat{Y}_j^{max}$ dans celles observées Y_j, c'est-à-dire, pour le degré α considéré :

$$[\hat{Y}_j^{max}]_\alpha \subseteq [Y_j]_\alpha, j = 1, ..., M \tag{1.49}$$

Dans ce cas, le modèle, également nommé modèle de nécessité dans [18], est identifié en maxi-

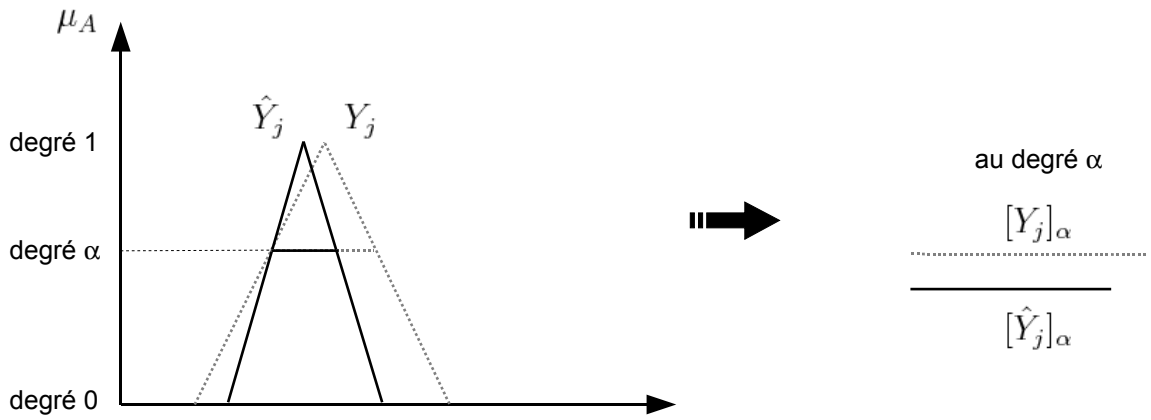


FIG. 1.14: Le problème Maximal

misant le critère choisi sous les contraintes (1.49) et (1.47). Ainsi, il s'agit de déterminer les paramètres du modèle tels que sa sortie soit la plus imprécise, tout en étant englobée dans les sorties observées au degré α fixé.

Le problème Conjonctif : Il s'agit dans ce cas de garantir par contraintes que l'intersection des sorties observées Y_j et prédites $\hat{Y}_j^{conj}$ n'est pas vide, c'est-à-dire, au degré α considéré :

$$[\hat{Y}_j^{conj}]_\alpha \cap [Y_j]_\alpha \neq \emptyset, j = 1, ..., M \tag{1.50}$$

Dans ce cas, le modèle est identifié en minimisant le critère choisi sous les contraintes (1.50) et

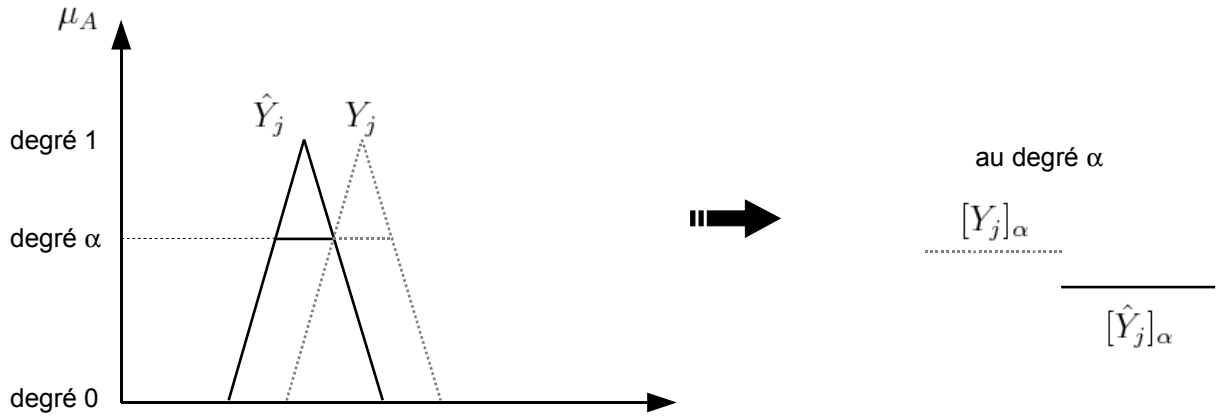


FIG. 1.15: Le problème Conjonctif

(1.47). Ainsi, il s'agit de déterminer les paramètres du modèle tels que sa sortie soit la moins imprécise, tout en ayant une intersection non nulle avec la sortie observée au degré α fixé.

Ces différentes formulations de la relation de la sortie du modèle identifié aux données observées ont quelques particularités importantes à noter et à mettre en perspective de la vision que nous avons retenue de la régression paramétrique linéaire floue.

Un point important à souligner concerne l'existence d'une solution, c'est-à-dire d'un modèle linéaire, pour chacun des problèmes d'optimisation dans le cas où le critère J_2 est utilisé [62]. Si l'on considère que l'utilisateur a fixé un degré $\alpha \in [0, 1[$ (le degré 1 est particulier, ce point sera plus largement étudié par la suite), il existe toujours une solution au problème minimal ([61], [66]) et au problème conjonctif [62]. Par contre, le problème maximal n'a pas toujours de solution. Il est montré dans [62] que l'existence d'une solution au problème maximal est liée à la caractéristique de la solution optimale du problème conjonctif. Ainsi, le problème maximal a une solution si et seulement si :

$$J_2(\mathbf{R}_\mathbf{A}^{conj}) = 0 \tag{1.51}$$

c'est-à-dire si la valeur optimale de critère pour le problème conjonctif est nulle. Ainsi, il est possible que les données ne permettent pas l'obtention d'un modèle dont la sortie est incluse au degré α dans toutes les observations.

De par la nature des relations aux données observées imposées par contraintes, l'imprécision des différents modèles identifiés (modèle de possibilité, de nécessité, ou conjonctif) diffère. En effet, dans le problème minimal, la totalité des imprécisions des sorties observées est capturée dans le modèle identifié, ce qui n'est pas le cas pour les autres formulations du problème de régression. Ainsi, le modèle de possibilité identifié est plus imprécis que le modèle de nécessité (lorsqu'il existe) et que le modèle conjonctif, qui ne prennent en compte qu'une partie de l'imprécision des sorties observées. Pour toute valeur fixée de $\alpha \in [0, 1[$, cela se traduit par les relations suivantes [62] :

$$\begin{array}{l} \text{Si } J_2(\mathbf{R}_\mathbf{A}^{conj}) = 0, \text{ alors } J_2(\mathbf{R}_\mathbf{A}^{max}) \leq J_2(\mathbf{R}_\mathbf{A}^{min}) \\ J_2(\mathbf{R}_\mathbf{A}^{conj}) \leq J_2(\mathbf{R}_\mathbf{A}^{min}) \end{array} \tag{1.52}$$

où $\mathbf{R}_\mathbf{A}^{min}$, $\mathbf{R}_\mathbf{A}^{conj}$ et $\mathbf{R}_\mathbf{A}^{max}$ (lorsqu'il existe) représentent respectivement les paramètres optimaux des problèmes minimal, conjonctif et maximal.

Finalement, lorsque le problème maximal admet une solution pour un α fixé, d'après la formulation des problèmes d'optimisation, il est évident que les contraintes d'inclusion suivantes sont satisfaites [61] :

$$[\hat{Y}_j^{max}]_\alpha \subseteq [Y_j]_\alpha \subseteq [\hat{Y}_j^{min}]_\alpha, j = 1, ..., M \tag{1.53}$$

La figure 1.16 illustre ces différents résultats théoriques dans le cas d'une unique entrée x et pour une valeur de α fixée. Les segments verticaux représentent les intervalles de sortie observés (intervalles conventionnels obtenus par α-coupe des intervalles flous initialement disponibles). La figure 1.16(a) concerne le cas où les trois problèmes admettent une solution. Les droites en trait plein, en tirets, et en alternance plein/pointillés représentent respectivement l'enveloppe de la sortie des modèles admissibles pour les problèmes minimal, maximal et conjonctif.

Conformément à (1.51), l'existence d'une solution au problème maximal induit que la solution du problème conjonctif est un modèle linéaire précis. De plus, la propriété d'inclusion (1.53) est bien vérifiée. Dans la figure 1.16(b), la simple modification de la deuxième sortie observée (segment vertical en trait gras) aboutit à l'inexistence d'une solution au problème maximal.

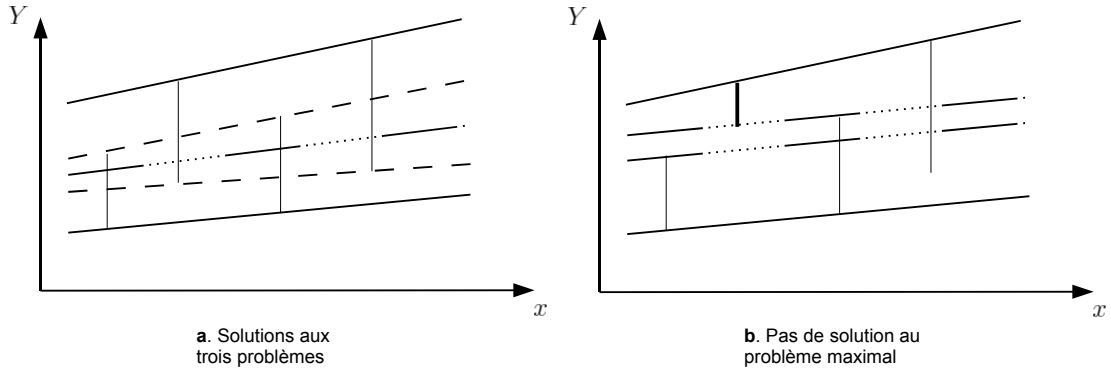


FIG. 1.16: Condition d'existence d'une solution aux différents problèmes

La potentielle inexistence du modèle maximal va à l'encontre de notre vision de la régression, puisqu'en aucun cas la technique régressive ne doit imposer une quelconque limitation sur les données à considérer, l'analyse de celles-ci étant l'objectif ultime de l'utilisateur. L'utilisation qui sera faite du modèle identifié est primordiale également pour déterminer quel problème doit être considéré. En effet, si l'utilisateur a pour objectif d'utiliser le modèle en prédiction une fois l'identification réalisée, il est important de prendre en considération l'intégralité des imprécisions observées. Ainsi, la sortie du modèle doit refléter celle des sorties observées via la contribution de ses paramètres imprécis. Dans ce cas, le modèle dit de possibilité, identifié à l'aide du problème minimal, sera le plus adéquat dans la mesure où rien de ce qui a été observé n'est négligé.

Synthèse 2 : Pour garantir l'obtention d'un modèle solution, nous nous focaliserons dans la suite sur la résolution du **problème minimal** qui s'exprime par :

$$\min_{\mathbf{R_A},\mathbf{M_A}} \; J_2(\mathbf{R_A}) \tag{1.54}$$

sous les contraintes :

$$R_{A_i} \geq 0, i = 0, ..., N \tag{1.55}$$

et

$$[Y_j]_\alpha \subseteq [\hat{Y}_j]_\alpha, j = 1, ..., M \tag{1.56}$$

en considérant les données décrites dans la synthèse 1.

1.4.2 Extension des moindres carrés

Cette deuxième approche de la régression floue, introduite par Diamond [17], repose sur la notion d'adéquation entre sortie du modèle et sorties observées. Dans ce contexte, l'identification d'un modèle flou fournissant une bonne représentation des données est basée sur l'utilisation d'une distance entre intervalles flous.

Si l'on considère deux intervalles flous triangulaires $A = (k_A, [S_A^-, S_A^+])$ et $B = (k_B, [S_B^-, S_B^+])$, la distance D_2 entre A et B définie par Diamond est la suivante :

$$D_2(A,B)^2 = (S_A^- - S_B^-)^2 + (k_A - k_B)^2 + (S_A^+ - S_B^+)^2 \tag{1.57}$$

Cette distance prend en considération les distances aux trois points d'intérêt des intervalles flous triangulaires, c'est-à-dire les bornes inférieure et supérieure du support, ainsi que la valeur modale. Dans [18], Diamond introduit une distance D_s plus simple à exploiter mais équivalente à D_2 dans le cas où les intervalles flous triangulaires sont symétriques. Ainsi, pour $A = [S_A^-, S_A^+]$ et $B = [S_B^-, S_B^+]$ deux intervalles flous triangulaires symétriques (cf convention des notation (1.28)),

$$D_s(A,B)^2 = (S_A^- - S_B^-)^2 + (S_A^+ - S_B^+)^2 \tag{1.58}$$

Dans le cadre de la synthèse 1, l'approche par moindres carrés consiste à minimiser la somme des distances entre les sorties observées et prédites. Cela se traduit par le problème d'optimisation

$$\min_{\mathbf{S_A^+}, \mathbf{S_A^-}, \mathbf{k_A}} \sum_{j=1}^{M} D_2(\hat{Y}_j, Y_j)^2 \tag{1.59}$$

dans le cas où tous les intervalles flous manipulés sont triangulaires et par le problème simplifié

$$\min_{\mathbf{S_A^+}, \mathbf{S_A^-}} \sum_{j=1}^{M} D_s(\hat{Y}_j, Y_j)^2 \tag{1.60}$$

lorsque les intervalles flous triangulaires sont symétriques.

L'existence d'une solution bien définie aux problèmes (1.59) ou (1.60) nécessite que les données soient compatibles avec la forme du modèle fixée. Dans [18], Diamond établit la condition de compatibilité des données avec un modèle flou linéaire dans le cas d'une seule variable d'entrée x. Pour le problème (1.60), cette condition dite de cohésion du jeu de données s'exprime :

$$\sum_{j=1}^{M} x_j (S_{Y_j}^+ - \bar{S_Y^+}) \geq \sum_{j=1}^{M} x_j (S_{Y_j}^- - \bar{S_Y^-}) \geq 0 \tag{1.61}$$

avec :

$$\begin{cases} \bar{S_Y^+} = 1/M \cdot \sum_{j=1}^{M} S_{Y_j}^+ \\ \bar{S_Y^-} = 1/M \cdot \sum_{j=1}^{M} S_{Y_j}^- \end{cases} \tag{1.62}$$

L'inégalité (1.61) traduit le fait que les tendances des bornes des supports des sorties observées, c'est-à-dire $S^-_{Y_j}$ et $S^+_{Y_j}$, $j = 1, ..., M$ ne doivent pas être opposées. Plus précisément, les supports des sorties observées doivent avoir une incertitude croissante pour des entrées positives, et décroissante pour des entrées négatives. Si la condition (1.61) est respectée, les paramètres du modèle identifié dans le problème (1.60) seront des intervalles flous bien définis, c'est-à-dire que leur Radius sera positif. Si les données ne respectent pas cette condition, les paramètres du modèle identifié ne seront pas des intervalles flous valides.

Dans l'approche possibiliste, conduisant à l'optimisation d'un critère linéaire sous contraintes, une condition de positivité des Radius est intégrée au jeu de contraintes, garantissant ainsi l'identification de paramètres flous bien définis. Dans cette approche de généralisation des moindres carrés, la condition formulée par Diamond [18] concerne les données, et non la technique d'identification en elle-même. Ainsi, dans le cas où les données ne respectent pas la contrainte de cohésion, aucun modèle valable ne peut être identifié. De notre point de vue, cette restriction sur les données qu'il est possible de représenter est en contradiction avec le choix initial d'un modèle régressif imprécis.

Pour remédier à cela, D'Urso [19, 20] envisage deux solutions. La première consiste à annuler le Radius des paramètres identifiés lorsque celui-ci est négatif. Ainsi, les paramètres en question deviennent précis, tandis que ceux ayant des Radius positifs sont inchangés. Dans l'absolu, cette méthode présente l'avantage de garantir l'identification d'un modèle flou défini, quelles que soient les données (la condition de cohésion (1.61) n'est plus nécessaire), sans changer la technique d'identification (critère quadratique inchangé). Cependant, un inconvénient majeur est que certains paramètres sont modifiés après identification. Ainsi, il n'est plus garanti que le modèle final soit optimal, du point de vue de son adéquation aux données. En effet, si la minimisation de la distance conduit à l'obtention de paramètres ayant un Radius négatif, la contribution de ceux-ci au Radius des sorties estimées ne peut être négligée sans perdre l'effet de compensation des Radius positifs des autres paramètres. Il est alors possible que le modèle modifié ait finalement une imprécision plus importante que nécessaire, ce qui nuit à sa qualité.

L'autre alternative proposée par D'Urso est l'ajout des contraintes (1.47) à la méthode d'identification par moindres carrés. Ainsi, la minimisation de la distance est réalisée sous contraintes de façon à garantir l'obtention de paramètres flous bien définis. Le problème d'identification est alors défini comme l'optimisation d'un critère quadratique basé sur la distance entre sorties observées et prédites, sous un ensemble de contraintes linéaires.

1.5 La régression paramétrique linéaire floue : limites et développements

Suite aux approches originales proposées par Tanaka et Diamond [18], de nombreux travaux ont été menés sur la régression linéaire floue dans l'optique de pallier certaines insuffisances fréquemment constatées dans la bibliographie. Ainsi, selon Redden et Woodall [51] ou encore Chang et Ayyub [48], il est essentiel d'étudier les aspects liés à l'imprécision trop importante des modèles linéaires flous identifiés ainsi que ceux relatifs aux relations entre données observées et prédictions.

Ces deux points sont discutés dans cette fin de chapitre dont l'objectif est de porter un regard critique sur les systèmes régressifs flous et d'exhiber leurs principaux défauts. Sans être exhaustif, quelques pistes d'amélioration issues de la littérature récente du domaine sont également évoquées de façon à positionner notre contribution qui sera présentée au chapitre 2.

1.5.1 Imprécision importante des modèles régressifs linéaires flous

Dans l'approche retenue de régression possibiliste (cf. synthèse 2), une des limites fondamentales concerne l'imprécision du modèle identifié, souvent jugée trop importante. Cette dernière dépend en premier lieu des données d'identification utilisées, et plus particulièrement de :

- leur amplitude,
- leur qualité.

La première origine de l'imprécision potentiellement trop élevée des modèles linéaires flous est liée à l'amplitude des entrées observées ([51]). En effet, si l'on considère l'expression de la sortie d'un modèle linéaire flou de la forme (1.9), son Midpoint et son Radius sont donnés par les équations (1.43) et (1.44). Le fait de dissocier le Midpoint et le Radius de la sortie du modèle permet de déterminer leur évolution en fonction de la variation de l'amplitude de l'entrée.

En ce qui concerne le Midpoint de la sortie, sa variation en fonction de chacune des composantes de l'entrée est donnée par :

$$\frac{dM_{\hat{Y}}}{dx_i} = M_{A_i}, i = 1, ..., N \tag{1.63}$$

Ainsi, l'équation (1.63) montre que le Midpoint de la sortie peut avoir une variation quelconque lorsque l'amplitude de l'entrée augmente. En effet, cette variation dépend du signe des Midpoints des paramètres A_i et ceux-ci peuvent être aussi bien positifs que négatifs. Par conséquent, la tendance de la sortie d'un modèle linéaire flou, modélisée par son Midpoint, peut être aussi bien croissante que décroissante relativement à chaque entrée.

En ce qui concerne le Radius, on obtient de la même manière :

$$\frac{dR_{\hat{Y}}}{dx_i} = R_{A_i} \cdot signe(x_i), i = 1, ..., N \tag{1.64}$$

Cette équation (1.64) montre que le Radius de la sortie du modèle a une variation dépendant du signe de l'entrée considérée. Sachant que les Radius des paramètres sont positifs puisque ces derniers sont des intervalles flous triangulaires bien définis, l'imprécision de la sortie du modèle ne peut qu'augmenter relativement à une entrée positive. A contrario, lorsque l'entrée est négative, l'imprécision de la sortie du modèle ne peut que décroître par rapport à cette dernière.

Cette limitation sur le sens de variation autorisé de l'imprécision de la sortie du modèle par rapport aux entrées est due au fait que les paramètres sont des intervalles flous. Par conséquent, elle peut être considérée comme une caractéristique intrinsèque des modèles linéaires flous de la forme (1.9).

Plusieurs solutions potentielles ont été proposées afin de remédier à cet inconvénient majeur des modèles linéaires flous, autant au niveau des techniques d'identification que de la structure même du modèle recherché.

En ce qui concerne les méthodes d'identification employées, nous avons vu que des contraintes garantissant la positivité des Radius des paramètres, et par propagation, de celui de la sortie, sont introduites dans les méthodes fondamentales, aussi bien dans le cadre de la régression dite possibiliste que dans l'approche basée sur l'extension des moindres carrés. Or, dans ce cas, l'imprécision de la sortie ne peut qu'augmenter avec l'amplitude de l'entrée, selon l'équation (1.64). Afin de remédier à cela, Wang et al. [68] proposent de ne pas introduire de contraintes sur les Radius, autorisant ainsi l'identification de paramètres ayant un Radius négatif si nécessaire. Ainsi, il est possible d'obtenir un modèle dont l'imprécision de la sortie peut être aussi bien croissante que décroissante quel que soit le signe de l'entrée. Dans une optique similaire, Nasrabadi et al. [45] et Modarres et al. [43] proposent d'introduire des contraintes garantissant l'obtention d'un modèle dont l'imprécision de la sortie est décroissante, quels que soient les signes des Radius des paramètres ainsi identifiés.

Ainsi, les techniques d'identification sont adaptées de manière à agir indirectement sur la structure du modèle, en permettant une plus grande flexibilité au niveau des paramètres identifiés. Cependant, ces approches présentent l'inconvénient majeur de ne plus garantir l'obtention de modèles flous au sens strict du terme, c'est-à-dire dont les paramètres sont des intervalles flous correctement définis. Ainsi, de notre point de vue, ces approches sortent du cadre formel de la régression linéaire floue.

Bien évidemment, il est également possible d'agir directement sur la structure du modèle que l'utilisateur détermine, et de développer les techniques d'identification appropriées. Nous avons vu que le phénomène d'augmentation de l'imprécision de la sortie d'un modèle linéaire flou en

rapport avec l'amplitude des entrées est dû aux paramètres du modèle, qui sont supposés être des intervalles flous. Ainsi, légitimement, il a été proposé de s'affranchir de cet aspect.

Ainsi, Lu [42] propose de chercher à identifier un modèle dont les coefficients sont précis, mais auquel on rajoute un terme d'erreur dont l'imprécision diminue en fonction de l'amplitude de l'entrée. Cette approche se retrouve dans [37], la méthode d'identification étant alors précédée d'une étape de défuzzification afin de faciliter l'identification des paramètres précis, ou encore dans [13], où la défuzzification est faite simultanément à l'identification proprement dite. Toujours dans l'idée d'identifier un tel modèle, une approche "inverse" a été proposée dans [14], où il s'agit de défuzzifier les paramètres du modèle identifié sur les données imprécises, avant ensuite d'intégrer au modèle le terme d'erreur, qui sera dans ce cas propre à chaque estimation.

Cependant, ces approches présentent l'inconvénient majeur de changer la structure du modèle à identifier, qui dans ce cas ne correspond plus à la forme présentée dans l'équation (1.9). Ainsi, on perd la notion d'imprécision vue comme une caractéristique intrinsèque du modèle, faisant partie intégrante de ses paramètres. De plus, le terme d'erreur peut potentiellement avoir une dispersion négative, et dans ce cas, celle-ci est annulée afin de respecter la définition d'ensemble flou [42]. Enfin, l'utilisation en prédiction du modèle identifié à l'aide de l'approche proposée dans [14] est délicate, le terme d'erreur étant défini pour chacun des échantillons observés, et non de manière globale.

Le deuxième point essentiel pour l'obtention d'un modèle flou d'imprécision raisonnable concerne la qualité des données d'identification. En effet, dans l'approche retenue de la synthèse 2, la sortie du modèle recherché doit contenir l'intégralité des données.

Des contraintes d'inclusion étant considérées, elles doivent bien entendu être respectées pour l'ensemble des données observées. Or, il est possible que l'ensemble d'apprentissage comporte des points aberrants, c'est-à-dire non représentatifs de la tendance globale décrite par la majorité des données, tendance devant au final être représentée au mieux par le modèle identifié. Ainsi, ces points aberrants, et notamment leur imprécision, doivent être également inclus dans la sortie prédite du modèle, afin de garantir le respect des contraintes. Par conséquent, les paramètres du modèle identifié peuvent être fortement influencés par quelques points aberrants non significatifs, ce qui est une limite forte à l'utilisation de modèles linéaires flous de possibilité en prédiction, et ce pour deux raisons principales :

- les paramètres identifiés en présence de points aberrants peuvent être fortement biaisés, la sortie du modèle peut donc avoir une tendance générale peu représentative de la majorité des données ;
- l'imprécision de la sortie du modèle est dégradée le cas échéant afin de respecter l'inclusion de points très minoritaires.

On remarquera ici que si d'autres types de relation entre sortie du modèle et données observées sont considérées, l'influence des points aberrants reste un phénomène important à étudier. En effet, dès que l'identification est réalisée à l'aide d'un problème d'optimisation sous contraintes, elle est sensible aux points aberrants, qui vont saturer les contraintes de manière inappropriée. De plus, bien que nous nous focalisons sur l'approche possibiliste, ce phénomène sera également existant dans les approches basées sur la programmation quadratique.

Une solution à envisager est bien entendu de chercher à supprimer ces points aberrants. Si l'on considère que les données d'apprentissage n'ont pas été prétraitées dans l'optique de supprimer les éventuels points aberrants, il peut donc être nécessaire de les détecter lors de la phase d'identification des paramètres afin de ne pas les prendre en compte. Dans cette optique, Hung et Yang. [33] proposent une méthode basée sur l'omission de points lors de l'identification. Dans le cadre d'une identification possibiliste, l'objectif est d'optimiser l'imprécision de la sortie du modèle recherché au sens de l'équation (1.46). Afin de détecter si un point observé est aberrant, chacune des données considérées est successivement retirée de l'ensemble des observations. Si l'amélioration que cette omission induit au niveau du critère est considérée comme significative, le point est considéré comme étant aberrant, et non retenu pour l'identification du modèle final. Cette méthode, quoique efficace, présente l'inconvénient majeur de nécessiter la répétition du processus d'identification autant de fois qu'il y a de données considérées pour l'apprentissage, puisque chacune d'entre elles est successivement étudiée. Or, si l'approche possibiliste est relativement simple à mettre en oeuvre et de complexité limitée, pour un grand nombre de données issues d'une application réelle, les temps de calcul peuvent devenir contraignants.

Une autre approche à envisager pour minimiser l'influence d'éventuels points aberrants est de chercher à identifier un modèle flou linéaire par morceaux, comme proposé par Yu et al. [72, 73]. Ainsi, les auteurs jouent ici sur la structure du modèle en lui-même. Chacun des sous-modèles linéaires composant le modèle global est identifié sur un sous ensemble des données d'identification. Celles-ci doivent donc être segmentée, et deux approches peuvent être distinguées :

- la segmentation peut être faite en amont de la phase d'identification des paramètres des sous-modèles proprement dite, celle-ci étant donc réalisée sur les sous ensembles de données définis par les points de rupture ainsi obtenus [72]
- ces points de rupture peuvent être recherchés lors de la phase d'identification, dans un processus de détection automatique [73].

Ainsi, en segmentant les données et en identifiant des sous-modèles linéaires sur chacun des sous ensembles obtenus, il est possible de limiter l'effet des points aberrants. En effet, ceux-ci introduisent des changements significatifs dans les tendances globales des données observées, et ils seront donc considérés comme des points de rupture. Les sous-modèles alors identifiés seront donc mieux représentatifs de la tendance globale des données sur chacun des segments ainsi déterminés. Dans ce cadre de modèles linéaires par morceaux, Roychowdhury et al. [53] se sont penchés sur la question de la continuité entre des sous-modèles linéaires identifiés sur des segments possiblement disjoints. Il est proposé de modéliser le comportement du modèle

entre ces segments afin d'assurer sa continuité en utilisant un ensemble de règles floues décrivant notamment les valeurs prises par le modèle aux bornes de cette plage de données absentes.

1.5.2 Relation entre données observées et prédictions

Dans la régression linéaire floue possibiliste (cf synthèse 2), le deuxième point fondamental outre la minimisation de l'imprécision du modèle est le respect de l'inclusion des observations dans les sorties de celui-ci, sorties supposées être des intervalles flous triangulaires.

Les approches étudiées sont basées sur l'utilisation des α-coupes afin de ne manipuler que des intervalles conventionnels. Ainsi, l'expression de la relation d'inclusion se fait sur des intervalles, à l'aide de l'équation (1.37). Cependant, l'objectif final est de fournir à l'utilisateur un modèle flou, il est donc nécessaire de reconstruire les paramètres flous identifiés à l'α-coupe choisie. Cette reconstruction est aisée, de par le fait que des intervalles flous triangulaires symétriques sont totalement définis par une unique α-coupe. Elle est par contre inexacte dans la mesure où les paramètres reconstruits ne sont pas optimaux vis-à-vis du problème d'optimisation initial.

Le problème majeur de cette approche est que l'inclusion, bien qu'imposée par contraintes au degré α choisi, ne peut pas être garantie à tous niveaux, c'est-à-dire $\forall \alpha \in [0, 1]$. En effet, en imposant l'inclusion à un degré α_0, celle-ci sera également obtenue pour les degrés inférieurs, c'est-à-dire $\forall \alpha_1 \leq \alpha_0$, mais pas nécessairement aux degrés supérieurs, c'est-à-dire $\forall \alpha_2 > \alpha_0$ (figure 1.17)

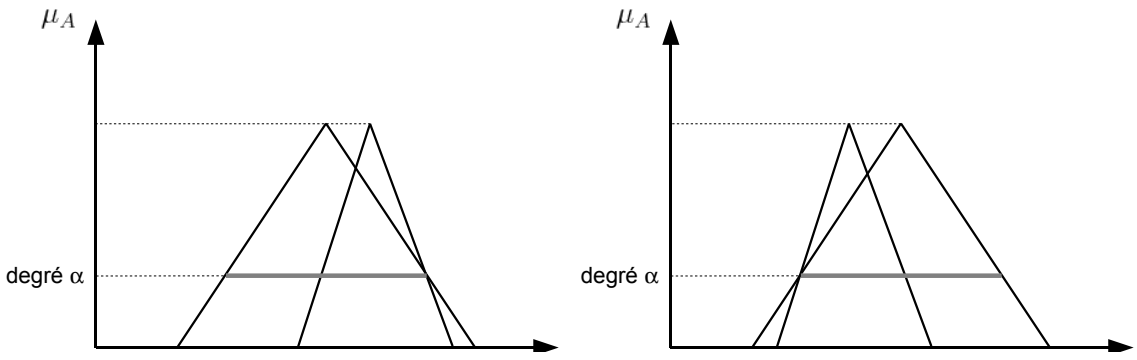


FIG. 1.17: Les limites de l'inclusion par α-coupes

Par conséquent, le choix du degré α considéré pour l'identification est un problème crucial [55], qu'il est peu approprié d'imposer à un utilisateur non spécialiste des techniques régressives floues. En effet, la réponse la plus évidente semble être de choisir un α_0 élevé pour l'identification, mais après reconstruction, cela conduit à l'obtention de paramètres (et donc d'un modèle) fortement imprécis [18]. De plus, l'identification ne peut pas être réalisée en considérant $\alpha_0 = 1$,

car l'obtention d'un modèle serait alors conditionnée au fait que les données observées soient alignées, ce qui n'est pas réaliste.

Indépendamment de l'inclusion, l'optimalité d'une solution identifiée au niveau α_0 n'est pas garantie à un niveau $\alpha' \neq \alpha_0$. Dans [66], Tanaka et al. explicite les formules de passage entre la solution optimale au niveau α_0 à celle optimale au niveau $\alpha' \neq \alpha_0$. Malheureusement la transformation proposée ne concerne que le cas de données observées précises.

Des travaux ont donc été menés dans l'optique de résoudre cet inconvénient majeur. Ils concernent essentiellement la technique d'identification mise en oeuvre, la structure du modèle linéaire étant conservée.

La plupart de ces approches sont basées sur la recherche d'un compromis entre inclusion et imprécision inhérente. Ainsi, dans [45], une solution itérative est proposée en vue de déterminer le bon compromis entre degré d'inclusion désiré et imprécision acceptable après reconstruction des intervalles flous. Dans [57], le degré α_0 est introduit dans le critère représentant l'imprécision de la sortie du modèle, afin de déterminer une solution optimale simultanément de l'imprécision et du degré d'inclusion. Dans [31], le critère linéaire à minimiser est basé sur les déviations entre les prédictions et les observations, l'inclusion totale n'étant alors plus nécessairement recherchée au degré α_0 considéré pour l'identification.

Cependant, ces approches, quoique permettant de ne plus impliquer l'utilisateur en amont de l'identification pour le choix du degré α_0 auquel elle sera menée, ne permettent pas d'obtenir un modèle respectant l'inclusion des observations dans les prédictions en terme d'intervalles flous. Qui plus est, la mise en place de tels critères complexifie fortement la méthode d'optimisation.

Outre le fait que l'inclusion à tous niveaux d'α-coupes est problématique, elle doit également concerner l'ensemble des données observées. Par conséquent, la qualité du jeu de données, et l'éventuelle présence de points aberrants en son sein, va également influer sur la saturation des contraintes d'inclusion, et de facto sur la qualité du modèle optimal obtenu.

Pour remédier à cela, Nasrabadi et al. [46] proposent d'intégrer des variables de relâchement dans le critère à optimiser dans une optique de programmation multi-objectifs. Ainsi, il est possible de ne pas respecter l'inclusion au niveau des points aberrants, en recherchant le meilleur compromis entre inclusion à maximiser et qualité du modèle à préserver. On remarquera ici cependant qu'il est nécessaire d'introduire dans le critère des coefficients de pondération, selon l'importance à accorder à l'un ou l'autre des objectifs. Ces coefficients doivent être déterminés a priori par l'utilisateur, ce qui nuit à la facilité d'utilisation de la méthode proposée. Qui plus est, la sortie du modèle identifié ne respecte pas l'inclusion totale des données qui est l'objectif d'un modèle de possibilité dans l'optique d'une utilisation en prédiction.

1.6 Conclusion

Dans un cadre de régression paramétrique linéaire en environnement imprécis, l'objectif est de fournir à un utilisateur une technique régressive adaptée permettant l'identification sur un jeu de données de la forme (1.8) de modèles linéaires (1.9) dont les paramètres sont des intervalles flous triangulaires.

L'approche possibiliste, reposant sur le principe de la minimisation de l'imprécision de la sortie du modèle est retenue. Qui plus est, la sortie du modèle recherché doit englober l'intégralité des informations disponibles dans les observations. Le problème minimal sera donc privilégié. De plus, afin de ne manipuler que des intervalles conventionnels afin de s'affranchir de tout calcul flou, le principe des α-coupes est adopté. Ainsi, l'identification est réalisée en optimisant un critère linéaire (1.54) sous un ensemble de contraintes d'inclusion des observations dans les prédictions (1.56), sachant que les paramètres obtenus doivent être des intervalles flous (1.55).

Cependant, cette approche présente des insuffisances majeures. En effet, l'imprécision du modèle, que l'on cherche à minimiser, est fortement impactée par l'amplitude des données observées, ainsi que par leur qualité, notamment au travers d'éventuels points aberrants. De plus, l'inclusion recherchée au travers de l'introduction de contraintes dans le problème d'optimisation ne peut être garantie au sens des intervalles flous, lorsque les α-coupes sont utilisées afin de simplifier la résolution de ce problème.

Dans la suite de ce travail, nous nous pencherons donc sur ces deux points essentiels de l'approche adoptée, c'est-à-dire minimisation de l'imprécision et garantie de l'inclusion, en ayant pour objectif principal de ne pas nuire à l'une pour favoriser l'autre.

Chapitre 2

La régression dans un environnement imprécis : propositions

Sommaire

2.1 Introduction

Dans ce chapitre, l'objectif est de présenter les différentes contributions apportées aux techniques régressives paramétriques linéaires dans un contexte d'imprécision. En cohérence avec les fonctionnalités que nous avons attribuées aux systèmes régressifs imprécis dans le chapitre précédent, une distinction sera faite entre les points concernant la structure du modèle fournie par l'utilisateur, et la technique d'identification à proprement parler, utilisée pour déterminer les paramètres de ce modèle.

Dans le chapitre précédent, deux insuffisances essentielles de la régression floue ont été mises en évidence. Elles concernent le non respect de l'inclusion totale des observations dans les prédictions et le manque de représentativité des modèles linéaires identifiés.

Notre objectif est maintenant de présenter les solutions retenues pour remédier à ces problèmes. Ces solutions peuvent être recherchées à deux niveaux, d'une part au niveau du choix de la structure du modèle linéaire à identifier, et d'autre part au niveau de la technique d'identification proprement dite.

Afin de clarifier les concepts développés, ceux-ci seront dans un premier temps présentés dans un cadre d'étude simplifié. Ainsi, le jeu de données considéré est composé d'échantillons n'ayant qu'une unique variable d'entrée précise et une sortie observée triangulaire floue. Il se présente donc sous la forme :

$$(x_j, Y_j), j = 1, ..., M \tag{2.1}$$

avec, selon l'équation (1.24) :

$$Y_j = (k_{Y_j}, (M_{S_{Y_j}}, R_{S_{Y_j}})) \tag{2.2}$$

Par conséquent, le modèle mono-entrée mono-sortie à identifier est de la forme :

$$\hat{Y} = A_0 \oplus A_1 \odot x \tag{2.3}$$

où les paramètres A_0 et A_1 sont des intervalles flous.

Une fois les solutions proposées introduites et discutées, leur généralisation au cas des données multi-entrées sera présentée. Diverses extensions seront également introduites.

2.2 La recherche de l'inclusion

Comme introduit dans le chapitre précédent, la recherche de l'inclusion des observations dans les prédictions, et ce, pour tout degré $\alpha \in [0, 1]$ est un problème épineux.

Quelques points fondamentaux peuvent résumer l'origine de la difficulté rencontrée.

La sortie du modèle identifié est supposée être de même nature que les observations, c'est-à-dire modélisée par des intervalles flous triangulaires symétriques. Cela permet de retenir le principe des α-coupes dans la formulation du problème d'optimisation, les intervalles flous triangulaires symétriques étant intégralement définis par une unique α-coupe. Par conséquent, les contraintes permettant d'exprimer la recherche de l'inclusion sont exprimées à un degré α fixé. Or, nous avons vu dans la section 1.5.2 que cela ne peut garantir l'inclusion des α-coupes de niveau supérieur. De plus, imposer les contraintes d'inclusion à un degré élevé nuit à l'imprécision du modèle, et la recherche de l'inclusion au degré maximal $\alpha = 1$ est impossible, sauf dans le cas très particulier de données colinéaires.

Adapter la technique d'identification pour résoudre ce problème ne semble donc pas être une solution acceptable. Ainsi, pour remédier à cela, nous proposons de travailler sur la structure même du modèle recherché. Il est ainsi proposé [3] de chercher à identifier un modèle dont les paramètres, et donc par propagation la sortie, sont des intervalles flous trapézoïdaux.

Ce choix est motivé par plusieurs raisons. Tout d'abord, en faisant ce choix d'intervalles flous trapézoïdaux, il est possible de conserver le formalisme lié aux intervalles conventionnels, et leur manipulation grâce à l'arithmétique des intervalles. En effet, la fonction d'appartenance étant toujours linéaire, l'utilisation des α-coupes est toujours possible. Ainsi, comme introduit dans la section 1.3.2, la sortie du modèle sera intégralement définie par ses intervalles Noyau et Support, c'est-à-dire, dans l'espace Midpoint / Radius :

$$\hat{Y} = ((M_{K_{\hat{Y}}}, R_{K_{\hat{Y}}}), (M_{S_{\hat{Y}}}, R_{S_{\hat{Y}}})) \tag{2.4}$$

avec :

$$\forall x : \begin{cases} M_{K_{\hat{Y}}} = M_{K_{A_0}} + M_{K_{A_1}}.x \\ R_{K_{\hat{Y}}} = R_{K_{A_0}} + R_{K_{A_1}}.|x| \\ M_{S_{\hat{Y}}} = M_{S_{A_0}} + M_{S_{A_1}}.x \\ R_{S_{\hat{Y}}} = R_{S_{A_0}} + R_{S_{A_1}}.|x| \end{cases} \tag{2.5}$$

Le deuxième point essentiel de cette approche est que le Noyau de la sortie du modèle n'est pas réduit à un point. Les sorties observées, représentées par des intervalles flous triangulaires, sont quant à elles unimodales. Ainsi, il est possible d'obtenir l'inclusion des sorties observées dans celles prédites au degré $\alpha = 1$:

$$[Y_j]_{\alpha=1} \subseteq [\hat{Y}_j]_{\alpha=1} \tag{2.6}$$

Bien entendu, il est également possible d'obtenir cette inclusion au degré $\alpha = 0$, de la même manière que lorsque l'on considère un modèle dont les paramètres sont des intervalles flous triangulaires :

$$[Y_j]_{\alpha=0} \subseteq [\hat{Y}_j]_{\alpha=0} \tag{2.7}$$

Combinée à la linéarité de la fonction d'appartenance, l'obtention de l'inclusion à ces deux niveaux de α-coupes permet de garantir que l'inclusion est réalisée au sens des intervalles flous, c'est-à-dire pour tout degré $\alpha \in [0,1]$:

$$[Y_j]_\alpha \subseteq [\hat{Y}_j]_\alpha, \forall \alpha \in [0,1] \tag{2.8}$$

2.2.1 La méthode d'identification

Dans l'approche possibiliste sur laquelle nous nous focalisons, il est nécessaire de définir la représentation de l'imprécision d'un modèle linéaire flou trapézoïdal, qui sera donc le critère à optimiser, ainsi que les contraintes à respecter pour garantir l'inclusion d'une part, et l'obtention de paramètres flous bien définis d'autre part. Sachant que la dimension verticale, c'est-à-dire le niveau α des α-coupes doit être considérée, il n'est plus réellement intéressant de se limiter à des sorties observées sous forme d'intervalles flous triangulaires symétriques. Par conséquent, la technique d'identification présentée par la suite le sera dans le cas plus général d'observations sous forme d'intervalles flous triangulaires non nécessairement symétriques.

2.2.1.1 Le critère

Dans le critère introduit par Tanaka pour un modèle linéaire flou triangulaire symétrique (1.46), l'imprécision du modèle est représentée par la somme des Radius des sorties triangulaires prédites, une unique α-coupe (classiquement, le Support des intervalles flous triangulaires considérés) étant suffisante pour parfaitement définir les paramètres et la sortie du modèle.

Si un modèle flou trapézoïdal est recherché, il est nécessaire de prendre en considération la dimension verticale en manipulant non plus une, mais deux α-coupes. L'extension naturelle est alors de définir le critère comme la somme des aires des sorties trapézoïdales prédites.

L'aire de la sortie trapézoïdale prédite pour l'entrée x_j est donnée [70] dans l'espace des bornes par :

$$aire(\hat{Y}(x_j)) = \frac{K^+_{\hat{Y}_j} + S^+_{\hat{Y}_j}}{2} - \frac{K^-_{\hat{Y}_j} + S^-_{\hat{Y}_j}}{2} \tag{2.9}$$

Il est également possible de l'exprimer dans l'espace Midpoint / Radius par :

$$aire(\hat{Y}(x_j)) = R_{K_{\hat{Y}_j}} + R_{S_{\hat{Y}_j}} \tag{2.10}$$

Par conséquent, le critère J_2 (équation (1.46)) étendu au cas des modèles trapézoïdaux J_{2Trap} est défini comme suit :

$$J_{2Trap}(\mathbf{R_{K_A}}, \mathbf{R_{S_A}}) = \sum_{j=1}^{M} aire(\hat{Y}(x_j)) \tag{2.11}$$

où $\mathbf{R_{K_A}}$ et $\mathbf{R_{S_A}}$ représentent les vecteurs regroupant les Radius respectivement des noyaux et des supports des différents paramètres.

En introduisant dans l'expression (2.11) les relations (2.10) et (2.5), le critère à optimiser est donc :

$$J_{2Trap}(\mathbf{R_{K_A}}, \mathbf{R_{S_A}}) = M \cdot (R_{S_{A_0}} + R_{K_{A_0}}) + (R_{S_{A_1}} + R_{K_{A_1}}) \cdot \sum_{j=1}^{M} |x_j| \tag{2.12}$$

2.2.1.2 Les contraintes

L'optimisation du critère J_{2Trap} (2.12) doit se faire sous un ensemble de contraintes, permettant de rechercher l'inclusion désirée et de garantir l'obtention de paramètres sous la forme d'intervalles flous bien définis.

Comme précisé précédemment, garantir l'inclusion des observations dans les prédictions aux degrés $\alpha = 0$ et $\alpha = 1$ permet de la garantir au final à tout degré $\alpha \in [0, 1]$. Ainsi, dans le problème d'optimisation, les contraintes (2.6) et (2.7) doivent être introduites.

- Inclusion des noyaux :
  Il faut garantir l'inclusion du noyau (réduit à un point) k_{Y_j} des sorties observées dans l'intervalle noyau $[K^-_{\hat{Y}_j}, K^+_{\hat{Y}_j}]$ des sorties prédites, et ce, $\forall j = 1, ..., M$:

  $$k_{Y_j} \in [K^-_{\hat{Y}_j}, K^+_{\hat{Y}_j}] \Leftrightarrow |M_{K_{\hat{Y}_j}} - k_{Y_j}| \leq R_{K_{\hat{Y}_j}} \tag{2.13}$$

  où $M_{K_{\hat{Y}_j}}$ et $R_{K_{\hat{Y}_j}}$ sont définis dans l'équation (2.5).
- Inclusion des supports :
  Il faut garantir l'inclusion de l'intervalle support $[S^-_{Y_j}, S^+_{Y_j}]$ des sorties observées dans l'intervalle support $[S^-_{\hat{Y}_j}, S^+_{\hat{Y}_j}]$ des sorties prédites, et ce, $\forall j = 1, ..., M$:

  $$[S^-_{Y_j}, S^+_{Y_j}] \subseteq [S^-_{\hat{Y}_j}, S^+_{\hat{Y}_j}] \Leftrightarrow |M_{S_{\hat{Y}_j}} - M_{S_{Y_j}}| \leq R_{S_{\hat{Y}_j}} - R_{S_{Y_j}} \tag{2.14}$$

  où $M_{S_{\hat{Y}_j}}$ et $R_{S_{\hat{Y}_j}}$ sont définis dans l'équation (2.5).

D'autres contraintes permettant de garantir l'obtention d'intervalles flous correctement définis sont également à considérer.

- Inclusion du noyau dans le support :
  Il est nécessaire de s'assurer que la sortie du modèle est représentée par un intervalle flou trapézoidal, et donc de garantir l'inclusion des noyaux $[K^-_{\hat{Y}_j}, K^+_{\hat{Y}_j}]$ des sorties prédites dans leur support $[S^-_{\hat{Y}_j}, S^+_{\hat{Y}_j}]$. Cela peut être fait en garantissant cette propriété d'inclusion au niveau des paramètres A_0 et A_1, c'est-à-dire, pour $i = \{0, 1\}$:

$$[K_{A_i}^-, K_{A_i}^+] \subseteq [S_{A_i}^-, S_{A_i}^+] \Leftrightarrow |M_{S_{A_i}} - M_{K_{A_i}}| \leq R_{S_{A_i}} - R_{K_{A_i}} \tag{2.15}$$

- Positivité des Radius des paramètres
  Il est nécessaire de s'assurer que les paramètre A_0 et A_1 soient des intervalles flous, c'est-à-dire que les Radius des intervalles aux deux α-coupes considérées doivent être positifs ou nuls :

$$R_{K_{A_0}} \geq 0 \text{ , et } R_{S_{A_0}} \geq 0 \tag{2.16}$$

$$R_{K_{A_1}} \geq 0 \text{ , et } R_{S_{A_1}} \geq 0 \tag{2.17}$$

Synthèse 3 : Afin de garantir l'inclusion des observations dans les sorties prédites d'un modèle régressif linéaire flou, il est proposé d'utiliser un modèle dont les paramètres, et par propagation, la sortie, sont des intervalles flous trapézoïdaux. Le problème d'optimisation correspondant est exprimé par :

$$\min_{\mathbf{R_{K_A}}, \mathbf{R_{S_A}}, \mathbf{M_{K_A}}, \mathbf{M_{S_A}}} J_{2Trap}(\mathbf{R_{K_A}}, \mathbf{R_{S_A}}) \tag{2.18}$$

sous les contraintes d'inclusion :

$$[Y_j]_{\alpha=1} \subseteq [\hat{Y}_j]_{\alpha=1} \tag{2.19}$$

$$[Y_j]_{\alpha=0} \subseteq [\hat{Y}_j]_{\alpha=0} \tag{2.20}$$

et les contraintes assurant l'obtention d'intervalles flous (2.15), (2.16) et (2.17).

2.2.2 Discussion

Cette approche visant à identifier un modèle trapézoïdal permet d'obtenir l'inclusion à tout degré α. Il est intéressant de constater ici qu'elle peut être vue comme une généralisation à moindre coût de l'approche classique consistant à déterminer un modèle triangulaire symétrique (cf. synthèse 2). Cela peut être souligné aussi bien au niveau de la structure du modèle, de la méthode d'identification, et de l'interprétation du modèle obtenu.

En ce qui concerne la structure du modèle, le fait de considérer des coefficients flous trapézoïdaux et non plus triangulaires symétriques revient à introduire deux nouveaux paramètres par coefficient, en l'occurrence, le Midpoint et le Radius de leur noyau. Cela permet

donc de disposer d'un noyau sous forme d'intervalle, et non plus réduit à un point, et donc d'obtenir l'inclusion des observations dans les prédictions au degré $\alpha = 1$. Ainsi, l'introduction de ces deux nouveaux paramètres permet d'une part de positionner le noyau de la sortie du modèle (via le Midpoint), et d'autre part de quantifier l'imprécision au niveau du noyau des paramètres du modèle (via le Radius). Ce dernier point est en accord avec notre vision de l'imprécision comme caractéristique intrinsèque du modèle. Plutôt que de voir cela comme un ajout d'imprécision dans la structure du modèle, nous préférerons donc parler de meilleure représentation.

L'ajout de deux paramètres supplémentaires a bien entendu une influence sur la méthode d'identification proposée, aussi bien au niveau du critère que des contraintes. On remarquera préalablement que cela n'impacte en rien l'approche par α-coupes permettant la manipulation d'intervalles conventionnels au travers de l'arithmétique associée. Cependant, l'utilisateur n'a plus à déterminer la valeur de α à laquelle réaliser l'identification, les deux niveaux considérés étant fixés définitivement comme étant le support et le noyau des intervalles flous considérés. Ce point est important, car synonyme de simplicité de mise en oeuvre, priorité mise en avant dans notre vision de la problématique de la régression linéaire floue.

La représentation retenue de l'imprécision (équation (2.11)) permet de considérer la dimension verticale de la sortie du modèle en minimisant l'aire des intervalles flous trapézoïdaux. Le critère finalement obtenu (2.12) reste linéaire et généralise celui utilisé pour identifier un modèle triangulaire symétrique (1.46). En effet, il s'agit au final de minimiser une somme pondérée des Radius des intervalles considérés, non plus à un unique degré, mais aux deux considérés simultanément. Si le Radius des noyaux des paramètres est nul, le critère (2.12) a la même expression que le critère (1.46). Ainsi, les modifications apportées au niveau du critère semblent être assez facilement assimilables par un utilisateur potentiellement intéressé par la mise en oeuvre de l'approche trapézoïdale.

En ce qui concerne les contraintes, la philosophie générale est inchangée. La seule différence notable est que l'inclusion doit être respectée à deux niveaux distincts, ce qui induit un plus grand nombre de contraintes. En effet, selon la synthèse 2 concernant l'identification de modèles triangulaires, une unique contrainte d'inclusion est associée à chaque exemple alors qu'il en faut deux dans le cas d'un modèle trapézoïdal (synthèse 3). De la même manière, le nombre de contraintes associées à l'obtention d'intervalles bien définis est doublé. Enfin, l'introduction de la contrainte (2.15) pour chaque paramètre permet de lever l'hypothèse de symétrie sur la sortie du modèle.

Globalement, en terme de complexité du problème d'optimisation, le coût reste somme toute modéré : facteur deux sur le nombre de paramètres à identifier et sur le nombre de contraintes à satisfaire.

Le point le plus délicat se situe au niveau de l'interprétation du modèle et de son usage en

prédiction. En effet, la différence de forme entre sorties prédites (trapèzes) et sorties observées (triangles) peut paraître contre-intuitive. Certains éléments de réflexion permettent cependant de la justifier. Tout d'abord, dans le cas particulier où la valeur modale des sorties observées suit parfaitement un modèle linéaire précis, le noyau du modèle identifié sera lui aussi précis. Autrement dit, le modèle identifié sera triangulaire si les observations le justifient. Dans le cas contraire, le noyau du modèle identifié reflète la variabilité des modes observés et le support leur imprécision.

Il est intéressant pour clore cette discussion d'introduire les travaux de Lee et Tanaka [40], [64], à notre connaissance les seuls abordant l'identification d'un modèle trapézoïdal. Ces travaux se focalisent sur le cas où les observations sont des intervalles conventionnels, éventuellement obtenus par α-coupe à un niveau α fixé d'observations floues. Les auteurs proposent d'identifier deux modèles distincts dont les paramètres sont des intervalles conventionnels :

- un modèle dit inférieur ;
- un modèle dit supérieur,

puis d'unifier ces deux modèles pour construire un modèle flou trapézoïdal. Le modèle inférieur est alors utilisé comme noyau du modèle final, alors que le modèle supérieur est associé au support de ce dernier. Deux principes différents sont proposés pour garantir l'inclusion du modèle inférieur (noyau) dans le modèle supérieur (support).

Le premier exploité dans [64] consiste à identifier un modèle de possibilité pour le modèle supérieur, et un modèle de nécessité pour le modèle inférieur (cf section 1.4.1.3 du chapitre I). Si le modèle de nécessité existe, il est effectivement inclus dans le modèle de possibilité. Si ce n'est pas le cas, Tanaka et Lee préconisent l'utilisation de modèles polynomiaux.

Le second principe proposé dans [40] consiste à identifier deux modèles de possibilité, mais avec deux jeux de données différents. Le modèle supérieur est alors identifié avec l'intégralité des échantillons disponibles, alors que seul un sous-ensemble de ceux-ci est utilisé pour l'identification du modèle inférieur. La sélection de ce sous-ensemble est réalisé selon une approche statistique par quantile.

Dans les deux cas, il est difficile de donner une interprétation au modèle flou final. Si le noyau de la sortie peut être vu comme le cas le plus optimiste, et son support comme le cas le plus défavorable, aucune signification ne peut être attribuée aux coupes de niveau intermédiaire.

Notre approche est donc fondamentalement différente de celles proposées par Tanaka et Lee. En effet, l'introduction de paramètres trapézoïdaux dans le modèle flou permet de garantir l'inclusion à tous niveaux d'α-coupe, et ainsi de donner un sens relatif à un niveau d'inclusion pour α variant de 0 à 1.

2.2.3 Exemple illustratif

L'objectif de cette section est d'illustrer sur un exemple simple les bénéfices de l'adoption des modèles trapézoïdaux en terme d'inclusion des observations dans les prédictions.

2.2.3.1 Les indicateurs de performance considérés

Afin de comparer différents modèles linéaires flous, il est nécessaire de disposer d'indicateurs permettant de refléter la qualité des modèles identifiés. Deux axes principaux peuvent être dégagés, selon que l'on s'attache aux caractéristiques intrinsèques du modèle, ou à l'adéquation de sa sortie aux données observées.

L'objectif essentiel de l'approche retenue en régression linéaire floue est la minimisation de l'imprécision de la sortie du modèle, qui devra donc être considérée comme une grandeur importante pour la comparaison de modèles linéaires flous. Pour ce faire, l'indicateur retenu est l'expression de l'imprécision de la sortie du modèle défini au préalable comme critère d'optimisation. Sachant que l'on cherche à caractériser aussi bien des modèles flous triangulaires que trapézoïdaux, il faut considérer l'indicateur le plus général, c'est-à-dire J_{2Trap} (équation (2.11)).

En ce qui concerne l'adéquation de la sortie du modèle aux données, deux aspects peuvent être mis en avant dans l'étude comparative de modèles régressifs.

Premièrement, la notion de distance, indicateur utilisé comme critère dans les approches basées sur une extension des moindres carrés au contexte imprécis, peut être considérée. Il s'agira donc ici de quantifier l'erreur quadratique entre observations et prédictions, en utilisant une distance entre intervalles flous.

Dans le premier chapitre, la distance de Diamond a été restreinte au cas des intervalles flous triangulaires. Lorsqu'il s'agit d'évaluer une distance entre la sortie d'un modèle trapézoïdal et des données triangulaires, il est nécessaire d'utiliser une distance adaptée aux intervalles flous trapézoïdaux. Dans [17], Diamond propose la distance suivante :

$$D(A,B) = (S_A^- - S_B^-)^2 + (K_A^- - K_B^-)^2 + (K_A^+ - K_B^+)^2 + (S_A^+ - S_B^+)^2 \tag{2.21}$$

où A et B sont deux intervalles trapézoïdaux. Dans notre cas, on obtient alors comme expression de l'indicateur de distance :

$$Distance = \sum_{j=1}^{M} D(\hat{Y}_j, Y_j) = \sum_{j=1}^{M} (S_{\hat{Y}_j}^- - S_{Y_j}^-)^2 + (K_{\hat{Y}_j}^- - k_{Y_j})^2 + (K_{\hat{Y}_j}^+ - k_{Y_j})^2 + (S_{\hat{Y}_j}^+ - S_{Y_j}^+)^2 \tag{2.22}$$

Il peut également être intéressant de quantifier l'erreur quadratique entre les valeurs modales des observations et des prédictions. Dans le cas de l'approche triangulaire (synthèse 2), elle est

définie par :

$$erreur = \sum_{j=1}^{M} (k_{Y_j} - k_{\hat{Y}_j})^2 \tag{2.23}$$

Dans le cas de l'approche trapézoïdale (synthèse 3), elle est définie en ne conservant que les Midpoints des intervalles noyaux, soit :

$$erreur = \sum_{j=1}^{m} (k_{Y_j} - M_{K_{\hat{Y}_j}})^2 \tag{2.24}$$

Le deuxième point à considérer est un indicateur susceptible de refléter l'inclusion des observations dans les prédictions, et plus particulièrement le "taux" d'inclusion. En effet, celui-ci est une caractéristique essentielle que nous cherchons à améliorer dans ces travaux. Pour le quantifier, nous considérons l'indicateur de compatibilité entre observations et prédictions défini dans [27] comme suit :

$$comp(\hat{Y}_j, Y_j) = \max_x \min(\mu_{\hat{Y}_j}(x), \mu_{Y_j}(x))) \tag{2.25}$$

Cette compatibilité est le degré α maximum correspondant à une intersection non vide des α-coupes. Il est plus simple de calculer un "taux" d'intersection entre intervalles flous qu'un "taux" d'inclusion. Par ailleurs, ce choix de simplicité n'influe pas sur la méthode de régression, puisque cet indicateur est uniquement utilisé pour la comparaison de modèles identifiés.

Afin d'obtenir un indicateur exploitable dans une comparaison de modèles régressifs, nous considérerons finalement la moyenne de ce degré de compatibilité pour tous les échantillons, de manière similaire à [27] c'est-à-dire :

$$Compatibilite = \frac{\sum_{j=1}^{M} comp(\hat{Y}_j, Y_j)}{M} \tag{2.26}$$

Ainsi, une valeur de compatibilité égale à 1 est une condition nécessaire mais non suffisante à l'inclusion totale. Cet indicateur est donc utilisé pour refléter un défaut d'inclusion plus ou moins important.

2.2.3.2 Les données et les modèles identifiés

Le comparatif en lui même est mené sur un jeu de données simple fréquemment rencontré dans la littérature ([31], [64]) et présenté dans le tableau 2.1. On remarquera ici que la représentation des intervalles par leurs bornes a été adoptée, permettant une meilleure mise en lumière des points détaillés par la suite. Le jeu de données comporte donc $M = 8$ échantillons. Chacun d'entre eux est constitué d'une entrée précise et d'un intervalle flou triangulaire de sortie. Celui-ci est défini par la valeur ponctuelle de son noyau, et son intervalle de support. Dans notre cadre d'étude, l'objectif de la technique de régression linéaire floue est d'identifier un modèle d'imprécision minimale englobant l'ensemble des observations.

j	x_j	Y_j	α-coupes de niveau $\alpha = 0.9$
1	0.1	$(2.25, [1.5, 3])$	$[2.175, 2.325]$
2	0.2	$(2.875, [2, 3.75])$	$[2.788, 2.962]$
3	0.3	$(2.5, [1.5, 3.5])$	$[2.4, 2.6]$
4	0.4	$(4.25, [2.5, 6])$	$[4.075, 4.425]$
5	0.5	$(4.0, [2.5, 5.5])$	$[3.85, 4.15]$
6	0.6	$(5.25, [4, 6.5])$	$[5.125, 5.375]$
7	0.7	$(7.5, [5.5, 9.5])$	$[7.3, 7.7]$
8	0.8	$(8.5, [7, 10])$	$[8.35, 8.65]$

TAB. 2.1: Le jeu de données observées et α-coupes de niveau $\alpha = 0.9$

Les modèles triangulaire et trapézoïdal sont identifiés en minimisant le critère J_2 pour le premier, et J_{2Trap} pour le second, sous l'ensemble des contraintes adéquates. L'identification du modèle triangulaire se fait sur les α-coupes de niveau $\alpha = 0$ des sorties observées. Les paramètres obtenus pour chacun des modèles sont présentés dans les tableaux 2.2 et 2.3, selon qu'une représentation dans l'espace des bornes ou dans l'espace Midpoint / Radius est adoptée.

	Modèle triangulaire ($\alpha = 0$)	Modèle trapézoïdal
A_0	$(0.96, [0, 1.92])$	$([0.25, 1.36], [0, 1.92])$
A_1	$(7.92, [5, 10.83])$	$([7.5, 8.93], [5, 10.83])$

TAB. 2.2: Paramètres des modèles obtenus (représentation dans l'espace des bornes)

	Modèle triangulaire ($\alpha = 0$)	Modèle trapézoïdal
A_0	$(0.96, 0.96)$	$((0.805, 0.555), (0.96, 0.96))$
A_1	$(7.915, 2.915)$	$((8.215, 0.715), (7.915, 2.915))$

TAB. 2.3: Paramètres des modèles obtenus (représentation dans l'espace Midpoint/Radius)

Les différents indicateurs introduits dans la section précédente sont ensuite calculés pour chacun des deux modèles identifiés et présentés dans le tableau 2.4.

Enfin, une représentation des modèles identifiés est proposée sur les figures 2.1 et 2.2. Celles-

	Modèle triangulaire ($\alpha = 0$)	Modèle trapézoïdal
J_{2Trap}	18.29	25.17
$Distance$	35.76	48.08
$erreur$	4.57	4.43
$Compatibilite$	0.83	1

TAB. 2.4: Indicateurs associés aux modèles identifiés

ci présentent les données triangulaires floues (intervalles verticaux pour le support et cercle pour le noyau ponctuel) ainsi que les supports (traits pleins) et noyaux (traits discontinus) des modèles identifiés.

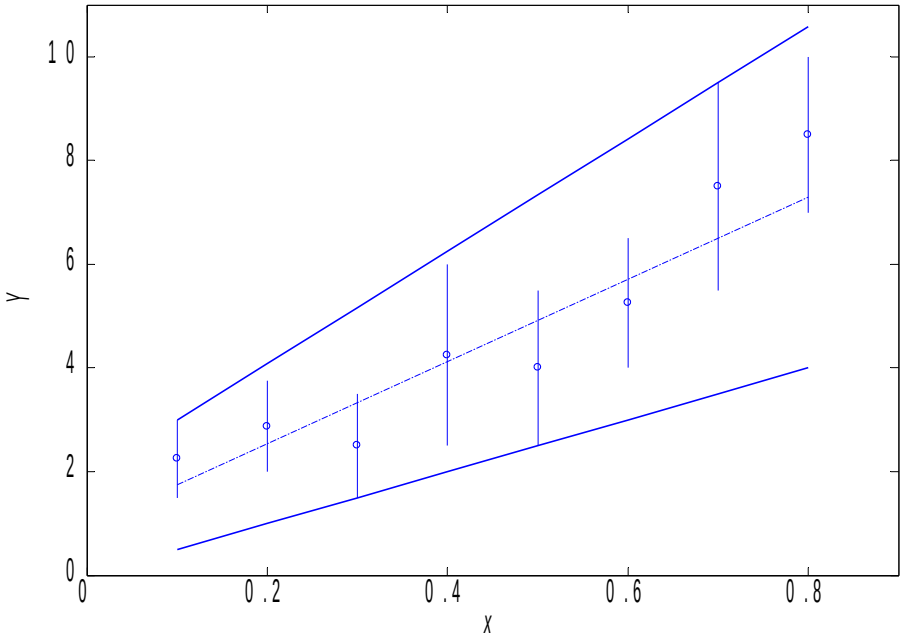


FIG. 2.1: Modèle flou triangulaire identifié ($\alpha = 0$)

On constate l'égalité des supports pour les deux modèles identifiés (cf. tableau 2.2) ainsi que l'inclusion des observations dans les prédictions au niveau des supports (garantie par contraintes dans les deux cas).

La différence notable entre les deux modèles se situe au niveau des noyaux. Dans le cas du modèle triangulaire (cf. figure 2.1), le noyau précis de la sortie correspond à la droite d'équation (2.27) :

$$k_{\hat{Y}} = 0.96 + 7.92 \cdot x \tag{2.27}$$

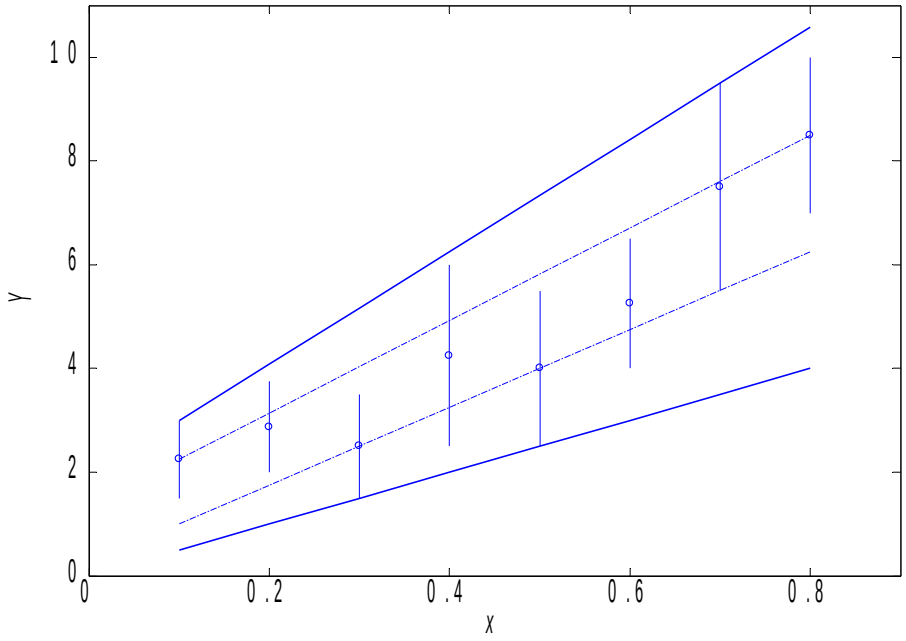


FIG. 2.2: Modèle flou trapézoïdal identifié

qui, si elle offre une bonne approximation de la tendance globale des valeurs modales observées, ne permet bien évidemment pas l'inclusion. A contrario, le noyau de la sortie du modèle trapézoïdal (cf. figure 2.2) est un intervalle, respectant l'inclusion de tous les noyaux des observations. Son Midpoint est donné par l'équation (2.28) :

$$M_{K_{\hat{Y}}} = 0.8 + 8.21 \cdot x \tag{2.28}$$

L'indicateur de compatibilité du modèle trapézoïdal, de valeur $Compatibilite = 1$, illustre le fait que l'inclusion est bien respectée au niveau $\alpha = 1$, ce qui est suffisant pour garantir l'inclusion à tout niveau α. Une vision tri-dimensionnelle des modèles triangulaire (figure 2.3) et trapézoïdal (figure 2.4) et du jeu de données met clairement en évidence une inclusion totale dans le cas trapézoïdal et uniquement partielle dans le cas triangulaire.

Cet état de fait est encore plus visible si l'on examine un échantillon particulier, par exemple le premier ($j = 1$) du tableau 2.1. Dans ce cas, la sortie prédite avec le modèle triangulaire est :

$$\hat{Y}_1 = (1.75, [0.5, 3]) \tag{2.29}$$

alors que celle prédite avec le modèle trapézoïdal est :

$$\hat{Y}_1 = ([1, 2.25], [0.5, 3]) \tag{2.30}$$

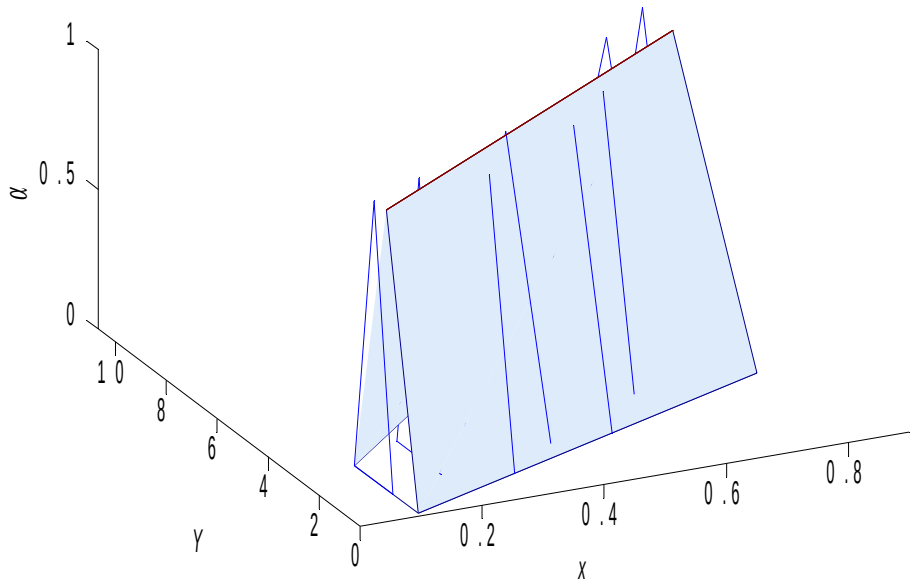


FIG. 2.3: Modèle flou triangulaire identifié - vue tridimensionnelle

Les figures 2.5 et 2.6 représentent l'observation $j = 1$ ainsi que les prédictions triangulaire et trapézoïdale (2.29) et (2.30). La figure 2.5 met en évidence un indice de compatibilité de 0.75 pour le modèle triangulaire, alors que l'inclusion n'est respectée qu'au niveau $\alpha = 0$ (saturation de la contrainte d'inclusion relative à la borne maximale du support). Cet exemple particulier illustre clairement qu'une valeur de compatibilité globale relativement élevée (0.83) mais différente de 1 ne garantit en rien un niveau d'inclusion supérieur à 0.

En dehors de la compatibilité, les indicateurs du tableau 2.4 reflètent une qualité moins bonne du modèle trapézoïdal (plus imprécis, moins en adéquation aux données) que du modèle triangulaire. Comme attendu, l'inclusion totale s'obtient au détriment des autres caractéristiques du modèle.

Pour une comparaison réaliste en terme d'imprécision des modèles triangulaire et trapézoïdal, il est donc nécessaire de pouvoir garantir un niveau minimum d'inclusion similaire pour les deux types de modèles. Malheureusement, comme discuté dans le premier chapitre, il n'est généralement pas possible de garantir une inclusion au niveau $\alpha = 1$ avec un modèle triangulaire, comme c'est le cas avec un modèle trapézoïdal. Par contre, le choix d'un α élevé (mais inférieur strictement à 1) dans le problème minimal (synthèse 2) garantit une inclusion à ce niveau α. Pour augmenter le niveau d'inclusion garanti du modèle triangulaire, celui-ci est maintenant identifié avec des contraintes d'inclusion sur les α-coupes au niveau $\alpha = 0.9$ (cf. tableau 2.1), et non plus

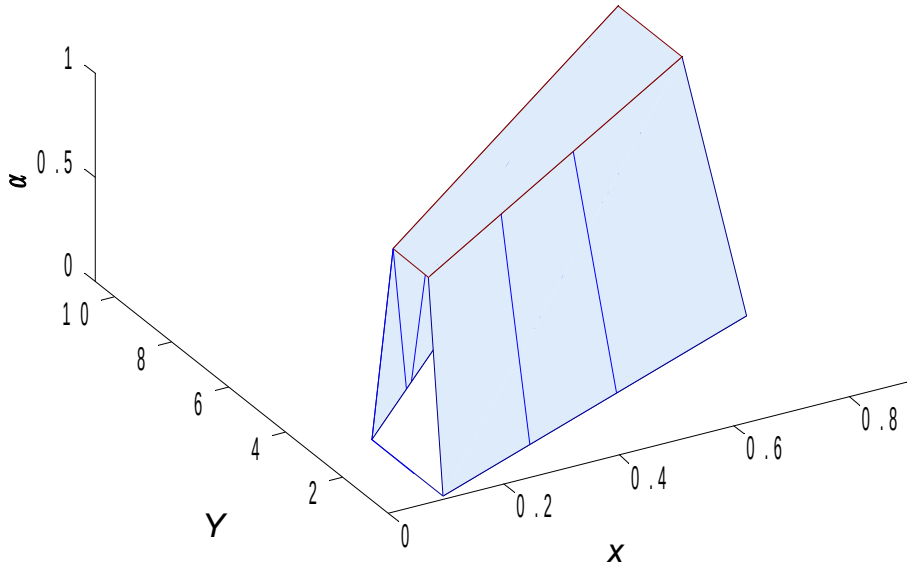


FIG. 2.4: Modèle flou trapézoïdal identifié - vue tridimensionnelle

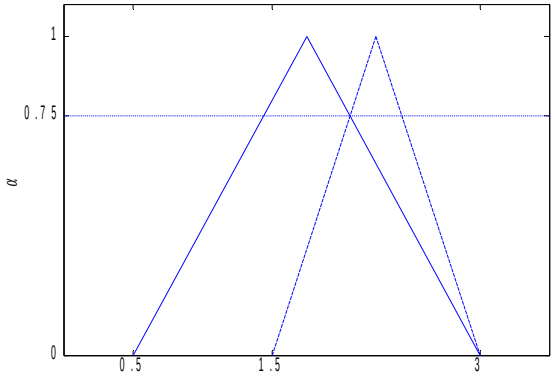


FIG. 2.5: Sorties observée et prédite ($j = 1$) pour le modèle triangulaire identifié à $\alpha = 0$

au niveau $\alpha = 0$ considéré précédemment. Les paramètres du modèle obtenu (représentation dans l'espace Midpoint / Radius), ainsi que les indicateurs de performance qui lui sont associés sont disponibles dans le tableau 2.5.

Comme précédemment, le cas particulier de l'observation $j = 1$ est détaillé. La sortie trian-

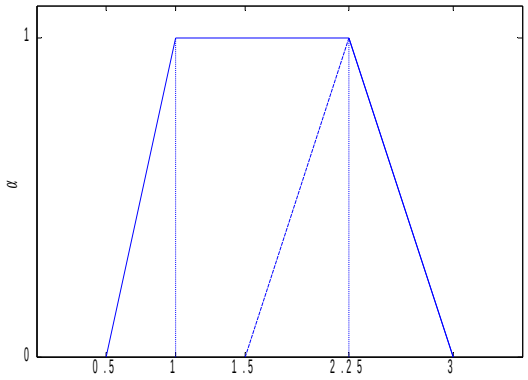


FIG. 2.6: Sorties observée et prédite ($j = 1$) pour le modèle trapézoïdal

	Modèle triangulaire ($\alpha = 0.9$)
A_0	$(0.82, 5.98)$
A_1	$(8.14, 8.53)$
$Compatibilite$	0.92
$Distance$	158.77
J_{2Trap}	79.98

TAB. 2.5: Paramètres du modèle triangulaire identifié à $\alpha = 0.9$ et indicateurs associés

gulaire identifiée est maintenant :

$$\hat{Y}_1 = (1.63, [-5.24, 8.51]) \tag{2.31}$$

Cette dernière est visualisée sur la figure 2.7 qui illustre un indice de compatibilité de 0.92.

Si, comme cela était recherché, l'inclusion a bien été augmentée (valeur de l'indicateur $compatibilite = 0.92$), il est évident à la vue de ces résultats que cela s'est fait au prix d'une détérioration très nette de l'imprécision du modèle (critère J_{2Trap}), et de son adéquation aux observations (critère $Distance$). Bien que l'inclusion ne soit toujours que partielle (cf.figure 2.7), tous les autres indicateurs sont à présent moins bons que ceux du modèle trapézoïdal identifié. Ainsi, maximiser l'inclusion en conservant un modèle triangulaire entraîne une dégradation très nette de la qualité de celui-ci, bien supérieure à la dégradation engendrée par l'identification d'un modèle trapézoïdal qui permet lui d'assurer l'inclusion totale.

Ces différentes considérations sur cet exemple simple valident le fait qu'il est préférable de considérer un modèle trapézoïdal si une inclusion importante des observations dans les

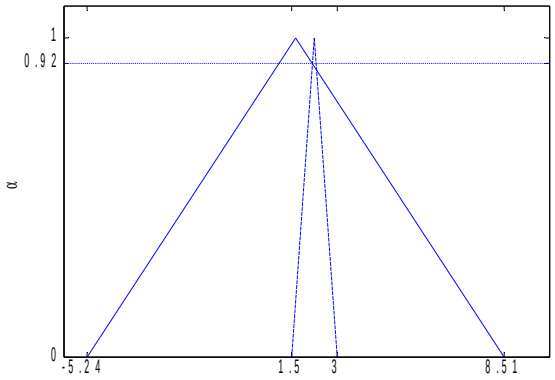


FIG. 2.7: Sorties observée et prédite ($j = 1$) pour le modèle triangulaire identifié à $\alpha = 0.9$

prédictions est recherchée. Par conséquent, dans la suite, nous retiendrons cette solution, et la technique d'identification adaptée, présentée dans la synthèse 3.

2.3 La recherche d'une meilleure représentativité

Le deuxième point important que nous cherchons à traiter concerne le manque de représentativité des modèles linéaires flous de la forme 2.3. En effet, de par la nature floue des paramètres et donc de la sortie, le Radius des intervalles flous considérés doit être positif. Cela se traduit, dans le cas où des paramètres triangulaires sont considérés, par une dépendance de la variation du Radius de la sortie au signe de l'entrée x, selon l'équation 1.64.

Dorénavant, des paramètres sous forme d'intervalles flous trapézoïdaux sont considérés. Ce problème se retrouve donc aux deux α-coupes considérées, l'évolution du Radius du support et du noyau de la sortie étant limitée par le signe de l'entrée x :

$$\frac{dR_{S_{\hat{Y}}}}{dx} = R_{S_{A_1}} \cdot signe(x) \tag{2.32}$$

$$\frac{dR_{K_{\hat{Y}}}}{dx} = R_{K_{A_1}} \cdot signe(x) \tag{2.33}$$

Cette caractéristique est intrinsèque aux modèles linéaires flous, car liée à la manipulation d'intervalles flous, dont les Radius ne peuvent être que positifs. Il est donc nécessaire, pour y remédier, de se focaliser sur le signe du coefficient pondérant le paramètre A_1, c'est-à-dire le signe de la variable d'entrée dans le cas d'un modèle (2.3). Or, celui-ci est déterminé par les observations, l'utilisateur n'a donc pas possibilité d'agir directement dessus.

Afin de s'affranchir de ces restrictions, nous proposons [3] d'ajouter un paramètre de réglage au modèle identifié de façon à pouvoir ajuster le signe du coefficient pondérant le paramètre A_1, quel que soit le signe de la variable d'entrée. Pour ce faire, il est proposé à l'utilisateur de chercher à identifier un modèle de la forme :

$$\hat{Y}(x) = A_0 \oplus A_1 \odot (x - shift) \tag{2.34}$$

le paramètre $shift$ étant un scalaire permettant de modifier à volonté le signe de la pondération de A_1.

Il est important ici de rappeler que le signe de la variable d'entrée originelle x du modèle est conditionné par les observations. Quelques constats peuvent être posés concernant le jeu de données observées (2.1) :

- Il est possible de déterminer sur le jeu fini de M données disponibles les entrées minimale et maximale, c'est-à-dire :

$$x_{min} = \min_{j=1,\dots,M} x_j \tag{2.35}$$

$$x_{max} = \max_{j=1,\dots,M} x_j \tag{2.36}$$

- Il est ainsi possible de fixer le domaine de définition du modèle à identifier comme étant l'intervalle D défini par :

$$D = [x_{min}, x_{max}] \tag{2.37}$$

  Toutes les observations appartiennent donc à cet intervalle D, qui peut donc être vu comme la plage de validité du modèle identifié.

Il est maintenant possible d'étudier la variation du Radius de la sortie d'un modèle linéaire flou trapézoïdal de la forme (2.34), l'entrée x variant sur son domaine de définition D. Le Radius du support de la sortie est donné par :

$$\forall x \in D,\ R_{S_{\hat{Y}}} = R_{S_{A_0}} + R_{S_{A_1}} |x - shift| \tag{2.38}$$

et sa dérivée par rapport à x devient :

$$\frac{dR_{S_{\hat{Y}}}}{dx} = R_{S_{A_1}} \cdot signe(x - shift) \tag{2.39}$$

En agissant sur la valeur du paramètre $shift$, il devient donc possible d'obtenir une variation quelconque du Radius (2.38), et ce, quel que soit le signe de x :

- Si $x - shift \geq 0\ \forall x \in D$, c'est-à-dire que l'on impose $shift \leq x_{min}$, alors le Radius du support de la sortie sera croissant sur le domaine de définition D.
- Si $x - shift \leq 0\ \forall x \in D$, c'est-à-dire que l'on impose $shift \geq x_{max}$, alors le Radius du support de la sortie sera décroissant sur le domaine de définition D.

Il est important de remarquer que le choix du comportement du Radius du support de la sortie du modèle conditionne celui du noyau. En effet, le décalage $shift$ s'applique également à celui-ci, et sachant qu'il est unique, des variations différentes à deux niveaux distincts ne peuvent être obtenues. De plus, dans le cadre de la régression linéaire, le modèle présente une variation unique sur l'ensemble de son domaine de définition.

Un autre point important à souligner est que le choix de la valeur du décalage va également être déterminant sur le positionnement de l'origine du modèle décalé. Il est donc particulièrement intéressant de chercher à positionner cette origine sur une des bornes du domaine de définition du modèle.

Enfin, afin de faciliter la prise en main de modèles linéaires flous introduisant un paramètre $shift$ de décalage des entrées, la valeur de celui-ci doit être facilement interprétable. Ainsi, il semble là encore important de le lier à des informations pré-existantes à l'identification.

Pour toutes ces raisons, un modèle dont le support de la sortie présente un Radius croissant sur son domaine D sera défini avec une valeur de décalage $shift = x_{min}$, tandis qu'un modèle dont le support de la sortie présente un Radius décroissant sur D sera défini avec une valeur de décalage $shift = x_{max}$ (cf. tableau récapitulatif 2.6). Ainsi, seuls deux choix de valeur de décalage sont admissibles, en accord avec le domaine de définition du modèle.

Variation du Radius du support de la sortie	$\nearrow$	$\searrow$
Modèle considéré	$A_0 \oplus A_1 \odot (x - x_{min})$	$A_0 \oplus A_1 \odot (x - x_{max})$

TAB. 2.6: Récapitulatif des modèles définis selon la variation du Radius souhaitée

2.3.1 La méthode d'identification associée

L'introduction du paramètre supplémentaire de décalage $shift$ doit être prise en compte dans la technique d'identification d'un modèle (2.34) sur un jeu de données quelconques (2.1). Deux points sont donc à considérer, le choix de la valeur adéquate du paramètre $shift$ en accord avec les valeurs admissibles du tableau 2.6, et l'identification proprement dite des paramètres A_0 et A_1 du modèle.

Le paramètre $shift$ doit être déterminé de telle sorte que la sortie du modèle puisse représenter au mieux la variation de l'imprécision sur l'ensemble des sorties triangulaires observées. Ainsi, il est nécessaire de considérer ces dernières pour choisir le paramètre $shift$.

Considérons que l'utilisateur a à sa disposition un jeu de M données (2.1), les entrées précises

x_j étant ordonnées par ordre croissant. Ce jeu de données permet de construire de manière immédiate le domaine D du modèle recherché, selon l'équation (2.37). La meilleure tendance de l'imprécision du modèle sera ensuite déterminée en comparant le Radius moyen des sorties correspondant aux entrées minimales, R_{init}, au Radius moyen des sorties associées aux entrées maximales, R_{final}. Ces grandeurs sont calculées sur les k premières et dernières sorties observées, avec $k < M$, de la manière suivante :

$$R_{init} = moyenne(R_{Y_1}, R_{Y_2}, ..., R_{Y_k}) \tag{2.40}$$

$$R_{fin} = moyenne(R_{Y_{M-k+1}}, ..., R_{Y_{M-1}}, R_{Y_M}) \tag{2.41}$$

La comparaison des valeurs R_{init} et R_{final} permet alors de définir la valeur appropriée de décalage $shift$:

- Si $R_{init} > R_{fin}$, alors le Radius des sorties observées est considéré comme globalement décroissant sur D, et le paramètre est fixé comme $shift = x_{max}$
- Si $R_{init} \leq R_{fin}$, alors le Radius des sorties observées est considéré comme globalement croissant sur D, et le paramètre est fixé comme $shift = x_{min}$

Une fois la valeur adéquate du décalage fixée, il est possible d'identifier les paramètres A_0 et A_1 du modèle, celui-ci étant donc à même de représenter toutes les tendances possibles des données observées. L'introduction de ce $shift$ dans la structure du modèle n'a aucun impact sur la formulation du problème d'optimisation. En effet, si l'on considère le changement de variable :

$$w = x - shift \tag{2.42}$$

c'est-à-dire $w_j = x_j - shift, \forall j = 1, ..., M$, pour les entrées observées, on retrouve le problème de la synthèse 3 formulé en terme de l'entrée w.

Synthèse 4 : Afin de permettre la représentation d'une variation quelconque de l'imprécision des données observées, un modèle flou trapézoïdal à entrée décalée, de la forme :

$$\hat{Y}(x) = A_0 \oplus A_1 \odot w \tag{2.43}$$

avec $w = x - shift$, est utilisé.

La première étape de l'identification d'un tel modèle consiste à déterminer la tendance de l'imprécision des données selon les équations (2.40) et (2.41) puis à fixer la valeur de décalage adéquate selon le tableau 2.6.

La seconde étape, relative à l'identification des paramètres trapézoïdaux A_0 et A_1 est similaire à la synthèse 3, une fois le changement de variable $w_j = x_j - shift, \forall j = 1, ..., M$ effectué sur les entrées observées.

2.3.2 Discussion

Identifier un modèle à décalage permet de représenter toutes tendances possibles des observations, aussi bien au niveau de leur Midpoint que de leur Radius. Cela est réalisable au prix de l'introduction d'un nouveau paramètre $shift$ dans la structure du modèle. Cela semble en contradiction avec notre volonté de conserver une structure de modèle régressif conventionnelle. Cependant, le modèle reste linéaire, et ce paramètre supplémentaire, précis, n'impacte que la variable d'entrée du modèle. Il peut d'ailleurs être occulté en utilisant le changement de variable (2.42).

De plus, la phase du choix du paramètre $shift$ le mieux approprié pour le modèle a été délibérément dissociée et positionnée en amont de la technique d'optimisation (première étape de la synthèse 4). Plusieurs raisons à cela peuvent être mises en avant. En effet, il est possible que l'utilisateur connaisse au préalable la tendance de l'imprécision des observations, au travers d'une représentation graphique du jeu de données par exemple, ou bien encore que la nature du système à l'étude impose une variation connue de l'imprécision de sortie. Ainsi, dans ce cas, il ne semble pas utile de complexifier le problème d'optimisation en y intégrant la détermination du décalage.

Il est intéressant aussi de remarquer qu'un choix inapproprié de décalage n'induit qu'une dégradation modérée du modèle identifié. En effet, dans ce cas, la sortie du modèle n'aura pas la variation d'imprécision présente dans les données et le Radius de son support ou de son noyau (cf. équation (2.5)) sera constant par saturation des contraintes assurant l'identification d'intervalles flous à Radius positifs. On retrouve une situation similaire à celle qui nous a conduit à introduire un décalage.

En ce qui concerne l'interprétation de ce type de modèle à décalage (2.34), l'introduction d'un paramètre supplémentaire n'est pas préjudiciable. En effet, la connaissance de la valeur du $shift$, couplée à celle du domaine de définition D du modèle, permet à l'utilisateur d'obtenir de manière immédiate, selon le tableau 2.6, la variation de l'imprécision de la sortie du modèle. Bien entendu, cette analyse doit ensuite être affinée par une étude plus fine des paramètres A_0 et A_1 du modèle.

2.3.3 Exemples illustratifs

L'objectif du premier exemple considéré dans cette section est de déterminer le gain engendré par l'identification de modèles trapézoïdaux à décalage au niveau de la représentativité de la tendance de l'imprécision des observations.

Comme indiqué dans la section 2.3, les modèles à décalage permettent de représenter certaines

variations de l'imprécision inaccessibles pour les modèles conventionnels. Or, dans l'exemple présenté dans le tableau 2.1, et selon le modèle trapézoïdal obtenu dans la section 2.2.3.2, l'imprécision des données est globalement croissante, et ce, pour des entrées positives. Ce cas n'est donc pas problématique. Afin de mettre en lumière les bénéfices des modèles à décalage, il est possible de modifier ces données d'imprécision croissante de façon à ce que les entrées deviennent négatives. Ainsi, en translatant les entrées d'une valeur de -0.9, on obtient le jeu de données du tableau 2.7, considéré par la suite pour l'identification des modèles, avec et sans décalage, sur le domaine $D = [-0.8, -0.1]$.

j	x_j	Y_j
1	-0.8	$(2.25, [1.5, 3])$
2	-0.7	$(2.875, [2, 3.75])$
3	-0.6	$(2.5, [1.5, 3.5])$
4	-0.5	$(4.25, [2.5, 6])$
5	-0.4	$(4.0, [2.5, 5.5])$
6	-0.3	$(5.25, [4, 6.5])$
7	-0.2	$(7.5, [5.5, 9.5])$
8	-0.1	$(8.5, [7, 10])$

TAB. 2.7: Le nouveau jeu de données observées

L'identification du modèle trapézoïdal conventionnel se fait selon la synthèse 3 alors que le modèle à décalage est déterminé selon la synthèse 4.

Sachant que l'imprécision des données est croissante, le décalage est fixé à la borne inférieure de D, c'est-à-dire $shift = -0.8$ et la variable décalée $w = x + 0.8$ est utilisée dans la formulation du problème d'optimisation.

Les modèles flous trapézoïdaux sans et avec décalage obtenus sont présentés dans le tableau 2.8, et une représentation en est proposée sur les figures 2.8 et 2.9.

	Modèle sans décalage	Modèle avec décalage
$shift$	Non	-0.8
A_0	$([7.57, 9.39], [6.07, 11.29])$	$([1, 2.25], [0.5, 3])$
A_1	8.93	$([7.5, 8.93], [5, 10.83])$

TAB. 2.8: Modèles obtenus (représentation dans l'espace des bornes)

Il est évident visuellement que l'utilisation d'un modèle à décalage sur le jeu de données considéré dans cet exemple est très avantageuse. En effet, une représentation identique à celle du

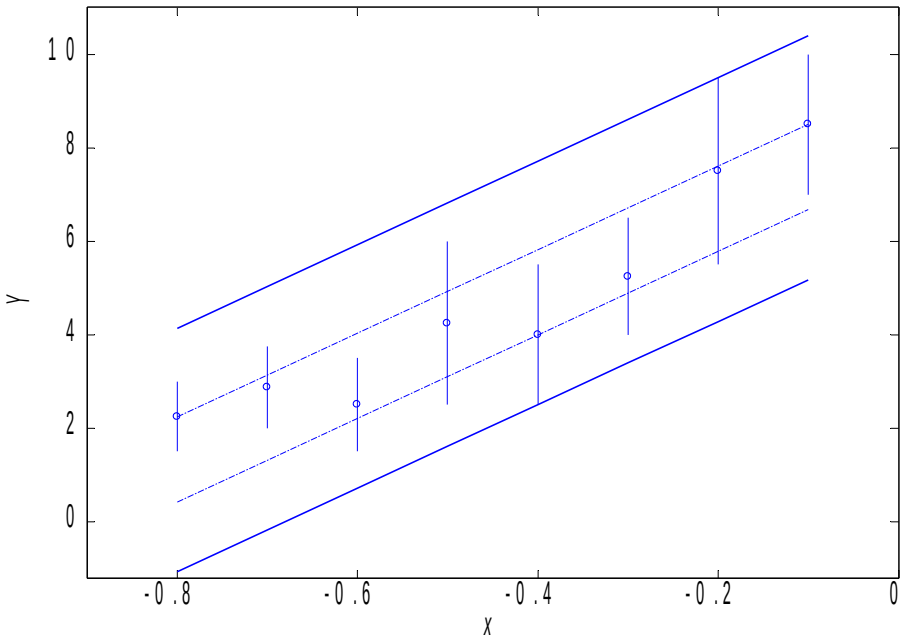


FIG. 2.8: Modèle flou trapézoïdal conventionnel identifié

	Modèle sans décalage	Modèle avec décalage
$shift$	Non	-0.8
$Distance$	58.83	48.08
J_{2Trap}	28.14	25.17

TAB. 2.9: Comparatif des modèles obtenus

cas des entrées positives est obtenue (même paramètre A_1 identifié, cf. tableau 2.2), permettant une représentation optimale des données. Ainsi, la sortie du modèle présente bien une imprécision croissante tant au niveau du support que du noyau. Par contre, le modèle sans décalage a une sortie d'imprécision constante, en cohérence avec le paramètre A_1 précis identifié (cf. tableau 2.8).

Ce gain en représentativité du modèle à décalage se traduit également par une amélioration des deux indicateurs considérés (cf. tableau 2.9). En effet, l'imprécision du modèle (critère J_{2Trap}) est diminuée de 10.5%, tandis que l'erreur quadratique de l'identification décroît de 18.3%. Le modèle à décalage obtenu est donc optimal au sens du critère d'imprécision défini, et présente une meilleure corrélation aux données d'identification.

Le second exemple présenté maintenant a pour objectif d'étudier l'influence de l'introduction

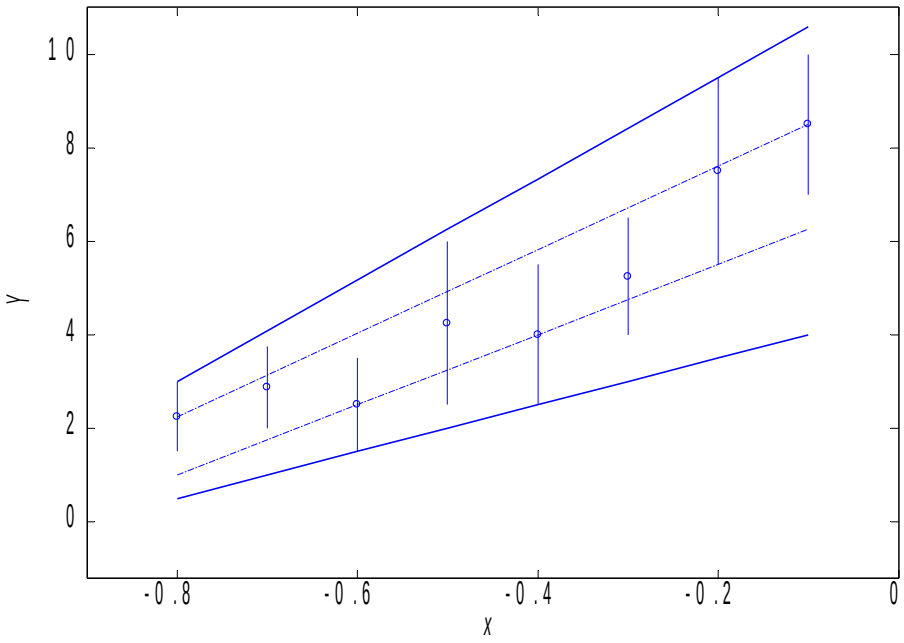


FIG. 2.9: Modèle flou trapézoïdal à décalage identifié

d'un décalage sur l'imprécision du modèle identifié. Pour ce faire, le jeu de données présenté dans le tableau 2.10 est considéré. Il s'agit en fait du jeu de données initial présenté dans le tableau 2.1 mais pour lequel les entrées ont été translatées d'une valeur de 2.

j	x_j	Y_j
1	2.1	$(2.25, [1.5, 3])$
2	2.2	$(2.875, [2, 3.75])$
3	2.3	$(2.5, [1.5, 3.5])$
4	2.4	$(4.25, [2.5, 6])$
5	2.5	$(4.0, [2.5, 5.5])$
6	2.6	$(5.25, [4, 6.5])$
7	2.7	$(7.5, [5.5, 9.5])$
8	2.8	$(8.5, [7, 10])$

TAB. 2.10: Le jeu de données observées

Deux modèles trapézoïdaux sont identifiés, l'un sans décalage, selon la méthodologie de la synthèse 3, et le second avec un décalage, selon la synthèse 4. Dans ce dernier cas, le domaine de définition du modèle est $D = [2.1, 2.8]$. Sachant que l'imprécision des observations est croissante

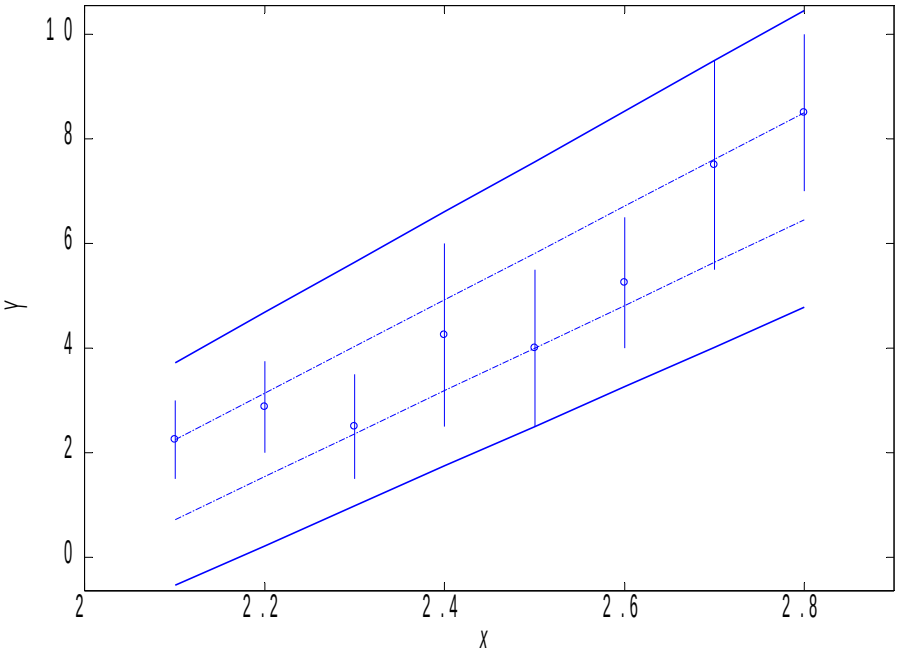


FIG. 2.10: Modèle flou trapézoïdal sans décalage identifié

	Modèle sans décalage	Modèle avec décalage
$shift$	Non	2.1
A_0	-16.5	$([1, 2.25], [0.5, 3])$
A_1	$([8.2, 8.93], [7.6, 9.63])$	$([7.5, 8.93], [5, 10.83])$
$Distance$	52.35	48.08
J_{2Trap}	27.05	25.17

TAB. 2.11: Modèles obtenus (représentation dans l'espace des bornes) et indicateurs associés

sur D, la valeur de décalage est fixée à $shift = 2.1$. Les paramètres et indicateurs associés à chacun de ces modèles sont présentés dans le tableau 2.11, tandis qu'une représentation en est proposée dans les figures 2.10 et 2.11.

Le paramètre A_1 du modèle trapézoïdal à décalage identifié est identique à celui du modèle obtenu sur le jeu de donnée initial (cf. tableau 2.2). Bien que les entrées aient été translatées, augmentant ainsi leur amplitude, l'imprécision des sorties du modèle, tout comme celle des observations, est restée inchangée. Ainsi, retrouver un paramètre A_1 identique est un résultat positif, la représentation des données par le modèle étant donc optimale.

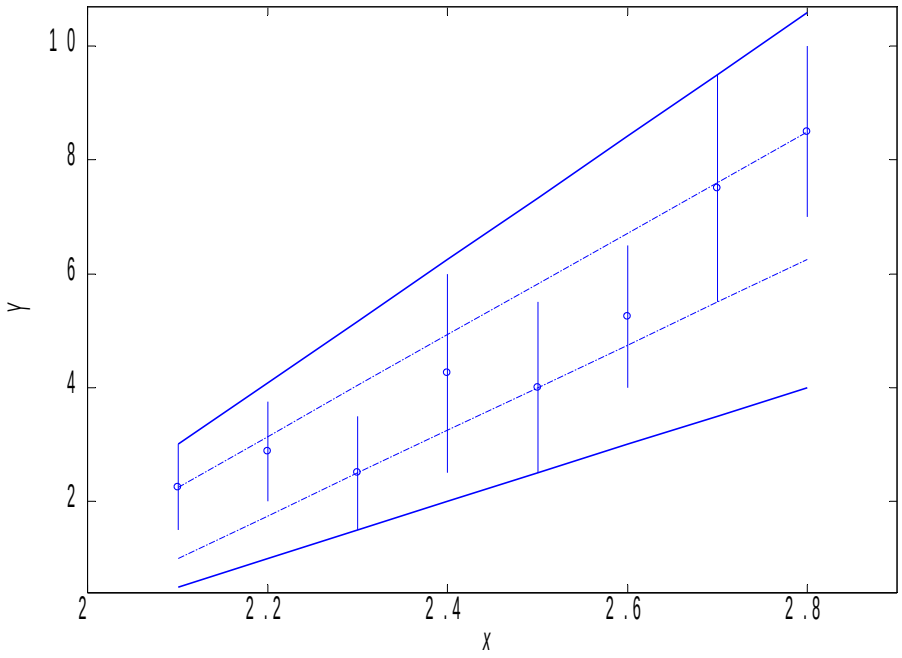


FIG. 2.11: Modèle flou trapézoïdal avec décalage identifié

En ce qui concerne le paramètre A_0, il correspond à la sortie du modèle évaluée au point $w = 0$, soit $x = 2.1$. L'intervalle flou trapézoïdal ainsi défini est le même que celui correspondant à la sortie du modèle identifié sur le jeu de donnée initial (cf. tableau 2.2) évaluée au point $x = 0.1$ (cf. équation (2.30)). Sachant que l'unique différence entre les deux jeux de données est la translation de valeur 2 des entrées, il est clair que les deux modèles sont totalement similaires, et optimaux au sens des indicateurs.

Si on considère les paramètres du modèle sans décalage, on voit que le paramètre A_0 est précis, le paramètre A_1 étant modifié en conséquence afin de respecter les inclusions des observations dans les prédictions. Cela a pour effet de détériorer l'adéquation de la sortie aux données (indicateur *Distance*), et d'augmenter l'imprécision du modèle (indicateur J_{2Trap}). Le modèle optimal n'est donc pas retrouvé. Cela s'explique par le fait que l'origine du modèle non décalé est conservée au point $x = 0$, la sortie du modèle en ce point étant le paramètre A_0 (cf. figure 2.12). Dans ce cas, il est clair visuellement que ce ne sont pas seulement les contraintes d'inclusion des observations dans les prédictions qui sont saturées, mais également la contrainte assurant l'obtention d'un paramètre A_0 à Radius positif. La saturation de cette contrainte (correspondant donc à un paramètre précis de Radius nul) entraine donc une altération de la valeur optimale du critère, et par conséquent, des paramètres optimaux.

Cet exemple illustre donc l'importance du positionnement du zéro sur une des deux bornes

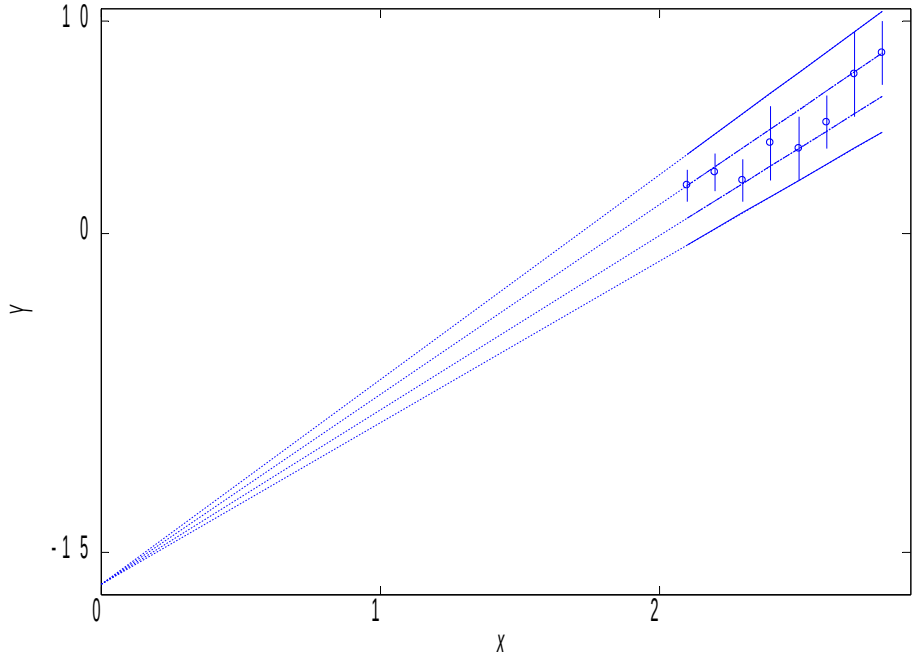


FIG. 2.12: Modèle flou trapézoïdal sans décalage identifié - visualisation du paramètre A_0

du domaine de définition du modèle identifié à l'aide du décalage de l'entrée, permettant ainsi de s'affranchir de l'amplitude des entrées, tout en garantissant l'obtention d'un modèle flou bien défini sur son domaine.

Pour résumer les bénéfices de l'identification de modèles trapézoïdaux à entrée décalée illustrés dans cette section, plusieurs points principaux sont à rappeler. Tout d'abord, en fixant une valeur de décalage $shift$ appropriée, il est possible de représenter toutes tendances possibles des données, aussi bien au niveau des valeurs modales que de l'imprécision, sans perte de linéarité du modèle. Ce paramètre est par ailleurs totalement invisible dans la méthode d'identification, qui reste inchangée par simple changement de variable. Enfin, ce décalage, en fixant le zéro sur des bornes du domaine de définition du modèle, permet de s'affranchir de l'impact négatif de l'amplitude des données sur la valeur optimale du critère.

Par conséquent, dans la suite, seuls des modèles linéaires flous trapézoïdaux à entrée décalée seront considérés.

2.4 Etude du critère

Les techniques d'identification considérées jusqu'à présent, qu'elles concernent des modèles flous trapézoïdaux avec ou sans décalage, ont en point commun d'être structurées de manière identique. Dans tous les cas, il s'agit de minimiser un critère linéaire, représentant l'imprécision du modèle, et de respecter une relation aux données (plus particulièrement l'inclusion des observations dans les prédictions) au travers d'un certain nombre de contraintes.

Quel que soit le type de modèle considéré, les contraintes appropriées mises en place sont linéaires, et expriment, au travers de l'utilisation des α-coupes de niveaux 0 et 1, la relation d'inclusion de manière simple et efficace. Ainsi, ces contraintes font le lien entre les données d'apprentissage, et les prédictions du modèle, permettant à ce dernier d'offrir une représentation de qualité des données.

Or, dans la définition retenue du critère J_{2Trap} (équation (2.12)), il apparait également une dépendance aux données d'apprentissage. En effet, ce critère est défini comme étant la somme des imprécisions des sorties du modèle évaluées aux entrées observées. Deux conséquences découlent de cette dépendance du critère J_{2Trap} aux entrées.

Premièrement, il est intéressant de se pencher sur l'étude de l'expression normalisée du critère, et du mécanisme d'optimisation mis en oeuvre. Selon la synthèse 4, si des modèles trapézoïdaux à décalage sont considérés, ce qui est le cas ici, toutes les entrées sont décalées dans une étape préalable à l'identification proprement dite. Par conséquent, les concepts introduits dans la suite le seront pour la variable d'entrée décalée $w = x - shift$.

Dans le cas d'un modèle de la forme (2.43) à paramètres trapézoïdaux, cette expression est donnée par :

$$J_{2Trap} = R_{S_{A_0}} + R_{K_{A_0}} + (R_{S_{A_1}} + R_{K_{A_1}}) \cdot \frac{\sum_{j=1}^{M} |w_j|}{M} \tag{2.44}$$

Il a été discuté en section 2.2.2 que l'introduction de paramètres trapézoïdaux nécessite la minimisation de l'imprécision à deux niveaux, les principes mis en avant dans la suite concerneront donc la minimisation du Radius du support de la sortie du modèle, les phénomènes observés étant identiques pour le noyau.

La visualisation des imprécisions des supports des paramètres A_0 et A_1 se fait donc en deux points distincts, comme illustré sur la figure 2.13 pour le cas particulier d'entrées décalées positives. Le Radius du support de A_0 se retrouve donc à l'origine, il correspond à l'imprécision de la sortie du modèle pour une entrée nulle. La minimisation du critère (2.44) revient à minimiser la valeur de $R_{S_{A_0}} + R_{S_{A_1}} \cdot |\bar{w}|$, où :

$$|\bar{w}| = \frac{\sum_{j=1}^{M} |w_j|}{M} \tag{2.45}$$

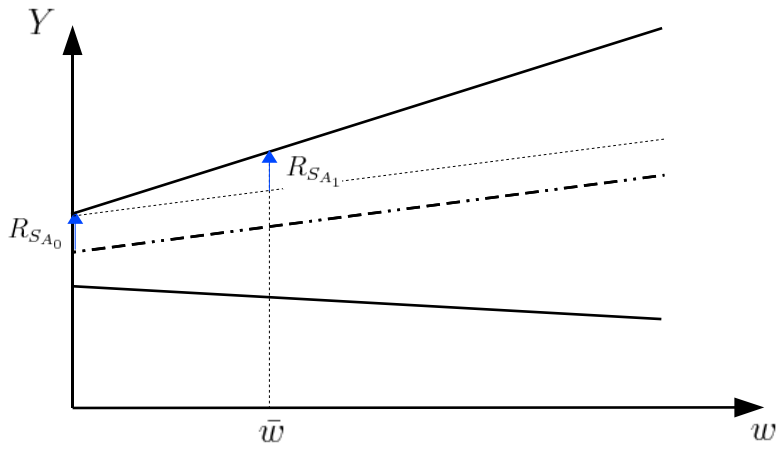


FIG. 2.13: Une vue schématique du principe d'optimisation de J_{2Trap}

Bien évidemment, cette optimisation doit être faite sous contraintes. Celles-ci, lorsqu'elles saturent, permettent donc de limiter la minimisation de l'imprécision. Ces contraintes sont donc présentes uniquement aux points observés.

Le fait que le point considéré corresponde à la moyenne des entrées considérées est sujet à caution. En effet, selon la valeur de l'entrée de chacun des échantillons, ce point est variable. Ainsi, le choix des entrées considérées pour l'identification va influencer non seulement le positionnement des contraintes, de manière tout à fait logique, mais également le point moyen où est réalisée l'optimisation. La valeur optimale du critère J_{2Trap}, qui est recherchée lors de l'identification des paramètres, est donc fortement dépendante des données observées.

La deuxième conséquence de la dépendance du critère J_{2Trap} aux entrées observées est la manque de robustesse que cela induit au niveau des modèles identifiés. En effet, la valeur de la moyenne des entrées présente dans l'expression du critère va changer selon les points considérés dans le jeu de données d'identification, selon leur nombre ou leur éventuelle redondance. Ainsi, un modèle optimal identifié sur un certain jeu de données pourra ne pas être retrouvé à la suite d'une identification sur un ensemble de données dans lequel des données auraient été rajoutées, même si l'influence de celles-ci sur la saturation des contraintes était nulle. Or, là encore, seules les contraintes doivent établir la relation entre les données et la sortie évaluée du modèle, le critère devant donc représenter l'imprécision de ce dernier le plus généralement possible.

2.4.1 Un nouveau critère

L'objectif ici est de déterminer une représentation de l'imprécision de la sortie d'un modèle linéaire flou de telle manière que, utilisée en tant que critère d'optimisation d'un problème

régressif, elle soit indépendante des données d'identification.

Il est important de rappeler plusieurs points dont on ne peut s'affranchir :

- Les modèles retenus sont à paramètres trapézoïdaux, il est donc nécessaire de considérer la dimension verticale dans la définition de l'imprécision de la sortie du modèle évaluée en une entrée précise. Cette imprécision est donc définie comme étant l'aire de l'intervalle flou trapézoïdal de sortie, définie par l'équation (2.9).
- Le jeu fini de données observées permet de définir un intervalle D selon l'équation (2.37), intervalle qui est assimilé au domaine de définition du modèle, c'est-à-dire sa plage de validité dans une utilisation en prédiction.
- Les entrées décalées seront par conséquent toutes incluses dans le domaine décalé $D_w = [D_w^-, D_w^+]$ dont l'une des bornes est le zéro du modèle.

Il est possible, afin de s'affranchir des inconvénients présentés préalablement, de chercher une représentation plus globale de l'imprécision du modèle. Celle-ci ne doit donc plus être évaluée aux entrées des observations, mais définie sur l'ensemble du domaine de définition du modèle, c'est-à-dire sur le domaine D_w.

Ainsi, l'imprécision du modèle n'est plus définie comme la somme des imprécisions de la sortie évaluée en chacune des entrées, c'est-à-dire l'aire de l'intervalle flou trapézoïdal, mais comme l'intégration de cette dernière sur l'ensemble des entrées possibles, c'est-à-dire l'intervalle D_w. Cela revient finalement à considérer comme imprécision globale du modèle le volume défini par l'évolution de la sortie sur la plage de fonctionnement. Ce volume est donc défini par :

$$J_{volume} = \int_{D_w^-}^{D_w^+} aire(\hat{Y}(w))dw \tag{2.46}$$

Soit, selon l'expression (2.10) :

$$J_{volume} = \int_{D_w^-}^{D_w^+} R_{K_{\hat{Y}(w)}} + R_{S_{\hat{Y}(w)}} dw \tag{2.47}$$

c'est-à-dire :

$$J_{volume} = \int_{D_w^-}^{D_w^+} (R_{K_{A_0}} + R_{S_{A_0}}) + (R_{K_{A_1}} + R_{S_{A_1}}) \cdot w \cdot signe(w)dw \tag{2.48}$$

Or, le signe de la variable w est constant sur le domaine décalé D_w, car les deux valeurs de décalage admissibles imposent le zéro sur une des bornes du domaine de définition initial. Par conséquent, il est possible d'écrire :

$$J_{volume} = (R_{K_{A_0}} + R_{S_{A_0}}) \cdot (D_w^+ - D_w^-) + \Delta_w \cdot (R_{K_{A_1}} + R_{S_{A_1}}) \int_{D_w^-}^{D_w^+} w \cdot dw \tag{2.49}$$

où :

$$\Delta_w = signe(w) \tag{2.50}$$

On obtient alors :

$$J_{volume} = 2 \cdot R_{D_w} \cdot (R_{K_{A_0}} + R_{S_{A_0}}) + 1/2 \cdot \Delta_w \cdot (D_w^{+2} - D_w^{-2}) \cdot (R_{K_{A_1}} + R_{S_{A_1}}) \tag{2.51}$$

Soit :

$$J_{volume} = 2 \cdot R_{D_w} \cdot (R_{K_{A_0}} + R_{S_{A_0}}) + 2 \cdot R_{D_w} \Delta_w \cdot M_{D_w} \cdot (R_{K_{A_1}} + R_{S_{A_1}}) \tag{2.52}$$

En simplifiant cette expression, le critère considéré est donc :

$$J_{volume} = R_{S_{A_0}} + R_{K_{A_0}} + (R_{S_{A_1}} + R_{K_{A_1}}) \cdot M_{D_w} \cdot \Delta_w \tag{2.53}$$

L'expression de J_{volume}, normalisée dans l'équation (2.53), est semblable à celle de J_{2Trap} pour plusieurs raisons. La grandeur M_{D_w}, c'est-à-dire le Midpoint de l'intervalle D_w correspondant au domaine de définition décalé, est une grandeur numérique connue lors de la phase d'identification. Ainsi, le critère reste linéaire en les paramètres d'optimisation, que sont les Radius et les Midpoints des paramètres du modèle. On remarquera que là encore, les Midpoints n'apparaissent pas dans l'expression du critère, mais la minimisation de celui-ci devant être réalisée sous l'ensemble de contraintes appropriées, ils sont tout de même présents dans le problème d'optimisation.

Tout comme le critère J_{2Trap}, le critère J_{volume} défini pour un modèle trapézoïdal fait intervenir la dimension verticale dans le problème d'optimisation, puisqu'il est défini comme une combinaison linéaire des Radius des paramètres aussi bien au niveau de leur support que de leur noyau. On remarquera ici que l'obtention du critère J_{volume} pour des modèles triangulaires est immédiate, puisque dans ce cas le Radius des noyaux des paramètres n'est plus introduit dans son expression, étant considéré comme nul afin d'obtenir des valeurs modales ponctuelles. Dans ce cas particulier [7], le critère est donné par :

$$J_{volume\_triangulaire} = R_{S_{A_0}} + R_{S_{A_1}} \cdot M_{D_w} \cdot \Delta_w \tag{2.54}$$

La principale différence entre les critères J_{2Trap} et J_{volume} se situe au niveau de la dépendance aux données observées [9]. En effet, les entrées n'apparaissent plus explicitement au travers de leur moyenne dans J_{volume}, c'est-à-dire que l'imprécision est définie indépendamment de celles-ci. Seul intervient le domaine de définition du modèle, quelles que soient les données et leur répartition au sein de celui-ci. Dans l'expression normalisée (2.53), et comme illustré sur la figure 2.14, il apparait que la minimisation de l'imprécision de la sortie du modèle se fait au point milieu de l'intervalle D_w, c'est-à-dire M_{D_w}.

Ainsi, à plage de fonctionnement déterminée identique, il n'y aura pas d'influence de la répartition des données observées sur celle-ci. En d'autres termes, si les données correspondant aux entrées extrêmes D_w^- et D_w^+ restent inchangées, l'intervalle D_w est défini de manière identique, et par conséquent le point de minimisation M_{D_w} est le même. Ainsi, quelle que soit la

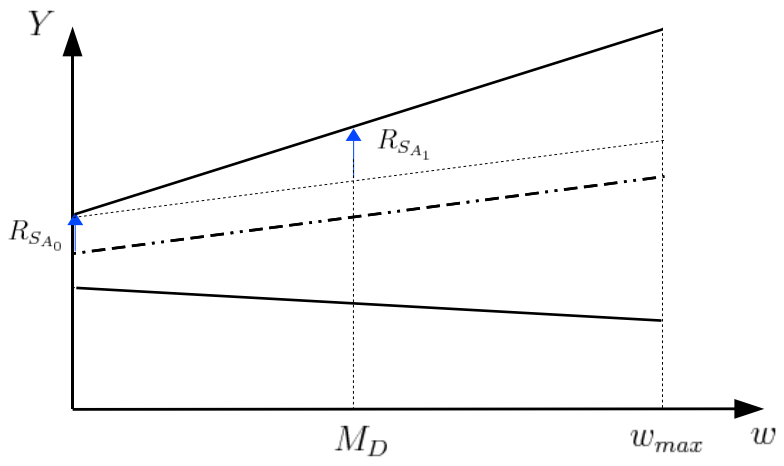


FIG. 2.14: Une vue schématique du principe d'optimisation de J_{volume}

répartition des données sur la plage d'étude, quel que soit leur nombre, il n'y a pas d'influence sur la structure du critère linéaire, et par conséquent sur son optimum correspondant aux meilleurs paramètres du modèle recherché.

Il en va de même si de nouvelles données ne saturant pas les contraintes sont insérées dans le jeu des échantillons observés, puisque si les bornes de l'intervalle D_w restent identiques, le critère et donc sa valeur optimale ne seront pas influencés. Ainsi, le modèle obtenu, considéré comme le meilleur au sens de son imprécision globale sur son domaine de définition sera retrouvé. L'utilisation du critère J_{volume} garantit donc une certaine "robustesse" à la technique d'identification linéaire considérée.

On remarquera ici que si les observations sont équi-réparties sur la plage de fonctionnement, c'est-à-dire sur le domaine D_w, alors dans ce cas, $M_{D_w} = \bar{w}$. Ainsi, dans les cas les plus favorables (mais aussi les plus souvent rencontrés dans la littérature), les deux critères sont rigoureusement identiques. L'amélioration apportée par l'optimisation du critère J_{volume} n'est donc effective que sur des cas particuliers de jeux de données, dans lesquels les mesures effectuées n'ont pas été homogènes.

Synthèse 5 : Afin d'identifier un modèle linéaire flou trapézoïdal à entrée décalée, une fois le changement de variable effectué selon la synthèse 4, il est proposé d'optimiser le critère défini par :

$$\min_{\mathbf{R_{K_A}},\mathbf{R_{S_A}},\mathbf{M_{K_A}},\mathbf{M_{S_A}}} J_{volume}(\mathbf{R_{K_A}},\mathbf{R_{S_A}}) \tag{2.55}$$

avec :

$$J_{volume} = R_{S_{A_0}} + R_{K_{A_0}} + (R_{S_{A_1}} + R_{K_{A_1}}) \cdot M_{D_w} \cdot \Delta_w \tag{2.56}$$

sous les contraintes d'inclusion inchangées de la synthèse 3.

2.4.2 Exemples illustratifs

Dans cette section, l'impact de l'utilisation du critère J_{volume} en lieu et place de J_{2Trap} est illustré sur deux exemples simples, l'un concernant la répartition des données sur le domaine d'identification, et l'autre concernant l'ajout de données dans cet ensemble d'échantillons.

Pour mener à bien cette étude, un jeu synthétique de données est construit à partir du modèle triangulaire défini par les paramètres :

$$\begin{cases} A_0 = (3, [1, 5]) \\ A_1 = (2, [1, 3]) \end{cases} \tag{2.57}$$

Les données générées, présentées dans le tableau 2.12, ne sont pas équi-réparties sur la plage de fonctionnement du modèle définie par l'intervalle $D = [0, 4]$, mais une majorité de mesures est concentrée aux alentours de la borne minimale de cet intervalle. Par ailleurs, les imprécisions étant croissantes, le décalage peut être ignoré puisque $shift = 0$. La répartition des entrées conduit à $\bar{x} = 1$ alors que $M_D = 2$. Ainsi, dans le critère J_{volume} l'importance de l'imprécision du paramètre A_1 relativement à celle de A_0 est doublée par rapport au critère J_{2Trap}.

Sur ce jeu de données idéal, où toutes les contraintes peuvent être saturées simultanément, le modèle initial est retrouvé, quel que soit le critère considéré dans la technique d'identification (tableau 2.13 et figure 2.15). En effet, le modèle de génération des données initial est bien évidemment celui d'imprécision minimale, quelle que soit la représentation choisie de cette dernière. Les valeurs des critères J_{2Trap} et J_{volume} du tableau 2.13 ne sont cependant pas comparables dans la mesure où pour J_{2Trap} la définition non normalisée utilisée jusqu'à présent est conservée (le critère normalisé est donné entre parenthèses).

j	x_j	Y_j
1	0	$(3, [1, 5])$
2	0.1	$(3.2, [1.1, 5.3])$
3	0.2	$(3.4, [1.2, 5.6])$
4	0.3	$(3.6, [1.3, 5.9])$
5	0.4	$(3.8, [1.4, 6.2])$
6	0.5	$(4, [1.5, 6.5])$
7	1	$(5, [2, 8])$
8	2.5	$(8, [3.5, 12.5])$
9	4	$(11, [5, 17])$

TAB. 2.12: Le jeu de données observées

	J_{2Trap}	J_{volume}
A_0	$(3, [1, 5])$	$(3, [1, 5])$
A_1	$(2, [1, 3])$	$(2, [1, 3])$
J_{2Trap}	27 (3)	27 (3)
J_{volume}	4	4

TAB. 2.13: Modèles obtenus selon les deux critères étudiés - données de base

Pour éliminer l'effet masquant des contraintes sur la différence de critères, un jeu de données modifiées est construit (tableau 2.14) en diminuant l'imprécision des sorties correspondant aux entrées minimales (échantillons $j = \{1, 2, 3, 4, 5, 6\}$).

j	x_j	Y_j
1	0	$(3, [2, 4])$
2	0.1	$(3.2, [2.1, 4.3])$
3	0.2	$(3.4, [2.2, 4.6])$
4	0.3	$(3.6, [2.3, 4.9])$
5	0.4	$(3.8, [2.4, 5.2])$
6	0.5	$(4, [2.5, 5.5])$
7	1	$(5, [2, 8])$
8	2.5	$(8, [3.5, 12.5])$
9	4	$(11, [5, 17])$

TAB. 2.14: Le jeu modifié de données observées

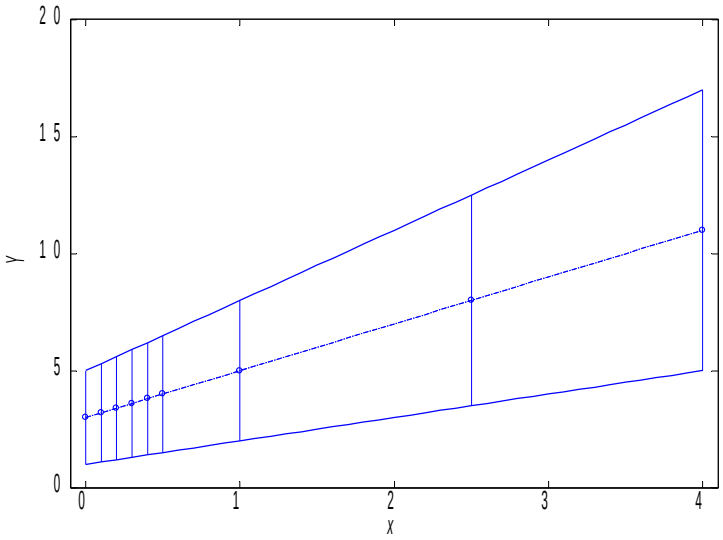


FIG. 2.15: Modèle obtenu selon les deux critères optimisés

Les modèles identifiés à partir de ce nouveau jeu de données sont présentés dans le tableau 2.15 et sur les figures 2.16 et 2.17. Si le modèle initial de paramètres (2.57) est retrouvé avec le critère J_{volume}, il n'en est pas de même avec le critère J_{2Trap}.

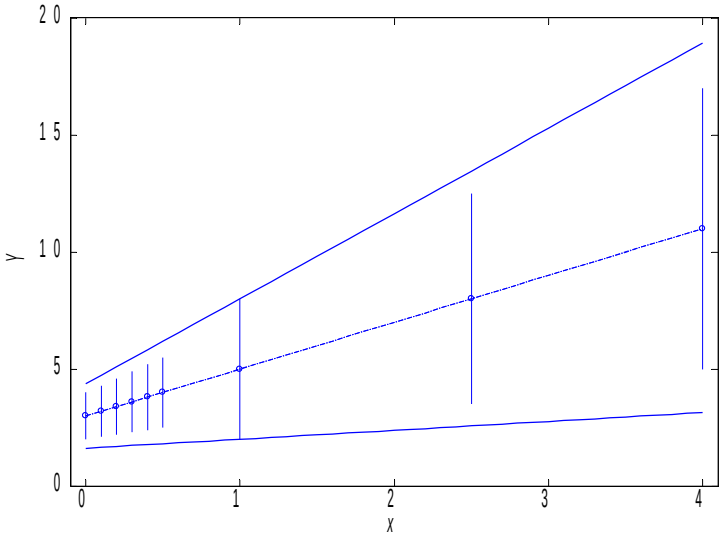


FIG. 2.16: Modèle obtenu par optimisation de J_{2Trap}

Avec le critère J_{2Trap}, la minimisation de l'imprécision de sortie est réalisée au point $\bar{x} = 1$.

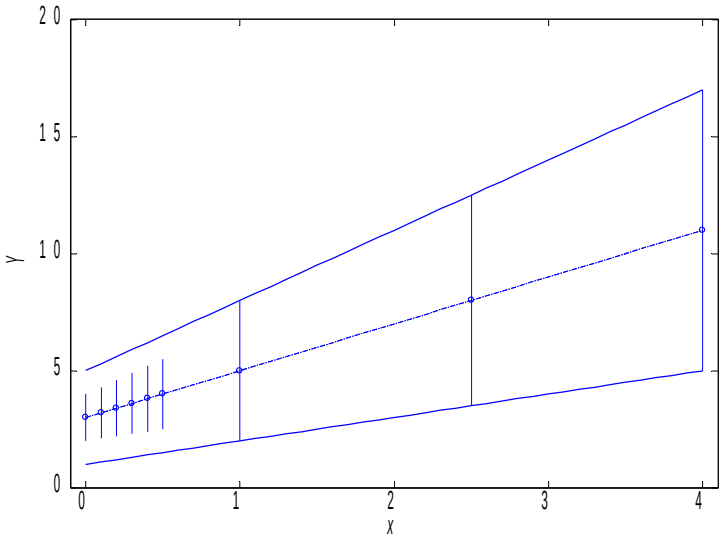


FIG. 2.17: Modèle obtenu par optimisation de J_{volume}

	J_{2Trap}	J_{volume}
A_0	$(3, [1.62, 4.36])$	$(3, [1, 5])$
A_1	$(2, [0.38, 3.641])$	$(2, [1, 3])$
J_{2Trap}	27 (3)	27 (3)
J_{volume}	4.62	4

TAB. 2.15: Modèles obtenus selon les deux critères étudiés - données modifiées

On vérifie sur la figure 2.16 qu'en ce point, l'imprécision minimale autorisée a bien été atteinte puisque la contrainte d'inclusion correspondante est saturée. Pour atteindre cette solution, le critère J_{2Trap} donnant un poids similaire à l'imprécision des paramètres A_0 et A_1, l'imprécision sur A_0 est diminuée en accord avec le relâchement des contraintes aux points 1 à 6, et l'augmentation sur A_1 augmentée en compensation.

A contrario, si c'est le critère J_{volume} qui est optimisé, la minimisation est faite au point $M_D = 2$, ce qui se traduit par le fait que l'imprécision de A_1 a deux fois plus d'importance que celle de A_0 dans la valeur de critère. En conséquence, la réduction de l'imprécision du paramètre A_0 est moins significative qu'avec le critère J_{2Trap}.

Ce premier exemple a permis d'illustrer la différence de comportement des critères J_{2Trap} et J_{volume}. Le deuxième a pour but d'étudier leur impact sur la qualité des modèles identifiés lors de l'ajout de données d'identification.

Pour ce faire, le jeu de données simple présenté dans le tableau 2.1 est de nouveau considéré. Afin de faciliter la lecture, ces données sont ici rappelées dans le tableau 2.16.

j	x_j	Y_j
1	0.1	$(2.25, [1.5, 3])$
2	0.2	$(2.875, [2, 3.75])$
3	0.3	$(2.5, [1.5, 3.5])$
4	0.4	$(4.25, [2.5, 6])$
5	0.5	$(4.0, [2.5, 5.5])$
6	0.6	$(5.25, [4, 6.5])$
7	0.7	$(7.5, [5.5, 9.5])$
8	0.8	$(8.5, [7, 10])$

TAB. 2.16: Le jeu de données observées

Un modèle linéaire flou trapézoïdal à décalage est identifié, en utilisant chacun des deux critères J_{2Trap} et J_{volume} dans le problème d'optimisation. Le domaine de définition considéré est de manière immédiate $D = [0.1, 0.8]$, et les imprécisions étant croissantes sur celui-ci, la valeur de décalage est fixée à $shift = 0.1$. Le modèle recherché est donc de la forme :

$$\hat{Y}(x) = A_0 \oplus A_1 \odot (x - 0.1) \tag{2.58}$$

Les données étant équi-réparties, les deux critères ont donc un fonctionnement similaire, et les

	J_{2Trap}	J_{volume}
$shift$	0.1	0.1
A_0	$([1, 2.25], [0.5, 3])$	$([1, 2.25], [0.5, 3])$
A_1	$([7.5, 8.93], [5, 10.83])$	$([7.5, 8.93], [5, 10.83])$
J_{2Trap}	25.17 (2.8)	25.17 (2.8)
J_{volume}	3.15	3.15
$Distance$	48.08	48.08

TAB. 2.17: Modèles obtenus selon les deux critères

modèles identifiés sont identiques (cf. tableau 2.17). Une représentation en est fournie sur la figure 2.18.

La deuxième phase de cette étude concerne l'ajout de données d'identification. Ces nouvelles données sont choisies de telle manière qu'elles soient incluses dans la sortie du modèle identifié initialement. Ainsi, leur influence au niveau des contraintes saturées est nulle et elles ne devraient

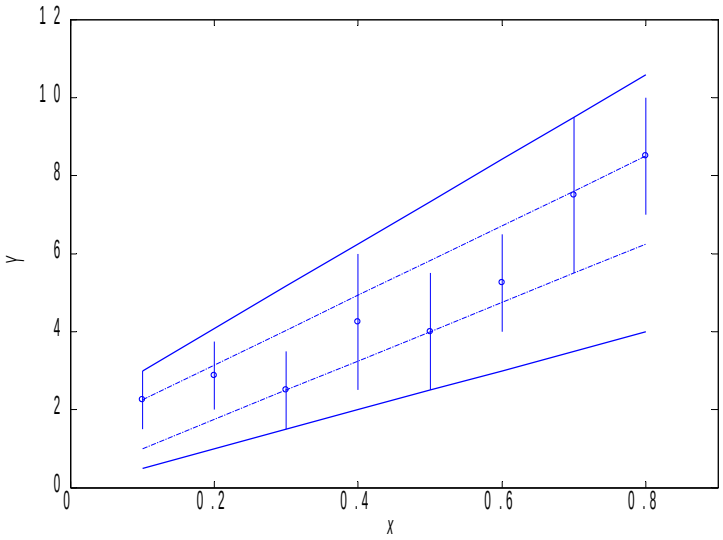


FIG. 2.18: Modèle obtenu par optimisation de J_{volume} ou J_{2Trap}

pas avoir d'influence sur la sortie du modèle. Ces nouveaux échantillons sont présentés dans le tableau 2.18.

j	x_j	Y_j
9	0.64	$(5.5, [4.9, 6.1])$
10	0.65	$(5.5, [4.9, 6.1])$
11	0.66	$(5.5, [4.9, 6.1])$

TAB. 2.18: Les échantillons ajoutés

Sur ce jeu de données étendu (tableaux 2.16 et 2.18), deux modèles sont là encore identifiés, un pour chacun des deux critères à l'étude. On remarquera ici que l'ajout de ces données ne modifie pas le domaine de définition du modèle, qui reste $D = [0.1, 0.8]$. Les paramètres des modèles obtenus, ainsi que leurs indicateurs de performance sont présentés dans le tableau 2.19. Une représentation en est également fournie sur les figures 2.19 et 2.20.

Les modèles identifiés sont donc différents selon le critère choisi. L'optimisation de J_{volume} conduit à l'obtention du modèle initial, donc celui d'imprécision globale minimale. Par contre, l'optimisation de J_{2Trap} conduit à un modèle de paramètres différents. Cela s'explique par le fait que l'ajout de ces données au sein du jeu d'échantillons a modifié la moyenne des entrées et par conséquent la forme du critère J_{2Trap}.

	J_{2Trap}	J_{volume}
$shift$	0.1	0.1
A_0	$([0.43, 2.25], [-1.071, 3])$	$([1, 2.25], [0.5, 3])$
A_1	$(8.93, [8.93, 10.83])$	$([7.5, 8.93], [5, 10.83])$
J_{2Trap}	36.65 (3.05)	36.78 (3.065)
J_{volume}	3.28	3.15
$Distance$	86.44	89.51

TAB. 2.19: Modèles obtenus selon les deux critères

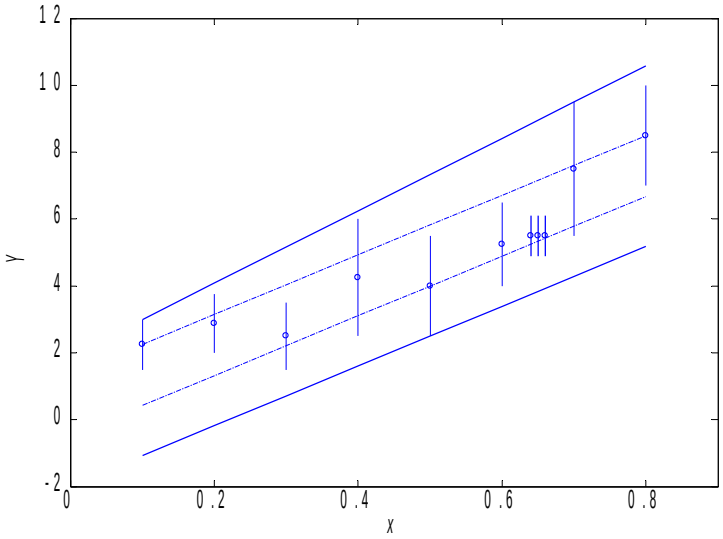


FIG. 2.19: Modèle obtenu par optimisation de J_{2Trap}

Sa valeur optimale en est changée, et les paramètres du modèle correspondant également. Ainsi, bien que les données aient été choisies de telle manière que leur influence au niveau du modèle soit nulle (inclusion parfaite dans le modèle initial), elles modifient l'expression du critère J_{2Trap} et impactent négativement la robustesse de l'identification vis-à-vis de l'ajout de données non conflictuelles.

Ainsi, il semble plus judicieux de privilégier le critère J_{volume}, c'est-à-dire la recherche du modèle le moins imprécis sur l'ensemble de son domaine de définition D.

La dernière phase de cette étude concerne l'impact de données redondantes, et non plus simplement non conflictuelles, sur les modèles identifiés avec les deux critères. Pour cela, l'ensemble des observations initiales (tableau 2.16) est à nouveau considéré mais cette fois la dernière données, c'est-à-dire $(8.5, [7, 10])$, est dupliquée trois fois. On remarquera que là encore, le do-

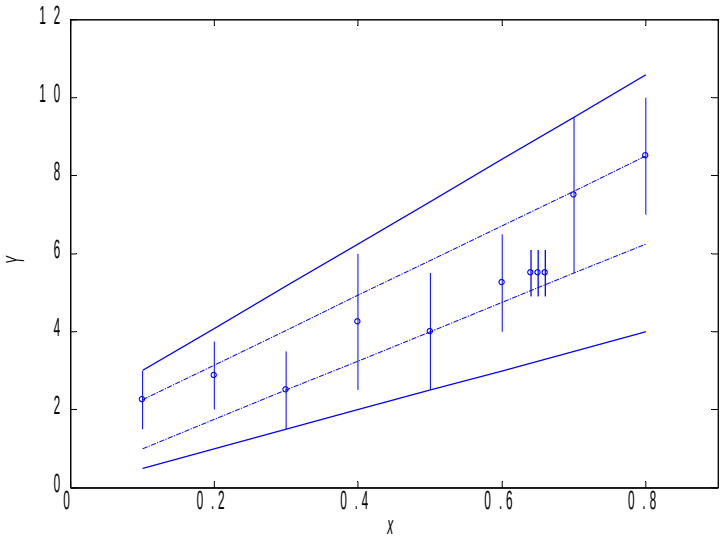


FIG. 2.20: Modèle obtenu par optimisation de J_{volume}

maine de définition du modèle reste inchangé, soit $D = [0.1, 0.8]$.

Les modèles identifiés sur le jeu de données incluant les données redondantes en optimisant chacun des deux critères sont présentés dans le tableau 2.20.

	J_{2Trap}	J_{volume}
A_0	$([0.43, 2.25], [-1.071, 3])$	$([1, 2.25], [0.5, 3])$
A_1	$(8.93, [8.93, 10.83])$	$([7.5, 8.93], [5, 10.83])$
J_{2Trap}	37.08 (3.09)	38.42 (3.2)
J_{volume}	3.28	3.15
$Distance$	71.05	91.29

TAB. 2.20: Modèles obtenus selon les deux critères

A nouveau, les modèles identifiés sont différents selon le critère mis en oeuvre dans la technique d'optimisation. La minimisation du critère J_{volume} conduit de nouveau à l'obtention du modèle initialement identifié sur le jeu de données de base, ce qui n'est pas le cas pour le critère J_{2Trap}. En effet, comme pour l'ajout de données non conflictuelles, la redondance de la dernière donnée a modifié la moyenne des entrées et ainsi la forme du critère J_{2Trap}, en attribuant un poids relatif plus élevé au paramètre A_1. Ainsi, comme sur l'exemple précédent, la valeur optimale et les paramètres du modèle correspondant sont modifiés.

Dans le cas présent, les trois données ajoutées dans les échantillons sont issues de la re-

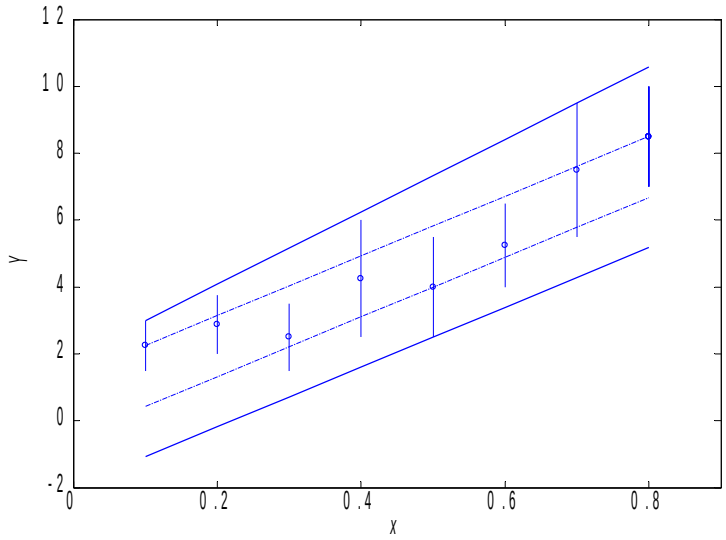


FIG. 2.21: Modèle obtenu par optimisation de J_{2Trap}

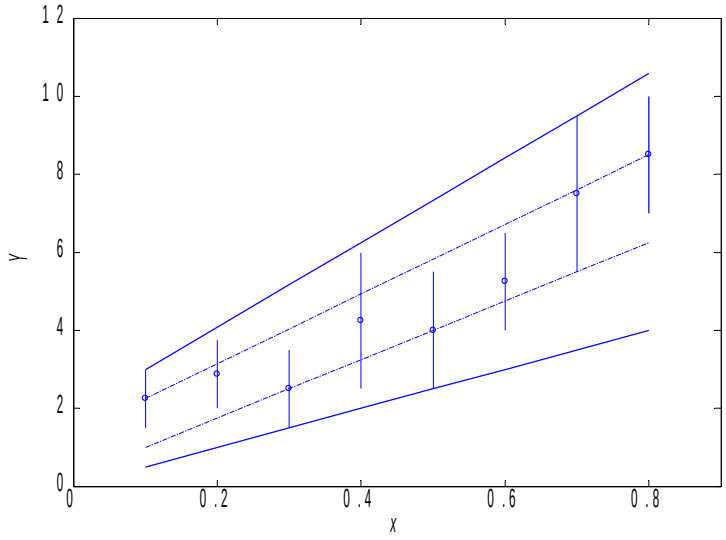


FIG. 2.22: Modèle obtenu par optimisation de J_{volume}

dondance d'une donnée déjà présente dans le jeu initial. Ainsi, cela peut être assimilé à une mesure répétée, qui n'apporte aucun supplément d'information, notamment en ce qui concerne l'imprécision du modèle. Par conséquent, cette redondance ne devrait pas modifier le résultat de l'identification. Ainsi, le fait de retrouver le modèle initial valide l'utilisation du critère J_{volume} bien que les indicateurs *Distance* et J_{2Trap} semblent considérablement dégradés.

Cependant, si l'on élimine les données redondantes pour l'évaluation des indicateurs de performance, tout en les conservant pour l'identification, les résultats du tableau 2.21 sont obtenus. Ils mettent en évidence que le modèle obtenu en optimisant J_{volume} est plus performant, que ce soit en terme d'imprécision (globale ou aux points observés) qu'en terme d'adéquation.

	J_{2Trap}	J_{volume}
J_{2Trap}	26.24	25.17
J_{volume}	3.28	3.15
$Distance$	50.13	48.08

TAB. 2.21: Indicateurs obtenus sans la redondance

Pour résumer, ces exemples mettent en évidence que le critère linéaire représentant l'imprécision de la sortie du modèle sur l'intégralité de son domaine de définition, J_{volume}, a pour propriété d'être indépendant des données observées. Cela permet de s'affranchir de la sensibilité aux données redondantes potentiellement présentes dans le jeu d'échantillons, mais également d'être moins sensible à la répartition des observations sur la plage d'étude. Par conséquent, l'introduction de ce critère dans une technique régressive linéaire floue permet d'améliorer la "robustesse" de cette dernière.

2.5 Les extensions possibles

Les différents principes retenus, résumés par la synthèse 5, ont été introduits pour des modèles linéaires à une seule entrée. Il est bien évidemment possible de les étendre à des modèles plus complexes pour répondre à des problèmes régressifs plus évolués. Ainsi, dans un premier temps, l'identification de modèles régressifs linéaires par morceaux sera introduite, afin de permettre la représentation de jeux de données où différentes tendances de variations sont présentes. Dans un second temps, une généralisation à des modèles multi-entrées sera proposée.

2.5.1 La régression par morceaux

2.5.1.1 Présentation

Les modèles linéaires introduits précédemment permettent de représenter des variations quelconques de tendances des données observées, aussi bien pour les valeurs modales que pour les imprécisions. Cependant, du fait de la structure même des modèles, dont la forme mathématique est une fonction affine, la variation représentée ne peut être qu'unique sur l'ensemble des données.

Or, dans une application réelle, il est rare que les données présentent une évolution rigoureusement linéaire, et identique sur l'ensemble de la plage d'étude.

Afin de remédier à cela, il est possible de chercher à identifier non plus un unique modèle sur les données, mais une collection de sous-modèles, toujours linéaires, chacun d'entre eux permettant de représenter une variation différente des autres [8].

Le modèle global est recherché sous la forme :

$$\hat{Y}(x) = \sum_{k=1}^{S}{}^{\oplus} \left[A_{k0} \oplus A_{k1} \odot (x - shift_k)\right] \odot 1_{[x^k_{min}, x^k_{max}]} \tag{2.59}$$

où :

- $\sum^{\oplus}$ représente la somme de plusieurs intervalles flous ;
- A_{k0} et A_{k1}, $k = 1, ..., S$, sont des intervalles flous trapézoïdaux ;
- $shift_k \in \{x^k_{min}, x^k_{max}\}$ est une valeur réelle de décalage appliqué sur l'entrée x ;
- $1_{[x^k_{min}, x^k_{max}]}$ représente la fonction égale à 1 sur l'intervalle $[x^k_{min}, x^k_{max}]$, et nulle partout ailleurs.

Le modèle global est donc composé de S sous-modèles linéaires. Afin que chacun d'entre eux bénéficie des avantages présentés dans les sections précédentes, ces sous-modèles sont bien entendu à paramètres flous trapézoïdaux et à entrée décalée, permettant donc d'obtenir l'inclusion totale recherchée et de représenter tout type de variation de l'imprécision.

Les sous-modèles sont ici indépendants les uns des autres. Ainsi, chacun d'entre eux a son propre domaine de définition, sa propre valeur de décalage, et ses propres paramètres, totalement indépendants de ceux des autres sous-modèles. Par conséquent, ils peuvent être identifiés puis utilisés en prédiction indépendamment les uns des autres.

La structure générale du modèle étant maintenant fixée, il reste à définir une stratégie d'identification. L'approche proposée se décompose en deux phases, la première dédiée à la segmentation des données d'identification, la seconde à l'identification des paramètres des sous-modèles.

L'objectif de la première phase de segmentation des données est d'obtenir les jeux d'échantillons pour chacun des sous-modèles. On suppose ici que le nombre S de segments est déterminé a priori. Chacun des sous-modèles pouvant représenter une variation quelconque des valeurs modales, ainsi qu'une variation quelconque des imprécisions, il est nécessaire ici de segmenter les données selon les changements de tendance de ces deux grandeurs. Le principe est illustré sur la figure 2.23.

La première étape de la segmentation va donc concerner les valeurs modales des observations, permettant ainsi de dégager les principales tendances des données. Sur chacun des domaines ainsi obtenus, une deuxième segmentation est réalisée, cette fois sur les variations des imprécisions.

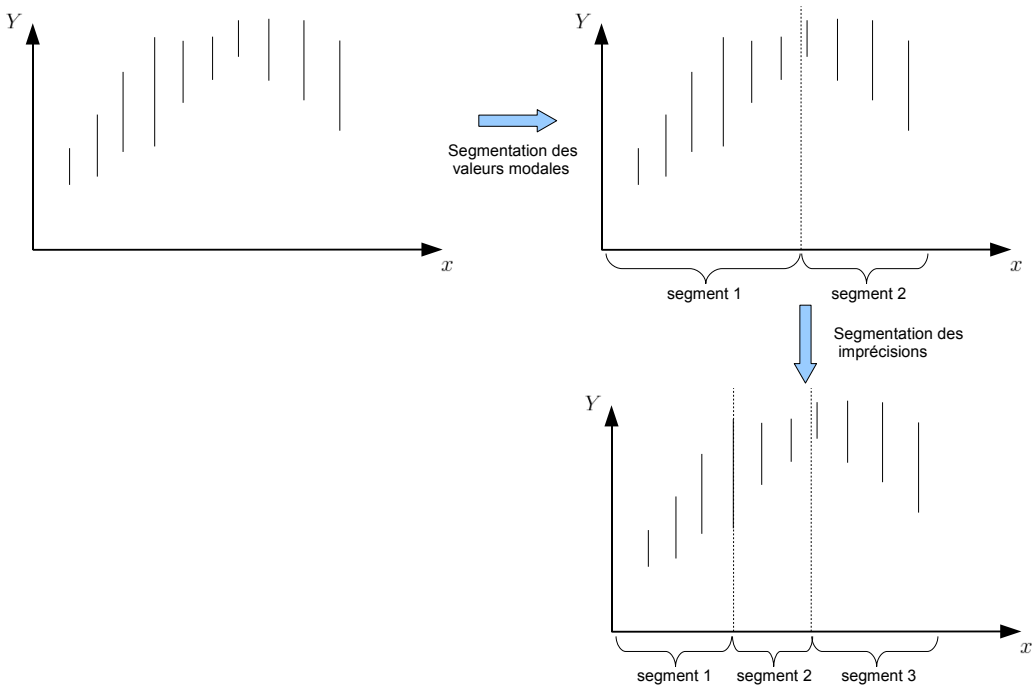


FIG. 2.23: Une vue schématique du principe de segmentation

Il est important de remarquer ici que, si l'on cherche à segmenter un ensemble d'observations formalisées par des intervalles flous, les deux segmentations successives sont réalisées sur des valeurs réelles précises. En effet, aussi bien les Noyaux k_{Y_j} que les Radius $R_{S_{Y_j}}$ sont des nombres précis. Ainsi, l'algorithme utilisé dans cette phase n'a pas à être développé dans un environnement imprécis, une approche classique, comme celle proposée par Keogh et al. [38] est tout à fait appropriée et utilisée dans ces travaux.

Suite à cette phase de segmentation, l'utilisateur dispose de S jeux d'observations, chacun d'entre eux permettant d'identifier un sous-modèle linéaire flou. La phase d'identification est alors réalisée en utilisant la technique régressive présentée auparavant (cf. synthèse 5).

Au final, l'utilisateur dispose d'une collection de sous-modèles linéaires flous, chacun étant d'imprécision minimale sur son domaine, respectant l'inclusion des observations dans les prédictions, et représentant des variations quelconques des valeurs modales et des imprécisions. Ces sous-modèles optimaux sont regroupés dans le modèle global (2.59).

Il est important ici de souligner que la stratégie proposée de décomposition d'un modèle global en sous-modèles locaux n'assure en rien l'optimalité du modèle global. Elle permet néanmoins d'aborder simplement le cas où un unique modèle linéaire ne permet pas une représentation utile de l'information. Il est clair que la continuité entre sous-modèles n'est généralement pas assurée. D'une part, leurs domaines de définition respectifs sont potentiellement disjoints, voire

éloignés, d'autre part il est probable que les sorties estimées aux points de jonction entre deux sous-modèles soient très différentes dans la mesure où c'est justement une différence sur les observations correspondantes qui a dirigé la segmentation.

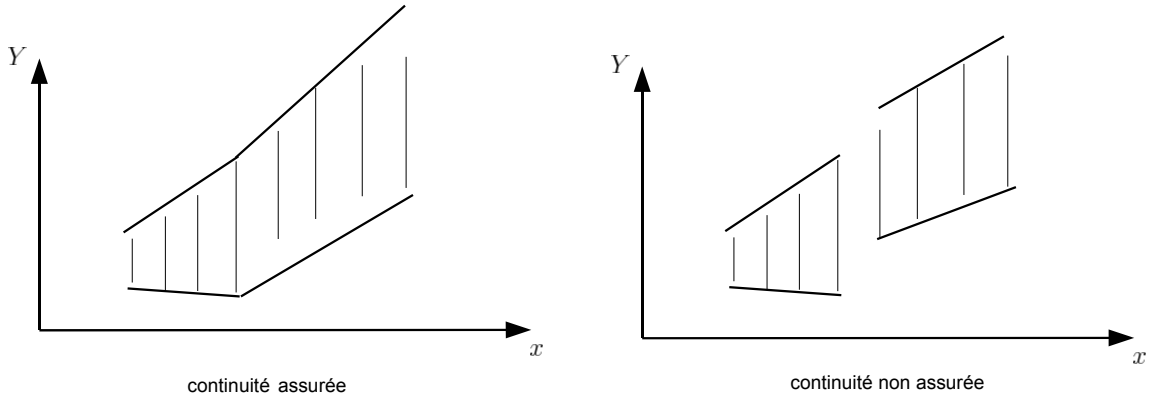


FIG. 2.24: Continuité du modèle global ?

La recherche de continuité du modèle global n'est pas abordée dans ces travaux. Une approche simple, illustrée à la figure 2.24, consisterait à rajouter, pour l'identification du modèle $k+1$, une contrainte d'égalité avec la sortie produite par le modèle k préalablement identifié. Cette façon de faire va cependant à l'encontre de la philosophie générale de notre approche, en dégradant fortement la qualité du modèle global en terme d'imprécision. Une approche plus réaliste nécessiterait de mettre en place un mécanisme d'interpolation entre sous-modèles adjacents.

Pour résumer, l'approche de régression linéaire floue par morceaux présentée ici permet d'obtenir sur un jeu de données quelconques un modèle global de bonne qualité, quoique généralement non continu. En effet, chacun des sous-modèles à décalage le composant présente une bonne représentativité, tout en garantissant l'inclusion recherchée et une imprécision optimale sur son domaine de définition.

Du point de vue de l'utilisateur, cette approche est aisée à prendre en main. En effet, la technique d'identification s'applique sur chacun des sous-ensembles d'observations, indépendamment les uns des autres. De plus, ces sous-ensembles sont obtenus lors d'une phase de segmentation dissociée de l'identification proprement dite. Par conséquent, cela ne complexifie pas outre mesure le processus global d'identification. Le dernier point concerne l'indépendance des sous-modèles qui restent donc utilisables de manière autonome.

2.5.1.2 Exemple

Afin d'illustrer les bénéfices de l'identification d'un modèle linéaire par morceaux, le jeu de données présentées dans le tableau 2.22 est considéré ([63], [71]). Celui-ci est composé de cinq

j	x_j	Y_j
1	5	$(7, [3, 11])$
2	8	$(9, [8, 10])$
3	11	$(10, [9, 11])$
4	14	$(11, [7, 15])$
5	17	$(13, [4, 22])$

TAB. 2.22: Le jeu de données observées

échantillons, le domaine de définition du modèle étant dans ce cas défini par $D = [5, 17]$.

Si la variation des valeurs modales des sorties observées en fonction de l'amplitude des entrées est considérée, comme présentée sur la figure 2.25.a, il est clair qu'une unique tendance se dégage (valeurs modales croissantes). Il n'est donc pas nécessaire de segmenter les données selon celles-ci. Cependant, si un unique modèle linéaire flou est identifié (les paramètres optimaux obtenus sont présentés dans le tableau 2.23), il est visible sur la figure 2.26 que le modèle identifié est fortement imprécis. Ainsi, s'il permet une bonne représentation de la tendance croissante des

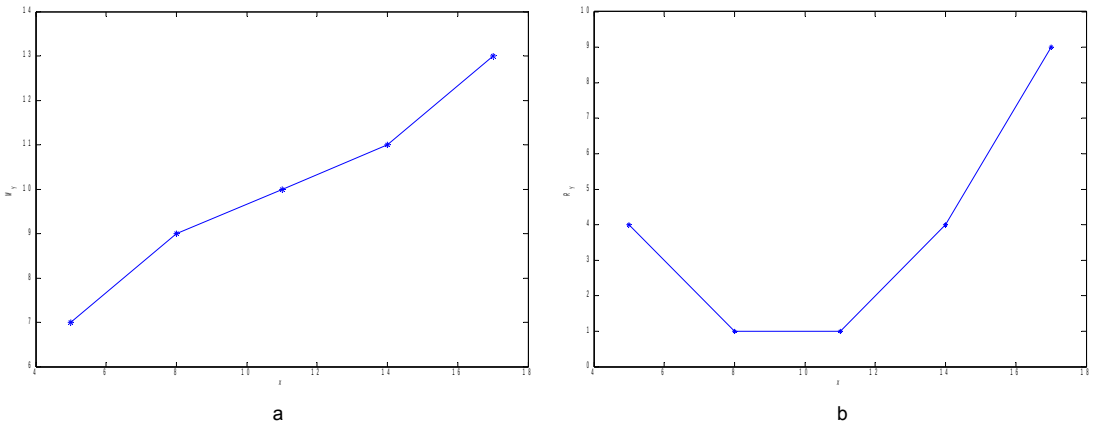


FIG. 2.25: Les variations du Midpoint des sorties observées (a) et du Radius (b)

valeurs modales des observations, il y a une perte complète d'informations quant à l'évolution de l'imprécision des données. En effet, les intervalles supports des échantillons extrêmes étant ceux d'imprécision maximale, ce sont eux qui vont saturer les contraintes d'inclusion. Par conséquent,

A_0	$([7, 7.67], [3, 11])$
A_1	$(0.44, [0.08, 11])$
$shift$	5
Domaine de définition	$[5, 17]$

TAB. 2.23: Modèle unique identifié

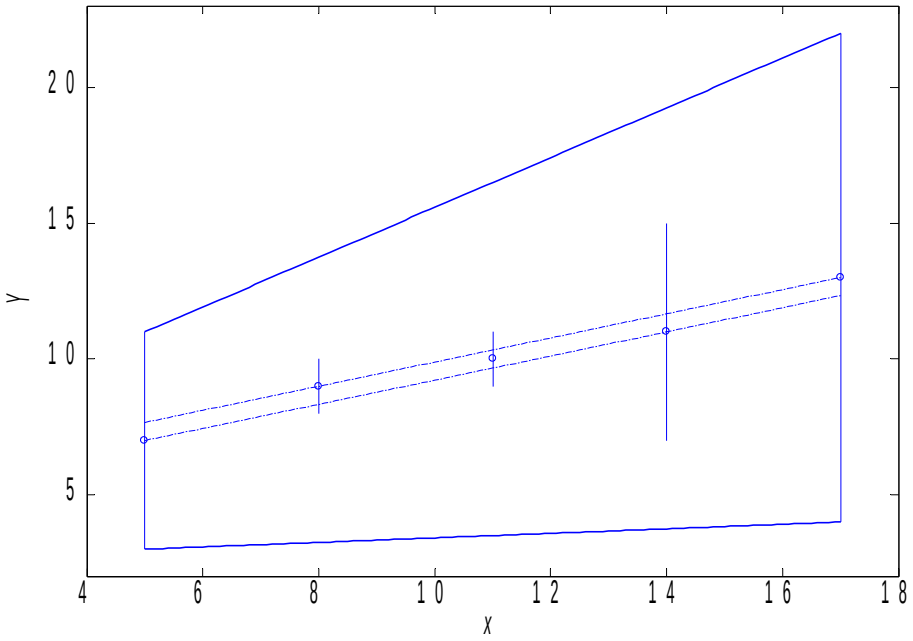


FIG. 2.26: Modèle linéaire unique identifié

les imprécisions plus faibles des données intermédiaires ne sont pas du tout représentées, car totalement transparentes au niveau des contraintes qu'elles engendrent.

Il est clair sur la figure 2.25.b que deux tendances globales de variation des imprécisions peuvent être dégagées. En effet, elles sont d'abord croissantes, puis décroissantes, le point de rupture étant dans ce cas l'observation correspondant à l'entrée $x = 11$.

Par conséquent, un modèle linéaire par morceaux est identifié. Il est composé de deux sous-modèles indépendants dont les paramètres sont fournis dans le tableau 2.24. Une représentation en est proposée sur la figure 2.27. La décomposition du modèle global en deux sous-modèles indépendants a permis de représenter différentes variations d'imprécisions. Ainsi, en fixant une valeur de décalage correspondant à la borne maximale de son domaine pour le premier sous-modèle, il est possible de représenter la variation décroissante de l'imprécision sur les premières

	sous-modèle 1	sous-modèle 2
A_0	$([10, 10.18], [9, 11])$	$([9.84, 10], [9, 11])$
A_1	$([0.39, 0.5], [0, 1])$	$([0.39, 0.5], [-0.83, 1.83])$
$shift$	11	11
Domaine de définition	$[5, 11]$	$[11, 17]$

TAB. 2.24: Modèle linéaire par morceaux identifié

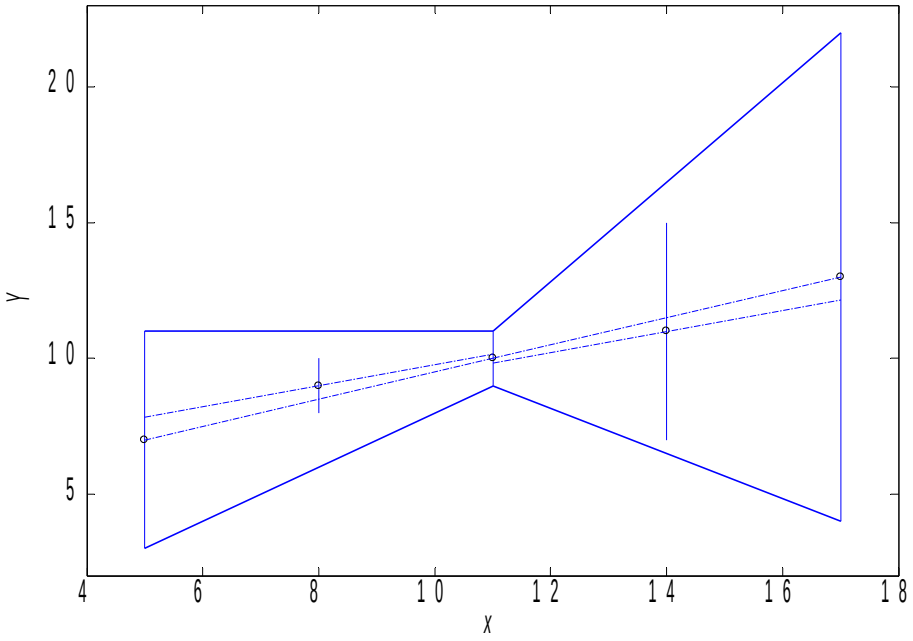


FIG. 2.27: Modèle linéaire par morceaux identifié

données positives. De manière identique, en fixant un décalage correspondant à la borne minimale du domaine de définition du second sous-modèle, l'imprécision croissante des données suivantes est bien représentée.

Il est intéressant de remarquer que pour chacun des sous modèles la donnée pour $x = 11$ sature les contraintes d'inclusion aussi bien en ce qui concerne la borne inférieure que supérieure du support. Par conséquent, la sortie évaluée des deux sous-modèles en ce point est identique au niveau des supports, ce qui permet d'assurer la continuité du modèle global au niveau $\alpha = 0$. Cependant, cette continuité n'est pas assurée pour les noyaux des sorties prédites.

2.5.2 La régression multi-entrées

Chacun des concepts présentés préalablement l'a été dans le cas de modèles linéaires mono-entrée. Or dans beaucoup de situations, la variable de sortie mesurée ne dépend pas que d'une unique variable indépendante, mais de plusieurs. Afin de généraliser l'approche présentée auparavant, plusieurs points sont à considérer. Il faut formaliser le jeu de données observées, définir la structure du modèle recherché, et adapter en conséquence la technique d'identification.

Dans un problème de régression multi-entrées, un jeu de M données est de la forme :

$$(\mathbf{x_j}, Y_j), j = 1, ..., M \tag{2.60}$$

Le vecteur d'entrées $\mathbf{x_j}$ comporte N composantes précises :

$$\mathbf{x_j} = [x_{1j}, x_{2j}, ..., x_{Nj}], j = 1, ..., M \tag{2.61}$$

Chacune des composantes $x_i, i = 1, ..., N$ du vecteur d'entrée est par ailleurs définie sur son propre domaine $D_i = [x_i^{min}, x_i^{max}]$. L'intervalle D_i représente donc l'ensemble des valeurs précises accessibles pour la i^{eme} composante de l'entrée.

Enfin, à chaque vecteur d'entrées est associée une sortie imprécise Y_j sous forme d'un intervalle flou triangulaire.

Considérant ce jeu de données observées, la structure du modèle recherché doit être adaptée. Celui-ci doit rester linéaire en ses paramètres. Ceux-ci sont recherchés sous forme d'intervalles flous trapézoïdaux, afin de conserver les propriétés liées à l'inclusion précédemment exposées. De même, il est intéressant de chercher à généraliser les bénéfices des modèles à entrée décalée en terme de représentativité de l'imprécision. Pour ce faire, sachant que les composantes du vecteur d'entrées sont considérées comme indépendantes, des décalages sont introduits pour chacune d'elles. Par conséquent, le modèle recherché est défini par :

$$\hat{Y}(x) = A_0 \oplus A_1 \odot (x_1 - shift_1) \oplus ... \oplus A_N \odot (x_N - shift_N) \tag{2.62}$$

avec :

- les paramètres $A_i, i = 0, ..., N$ sont des intervalles flous trapézoïdaux
- les décalages $shift_i, i = 1, ..., N$ sont définis tels que $shift_i \in \{x_i^{min}, x_i^{max}\}$

La sortie du modèle est donc, par propagation des imprécisions, un intervalle flou trapézoïdal, défini par son support et son noyau (cf. équation (2.4)). Bien entendu, dans le cadre d'un modèle multi-entrées, il faut considérer les expressions étendues des Midpoint et Radius des intervalles

concernés. En introduisant les changements de variable $w_i = x_i - shift_i$, on obtient :

$$\forall \mathbf{x} : \begin{cases} M_{K_{\hat{Y}}} = M_{K_{A_0}} + \sum_{i=1}^{N} M_{K_{A_i}}.w_i \\ R_{K_{\hat{Y}}} = R_{K_{A_0}} + \sum_{i=1}^{N} R_{K_{A_i}}.|w_i| \\ M_{S_{\hat{Y}}} = M_{S_{A_0}} + \sum_{i=1}^{N} M_{S_{A_i}}.w_i \\ R_{S_{\hat{Y}}} = R_{S_{A_0}} + \sum_{i=1}^{N} R_{S_{A_i}}.|w_i| \end{cases} \tag{2.63}$$

Le jeu de données observées et la structure du modèle à identifier ayant changé, il est nécessaire dans la suite d'adapter la technique régressive à utiliser en vue de l'identification des paramètres du modèle.

Comme pour un modèle mono-entrée, le fait d'introduire des décalages sur les composantes du vecteur d'entrées nécessite une première phase en amont de l'identification proprement dite visant à déterminer les valeurs appropriées de chacun des $shift_i, i = 1, ..., N$. A chacune des entrées est associé un domaine de définition D_i, construit de manière immédiate selon les observations. Il est donc possible d'étudier la variation de l'imprécision des sorties observées en fonction de la composante x_i sur le domaine D_i. Cela est fait de manière identique à ce qui a été développé pour un modèle mono-entrée (section 2.3.1).

Une fois chacune des valeurs de décalage $shift_i$ déterminée, l'ensemble des données est translaté en effectuant les changement de variables $w_{ij} = x_{ij} - shift_i, i = 1, ..., N, j = 1, ..., M$.

L'étape suivante concerne l'identification proprement dite des paramètres du modèle. Selon la philosophie de notre approche, cela doit être fait en minimisant un critère linéaire représentant l'imprécision globale du modèle sur son domaine de définition, le tout sous un ensemble de contraintes permettant de garantir l'inclusion des observations dans les prédictions, et l'obtention de paramètres flous bien définis.

Le critère à optimiser doit être une généralisation de celui proposé dans l'équation (2.46), c'est-à-dire la minimisation de l'imprécision de la sortie du modèle sur l'intégralité du domaine de définition. Or, dans le cas multi-entrées, ce domaine de définition D est défini comme étant le produit cartésien de l'ensemble des domaine de variation de chacune des composantes du vecteur d'entrées :

$$D = \bigotimes_{i=1}^{N} D_i \tag{2.64}$$

Par conséquent, l'imprécision du modèle est définie comme étant l'intégration de l'imprécision de la sortie évaluée (aire couverte par l'intervalle flou trapézoïdal correspondant (2.10)) sur

l'ensemble des domaines décalés D_{w_i}, soit :

$$volume = \int_{w_{min}^1}^{w_{max}^1} ... \int_{w_{min}^N}^{w_{max}^N} aire(\hat{Y}(w_1, ..., w_N))dw_N...dw_1 \tag{2.65}$$

En introduisant dans l'expression (2.10) de l'aire de la sortie trapézoïdale $aire(\hat{Y}(w_1, ..., w_N))$ l'équation (2.63), le critère considéré est finalement :

$$J_{volume} = R_{S_{A_0}} + R_{K_{A_0}} + \sum_{i=1}^{N}(R_{S_{A_i}} + R_{K_{A_i}}) \cdot |M_{D_{w_i}}| \tag{2.66}$$

L'extension des contraintes à considérer au cas multi-entrées est immédiate. En effet, il suffit d'introduire l'expression de la sortie du modèle (2.63) dans les contraintes assurant l'inclusion des observations dans les prédictions au niveau des noyaux (équation (2.13)), des supports (équation (2.14)), ainsi que dans les contraintes garantissant l'inclusion du noyau dans le support des paramètres (équation (2.15), $i = 0, ..., N$). Bien entendu, les contraintes de positivité des Radius des paramètres doivent concerner tous ceux-ci, c'est-à-dire :

$$R_{K_{A_i}} \geq 0 \text{ , et } R_{S_{A_i}} \geq 0, \forall i = 0, ..., N \tag{2.67}$$

Il est clair que le critère généralisé, ainsi que les contraintes, sont encore linéaires. L'identification d'un modèle trapézoïdal décalé et la présence des imprécisions aux deux niveaux (support et noyau) dans le critère permettent de respecter l'inclusion totale, tout en garantissant l'obtention d'un modèle d'imprécision minimale sur son domaine.

En ce qui concerne plus particulièrement le critère, chacune des composantes du vecteur d'entrées est considérées au travers du Midpoint de son domaine de variation $M_{D_{w_i}}$. Ainsi, le critère est encore indépendant de la répartition des données, de leur nombre, ou de leur éventuelle redondance. L'extension au cas multi-entrées du critère J_{volume} permet donc de conserver les améliorations en terme de robustesse introduite dans le cadre d'un modèle simple à une unique variable dépendante.

Il est possible d'étendre la régression par morceaux au cas multi-entrées [6]. Il est en effet envisageable de chercher à identifier un modèle global composé de plusieurs sous-modèles linéaires de la forme (2.62). Il a été discuté dans la section 2.5.1 qu'en considérant des sous-modèles indépendants et en dissociant la phase de segmentation des données de la phase d'identification des sous-modèles, le problème global était facilement soluble dans le cas mono-entrée. Il en va de même dans le cas sont multi-entrées. La difficulté majeure réside en fait dans la phase de segmentation des données.

Dans le cas mono-entrée, il a été proposé d'effectuer la segmentation d'abord sur la valeur modale des sorties observées, puis sur leur imprécision, dans l'optique de différencier les comportements croissants et décroissants par rapport à la variable d'entrée. Dans le cas multi-entrées,

cette stratégie peut être appliquée indépendamment sur chaque entrée. Le produit cartésien des diverses segmentations obtenues constitue alors la segmentation finale des données multi-dimensionnelles.

Cependant, il est clair que plus le nombre d'entrées est élevé, plus le risque d'une explosion combinatoire est grand. De plus, cette approche de la segmentation nécessite d'avoir un jeu de données initial homogène et sans incomplétude trop importante. Dans le contexte de la régression multi-entrées par morceaux, non traité dans cette thèse, l'utilisation d'une méthode adaptée de clustering des données serait probablement très profitable [23].

2.6 Conclusion

Dans le premier chapitre, plusieurs limites de la régression linéaire floue en environnement imprécis ont été soulignées dans le cadre de l'approche de Tanaka. Dans ce second chapitre, nous avons proposé différentes modifications des modèles linéaires flous régressifs permettant de lever certaines limites mentionnées. Trois contributions ont ainsi été développées.

La première concerne la structure du modèle à identifier. En cherchant des paramètres sous forme d'intervalles flous trapézoïdaux, et non plus triangulaires, il a été montré que le problème lié à la recherche de l'inclusion au sens des intervalles flous était surmonté. Cela est donc possible au prix de l'ajout de deux paramètres au niveau des noyaux des coefficients, paramètres ne complexifiant pas l'interprétation par l'utilisateur du modèle obtenu (cf. synthèse 3).

La seconde proposition concerne l'introduction d'un décalage sur les variables d'entrée dans la formulation du modèle. Cette modification permet d'améliorer la représentativité du modèle en ce qui concerne la tendance de l'imprécision des observations. Cela est réalisé sans perte de linéarité, et toujours au prix d'un unique paramètre dont l'obtention et l'interprétation sont immédiates (cf. synthèse 4).

La troisième proposition concerne la formulation d'un nouveau critère d'identification. Ainsi, il est possible de définir la notion d'imprécision globale du modèle sur son domaine de définition, cette définition fournissant un critère linéaire dont la "robustesse" est améliorée (cf. synthèse 5).

Ces trois propositions, initialement présentées et discutées pour des modèles simples ne comportant qu'une unique variable indépendante ont été généralisées pour des modèles comportant un nombre fini quelconque d'entrées. L'utilisation des concepts introduits est ensuite discutée dans le contexte de l'identification de modèles régressifs par morceaux.

Chapitre 3

La régression dans un environnement imprécis : applications

Sommaire

3.1 Introduction

Dans ce chapitre, nous présentons quelques résultats de simulation numérique, sur des exemples fréquemment utilisés dans la littérature de la régression imprécise, pour la validation des algorithmes d'identification.

La méthode développée dans les chapitres précédents est analysée et comparée à travers une série de tests. Ces tests visent notamment à évaluer qualitativement les performances de la méthode proposée.

Pour ce faire, la première section sera consacrée à l'identification de modèles polynomiaux. Cette étude permettra de mettre en évidence la problématique liée au choix de la structure adéquate du modèle à identifier, point crucial dans une approche paramétrique.

Dans un second temps, la problématique liée aux données multi-entrées sera abordée, permettant ainsi de mettre en évidence l'intérêt de l'identification de modèles de structure complexe. L' exemple traité permettra également d'éprouver la qualité des modèles obtenus lors d'une phase de validation sur des données de test, selon une procédure exploitée en apprentissage.

Enfin, la dernière section sera consacrée à l'identification de modèles linéaires dynamiques sur des données réelles, en l'occurrence issues du domaine de l'économétrie. Cette section a pour objectif de mettre en évidence les éventuelles faiblesses de la méthode proposée et d'ouvrir sur des perspectives concernant l'identification de modèles à entrées imprécises.

3.2 L'identification de modèles polynomiaux

L'objectif de cette section est d'exploiter les performances de la méthode de régression proposée dans l'identification des modèles régressifs flous polynomiaux d'ordre N. Si ces derniers conservent la propriété de linéarité en les paramètres, ils ont la possibilité de bien représenter et de mieux exhiber la relation entrées-sortie dans un jeu de données complexes.

Supposons un modèle régressif polynomial d'ordre N donné par l'équation (3.1) suivante :

$$Y = A_0 \oplus \sum_{i=1}^{N} A_i \odot (x^i - shift^i) \tag{3.1}$$

les entrées du modèle (et les décalages associés) étant définis comme les puissances successives de l'entrée x. Ce type de modèle polynomial est assimilable à un modèle linéaire multi-entrées. En effet, si on pose $\tilde{x}_i = x^i - shift^i$, le modèle (3.1) devient le suivant :

$$Y = A_0 \oplus \sum_{i=1}^{N} A_i \odot \tilde{x}_i \tag{3.2}$$

En d'autres termes, le modèle polynomial mono-entrée mono-sortie (SISO) est transformé en un système multi-entrées mono-sortie (MISO). Cette formulation (3.2) permet de transformer l'identification d'un modèle polynomial en celle d'un système linéaire. Dans ce cadre, il est important d'insister sur le fait que cette transformation est une réécriture purement analytique qui n'a pas de signification physique, dictée avant tout par des considérations d'ordre pratique concernant l'implémentation.

Si l'ordre du système est bien choisi, ces modèles peuvent représenter de façon pertinente les données entrées-sorties d'un système par un jeu de paramètres relativement restreint. La simplicité de calcul des paramètres et la transposition directe de la méthode proposée en font d'excellents candidats.

Pour l'identification de ces modèles, le problème prépondérant à résoudre réside dans la détermination de l'ordre du modèle. En d'autres termes, il s'agit de trouver le nombre suffisant ou adéquat de paramètres pour représenter les données. Dans ce contexte, il est évident qu'un ordre trop élevé risque d'alourdir le problème de régression (nombre de paramètres très important) sans apporter une amélioration pertinente de la qualité du modèle. Dans le cas inverse, un nombre de paramètres trop faible risque de mal représenter les données.

D'une manière générale, la démarche souvent exploitée pour résoudre cette problématique consiste à comparer des modèles d'ordre différent à partir d'un ensemble de données pour choisir le modèle le plus adéquat. Dans ce contexte, des méthodes ont été formalisées pour quantifier a posteriori la performance et la qualité d'un modèle en fonction de sa structure et/ou du nombre de ses paramètres. Dans ce cadre, une fonction de coût est alors définie et exploitée. Cette dernière exprime à la fois le critère associé au modèle ainsi qu'une fonction de complexité du modèle notée C pénalisant les ordres élevés. Autrement dit, pour un critère $J(A_1, ..., A_N)$, cette fonction s'exprime sous la forme suivante :

$$Cout(J(A_1, ..., A_N), N, M) = C(N, M) \cdot J(A_1, ...A_N) \tag{3.3}$$

où J représente le critère d'identification, M est le nombre de données collectées et N quantifie l'ordre du système.

Dans le cadre de la théorie de l'information, Akaike [1] et Rissanen [52] ont proposé des fonctions de coût transposables dans le contexte de notre identification. Ces dernières sont données par les équations suivantes :

- Méthode d'Akaike (AIC : critère d'information d'Akaike) :

$$Cout_{AIC}(J(A_1, ..., A_N), N, M) = M \cdot ln(J(A_1, ...A_N)) + 2 \cdot N \tag{3.4}$$

- Méthode de Rissanen (MDL : critère de la longueur de description minimum) :

$$Cout_{MDL}(J(A_1, ..., A_N), N, M) = ln(J(A_1, ...A_N)) + \frac{N}{M} \cdot ln(M) \tag{3.5}$$

Comme illustré sur la figure 3.1, l'ordre optimal du modèle correspond à celui pour lequel le critère passe par un minimum. Les deux méthodes donnent souvent des résultats comparables. Cependant, il faut attirer l'attention du lecteur sur le fait que la technique d'Akaike a tendance à surdimensionner l'ordre du modèle par rapport à celle de Rissanen.

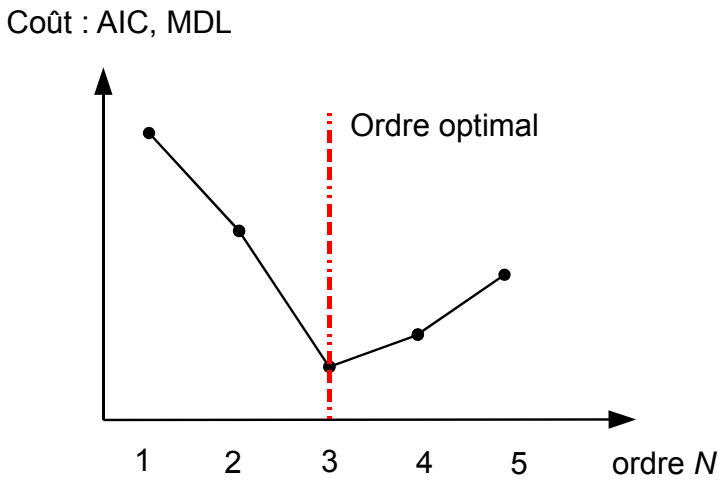


FIG. 3.1: Détermination de l'ordre optimal du modèle selon les méthodes d'Akaike ou de Rissanen

Il est important d'insister sur le fait que l'identification d'un nombre réduit et optimal de paramètres pertinents présente plusieurs avantages. Elle permet d'abord de réduire le temps de calcul et la complexité d'optimisation mise en oeuvre. Elle facilite également l'interprétation et l'exploitation du modèle.

Dans un premier temps, le jeu de données considéré est celui introduit dans [39] et [69] et présenté dans le tableau 3.1. Les entrées représentent la vitesse d'avancée d'une meule sur une surface à polir (unité : 10 mm / min). Les sorties mesurées associées donnent un indice de la rugosité finale de la surface. L'objectif est donc de déterminer un modèle exprimant la variation de la rugosité en fonction de la vitesse de polissage. Les entrées disponibles sont précises, tandis que les sorties sont données sous forme d'intervalles flous triangulaires symétriques, définis par $Y_j = (M_{Y_j}, R_{Y_j})$. D'un point de vue pratique, ceux-ci sont construits à partir d'intervalles conventionnels. Ces derniers ont pour bornes les mesures minimale et maximale de la rugosité du plan usiné, l'expérimentation étant répétée 3 fois [39]. Cela semble être le protocole le plus exploité et le plus accessible pour obtenir des intervalles flous dans le cadre d'applications industrielles où l'imprécision est prise en considération.

L'ordre optimal du modèle recherché étant inconnu, nous identifions tout d'abord un modèle

j	x_j	Y_j
1	0.75	$(0.290, 0.020)$
2	1.00	$(0.240, 0.050)$
3	1.25	$(0.240, 0.040)$
4	1.50	$(0.280, 0.035)$
5	1.75	$(0.280, 0.050)$
6	2.00	$(0.235, 0.035)$
7	2.25	$(0.230, 0.060)$
8	2.50	$(0.330, 0.130)$
9	2.75	$(0.275, 0.075)$
10	3.00	$(0.300, 0.080)$
11	3.25	$(0.335, 0.075)$
12	3.50	$(0.275, 0.055)$
13	3.75	$(0.400, 0.100)$
14	4.00	$(0.455, 0.105)$
15	4.25	$(0.420, 0.080)$
16	4.50	$(0.485, 0.115)$
17	4.75	$(0.500, 0.100)$
18	5.00	$(0.650, 0.240)$
19	5.25	$(0.640, 0.160)$

TAB. 3.1: Le jeu de données observées

linéaire flou trapézoïdal, de la forme :

$$Y = A_0 \oplus A_1 \odot (x - shift) \tag{3.6}$$

Selon le tableau 3.1, la plage de variation des entrées est déterminée comme étant l'intervalle $D = [0.75, 5.25]$. De plus, comme illustré sur la figure 3.2, l'imprécision de la sortie est clairement croissante sur ce domaine D. Par conséquent, la valeur de décalage est fixée comme étant $shift = 0.75$.

Les paramètres du modèle identifié sont présentés dans le tableau 3.2, aussi bien dans l'espace des bornes que dans l'espace Midpoint / Radius. Une représentation des données et du modèle identifié est proposée sur la figure 3.3.

La figure 3.3 illustre bien le manque de représentativité d'un modèle linéaire par rapport au jeu de données. Il est donc nécessaire de chercher à identifier un modèle d'ordre supérieur, c'est-à-dire un modèle de la forme (3.1). L'ordre adéquat étant inconnu, plusieurs identifications successives sont réalisées, de l'ordre 2 à l'ordre 5. Dans le processus d'implantation, les plages

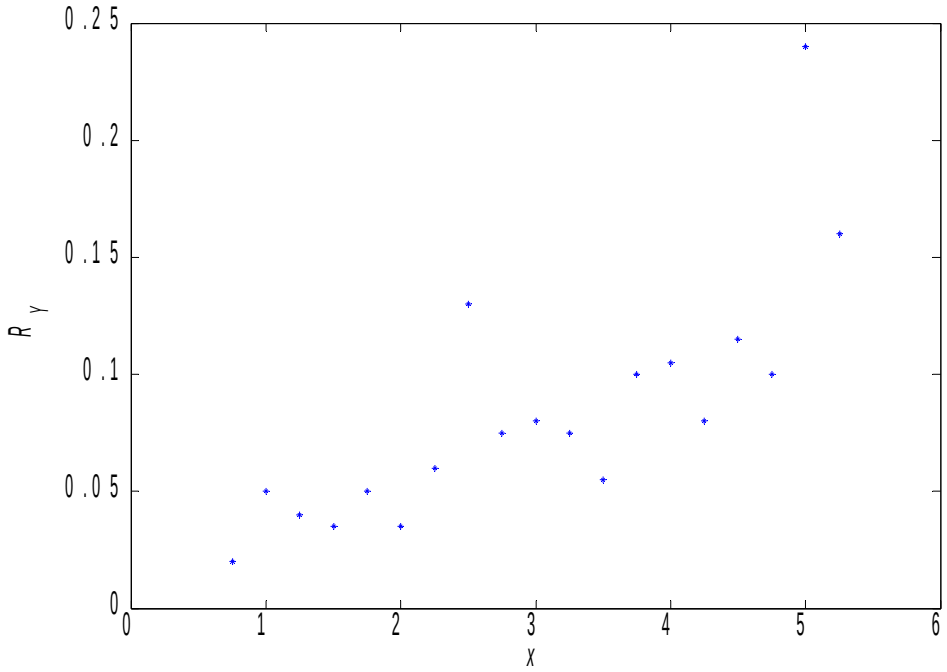


FIG. 3.2: Variation du Radius de la sortie

	Espace des bornes	Espace Midpoint / Radius
A_0	$([0.165, 0.290], [0.110, 0.310])$	$((0.2275, 0.0625), (0.21, 0.10))$
A_1	$([0.04, 0.085], [0.04, 0.136])$	$((0.0625, 0.0225), (0.088, 0.048))$
J_{volume}	0.321	

TAB. 3.2: Paramètres du modèle linéaire d'ordre 1 identifié

de variation de l'entrée $x^i, i = 1, ..., 5$ sont données par $D_i = [0.75^i, 5.25^i], i = 1, ..., 5$. Pour chacun des modèles, la valeur du critère d'optimisation J_{volume} est considérée (tableau 3.3). Il est ensuite possible d'établir l'évolution des coûts AIC et MDL en fonction de l'ordre des modèles identifiés, comme présenté sur la figure 3.4. Les deux méthodes indiquent clairement que le modèle optimal est celui d'ordre 2. Par conséquent, le modèle retenu est de la forme :

$$Y = A_0 \oplus A_1 \odot (x - shift) \oplus A_2 \odot (x^2 - shift^2) \tag{3.7}$$

Les paramètres du modèle identifié sont présentés dans le tableau 3.4, tandis qu'une représentation de sa sortie est proposée sur la figure 3.5.

Il est clair sur la représentation graphique qu'un modèle d'ordre 2 est beaucoup plus approprié pour modéliser la relation entre la rugosité du plan usiné et la vitesse de la meule. Cela se traduit

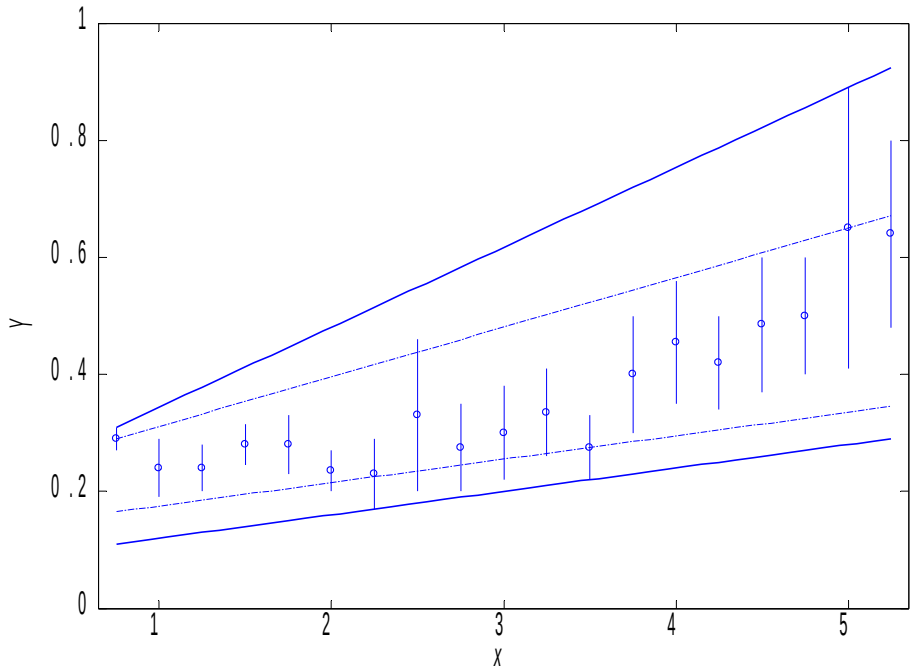


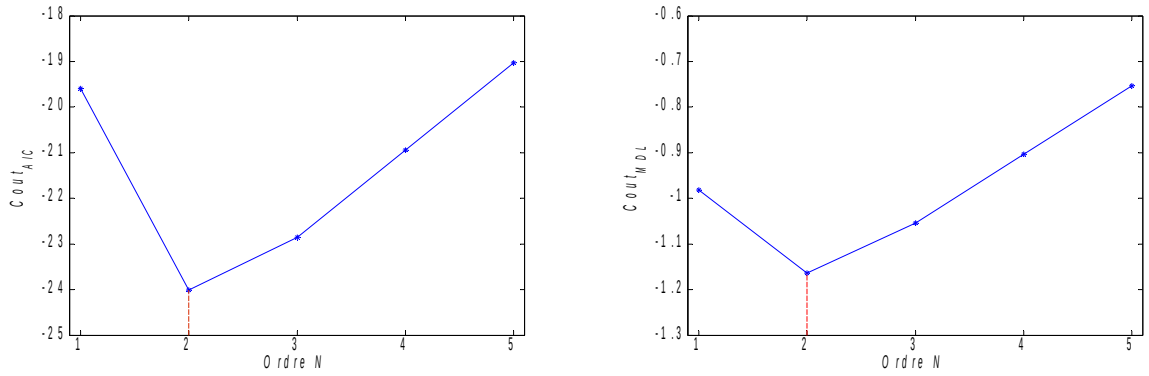
FIG. 3.3: Représentation du modèle d'ordre 1 identifié

Ordre N du modèle	Valeur du critère J_{volume}
$N = 1$	0.321
$N = 2$	0.229
$N = 3$	0.219
$N = 4$	0.218
$N = 5$	0.217

TAB. 3.3: Valeur du critère d'optimisation selon l'ordre N du modèle identifié

notamment par une nette amélioration du critère J_{volume}. Ainsi, l'imprécision globale du modèle d'ordre 2 est inférieure de 29% à celle d'un modèle d'ordre 1. Il est intéressant de constater ici que le fait de rajouter un paramètre potentiellement imprécis au modèle (le paramètre A_2 ici) peut permettre de diminuer l'imprécision globale du modèle. En effet, cela permet de rajouter un degré de liberté dans le cadre de l'optimisation, permettant une meilleure adaptation aux données, et diminuant par conséquent l'imprécision de la sortie du modèle sur le domaine de définition.

Par conséquent, le choix de la structure du modèle à identifier est un point important. Il s'agit de déterminer le bon compromis entre complexité du modèle, c'est-à-dire le nombre de



a b

FIG. 3.4: Coûts en fonction de l'ordre du modèle (a. AIC et b. MDL)

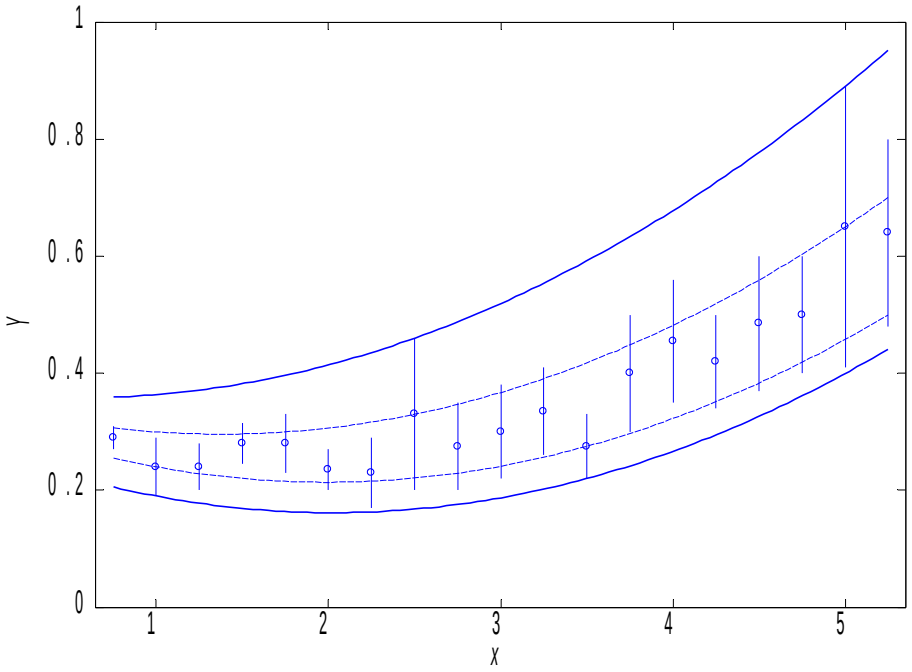


FIG. 3.5: Représentation du modèle d'ordre 2 identifié

paramètres à considérer, et gain réel en terme d'imprécision du modèle et de représentativité des données.

Les résultats obtenus ici peuvent également être mis en perspectives de ceux présentés dans [39] et [69].

Dans [69], le modèle proposé est non paramétrique, et identifié à l'aide d'une technique

	Espace des bornes	Espace Midpoint / Radius
A_0	$([0.255, 0.306], [0.205, 0.359])$	$((0.2805, 0.0255), (0.2822, 0.0768))$
A_1	$([-0.107, -0.074], [-0.109, -0.029])$	$((-0.0904, 0.0167), (-0.0694, 0.0397))$
A_2	0.0269	0.0269
J_{volume}	0.229	

TAB. 3.4: Paramètres du modèle linéaire d'ordre 2 identifié

de lissage adaptée à la manipulation d'intervalles flous triangulaires symétriques. Dans ce cas, aucun modèle explicite n'est fourni, seule une représentation graphique de l'évolution de la sortie évaluée est donnée. Cette représentation met en évidence le fait que le modèle obtenu permet une bonne représentation des sorties observées, l'inclusion des observations dans les prédictions n'étant pas recherchée. Ainsi, à la vue de ces points, notre approche présente le principal avantage de fournir un modèle paramétrique simple (polynôme d'ordre 2) dont la sortie, outre le fait qu'elle respecte l'inclusion totale des observations dans les prédictions, représente de façon tout à fait pertinente (et relativement équivalente à celle de [69]) l'évolution des mesures.

Dans [39], l'objectif de la méthode présentée est d'identifier simultanément deux modèles polynomiaux d'ordre 2 à paramètres triangulaires non symétriques. L'un des modèles, appelé modèle d'approximation supérieure, correspond au problème minimal (inclusion des observations dans les prédictions), l'autre, appelé modèle d'approximation inférieure, correspond au problème maximal (inclusion des prédictions dans les observations) (cf. section 1.4.1.3). Des contraintes sur la structure des modèles identifiés sont fixées dans [39], notamment :

- Les valeurs modales des paramètres triangulaires A_0, A_1 et A_2 sont les mêmes pour les deux modèles.
- Les dispersions gauche et droite (celles-ci étant distinctes dans le cas de paramètres flous non symétriques) des paramètres du modèle minimal sont définies comme étant celles des paramètres du modèle maximal augmentées d'une variable d'optimisation supplémentaire.

Dans ce contexte, la formulation du problème d'optimisation utilise un critère quadratique. Celui-ci est composé d'une distance permettant de positionner les valeurs modales, et d'un terme permettant la minimisation des imprécisions. L'optimisation est réalisée sous un ensemble de contraintes d'inclusions adéquates pour chaque modèle. Ces contraintes sont définies à un niveau α fixé (dans le cas présent, $\alpha = 0$).

Nous nous focaliserons ici sur le modèle minimal identifié dans [39], qui correspond à l'approche que nous avons retenue dans cette étude. Les paramètres du modèle de Lee et al. sont présentés dans le tableau 3.5, sous la forme retenue jusqu'à présent.

Selon les tableaux 3.4 et 3.5, il est clair que les modèles obtenus sont très proches en ce qui concerne les supports des paramètres. En effet, dans [39], les contraintes d'inclusion sont définies

	Espace des bornes	Espace Midpoint / Radius
A_0	$(0.375, [0.278, 0.390])$	$(0.375, (0.3344, 0.056)$
A_1	$(-0.1137, [-0.117, -0.0442])$	$(-0.1137, (-0.0806, 0.0367))$
A_2	$(0.0288, [0.0286, 0.0288])$	$(0.0288, (0.0287, 0.0001)$
J_{volume}	0.168	

TAB. 3.5: Paramètres du modèle linéaire (problème minimal) d'ordre 2 identifié dans [39]

à $\alpha = 0$, celles-ci sont donc identiques à celles définies dans notre approche au niveau des supports des intervalles trapézoïdaux manipulés. Ainsi, on retrouve par exemple un paramètre A_2 précis dans notre cas, et de Radius très faible dans [39].

Deux points principaux expliquent le fait que les paramètres identifiés sont similaires mais pas identiques. Premièrement, l'origine des modèles utilisés n'est pas identique. En effet, de par l'introduction des décalages sur les variables d'entrée, l'origine de notre modèle est définie en $x = 0.75$, alors que pour celui identifié dans [39], elle est définie pour $x = 0$. Comme discuté dans le chapitre précédent, cela a un impact important sur le paramètre A_0.

De plus, dans notre approche, les Midpoints des paramètres à optimiser ne sont pas introduits dans le critère linéaire, contrairement au critère utilisé par Lee et al., dans lequel les valeurs modales des paramètres triangulaires sont présents, au travers d'un terme de distance aux observations à minimiser. Par conséquent, le positionnement optimal des valeurs modales n'est pas similaire dans les deux approches étudiées.

Le critère J_{volume} est plus petit pour le modèle de [39] puisqu'à support identique, l'imprécision globale d'un modèle triangulaire est inférieure à celle d'un modèle trapézoïdal. Le meilleur critère J_{volume} obtenu par Lee et al. n'est donc pas attribuable à un support plus précis, mais au choix d'un modèle triangulaire. Il s'agit donc d'une nouvelle illustration du fait que la recherche de l'inclusion totale des observations dans les prédictions va de pair avec une augmentation de l'imprécision globale de la sortie du modèle sur son domaine de définition.

Pour terminer ce comparatif, rappelons que le problème maximal n'a pas toujours de solution, c'est-à-dire qu'il est possible d'avoir à considérer des jeux de données pour lesquels il est impossible de déterminer un modèle dont la sortie est incluse dans l'ensemble des observations. Par conséquent, chercher à identifier de manière simultanée les modèles minimal et maximal comme dans la méthode proposée dans [39], c'est-à-dire en définissant les paramètres du premier en fonction de ceux du second, peut éventuellement conduire à un blocage.

Afin de mieux apprécier les potentialités de notre méthode, un jeu de données plus complexes

est considéré, [21] et [16]. Ce dernier est présenté dans le tableau 3.6. Les entrées disponibles

j	x_j	Y_j	j	x_j	Y_j
1	1	$(9.000, 5.000)$	16	16	$(9.500, 3.500)$
2	2	$(6.500, 4.500)$	17	17	$(13.000, 7.000)$
3	3	$(9.500, 6.500)$	18	18	$(16.000, 6.000)$
4	4	$(10.000, 9.000)$	19	19	$(20.500, 6.500)$
5	5	$(12.500, 8.500)$	20	20	$(29.000, 8.000)$
6	6	$(20.000, 6.000)$	21	21	$(29.500, 8.500)$
7	7	$(18.500, 6.500)$	22	22	$(31.500, 5.500)$
8	8	$(21.000, 7.000)$	23	23	$(35.500, 2.500)$
9	9	$(26.000, 6.000)$	24	24	$(45.000, 4.000)$
10	10	$(28.000, 7.000)$	25	25	$(45.000, 6.000)$
11	11	$(25.500, 6.500)$	26	26	$(42.500, 4.500)$
12	12	$(26.000, 5.000)$	27	27	$(50.000, 6.000)$
13	13	$(5.000, 3.000)$	28	28	$(48.500, 4.500)$
14	14	$(7.000, 4.000)$	29	29	$(49.000, 5.000)$
15	15	$(9.500, 7.500)$	30	30	$(48.500, 4.500)$

TAB. 3.6: Le jeu de données observées

sont précises, tandis que les sorties sont données sous forme d'intervalles flous triangulaires symétriques, définis par $Y_j = (M_{Y_j}, R_{Y_j})$.

Le domaine de variation de l'entrée du modèle est définie comme étant l'intervalle $D = [1, 30]$. Si l'on considère les $k = 5$ premières et dernières données, il est possible d'obtenir l'imprécision moyenne des sorties initiale R_{init} et finale R_{fin}. Les valeurs obtenues sont $R_{init} = 6.7$ et $R_{fin} = 5.08$, ce qui implique donc $R_{init} > R_{fin}$. Par conséquent, l'imprécision des sorties sur le domaine est globalement croissante, la valeur de décalage appropriée est donc fixée à $shift = 30$.

L'ordre adéquat du modèle le mieux à même de représenter la relation entre les entrées et la sortie de ce jeu de données est inconnu. Par conséquent, des modèles d'ordre allant de $N = 1$ à $N = 5$ sont identifiés successivement. Chacun d'entre eux est représenté sur la figure 3.6. Les valeurs optimales du critère d'identification, J_{volume}, ainsi que les indicateurs d'adéquation aux données $Distance$, sont regroupés dans le tableau 3.7.

Les fonctions de coût obtenues selon les méthodes d'Akaike et de Rissanen sont ensuite déterminées pour les ordres de modèle successifs $N = 1, ..., 5$. Leur évolution selon N est représentée sur la figure 3.7.

Selon les deux méthodes de calcul de coût, le modèle optimal est d'ordre 4 (figure 3.7). Cela

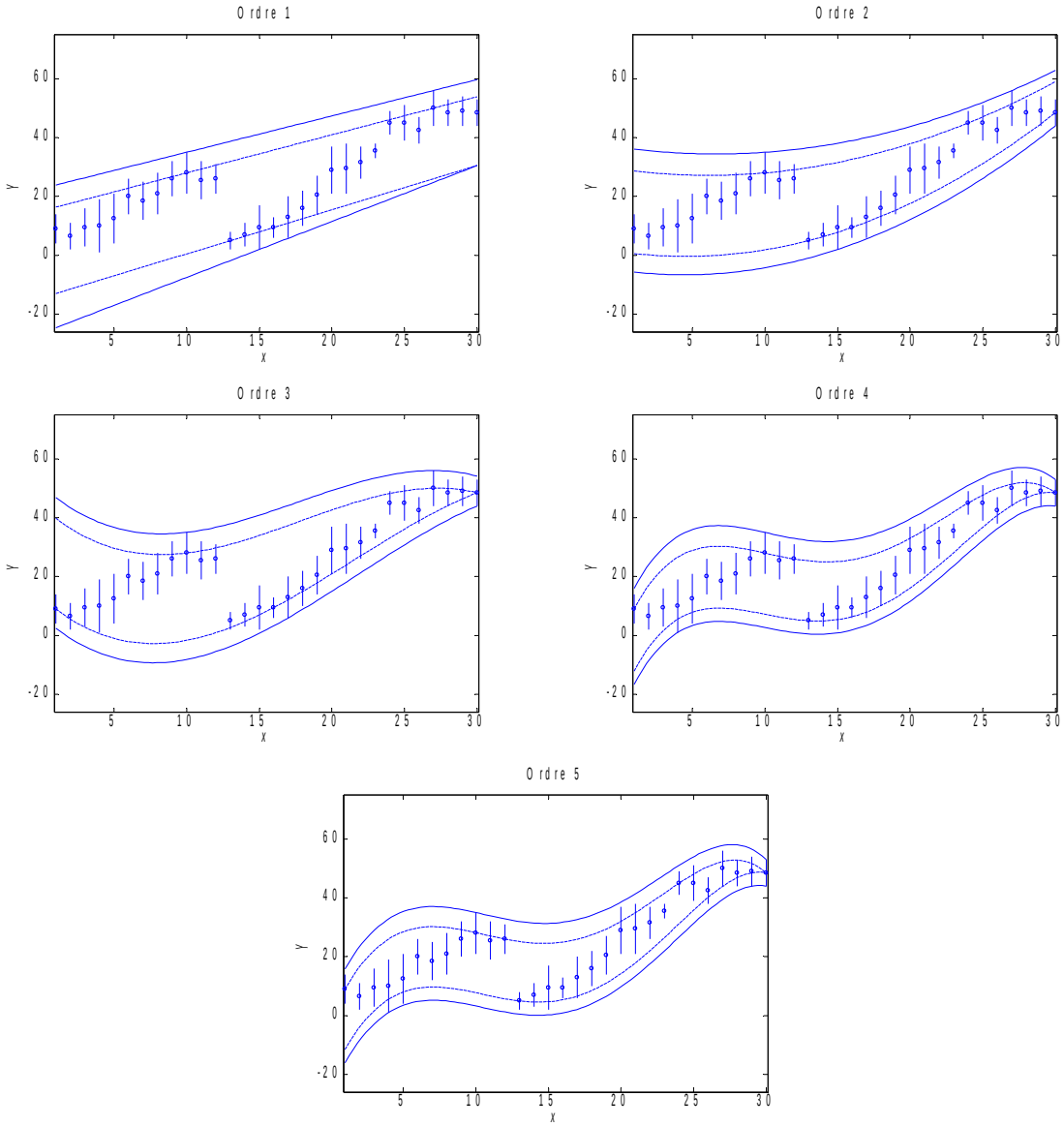


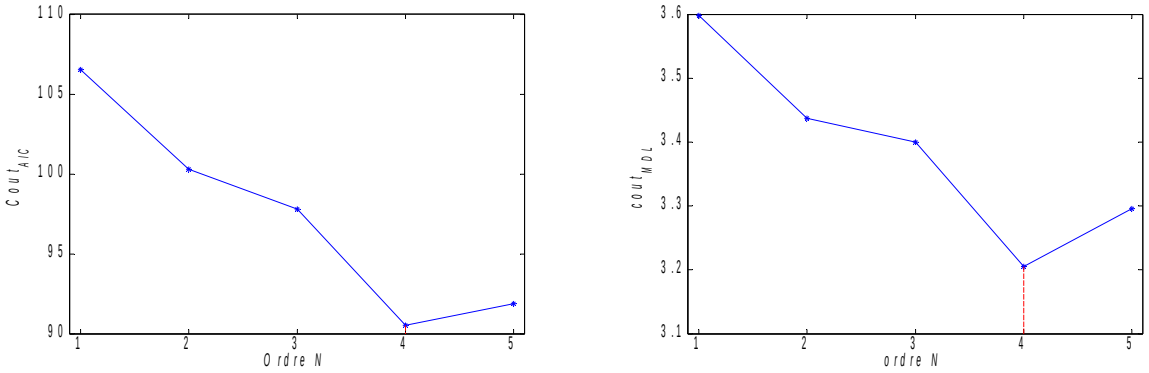
FIG. 3.6: Représentation des modèles identifiés d'ordre 1 à 5

conforte de manière rigoureuse le choix qui aurait pu être fait en considérant uniquement les représentations graphiques de la sortie des modèles d'ordre successif, comme illustré sur la figure 3.6. On remarquera notamment que le faible apport en terme de qualité de représentation des données d'un modèle d'ordre 5, mais également en terme d'amélioration du critère coût sont corrélés.

Le modèle finalement retenu est donc d'ordre 4, la valeur de décalage ayant été fixée à $shift = 30$. Ses paramètres ainsi que la valeur du critère J_{volume} et de l'indicateur $Distance$ sont synthétisés dans le tableau 3.8.

Ordre N du modèle	Valeur du critère J_{volume}	Valeur de l'indicateur *Distance*
$N = 1$	32.606	32.67
$N = 2$	24.79	27.18
$N = 3$	21.331	30.71
$N = 4$	15.666	20.548
$N = 5$	15.32	20.409

TAB. 3.7: Valeur du critère d'optimisation et de l'indicateur *Distance* selon l'ordre N du modèle identifié



a b

FIG. 3.7: Coûts en fonction de l'ordre du modèle (a. AIC et b. MDL)

A titre informatif, le modèle d'ordre optimal $N = 4$, mais de valeur de décalage non appropriée, c'est-à-dire $shift = 1$ est également présenté dans le tableau 3.8, tandis qu'une représentation en est fournie sur la figure 3.8.

Bien que l'ordre soit optimal, le modèle obtenu n'offre pas une représentation des plus pertinentes des données. Il est donc clair que la combinaison d'un modèle linéaire d'ordre adéquat déterminé à l'aide des méthodes de calcul de coût, couplé à un choix de valeur de décalage des entrées pertinent permet d'obtenir un modèle performant.

Par rapport à l'approche présentée dans [16], le modèle obtenu ici présente l'avantage d'être paramétrique. Certes, la représentation des données est moins performante, phénomène notamment dû à la présence de la cassure importante entre les données, mieux représentée dans le cadre de l'approche non paramétrique de [16], mais il est possible d'interpréter le modèle au travers de ses paramètres.

On remarquera également que l'ordre du modèle optimal, c'est-à-dire $N = 4$, est iden-

	Modèle pour $shift = 30$	Modèle pour $shift = 1$
A_0	$(48.5, [44, 53])$	$([-11.817, 9], [-14.816, 16])$
A_1	12.04	10.179
A_2	1.5237	-1.1888
A_3	0.07271	0.05228
A_4	$([-0.00111, -0.00109], [-0.00112, -0.00109])$	-0.00071
J_{volume}	15.666	25.817
$Distance$	20.548	23.637

TAB. 3.8: Paramètres des modèles d'ordre 4 obtenus pour $shift = 30$ et $shift = 1$

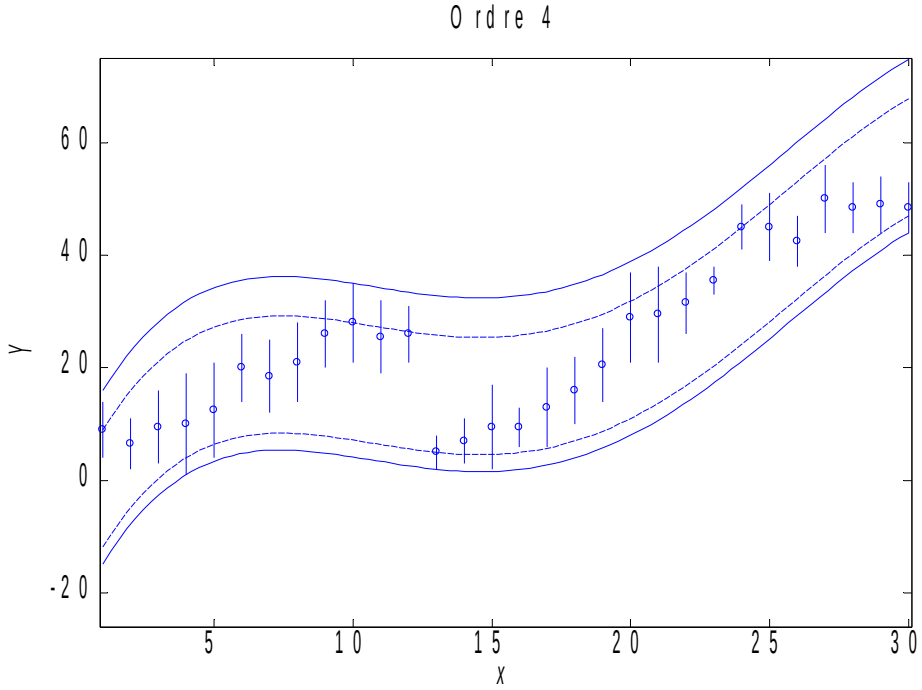


FIG. 3.8: Représentation du modèle identifié d'ordre 4 pour une valeur erronée de décalage

tique à celui retenu dans l'approche dévelopée dans [21]. Les principales distinctions entre ces deux modèles viennent de la recherche de l'inclusion des observations dans les prédictions, contrainte non introduite dans [21], et de l'introduction du décalage sur l'entrée, permettant la représentation de la décroissance de l'imprécision selon l'amplitude de l'entrée.

Après avoir montré l'intérêt que présente l'identification d'un modèle polynomial, il est possible d'aborder le problème de la présence éventuelle de points aberrants dans un ensemble

d'observations. La définition de la notion de point aberrant est particulièrement complexe. Dans [50], plusieurs définitions de différents auteurs sont rappelées ([25], [2]). Celles-ci ont en commun de mettre en évidence la subjectivité liée à la notion même de donnée aberrante.

Sur la figure 3.2, deux sorties mesurées présentent une imprécision (Radius) très distincte de celle des autres observations, et peuvent donc être considérées comme aberrantes. De plus, on voit sur la figure 3.5 que ces données correspondent aux contraintes d'inclusion saturées au niveau des supports. D'un point de vue plus général, il est envisageable que des données soient considérées comme aberrantes également au niveau de leur valeur modale, ou encore au niveau de leur imprécision et de leur valeur modale simultanément. Or, en considérant des modèles trapézoïdaux, les contraintes d'inclusion sont définies au niveau des supports et des noyaux. Par conséquent, dans tous les cas, ces points aberrants satureront là encore les contraintes à l'α-coupe concernée.

Il est possible d'adapter la technique d'identification avec détection des points aberrants par omission de données présentée dans [33]. Ainsi, il n'est pas nécessaire de retirer successivement chacune des données lors des identifications afin de quantifier la dégradation qu'elles entrainent en terme d'imprécision pour déterminer si elles doivent être considérées comme aberrantes (et par conséquent non utilisées dans l'identification du modèle finalement retenu). Il peut être envisagé de ne retirer que les données saturant les contraintes aux deux niveaux support et noyau, limitant ainsi le nombre d'identifications successives à réaliser, point contraignant dans le cas d'un jeu de données comportant un nombre conséquent d'échantillons.

3.3 L'identification de modèles multilinéaires

Dans cette section, une étude expérimentale, à travers un système multi-entrées mono-sortie donné dans la littérature [11], sera menée pour mettre en évidence l'influence du choix de la structure du modèle sur la qualité de l'identification.

Dans cette application, le jeu de données, composé de 100 échantillons, présenté dans [11] est considéré (cf. annexe A). La taille du modèle est supposée connue. En d'autres termes, le nombre des entrées est disponible ou a été déterminé a priori. L'objectif est alors d'identifier à l'aide de notre méthode un modèle régressif le plus représentatif de la relation entrées-sortie entre les cinq variables d'entrée $x_1, ..., x_5$ et la variable de sortie y.

Le protocole de génération de ces données détaillé dans [11] est basé sur l'utilisation d'un modèle mathématique précis que nous supposerons ici inconnu. Certains échantillons de sortie (au nombre de 10, dénommés points aberrants dans [11] et dans l'annexe A) sont bruités, afin d'étudier l'impact de ceux-ci sur le modèle identifié. Le bruit additif est une valeur aléatoire de distribution normale dont l'amplitude est comprise entre 0 et 2.

D'une manière générale et comme supposé dans la plupart des méthodes d'identification, le modèle obtenu est validé sur un ensemble de test. Cet ensemble peut être généré à partir des données aléatoires lorsque le système le permet. Dans d'autres situations, les données disponibles auront été divisées au préalable en deux sous-ensembles, un ensemble d'apprentissage et un ensemble de test. C'est cette approche qui sera exploitée dans cette application. En effet, sur les 100 échantillons disponibles, seuls 90 (dont les 10 bruités) sont utilisés pour l'identification, 10 étant conservés comme données de test et de validation. Il est important d'attirer l'attention du lecteur sur le fait qu'en présence d'un nombre très réduit de données, il est possible de faire appel à des outils de validation statistiques et de construction d'intervalles de confiance.

Dans un premier temps un système linéaire de la forme

$$Y = A_0 \oplus \sum_{i=1}^{5} A_i \odot (x_i - shift_i) \tag{3.8}$$

est adopté dans le processus d'identification.

D'après les données de l'annexe A, il est clair que toutes les entrées sont comprises entre 0 et 1. Dans ce cas, le domaine de définition pour chacune des entrées est défini par l'intervalle $D_i = [0, 1], i = 0, ..., 5$.

L'étude empirique des variations de la sortie selon chacune des entrées (cf. figure 3.9) permet de constater qu'elle peut être considérée comme étant croissante selon les entrées x_1, x_2, x_3 et x_5, et décroissante selon x_4. Par conséquent, dans un premier temps, les décalages retenus sont définis par $shift_i = 0, i = \{1, 2, 3, 5\}$ et $shift_4 = 1$.

Les paramètres du modèle de la forme (3.8) ainsi obtenus sont présentés dans le tableau 3.9.

	Espace des bornes	Espace Midpoint / Radius
A_0	$[-2.9212, -1.8606]$	$(-2.3909, 0.5303)$
A_1	5.0114	5.0114
A_2	5.7704	5.7704
A_3	$[4.8272, 7.5386]$	$(6.1829, 1.3557)$
A_4	$[0.8991, 1.9578]$	$(1.4284, 0.5293)$
A_5	$[0.63901, 1.4114]$	$(1.0252, 0.3862)$
J_{volume}	3.3319	

TAB. 3.9: Paramètres du modèle identifié

Les sorties observées étant précises, les paramètres obtenus (et par conséquent la sortie du modèle) sont réduits à des intervalles, c'est-à-dire que le noyau et le support sont identiques.

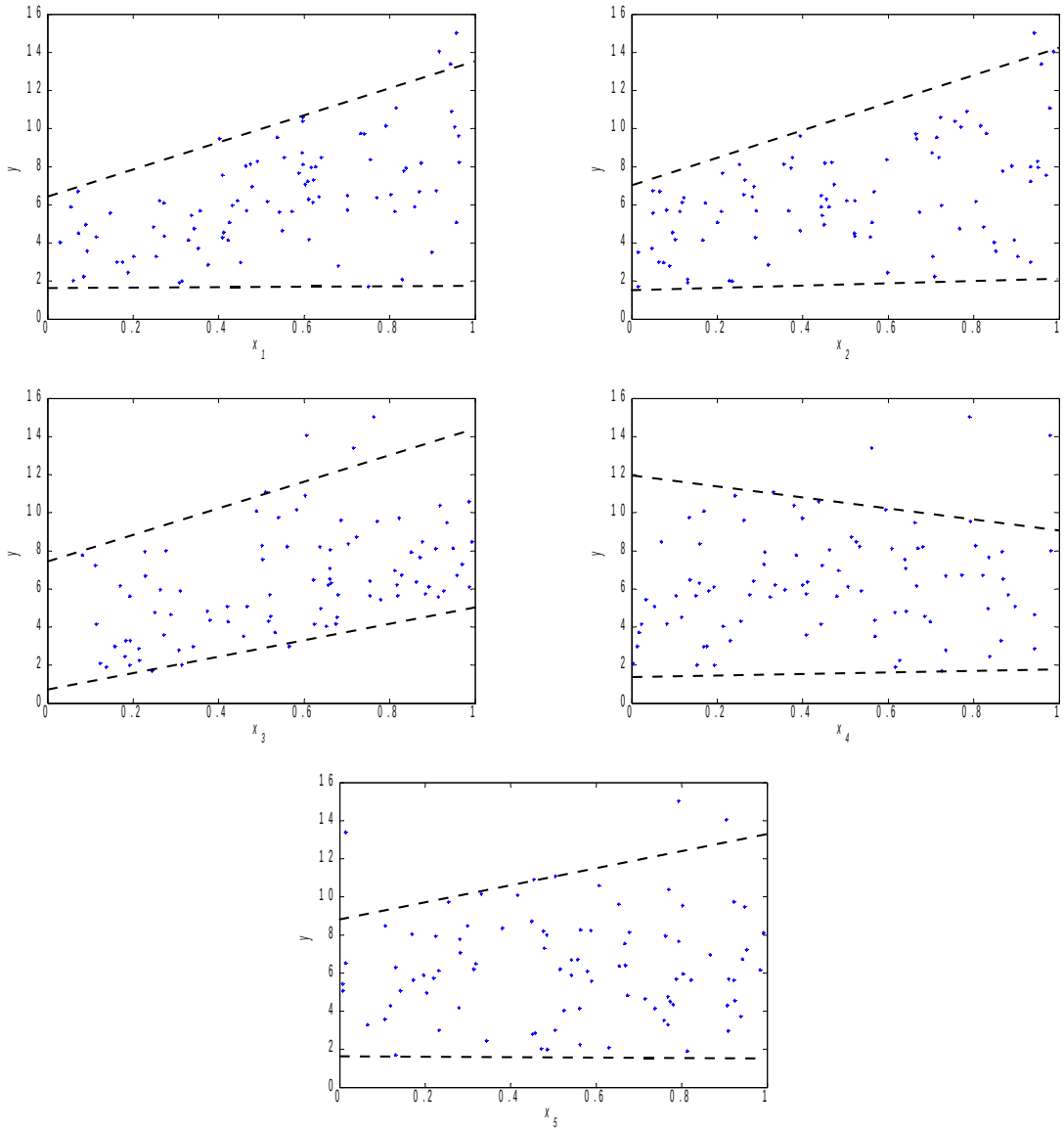


FIG. 3.9: Evolution de la sortie selon les composantes du vecteur d'entrées

Cela valide le fait que l'approche proposée concernant l'identification de modèles trapézoïdaux généralise l'approche par intervalles, dans le cas où l'imprécision des observations est nulle.

Après l'identification du modèle, une phase de validation de ce dernier est nécessaire. Dans ce contexte, la même stratégie de validation que celle détaillée dans [11] est exploitée. Il s'agit, pour chacune des 10 données de validation, de déterminer l'erreur entre observations et prédictions.

Dans [11], la sortie du modèle identifié est un intervalle flou triangulaire symétrique. Dans la phase de test réalisée sur les 10 données, l'erreur relative entre la sortie y et la valeur modale

$M_{S_{\hat{Y}}}$ de la sortie évaluée du modèle identifié est quantifiée. Cette erreur relative est définie par :

$$erreur\ relative = \frac{M_{S_{\hat{Y}}} - y}{y} \times 100 \tag{3.9}$$

L'erreur relative moyenne pour l'ensemble des données de test est également considérée.

Par ailleurs, l'inclusion de la sortie y dans l'intervalle de sortie estimé (prédiction) $\hat{Y}$ à l'aide du modèle identifié est vérifiée, cet aspect étant un point essentiel de notre approche.

Il est possible de comparer les résultats de validation de notre modèle avec ceux présentés dans [11] pour deux modèles triangulaires symétriques à entrées non décalées.

Le premier, nommé dans la suite "modèle FR-Tanaka" est obtenu selon la méthode de Tanaka [66], ses paramètres sont présentés dans le tableau 3.10.

	Espace des bornes	Espace Midpoint / Radius
A_0	$[-50.4622, 36.4618]$	$(-7.1502, 43.612)$
A_1	8.6767	8.6767
A_2	7.0221	7.0221
A_3	4.448	4.448
A_4	1.3817	1.3817
A_5	5.6052	5.6052

TAB. 3.10: Paramètres du modèle FR -Tanaka

Le second, nommé dans la suite "modèle FR-Peters" est obtenu selon la méthode de Peters [49], ses paramètres sont présentés dans le tableau 3.11. Précisons ici que celle-ci n'est pas basée sur la recherche d'une relation d'inclusion entre observations et prédictions, mais sur une notion de proximité entre elles, et permet de détecter les points aberrants, ceux-ci étant par définition les plus "éloignés" selon [11]

Les résultats de la validation du modèle obtenu à l'aide de notre méthode sont présentés dans le tableau 3.12, ceux issus du modèle FR-Tanaka dans le tableau 3.13 et ceux issus du modèle FR-Peters dans le tableau 3.14.

En ce qui concerne les erreurs relatives, on constate que le modèle obtenu présente la valeur moyenne des erreurs la plus basse. Le fait d'avoir introduit un décalage approprié sur l'entrée x_4 contribue à cette amélioration, en permettant une meilleure représentation de la sortie selon cette entrée. On remarquera cependant que les valeurs bruitées n'étant pas similaires à celles présentées dans [11], l'erreur relative moyenne pour notre modèle est du même ordre de grandeur que celle obtenue pour le modèle FR-Peters.

	Espace des bornes	Espace Midpoint / Radius
A_0	$[-5.4449, -3.2449]$	$(-4.3449, 1.1)$
A_1	5.4541	5.4541
A_2	6.0718	6.0718
A_3	7.1213	7.1213
A_4	0.85327	0.85327
A_5	1.078	1.078

TAB. 3.11: Paramètres du modèle FR -Peters

j	y	$M_{\hat{Y}}$	Erreur relative (%)	$\hat{Y}$	Inclusion
1	1.3915	1.5300	9.9503	$[0.2675, 1.8734]$	Oui
5	1.1886	-0.0340	102.8584	$[-1.9941, 0.4201]$	Non
26	1.7484	2.2275	27.4029	$[1.6781, 2.9302]$	Oui
32	2.3165	2.0436	11.7818	$[0.5660, 2.9745]$	Oui
33	4.0561	3.7446	7.6806	$[3.0730, 5.3822]$	Oui
34	5.1463	5.6391	9.5763	$[3.8384, 5.7367]$	Oui
41	2.2748	2.7977	22.9880	$[1.4329, 3.4925]$	Oui
67	1.4327	0.4190	70.7555	$[-1.3283, 1.3375]$	Non
91	2.8960	5.1179	76.7243	$[3.8305, 5.4942]$	Non
93	2.7117	2.4326	10.2926	$[1.5477, 3.5363]$	Oui
Moyenne			35.0011		

TAB. 3.12: Test sur les données de validation - modèle obtenu

En ce qui concerne l'inclusion des sorties observées dans les intervalles prédits, on voit dans le tableau 3.12 que sur les 10 données de validation, 3 ne respectent pas cette inclusion (indices $j = 5, 67, 91$). Cette violation de l'inclusion apparait pour les données dont l'erreur relative entre la prédiction et l'observation est très élevée. Cela montre que le modèle identifié, de par sa structure non adaptée, échoue à représenter de manière correcte les données d'identification, et par conséquent à être performant sur les tests de validation. On remarquera que dans le cas du modèle FR-Peters (tableau 3.14), ces trois mêmes données ne respectent pas là encore l'inclusion, ainsi qu'une donnée supplémentaire ($j = 32$). Ce phénomène s'explique par le fait que la recherche de l'inclusion n'est pas un objectif de cette approche.

Pour le modèle FR-Tanaka (tableau 3.13), l'inclusion est respectée pour toutes les données de validation. Cependant, les intervalles de sortie évalués sont larges, et donc peu représentatifs des données. Ce point explique d'ailleurs la valeur importante des erreurs relatives, les Midpoints de ces intervalles étant peu en accord avec les observations.

j	y	$M_{\hat{Y}}$	Erreur relative (%)	$\hat{Y}$	Inclusion
1	1.3915	2.1465	54.2585	$[-40.5236, 46.7004]$	Oui
5	1.1886	1.5744	32.4573	$[-45.4422, 41.7818]$	Oui
26	1.7484	0.6220	64.4257	$[-39.3759, 47.8481]$	Oui
32	2.3165	3.6046	55.6067	$[-41.8156, 45.4084]$	Oui
33	4.0561	3.0201	25.5405	$[-40.5268, 46.6972]$	Oui
34	5.1463	9.3383	81.4570	$[-35.4950, 51.7290]$	Oui
41	2.2748	4.6264	103.3776	$[-39.2658, 47.9582]$	Oui
67	1.4327	1.7072	19.1602	$[-45.6398, 41.5842]$	Oui
91	2.8960	6.2691	116.4729	$[-36.5643, 50.6597]$	Oui
93	2.7117	2.4022	11.4142	$[-40.8477, 46.3763]$	Oui
Moyenne			56.418		

TAB. 3.13: Test sur les données de validation - modèle FR-Tanaka

j	y	$M_{\hat{Y}}$	Erreur relative (%)	$\hat{Y}$	Inclusion
1	1.3915	0.7761	44.2273	$[-0.6600, 1.5400]$	Oui
5	1.1886	-0.6491	154.6114	$[-2.6705, -0.4705]$	Non
26	1.7484	1.5243	12.8155	$[0.6259, 2.8259]$	Oui
32	2.3165	1.7202	25.7395	$[0.0997, 2.2997]$	Non
33	4.0561	3.7278	8.0949	$[2.6742, 4.8742]$	Oui
34	5.1463	5.1383	0.1564	$[3.2879, 5.4879]$	Oui
41	2.2748	2.3564	3.5854	$[0.8592, 3.0592]$	Oui
67	1.4327	0.0687	95.2046	$[-1.8168, 0.3832]$	Non
91	2.8960	4.5491	57.0830	$[3.0985, 5.2985]$	Non
93	2.7117	2.1251	21.6329	$[0.9148, 3.1148]$	Oui
Moyenne			42.315		

TAB. 3.14: Test sur les données de validation - modèle FR-Peters

Cette analyse montre que notre approche donne des résultats similaires à ceux obtenus avec la méthode FR-Peters. Sur cet exemple, la recherche de l'inclusion n'apporte pas de contrepartie négative en terme d'erreur relative lors de la phase de validation par rapport à la méthode de FR-Peters qui est basée sur une notion de proximité. Or, cette notion de proximité impose à l'utilisateur de déterminer a priori un certain nombre de paramètres à insérer dans le problème d'optimisation. Selon [11], ces paramètres découlent de la connaissance experte que doit avoir l'utilisateur du jeu de données et de ses objectifs en terme de représentation des données. Notre approche représente donc une bonne alternative dans l'optique de l'obtention

d'un modèle linéaire flou.

La deuxième phase de l'étude menée sur cet exemple consiste à considérer un autre modèle à identifier, c'est-à-dire un modèle certes toujours linéaire en ses paramètres, mais dont les entrées ne sont plus nécessairement $x_1, ..., x_5$.

Comme discuté précédemment, la méthode d'identification proposée reste applicable sur tout type de système où l'expression de la sortie est linéaire en les paramètres (modèle linéaire, modèle polynomial, modèle multilinéaire, ... etc). Cependant, aucune méthodologie générale permettant de définir la structure du modèle utilisé dans le processus d'identification n'est proposée. Seule une expertise et/ou une étude expérimentale permettant de juger de la pertinence et de la capacité du modèle à pouvoir caractériser et représenter correctement la relation entrées-sorties sont exploitées. Il est clair que la performance de notre méthode est tributaire d'un choix pertinent de la structure du modèle employé. Autrement dit, on suppose que le choix de la structure du système a fait l'objet d'un travail préalable.

D'un point de vue pratique, le choix de la structure est souvent guidé par des considérations empiriques qui consistent à essayer plusieurs structures et à adopter celle qui donne la meilleure performance. Ce raisonnement heuristique et/ou expérimental donne satisfaction dans un certain nombre d'applications. Cependant, il est difficile d'évaluer dans quelle mesure ces modèles sont réellement représentatifs des données. En effet, l'inconvénient majeur de cette approche réside dans le manque de généralisation. Si l'expertise n'est pas efficace ou si les données expérimentales ne balayent pas l'intégralité de l'espace des entrées-sorties, l'utilisation de notre technique d'identification peut s'avérer être un mauvais choix.

Dans la littérature, plusieurs techniques ont été proposées pour déterminer la structure d'un modèle. Par exemple, dans [29] est présenté le concept de régression linéaire séquentielle, visant à prendre en considération les interactions successives possibles entre les entrées. Dans [67], une méthode de sélection de la structure de modèle la plus pertinente basée sur l'utilisation d'un algorithme génétique est introduite.

Ainsi, il est possible d'incorporer une technique d'identification de la structure du modèle en amont de l'utilisation de notre méthode. Dans ce contexte, la technique proposée dans [11] peut être retenue pour déterminer la structure adéquate du modèle. Cette méthode repose sur l'utilisation d'un algorithme génétique visant à déterminer la structure paramétrique du modèle.

Le modèle à identifier est donc dorénavant de la forme :

$$y = A_1 \odot (x_1.x_2) \oplus A_2 \odot x_3 \oplus A_3 \odot (x_3.x_4) \oplus A_4 \odot (x_5^2) \tag{3.10}$$

soit,

$$y = A_1 \odot \tilde{x_1} \oplus A_2 \odot \tilde{x_2} \oplus A_3 \odot \tilde{x_3} \oplus A_4 \odot \tilde{x_4} \tag{3.11}$$

avec $\tilde{x_1} = x_1.x_2$, $\tilde{x_2} = x_3$, $\tilde{x_3} = x_3.x_4$ et $\tilde{x_4} = x_5^2$.

On remarquera ici que cette expression du modèle recherché est par ailleurs tout à fait en accord avec le modèle générateur, qui s'il était supposé inconnu dans la première partie de cet exemple, est en fait donné par :

$$y = 10.x_1.x_2 + 5.x_3 + 2.x_3.x_4 + x_5^2 \tag{3.12}$$

On remarquera que les décalages ne sont pas représentés dans l'expression (3.10). En effet, chacune des entrées de ce modèle est définie sur le domaine $D = [0, 1]$, et de par la nature du modèle générateur, les valeurs appropriées de décalage pour chacune de ces entrées est $shift = 0$. Les paramètres du modèle obtenu à l'aide de notre méthode sont présentés dans le tableau 3.15, tandis qu'une représentation en est proposée sur la figure 3.10.

	Espace des bornes	Espace Midpoint / Radius
A_1	9.9999	9.9999
A_2	$[5, 9.4373]$	$(7.2186, 2.2186)$
A_3	2	2
A_4	0.9985	0.9985

TAB. 3.15: Paramètres du modèle multilinéaire obtenu

Sur la figure 3.10, les sorties précises (donc ponctuelles) observées sont présentées, ainsi que l'intervalle de sortie du modèle identifié évalué pour chacun des échantillons considérés. Par ailleurs, la dimension 4 du vecteur d'entrée impose la représentation indicielle, bien que celle-ci soit peu intéressante en terme d'exploitation.

A titre de comparaison, les paramètres du modèle (nommé par la suite modèle GP - FR) obtenu dans [11] à l'aide d'une méthode couplant algorithme génétique et technique régressive de Peters sont fournis dans le tableau 3.16.

	Espace des bornes	Espace Midpoint / Radius
A_1	$[9.7015, 11.6005]$	$(10.651, 0.9495)$
A_2	[4.76977.0443]	$(5.907, 1.1373)$
A_3	$[0.8194, 0.8948]$	$(0.8571, 0.0377)$
A_4	$[0.6312, 1.7090]$	$(1.1701, 0.5389)$

TAB. 3.16: Paramètres du modèle GP - FR

La première constatation découlant des tableaux 3.15 et 3.16 est que les paramètres des deux modèles obtenus selon les deux approches distinctes ont des paramètres du même ordre de

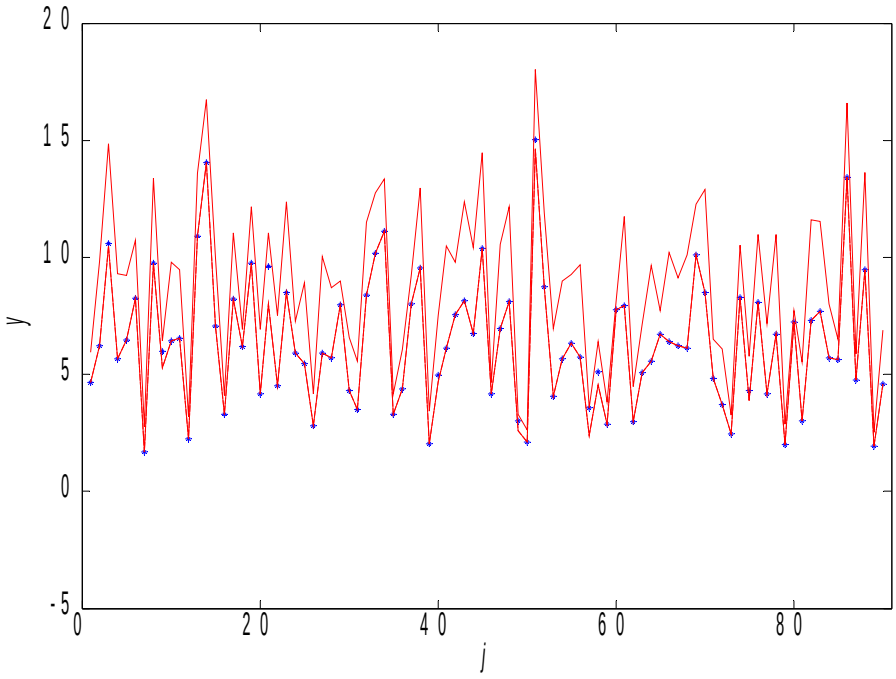


FIG. 3.10: Modèle multilinéaire obtenu

grandeur que ceux du modèle générateur. Bien évidemment, de par l'introduction de données bruitées dans le jeu d'observations, les paramètres exacts ne sont pas retrouvés.

Dans le modèle obtenu selon notre approche, c'est le paramètre A_2 qui est fortement impacté pas ces données bruitées, sa valeur de Midpoint étant $M_{S_{A_2}} = 7.2186$, la valeur exacte étant rappelons le $a_2 = 5$. Dans le cas du modèle GP - FR, c'est le paramètre A_3 qui est le plus dégradé. Ces différences s'expliquent évidemment par le fait que les techniques d'identifications ne sont pas basées sur le même objectif.

Un point intéressant concernant le modèle que nous obtenons est que si le paramètre A_2 a une valeur de Midpoint différente de celle du modèle générateur, sa borne inférieure est elle égale à 5. Les autres paramètres obtenus étant très proches de ceux du modèle générateur, et précis, on remarque donc que la borne inférieure de la sortie du modèle représente de manière quasiment parfaite les données, et donne une expression tout à fait pertinente du modèle générateur. Cela se voit très bien sur la figure 3.10. En effet, les données non bruitées sont parfaitement représentées par la borne inférieures de la sortie du modèle identifié, les données bruitées impactant uniquement la borne supérieure (et bien entendu, la valeur du Midpoint de la sortie, donc des paramètres, phénomène bien visible avec la représentation Midpoint / Radius de ceux-ci dans le tableau 3.15).

Dans le cas du modèle GP - FR (tableau 3.16), ce sont les Midpoints des intervalles définissant les paramètres qui sont les plus proches des valeurs initiales du modèle générateur. Cependant, ils n'en fournissent pas une expression aussi pertinente que dans le cas du modèle défini à l'aide des bornes inférieures des paramètres obtenu à l'aide de notre approche.

Les deux modèles sont ensuite validés sur les données de test. Les résultats de cette phase de validation sont présentés dans les tableaux 3.17 et 3.18.

j	y	$M_{\hat{Y}}$	Erreur relative (%)	$\hat{Y}$	Inclusion
1	1.3915	1.4240	2.3379	[1.3911, 1.4570]	Oui
5	1.1886	1.2257	3.1211	[1.1872, 1.2642]	Oui
26	1.7484	1.7770	1.6330	[1.7484, 1.8056]	Oui
32	2.3165	2.4808	7.0925	[2.3159, 2.6457]	Oui
33	4.0561	4.2129	3.8655	[4.0561, 4.3697]	Oui
34	5.1463	5.1715	0.4894	[5.1450, 5.1980]	Oui
41	2.2748	2.4392	7.2257	[2.2743, 2.6040]	Oui
67	1.4327	1.5108	5.4490	[1.4318, 1.5897]	Oui
91	2.8960	2.9482	1.8023	[2.8955, 3.0009]	Oui
93	2.7117	2.7713	2.1977	[2.7117, 2.8309]	Oui
Moyenne			3.5214		

TAB. 3.17: Test sur les données de validation - modèle multilinéaire obtenu

j	y	$M_{\hat{Y}}$	Erreur relative (%)	$\hat{Y}$	Inclusion
1	1.3915	1.5011	7.8742	[1.2456, 1.7566]	Oui
5	1.1886	1.3638	14.7368	[0.8258, 1.9018]	Oui
26	1.7484	1.8557	6.1362	[1.6828, 2.0286]	Oui
32	2.3165	2.5158	8.6049	[2.0809, 2.9507]	Oui
33	4.0561	4.3602	7.4963	[3.9282, 4.7921]	Oui
34	5.1463	5.5580	8.0006	[4.7108, 6.4053]	Oui
41	2.2748	2.4373	7.1441	[2.0380, 2.8366]	Oui
67	1.4327	1.6051	12.0345	[1.1705, 2.0397]	Oui
91	2.8960	3.1009	7.0766	[2.6903, 3.5116]	Oui
93	2.7117	2.8964	6.8121	[2.6060, 3.1869]	Oui
Moyenne			8.5916		

TAB. 3.18: Test sur les données de validation - modèle GP-FR

Il est clair que les erreurs relatives moyennes évaluées pour les deux modèles sont bien meilleures que dans le cas où les interactions entre les entrées ne sont pas prises en compte. Dans le tableau 3.17, l'erreur relative très faible montre que la dégradation du Midpoint du paramètre A_2 ne nuit pas excessivement à la pertinence des sorties évaluées. Cela est notamment dû au fait que les autres paramètres sont identiques à ceux du modèle générateur.

En ce qui concerne les inclusions, elle est dans le cas présent, et pour les deux modèles, respectées pour les 10 données de test. Là encore, l'identification d'un modèle de structure adéquate est donc prépondérante.

De plus, dans le cas de notre approche, les sorties générées correspondent aux bornes inférieures des intervalles de sorties de notre modèle évaluées aux entrées correspondantes. Ces sorties de test étant non bruitées, elles sont donc parfaitement représentées par la borne inférieure de la sortie de notre modèle.

Pour conclure cette étude, si les interactions entre les entrées sont prises en considération, c'est-à-dire si la structure du modèle à identifier est définie de manière correcte a priori ou automatiquement à l'aide d'une méthode adaptée (par exemple, dans le cas présent, un algorithme génétique), notre approche permet d'obtenir une bonne approximation du modèle générateur, en dépit de la présence de données bruitées.

3.4 L'identification de modèles dynamiques

Dans les exemples précédents, le potentiel et la validité de l'approche de régression proposée ont été expérimentés sur des systèmes dont le transfert entre les entrées et la sortie est statique. Dans cette dernière application, nous allons appliquer notre méthode à l'identification d'un type de systèmes bien connus en automatique et en théorie des systèmes, à savoir les systèmes dynamiques. L'apport des modèles régressifs imprécis et leurs limites seront étudiés sur des données boursières afin de mettre en évidence les conditions d'applicabilité de la méthode sur des données réelles et ouvrir la voie à des perspectives pour son évolution.

Dans cette application, l'approche d'identification dite "boîte noire" est exploitée sur des données boursières imprécises manipulées par des acteurs du monde de la finance, en l'occurrence l'évolution de l'indice Dow Jones. Rappelons ici que le Dow Jones est l'indice de référence de la bourse de New York, basé sur 30 valeurs industrielles. Les évolutions de son cours sont extrêmement suivies car très influentes sur les autres places financières mondiales.

D'une manière générale, les acteurs du monde de la finance sont particulièrement intéressés par l'obtention de modèles dans le cadre de l'économétrie. Ceux-ci doivent permettre d'établir des évolutions d'indices, des prévisions à plus ou moins long terme de rentabilité ou de risque.

De plus, les données mises en jeu dans ces modèles ont pour caractéristique d'être soumises à de fortes variabilités, dont les origines ne peuvent pas toujours être quantifiées, de par l'intervention d'acteurs humains aux comportements difficilement modélisables et prédictibles.

Il est clair que ce type de système ne peut pas être inféré directement à partir de lois physiques régissant son comportement. Dans ce cas, il est nécessaire de recourir à un modèle comportemental estimé directement à partir des mesures entrées-sorties. C'est cette approche qui est adoptée ici, sachant qu'aucune connaissance a priori des mécanismes physiques liant les entrées et sorties n'est supposée disponible. Autrement dit, il s'agit, à partir d'observations entrées-sortie imprécises collectées sur ce système, d'estimer les paramètres du modèle qui pourrait reproduire au mieux son comportement.

Les données utilisées concernent l'évolution du Dow Jones au cours de l'année 2009. Les valeurs de cotation de l'indice sur cette période sont présentées en annexe B. Pour chaque jour de cotation (au nombre de 252), les cours à l'ouverture et à la fermeture de la séance sont disponibles, ainsi que les valeurs minimale et maximale atteintes par l'indice en cours de séance. Dans ce contexte, il est possible de disposer, pour chaque journée de cotation, d'un intervalle de sortie d'indice $j = 1, ..., 252$ défini comme suit :

$$Y_j = [Y_j^-, Y_j^+] = (M_{Y_j}, R_{Y_j}) \tag{3.13}$$

où Y_j^- et Y_j^+ représentent les valeurs minimale et maximale atteintes par l'indice Dow Jones en séance.

Il est clair que ces données sont donc des intervalles conventionnels. Par conséquent, le modèle recherché doit être de même nature, ses paramètres seront donc recherchés sous forme d'intervalles. Or, il a été discuté précédemment que notre méthode visant à identifier des modèles flous trapézoïdaux est une généralisation de l'approche par intervalle. Ainsi, dans la suite, seuls les supports des intervalles flous identifiés seront considérés.

A partir de ces données collectées, l'objectif est d'identifier un modèle permettant d'évaluer la cotation de l'indice au jour j en fonction du cours en vigueur les jours précédents. Dans la littérature, cette identification a été traitée à travers des techniques d'apprentissage basées par exemple sur les réseaux de neurones [16]. Dans cette approche, des réseaux de neurones basés sur la manipulation d'intervalles sont utilisés dans l'optique de fournir deux modèles représentatifs des variations du Dow Jones sur une année, l'un pour la valeur minimale et l'autre pour la valeur maximale de chaque jour en fonction des valeurs des jours précédents.

Cependant, une critique que l'on pourrait faire sur cette méthode [16] est qu'elle n'est pas assez transparente, dans le sens où son interprétabilité est limitée. Or, la finalité d'un modèle est de permettre l'analyse et l'interprétation du système modélisé. Le modèle doit donc, tout en étant relativement pertinent, rester simple à appréhender. Il se trouve justement que la plupart des outils efficaces et bien maitrisés pour l'analyse et la synthèse concernent les modèles

paramétriques linéaires.

Ainsi, étant données les sorties collectées de l'indice Dow Jones sous forme d'intervalles Y_j, nous nous intéressons à l'existence d'un modèle paramétrique de la forme :

$$Y_j = A_0 \oplus \sum_{i=1}^{N}{}^{\oplus} A_i \otimes (Y_{j-i} \ominus shift_i) \tag{3.14}$$

Dans l'équation (3.14), les opérateurs $\otimes$ et $\ominus$ font référence à la multiplication et à la soustraction entre intervalles. Le modèle (3.14) n'est en fait qu'une représentation d'un modèle ARMA (cf. figure 3.11) avec des commandes nulles (le système est donc libre) et où A_0 représente un offset sur la sortie.

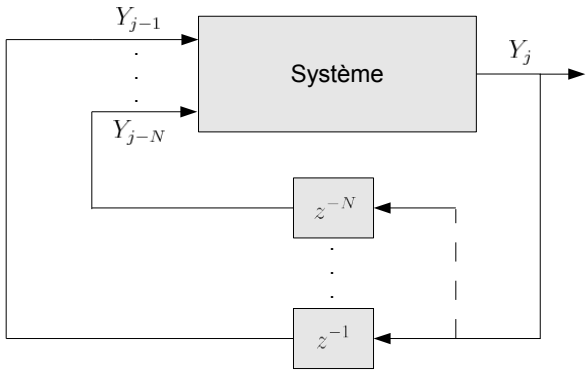


FIG. 3.11: Schéma explicatif du système considéré

La première limite de la méthode d'identification proposée concerne son implantation dans un contexte où les entrées sont imprécises. Or, elle a été définie dans le cadre des entrées précises. En d'autres termes, la propagation des imprécisions des sorties dans le mécanisme de rebouclage n'est pas pris en considération. En effet, ce sont les sorties aux jours précédents qui sont utilisées comme entrées du modèle régressif. Pour résoudre ce problème, une opération intuitive de sélection est effectuée pour transformer les entrées imprécises (intervalles) en entrées précises. Cette opération consiste simplement à ne conserver que le Midpoint de chaque sortie. Dans ce cas, le modèle (3.14) devient le suivant :

$$Y_j = A_0 \oplus \sum_{i=1}^{N}{}^{\oplus} A_i \otimes (M_{Y_{j-i}} - shift_i) \tag{3.15}$$

Il est clair que le mécanisme de sélection proposé est très restrictif en terme de représentation des imprécisions dans le système. En effet, l'information imprécise attachée aux entrées est négligée. Par voie de conséquence, cela impactera les paramètres du modèle et ainsi, la sortie du modèle. La qualité de ce dernier sera donc dégradée. Pour remédier à ce problème, une extension de la méthode proposée aux entrées imprécises est nécessaire.

Dans un premier temps, la méthode d'identification est appliquée pour déterminer un modèle sous la forme (3.15) avec des entrées précises. Ensuite, une extension pour des entrées imprécises est proposée pour identifier des modèles de la forme (3.14).

3.4.1 Modèle Entrées précises - Sortie imprécise

Comme expliqué dans le premier exemple, dans l'identification d'un modèle de la forme (3.15) à partir d'un ensemble d'observations, il est nécessaire de déterminer l'ordre optimal du système. Dans le cas présent, il s'agit dont de déterminer combien de jours précédents doivent être conservés pour évaluer la sortie du modèle au jour j. Pour ce faire, la méthode d'Akaike sera considérée.

Des modèles d'ordre $N = 1$ à $N = 5$ sont donc successivement identifiés. Plusieurs étapes sont bien entendu nécessaires dans ce processus d'identification :

- Dans un premier temps, les domaines de définition pour chacune des composantes du vecteur d'entrée $D_i = [D_i^-, D_i^+], i = 1, ..., N$ sont déterminés.
- Ensuite, il est nécessaire de fixer les valeurs de décalages appropriées. Pour ce faire, il faut étudier la variation de l'imprécision de la sortie selon les composantes du vecteur d'entrée. Selon la figure 3.12, l'imprécision des sorties observées est décroissante selon l'augmentation des amplitudes des entrées. Or, celles-ci sont exprimées en fonction des cours boursiers des jours précédents. En fait, l'imprécision de la sortie diminue avec l'amplitude de celle-ci : plus l'indice boursier est élevé moins il est imprécis, c'est-à-dire qu'il est soumis à des variations de moindre amplitude en cours de séance. Dans ce cas, les décalages sont fixés à la valeur maximale du domaine de définition, soit $shift_i = D_i^+, i = 1, ..., N$.
- La dernière étape consiste à identifier les paramètres du modèle (3.15) à l'ordre N considéré, en minimisant l'imprécision de la sortie sur le domaine de définition tout en respectant les contraintes d'inclusion des observations dans les prédictions.

Pour des modèles d'ordre $N = 1$ à $N = 5$, la valeur optimale du critère d'identification J_{volume} est présentée dans le tableau 3.19.

Ordre N du modèle	Valeur du critère J_{volume}
$N = 1$	631
$N = 2$	625
$N = 3$	615
$N = 4$	606
$N = 5$	604

TAB. 3.19: Valeur du critère d'optimisation selon l'ordre N du modèle identifié

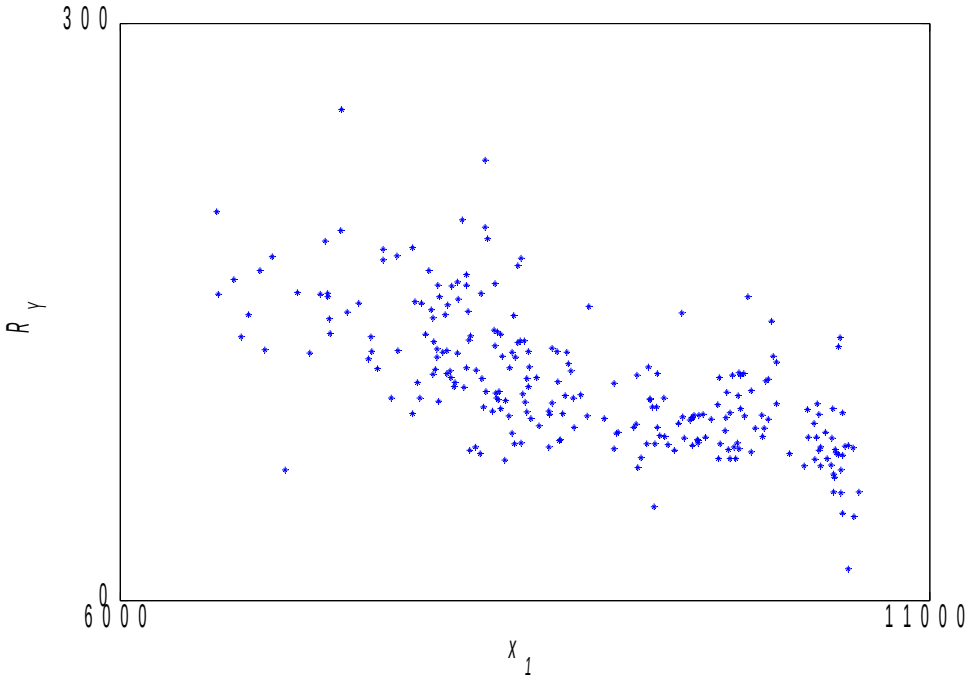


FIG. 3.12: Evolution de l'imprécision des sorties observées

La figure 3.13 expose l'évolution du critère coût évalué selon la méthode d'Akaike en fonction de l'ordre N du modèle. L'ordre optimal correspondant au minimum local, il est approprié de chercher à identifier un modèle d'ordre 4, c'est-à-dire un modèle exprimant la valeur de l'indice Dow Jones en fonction de la cotation des 4 jours précédents.

Les paramètres du modèle identifié sont présentés dans le tableau 3.20. Une représentation indicielle de la sortie du modèle est fournie sur la figure 3.14, sur laquelle les données observées sont les intervalles en pointillés, la sortie du modèle étant en traits pleins.

	Espace des bornes	Espace Midpoint / Radius
A_0	$[10321, 10788]$	$(10555, 233)$
A_1	0.807	0.807
A_2	0.242	0.242
A_3	$[-0.275, -0.118]$	$(-0.1965, 0.0785)$
A_4	0.145	0.145

TAB. 3.20: Paramètres du modèle à entrées précises identifié sur le cours du Dow Jones

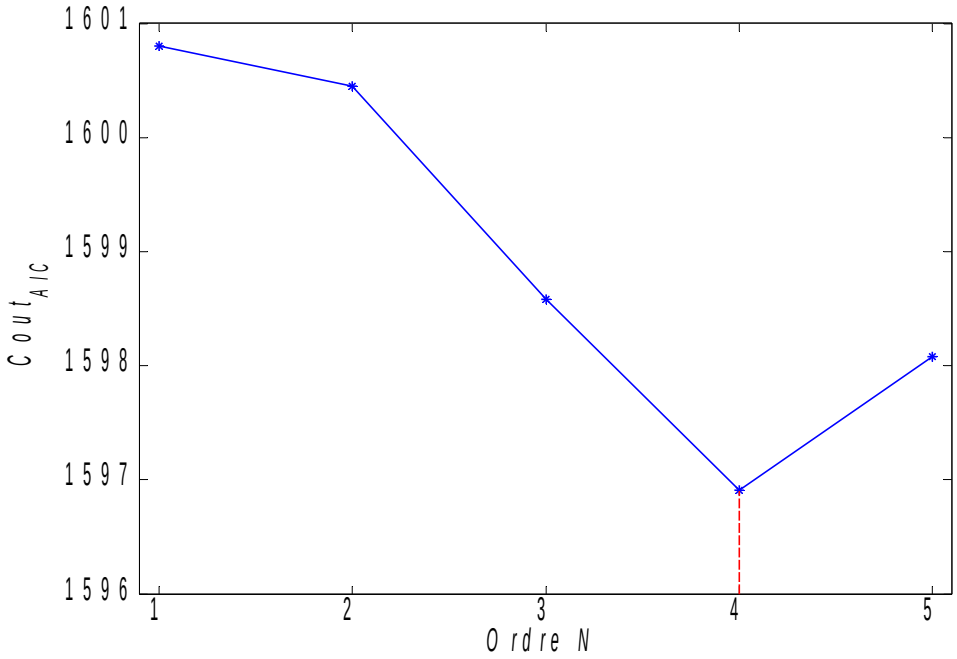


FIG. 3.13: Coût en fonction de l'ordre du modèle - méthode d'Akaike

Le modèle obtenu permet une bonne représentation de l'évolution du cours du Dow Jones sur l'année 2009. Une analyse plus fine des paramètres optimaux obtenus peut également être menée, selon le tableau 3.20.

Le paramètre A_0 représente le cours au dernier jour de l'année 2009. En effet, les valeurs de décalage introduites sur les composantes du vecteur d'entrée correspondent aux bornes maximales des domaines de définition, et l'indice étant globalement à la hausse sur l'année 2009, ces bornes sont les derniers cours. Ainsi, en permettant le positionnement de l'origine du modèle sur une des bornes de sa plage de validité, les décalages introduits permettent une interprétation concrète de ce paramètre A_0.

Le paramètre A_1 est plus important en valeur absolue que les autres paramètres A_2, A_3 et A_4, traduisant le fait relativement intuitif que la cotation du jour précédent a un impact plus élevé que celles des jours antérieurs sur le cours du jour considéré.

Le fait que le paramètre A_3 soit imprécis s'interprète par le fait que l'imprécision du cours du Dow Jones, c'est-à-dire l'amplitude de ses variations en cours de séance au jour j est corrélée à la valeur du cours moyen au jour $j-3$. Néanmoins, les enseignements à tirer d'une telle constatation sont à la discrétion d'experts financiers. Plus prosaïquement, le fait que la valeur du Midpoint de A_3 soit négative montre que la sortie au jour j est impactée négativement par la

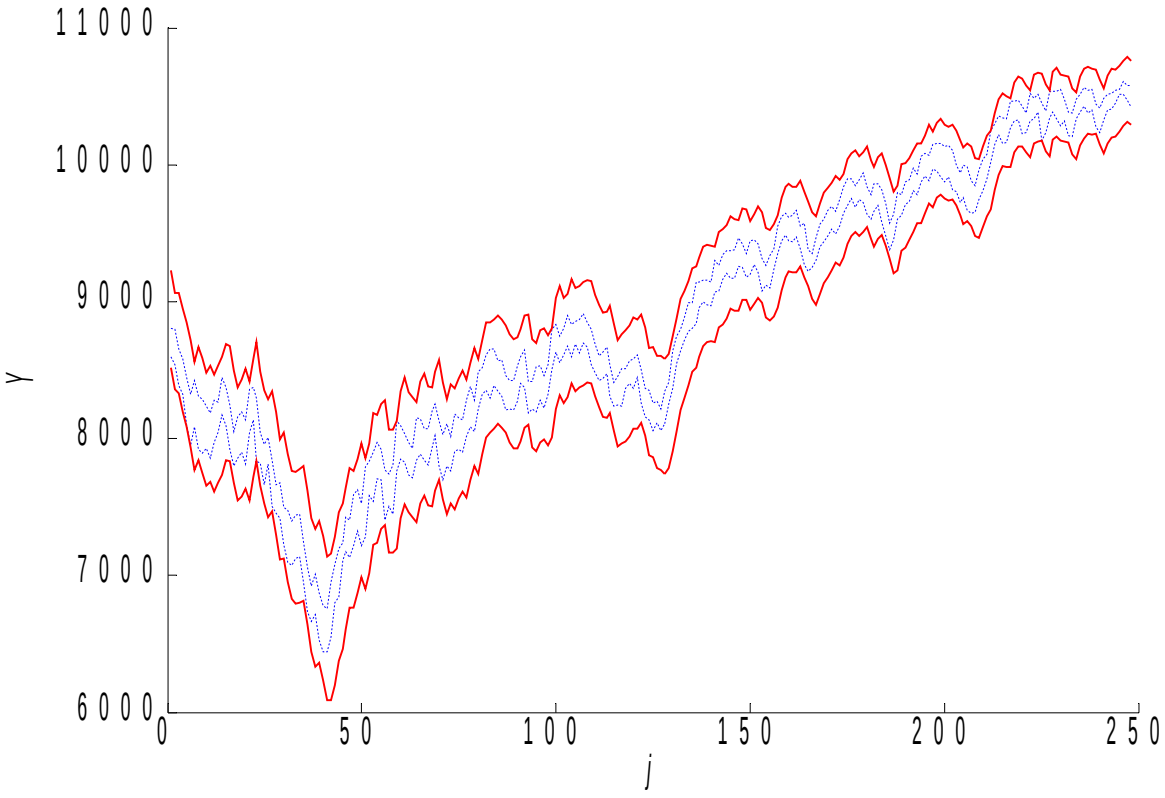


FIG. 3.14: Evolution de l'indice Dow Jones sur l'année 2009 - observations (pointillés) et prédictions (traits pleins) - modèle à entrées précises

valeur de l'indice boursier au jour $j-3$. Peut-être est-il possible d'expliquer cela par le fait que sur l'année 2009, les marchés ont été relativement agités, et que les longues périodes consécutives de hausse ou de baisse du Dow Jones ont été rares.

Cette phase d'interprétation des paramètres du modèle obtenu grâce à notre approche met en évidence un de ses avantages principaux par rapport aux approches non paramétriques telles que celle introduite dans [16].

3.4.2 Modèle Entrées imprécises - Sortie imprécise

Dans l'approche précédente, seul le cours médian de l'indice aux jours précédents est considéré comme entrées du système. Or, celui-ci produit des sorties imprécises, ces paramètres étant des intervalles. Par conséquent, la sélection opérée lors du rebouclage entraîne une perte d'une partie de l'information.

Il est donc nécessaire de chercher à identifier un modèle dont les entrées sont imprécises, afin de remédier à cette perte d'information en permettant une propagation de l'intégralité de celle-ci. Il est donc indispensable dans un premier temps d'adapter la technique d'identification, avant de l'appliquer ensuite sur les données financières à l'étude.

Nous attirons l'attention du lecteur que les concepts présentés dans la suite ouvrent sur des perspectives de travaux futurs, les différents points problématiques qu'ils soulèvent étant mis en avant.

L'extension de la méthode proposée aux entrées imprécises, comme dans le cas des modèles à entrées précises, repose sur deux principes fondamentaux, c'est-à-dire sur la minimisation de l'imprécision de la sortie du modèle sur l'intégralité de son domaine de définition, et le respect de l'inclusion des observations dans les prédictions [4], [5].

Par souci de simplicité et dans l'optique de mettre en exergue les points essentiels à discuter, nous considérons dans un premier temps un modèle linéaire de la forme :

$$\hat{Y}(X) = A_0 \oplus A_1 \otimes (X \ominus shift) \tag{3.16}$$

En considérant l'ensemble des observations, il est possible de déterminer l'intervalle D correspondant à la plage de variations des entrées. Celles-ci étant imprécises, l'intervalle D est donné par :

$$D = [\min_j X_j^-, \max_j X_j^+], j = 1, ..., M \tag{3.17}$$

L'entrée du modèle (3.16) peut donc être tout intervalle inclus dans D. Nous définissons donc l'ensemble des intervalles inclus dans D comme le domaine de définition du modèle (3.16). Ce domaine de définition est noté Δ_D, et illustré sur la figure 3.15 dans l'espace Midpoint / Radius. Le domaine de définition du modèle est donc défini par :

$$\Delta_D = \{(M, R) \in \Re^2 \mid 0 \leq R \leq R_D, M_1(R) \leq M \leq M_2(R)\} \tag{3.18}$$

avec :

$$\begin{cases} M_1(R) = M_D - R_D + R \\ M_2(R) = M_D + R_D - R \end{cases} \tag{3.19}$$

L'élément le plus imprécis du domaine de définition est l'intervalle D lui même (Radius maximal), c'est-à-dire le sommet du triangle symbolisant ce domaine. De plus, la base du triangle correspond à l'ensemble des valeurs précises du domaine de définition. Ce dernier permet donc de considérer aussi bien des entrées imprécises, c'est-à-dire tous les intervalles inclus dans D que des entrées précises, c'est-à-dire l'ensemble des réels appartenant à D.

En fixant une valeur de décalage $shift$ selon une des bornes de l'intervalles D, selon la variation de l'imprécision de la sortie du modèle, le problème d'identification revient finalement

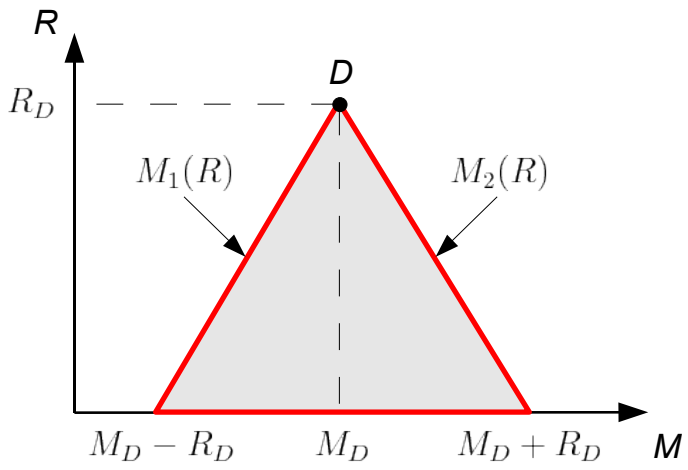


FIG. 3.15: Représentation du domaine de définition Δ_D d'un modèle à entrée imprécise

à déterminer les paramètres A_0 et A_1 du modèle de la forme :

$$\hat{Y}(W) = A_0 \oplus A_1 \otimes W \tag{3.20}$$

avec $W = X \ominus shift$ l'entrée décalée définie sur Δ_{D_W}.

Il est important de remarquer que le fait d'imposer au préalable ce changement de variable mettant en jeu un décalage $shift$ permet d'obtenir un domaine de définition Δ_{D_W} sur lequel tous les intervalles d'entrées sont de signe constant.

Dans l'approche régressive considérée, deux points sont à étudier, le critère d'optimisation d'une part, et les contraintes d'inclusion d'autre part. Il est donc nécessaire d'exprimer la sortie du modèle (3.20) sur le domaine Δ_{D_W}. Celle-ci est donnée par :

$$\hat{Y}(W) = [\hat{Y}^-(W), \hat{Y}^+(W)] = [M_{\hat{Y}(W)} - R_{\hat{Y}(W)}, M_{\hat{Y}(W)} + R_{\hat{Y}(W)}] \tag{3.21}$$

Selon l'expression 3.21 de la sortie du modèle, le problème d'identification va donc dépendre du signe des paramètres, qu'il faut supposer déterminé au préalable. Ainsi, dans la suite, nous supposerons le signe de A_1 connu a priori, afin de pouvoir expliciter les différents mécanismes mis en jeu dans la technique d'identification. De plus, seul le cas particulier d'un paramètre A_1 et d'une entrée W positifs sera détaillé, les autres cas de figure possible étant traités de manière totalement similaires. Dans ce cas, la sortie (3.21) du modèle est définie par :

$$\begin{cases} M_{\hat{Y}(W)} = M_{A_0} + M_{A_1} M_W + R_{A_1} R_W \\ R_{\hat{Y}(W)} = R_{A_0} + M_{A_1} R_W + R_{A_1} M_W \end{cases} \tag{3.22}$$

Pour déterminer le critère d'optimisation J associé à un tel modèle, l'imprécision de sa sortie, c'est-à-dire son Radius, est intégrée sur l'ensemble du domaine de définition, c'est-à-dire :

$$J = \int \int_{\Delta_{D_W}} R_{\hat{Y}}(M_W, R_W) dM_W dR_W \tag{3.23}$$

Les détails du calcul du critère sont présentés dans l'annexe C. Le critère finalement obtenu est donné par :

$$J = R_{A_0} + M_{D_W}.R_{A_1} + \frac{1}{3} R_{D_W}.M_{A_1} \tag{3.24}$$

Tout comme dans l'approche concernant l'identification de modèles à entrées précises, ce critère doit être optimisé sous un ensemble de contraintes. Classiquement, ces dernières permettent de respecter l'inclusion des observations Y_j dans les prédictions $\hat{Y}_j$. Ces contraintes sont obtenues en introduisant l'expression (3.21) dans la relation d'inclusion de deux intervalles (1.37) pour l'ensemble des échantillons $j = 1, ..., M$.

Les contraintes d'optimisation doivent également garantir l'obtention de paramètres sous formes d'intervalles bien définis, il est donc nécessaire d'introduire les contraintes de positivité des Radius de chacun des paramètres.

De plus, l'hypothèse de travail retenue dans l'expression de la sortie soit également être respectée. Par conséquent, il est indispensable d'introduire une contrainte imposant le signe du paramètre A_1. Dans le cas présent, celui-ci doit par exemple être positif, cette contrainte se traduit donc par la relation :

$$M_{A_1} \geq R_{A_1} \geq 0 \tag{3.25}$$

On remarquera que dans le cas d'une généralisation à un modèle multi-entrées, le critère est défini en intégrant l'expression du Radius de la sortie du modèle sur les domaines de définition associés à chacune des entrées. Le critère sera donc additif, comme cela était le cas pour les entrées précises. Ainsi, chacun des Radius et Midpoint des paramètres est pondéré par les grandeurs caractéristiques du domaine de définition associé.

Enfin, dans le cas d'un modèle à entrées précises, son domaine de définition est obtenu en considérant un domaine Δ_{D_W} pour lequel $R_{D_W} = 0$. Dans ce cas, le critère est défini par :

$$J = R_{A_0} + M_{D_W}.R_{A_1} \tag{3.26}$$

Cette expression est équivalente au critère (2.54) dans le cas d'une entrée positive (cas particulier ici explicité). De même, en considérant des entrées à Radius nul, donc précises, dans la contrainte d'inclusion définie pour le cas des entrées imprécises, celle-ci est équivalente aux contraintes (2.14) définies préalablement pour un modèle à entrées précises. Par conséquent, l'approche développée ici pour des modèles à entrées imprécises est une généralisation de la technique proposée dans le chapitre II concernant l'identification de modèles à entrées précises.

Les concepts introduits précédemment sont maintenant appliqués à l'identification d'un modèle à entrées imprécises de la forme (3.14).

Par rapport à l'étude précédente concernant l'identification d'un modèle à entrées précises, les domaines de définition $D_i, i = 1, ..., N$ sont inchangés, ainsi que les valeurs de décalages appropriées. Par conséquent, les entrées décalées seront toutes négatives.

De plus, l'étude précédente fournit un modèle optimal dont les paramètres A_1, A_2 et A_4 sont positifs, tandis que le paramètre A_3 est négatif. Cela nous permet donc de disposer d'une connaissance a priori sur les signes des paramètres à imposer dans la technique de régression développée pour un modèle à entrées imprécises.

Par application de la méthode d'identification, les paramètres du modèle sont identifiés et présentés dans le tableau 3.21. Une représentation indicielle de la sortie du modèle est fournie sur la figure 3.16, sur laquelle les données observées sont les intervalles en pointillés, la sortie du modèle étant en traits pleins.

	Espace des bornes	Espace Midpoint / Radius
A_0	$[10435, 10736]$	$(10586, 150.5)$
A_1	0.755	0.755
A_2	0.219	0.219
A_3	$[-0.0436, -0.0344]$	$(-0.039, 0.0046)$
A_4	$[0, 0.0958]$	$(0.0479, 0.0479)$

TAB. 3.21: Paramètres du modèle à entrées imprécises identifié sur le cours du Dow Jones

Il est peu aisé à la lecture du tableau 3.21 ou en considérant la représentation indicielle 3.16 de quantifier l'amélioration de la qualité du modèle induite par la propagation des entrées imprécises. Certes, sur la figure, il est possible de constater une diminution de l'imprécision de la sortie prédite, sans pour autant qu'elle soit prépondérante.

Pour remédier à cela, deux indicateurs sont considérés pour chacun des deux modèles (entrées précises et entrées imprécises), la distance (au sens de Diamond) entre les intervalles prédits et les observations d'une part, et l'imprécision totale de leurs paramètres (somme des Radius) d'autre part. Ces indicateurs sont synthétisés dans le tableau 3.22

A partir de ces résultats, il est clair qu'en propageant l'imprécision des entrées dans l'évaluation de la sortie du modèle lors de l'identification de ses paramètres, ces derniers voient leur imprécision diminuer. Par conséquent, le modèle ainsi obtenu permet de représenter l'imprécision du "système" d'une part, mais également l'information imprécise des entrées, per-

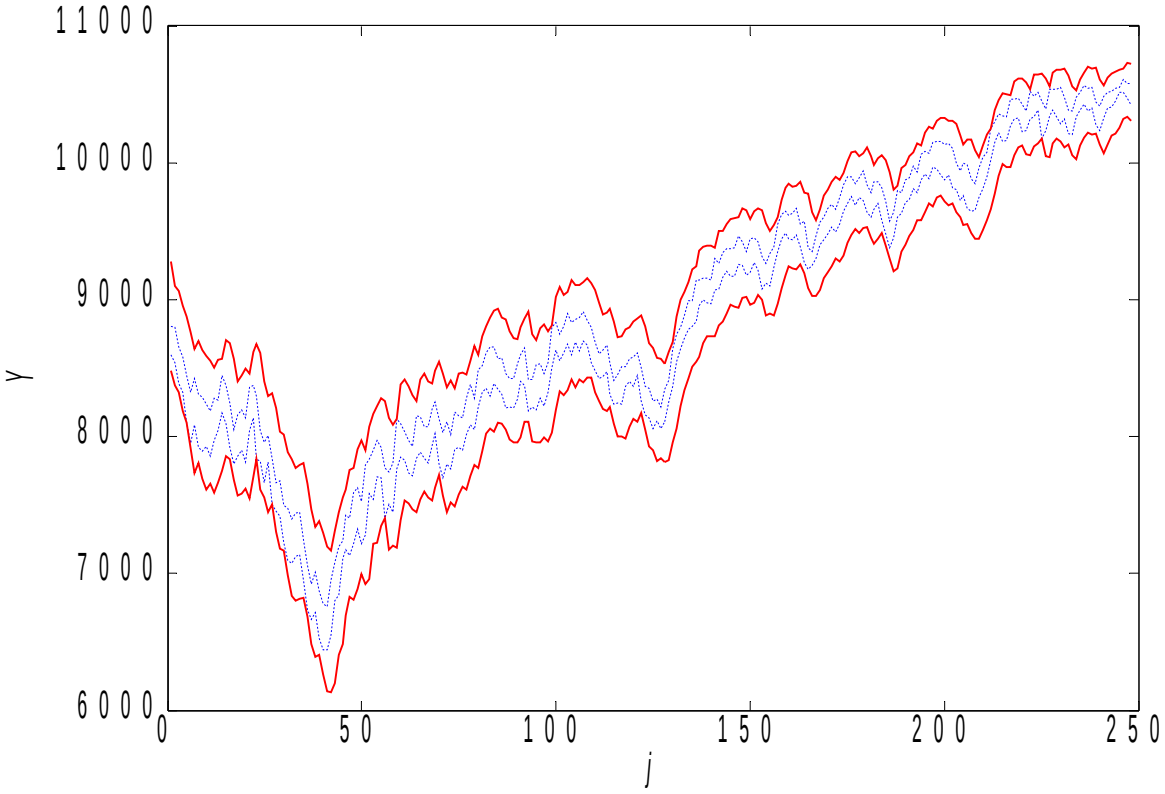


FIG. 3.16: Evolution de l'indice Dow Jones sur l'année 2009 - observations (pointillés) et prédictions (traits pleins) - modèle à entrées imprécises

	Entrées précises	Entrées imprécises
$\sum_{i=0}^{4} R_{A_i}$	233.0785	150.5525
$Distance$	154720	150840

TAB. 3.22: Indicateurs associés à chacun des modèles identifiés

mettant une amélioration de sa qualité. Cela se traduit également par une diminution de la distance entre les prédictions et les observations. On remarquera que cette diminution est relativement faible, en lien avec le fait que les imprécisions sont relativement faibles par rapport aux amplitudes globales.

Pour conclure, il est important de noter que si l'utilisation de modèles à entrées imprécises est validée de par l'amélioration (même minime) en terme de qualité de la représentation de l'évolution de l'indice Dow Jones sur l'année 2009 qu'ils engendrent, l'hypothèse de base fondamentale consistant à identifier un modèle linéaire n'en reste pas moins sujette à caution. Il

semble utopique de chercher à modéliser un système aussi complexe qu'un indice boursier en fonction de ses variations antérieures, sans tenir compte des perturbations extérieures ou des interventions des multiples agents humains intervenant dans le milieu de la finance. Certes, propager l'imprécision au sein d'un modèle dynamique semble avantageux, mais l'utilisation même d'un tel modèle paramétrique doit être vue comme une première approximation dont la finalité ne saurait être une performance importante en terme de représentation de l'évolution d'un indice boursier.

Conclusions et perspectives

Dans ce mémoire, nous nous sommes intéressés à l'identification de modèles régressifs dans un environnement imprécis. Dans ce contexte, l'imprécision est considérée comme une caractéristique intrinsèque du modèle, et elle est par conséquent introduite et manipulée tout au long du processus d'identification.

Le choix de modèles régressifs paramétriques linéaires se justifie par le contrôle qu'il est possible d'exercer sur la structure du modèle, sur le nombre de paramètres, et l'interprétation facilitée grâce à l'expression analytique simple obtenue.

Dans un environnement imprécis, cela conduit à distinguer différents types de modèles, selon que les observations d'entrée et de sorties sont imprécises ou non. Dans ce contexte, nous nous sommes particulièrement focalisés sur le cadre de modèles à paramètres imprécis, permettant ainsi d'expliciter les relations incertaines entre les sollicitations (entrées) du système supposées parfaitement déterminées, et les réactions (sorties) mesurées qui sont quant à elles imprécises.

Le formalisme des ensembles flous a été retenu en vue de modéliser les imprécisions à tous niveaux. En considérant le principe des α-coupes réalisées sur les intervalles flous, il est possible de formaliser dans un cadre unifié les modèles régressifs et les techniques d'identification associées dans un environnement imprécis.

Deux aspects principaux de ces dernières ont retenu notre attention : l'imprécision qu'elles induisent sur les modèles identifiés, et la relation entre les observations et les prédictions. L'objectif fixé est de pouvoir obtenir des modèles d'imprécision minimale, mais assurant néanmoins l'inclusion des observations dans les prédictions. Autrement dit, il faut pouvoir améliorer l'une sans nuire à l'autre.

Pour remplir cet objectif, la structure des modèles régressifs linéaires flous a été modifiée, en introduisant trois paramètres supplémentaires. Deux d'entre eux permettent de considérer des intervalles flous trapézoïdaux et non plus triangulaires, permettant ainsi d'assurer l'inclusion recherchée, sans pour autant augmenter de manière inappropriée l'imprécision du modèle ni

se départir de l'usage de l'arithmétique des intervalles dans le processus d'identification. La sortie d'un modèle trapézoïdal permet donc de représenter, dans un modèle unique, aussi bien la distribution des valeurs modales des observations, au niveau du noyau, que celle de leur imprécision, au niveau du support.

Le dernier paramètre, précis, et lié au domaine de définition du modèle, a un double objectif. L'introduction d'un décalage sur les entrées du modèle permet de positionner de manière rigoureuse son origine, limitant ainsi l'impact de l'amplitude des données observées sur les paramètres optimaux. Le décalage permet par ailleurs de disposer au final de modèles linéaires capables de représenter une variation quelconque de l'imprécision du système modélisé.

La technique d'identification de tels modèles a par ailleurs été revisitée. Une nouvelle définition de l'imprécision de la sortie du modèle est en effet introduite comme critère d'optimisation. L'imprécision est évaluée sur l'ensemble du domaine de définition du modèle, et non plus uniquement aux points d'identification. Cela permet de diminuer l'influence de données non conflictuelles ou redondantes au sein de l'ensemble des observations. Ce critère permet également de s'affranchir de la répartition des observations sur le domaine de définition du modèle. Ce dernier voit donc son imprécision être potentiellement diminuée et donc sa qualité augmentée.

Tous ces concepts permettent au final de disposer de modèles régressifs linéaires flous identifiés dont l'imprécision est optimisée, et dont la sortie englobe intégralement l'ensemble des observations. Dans un contexte d'identification de systèmes, ils permettent, en étant interprétables, de fournir une représentation pertinente d'un système imprécis, en intégrant les incertitudes dans la structure même du modèle.

Lorsque des jeux de données traduisant le comportement de systèmes complexes sont considérés, il est bénéfique de chercher à identifier des modèles plus évolués. Ils peuvent ainsi être linéaires par morceaux, polynomiaux, ou encore multilinéaires. S'il est possible d'adapter la méthode proposée pour identifier les paramètres d'un modèle régressif linéaire dédié à chacun de ces cas, il n'en reste pas moins que certains points problématiques subsistent.

Dans une approche par morceaux, le segmentation des données est une étape cruciale, dont le résultat impacte fortement la qualité de l'identification. Or, notamment dans le cas de données multi-entrées, ce processus est lourd et complexe. De plus, déterminer le nombre optimal de segments est également un enjeu important. Par conséquent, une attention particulière doit être portée sur ce processus.

Dans le cadre de la régression polynomiale ou multilinéaire, c'est le choix de la forme analytique adéquate du modèle à identifier qui est crucial. Là encore, ce choix a un fort impact sur la qualité du modèle obtenu. Dans une approche paramétrique, cette étape de décision est à la charge de l'utilisateur, bien que des méthodes existent pour l'assister. Néanmoins, leur coût en

terme de calculs reste élevé, et donc peu satisfaisant.

L'identification de modèles dynamiques doit également être envisagée. En environnement imprécis, survient alors la problématique liée à la propagation des imprécisions dans le rebouclage. Il est alors nécessaire de considérer et d'identifier des modèles à entrées imprécises. Une adaptation de notre méthode permet néanmoins de gérer ce cas de figure. Cependant, cette extension n'a été ici développée que dans l'optique de manipuler des intervalles conventionnels, et non flous. Or, dans la définition des modèles flous dynamiques, il est nécessaire de considérer le produit entre intervalles flous, ayant pour conséquence d'entraîner la perte de linéarité des fonctions d'appartenance. Ainsi, un des avantages essentiels de notre approche, permettant l'obtention de modèles flous en ne considérant que deux α-coupes, et donc en ne manipulant que des intervalles conventionnels, n'est plus valide. Ce point mérite donc également que l'on s'y attarde.

Pour finir, rappelons que deux problématiques ont été à l'origine de ces travaux : la représentation et la manipulation des imprécisions intervenant dans un contexte d'identification de modèles régressifs linéaires flous pertinents et interprétables. Cette étude a permis d'y répondre, au travers de la proposition d'une méthode permettant l'identification de modèles flous variés et adaptés à un grand nombre de systèmes à modéliser.

Annexe A

Le jeu de données bruitées

j	Type	x_1	x_2	x_3	x_4	x_5	y
1	Test	0.36031	0.28594	0.014864	0.8952	0.51015	1.3915
2	Identification	0.54851	0.39413	0.28819	0.94239	0.71396	4.6557
3	Identification	0.26177	0.50301	0.81673	0.33508	0.51521	6.2132
4	Point aberrant	0.59734	0.72198	0.98548	0.43736	0.60587	**10.5848**
5	Test	0.049278	0.30621	0.017363	0.47116	0.9667	1.1886
6	Identification	0.57106	0.11216	0.81939	0.14931	0.82212	5.658
7	Identification	0.70086	0.44329	0.62114	0.13586	0.31775	6.4823
8	Identification	0.96229	0.46676	0.56022	0.5325	0.5877	8.2347
9	Identification	0.75052	0.014669	0.24403	0.72579	0.1302	1.7014
10	Identification	0.73999	0.66405	0.82201	0.3987	0.25435	9.7441
11	Point aberrant	0.43187	0.72406	0.26321	0.35842	0.80303	**5.9823**
12	Identification	0.63427	0.28163	0.75363	0.28528	0.66785	6.4305
13	Identification	0.80303	0.26182	0.65964	0.86864	0.013626	6.5469
14	Identification	0.083881	0.70847	0.21406	0.62641	0.56158	2.2481
15	Identification	0.94546	0.78386	0.60212	0.24117	0.45456	10.919
16	Identification	0.91594	0.98616	0.60494	0.97808	0.90495	14.06
17	Identification	0.60199	0.47334	0.6595	0.6405	0.28216	7.0714
18	Identification	0.25356	0.90282	0.18336	0.22985	0.065034	3.2945
19	Identification	0.87345	0.45106	0.63655	0.68134	0.47659	8.2171
20	Identification	0.5134	0.80452	0.17031	0.66582	0.98371	6.1764

j	Type	x_1	x_2	x_3	x_4	x_5	y
21	Identification	0.73265	0.82886	0.5396	0.13472	0.92235	9.7668
22	Identification	0.42223	0.16627	0.62339	0.022493	0.5612	4.162
23	Point aberrant	0.96137	0.39391	0.68589	0.2622	0.65232	**9.6279**
24	Identification	0.072059	0.52076	0.67735	0.11652	0.77268	4.5169
25	Identification	0.55341	0.71812	0.87683	0.069318	0.10618	8.4911
26	Test	0.29198	0.56919	0.012891	0.85293	0.0010734	1.7484
27	Identification	0.85796	0.46081	0.3104	0.18033	0.54176	5.911
28	Identification	0.33576	0.44531	0.77908	0.032419	0.0068578	5.4411
29	Identification	0.6802	0.087745	0.3073	0.73393	0.45134	2.7881
30	Identification	0.053444	0.44348	0.92668	0.53652	0.19566	5.9031
31	Identification	0.35666	0.3663	0.67872	0.27603	0.78714	5.6943
32	Test	0.4983	0.30253	0.074321	0.36846	0.61856	2.3165
33	Test	0.43444	0.85184	0.070669	0.012886	0.015521	4.0561
34	Test	0.56246	0.75948	0.01193	0.88921	0.89085	5.1463
35	Identification	0.61662	0.94976	0.22715	0.86602	0.7617	7.9658
36	Point aberrant	0.11334	0.55794	0.51625	0.25425	0.90704	**4.3186**
37	Identification	0.89825	0.014233	0.4582	0.56948	0.75857	3.5161
38	Identification	0.75455	0.59618	0.7032	0.15926	0.38073	8.3834
39	Identification	0.79112	0.81621	0.58248	0.59436	0.33111	10.172
40	Identification	0.81495	0.97709	0.50921	0.3311	0.50408	11.1
41	Test	0.67	0.22191	0.07429	0.65861	0.56457	2.2748
42	Identification	0.20088	0.70368	0.19324	0.86363	0.7672	3.3021
43	Identification	0.27309	0.52206	0.3796	0.56762	0.77987	4.3628
44	Identification	0.62623	0.9329	0.27643	0.98048	0.4841	8.0007
45	Identification	0.53685	0.71335	0.77088	0.79183	0.80221	9.5484
46	Identification	0.059504	0.22804	0.31393	0.15259	0.47101	2.023
47	Point aberrant	0.088962	0.44964	0.63819	0.83303	0.20276	**4.9731**
48	Identification	0.27131	0.1722	0.98657	0.19186	0.57961	6.1146
49	Identification	0.40907	0.96882	0.50288	0.63899	0.6665	7.5644
50	Identification	0.47404	0.35572	0.9477	0.669	0.67677	8.1508

j	Type	x_1	x_2	x_3	x_4	x_5	y
51	Identification	0.90899	0.049047	0.82803	0.77209	0.94251	6.7529
52	Identification	0.59625	0.75534	0.91756	0.37982	0.77015	10.382
53	Identification	0.32896	0.89481	0.11308	0.44159	0.7374	4.1526
54	Identification	0.47819	0.28615	0.81213	0.48306	0.86626	6.964
55	Identification	0.59717	0.2512	0.90826	0.60811	0.99095	8.128
56	Point aberrant	0.16145	0.93274	0.15638	0.176	0.50393	**3.0023**
57	Identification	0.82947	0.13098	0.12212	0.002026	0.62909	2.0933
58	Point aberrant	0.95612	0.94082	0.76267	0.79022	0.79261	**15.0394**
59	Identification	0.59555	0.70185	0.7218	0.51361	0.44865	8.7316
60	Identification	0.028748	0.84768	0.65164	0.21323	0.52436	4.0547
61	Identification	0.81212	0.20927	0.75402	0.10345	0.17147	5.655
62	Identification	0.61011	0.45509	0.66316	0.15734	0.13067	6.3181
63	Identification	0.70149	0.081074	0.88349	0.40751	0.21878	5.7541
64	Point aberrant	0.092196	0.85112	0.27216	0.40776	0.10548	**3.5862**
65	Point aberrant	0.42489	0.56205	0.41943	0.052693	0.14143	**5.0938**
66	Identification	0.37558	0.3193	0.21299	0.94182	0.45697	2.8742
67	Test	0.16615	0.3749	0.0356	0.14997	0.78813	1.4327
68	Identification	0.83315	0.8678	0.081164	0.38437	0.28106	7.7773
69	Identification	0.83864	0.37218	0.85057	0.31106	0.22479	7.9538
70	Identification	0.45161	0.07369	0.3402	0.16853	0.90887	2.9745
71	Identification	0.9566	0.19984	0.46615	0.89665	0.007329	5.0784
72	Identification	0.14715	0.049493	0.91376	0.32272	0.58874	5.578
73	Identification	0.86993	0.56671	0.22858	0.734	0.54212	6.7023
74	Identification	0.76944	0.12192	0.86204	0.4109	0.65352	6.3838
75	Identification	0.44416	0.52211	0.65662	0.39979	0.31343	6.2254
76	Identification	0.62062	0.11706	0.89118	0.50552	0.23116	6.1369
77	Identification	0.95169	0.76992	0.48814	0.16931	0.41606	10.106
78	Identification	0.64001	0.37506	0.99265	0.52475	0.2988	8.4947
79	Identification	0.24733	0.82339	0.37333	0.6412	0.67244	4.8341
80	Identification	0.3527	0.046636	0.53138	0.016197	0.93826	3.7189
81	Identification	0.18786	0.59791	0.18132	0.83685	0.34315	2.4511

j	Type	x_1	x_2	x_3	x_4	x_5	y
82	Identification	0.49064	0.94915	0.50194	0.80346	0.56296	8.2901
83	Point aberrant	0.40927	0.2888	0.42219	0.69778	0.11889	**4.2938**
84	Identification	0.46353	0.88883	0.66043	0.46189	0.16902	8.0608
85	Identification	0.61094	0.10159	0.67365	0.082613	0.2789	4.178
86	Identification	0.071168	0.065315	0.95733	0.82072	0.55681	6.7146
87	Identification	0.31428	0.2343	0.19187	0.19302	0.48559	2.0056
88	Identification	0.60838	0.9331	0.11122	0.44535	0.95222	7.2387
89	Identification	0.17502	0.063128	0.56505	0.012958	0.23192	3.0042
90	Identification	0.62103	0.26422	0.96917	0.30874	0.47866	7.3143
91	Test	0.24596	0.99953	0.023744	0.87535	0.52652	2.896
92	Identification	0.58736	0.21199	0.87022	0.83526	0.79272	7.6784
93	Test	0.50605	0.49841	0.026877	0.3331	0.19301	2.7117
94	Identification	0.46478	0.29049	0.51953	0.88071	0.9096	5.6903
95	Identification	0.54142	0.67275	0.19229	0.47969	0.9222	5.6388
96	Identification	0.94233	0.95799	0.71569	0.56082	0.013266	13.409
97	Identification	0.34176	0.76655	0.25067	0.61591	0.76755	4.771
98	Identification	0.4018	0.66612	0.93386	0.6619	0.94734	9.4795
99	Identification	0.30769	0.13094	0.13719	0.61663	0.81331	1.9195
100	Identification	0.41157	0.095413	0.52162	0.68514	0.92383	4.569

Annexe B

Le cours du Dow Jones sur l'année 2009

indice	*Date*	*Open*	*High*	*Low*	*Close*
1	02/01/09	8772.25	9080.57	8725.10	9034.69
2	05/01/09	9027.13	9093.47	8841.70	8952.89
3	06/01/09	8954.57	9175.19	8868.07	9015.10
4	07/01/09	8996.94	8996.94	8690.45	8769.70
5	08/01/09	8769.94	8807.14	8593.52	8742.46
6	09/01/09	8738.80	8800.45	8541.75	8599.18
7	12/01/09	8599.26	8653.97	8391.85	8473.97
8	13/01/09	8474.61	8584.68	8325.59	8448.56
9	14/01/09	8446.01	8446.01	8097.95	8200.14
10	15/01/09	8196.24	8326.06	7949.65	8212.49
11	16/01/09	8215.67	8424.59	8086.01	8281.22
12	20/01/09	8279.63	8309.02	7920.66	7949.09
13	21/01/09	7949.17	8286.40	7890.63	8228.10
14	22/01/09	8224.43	8239.33	7925.75	8122.80
15	23/01/09	8108.79	8187.88	7856.86	8077.56
16	26/01/09	8078.04	8278.12	7971.15	8116.03
17	27/01/09	8117.39	8264.10	8042.60	8174.73
18	28/01/09	8175.93	8446.33	8175.93	8375.45
19	29/01/09	8373.06	8373.06	8092.14	8149.01
20	30/01/09	8149.01	8243.95	7924.88	8000.86

indice	*Date*	*Open*	*High*	*Low*	*Close*
21	02/02/09	8000.62	8053.43	7796.17	7936.83
22	03/02/09	7936.99	8157.13	7855.19	8078.36
23	04/02/09	8070.32	8197.04	7899.79	7956.66
24	05/02/09	7954.83	8138.65	7811.70	8063.07
25	06/02/09	8056.38	8360.07	8044.03	8280.59
26	09/02/09	8281.38	8376.56	8137.70	8270.87
27	10/02/09	8269.36	8293.17	7835.83	7888.88
28	11/02/09	7887.05	8042.36	7820.14	7939.53
29	12/02/09	7931.97	7956.02	7662.04	7932.76
30	13/02/09	7933.00	8005.96	7811.38	7850.41
31	17/02/09	7845.63	7845.63	7502.59	7552.60
32	18/02/09	7546.35	7661.56	7451.37	7555.63
33	19/02/09	7555.23	7679.01	7420.63	7465.95
34	20/02/09	7461.49	7500.44	7226.29	7365.67
35	23/02/09	7365.99	7477.10	7092.64	7114.78
36	24/02/09	7115.34	7396.34	7077.35	7350.94
37	25/02/09	7349.58	7442.13	7123.94	7270.89
38	26/02/09	7269.06	7451.13	7135.25	7182.08
39	27/02/09	7180.97	7244.61	6952.06	7062.93
40	02/03/09	7056.48	7056.48	6736.69	6763.29
41	03/03/09	6764.81	6922.59	6661.74	6726.02
42	04/03/09	6726.50	7012.19	6715.11	6875.84
43	05/03/09	6874.01	6874.01	6531.28	6594.44
44	06/03/09	6595.16	6776.44	6443.27	6626.94
45	09/03/09	6625.74	6758.44	6440.08	6547.05
46	10/03/09	6547.01	6951.50	6547.01	6926.49
47	11/03/09	6923.13	7078.22	6804.55	6930.40
48	12/03/09	6932.39	7198.25	6840.79	7170.06
49	13/03/09	7219.20	7241.98	7106.34	7223.98
50	16/03/09	7225.33	7428.75	7171.41	7216.97

indice	*Date*	*Open*	*High*	*Low*	*Close*
51	17/03/09	7218.00	7407.41	7129.60	7395.70
52	18/03/09	7395.70	7592.03	7218.24	7486.58
53	19/03/09	7489.68	7624.45	7325.13	7400.80
54	20/03/09	7402.31	7524.81	7215.77	7278.38
55	23/03/09	7279.25	7789.24	7279.25	7775.86
56	24/03/09	7773.47	7837.11	7585.98	7660.21
57	25/03/09	7659.81	7897.48	7539.54	7749.81
58	26/03/09	7752.36	7969.00	7709.19	7924.56
59	27/03/09	7922.57	7922.57	7695.97	7776.18
60	30/03/09	7773.31	7773.31	7406.85	7522.02
61	31/03/09	7523.77	7744.24	7502.98	7608.92
62	01/04/09	7606.13	7804.77	7450.74	7761.60
63	02/04/09	7763.99	8129.33	7763.99	7978.08
64	03/04/09	7980.63	8090.71	7850.33	8017.59
65	06/04/09	8016.16	8037.42	7830.66	7975.85
66	07/04/09	7968.92	7968.92	7733.56	7789.56
67	08/04/09	7788.68	7925.36	7715.09	7837.11
68	09/04/09	7839.89	8150.44	7839.89	8083.38
69	13/04/09	8082.02	8146.86	7888.96	8057.81
70	14/04/09	8057.41	8076.05	7840.53	7920.18
71	15/04/09	7914.92	8069.92	7808.19	8029.62
72	16/04/09	8029.14	8201.81	7933.08	8125.43
73	17/04/09	8125.43	8251.20	8024.92	8131.33
74	20/04/09	8128.94	8128.94	7801.58	7841.73
75	21/04/09	7841.73	8027.54	7699.79	7969.56
76	22/04/09	7964.78	8111.02	7802.46	7886.57
77	23/04/09	7886.81	8015.36	7762.80	7957.06
78	24/04/09	7957.45	8182.30	7905.60	8076.29
79	27/04/09	8073.82	8152.27	7920.42	8025.00
80	28/04/09	8023.56	8136.74	7898.75	8016.95

indice	*Date*	*Open*	*High*	*Low*	*Close*
81	29/04/09	8018.31	8278.12	8018.31	8185.73
82	30/04/09	8188.51	8383.81	8083.62	8168.12
83	01/05/09	8167.41	8278.28	8047.54	8212.41
84	04/05/09	8213.60	8488.87	8213.60	8426.74
85	05/05/09	8425.55	8520.80	8321.37	8410.65
86	06/05/09	8403.48	8608.26	8350.12	8512.28
87	07/05/09	8513.56	8651.51	8296.04	8409.85
88	08/05/09	8410.73	8657.96	8388.11	8574.65
89	11/05/09	8569.23	8569.23	8347.41	8418.77
90	12/05/09	8419.17	8574.88	8306.47	8469.11
91	13/05/09	8461.80	8461.80	8208.74	8284.89
92	14/05/09	8285.92	8427.93	8218.94	8331.32
93	15/05/09	8326.22	8422.28	8206.67	8268.64
94	18/05/09	8270.15	8534.66	8270.15	8504.08
95	19/05/09	8502.48	8594.16	8402.61	8474.85
96	20/05/09	8471.82	8645.85	8376.40	8422.04
97	21/05/09	8416.07	8416.07	8185.25	8292.13
98	22/05/09	8292.21	8415.75	8218.86	8277.32
99	26/05/09	8275.33	8523.59	8194.33	8473.49
100	27/05/09	8473.65	8534.66	8280.82	8300.02
101	28/05/09	8300.50	8463.70	8221.65	8403.80
102	29/05/09	8404.04	8541.27	8323.91	8500.33
103	01/06/09	8501.53	8797.58	8501.53	8721.44
104	02/06/09	8721.60	8832.16	8635.25	8740.87
105	03/06/09	8740.07	8750.83	8556.90	8675.24
106	04/06/09	8665.72	8802.59	8609.17	8750.24
107	05/06/09	8751.75	8900.48	8673.41	8763.13
108	08/06/09	8759.35	8832.13	8593.84	8764.49
109	09/06/09	8764.83	8854.80	8688.99	8763.06
110	10/06/09	8763.66	8871.36	8625.21	8739.02

indice	*Date*	*Open*	*High*	*Low*	*Close*
111	11/06/09	8736.23	8911.11	8697.99	8770.92
112	12/06/09	8770.01	8850.95	8671.61	8799.26
113	15/06/09	8798.50	8798.50	8540.87	8612.13
114	16/06/09	8612.44	8688.69	8483.58	8504.67
115	17/06/09	8504.36	8602.99	8421.46	8497.18
116	18/06/09	8496.73	8634.28	8438.61	8555.60
117	19/06/09	8556.96	8665.26	8476.02	8539.73
118	22/06/09	8538.52	8538.52	8306.66	8339.01
119	23/06/09	8340.44	8413.22	8239.17	8322.91
120	24/06/09	8323.51	8456.83	8246.20	8299.86
121	25/06/09	8299.25	8512.60	8236.07	8472.40
122	26/06/09	8468.54	8509.73	8364.17	8438.39
123	29/06/09	8440.13	8569.59	8406.57	8529.38
124	30/06/09	8528.93	8584.17	8369.99	8447.00
125	01/07/09	8447.53	8610.32	8447.00	8504.06
126	02/07/09	8503.00	8503.00	8260.41	8280.74
127	06/07/09	8279.30	8364.02	8156.49	8324.87
128	07/07/09	8324.95	8355.48	8138.51	8163.60
129	08/07/09	8157.02	8259.05	8057.94	8178.41
130	09/07/09	8179.01	8273.48	8117.27	8183.17
131	10/07/09	8182.49	8216.65	8057.57	8146.52
132	13/07/09	8146.82	8348.08	8106.16	8331.68
133	14/07/09	8331.37	8407.48	8255.27	8359.49
134	15/07/09	8363.95	8643.04	8363.95	8616.21
135	16/07/09	8612.66	8750.28	8543.97	8711.82
136	17/07/09	8711.89	8797.97	8638.81	8743.94
137	20/07/09	8746.05	8884.43	8717.26	8848.15
138	21/07/09	8848.15	8991.07	8780.82	8915.94
139	22/07/09	8912.39	8993.48	8802.13	8881.26
140	23/07/09	8882.31	9143.05	8837.95	9069.29

indice	*Date*	*Open*	*High*	*Low*	*Close*
141	24/07/09	9066.11	9144.48	8955.77	9093.24
142	27/07/09	9093.09	9154.23	8996.58	9108.51
143	28/07/09	9106.92	9154.76	8980.03	9096.72
144	29/07/09	9092.34	9141.23	8967.26	9070.72
145	30/07/09	9072.84	9298.13	9072.84	9154.46
146	31/07/09	9154.61	9264.65	9081.30	9171.61
147	03/08/09	9173.65	9342.11	9162.09	9286.56
148	04/08/09	9285.05	9370.30	9207.21	9320.19
149	05/08/09	9315.36	9374.38	9173.20	9280.97
150	06/08/09	9277.19	9378.01	9168.44	9256.26
151	07/08/09	9258.45	9466.89	9258.45	9370.07
152	10/08/09	9368.41	9420.56	9249.99	9337.95
153	11/08/09	9334.33	9351.86	9180.23	9241.45
154	12/08/09	9236.06	9442.47	9199.80	9361.61
155	13/08/09	9362.29	9448.97	9269.26	9398.19
156	14/08/09	9398.04	9425.17	9214.47	9321.40
157	17/08/09	9313.85	9313.85	9078.28	9135.34
158	18/08/09	9134.36	9262.08	9124.08	9217.94
159	19/08/09	9208.68	9333.34	9099.14	9279.16
160	20/08/09	9278.55	9385.72	9237.52	9350.05
161	21/08/09	9347.86	9549.19	9347.86	9505.96
162	24/08/09	9506.18	9625.89	9442.17	9509.28
163	25/08/09	9509.21	9646.53	9485.70	9539.29
164	26/08/09	9538.61	9613.65	9446.71	9543.52
165	27/08/09	9541.63	9629.98	9440.43	9580.63
166	28/08/09	9582.74	9666.71	9476.63	9544.20
167	31/08/09	9542.91	9552.97	9389.27	9496.28
168	01/09/09	9492.32	9573.67	9275.15	9310.60
169	02/09/09	9306.21	9378.77	9223.08	9280.67
170	03/09/09	9282.03	9350.27	9252.93	9344.61

indice	*Date*	*Open*	*High*	*Low*	*Close*
171	04/09/09	9345.36	9465.37	9302.28	9441.27
172	08/09/09	9440.13	9564.45	9402.80	9497.34
173	09/09/09	9496.59	9604.43	9435.45	9547.22
174	10/09/09	9546.54	9666.55	9479.20	9627.48
175	11/09/09	9625.44	9698.67	9532.11	9605.41
176	14/09/09	9598.08	9662.10	9492.96	9626.80
177	15/09/09	9626.42	9745.91	9553.80	9683.41
178	16/09/09	9683.71	9837.05	9648.95	9791.71
179	17/09/09	9789.82	9896.38	9706.23	9783.92
180	18/09/09	9784.75	9898.57	9751.27	9820.20
181	21/09/09	9818.61	9846.12	9688.40	9778.86
182	22/09/09	9779.61	9890.71	9742.96	9829.87
183	23/09/09	9830.63	9937.72	9724.90	9748.55
184	24/09/09	9749.99	9836.82	9637.53	9707.44
185	25/09/09	9706.68	9781.73	9605.19	9665.19
186	28/09/09	9663.23	9861.39	9658.09	9789.36
187	29/09/09	9789.74	9861.99	9705.10	9742.20
188	30/09/09	9741.83	9817.17	9583.04	9712.28
189	01/10/09	9711.60	9714.70	9482.98	9509.28
190	02/10/09	9507.62	9571.71	9378.77	9487.67
191	05/10/09	9488.73	9640.33	9449.81	9599.75
192	06/10/09	9601.26	9793.37	9601.26	9731.25
193	07/10/09	9725.69	9782.56	9634.96	9725.58
194	08/10/09	9728.22	9872.50	9709.78	9786.87
195	09/10/09	9786.04	9890.41	9731.32	9864.94
196	12/10/09	9865.24	9978.07	9814.45	9885.80
197	13/10/09	9883.98	9935.53	9780.90	9871.06
198	14/10/09	9873.55	10064.98	9873.55	10015.86
199	15/10/09	10014.88	10087.43	9916.93	10062.94
200	16/10/09	10061.36	10072.62	9884.51	9995.91

indice	*Date*	*Open*	*High*	*Low*	*Close*
201	19/10/09	9996.67	10146.61	9967.49	10092.19
202	20/10/09	10092.42	10157.26	9952.98	10041.48
203	21/10/09	10038.84	10157.94	9909.83	9949.36
204	22/10/09	9946.18	10133.08	9879.07	10081.31
205	23/10/09	10099.90	10138.59	9908.70	9972.18
206	26/10/09	9972.33	10107.99	9817.55	9867.96
207	27/10/09	9868.34	9994.55	9802.36	9882.17
208	28/10/09	9881.11	9940.89	9723.31	9762.69
209	29/10/09	9762.91	9996.67	9762.91	9962.58
210	30/10/09	9961.52	9980.19	9664.89	9712.73
211	02/11/09	9712.13	9883.68	9647.06	9789.44
212	03/11/09	9787.47	9844.84	9649.78	9771.91
213	04/11/09	9767.30	9962.35	9745.76	9802.14
214	05/11/09	9807.80	10043.75	9807.80	10005.96
215	06/11/09	10001.35	10077.08	9898.49	10023.42
216	09/11/09	10020.62	10248.93	10020.62	10226.94
217	10/11/09	10223.01	10300.33	10148.12	10246.97
218	11/11/09	10247.42	10357.38	10217.19	10291.26
219	12/11/09	10289.82	10341.21	10157.64	10197.47
220	13/11/09	10197.85	10332.29	10162.93	10270.47
221	16/11/09	10267.53	10465.83	10267.53	10406.96
222	17/11/09	10404.77	10465.76	10318.69	10437.42
223	18/11/09	10426.27	10471.28	10330.33	10426.31
224	19/11/09	10425.33	10425.33	10226.41	10332.44
225	20/11/09	10327.91	10377.41	10237.60	10318.16
226	23/11/09	10320.13	10524.40	10320.13	10450.95
227	24/11/09	10451.25	10488.66	10335.62	10433.71
228	25/11/09	10432.96	10513.60	10385.65	10464.40
229	27/11/09	10452.23	10452.23	10179.33	10309.92
230	30/11/09	10309.77	10394.34	10238.05	10344.84

indice	*Date*	*Open*	*High*	*Low*	*Close*
231	01/12/09	10343.82	10537.03	10343.82	10471.58
232	02/12/09	10470.44	10537.63	10386.03	10452.68
233	03/12/09	10455.63	10533.55	10338.49	10366.15
234	04/12/09	10368.57	10549.04	10285.44	10388.90
235	07/12/09	10386.86	10478.23	10321.11	10390.11
236	08/12/09	10385.42	10385.42	10216.44	10285.97
237	09/12/09	10282.85	10377.11	10207.29	10337.05
238	10/12/09	10336.00	10479.06	10332.14	10405.83
239	11/12/09	10403.41	10516.47	10385.42	10471.50
240	14/12/09	10471.28	10566.88	10431.60	10501.05
241	15/12/09	10499.31	10542.09	10380.96	10452.00
242	16/12/09	10449.81	10552.75	10401.90	10441.12
243	17/12/09	10439.99	10439.99	10279.39	10308.26
244	18/12/09	10309.39	10412.55	10237.75	10328.89
245	21/12/09	10330.10	10489.41	10330.10	10414.14
246	22/12/09	10414.67	10511.56	10399.33	10464.93
247	23/12/09	10464.32	10520.93	10409.00	10466.44
248	24/12/09	10467.12	10541.26	10450.95	10520.10
249	28/12/09	10517.91	10550.78	10517.91	10547.07
250	29/12/09	10547.83	10605.65	10518.59	10545.41
251	30/12/09	10544.36	10583.28	10470.75	10548.51
252	31/12/09	10548.51	10578.74	10420.56	10428.05

Annexe C

Calcul du critère d'identification d'un modèle à entrées imprécises

Dans le cas d'un modèle dont le paramètre A_1 et l'entrée W (définie dans le domaine Δ_{D_W}) sont des intervalles positifs, le critère à optimiser est donné par :

$$J = \int\int_{\Delta_{D_W}} R_{\hat{Y}}(M_W, R_W) dM_W dR_W \tag{C.1}$$

avec :

$$R_{\hat{Y}}(M_W, R_W) = R_{A_0} + M_{A_1} R_W + R_{A_1} M_W \tag{C.2}$$

En considérant les définitions (3.18) et (3.19) du domaine Δ_{D_W}, on obtient :

$$J = \int_0^{R_{D_W}} \int_{M_1(R_W)}^{M_2(R_W)} R_{\hat{Y}}(M_W, R_W) dM_W dR_W \tag{C.3}$$

Afin d'alléger les notations, la référence à l'entrée W ne sera plus apparente dans la suite, c'est-à-dire que l'on notera $R_{D_W} = R_D$, $M_W = M$ et $R_W = R$.

Il est possible d'écrire :

$$J = \int_0^{R_D} I(R_{\hat{Y}}, M) dR \tag{C.4}$$

en posant :

$$I(R_{\hat{Y}}, M) = \int_{\Delta_D} R_{\hat{Y}}(M, R) dM \tag{C.5}$$

En introduisant l'expression du Radius (3.22) dans cette intégrale (C.5), il vient :

$$I(R_{\hat{Y}}, M) = [(R_{A_0} + M_{A_1}.R).M + \frac{1}{2} R_{A_1}.M^2]_{M=M_1(R)}^{M=M_2(R)} \tag{C.6}$$

Sachant que l'on a, en considérant l'expression (3.19) :

$$\begin{cases} M_2(R) - M_1(R) = 2.R_D - 2.R \\ M_2(R)^2 - M_1(R)^2 = -4.M_D.(R - R_D) \end{cases} \tag{C.7}$$

il en découle :

$$I(R_{\hat{Y}}, M) = (R_{A_0} + M_{A_1}.R).(2R_D - 2R) + \frac{1}{2}R_{A_1}.(-4M_D.(R - R_D)) \tag{C.8}$$

En regroupant les termes en fonction de leur ordre selon la variable R, il est possible d'écrire :

$$I(R_{\hat{Y}}, M) = (2R_D R_{A_0} + 2M_D R_D R_{A_1}) + (-2R_{A_0} + 2R_D M_{A_1} - 2M_D R_{A_1}).R - 2M_{A_1}.R^2 \tag{C.9}$$

En introduisant cette expression (C.9) dans l'intégrale (C.4), on obtient :

$$J = [(2R_D R_{A_0} + 2M_D R_D R_{A_1}).R + \frac{1}{2}(-2R_{A_0} + 2R_D M_{A_1} - 2M_D R_{A_1}).R^2 - \frac{2}{3}M_{A_1}.R^3]_{R=0}^{R=R_D} \tag{C.10}$$

Soit, en simplifiant par la grandeur R_D^2 :

$$J = 2R_{A_0} + 2M_D R_{A_1} - R_{A_0} + R_D M_{A_1} - M_D R_{A_1} - \frac{2}{3}R_D M_{A_1} \tag{C.11}$$

De cette expression, on obtient de manière immédiate l'expression du critère (3.24), en réintroduisant les notations complètes.

Bibliographie

[1] H. Akaike. A look at the statistical model identification. *IEEE Transactions on Automatic Control AC*, 19(6) :716–723, 1974.

[2] V. Barnett and T. Lewis. *Outliers in statistical data*. Wiley New York, 1994.

[3] A Bisserier, R Boukezzoula, and S Galichet. Linear fuzzy regression using trapezoidal fuzzy intervals. In *IPMU Conference*, pages 181–188, July 2008.

[4] A Bisserier, R Boukezzoula, and S Galichet. An interval approach for fuzzy linear regression with imprecise data. In *IFSA - EUSFLAT Conference*, pages 1305–1310, July 2009.

[5] A Bisserier, R Boukezzoula, and S Galichet. Représentation, identification et propagation des incertitudes dans un contexte de régression linéaire. In *Rencontres francophones sur la Logique Floue et ses Applications (LFA 2009)*, pages 51–58, Novembre 2009.

[6] A. Bisserier, R. Boukezzoula, and S. Galichet. Linear fuzzy regression using trapezoidal fuzzy intervals. *Foundations of Reasoning under Uncertainty*, pages 1–22, 2010.

[7] A Bisserier, S Galichet, and R Boukezzoula. Une vision de la régression linéaire floue au travers de l'arithmétique des intervalles. In *Rencontres francophones sur la Logique Floue et ses Applications (LFA 2007)*, pages 57–64, Novembre 2007.

[8] A. Bisserier, S. Galichet, and R. Boukezzoula. Fuzzy piecewise linear regression. In *IEEE International Conference on Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence)*, pages 2089–2094, 2008.

[9] A Bisserier, F. Megri, S Galichet, and R Boukezzoula. Etude expérimentale de la robustesse des techniques linéaire et quadratique de régression floue. In *Rencontres francophones sur la Logique Floue et ses Applications (LFA 2008)*, pages 86–93, Octobre 2008.

[10] R. Boukezzoula, S. Galichet, and L. Foulloy. MIN and MAX Operators for Fuzzy Intervals and Their Potential Use in Aggregation Operators. *IEEE Transactions on Fuzzy Systems*, 15(6) :1135–1144, 2007.

[11] KY Chan, CK Kwong, and TC Fogarty. Modeling manufacturing processes using a genetic programming-based fuzzy regression with detection of outliers. *Information Sciences*, 180(4) :506–518, 2010.

[12] S. Chatterjee and A.S. Hadi. *Regression analysis by example*. Wiley-Interscience, 2006.

[13] L.H. Chen and C.C. Hsueh. A Mathematical Programming Method for Formulating a Fuzzy Regression Model Based on Distance Criterion. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(3) :705–712, 2007.

[14] S.P. Chen and J.F. Dang. A variable spread fuzzy linear regression model with higher explanatory power and forecasting accuracy. *Information Sciences*, 178(20) :3973–3988, 2008.

[15] C.C. Chuang. Fuzzy weighted support vector regression with a fuzzy partition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(3) :630–640, 2007.

[16] C.C. Chuang. Extended support vector interval regression networks for interval input-output data. *Information Sciences*, 178(3) :871–891, 2008.

[17] P. Diamond. Fuzzy least squares. *Information Sciences : an International Journal*, 46(3) :141–157, 1988.

[18] P. Diamond and H. Tanaka. Fuzzy regression analysis. In *Fuzzy sets in decision analysis, operations research and statistics*, page 387. Kluwer Academic Publishers, 1999.

[19] P. D'Urso. Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. *Computational Statistics and Data Analysis*, 42(1-2) :47–72, 2003.

[20] P. D'Urso and T. Gastaldi. A least-squares approach to fuzzy linear regression analysis. *Computational Statistics and Data Analysis*, 34(4) :427–440, 2000.

[21] P. D'Urso and T. Gastaldi. An orderwise polynomial regression procedure for fuzzy data. *Fuzzy Sets and Systems*, 130(1) :19, 2002.

[22] R.L. Eubank. *Spline smoothing and nonparametric regression*. M. Dekker New York, 1988.

[23] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2) :205–217, 2003.

[24] J. Fox. *Applied regression analysis, linear models, and related methods*. Sage Publications, 1997.

[25] F.E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1) :1–21, 1969.

[26] R.F. Gunst and R.L. Mason. *Regression analysis and its application : a data-oriented approach*. CRC Press, 1980.

[27] P.Y. Hao and J.H. Chiang. Fuzzy Regression Analysis by Support Vector Learning Approach. *IEEE Transactions on Fuzzy Systems*, 16(2) :428–441, 2008.

[28] B. Heshmaty and A. Kandel. Fuzzy linear regression and its applications to forecasting in uncertain environment. *Fuzzy Sets and Systems*, 15(2) :159–191, 1985.

[29] RR Hocking. A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, 32(1) :1–49, 1976.

[30] R.R. Hocking. *Methods and applications of linear models : Regression and the analysis of variance*. Wiley-Interscience, 2005.

[31] M. Hojati, CR Bector, and K. Smimou. A simple method for computation of fuzzy linear regression. *European Journal of Operational Research*, 166(1) :172–184, 2005.

[32] D.H. Hong and C. Hwang. Interval regression analysis using quadratic loss support vector machine. *IEEE Transactions on Fuzzy Systems*, 13(2) :229–237, 2005.

[33] W.L. Hung and M.S. Yang. An omission approach for detecting outliers in fuzzy regression models. *Fuzzy Sets and Systems*, 157(23) :3109–3122, 2006.

[34] C. Hwang, D.H. Hong, and K. Ha Seok. Support vector interval regression machine for crisp input and output data. *Fuzzy Sets and Systems*, 157(8) :1114–1125, 2006.

[35] H. Ishibuchi and M. Nii. Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks. *Fuzzy Sets and Systems*, 119(2) :273–290, 2001.

[36] S. Jozsef. On the effect of linear data transformations in possibilistic fuzzy linear regression. *Fuzzy Sets and Systems*, 45(2) :185–188, 1992.

[37] C. Kao and C.L. Chyu. A fuzzy linear regression model with better explanatory power. *Fuzzy Sets and Systems*, 126(3) :401–409, 2002.

[38] E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time series. In *Proceedings of IEEE International Conference on Data Mining*, pages 289–296, 2001.

[39] H. Lee and H. Tanaka. Fuzzy approximations with non-symmetric fuzzy parameters in fuzzy regression analysis. *Journal of the Operations Research Society of Japan-Keiei Kagaku*, 42(1) :98–112, 1999.

[40] H. Lee and H. Tanaka. Upper and lower approximation models in interval regression using regression quantile techniques. *European Journal of Operational Research*, 116(3) :653–666, 1999.

[41] L. Ljung. Perspectives on system identification. In *Proceedings of 17th IFAC World Congress*, pages 7172–7184, 2008.

[42] J. Lu and R. Wang. An enhanced fuzzy linear regression model with more flexible spreads. *Fuzzy Sets and Systems*, 2009.

[43] M. Modarres, E. Nasrabadi, and MM Nasrabadi. Fuzzy linear regression models with least square errors. *Applied Mathematics and Computation*, 163(2) :977–989, 2005.

[44] RE Moore. Interval analysis (Prentice-Hall Series in Automatic Computation). 1966.

[45] M.M. Nasrabadi and E. Nasrabadi. A mathematical-programming approach to fuzzy linear regression analysis. *Applied Mathematics and Computation*, 155(3) :873–881, 2004.

[46] M.M. Nasrabadi, E. Nasrabadi, and A.R. Nasrabady. Fuzzy linear regression analysis : a multi-objective programming approach. *Applied Mathematics and Computation*, 163(1) :245–251, 2005.

[47] N.J. Nilsson. *Learning machines*. McGraw-Hill New York, 1965.

[48] Y.H. O. Chang and B. M. Ayyub. Fuzzy regression methods–a comparative assessment. *Fuzzy Sets and Systems*, 119(2) :187–203, 2001.

[49] G. Peters. Fuzzy linear regression with fuzzy intervals. *Fuzzy Sets and Systems*, 63(1) :45–55, 1994.

[50] V. Planchon. Traitement des valeurs aberrantes : concepts actuels et tendances générales. *Biotechnol. Agron. Soc. Environ*, 9(1) :19–34, 2005.

[51] D.T. Redden and W.H. Woodall. Further examination of fuzzy linear regression. *Fuzzy sets and systems*, 79(2) :203–211, 1996.

[52] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5) :465–471, 1978.

[53] S. Roychowdhury and W. Pedrycz. Modeling temporal functions with granular regression and fuzzy rules. *Fuzzy Sets and Systems*, 126(3) :377–387, 2002.

[54] M. Sakawa and H. Yano. Multiobjective fuzzy linear regression analysis for fuzzy input-output data. *Fuzzy Sets and Systems*, 47(2) :173–181, 1992.

[55] D.A. Savic and W. Pedrycz. Evaluation of fuzzy linear regression models. *Fuzzy Sets and Systems*, 39(1) :51–63, 1991.

[56] M. Serrurier and H. Prade. A general framework for imprecise regression. In *IEEE International Fuzzy Systems Conference, FUZZ-IEEE 2007*, pages 1597–1602, 2007.

[57] H. Shakouri G and R. Nadimi. A novel fuzzy linear regression model based on a non-equality possibility index and optimum uncertainty. *Applied Soft Computing Journal*, 9(2) :590–598, 2009.

[58] P. Smets. Varieties of ignorance and the need for well-founded theories. *Information Sciences*, 57(58) :135–144, 1991.

[59] K. Takezawa. *Introduction to nonparametric regression*. Wiley-Interscience, 2005.

[60] P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.

[61] H. Tanaka. Fuzzy data analysis by possibilistic linear models. *Fuzzy Sets and Systems*, 24(3) :363–375, 1987.

[62] H. Tanaka, I. Hayashi, and J. Watada. Possibilistic linear regression analysis for fuzzy data. *European Journal of Operational Research*, 40(3) :389–396, 1989.

[63] H. Tanaka and H. Ishibuchi. Identification of possibilistic linear systems by quadratic membership functions of fuzzy parameters. *Fuzzy sets and Systems*, 41(2) :145–160, 1991.

[64] H. Tanaka and H. Lee. Interval regression analysis by quadratic programming approach. *IEEE transactions on fuzzy systems*, 6(4) :473–481, 1998.

[65] H. Tanaka, S. Uejima, and K. Asai. Linear regression analysis with fuzzy model. *IEEE TRANS. SYS. MAN AND CYBER.*, 12(6) :903–907, 1982.

[66] H. Tanaka and J. Watada. Possibilistic linear systems and their application to the linear regression model. *Fuzzy Sets and Systems*, 27(3) :275–289, 1988.

[67] J.M. Vesin and R. Grüter. Model selection using a simplex reproduction genetic algorithm. *Signal Processing*, 78(3) :321–327, 1999.

[68] H.F. Wang and R.C. Tsaur. Resolution of fuzzy regression model. *European Journal of Operational Research*, 126(3) :637–650, 2000.

[69] N. Wang, W.X. Zhang, and C.L. Mei. Fuzzy nonparametric regression based on local linear smoothing technique. *Information sciences*, 177(18) :3882–3900, 2007.

[70] R.R. Yager. Using trapezoids for representing granular objects : applications to learning and OWA aggregation. *Information Sciences*, 178(2) :363–380, 2008.

[71] M.S. Yang and H.H. Liu. Fuzzy least-squares algorithms for interactive fuzzy linear regression models. *Fuzzy Sets and Systems*, 135(2) :305–316, 2003.

[72] Jing-Rung Yu, Gwo-Hshiung Tzeng, and Han-Lin Li. A general piecewise necessity regression analysis based on linear programming. *Fuzzy Sets and Systems*, 105 :429–436, 1999.

[73] J.R. Yu, G.H. Tzeng, and H.L. Li. General fuzzy piecewise regression analysis with automatic change-point detection. *Fuzzy sets and systems*, 119(2) :247–258, 2001.

[74] LA Zadeh. Fuzzy Sets, Information and Control, Vol. 8. *No*, 3 :338–353, 1965.

Publications de l'auteur

Revue d'audience internationale

1. A. Bisserier, R. Boukezzoula and S. Galichet, "A revisited Approach for Linear Fuzzy Regression Using Trapezoidal Fuzzy Intervals", Information Sciences, vol. 180, Issue 19, 2010, pp 3653-3673.

2. A. Bisserier, R. Boukezzoula and S. Galichet, "Linear Fuzzy Regression Using Trapezoidal Fuzzy Intervals", Journal of Uncertain Systems, vol. 4, N. 1, 2010, pp 59-72.

Contribution à ouvrage

3. A. Bisserier, R. Boukezzoula and S. Galichet, "Linear Fuzzy Regression Using Trapezoidal Fuzzy Intervals", Series on Studies in Fuzziness and Soft Computing, vol. 249, Eds. B. Bouchon-Meunier et al., Springer Verlag, 2010, pp 1-22.

Conférences d'audience internationale avec actes

4. L. Tozzi, A. Evsukoff, A. Bisserier, R. Boukezzoula and S. Galichet, "Combining Climate Temperature Models through Fuzzy Interval Regression : Application to La Plata Basin", FUZZ-IEEE Conference, CD-ROM, Barcelona, Spain, July 2010, 6 pages.

5. A. Bisserier, R. Boukezzoula and S. Galichet, "An Interval Approach for Fuzzy Linear Regression with Imprecise Data", IFSA / EUSFLAT Conference, Lisbon, Portugal, July 2009, most, pp 1305-1310.

6. A. Bisserier, R. Boukezzoula and S. Galichet, "Linear Fuzzy Regression Using Trapezoidal Fuzzy Intervals", IPMU Conference, Malaga, Spain, July 2008, pp 181-188.

7. A. Bisserier, S. Galichet and R. Boukezzoula, "Fuzzy Piecewise Linear Regression", FUZZ-IEEE Conference, Hong-Kong, China, June 2008, pp 2089-2094.

Conférences d'audience nationale et francophone avec actes

8. A. Bisserier, R. Boukezzoula et S. Galichet, "Représentation, identification et propagation des incertitudes dans un contexte de régression linéaire", LFA 2009, Annecy, France, novembre 2009, pp 51-58.

9. L. Tozzi, A. Bisserier, A. Evsukoff, R. Boukezzoula and S. Galichet, "Combining Climate Temperature Models through Fuzzy Interval Regression : Application to La Plata Basin", CILAMSE 2009, Armacao de Buzios, Brésil, novembre 2009, pp 1-6.

10. A. Bisserier, F. Megri, R. Boukezzoula et S. Galichet, "Etude expérimentale de la robustesse des techniques linéaire et quadratique de régression floue", LFA 2008, Lens, France, octobre 2008, pp 86-93.

11. A. Bisserier, S. Galichet et R. Boukezzoula , "Une vision de la régression linéaire floue au travers de l'arithmétique des intervalles", LFA 2007, Nîmes, France, novembre 2007, pp 57-64.

Publications soumises

12. R. Boukezzoula, S. Galichet and A. Bisserier, "A Midpoint-Radius Interval Approach for Fuzzy Linear Regression with Imprecise Data", International Journal of Approximate Reasoning, soumis.

Mots-clé : Régression linéaire floue, Régression à base d'intervalles, Identification de modèles polynomiaux imprécis, Identification de modèles dynamiques imprécis

Résumé : L'identification de systèmes est un terme regroupant l'ensemble des techniques permettant l'identification de modèles mathématiques à l'aide d'observations. Dans ce cadre, les techniques régressives sont couramment utilisées. Cette thèse présente une étude de la régression paramétrique linéaire en environnement imprécis. Les mesures et les paramètres du modèle sont alors considérés comme imprécis, et modélisés à l'aide de la théorie des ensembles flous, les entrées étant précises.
Les techniques de régression floues existantes présentent deux limites principales. D'une part, l'imprécision des modèles obtenus est trop importante, du fait notamment que sa variation est imposée par le signe de l'entrée. D'autre part, l'inclusion n'est pas respectée lors de la recherche d'un modèle flou triangulaire devant englober les observations.
Dans ce contexte, différentes améliorations sont proposées et illustrées. L'inclusion est obtenue en identifiant un modèle flou trapézoïdal. L'ajout d'un terme de décalage des entrées permet de rendre l'imprécision de la sortie du modèle indépendante du signe de l'entrée, tout en conservant la linéarité de la structure du modèle. Enfin, un critère d'identification représentant l'imprécision globale du modèle sur son domaine de définition est introduit. Il est alors possible d'identifier un modèle dont la précision et la représentativité sont améliorées. Ces différents concepts sont étendus à l'identification de modèles linéaires par morceaux et multi-entrées.
Le potentiel de la méthode proposée est testé sur des jeux de données réalistes, concernant l'identification de modèles polynomiaux d'ordre adéquat et de modèles multilinéaires. L'identification de modèles dynamiques sur les fluctuations d'un indice boursier permet également d'introduire les problématiques liées aux modèles régressifs flous à entrées imprécises.

Abstract : System identification is a term gathering tools that identify mathematical models from observations. Within this framework, regression techniques are frequently used. This Ph. D. thesis deals with the study of parametrical linear regression in an imprecise context. So, measurements and model parameters are imprecise and represented using fuzzy set theory, while inputs are crisp numbers.
Existing fuzzy regression techniques present two main limits. On the one hand, the imprecision of identified models is too important, mainly due to the link between imprecision variation and input sign. On the other hand, inclusion is not guaranteed even when a triangular fuzzy model, which should include observations, is identified.
In this context, several improvements are introduced and illustrated. Inclusion is guaranteed by the identification of trapezoidal fuzzy models. By applying a shift term to inputs, the model output imprecision becomes independent of input sign, while model linear structure is preserved. Lastly, an optimization criterion which represents the global fuzziness of the model on its definition domain is introduced. It is then possible to improve the precision of the identified model as well as its representativeness. All these concepts are extended to piecewise and multi-inputs linear model identification.
The potential of the proposed method is tested on realistic data sets, concerning the identification of polynomial models with appropriate order and multi-linear models. By identifying dynamical models from variations of a market index, problems related to fuzzy regressive models with imprecise inputs are also introduced.