Applications de techniques avancées de contrôle des procédés en industrie du semi-conducteur.

Nader Jedidi

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

N° d'ordre : **535 M**

THÈSE

présentée par

Nader JEDIDI

pour obtenir le grade de
Docteur de l'École Nationale Supérieure des Mines de Saint-Étienne

Spécialité : Microélectronique

APPLICATIONS DE TECHNIQUES AVANCEES DE CONTROLE DES PROCEDES EN INDUSTRIE DU SEMI-CONDUCTEUR

Soutenue à Gardanne, le 05 Octobre 2009

Membres du jury

Président	Jean-François LAFAY	Professeur, Ecole Centrale de Nantes
Rapporteurs	René-Louis INGLEBERT	Professeur, Université J. Fourier, Polytech Grenoble
	Didier GOGUENHEIM	Directeur de la Recherche, ISEN, Toulon
Examinateurs	Philippe CASTAGLIOLA	Professeur, Université de Nantes
	Pascal SALLAGOITY	Docteur, STMicroelectronics, Rousset
	Jacques PINATON	Ingénieur, STMicroelectronics, Rousset
Directeur de thèse	Stéphane DAUZERE-PERES	Professeur, Ecole des Mines de Saint-Etienne
Co-encadrante	Agnès ROUSSY	Maître-Assistant, Ecole des Mines de Saint-Etienne

Spécialités doctorales :	**Responsables :**
SCIENCES ET GENIE DES MATERIAUX	J. DRIVER Directeur de recherche – Centre SMS
MECANIQUE ET INGENIERIE	A. VAUTRIN Professeur – Centre SMS
GENIE DES PROCEDES	G. THOMAS Professeur – Centre SPIN
SCIENCES DE LA TERRE	B. GUY Maître de recherche – Centre SPIN
SCIENCES ET GENIE DE L'ENVIRONNEMENT	J. BOURGOIS Professeur – Centre SITE
MATHEMATIQUES APPLIQUEES	E. TOUBOUL Ingénieur – Centre G2I
INFORMATIQUE	O. BOISSIER Professeur – Centre G2I
IMAGE, VISION, SIGNAL	JC. PINOLI Professeur – Centre CIS
GENIE INDUSTRIEL	P. BURLAT Professeur – Centre G2I
MICROELECTRONIQUE	Ph. COLLOT Professeur – Centre CMP

Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou d'une HDR)

AVRIL	Stéphane	MA	Mécanique & Ingénierie	CIS
BATTON-HUBERT	Mireille	MA	Sciences & Génie de l'Environnement	SITE
BENABEN	Patrick	PR 2	Sciences & Génie des Matériaux	CMP
BERNACHE-ASSOLANT	Didier	PR 0	Génie des Procédés	CIS
BIGOT	Jean-Pierre	MR	Génie des Procédés	SPIN
BILAL	Essaïd	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR 2	Informatique	G2I
BOUCHER	Xavier	MA	Génie Industriel	G2I
BOUDAREL	Marie-Reine	MA	Génie Industriel	DF
BOURGOIS	Jacques	PR 0	Sciences & Génie de l'Environnement	SITE
BRODHAG	Christian	DR	Sciences & Génie de l'Environnement	SITE
BURLAT	Patrick	PR 2	Génie industriel	G2I
COLLOT	Philippe	PR 1	Microélectronique	CMP
COURNIL	Michel	PR 0	Génie des Procédés	DF
DAUZERE-PERES	Stéphane	PR 1	Génie industriel	CMP
DARRIEULAT	Michel	IGM	Sciences & Génie des Matériaux	SMS
DECHOMETS	Roland	PR 1	Sciences & Génie de l'Environnement	SITE
DESRAYAUD	Christophe	MA	Mécanique & Ingénierie	SMS
DELAFOSSE	David	PR 1	Sciences & Génie des Matériaux	SMS
DOLGUI	Alexandre	PR 1	Génie Industriel	G2I
DRAPIER	Sylvain	PR 2	Mécanique & Ingénierie	SMS
DRIVER	Julian	DR	Sciences & Génie des Matériaux	SMS
FEILLET	Dominique	PR 2	Génie Industriel	CMP
FOREST	Bernard	PR 1	Sciences & Génie des Matériaux	CIS
FORMISYN	Pascal	PR 1	Sciences & Génie de l'Environnement	SITE
FORTUNIER	Roland	PR 1	Sciences & Génie des Matériaux	SMS
FRACZKIEWICZ	Anna	DR	Sciences & Génie des Matériaux	SMS
GARCIA	Daniel	CR	Génie des Procédés	SPIN
GIRARDOT	Jean-Jacques	MR	Informatique	G2I
GOEURIOT	Dominique	MR	Sciences & Génie des Matériaux	SMS
GOEURIOT	Patrice	MR	Sciences & Génie des Matériaux	SMS
GRAILLOT	Didier	DR	Sciences & Génie de l'Environnement	SITE
GROSSEAU	Philippe	MR	Génie des Procédés	SPIN
GRUY	Frédéric	MR	Génie des Procédés	SPIN
GUILHOT	Bernard	DR	Génie des Procédés	CIS
GUY	Bernard	MR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HERRI	Jean-Michel	PR 2	Génie des Procédés	SPIN
INAL	Karim	MR	Microélectronique	CMP
KLÖCKER	Helmut	MR	Sciences & Génie des Matériaux	SMS
LAFOREST	Valérie	CR	Sciences & Génie de l'Environnement	SITE
LERICHE	Rodolphe	CR	Mécanique et Ingénierie	SMS
LI	Jean-Michel	EC (CCI MP)	Microélectronique	CMP
LONDICHE	Henry	MR	Sciences & Génie de l'Environnement	SITE
MOLIMARD	Jérôme	MA	Mécanique et Ingénierie	SMS
MONTHEILLET	Frank	DR 1 CNRS	Sciences & Génie des Matériaux	SMS
PERIER-CAMBY	Laurent	PR1	Génie des Procédés	SPIN
PIJOLAT	Christophe	PR 1	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR 1	Génie des Procédés	SPIN
PINOLI	Jean-Charles	PR 1	Image, Vision, Signal	CIS
STOLARZ	Jacques	CR	Sciences & Génie des Matériaux	SMS
SZAFNICKI	Konrad	CR	Sciences & Génie de l'Environnement	DF
THOMAS	Gérard	PR 0	Génie des Procédés	SPIN
VALDIVIESO	François	MA	Sciences & Génie des Matériaux	SMS
VAUTRIN	Alain	PR 0	Mécanique & Ingénierie	SMS
VIRICELLE	Jean-Paul	MR	Génie des procédés	SPIN
WOLSKI	Krzysztof	MR	Sciences & Génie des Matériaux	SMS
XIE	Xiaolan	PR 1	Génie industriel	CIS

Glossaire :

PR 0	Professeur classe exceptionnelle
PR 1	Professeur 1ère catégorie
PR 2	Professeur 2ème catégorie
MA(MDC)	Maître assistant
DR (DR1)	Directeur de recherche
Ing.	Ingénieur
MR(DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
IGM	Ingénieur général des mines

Centres :

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
SITE	Sciences Information et Technologies pour l'Environnement
G2I	Génie Industriel et Informatique
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

Dernière mise à jour le : 22 juin 2009

Table des matières

Introduction

La course ininterrompue à la performance des dispositifs submicroniques, et à la réduction des coûts, a servi de moteur à la miniaturisation des transistors dans l'industrie de la micro-électronique. Jusqu'alors et en accord avec la loi de Moore, les générations successives de circuits **MOS** (***M**etal **O**xyde **S**emiconductor*) suivent une loi d'échelle pour les grandeurs principales telles que la longueur de grille, l'épaisseur d'oxyde de grille et le dopage du canal [Schaller (1997)]. Tous les trois ans, à chaque nouvelle génération, les dimensions sont divisées par racine de deux. Actuellement, les transistors réalisés en production chez STMicroelectronics ont des longueurs de grille de 65 nm. Les générations 45 nm sont en cours de développement.

L'industrialisation des circuits **CMOS** (***C**omplementary **M**etal **O**xyde **S**emiconductor*) doit néanmoins faire face à plusieurs limitations, tant au niveau de l'architecture du transistor et des technologies nécessaires pour atteindre l'intégration souhaitée, qu'au niveau de la maîtrise des différentes étapes de fabrication. Dans un environnement de production, les procédés de réalisation du transistor souffrent en effet d'une fluctuation permanente, engendrée par plusieurs facteurs : la fluctuation des propriétés de la matière première (résine photosensible, substrat), l'évolution des équipements dans le temps (vieillissement, changement de température)...

Parallèlement, le coût élevé des équipements en microélectronique rend impossible leur amortissement sur une génération technologique, d'autant plus que la durée de vie des générations décroît rapidement. Ceci incite l'industriel à adapter des équipements d'ancienne génération à des procédés de fabrication plus exigeants. Certains songent à utiliser les machines avec des produits de technologies très différentes. Cette stratégie n'est pas sans conséquence sur la dispersion des caractéristiques physiques (dimensions géométriques) et électriques (courant, tension,...) des transistors qui se trouve amplifiée. Le rendement final (Yield) pourrait ainsi être dégradé et entraîner une réduction des bénéfices.

Au delà de la multiplicité des sources de variation des caractéristiques du produit final, la figure 1 illustre un autre élément à prendre en compte. La sensibilité des paramètres électriques aux variations des dimensions physiques croît à chaque nouveau nœud technologique. La fréquence de l'horloge en est un exemple. Dans ce contexte, le contrôle des étapes de fabrication élémentaires s'avère être un des points clés de la survie des technologies avancées car il permet de minimiser au final la dispersion des caractéristiques électriques du dispositif.

Avant d'introduire les différents chapitres, nous devons définir un terme récurrent qui est au centre de ces travaux de thèse : la variabilité ou la variation. L'industrie du semi-conducteur est une industrie où la fabrication des produits est réalisée par lots. Chaque lot contient 25 plaques en silicium monocristallin. Dans ce cadre, la

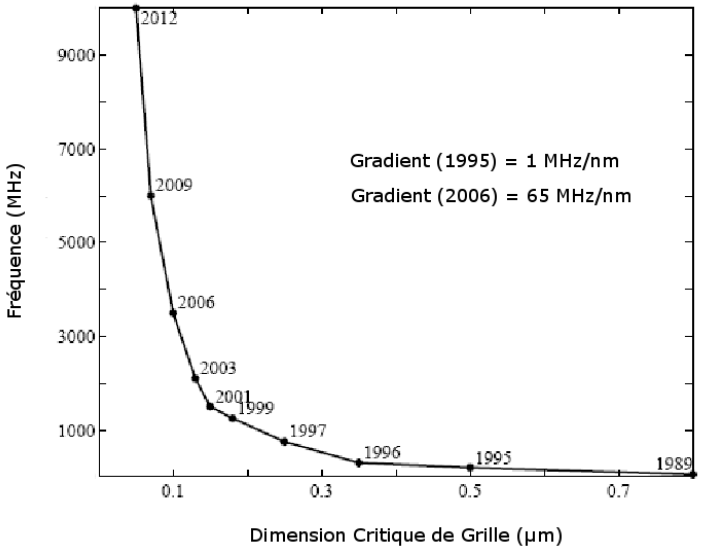

FIGURE 1 – *Evolution de la vitesse d'un microprocesseur en fonction de la dimension critique de grille (CD). Source : Ruegsegger (1998).*

variabilité d'un paramètre, qu'il soit physique ou électrique, est double : spatiale et temporelle [Stine et al. (1997)]. La variation spatiale est la somme d'une variation intra-plaque, due par exemple à un procédé de gravure à vitesse rayon-dépendante ou à un dépôt par centrifugation peu uniforme, et d'une variation intra-champs, intimement liée au procédé de photolithographie. La variation intra-champs a ses origines dans les aberrations de la lentille, l'architecture du circuit et les erreurs intra-réticule [1]. Enfin, la variabilité temporelle, lot à lot ou plaque à plaque, est due principalement à une variation de la matière première et aux dérives des procédés.

Le premier objectif de ces travaux est de définir une méthodologie statistique capable de satisfaire deux points :

⋈ estimer le poids des différentes composantes de variabilité (intra-plaque, plaque à plaque, lot à lot) des paramètres électriques critiques, à savoir le courant de saturation, le courant de fuite et la tension de seuil des transistors MOS courts $0.13\mu m$.

⋈ et indiquer, à travers une analyse statistique multi-variée, les étapes de fabrication critiques, principalement responsables des variabilités temporelle et spatiale des performances électriques du dispositif.

Cette méthodologie sera appliquée à un ensemble de produits logiques d'une technologie CMOS $0.13\mu m$ (H9), fabriqués sur le site de STMicroelectronics à Rousset. Les résultats obtenus pour cette analyse, détaillés chapitre 2, montrent une forte variabilité lot à lot du courant de saturation et du courant de fuite. Elle est due au premier ordre à la variabilité de la longueur de grille en poly-Si. Ce constat a été le point de départ pour le développement et la mise en place de boucles de régulation, connues sous le nom de boucles *Run to Run* dans la littérature anglo-saxonne.

Une régulation est un système de contrôle qui ajuste une ou plusieurs variables

1. La qualité du masque est en effet critique, car toute déviation de la dimension critique de la valeur prévue par le concepteur sera amplifié d'un facteur appelé **MEEF** (***M**ask **E**rror **E**nhancement **F**actor*).

du procédé (un débit de gaz, une dose d'énergie, etc) de manière à minimiser la dispersion d'une ou plusieurs caractéristiques du produit. Un premier type de régulation est dit avec retour (*Feed-Back*, FB). La commande à l'instant k est alors dépendante des déviations de la grandeur régulée, mesurées aux instants précédents (< k). Il s'agit d'une action en boucle fermée. L'étude de telles boucles et leur application dans les usines de fabrication des composants semi-conducteurs a commencé au début des années 90 avec les travaux de A.Ingolfsson and E.Sachs [Ingolfsson (1991), Sachs et Ingolfsson (1993), Sachs et al. (1995)]. Aujourd'hui, nous disposons d'un large panel de régulations qui s'étend du contrôle adaptatif au contrôle prédictif. Un état de l'art des principales méthodes de régulation et quelques une de leurs applications dans la fabrication des dispositifs micro-électroniques sera présenté dans le chapitre 3.

Un second type de régulation est celle dite de compensation (*Feed-Forward*, FF). La perturbation est alors prise en compte avant qu'elle ne prenne effet sur les paramètres régulés. Cette action, en boucle ouverte, anticipe l'effet de la perturbation : on parle alors d'action anticipatrice. Un exemple de boucle FF est celle qui compense la déviation de la longueur de la grille en ajustant la dose d'implantation des poches des transistors MOS. L'objectif est alors de réduire la dispersion lot à lot de la longueur effective du canal. Le chapitre 4 sera consacré au développement de cette régulation et aux résultats de son déploiement en production.

Enfin, dans le chapitre 5 de ce manuscrit, nous allons nous focaliser sur une thématique d'actualité, qui a émergé depuis la fin des années 90 [Miller et al. (1997)]. Il s'agit de prendre en compte d'une façon complètement automatisée la multitude des produits, des équipements de process et de métrologie, et des recettes utilisées dans la conception des boucles de régulation. En clair, l'objectif sera de concevoir un estimateur récursif en ligne qui mettrait à jour à chaque nouvelle mesure les contributions à la déviation de la variable de sortie de chaque source de variation qualitative (équipement, recette, produit,...). Cette démarche rompt avec la méthode actuellement utilisée qui consiste à mettre en place une boucle indépendante pour chaque couple (équipement de process, recette). L'évaluation de la performance de ce nouveau contrôleur sera réalisée à travers la simulation de l'atelier de photolithographie. Réalisées à partir des données de production, les simualtions permettront de comparer, à l'existant, différents algorithmes d'estimations candidats (filtre de Kalman, moindres carrés récursifs, etc) .

Eléments de l'industrie de fabrication des composants semi-conducteurs

1

Sommaire

La technologie **CMOS** (***C**omplementary **M**etal **O**xyde **S**emiconductor*) est aujourd'hui la plus répandue parmi toutes les technologies semi-conducteurs (SOI, BiCMOS,...). Elle englobe à elle seule 70% de la production mondiale de circuits intégrés. Un de ses atouts majeurs est un processus de fabrication simple et parfaitement maîtrisé. Dans ce chapitre, nous allons en décrire quelques étapes élémentaires. L'accent sera mis sur des opérations communément adoptées pour la fabrication des composants CMOS, à savoir la lithographie, la gravure, l'implantation ionique et le recuit d'activation. Ces étapes seront au centre du fonctionnement des régulations développées dans les chapitres suivants.

Parallèlement aux étapes élémentaires de fabrication, sont intégrées des méthodes de caractérisation en ligne de différents paramètres physiques et géométriques associés au produit. Elles renseignent sur le déroulement des opérations. Parmi ces outils de métrologie, nous allons nous intéresser particulièrement à la mesure de la dimension critique, une mesure qui alimentera les contrôleurs mentionnés ci-dessus. Une description détaillée des principes de fonctionnement de la scattérométrie et de la microscopie électronique à balayage sera en effet intégrée à ce chapitre.

En fin du processus de fabrication des plaques et avant la découpe et l'encapsulation des puces, une batterie de tests est réalisée afin de vérifier le bon fonctionnement du produit. La première partie du test est le test paramétrique (*Parametric Test* ou PT), où l'on doit contrôler certaines grandeurs électriques (courant de saturation, tension de seuil, etc). Le PT est un outil précieux pour qualifier les étapes de fabrication et identifier la cause en cas de défaillance. Par la suite est réalisé le test électrique final généralement appelé EWS (*Electrical Wafer Sorting*). La dernière partie de ce chapitre sera consacrée au test.

1.1 Procédés de Fabrication en Industrie du Semi-conducteur

1.1.1 La photolithographie [Levinson (2005), Benedicte (2001), Mack (2006)]

Il s'agit du procédé clé de la fabrication des circuits semi-conducteurs. Il consiste en une succession d'opérations élémentaires qui doivent permettre le transfert des motifs reproduits sur le masque dans le film de résine. Du fait des phénomènes de diffraction de la lumière lorsque les dimensions critiques à transférer dans la résine sont inférieures à la longueur d'onde d'exposition, la réduction des dimensions des circuits entraîne la réduction en lithographie de la longueur d'onde d'exposition. Les longueurs d'onde communément utilisées ces dernières années en production ont été :

- $\lambda = 436$ nm (raie g de la lampe à vapeur de mercure), pour des résolutions jusqu'à $0.7\,\mu$m.
- $\lambda = 365$ nm (raie i de la lampe à vapeur de mercure), pour des résolutions jusqu'à $0.35\,\mu$m.
- $\lambda = 248$ nm (laser excimère KrF), pour des résolutions jusqu'à $0.18\,\mu$m.
- $\lambda = 193$ nm (laser excimère ArF), pour des résolutions en deça de $0.13\,\mu$m.

Le procédé lithographique utilisé pour les deux premières longueurs d'onde (UV proche) est identique. Il utilise une résine qui fonctionne sur un principe de photosensibilisation directe : un photon induit une et une seule réaction photochimique qui modifie le comportement de dissolution de la résine dans le développeur. Le passage de la lithographie en UV proche à la lithographie en UV profonds (248 nm et 193 nm) s'accompagne de l'augmentation de l'énergie des photons avec la réduction de leur longueur d'onde. Par conséquent, pour une même dose d'énergie appliquée, le nombre de photons diminue. Il n'est alors plus possible de garder un mécanisme de déprotection basé sur le principe "un photon engendre une réaction de modification". Un nouveau concept a donc été proposé : l'amplification chimique.

Les résines à amplification chimique sont composées d'une matrice en polymère ($>$ 90 % de la formulation), d'un composant photogénérateur d'acide ainsi que de quelques additifs. Afin d'empêcher la dissolution du polymère, initialement soluble dans le développeur, des groupements chimiques inhibiteurs de dissolution lui sont greffés. Sous exposition UV, le composé photogénérateur d'acide (*PhotoAcid Generator*, **PAG**) va libérer un acide :

$$\text{PAGs} \xrightarrow{h\nu} \mathit{Acide} + (\text{Sous}-\text{produits}), \tag{1.1}$$

Il y a création d'une image latente du masque dans la résine, formé par la présence de molécules d'acide photogénérées dans les zones exposées. Cependant, à cette étape du procédé, aucun changement des propriétés de la résine n'a eu lieu dans les zones exposées. Le changement de vitesse de dissolution dans le développeur des zones exposées est réalisé au cours d'une seconde réaction chimique, dite de déprotection de la matrice polymère par catalyse d'acide. Lors de cette réaction, les groupements protecteurs sont clivés , entraînant une augmentation de la dissolution des zones exposées de la résine dans le développeur.

$$\text{Polymère}_{insoluble} + \text{Acide} \xrightarrow{\Delta} \text{Polymère}_{soluble} + \text{Acide}_{\text{regénéré}} + \text{Sous}-\text{produits} \quad (1.2)$$

La déprotection par catalyse acide d'une résine pour UV profonds nécessite très souvent une activation thermique et se déroule donc généralement durant le recuit après exposition (***P**ost **E**xposure **B**ake*, **PEB**). Une molécule d'acide est régénérée en fin de réaction et elle est susceptible d'aller déprotéger un autre site après diffusion dans le film de résine. Il a ainsi été évalué qu'une molécule d'acide photogénéré engendre la déprotection de 500 à 1000 sites protégés sur le polymère, provoquant ainsi un changement chimique important dans la résine exposée. Ces résines sont qualifiées de résines à amplification chimique.

Les différentes étapes nécessaires à la réalisation des motifs critiques (grille, STI, contact) en lithographie appliquée aux résines à amplification chimique 193 nm sont décrites brièvement ci-dessous.

Dépôt de la résine

La résine est déposée par centrifugation (*Spin Coating*). L'épaisseur e du film est fonction de la viscosité de la résine, des conditions atmosphériques de la salle blanche (température, degré d'hygrométrie) et surtout de la vitesse de rotation v. Le film de résine généré après dépôt est instable physiquement car très expansé du fait de l'élimination du solvant et du figeage rapide du film.

$$e = av^{-\frac{1}{2}}, \quad (1.3)$$

où a est une constante indépendante de la vitesse de la tournette $(nm \cdot rpm^{1/2})$.

Recuit de la résine

Ce recuit a pour but de stabiliser le film de résine expansé en permettant sa re-compaction accélérée au cours d'un recuit à température élevée de l'ordre de 100°C. En réduisant le volume libre présent dans le film, généré par le dépôt, il permet de densifier le film et de contrôler les phénomènes de diffusion des PAGs notamment.

Insolation

Suite au recuit de résine, la plaque est exposée dans un outil de lithographie par projection. Cette technique permet l'utilisation de masques dessinés à l'échelle de 4 à 5 par rapport aux dimensions réelles du circuit. L'insolation des étapes critiques des technologies sub 0.18μm est réalisée par répétition et balayage (voir Figure 1.1). L'image du réticule est projetée au travers d'une fente[1] au cours du balayage synchrone du masque et du substrat. Le passage d'un champ d'exposition à un autre s'effectue dans la continuité de l'étape d'exposition. Les avantages des outils d'exposition par balayage sont surtout l'utilisation d'une partie réduite de la lentille, à savoir le centre de la lentille, et la possibilité de travailler avec de très grands champs d'exposition dans le sens du balayage.

1. Le terme anglo-Saxon qui désigne la fente est ***slit***. Il s'agit de la partie utile de la lentille où l'on limite les problèmes de distorsion liés aux aberrations optiques

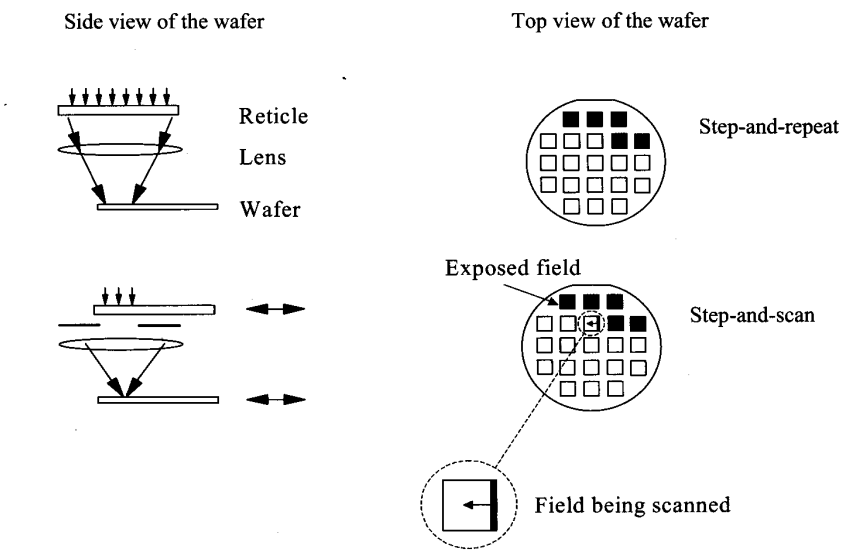

FIGURE 1.1 – *Deux configurations pour l'insolation par projection : step and repeat & step and scan. Source : Levinson (2005).*

Recuit Post Insolation

Le recuit PEB (*Post Exposure Bake*) est une phase capitale du procédé pour les résines à amplification chimique, car c'est au cours de celle-ci que la déprotection par catalyse acide de la matrice polymère est achevée. Le recuit PEB a aussi pour but l'homogénéisation de l'image latente transférée dans la résine, en accroissant la diffusion de l'acide photogénéré, ce qui permet de réduire les ondes stationnaires formés dans le film de résine lors de l'exposition. A la fin du PEB, la plaque est brutalement refroidie par le contact direct avec une coupelle froide, afin d'arrêter les réactions de déprotection. A ce stade, l'image du masque dans la résine est une image latente formée du polymère déprotégé dans les zones exposées et du polymère protégé dans les zones non exposées. A noter que le gradient de température du four PEB impacte notamment la dispersion dimensionnelle intra-plaque, ce qui rend cette étape particulièrement critique comme l'on démontre dans la section 2.4 du chapitre 2.

Développement

Suite à la déprotection des polymères dans les zones insolées, la résine peut être développée en mettant à nu les zones du substrat où la résine a été exposée. Les résines à amplification chimique requièrent un développement humide, avec l'utilisation d'un développeur aqueux basique. Cette étape est une succession dans l'ordre chronologique d'un procédé dynamique d'homogénéisation du développeur sur la plaque, d'un procédé statique où le développeur agit sur la résine et finalement d'un rinçage.

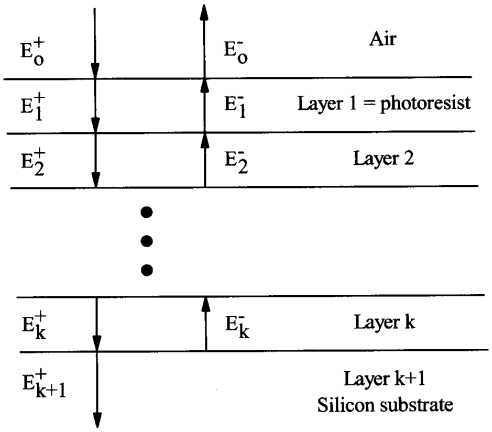

FIGURE 1.2 – *Configuration optique d'un empilement de films minces. Source : Levinson (2005).*

Contrôle de la réflectivité du substrat à 193 **nm : Couche anti-reflet**

Dans le film de résine, la lumière comporte une composante incidente et une composante réfléchie (voir figure 1.2). Deux conséquences de taille suivent ce constat [Levinson (2005)].

⋈ La résine est le siège d'un système d'ondes stationnaires, résultat des interférences entres les deux composantes. La période du système est de $\lambda/4n$ avec λ est égal à la longueur d'onde de la lumière d'exposition et n est la partie réelle de l'indice de réfraction de la résine. Cette périodicité est illustrée en figure 1.3, elle se superpose à la perte progressive de l'intensité lumineuse dans l'épaisseur du film, du fait de l'absorption optique de la résine. Ce phénomène induit une dégradation du profil de la résine après développement. En jouant sur la diffusion de l'acide photogénéré, le choix des conditions de recuit PEB permet de lisser les ondes stationnaires et éviter la rugosité périodique sur les flans de résine.

⋈ La dose d'énergie réellement déposée dans la couche de résine est dépendante de l'épaisseur de résine et des épaisseurs des couches sous-jacentes.

Afin de limiter la dégradation induite par la création d'un système d'ondes stationnaires dans la résine, l'emploi de couches anti-reflet dites BARC (*Bottom Anti Reflective Coating*) a été introduit : il s'agit de couches qui s'intercalent entre le substrat et la résine et qui agissent sur l'ensemble des réflexions à l'interface résine/substrat. Le BARC est optimisé de sorte qu'il absorbe un maximum de la radiation incidente d'exposition afin d'éliminer l'onde réfléchie dans la résine. Deux possibilités se présentent pour l'utilisation d'un BARC 193 nm :

⋈ Une couche anti-reflet organique. Elle se compose de polymères souvent assez proches de ceux composant la résine, auxquels sont greffés des groupements fortement absorbants à la radiation d'exposition. Elle est déposée à la tournette puis stabilisée au cours d'un recuit à haute température. La résine est ensuite déposée et traitée selon le processus décrit précédemment.

⋈ Une couche anti-reflet minérale de type SiOxNy. Le procédé l'utilisant est plus complexe, car il nécessite le dépôt par plasma de la couche inorganique sur le substrat.

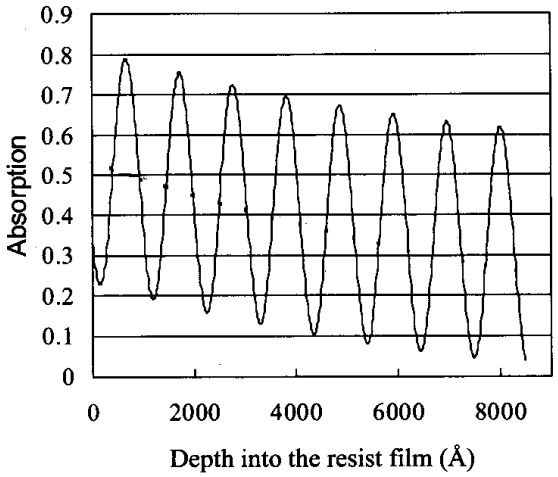

FIGURE 1.3 – *L'évolution de l'intensité de la lumière ($\lambda = 365\,nm$) à travers un film épais de résine déposé sur un substrat en silicium. Source : Levinson (2005)*

1.1.2 La gravure plasma [Pargon (2004)]

La gravure plasma consiste à transférer des motifs initialement définis par la lithographie dans les couches actives des dispositifs (semi-conducteurs, isolants ou conducteurs). Elle est de nature physico-chimique car elle met en jeu à la fois un bombardement ionique (une gravure physique ou mécanique) et une réaction chimique (gravure chimique) entre le gaz ionisé et la surface de la plaque.

La gravure physique

La gravure physique correspond au bombardement du substrat par les ions du plasma. Lorsque ces ions entrent en collision avec les atomes de la surface, ces derniers peuvent gagner assez d'énergie sous l'impact ionique pour être pulvérisés et quitter ainsi la surface de l'échantillon. La gravure en plasma de gaz rare tel l'argon est l'exemple type de gravure purement ionique et physique. La gravure physique présente le grand intérêt d'être anisotrope mais souffre généralement d'une absence de sélectivité : une faible différence entre la vitesse de pulvérisation des différents matériaux utilisés en micro-électronique.

La gravure chimique

Elle ne dépend que de l'interaction entre le matériau à graver et les espèces réactives générées dans le plasma. Le mécanisme d'une gravure chimique peut se décomposer en quatre étapes élémentaires décrites ci-après et appuyées sur l'exemple de la gravure du silicium par un plasma à base de CF_4.

– Création d'espèces chimiquement réactives dans le plasma

$$CF_4 + e^- \longrightarrow C\uparrow + 4F\uparrow + e^-, \tag{1.4}$$

– Adsorption des espèces réactives à la surface

$$Si + F\uparrow \longrightarrow SiF, \tag{1.5}$$

– Formation des produits de réaction volatils,

$$4\,\mathrm{F}\uparrow + \mathrm{Si} \longrightarrow \mathrm{SiF_4}\uparrow, \tag{1.6}$$

– Désorption et pompage hors de l'enceinte des produits de réaction.

Il est important que le produit de gravure (SiF_4 dans l'exemple) soit volatil et stable afin qu'il puisse rapidement quitter la surface, puis être évacué du milieu par le système de pompage. L'avantage de la gravure chimique est sa forte sélectivité. De plus, elle ne provoque pas de dommage en volume du matériau en cours de gravure. Par contre, c'est une gravure lente et totalement isotrope. Elle est donc inadaptée à la réalisation de motifs sub-micrométriques pour laquelle une attaque latérale sous le masque en résine est inacceptable.

La gravure ionique réactive RIE

La gravure ionique réactive se situe à mi-chemin entre gravure physique et gravure chimique. Les ions, se neutralisant à l'approche de la surface du substrat, deviennent alors des neutres réactifs hautement énergétiques qui participent directement à la gravure chimique de l'échantillon. La gravure anisotrope du matériau est possible car les ions sont accélérés quasi perpendiculairement au substrat. Cependant, la composante latérale de la gravure, due essentiellement à l'aspect chimique, n'est pas négligeable, et peut induire des distorsions dans les profils de gravure. Cette composante latérale de la gravure peut néanmoins être minimisée grâce à la passivation des flancs (voir Figure 1.4). En effet, en plus des ions et des neutres réactifs, le plasma peut aussi produire des molécules dites inhibitrices. Ces molécules peuvent alors s'adsorber sur les flancs des motifs en cours de gravure pour former une couche mince de passivation sur les flancs du matériau à graver et ainsi bloquer la gravure latérale, en isolant le matériau des espèces réactives du plasma.

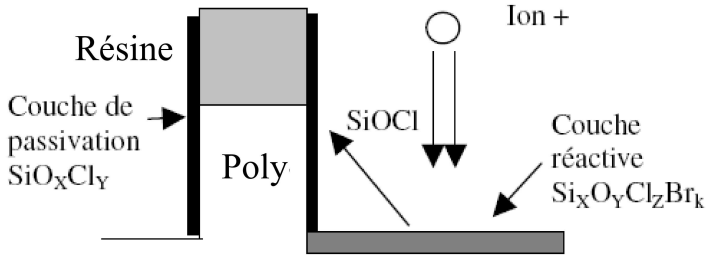

FIGURE 1.4 – *Formation des couches de passivation sur les flancs de la grille en silicium en chimie $HBr/Cl_2/O_2$: pulvérisation de produits SiOCl du fond des tranchées et redépôt direct sur les flancs des motifs.*

La gravure du poly-silicium

Du fait de l'importance de la grille dans la performance du dispositif, la gravure du poly-silicium est considérée comme critique. L'objet de ce paragraphe est d'en décrire les différentes étapes (voir Figure 1.5). Afin de graver du silicium, des chimies à base d'halogènes sont souvent utilisées car elles permettent la formation de produits de réaction volatils de type SiX_4 (où X= F, Cl, ou Br).

⋈ *La gravure de la couche d'anti-reflet (BARC)*
La couche anti-réflective doit être gravée afin de déboucher sur le poly-silicium.

⋈ *La gravure du polysilicium*
Cette gravure est réalisée aujourd'hui en quatre étapes successives qui sont : le perçage de l'oxyde natif (*BreakThrough* **BT**), la gravure principale (*Main-Etch* **ME**), l'atterrissage sur la couche sous-jacente d'oxyde mince (*Soft Landing* **SL**) et la sur-gravure (*Over-Etch* **OE**). Le rôle du BT est la destruction de l'oxyde natif en surface du silicium, oxyde natif qui s'est formé pendant le transport entre les étapes de lithographie et de gravure. L'étape de gravure principale consiste à graver en grande partie la grille et à s'arrêter peu avant de déboucher sur l'oxyde de silicium (50 nm de l'oxyde environ), grâce à l'interférométrie, une technique qui permet de suivre en temps réel l'épaisseur gravée. Nous utilisons aujourd'hui des chimies halogénées telles que $HBr/Cl_2/O_2$ ou $HBr/Cl_2/O_2/CF_4$. L'étape d'atterrissage sur l'oxyde a été introduite afin de protéger l'oxyde de grille lorsque le plasma arrive dessus. Les paramètres du plasma sont ajustés pour obtenir un bombardement ionique de faible énergie. Quant à l'étape de sur-gravure, elle permet de terminer la gravure des résidus de silicium sur l'oxyde de grille.

⋈ *Le retrait du masque*
Le retrait de la résine et du BARC, également appelé *stripping*, consiste à retirer, sélectivement vis-à-vis de l'oxyde de silicium ou du poly-silicium, la résine et la couche d'anti-reflet. Cette étape est réalisée par voie "sèche" avec un plasma d'oxygène pur, suivie d'une gravure humide dans une solution d'acide fluorhydrique diluée à quelques %.

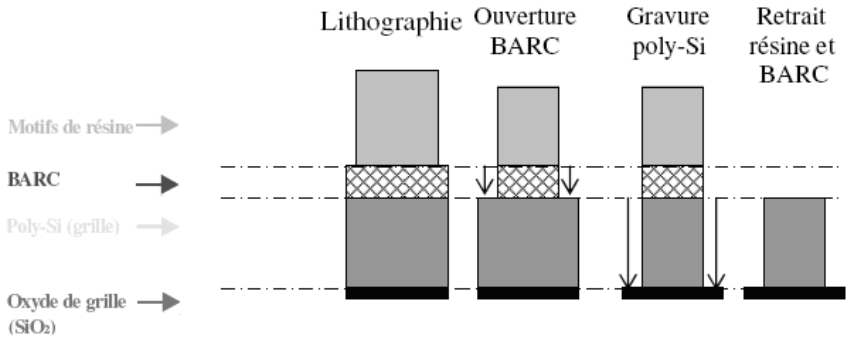

FIGURE 1.5 – *Etapes de fabrication d'une grille en poly-silicium avec un masque résine*

1.1.3 Le dopage par implantation ionique

L'implantation ionique consiste en l'introduction d'une faible concentration d'atomes de la colonne **III** (impuretés de type P) ou de la colonne **V** (impuretés de type N) dans le réseau cristallin du silicium, lui conférant ainsi des propriétés semi conductrices. D'un point de vue pratique, les atomes dopants sont vaporisés, ionisés, accélérés et projetés sur le matériau à doper dans lequel se produisent de nombreuses collisions avec les atomes du substrat. Les ions subissent une perte graduelle d'énergie jusqu'à ce qu'ils s'arrêtent à une certaine profondeur. La profondeur moyenne est

contrôlée en ajustant l'énergie de l'implantation. La dose d'impuretés introduites est fixée par le courant ionique et la durée du balayage.

Plus de 250 étapes dont une trentaine d'implantation sont actuellement nécessaires à la réalisation des transistors MOS. Citons à titre d'exemple l'implantation du canal qui permet d'ajuster la tension de seuil du transistor long. Aussi, l'implantation des poches qui vient se superposer à cette dernière pour réaliser un canal rétrograde, mieux adapté à la réduction des effets de canal court. Enfin, citons l'implantation des extensions, de la source et du drain qui sont des implantations à très forte dose. Les zones dopées sont ainsi dégénérées, afin de réduire les résistances d'accès.

Suite à l'étape de dopage par implantation ionique, un recuit rapide d'activation est nécessaire pour réparer les défauts ponctuels engendrés par l'implantation ionique et restaurer ainsi le caractère cristallin du substrat. Il permet aussi d'activer les dopants en les positionnant en sites substitutionnels dans le silicium, là où ils sont électriquement actifs. Actuellement, la formation des zones source/drain (S/D), des extensions (*Low Doped Drain*) et l'activation électrique de la grille en poly-silicium sont réalisées par implantation ionique suivie d'un recuit spike (voir Figure 1.6).

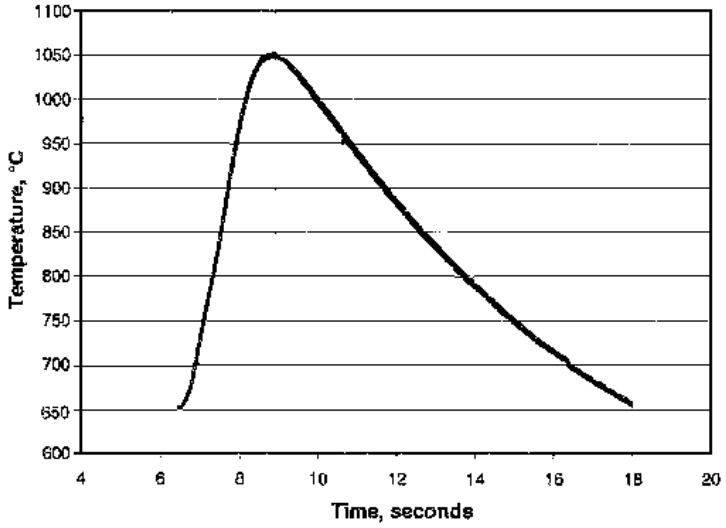

FIGURE 1.6 – *Exemple d'une recette de process de recuit spike. Nous y observons trois étapes principales : Une étape de stabilisation qui dure* $\sim$ 6 *secondes, suivie d'une rampe de température où l'on progresse de* 75 ℃/*s jusqu'à* 1050 ℃ *et enfin une descente en température. Source : rapport technique interne, STMicroelectronics.*

Le recuit spike est réalisé sur un équipement de recuit rapide **RTP** (*Rapid Thermal Process*). La plaque de silicium est chauffée de façon radiative sur la face avant par des lampes tungstènes ou halogènes disposées en nid d'abeilles. Elle est mise en rotation afin de garantir une bonne uniformité du procédé. Un balayage de gaz inerte (N_2 ou H_2) permet d'assurer une bonne conductivité thermique dans la chambre de procédé et de refroidir la plaque lors de la descente en température. La mesure de la température est assurée par des pyromètres situés radialement en face arrière, qui alimentent un système de régulation de type PID (commande proportionnelle

intégrale dérivée).

Lors du recuit, la présence de motifs ou de films sur la plaque peut modifier l'émissivité de celle-ci et ainsi perturber les pyromètres. Le contrôle de la température est très critique. Les rampes très agressives (jusqu'à 75°C/s en montée) et l'absence de plateau requièrent un contrôle de température efficace et rapide, afin de limiter les variations du budget thermique de lot à lot et de plaque à plaque. Une telle variation peut entraîner une variation des profondeurs des jonctions, et donc des paramètres électriques du transistor (longueur effective du canal, etc).

1.1.4 Vue globale du procédé CMOS [Quirk et Serda (2000)]

Nous allons décrire d'une façon simplifiée les étapes principales nécessaires à la réalisation d'un circuit en technologie CMOS, en prenant pour exemple la structure illustrée en figure 1.7.

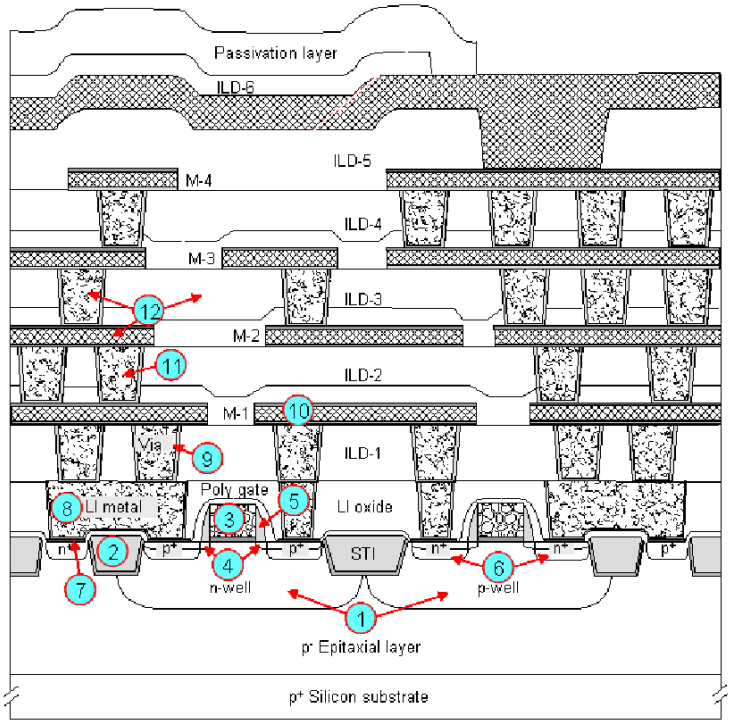

FIGURE 1.7 – *Vue en coupe schématique d'une portion d'un circuit intégré en technologie CMOS. Source : Quirk et Serda (2000).*

La figure 1.7 est une vue en coupe schématique d'une portion d'un circuit intégré réalisé par un procédé CMOS. La structure est formée à partir d'une plaquette de silicium monocristalline de type P sur laquelle est formée une couche épitaxiée de type P. Le transistor de type P est formé dans un caisson (1) de type N (Nwell). Le caisson N est délimité latéralement et en surface par des zones d'isolement en oxyde épais (2), formées par une technique d'isolation par tranchées peu profondes (***S**hallow **T**rench **I**solation*, STI). Le transistor MOS comprend de part et d'autre de la grille (3) des régions de drain et de source (6). De façon classique, cette structure comporte des espaceurs (5) et des zones d'extension à faible niveau de dopage (LDD) (4). D'une

façon équivalente, le transistor MOS à canal N est formé de façon complémentaire dans un caisson (1) de type P (Pwell).

Le procédé de fabrication de la zone active de la structure illustrée en figure 1.7 comprend les étapes principales suivantes : réalisation des tranchées d'isolation, implantation des caissons N & P, formation des grilles des transistors à canal N et P, implantation des régions faiblement dopées (LDD) de type N et P, formation des espaceurs et enfin implantation des régions de drain et de source.

La réalisation des contacts (7) et des interconnexions vise à connecter les transistors entre eux et à l'extérieur du circuit intégré. Les interconnexions s'étalent sur plusieurs niveaux qui sont reliés entre eux par des plots conducteurs verticaux appelés vias (9, 11). Les technologies CMOS actuelles sub $0.13\mu m$ utilisent pas moins de 6 niveaux. Elles utilisent un procédé dit Damascène qui comprend : le dépôt d'une couche d'un matériau diélectrique, la gravure de tranchées dans cette couche, le dépôt d'une couche métallique de cuivre de façon à remplir les tranchées et enfin le polissage mécanochimique (CMP) de l'excès de métal.

1.2 Métrologie de la Dimension Critique

Comme l'a relevé Ruegsegger [Ruegsegger (1998)] dans ses travaux sur la régulation de compensation, le bruit de la mesure nuit à la performance d'une régulation. Il peut entraîner l'amplification de la variabilité de la grandeur régulée au lieu de la minimiser. Dans un environnement de production où le risque n'est pas toléré, nous devons adopter l'outil de métrologie le plus à même à caractériser le processus de fabrication avec précision et exactitude. Dans ce paragraphe, nous nous proposons d'examiner les propriétés de deux méthodes de caractérisation de la dimension critique, à savoir la scattérométrie et la microscopie électronique à balayage (CD-SEM). Cette mesure sera au centre de la régulation développée dans le chapitre 4 (voir la section 4.3.1).

1.2.1 La microscopie électronique à balayage [Salaun (2004)]

Le microscope électronique à balayage est un outil classique qui a toujours accompagné l'industrie du semi-conducteur depuis son envol dans les années 1960. Il consiste à explorer la surface d'une structure de test à l'aide d'un faisceau d'électrons. Les images produites sont en fait formées par les électrons secondaires émis par l'échantillon en cours de son bombardement. La résolution de cette méthode de caractérisation dépend de la tension d'accélération des électrons incidents régissant la longueur d'onde associée. Elle dépend aussi de l'ouverture numérique du faisceau qui determine le diamètre du *spot*. L'équipement CD-SEM (*Scanning Electron Microscopy for Critical Dimension*) est relativement rapide et complètement automatisé. N'exigeant aucune contrainte particulière en ce qui concerne la structure de test, il sert à contrôler les dimensions de lignes, de contact, de via, etc. Il présente toutefois plusieurs inconvénients. Tout d'abord, la mesure est réalisée sous vide, incompatible avec une tendance vers des outils de métrologie compacts et intégrés. Un second inconvénient et non pas le moindre est une forte incertitude de mesure due à plusieurs éléments :

⋈ La dimension critique d'une ligne d'un matériau quelconque (Résine, Poly-Si,...) dépend du profil de la structure, appelé couramment *Side Wall Angle* (SWA). Ce dernier est défini sur la figure 1.8. Si le profil des structures de test demeure constant, le suivi du procédé en production via la mesure en microscopie électronique est pertinent. En revanche, pour un procédé mal au point (un équipement marginal) ou en cours de développement, le profil d'une ligne en poly-silicium à titre d'exemple, pourrait varier. La mesure SEM serait sujette dans ce cas à un biais et les variations constatées ne pourraient être interprétées aisément.

⋈ L'interaction électrons-échantillon conduit à des effets d'accumulation de charges à la surface. Ces charges induisent un champ électrique localisé qui peut modifier la trajectoire des électrons secondaires et déforment de ce fait le signal porteur de l'information [Salaun (2004), Apostolakis (1991)]. Ce phénomène est d'autant plus établi dans le cas des matériaux peu conducteurs, telle la résine photosensible. Dans le cadre de campagnes de répétabilité et de reproductibilité (R&R), les valeurs de la mesure tendent à augmenter proportionnellement à la dose totale d'électrons appliquée [Monahan et Khalessi (1992)].

⋈ Enfin, Il a été démontré que les résines 193nm souffrent d'un rétrécissement significatif ($> 3nm$) sous le bombardement des électrons, indépendamment de la taille de la structure mesurée [Habermas et al. (2002)]. Chih-Ming Ke *et al.* ont évalué l'impact de l'ensemble des paramètres du faisceau électronique (tension d'accélération, courant du faisceau, durée de la mesure,$\cdots$) afin d'en optimiser les consignes pour un rétrécissement minimal [Ke et al. (2002)].

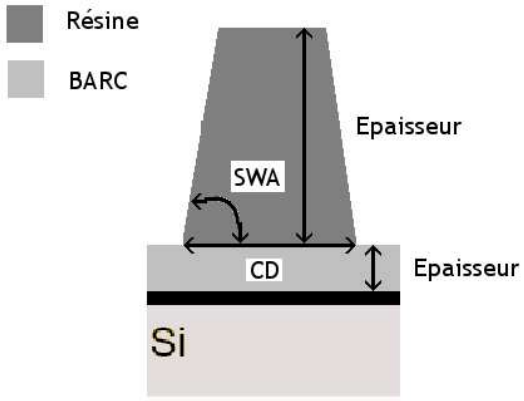

FIGURE 1.8 – *Le modèle choisi en scattérométrie pour décrire la grille à l'étape de photolithographie. Les paramètres flottants, optimisés pendant la mesure, sont : les épaisseurs du BARC et de la résine, Le profil de la ligne de résine, appelé couramment SWA (*Side Wall Angle*) et la dimension critique (CD).*

1.2.2 La scattérométrie [Bao (2003)]

Il s'agit d'une métrologie optique indirecte, fondée sur l'analyse de la lumière diffractée sur une structure de test. Pour qu'elle soit effective et parvenir à extraire les dimensions caractéristiques du motif, deux conditions *sine qua non* doivent être remplies :

1. Toute variation de la géométrie de la structure de test doit engendrer un changement de la réponse scattérométrique,

2. la relation qui lie la géométrie du motif à cette réponse doit être unique.

Cette manière de faire est commune à toute métrologie optique, dite *sensitivity-based*. L'ellipsométrie et la réflectométrie destinées à l'analyse des couches minces en font aussi partie .

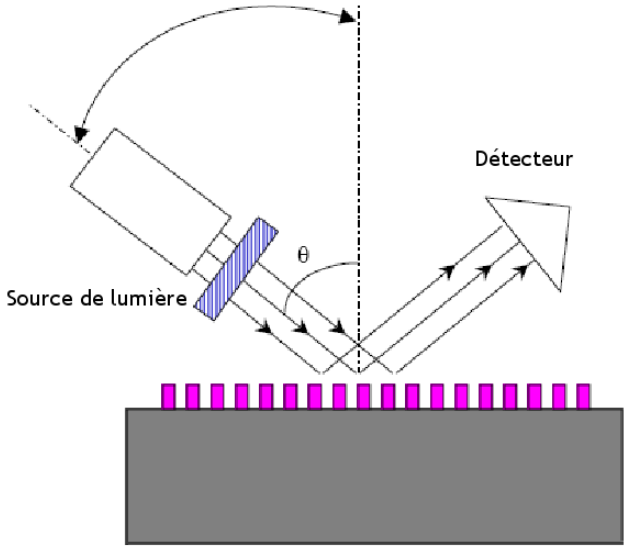

FIGURE 1.9 – *Configuration de la scattérométrie* $2-\theta$ *: Une seule longueur d'onde est utilisée ; l'angle d'incidence est variable. Seul l'ordre* 0 *de la lumière diffractée est collectée par le détecteur.*

D'un point de vue matériel, un scattéromètre est généralement une association de trois éléments : un équipement optique de métrologie, un simulateur électromagnétique de la réponse scattérométrique et des modules de calculs très avancés (plusieurs CPUs en parallèle).

⋈ Selon la technique de mesure adoptée ou retenue, l'outil optique de métrologie doit répondre à un certain cahier des charges. Dans le cas de la scattérométrie à angle-variable ou encore la **scattérométrie** $2-\theta$ [Raymond et al. (1996)], un réflectomètre à angle variable est nécessaire. L'ordre 0 diffracté est en effet mesuré pour différents angles d'incidence, comme le montre la figure 1.9. Quant à la scattérométrie spectroscopique, deployée par ailleurs sur le site de fabrication de STM à Rousset, elle nécessite un ellipsomètre spectroscopique. La mesure de l'ordre 0 est dans ce cas réalisée pour toute une gamme de longueurs d'onde ($250-800$ nm). L'incidence est normale et elle est gardée invariable.

⋈ Le simulateur est chargé de simuler la réponse optique d'un échantillon donné, se substituant ainsi à l'ellipsomètre spectroscopique. Pour ce faire, la structure réelle est tout d'abord modélisée par une figure géométrique, définie par un ensemble de paramètres (angles & longueurs). Une ligne de résine est par exemple souvent modélisée par un trapèze rectangle. La figure 1.8 en donne un aperçu. Par ailleurs, une grande variété de méthodes destinées à simuler ce type de problème existe. Un état de l'art très complet de ces méthodes est proposé par Junwei Bao dans son

manuscrit de thèse [Bao (2003)].

⋊ Les modules de calculs, associés à un algorithme d'optimisation, sont là pour extraire les dimensions de la structure de test et résoudre ce qu'on appelle le problème inverse, celui de retrouver les dimensions de l'échantillon à partir de la réponse optique mesurée. Plusieurs méthodes existent. La méthode de régression non-linéaire et la méthode des bibliothèques (*Library-based Method*) sont parmi les plus connues. La méthode de Régression est une méthode d'optimisation itérative qui va comparer à chaque itération le signal de réponse mesuré au signal simulé. En cas de discordance au sens d'un critère de convergence bien défini, un autre jeu de paramètres est simulé pour une nouvelle itération, jusqu'à minimiser une certaine fonction coût. De multiples algorithmes à optimisation locale ou globale ont été développés pour réaliser cet objectif. Nous pouvons citer la méthode Gauss-Newton, la méthode du recuit simulé, etc. Quant à la méthode des bibliothèques, illustrée en figure 1.10, elle consiste à construire (off-line) une bases de données de signatures scattérométriques. Chaque signature est issue de la simulation électromagnétique d'un jeu de paramètres géométriques unique. Afin de couvrir les variations d'un procédé tel la photolithographie de la grille où la struture de test est définie par 5 à 6 paramètres (épaisseur de la résine, épaisseur du BARC, le profil SWA, la dimension critique, etc), il faudrait quelques centaines de milliers de signatures scattérométriques, et quelques centaines d'heures pour les simuler tous. Une fois en production, le spectre mesuré par ellipsométrie spectroscopique est comparé aux spectres de la bibliothèque ; la réponse qui correspond le mieux sera retenue.

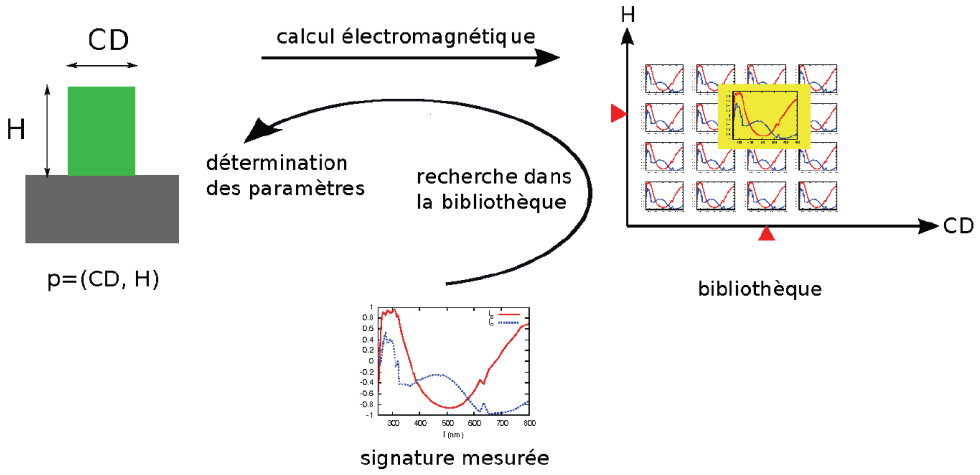

FIGURE 1.10 – *Illustration de la méthode des bibliothèques pour un modèle rectangle, paramétré par une largeur* ***CD*** *et une hauteur* ***H****.*

En pratique, la structure de test mesurée en scattérométrie forme un réseau périodique de motifs (lignes, tranchées, etc) et ceci pour plusieurs raisons. Tout d'abord, la dimension du spot lumineux ($> 35\mu m$) est largement supérieure aux dimensions critiques des technologies avancées sub $0.13\mu m$. Plusieurs lignes doivent être ainsi échantillonnées dans une seule mesure. Outre l'effet cumulatif de structures répétées qui améliore et renforce ***le contenu du signal*** (expression empreintée à Junwei Bao [Bao (2003)]), l'utilisation de réseaux permet d'extraire un profil moyen, insensible aux variations locales de la dimension critique. Enfin, d'un point de vue simulation,

la construction de la réponse optique se trouve plus aisée, dans le cas d'une structure périodique [Bao (2003)]. Du fait que les structures de test en réseaux sont implémentées naturellement dans des lignes de découpe, la scattérométrie est toutefois incapable de réaliser des mesures dans le circuit. La mesure de réseaux à différents pas [2] ou densités permet de contrebalancer cela (voir figure 1.11). Cette mesure multiple est plus représentative du circuit et donne une information complète en ce qui concerne le processus de fabrication.

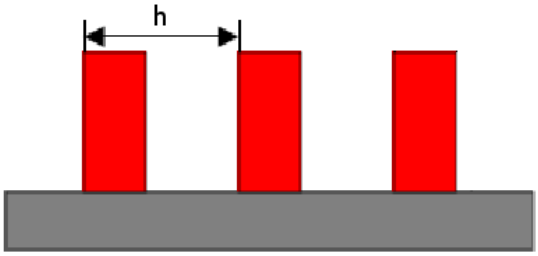

FIGURE 1.11 – *Vue schématique d'un réseau périodique de lignes ; h est le pas du réseau.*

Plusieurs travaux ont été réalisés pour investiguer l'application de la scattérométrie dans l'industrie du semi-conducteur, notamment en photolithographie et en gravure de grille. Sendelback *et al.* [Sendelbach et al. (2004)], à travers une analyse statistique rigoureuse, appelée **TMU** (***T**otal **M**easurement **U**ncertainty*), ont démontré que la mesure des lignes de poly-silicium en scattérométrie est une métrologie viable pour prédire les caractéristiques électriques du transistor. Sendelback *et al.* ainsi que Bao [Bao (2003)] ont aussi souligné son fort potentiel d'amélioration des régulations run-to-run : une forte productivité, une bonne précision, une information complète sur le profil et les couches sous-jacentes et de surcroît il serait techniquement possible de l'intégrer à l'équipement du procédé [Sendelbach et al. (2006)]. Ils ont fait savoir en revanche que la précision de la mesure est subordonnée à la qualité du modèle. Une mauvaise modélisation du motif de test (typiquement, une structure décrite par trop peu de variables flottantes) réduirait d'une manière significative la capacité de cette technique de mesure.

La scattérométrie pourrait potentiellement être utilisée pour qualifier et monitorer un large panel d'opérations de fabrication en microélectronique (grille, Contact, active,...). Sendelback *et al.* se sont focalisés sur la caractérisation de l'exposition et la gravure du polysilicium et ceci pour une technologie de 90 nm [Sendelbach et al. (2004)]. Junwei Bao [Bao (2003)] s'est intéressé dans ses travaux de thèse à plusieurs niveaux dont la gravure des tranchées (STI), la photolithographie et la gravure de la grille et la photolithographie et la gravure des lignes d'interconnexion. Il y compare à chaque fois des mesures réalisées en microscopie électronique à balayage et des mesures en scattérométrie basées sur la méthode des bibliothèques. La corrélation entre les deux moyens de métrologie est forte. Elle est affectée néanmoins par plusieurs facteurs.

⋈ L'erreur de la mesure en scattérométrie est due à l'erreur de modélisation, et aussi à l'incertitude issue de la discrétisation de la bibliothèque.

2. Le pas d'un réseau est la distance qui sépare les motifs.

⋊ En ce qui concerne la mesure SEM, le profil (SWA) semble avoir un impact important sur les moyens logiciels embarqués dans l'équipement (algorithmes de détection des contours). Toute variation de l'angle SWA pourrait induire un CD biaisé. Par ailleurs, le CD-SEM est conçu pour mesurer des motifs spécifiques (lignes, contacts, etc), ce qui rend la mesure sensible aux variations locales du CD. Au contraire, la mesure en scattérométrie rend compte d'un profil moyen.

Afin de comparer la précision des deux systèmes de mesure, Junwei Bao [Bao (2003)] a mené une étude de répétabilité et de reproductibilité [3]. Pour une structure en poly-silicium en technologie $0.18\mu m$, l'incertitude en scatérométrie a été 4 fois moins importante $(3Sigma = 0.75nm)$ que celle en microscopie $(3Sigma = 3.28nm)$. Dans cette même campagne de mesures, l'incertitude de la mesure en scattérométrie basée sur la méthode de régression a été évaluée, elle est de 0.25 nm. Pour toutes ces raisons, nous avons choisi la scattérométrie, comme la métrologie la plus appropriée pour la mise en production de la boucle de compensation entre la mesure de la dimension critique (CD) de la grille et l'implantation des poches (voir chapitre 4).

1.3 Le test

A la fin du processus de fabrication et en amont du découpage et la mise en boîtier des puces, les plaquettes sont soumises à un test électrique. Deux types de test sont réalisés : le test paramétrique (*Parametric Test*, **PT**) et le tri électrique des plaquettes (*Electrical Wafer Sort* **EWS**).

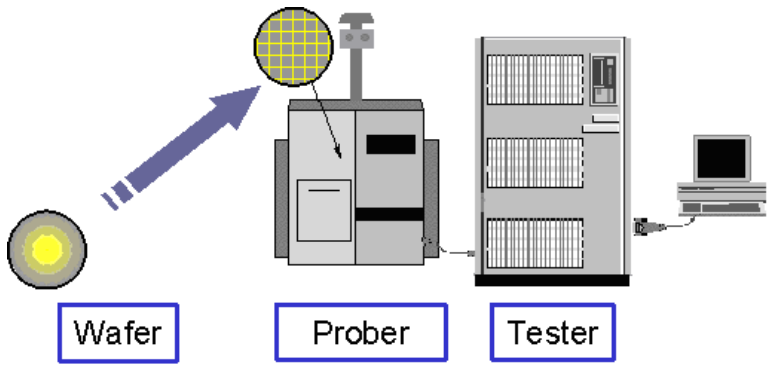

Figure 1.12 – *L'environnement du test paramétrique*

1.3.1 Le test paramétrique

Le test paramétrique est la dernière étape réalisée sur le site de fabrication. Il s'agit d'un outil d'analyse physique et de diagnostic qui permet de qualifier les étapes du processus de fabrication et, par conséquent, d'identifier la cause en cas de défaillance. Grâce à l'action combinée de deux machines le *Prober* (Sonde) et le *Tester* (Testeur), illustrées sur la figure 1.12, la procédure du test consiste en pratique

3. Il s'agit d'une analyse de variance dont l'objectif est de quantifier la contribution de plusieurs facteurs pouvant contribuer à la variabilité d'une mesure, et notamment la variabilité répétition (la précision de l'appareil de mesure), la variabilité temporelle, la variabilité humaine (l'effet opérateur) et la variabilité du processus de fabrication.

à poser des pointes sur les plots d'une structure de test (*Test Elementary Group*) et d'appliquer un signal pour en extraire une réponse. L'ensemble des réponses est divisé en 2 catégories : critique et complémentaire. Un paramètre est critique s'il est jugé garant des fonctionnalités du circuit. Citons à titre d'exemple les tensions de seuil, les courants de fuite, etc. Un paramètre est complémentaire s'il constitue simplement un supplément d'informations qui permettrait d'affiner un diagnostic pour reconnaître un défaut ou une dérive du processus de fabrication.

Les structures de test sont logées dans les chemins de découpe. Afin qu'elles soient représentatives du produit, elles sont conçues de formes et de natures très diverses : transistors à multiples dimensions, lignes résistives, diodes, peigne inter-digités, etc. Dans le chapitre 4, nous ferons souvent référence à trois paramètres critiques : le courant de saturation I_{on}, le courant de fuite I_{off} et la tension de seuil V_{th} des transistors MOS de longueur de grille nominale $L_{poly} = 0.13\mu$m, dits transistors courts [Mathieu (2001)]. La structure de test correspondante (voir figure 1.13) est un transistor court élémentaire, avec une grille isolée en poly-silicium. Nous entendons par isolé que la structure n'a pas de lignes en poly-Si dans son voisinage. Nous verrons plus loin dans la section 4.6 du chapitre 4 que cet aspect isolé aura une importance particulière dans la conception de la régulation de compensation Gravure-Implantation des poches.

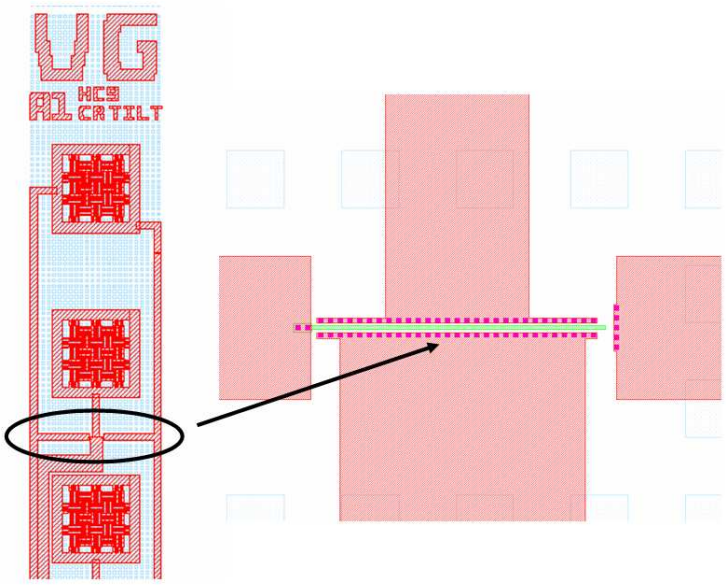

FIGURE 1.13 – *Une structure de test destinée à mesurer des paramètres critiques (I_{on}, I_{off} et V_{th}) de transistors courts. Le barreau vert est la grille en polysilicium dont les dimensions sont $L_{poly} = 0.13\mu m$ et $W = 10\mu m$.*

1.3.2 Le test EWS

Le tri électrique des plaquettes consiste en une batterie de tests ayant pour but de faire le tri entre les puces fonctionnelles et les puces défectueuses.

Le test structurel

Il consiste à vérifier directement l'existence ou non de certains défauts (court-circuit, circuit ouvert,...) parmi les plus probables, sans se préoccuper des fonctionnalités du circuit.

Le test fonctionnel

L'objet de cette méthode est de vérifier que le circuit fonctionne correctement et que toutes les spécifications sont réalisées. Des séquences de stimuli numériques sont envoyées ; on compare alors la réponse obtenue en sortie avec celle que l'on devrait avoir selon le cahier des charges du circuit.

Le test IDDQ (*Quiescent Chip Current*)

Afin de comprendre l'intérêt du test IDDQ, il est nécessaire de rappeler l'architecture d'un circuit CMOS. Un circuit CMOS est constitué d'un réseau de transistors NMOS en pull-down et d'un réseau de transistors PMOS en pull-up (voir figure 1.14). En régime statique et dans le cas dépourvu de défauts, un seul des transistors conduit. Le courant de repos dans un circuit CMOS est ainsi très faible. Une faible consommation est en effet un des atouts importants de la technologie CMOS.

En présence d'un quelconque défaut (particule déposée, court-circuit, etc), ce courant peut augmenter de plusieurs ordres de grandeurs . C'est ainsi qu'en observant le courant aux bornes de l'alimentation, que nous pouvons détecter si un circuit présente un défaut ou non. Au delà des erreurs dues au process qui vont causer une forte consommation en courant, le test **IDDQ** rend compte aussi d'autres problèmes et notamment la fuite de grille.

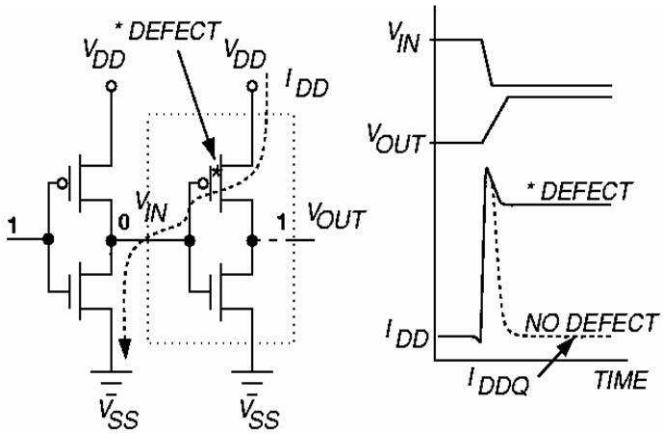

FIGURE 1.14 – *Formation d'un fort courant entre l'alimentation et la masse à cause d'un défaut au niveau de l'oxyde de grille*

En pratique, le test IDDQ permet d'écarter les pièces ne respectant pas les critères de consommation du circuit. Pour les produits destinés aux applications mobiles, comme les téléphones portables, qui nécessitent une très faible consommation, ces critères ont en effet une importance capitale. Par ailleurs, ce test est utilisé pour rejeter les pièces porteuses de défauts physiques. Ces défauts peuvent ne pas affecter

la fonctionnalité du circuit, mais générer des problèmes de fiabilité dans la vie du produit.

2 Analyse de Variabilité des Paramètres Électriques : Application à une technologie CMOS logique 0.13μm

Sommaire

Le test paramétrique (PT) de technologies logiques avancées rend compte d'une forte variabilité de certains paramètres critiques des transistors MOS, notamment le courant de saturation I_{on}, le courant de fuite I_{off} et la tension de seuil V_{th}. La figure 2.1 appuie ce constat. Nous y observons une forte variabilité lot à lot du courant de saturation des transistors courts NMOS. La dispersion paramétrique des produits logiques avancés fait partie des chiffres officiels communiqués aux clients. Des valeurs élevées véhiculent une image défavorable de la fabrication . Au delà de cette image, une forte variabilité paramétrique pourrait engendrer une dégradation du rendement de certains produits sensibles et est incompatible avec des spécifications de plus en plus serrées [1].

L'objectif de ce chapitre est d'identifier les principales caractéristiques physiques (épaisseur de l'oxyde de grille, l'hauteur de marche des tranchées d'isolation, etc), mesurées au fur et à mesure du processus de fabrication, et qui sont à l'origine des variations paramétriques. Nous avons réalisé une importante campagne de mesures qui s'inscrit dans cette finalité. Ces paramètres physiques que nous avons considérés répondent à deux critères : ils sont mesurés en ligne, et ils sont fortement suspectés [2] d'avoir une influence majeure sur les paramètres électriques des transitors MOS. Avant de décrire cette campagne de mesures et en détailler les résultats, nous nous

1. Ces spécifications sont fixées par l'**ITRS** (**I**nternational **T**echnology **R**oadmap for **S**emiconductors), **www.itrs.net**

2. Par les experts qui sont en l'occurrence les ingénieurs *Device*.

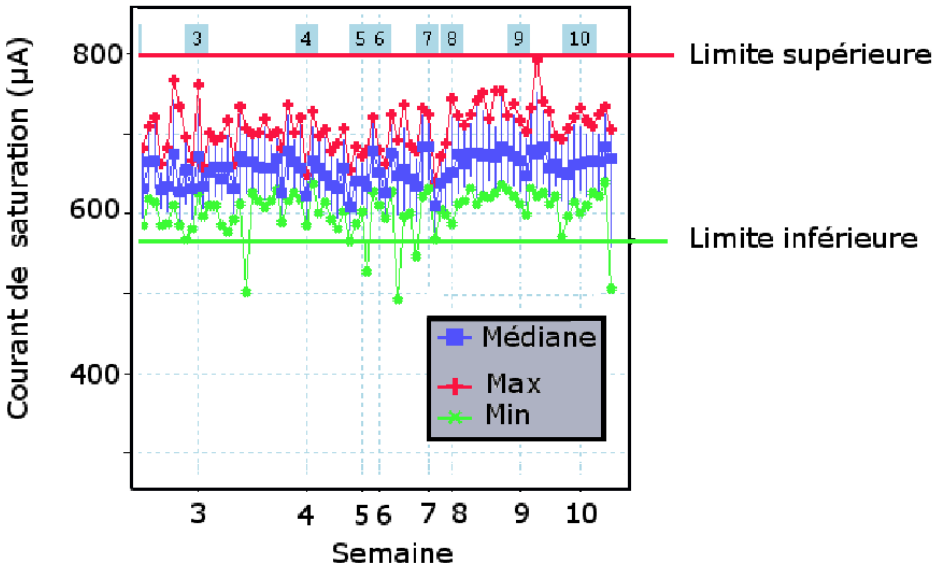

FIGURE 2.1 – *La tendance du courant de saturation I_{on} entre la semaine 3 et la semaine 10 de l'année 2006. la courbe bleue correspond aux moyennes des lots, tandis que les courbes rouge et verte correspondent respectivement aux valeurs individuelles maximales et minimales enregistrées par lot. A noter que toutes les plaques (25) sont mesurées, avec 5 points de mesure par plaque.*

sommes attelés, dans la première partie de ce chapitre, à présenter la méthodologie statistique utilisée et à en justifier le choix. Il s'agit d'une méthodologie double qui associe une régression PLS (*Partial Least Square*) à une analyse de variance hiérarchique [Herk et al. (2005)].

2.1 Analyse de variabilité : Méthodologie et outils

Au début de cette thèse, nous avons opté pour une campagne de mesures, dont l'objectif était de spécifier les sources de variabilité paramétrique, et de quantifier leur impact d'une façon précise. A partir d'un premier état de l'art des travaux qui se sont intéressés à ce même sujet et en prenant en compte les moyens matériels dont nous disposons, nous avons établi une liste de paramètres physiques qui seraient potentiellement de grande influence sur la variabilité des caractéristiques critiques (I_{on}, I_{off}, V_{th}). Suite à cette campagne et au collecte des données, nous avons appliqué une méthodologie d'analyse générique, inspirée des travaux d'un collectif d'ingénieurs spécialistes sur le site de Crolles [Herk et al. (2005)]. Cette méthodologie se déroule en 3 points :

i. Tout d'abord, expliquer des variables électriques (I_{on}, I_{off} & V_{th}) par un ensemble de caractéristiques physiques mesurées en ligne (longueur de grille, épaisseur de l'oxyde,...). Une régression de type PLS répond amplement à cet objectif. Ce choix sera justifié dans la section suivante,

ii. identifier ensuite les briques de process associées aux sources de variabilité les plus importantes. Nous nous sommes appuyés pour cela sur les valeurs de l'indice **VIP** (***V**ariable **I**mportance in **P**rojection*). Cet indice sera introduit plus loin dans ce chapitre,

iii. enfin, réaliser une analyse de variance pour dissocier la composante spatiale (intra-plaque) des composantes temporelles, des briques de process identifiées au préalable.

2.1.1 La régression PLS [Tenenhaus (1998)]

Afin de mieux comprendre la régression PLS, nous pouvons nous appuyer sur la régression linéaire multiple. Soit Y la variable réponse ou à expliquer et $\mathbf{X} = \{X1, ...XM\}$ l'ensemble de variables explicatives. La régression multiple consiste à identifier le vecteur $\Theta = \begin{bmatrix}\theta_1 & \theta_2 & \cdots & \theta_M\end{bmatrix}'$ tel que :

$$y_i = \sum_{j=1}^{M} \theta_j \, x_{i,j} + \omega_i \tag{2.1}$$

où ω_i sont les erreurs supposées indépendantes, identiquement distribuées (i.i.d), et de moyenne nulle. Suite à la collecte d'un nombre n de réalisations ($n > M$), nous pouvons écrire matriciellement :

$$Y = \mathbf{X} \begin{bmatrix}\theta_0 & \theta_1 & \cdots & \theta_M\end{bmatrix}' + \Omega \tag{2.2}$$

Avec Y et Ω sont des vecteurs à n dimensions et $\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,M} \\ x_{2,1} & \cdots & x_{2,M} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,M} \end{bmatrix}$ est une matrice $n \times M$.

La solution au problème de la régression multiple est donnée par la solution des moindres carrés [3], formulée ci-dessous (équation 2.3). Une interprétation utile

3. détaillée davantage dans le chapitre 5 de ce manuscrit

de cette formulation est l'interprétation géométrique. $\hat{\Theta}$ correspond en effet à une projection orthogonale du vecteur $\mathbf{Y}$ sur l'hyperplan formé par les variables X_i. Dans le cas de variables X_i colinéaires, ie fortement corrélées entre elles (la matrice $X'X$ est mal-conditionnée et $det(\mathbf{X}'\mathbf{X}) \simeq 0$), l'hyperplan, ainsi défini est instable ce qui engendre une inflation de l'écart-type des θ_i. Dans certains cas de multicolinéarité, la variance des θ_i est telle que les paramètres influents pourraient être déclarés insignifiants, ou encore une variable donnée apparaît dans l'équation du modèle avec un coefficient positif alors que son influence sur Y est négative.

$$\Theta = \begin{bmatrix}\theta_0 & \theta_1 & \cdots & \theta_M\end{bmatrix}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \tag{2.3}$$

Lorsque les variables explicatives sont nombreuses et en partie colinéaires, la régression multiple est alors inappropriée et inefficace. C'est là que la régression PLS prends tout son intérêt. La Régression PLS remplace l'espace initial des variables explicatives $\mathbf{X} = \{X1, ...XM\}$ par un espace de faible dimension, constitué d'un petit nombre de variables appelées variables latentes ou facteurs. Ces facteurs seront les nouvelles variables explicatives d'un modèle de régression linéaire classique. Ils sont orthogonaux (non corrélés), et sont des combinaisons linéaires des variables explicatives initiales.

D'autres alternatives à la régression multiple qui répondent à l'objectif d'expliquer une variable cible par des variables explicatives (régresseurs) existent. Je citerai notamment les arbres de décision, les réseaux de neurones, ou encore la technique MinCorr [4] [Casali et al. (2007)]. Bien que les arbres de décision ne soient pas perturbés par la colinéarité des régresseurs [Tufféry (2005)], elles présentent un inconvénient majeur pour la suite de notre étude : elles s'appuient sur un modèle non paramétrique et ne quantifient pas dès lors l'influence des X_i sur la réponse Y. L'utilisation des réseaux de neurones ne peut être justifiée sans passer par la case des modèles linéaires et nécessite en outre, au même titre que les arbres de décision, un nombre important d'observations. Nous avons ainsi retenu et mis en application la régression PLS.

Les variables latentes T_i de la régression PLS sont construites l'une après l'autre de façon itérative. La première variable latente $T_1 = \mathbf{X}w_1$ maximise $cov(T_1; Y)$; la variable $T_2 = \mathbf{X}w_2$ maximise $cov(T_2; Y)$ sous la contrainte $T_2 \perp T_1$, et ainsi de suite. A chaque étape h, on construit la régression de Y sur les composantes $\{T_1, ...T_h\}$. L'équation de la régression PLS est obtenue en exprimant cette équation en fonction des variables d'origine, comme le montre les équations suivantes.

$$Y = c_1\,T_1 + c_2\,T_2 + \cdots + c_h\,T_h + \Omega \tag{2.4}$$
$$= c_1\,\mathbf{X}w_1 + c_2\,\mathbf{X}w_2 + \cdots + c_h\mathbf{X}w_h + \Omega \tag{2.5}$$
$$= \mathbf{X}\Theta + \Omega \tag{2.6}$$

avec $\Theta = c_1\,w_1 + c_2\,w_2 + \cdots + c_h\,w_h$.

Le nombre h des composantes T_i est déterminé par validation croisée. Pour chaque nouvelle composante T_h, nous calculons l'indice de Stone-Geisser Q_h^2(voir équation 2.7) à partir du RSS (Residual Sum of Squares) et du PRESS (*Predicted Error Sum of*

4. **MinCorr** est en cours de développement par Christian Ernst, Groupe SFL, Centre de Microélectronique de Provence, Gardanne

Squares).

$$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}} \tag{2.7}$$

Avant de définir les critères RSS et PRESS, nous introduisons deux alternatives pour estimer la prédiction $\hat{Y} = Y - \Omega$ (Equation 2.4). Pour chaque échantillon i, la prédiction $\hat{y}_{hi}$ de y_i est obtenue en utilisant toutes les observations (au nombre de n) pour réaliser la régression PLS. Au contraire, la prédiction $\hat{y}_{h(-i)}$ de y_i est obtenue en excluant l'individu i des calculs du modèle. Les critères RSS et PRESS sont alors formulés par :

$$RSS_h = \sum_{i=1}^{n} (\hat{y}_{hi} - y_i)^2 \tag{2.8}$$

$$PRESS_h = \sum_{i=1}^{n} (\hat{y}_{h(-i)} - y_i)^2 \tag{2.9}$$

Dans le logiciel **SIMCA-P**[5], application centrée sur la régression linéaire PLS et qui en fournit une utilisation facile et une aide à l'interprétation, une composante T_h est déclarée significative si $Q_h^2 \geq 0.0975$. Un autre indicateur de la qualité du modèle est le paramètre R^2. Le R_h^2, donné par l'équation 2.10, représente la part expliquée par la régression de la somme des carrés des écarts à la moyenne. Enfin, pour déterminer les variables explicatives à exclure du modèle, nous pouvons nous appuyer sur un dernier indice : le VIP. L'indice VIP représente la contribution de chaque variable explicative à l'ajustement du modèle PLS. Si une variable explicative a un coefficient de régression relativement faible en valeur absolue et si son VIP est faible aussi (moins de 0.8) alors elle peut être exclue du modèle.

$$R_h^2 = \frac{\sum_{i=1}^{n} (\hat{y}_{hi} - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2.10}$$

2.1.2 L'analyse de variance

L'analyse de variance (ANOVA) est une technique de répartition de la variation totale des mesures observées lors d'une campagne d'expériences, entre les différentes sources de variation auxquelles elle peut être attribuée. Cette technique permet de vérifier si la variation due à une composante particulière quelconque est significative. L'analyse de variance se fait selon un modèle sous-jacent qui exprime la réponse comme somme de différents effets. L'équation 2.11 est un modèle d'analyse de variance à un seul facteur (l'exemple des chambres de gravure), qui peut être étendu à plusieurs sources de variabilité.

$$y_{ij} = \mu + \alpha_i + \omega_{ij} \tag{2.11}$$

Avec
- y_{ij} : La j-ème mesure ou observation du groupe i (exemple : L_{poly}),
- μ : la moyenne des observations,
- α_i : l'effet associé à l'appartenance au groupe i (chambre de gravure i),
- ω_{ij} : l'erreur aléatoire associée à la (ij)-ème observation, $\omega_{ij} \sim \textbf{iid}\, \mathcal{N}(0; \sigma^2)$.

Les effets pris en compte dans le modèle peuvent être des effets fixes ou des effets aléatoires. Les effets de deux équipements de photolithographie bien définis sont

5. http ://www.umetrics.com/

fixes, puisque l'on peut raisonnablement supposer que chaque équipement a un effet déterminé. En revanche, les effets associés aux lots, et aux plaques de la même façon, sont des effets aléatoires. Les lots appartiennent en effet à un échantillon choisi au hasard dans un plus grand ensemble de provenances. Les résultats sont transposés par contre à l'ensemble des échantillons. Une analyse de variance combinant des effets fixes et des effets aléatoires est appelée une analyse de variance à effets mixtes. Il est important enfin de noter que l'estimation d'un effet fixe revient à quantifier l'écart de la réponse y, dû à cet effet, alors que celle d'un effet aléatoire revient à estimer la variance de cette catégorie d'effet.

Dans le cadre de nos travaux, nous avons besoin d'une extension de l'anova classique : l'anova hiérarchique ou emboîtée. Elle est destinée à traiter le cas où un critère de classification (une catégorie d'effet) est subdivisé aléatoirement en deux ou plusieurs sous-groupes. Afin d'analyser les sources de variation d'un paramètre donné, tel le courant de saturation I_{on} ou la longueur L_{poly}, nous irons chercher un certain nombre de lots de production. Dans chacun de ces lots, nous testerons plusieurs plaques ($\leq$ 25) et, pour chacune de ces plaques, plusieurs structures réparties sur toute la superficie sont mesurées. Nous aurons donc bâti un modèle anova hiérarchique à trois niveaux emboîtés, tous aléatoires (voir équation 2.12).

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \delta_{k(ij)} + \omega_{ijk} \tag{2.12}$$

Avec
- y_{ijk} : La mesure du k-ème site de la plaque j appartenant au lot i,
- μ : la moyenne des observations,
- α_i : l'effet lot, $\alpha_i \sim$ **iid** $\mathcal{N}(0; \sigma^2_{lot})$,
- $\beta_{j(i)}$: l'effet plaque ou inter-plaque, $\beta_{j(i)} \sim$ **iid** $\mathcal{N}(0; \sigma^2_{plaque})$,
- $\delta_{k(ij)}$: l'effet site ou intra-plaque, $\delta_{k(ij)} \sim$ **iid** $\mathcal{N}(0; \sigma^2_{site})$,
- ω_{ijk} : les résidus, $\omega_{ijk} \sim$ **iid** $\mathcal{N}(0; \sigma^2)$.

Notons que dans la littérature où j'ai pu recenser quelques études de cas en fabrication de composants, que ce soit des académiques ou des chercheurs en milieu industriel, ils ont opté souvent pour un modèle où l'erreur résiduelle et l'effet site sont confondus [Littell et al. (1996), Czitrom et Spagon (1987), Drain (1997)]. Un seul terme y est alors associé. Ce choix s'expliquerait par l'hypothèse suivante : Indépendamment des causes de variation, la variabilité de la réponse y est transposée totalement en trois composantes : lot-à-lot, plaque-à-plaque et intra-plaque[6]. Par ailleurs, une analyse de variance basée sur le modèle 2.12 pourrait impliquer des temps de calcul infiniment longs dans certains cas[7]. L'analyse ne pourrait alors aboutir sans abandonner la séparation entre l'effet site (intra-plaque) et les résidus.

D'un point de vue logiciel, toutes les analyses de variances ont été réalisées sous SAS. SAS dispose en effet d'une commande très utile, facile d'utilisation, qui s'appelle **PROC MIXED**. Pour toute quête d'information concernant la mise en oeuvre de cette commande, je renvoie le lecteur à l'ouvrage très complet de Litell *et al.* [Littell et al. (1996)]. Nous avons aussi utilisé la commande **LME** du logiciel libre **R**.

6. Seul le bruit de la mesure ne peut être contenu dans ces dernières. Dans le cas d'une campagne de mesures sans répétition, cette incertitude est par ailleurs inaccessible.

7. Nous avons rencontré ce problème lors de l'étude de la variabilité des mesures **IDDQ**. Dans ce cas le nombre de sites mesurés par plaque était de l'ordre de quelques centaines de points

2.2 Analyse de variabilité : Application

L'objet de ce paragraphe est de décrire les résultats de la campagne de mesures lancée pour spécifier les sources de la variabilité paramétrique et quantifier leur impact d'une façon précise.

2.2.1 Choix des paramètres physiques pris en compte

Les paramètres physiques inclus dans cette campagne sont par ordre chronologique de la mesure :

⋊ La hauteur de marche STI[8] (e_m). Cette mesure est réalisée en microscopie à force atomique (AFM). La durée très longue d'une mesure individuelle en AFM nous a contraint à réduire le nombre de plaques à mesurer. Notons de surcroît sa nature très bruitée qui réduit encore davantage sa significativité.

⋊ l'épaisseur de l'oxyde (t_{ox}). La mesure est réalisée dans ce cas au moyen d'un éllipsomètre de production,

⋊ finalement, la longueur de grille en poly-Si (L_{poly}), son profil (SWA) et l'épaisseur de la couche de poly-Si t_{poly}. Tous les trois sont mesurés en-ligne sur un équipement de scattéromètrie,

Nous avons représenté certaines de ces variables sur le schéma ci-dessous (figure 2.2). Afin de réduire le nombre d'éventuelles sources de variabilité, nous avons veillé à réaliser la mesure sur un unique équipement de métrologie pour chacun des paramètres susmentionnés.

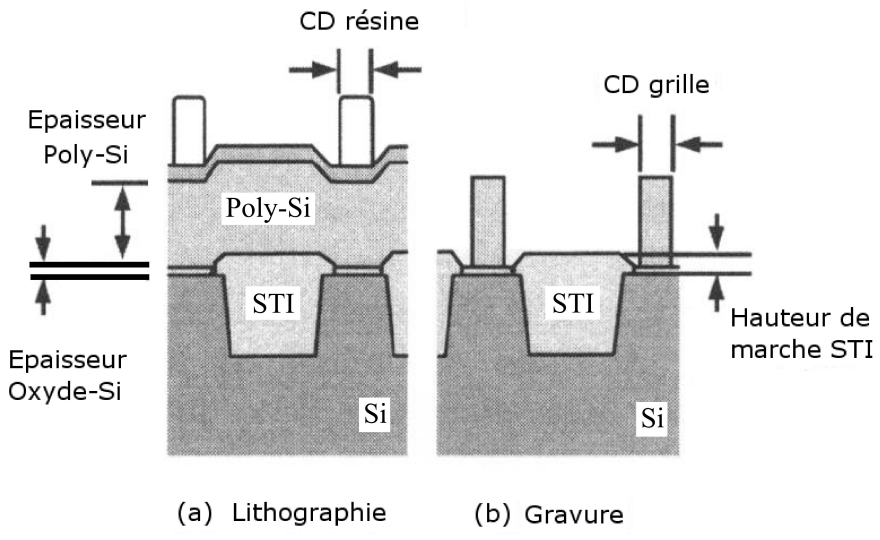

Figure 2.2 – *Représentation schématique des différentes caractéristiques physiques mesurées dans le cadre de la campagne de mesures.*

Cette campagne de mesures consistait à mesurer la totalité des plaques de 5 lots de production d'un même produit. Le nombre de sites mesurés sur chacune des plaques est relativement important comparé à celui des sites réalisés en production.

8. *Shallow Trench Isolation*

Il est égal à 15 pour L_{poly} et à 14 pour le t_{ox}. Comme nous l'avons indiqué précédemment, seule la hauteur de marche e_m fait exception à cette règle. Le nombre de plaques mesurées est seulement de 4 et le nombre de points par plaque est de 16.

Rappelons que toutes ces mesures sont réalisées sur des structures adaptées au paramètre à mesurer, qui sont logées dans les lignes de découpe. Que l'on veuille mesurer le L_{poly} ou le t_{ox}, il s'agit rarement de la même structure à tester. De la même façon, les structures de test paramétrique (PT) résident dans des lignes de découpes et elles sont différentes de celles mesurées en ligne. A partir de là, la régression PLS, destinée à corréler les variables à expliquer (I_{on}, I_{off} et V_{th}) aux variables explicatives (L_{poly}, t_{ox}, t_{poly}, SWA, e_m) sera effectuée au niveau plaque, ie : les données qualitatives sont des moyennes par plaque.

Variable latente	R2X	R2X(cum)	R2Y	R2Y(cum)	Q2	Q2(cum)
1	0.463	0.463	0.412	0.412	0.405	0.405
2	0.266	0.729	0.164	0.576	0.27	0.566
3	0.0929	0.822	0.0206	0.596	0.0265	0.577

TABLE 2.1 – *Tableau récapitulatif des parts de variance expliquée par un modèle à 3 variables latentes.*

2.3 RÉSULTATS

2.3.1 La régression PLS

La régression PLS fournit un tableau 2.1 permettant de vérifier comment chaque variable latente extraite rend compte de la variabilité des facteurs X_i et des réponses Y_i au sein du modèle. Nous pouvons constater ainsi que les trois premières composantes représentent près de 82% de la variabilité des X_i (L_{poly}, t_{ox}, SWA, t_{poly}, e_m) et expliquent 60% de la variabilité de l'ensemble des réponses (I_{on}, I_{off} et V_{th}). Le tableau 2.2 fournit les couples (R^2, Q^2) relatifs aux variables de réponses individuelles. La régression PLS explique, par d'exemple, 75% des variations du courant I_{off}. Notons aussi la forte valeur de l'indice Q^2 correspondant qui montre que le modèle a des bonnes capacités de prédiction. Les résultats sont néanmoins moins bons quand il s'agit du courant de saturation et de la tension de seuil ($R^2 \simeq Q^2 \simeq 50\%$).

Réponse	R^2	Q^2
I_{off}	75%	74%
I_{on}	53%	51%
V_{th}	50%	48%

TABLE 2.2 – *Tableau récapitulatif des indices R^2 et Q^2 des différentes variables de réponse.*

La figure 2.3 est le graphe le plus important d'une régression PLS. La proximité des projections des variables I_{on} et I_{off} nous indique que les deux réponses dépendent des mêmes variables explicatives. La projection de la longueur de grille L_{poly} laisse penser qu'elle est de loin le prédicteur le plus important pour l'ajustement des modèles du couple (I_{on}, I_{off}). En ce qui concerne la tension de seuil V_{th}, L_{poly} est moins influent et ceci aux dépens du profil de la grille SWA et de l'épaisseur de l'oxyde t_{ox}.

Rajoutons enfin que V_{th} est corrélée positivement au couple (L_{poly},SWA) et négativement au t_{ox}.

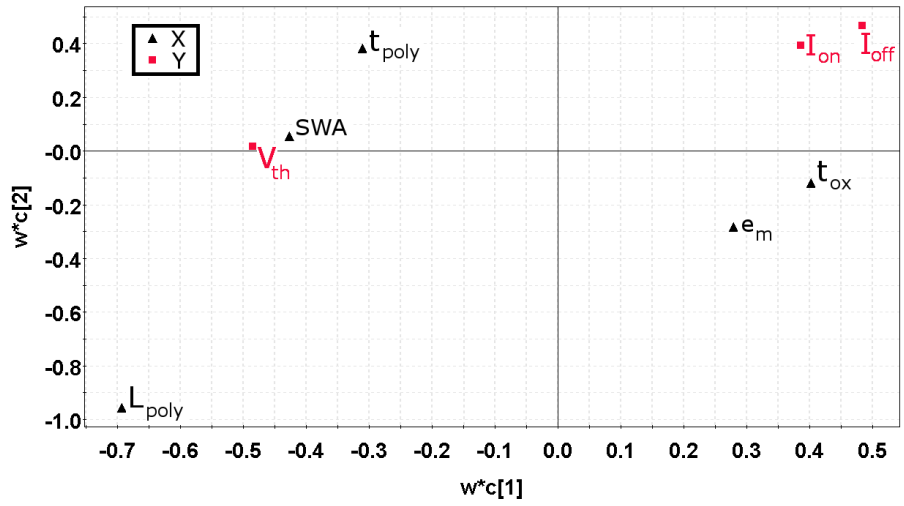

FIGURE 2.3 – *Diagramme des corrélations des variables explicatives et des réponses.*

Le diagramme des résidus normalisés (ou droite de Henry) permet de vérifier que les résidus suivent une loi gaussienne. Il permet en outre de détecter des situations d'anormalité (auto-corrélation, effets quadratiques non pris en compte,$\cdots$). Le diagramme de normalité du modèle V_{th} (figure 2.4) correspond au diagramme idéal, les points sont distribués selon une droite. Notons que lors d'une étude préalable, nous avons identifié trois points atypiques,que nous avons naturellement exclus. Les droites de Henry relatives au couple (I_{on}, I_{off}) témoignent d'une manière similaire de la normalité des résidus.

L'examen du diagramme des indices VIP (Tableau 2.4) et des coefficients de régression 2.3 nous donne le moyen d'identifier les facteurs non influents, à exclure du modèle, à savoir la hauteur de marche dont l'indice VIP est inférieur à 0.8 et l'épaisseur du poly-Si. La suppression de ces facteurs change très peu les résultats détaillés dans les paragraphes précédents. Par ailleurs, la sensibilité du courant de saturation aux variations de L_{poly} est estimée à $-5.7\mu\, A/nm$, un résultat plutôt rassurant car proche de la valeur fournie par le plan d'expérience que nous avons réalisé et qui sera décrit dans le chapitre 4.

En conclusion, nous constatons que la variabilité temporelle du couple (I_{on}, I_{off}), à savoir lot à lot et plaque à plaque, est expliquée pour une grande part par la variabilité de **la longueur de grille** L_{poly}. Il s'agit en effet d'une source de variation de premier ordre. L'épaisseur de l'oxyde de grille t_{ox} et le profil SWA contribuent d'une manière significative à la variabilité paramétrique et notamment à celle de la tension de seuil V_{th}, ils arrivent néanmoins en second plan. A partir de là, notre prochaine étape est d'estimer les proportions de chacune des composantes lot à lot, plaque à plaque et intra-plaque des variations paramétriques et de L_{poly}.

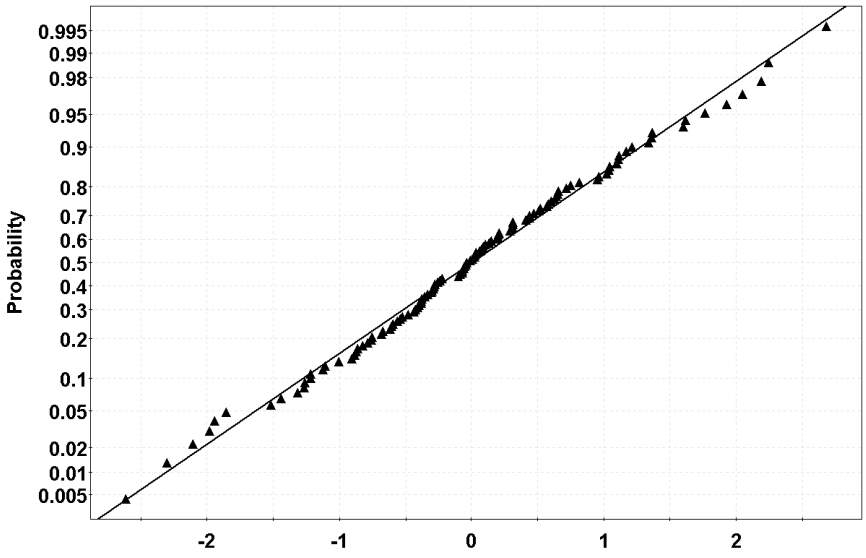

FIGURE 2.4 – *Droite de Henry relatifs aux résidus du modèle de V_{th}.*

Facteurs	**Coeff(I_{on})**	**Coeff(I_{off})**	**Coeff(V_{th})**
e_m	2.1E-01	2.7E-02	8.5E-02
t_{ox}	-3.2E-02	1.2E-01	-3.5E-01
L_{poly}	-6.9E-01	-7.9E-01	2.7E-01
t_{poly}	8.4E-02	3.3E-02	2.1E-01
SWA	-9.1E-02	-1.8E-01	2.6E-01

TABLE 2.3 – *Tableau récapitulatif des coefficients de régression réduits centrés du triplet (I_{on}, I_{off} et V_{th}).*

2.3.2 L'analyse de variance

L'impact de L_{poly} sur la dispersion paramétrique temporelle, somme des variations lot à lot et plaque à plaque, s'est avéré d'une importance majeure. Nous nous proposons dans ce paragraphe de réaliser une décomposition de la variation au moyen d'une analyse de variance hiérarchique. Elle vise à quantifier les effets lot et plaque des paramètres électriques critiques (I_{on}, I_{off} et V_{th}) et de la longueur de grille L_{poly}. Le modèle sous-jacent à cette étude correspond à l'équation 2.12.

Les figures 2.5, 2.6 et 2.7 montrent que la dispersion paramétrique intra-plaque est de loin la plus importante. Elle est de l'ordre de 50% de la variance totale. Le même constat est valable pour la variance intra-plaque de la longueur de grille L_{poly} et pointe ainsi du doigt un éventuel rapport de cause à effet entre la composante spatiale de L_{poly} et la composante spatiale paramétrique. Nous allons étudier cette question dans les prochains paragraphes.

Comparée à l'effet plaque, l'effet lot est plus ou moins important selon le para-

Facteurs	VIP
L_{poly}	1.52
t_{ox}	0.84
SWA	0.84
t_{poly}	0.83
e_m	0.77

TABLE 2.4 – *Tableau récapitulatif des indices VIP de l'ensemble des facteurs.*

mètre électrique étudié. Il faudrait tout de même rappeler que ces calculs sont basés sur les données de 5 lots, un nombre trop faible pour avoir une estimation précise de cet effet. Le même calcul réalisé avec les mesures d'un nombre important de lots [9] nous révèle une composante lot à lot significative et plus importante que la composante plaque à plaque. De la même façon, l'effet lot dans les variations de L_{poly} est plus considérable que l'effet plaque (figure 2.8).

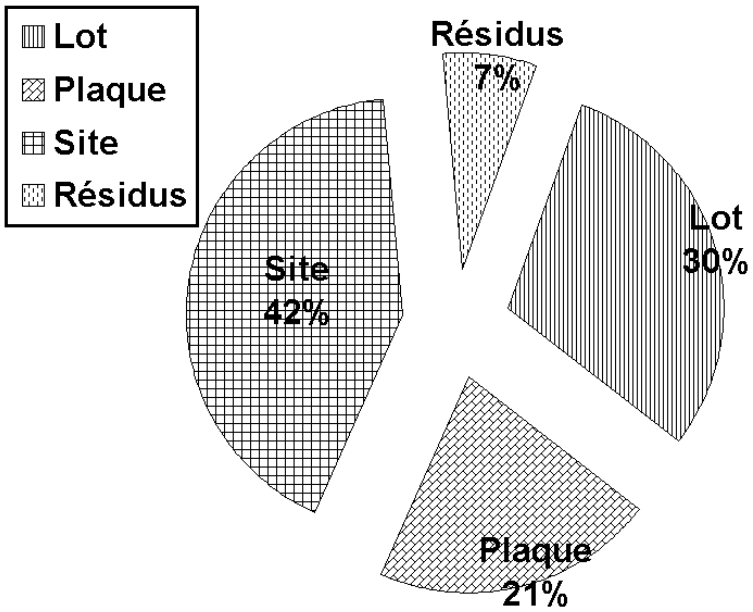

FIGURE 2.5 – *Décomposition de variance du courant de fuite I_{off}.*

Pour chacune des variables considérées, paramétriques ou physiques, nous n'avons pas manqué de construire les diagrammes des résidus normalisés correspondants. L'objectif était de vérifier des hypothèses essentielles à la conduite d'une analyse de variance : la normalité des résidus ainsi que des différents effets aléatoires. Nous n'avons pas jugé nécessaire de faire apparaître ces diagrammes dans ce manuscrit.

2.4 La variabilité paramétrique intra-plaque

Nous avons montré que la variabilité paramétrique intra-plaque est importante et dépasse 50% de la variabilité totale. De la même façon, nous avons vu que la compo-

9. Au PT, 100% des plaques sont mesurées. Le nombre de sites mesurés est de 5 par plaque

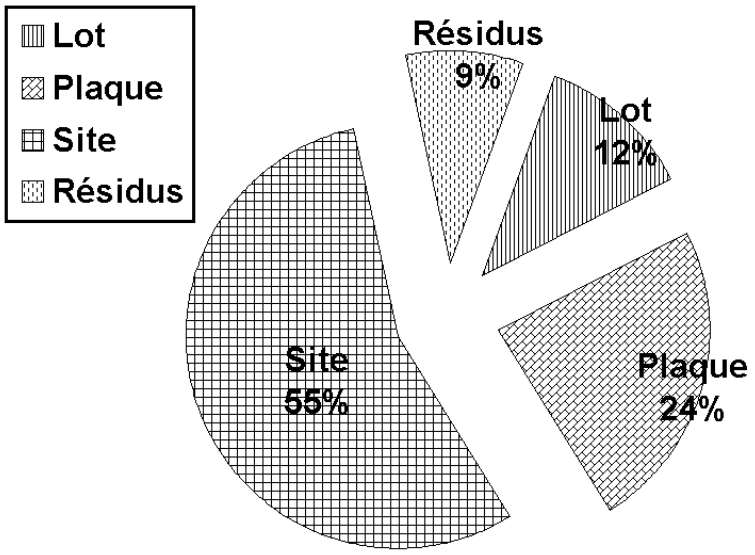

FIGURE 2.6 – *Décomposition de variance du courant de saturation I_{on}.*

sante intra-plaque est majoritaire dans les variations de L_{poly}. Nous nous proposons alors de comparer la cartographie de L_{poly} à celles des caractéristiques électriques critiques étudiées jusqu'alors. Ces cartographies sont réalisées à l'aide de Waferfit, une application logicielle développée par MASA [10]. Waferfit ne se contente pas de relier les mesures [11] à partir des coordonnées spatiales des structures de test, mais s'appuie sur des techniques d'interpolation statistiques afin de générer des cartographies plus précises.

Nous avons réalisé des cartographies de l'ensemble des plaques des 5 lots mesurés. Les paramètres considérés sont encore une fois L_{poly}, I_{on}, I_{off} et V_{th}. Le constat au sujet de la première variable L_{poly} est catégorique : nous avons noté la co-habitation de deux profils distincts, illustrés par les figures 2.9 et 2.10. 100% des plaques sont représentées par l'un ou l'autre de ce couple de profils. Les figures 2.11 et 2.12 correspondent à des cartographies du courant I_{on}. Le constat est sans appel : la signature mesurée au PT est très similaire (voire identique) à celle de la longueur de grille. Ce résultat est valable pour une grande majorité des plaques testées, aussi bien pour le courant I_{on} ou le courant I_{off}, et rend ainsi compte de l'impact décisif de la dispersion intra-plaque de L_{poly} sur la dispersion spatiale paramétrique.

Une analyse approfondie des causes probables à l'origine de ces profils distincts au niveau de la gravure de grille nous amène naturellement aux chambres de gravure mais aussi aux fours PEB. La résine $193nm$ utilisée présente en effet une forte sensibilité au PEB de l'ordre de $8\,nm/^{\circ}C$. Grâce à l'outil de traçabilité des plaques intégré à l'équipement de photolithographie, nous avons pu associer à chaque profil de L_{poly} une unité PEB différente. Il s'est avéré effectivement qu'un four parmi les trois qualifiés est atypique. Les cartographies de la longueur de résine L_{resist} en amont de la gravure (voir figures 2.13 et 2.14) confirment cette constatation. Le four atypique présente par ailleurs une dispersion plus élevée que les deux autres fours. Les tâches

10. www.bluekaizen.com
11. Le nombre de mesures est trop faible pour couvrir toute la surface de la plaque

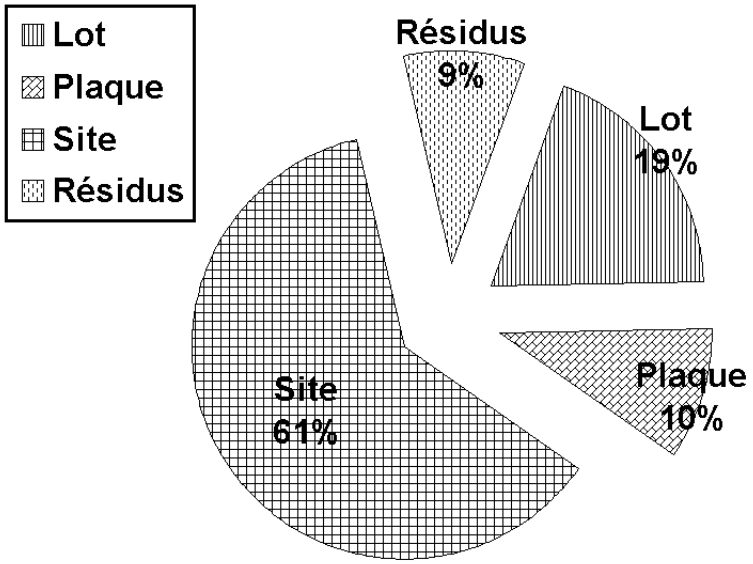

FIGURE 2.7 – *Décomposition de variance de la tension de seuil V_{th}.*

de qualité existantes ne permettaient pas de mettre en évidence cet aspect.

Ce résultat a été un argument de poids en faveur du remplacement les fours RHP (Rapid Hot Plate) existants par des fours S-RHP (Super Rapid Hot Plate). Les SRHP garantissent une bien meilleure uniformité en température, et par conséquent une bien plus faible variabilité dimensionnelle (voir figure 2.15).

2.5 CONCLUSION

L'objet de ce chapitre était d'identifier la cause de premier ordre à l'origine des variations paramétriques. Nous avons déployé pour cela une méthodologie double : une régression PLS suivie d'une analyse de variance hiérarchique. La régression PLS est une technique de modélisation linéaire, qui présente un grand nombre d'atouts vis-à-vis d'autres méthodes statistiques que nous aurions pu envisager. Un atout important est sa robustesse à l'éventuelle colinéarité des facteurs X_i pris en compte dans l'analyse.

L'application de cette méthodologie a démontré que la longueur de grille L_{poly} est la source majeure des variations lot à lot et plaque à plaque des paramètres critiques (I_{on}, I_{off} et V_{th}). L'étendue de la composante spatiale, aussi bien pour la variable L_{poly} que pour les caractéristiques électriques, nous a amené à déterminer la signature des paramètres susmentionnés, par le biais de l'application WaferFit. L'analyse des cartographies de la surface des plaques testées a mis en évidence une signature spatiale identique, synonyme d'un lien de cause à effet entre la longueur de grille L_{poly} et les caractéristiques électriques critiques.

Nous avons aussi démontré le poids important des unités PEB sur les signatures révélées. Grâce à ce résultat, les fours RHP ont été remplacés par des fours S-RHP plus récents. Cette action a permis de réduire le gradient de température des fours PEB et améliorer par conséquent l'uniformité intra-plaque des paramètres électriques

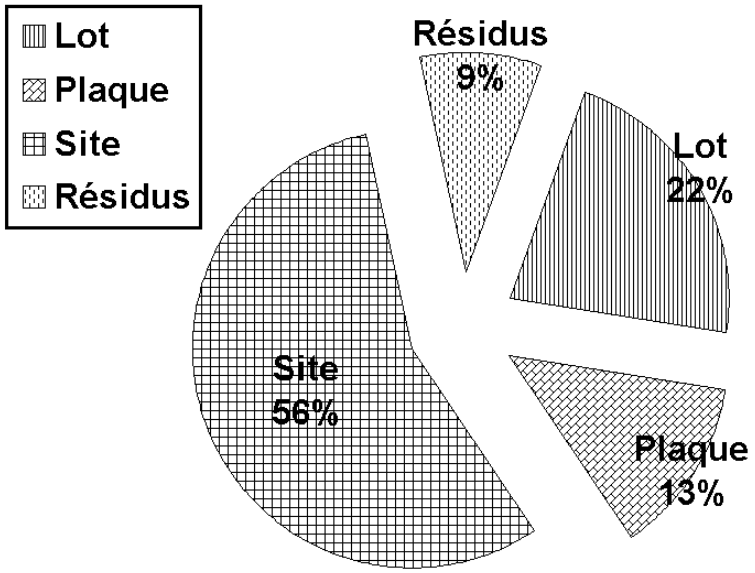

FIGURE 2.8 – *Décomposition de variance de la longueur de grille L_{poly}.*

critiques.

Bien que l'effet lot soit faible comparé à la composante intra-plaque ($\simeq$ 75%), la composante lot demeure assez considérable pour espérer un gain significatif par la mise en production de la boucle (FF) entre la gravure de grille et l'implantation des poches. Cette boucle, qui sera décrite en détails au chapitre 4, permet de compenser la variabilité lot à lot de L_{poly} en ajustant la dose d'implantation des poches.

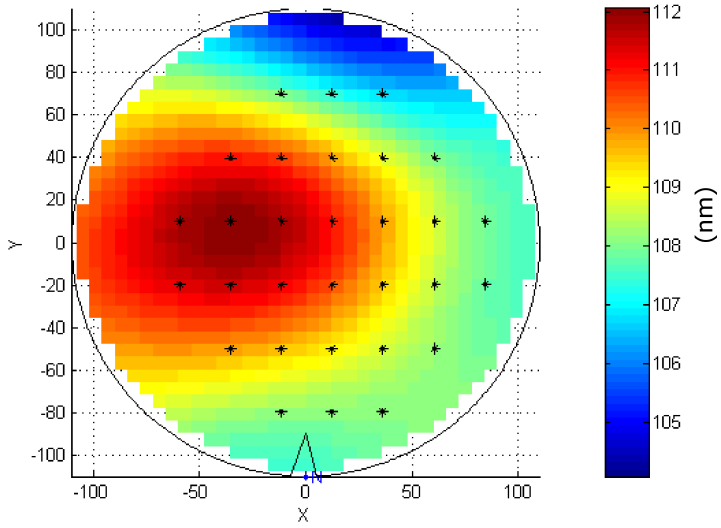

FIGURE 2.9 – *Cartographie du paramètre L_{poly} (nm). Il s'agit d'un premier profil récurrent commun à plusieurs plaques de lots différents.*

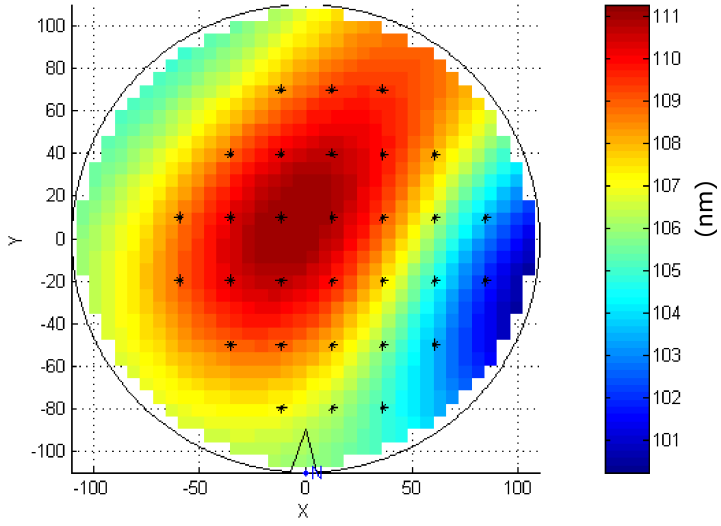

FIGURE 2.10 – *Cartographie du paramètre L_{poly} (nm). Il s'agit d'un second profil récurrent commun à plusieurs plaques de lots différents.*

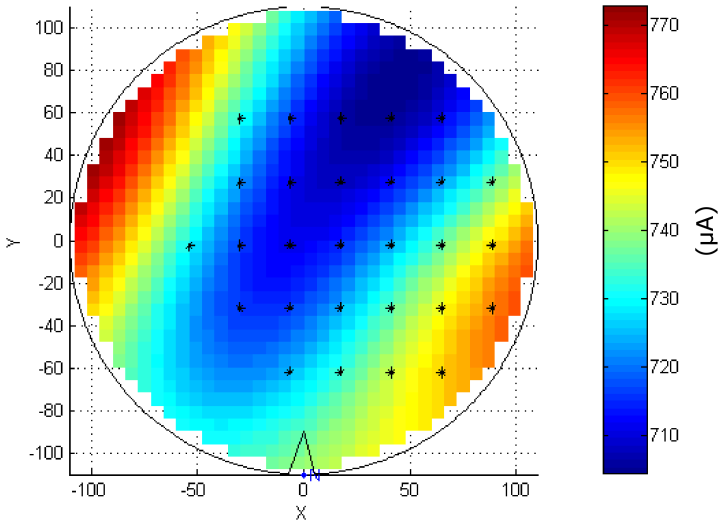

FIGURE 2.11 – *Cartographie du paramètre I_{on} (µA). Il s'agit d'un premier profil récurrent commun à plusieurs plaques de lots différents.*

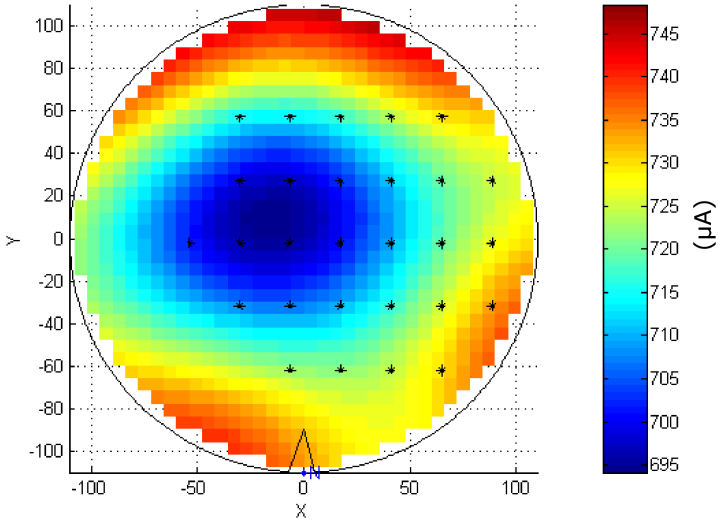

FIGURE 2.12 – *Cartographie du paramètre I_{on} (µA). Il s'agit d'un second profil récurrent commun à plusieurs plaques de lots différents.*

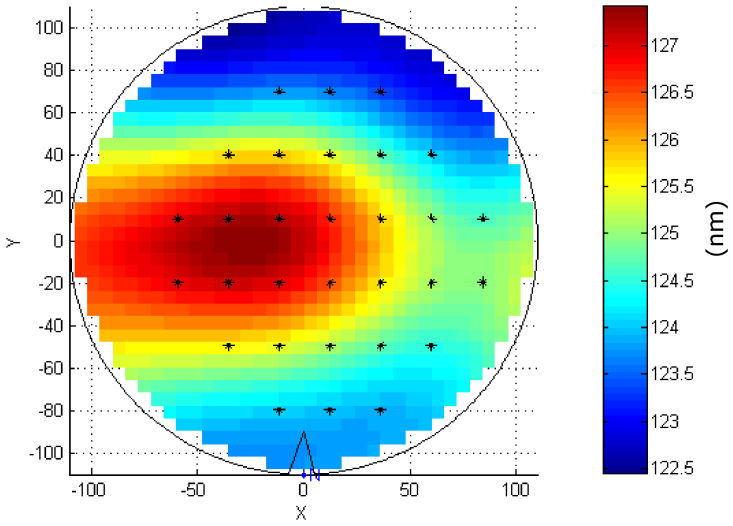

FIGURE 2.13 – *Cartographie du paramètre L_{resist} (nm). Il s'agit d'un premier profil récurrent commun à plusieurs plaques de lots différents.*

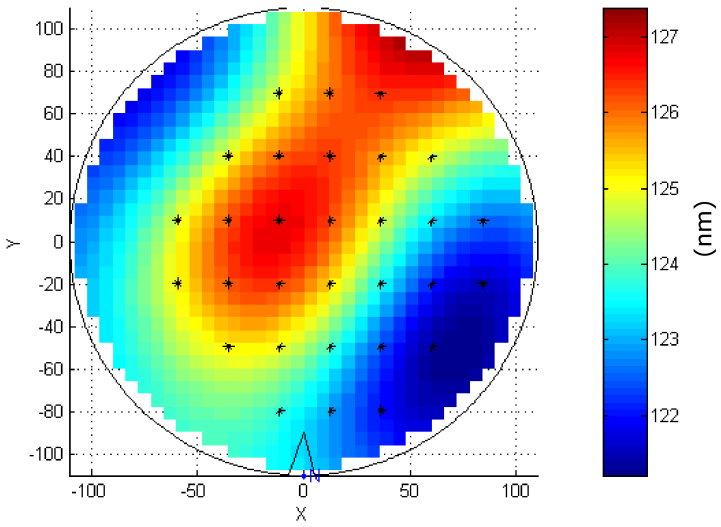

FIGURE 2.14 – *Cartographie du paramètre L_{resist} (nm). Il s'agit d'un second profil récurrent commun à plusieurs plaques de lots différents.*

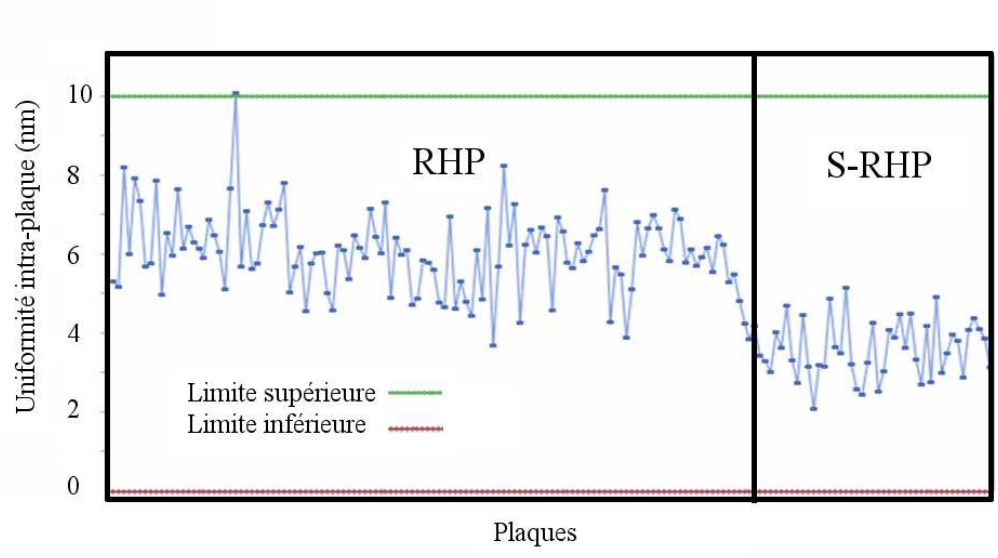

FIGURE 2.15 – *Evolution de l'uniformité dimensionnelle intra-plaque suite au remplacement des fours RHP par des fours S-RHP. L'uniformité, sur ce graphe, est mesurée par l'étendue du CD (valeur maximale - valeur minimale).*

3 Aperçu des Méthodes de Régulation : Application aux Procédés Industriels à Gain Statique

Sommaire

La littérature scientifique est très riche en méthodes de régulation (FB). Évidemment, elles ne sont pas toujours décrites en vue d'une application industrielle dans le monde du semi-conducteur. Elles sont, au contraire, souvent détaillées d'un point de vue automaticien pour des applications de l'industrie chimique généralement. Indépendamment du choix de la loi de commande (commande adaptative, prédictive, PID, etc), il est essentiel de faire une modélisation préalable du procédé et des perturbations. En réalité, le procédé et les perturbations sont indissociables, et une modélisation parfaite de cet ensemble n'est qu'illusion. L'origine diverse des perturbations en est la principale cause. Si nous prenons l'exemple du procédé lithographique en semi-conducteur, les perturbations, à l'origine des variations de la dimension critique d'une ligne de résine, peuvent être regroupées en 3 familles :

Les perturbations liées à la machine Les équipements de lithographie (Scanner/Track) sont équipés de boucles d'asservissement sophistiquées pour garantir que les paramètres de la recette de production sont respectés (Température des fours PEB, l'intensité du laser,...). Néanmoins, ces paramètres sont sujets à des dérives.

Les perturbations liées à la matière première D'un batch de résine à un autre, les caractéristiques de la résine (viscosité, concentration en PAG,...) peuvent changer.

Les perturbations liées à l'Environnement Telle la fluctuation de la pression atmosphérique, ou la température ambiante...

Toutes ces perturbations ne peuvent toutes être prises en compte dans un modèle mathématique, et ceci pour différentes raisons : un nombre démesuré de paramètres, capteurs de mesure trop coûteux ou inexistants, interactions et effets non linéaires, etc. Aussi, une telle démarche ne serait pas pertinente dans un monde industriel où temps et argent sont le plus souvent à épargner. Dans le cadre de ces travaux, nous postulons deux hypothèses :

H1 : L'évolution du ou des paramètres de sortie (Output) est la somme de deux composantes :

1. Une composante déterministe à gain statique,
2. et une fonction stochastique pour formaliser le bruit (les perturbations). Elle comprends tous les effets qui ne sont pas pris en compte dans le modèle déterministe et qui engendrent toutefois une variation des variables de sortie. Les effets du bruit de mesure, des erreurs de modélisation en font partie. Elle est évidemment non mesurable et de nature stochastique. Il s'agit souvent d'une combinaison d'éléments aléatoires (bruit gaussien blanc) et déterministes (décalage, marche, etc).

H2 : Les procédés qui feront l'objet d'application des différentes lois de commandes par la suite, sont linéaires et invariants.

L'avantage de ces hypothèses est surtout de pouvoir allier une approche automatique classique (stabilité, erreur en régime permanent,...) et une approche statistique (inflation de la variance,...) dans l'étude des différentes lois de commandes qui existent. Ces hypothèses sont souvent implicitement adoptées par les membres de la communauté scientifique qui ont travaillé sur le sujet. Muske et al, dans un article de synthèse qui traite du contrôle prédictif dans l'industrie chimique, a adopté une variante de H1 [Muske et Rawlings (1993)]. Le principe demeure le même sauf que la perturbation est injectée au niveau des variables d'entrée.

Il est essentiel de souligner que ce chapitre n'a pas pour vocation de développer des nouvelles approches de régulation. Nous y cherchons tout au plus à montrer l'intérêt de méthodes existantes dans le contexte d'une production de composants microélectroniques. Dans ce cadre, nous allons nous intéresser notamment à la commande **PID**, et développer certaines de ses caractéristiques en boucle fermée. L'inflation de la variance de la déviation sera aussi étudiée. Dans une seconde partie, nous allons présenter d'une manière succincte quelques contrôleurs plus évolués et plus récents, tels le contrôleur **MMSE** (Minimum Mean Square Error) et le contrôleur adaptatif. Nous clorons le chapitre par un état de l'art de quelques applications des boucles de régulation en photolithographie.

3.1 Modélisation du système

En parlant de système, nous sous-entendons qu'il s'agit d'un procédé de fabrication dans l'industrie microélectronique. Au sein de cette industrie, il existe un large panel de procédés aux dynamiques très différentes. A titre d'exemple, dans le cas du recuit rapide [Keyser et Donald (1999)], l'inertie thermique du système se traduit par une relation dynamique entre la puissance des lampes halogènes (variable d'entrée) et la température de la plaque (variable de réponse). A l'opposé, il est admis que la lithographie, **procédé d'application pour nos travaux**, est un procédé à gain statique.

Une représentation générale de type fonction de transfert Box-Jenkins est retranscrite ci-dessous [Box et al. (1994)]. Le modèle Box-Jenkins présente l'avantage d'être simple à interpréter : la variable de sortie est la somme d'une perturbation $P_k = \frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})}\varepsilon_k$ et d'un signal de transfert entrée-sortie. Nous nous plaçons dans le cas le plus simple (voir équation 3.1), où le système aurait une seule variable d'entrée (dose d'énergie d'exposition) et une seule variable de sortie (dimension critique de la résine).

$$y_k = \frac{\mathbf{B}(q^{-1})}{\mathbf{A}(q^{-1})}u_k + \frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})}\varepsilon_k \tag{3.1}$$

où **A**, **B**, **C** et **Q** sont des polynômes retard [1], u_k est la variable d'entrée, y_k est la variable de sortie. Dans le cadre de nos travaux, en l'occurrence dans le cas d'un procédé à gain statique, le polynôme $A(q^{-1})$ est un scalaire égal à 1 et le polynôme $B(q^{-1})$ est égal au gain du système β. L'équation 3.1 est alors réduite à l'équation 3.2.

$$y_k = \beta\, u_k + P_k \tag{3.2}$$

La fraction rationnelle $\frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})}$ modélise l'influence du bruit dans le système. Ce transfert entre la perturbation et la sortie est généralement impossible à identifier. Il ne demeure pas moins qu'il s'agit d'un paramètre important, comme l'a souligné De Keyser dans ses travaux [Keyser et Ionescu (2003)]. Une manière simple et surtout négligée de modéliser ce rapport est d'admettre qu'il est égal à 1. Les réalisations de la perturbation sont alors supposées être indépendantes et à moyenne nulle. Les procédés en fabrication des semi-conducteurs sont rarement de cette nature. Exposer les lots en lithographie à une dose d'énergie constante conduit à terme à une dimension critique hors spécifications, ce qui témoigne effectivement du caractère non stationnaire des perturbations. Dans le cas où cette hypothèse serait malgré tout vraie, nous savons, à travers l'expérience de l'entonnoir menée par Deming, que l'ajustement par retour de mesure (le bouclage, le feedback) d'un procédé sous contrôle statistique (perturbations aléatoires indépendantes) et centré ne peut qu'altérer la variance du procédé [Deming (2000), Castillo (2002), MacGregor (1990)].

De ce fait, ce modèle réduit à l'unité est souvent délaissé en faveur d'une modélisation qui colle davantage à la réalité. Le modèle 3.3, appelé **marche aléatoire**, en est un exemple. Il s'agit du modèle de prédilection dans plusieurs applications run-to-run [Moyne et al. (2000)], où la perturbation est décrite comme une série temporelle auto-corrélée [2] et non stationnaire. Son atout principal ; aucune démarche d'identification à réaliser. Dans le cas d'une distribution des ε_k en forme d'un dirac à

1. J'invite le lecteur à consulter l'annexe I pour une brève description des modèles ARIMA
2. Il existe une corrélation entre les réalisations successives de la perturbation P_k

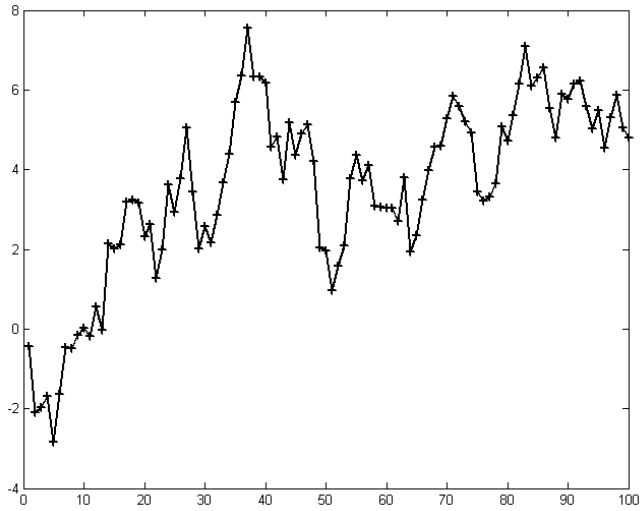

FIGURE 3.1 – *Illustration d'un processus de marche aléatoire*

moyenne nulle, la marche aléatoire est équivalente à un échelon à moyenne égale à P_0 [MacGregor et al. (1984)].

$$\frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})} = \frac{1}{1-q^{-1}} \Leftrightarrow P_{k+1} = P_k + \varepsilon_{k+1} \tag{3.3}$$

La marche aléatoire demeure toutefois un modèle fortement non stationnaire, à priori non adapté à tous les procédés élémentaires de l'industrie du semi-conducteur. Afin de palier à cet inconvénient et affiner davantage la modélisation, Box et Luceno ont opté pour un modèle de type $\mathbf{IMA}(1,1)$ [Box et Luceño (1997b)].

$$\frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})} = \frac{1-\theta q^{-1}}{1-q^{-1}} \Leftrightarrow P_{k+1} - P_k = \varepsilon_{k+1} - \theta\varepsilon_k \tag{3.4}$$

où θ est le paramètre de lissage, $|\theta| < 1$. Une formulation équivalente est

$$P_{k+1} = \varepsilon_{k+1} + (1-\theta)\sum_{i=1}^{k}\varepsilon_i \tag{3.5}$$

La perturbation peut ainsi être considérée comme une somme pondérée d'un bruit blanc obtenu pour une valeur unitaire de θ et une marche aléatoire fortement non stationnaire, obtenu dans le cas où θ est égal à 0. Le modèle $\mathbf{IMA}(1,1)$ occupe une place centrale dans la description du comportement de plusieurs perturbations non-stationnaires [Box et Kramer (1992)]. Un exemple rencontré en industrie du semi-conducteur où la perturbation se confondrait avec un processus $\mathbf{IMA}(1,1)$, est celui du polissage mécanochimique et mécanique (CMP). Il est détaillé dans [Castillo (2002)].

Par ailleurs, Macgregor, Harris et Wright ont démontré que les perturbations déterministes (shifts, dérives, etc), se produisant peu fréquemment à des intervalles aléatoires, peuvent être représentées par des des processus autorégressifs intégrateurs à moyenne mobile (**ARIMA**) [MacGregor et al. (1984)]. Par souci d'être le plus exhaustif possible et couvrir un plus grand nombre de perturbations, nous allons adopter par la suite le modèle $\mathbf{ARIMA}(1,1)$. Il s'agit en effet d'un modèle général, dont la

marche aléatoire et le modèle **IMA**(1,1) sont des cas particuliers, comme le montre l'expression 3.6. Dans une étude de comparaison entre différentes commandes classiques de l'industrie microélectronique , Del Castillo a opté pour ce même modèle pour décrire la perturbation [Castillo et Hurwitz (1997)].

$$\frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})} = \frac{1-\theta q^{-1}}{(1-q^{-1})(1-\omega\, q^{-1})} \tag{3.6}$$

où θ et ω sont deux paramètres caractéristiques de la perturbation ($|\theta| < 1$, $|\omega| \leq 1$).

En regard du contexte industriel de ces travaux de thèse et la complexité de la tâche, il n'est pas envisagé de mener une campagne d'essais pour valider tel ou tel modèle. L'object de ce chapitre est de comprendre globalement le choix des régulations existantes (mais aussi le non choix de certaines) à travers cette double approche statisticienne/automaticienne, mais aussi de saisir des pistes potentielles d'amélioration.

Suite au choix de la forme générale du modèle, à savoir la somme d'un signal de transfert entrée-sortie à gain statique et une perturbation de type **ARIMA**$(1,1)$, l'objectif de ce chapitre est de réaliser un contrôleur, capable de minimiser les déviations de la variable de sortie y_k et d'en optimiser un indice de performance. Cet indice sera définit dans la section suivante. Pour cela, nous allons dérouler une double approche,

⋈ Une première approche axée sur l'étude des caractéristiques du système bouclé d'un point de vue automatique : stabilité, comportement en régime transitoire et permanent suite à l'application d'une perturbation. Cette étape sera basée sur les outils classiques d'automatique.

⋈ Une seconde approche statistique où le but est d'estimer l'impact du correcteur sur des indices de qualité prédéfinis telle la variance de la variable de sortie.

Cette méthodologie est inspirée des travaux de David E. Hardt et Tsz-Sin-Siu, où ils ont appliqué cette double approche afin de concevoir des correcteurs de type PI pour deux procédés, à savoir le cintrage du métal, et le moulage par injection [Hardt et Siu (2002)].

3.1.1 L'approche automaticienne [Borne et Richard (1993), de Larminat (1996)]

Pour caractériser le système en boucle fermée, l'automaticien a recours à un ensemble de méthodes descrites brièvement en annexe **A**, et qui lui permet d'analyser trois éléments importants : la stabilité du système, son comportement en régime transitoire et son comportement en régime permanent. L'industrie du semi-conducteur est une industrie où la fabrication des dispositifs micro-électroniques est réalisée par lots [3]. De ce fait, nous allons nous intéresser particulièrement à l'étude de systèmes discrets, décrits généralement au moyen de l'opérateur z (voir annexe **A**). Rappelons que cette approche traitera exclusivement les perturbations déterministes (la perturbation en échelon, rampe, etc).

3. Les processus de fabrication sont discontinus.

3.1.2 L'approche statistique

En amont de la mise en place d'une boucle de régulation pour un procédé déterministe donné, l'étude d'un ensemble de caractéristiques (stabilité, erreur statique, ...) selon le paramétrage du contrôleur est une approche classique de la théorie de contrôle de base. La dimension stochastique des procédés rencontrés en fabrication des semiconducteurs rend cette approche insuffisante. L'objectif de l'asservissement n'est plus de maintenir une grandeur physique constante et égale à la consigne, mais de réaliser un double objectif : maintenir la moyenne des réalisations de la variable de sortie égale à la consigne et réduire aussi sa fluctuation statistique. De cet objectif est né le besoin de mettre en place un indice de qualité qui rend compte des moments du premier et du second ordre de la variable de sortie (moyenne et variance), considérée désormais comme une variable aléatoire et non une grandeur déterministe.

L'indice de capabilité C_{Pk}

Le C_{Pk} est un indicateur de qualité couramment utilisé par les industriels en fabrication des composants. Il rend compte de la dispersion et du centrage des réalisations d'un paramètre donné (voir expression 3.7).

$$C_{Pk} = \min\left(\frac{\bar{X} - lsi}{3\sigma}, \frac{lss - \bar{X}}{3\sigma}\right) \tag{3.7}$$

où $\bar{X}$ est la valeur moyenne des réalisations, σ est l'écart-type de la distribution et lsi et lss sont respectivement les limites de spécification inférieure et supérieure.

La Variance Asymptotique

Dans le cas d'un processus stochastique décrit par une équation aux différences, il est souvent aisé d'estimer la variance asymptotique du processus **AVar**, définie comme la limite finie ou infinie de la variance lorsque le nombre de réalisations augmente infiniment (voir équation 3.8). Il s'agit là d'un critère qui reconnaît d'une manière implicite l'existence d'un régime transitoire dont il faudrait tenir compte séparément. **AVar** ne prend pas en compte le centrage (la moyenne) des réalisations. Dans le cas de processus non-stationnaires, **AVar** est infinie, alors pour comparer différentes lois de commande susceptibles de générer des processus non stationnaires, il est plus instructif de s'appuyer sur un autre critère : l'Erreur Quadratique Moyenne (**M**ean **S**quare **E**rror **MSE**) relatif à un nombre fini de réalisations [Tsung et al. (1998)].

$$\mathbf{AVar} = \lim_{n\to\infty} Var(X_i)_{1\leq i\leq n} \tag{3.8}$$

où $(X_i)_{1\leq i\leq n}$ est une variable aléatoire avec n réalisations.

L'Erreur Quadratique Moyenne

Le critère de l'erreur quadratique moyenne est équivalent à la fonction de perte ou *loss quality function*, introduit par Taguchi pour évaluer la qualité d'un produit [Taguchi et al. (1988)]. Il est défini comme l'écart quadratique moyen par rapport à la consigne (voir équation 3.9).

$$\mathbf{MSE}(X) = \mathbf{E}[(X_i - \mathbf{T})^2_{1\leq i\leq n}] \tag{3.9}$$

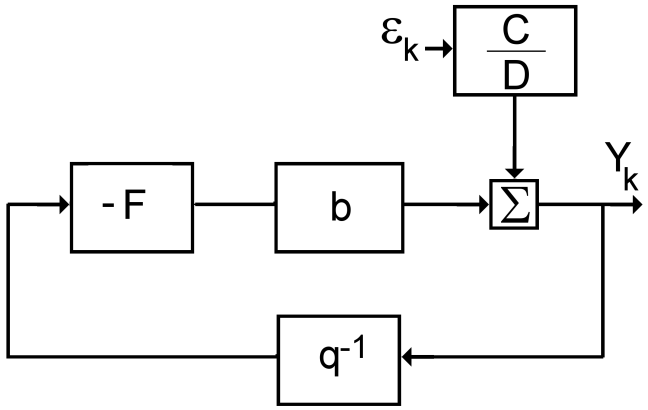

FIGURE 3.2 – *Schéma d'un système bouclé. Y_k est la déviation de la variable de sortie (consigne nulle), b est la gain du procédé, F est la fonction de transfert du contrôleur, $P_k = \frac{C}{D}\varepsilon_k$ est le modèle relatif à la perturbation stochastique et q^{-1} est l'opérateur de retard.*

où X est une variable aléatoire avec n réalisations, **T** est la consigne et **E** l'opérateur moyenne. A partir de cette définition, nous déduisons la formule 3.10. Cette nouvelle formulation nous indique que le **MSE** capture aussi bien les variations du procédé que sa déviation par rapport à la consigne.

$$\mathbf{MSE}(X) = Var(X) + (\bar{X} - \mathbf{T})^2 \tag{3.10}$$

Par la suite, nous allons revoir les différents contrôleurs qui ont été illustrés dans la littérature. Nous allons dérouler la double approche automaticienne/statisticienne pour la seule commande **PID** (**P**roportionnelle **I**ntégrale **D**érivée). En ce qui concerne les autres lois de commandes, nous allons nous contenter d'une description succincte de quelques propriétés.

3.2 La Commande Proportionnelle Intégrale Dérivée

Le système bouclé peut être représenté par le schéma 3.2 [Borne et Richard (1993), de Larminat (1996)]. Le retard pur incorporé dans la chaîne de retour s'explique simplement par le fait que la déviation[4] y_k relative au cycle[5] k servira au calcul de la commande du cycle $k+1$, voire des cycles futurs ($k+i, i \geq 1$). Dans le contexte d'une fabrication semi-conducteur, ce retard est très souvent d'ordre supérieur ($r > 1$), dépendant du degré de priorité du lot, de la charge des équipements de métrologie et de la disponibilité des opérateurs. Le paramètre mesuré (y_k) pourrait, par ailleurs, être échantillonné. Dans ces différents cas de figure, q^{-1} serait alors remplacé par q^{-r_k}, où r_k est le retard de la réponse y_k, et est variable dans le temps.

L'étude d'un tel système, aussi complexe, requiert des compétences en automatique avancée, et notamment dans le domaine des systèmes à retard. En regard du contexte industriel de ces travaux et la complexité de la tâche, elle pourrait éventuellement être réalisée dans le cadre d'une nouvelle thèse, à vocation plus académique. Dans ce chapitre, nous avons opté pour une hypothèse simplificatrice où la suite r_k

4. y_k est une déviation d'un paramètre physique donné, la consigne est sous-entendue nulle.
5. Un cycle est la traduction du mot anglosaxon run. Nous aurions pu utilisé aussi, de manière équivalente le mot lot.

est constante et égale à 1. La fonction de transfert du système en boucle ouverte est $\mathbf{T}_{BO} = b\,q^{-1}\,\mathbf{F}$. Le système en BF (boucle fermée) est alors formalisé ainsi :

$$y_k = \frac{1}{1 + T_{BO}} P_k \tag{3.11}$$

Il est important de noter que le transfert en boucle fermée, comme défini dans ce chapitre, correspond à celui entre la perturbation P_k et la déviation y_k, autrement la sensibilité de y_k à P_k. Dans la littérature automatique, le terme **transfert en BF** est souvent réservé à celui entre la consigne et la réponse y_k. Dans le cadre de nos travaux, le choix de ce transfert ne serait pas pertinent. Nous supposons, en effet, avoir une consigne constante (nulle). L'objectif du contrôleur est ainsi réduit au rejet des perturbations.

3.2.1 La commande proportionnelle

Approche automaticienne

Le régulateur à action proportionnelle, ou régulateur P, a une action simple et naturelle, puisqu'il construit une commande u_{k+1} proportionnelle à l'erreur y_k, comme le montre l'équation 3.12.

$$u_{k+1} = u_0 - k_P y_k \tag{3.12}$$

où k_P est le gain proportionnel et u_0 est une constante qui correspond à la commande nominale, désormais égale à 0. La fonction de transfert en boucle ouverte $\mathbf{T}_{BO}$ se réduit à $b\,k_P\,q^{-1}$. Nous posons $g = b\,k_P$. La fonction de transfert en boucle fermée est formulée ci dessous.

$$\mathbf{T}_{BF} = \frac{1}{1 + bk_P q^{-1}} = \frac{1}{1 + g\,q^{-1}} \tag{3.13}$$

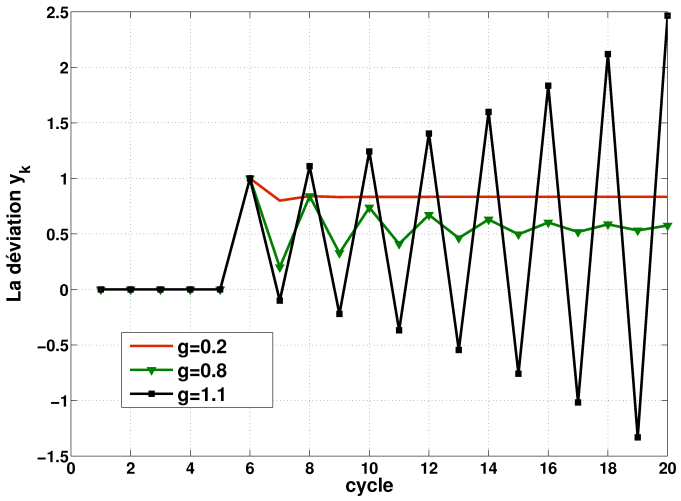

FIGURE 3.3 – *Réponse du système bouclé à une perturbation en échelon ; il s'agit d'une commande P à gain k_p variable. Pour $|g| > 1$, le système est instable et y_k diverge. Tandis que pour $|g| < 1$, nous constatons une erreur statique non nulle, dont l'amplitude est inversement proportionnelle à $|g|$.*

Lorsque le gain k_P augmente de o à $+\infty$, le pôle du système bouclé égal à $-g$ varie de o à $-\infty$. Pour garantir la stabilité en boucle fermée, il faudrait que g soit

strictement compris entre -1 et 1. Son choix aura par ailleurs un impact sur la durée du régime transitoire t_{rep}. t_{rep} est approximé, à l'instar des sytèmes continus, par une expression empirique donnée en annexe **A** et rappelé ci-dessous.

$$t_{rep}(2\%) = -\frac{4}{Ln(|p_m|)} \tag{3.14}$$

Où p_m est le pôle dominant du système bouclé, $p_m = \{p, |p| = \max(|p_i|)\}$. Pour $g = 0.5$, t_{rep} est de l'ordre de 13 cycles. Il est égal à 4 cycles pour $g = 0.1$. Le lecteur peut apprécier ce constat à travers la figure 3.3. Enfin, en ce qui concerne le biais ζ_k de la déviation y_k, ζ_k s'écrit à chaque cycle k :

$$\zeta_k = -\frac{1}{1 + g\,q^{-1}} P_k \tag{3.15}$$

Contrairement à la stabilité et à la rapidité de l'asservissement, caractéristiques indépendantes de la nature de la perturbation P_k et intrinsèques au système, l'erreur statique ζ_∞ en dépend. Pour une perturbation en échelon d'amplitude κ, l'erreur statique est non nulle. Elle est égale à

$$\zeta_\infty = \lim_{q\to 1}(1 - q^{-1})\zeta_k = \lim_{q\to 1}(1 - q^{-1})\,\frac{-1}{1 + g\,q^{-1}}\,\frac{\kappa}{1 - q^{-1}} = -\frac{\kappa}{1 + g} \neq 0 \tag{3.16}$$

L'erreur en régime permanent n'est pas nulle. Nous savons effectivement que pour rejeter une perturbation en échelon, nous devons appliquer une action intégrale en amont du point d'application de la perturbation, ce qui n'est pas le cas ici. Augmenter le gain k_P permet toutefois de minimiser $|\zeta_\infty|$, au détriment de la stabilité.

Approche statisticienne

Nous interrompons cette démarche d'automaticien et nous proposons de résoudre l'équation aux différences qui régit le système bouclé. La finalité de ce calcul est d'exprimer la variance de la déviation y_k en boucle fermée.

$$y_{n+1} + g\,y_n = P_{n+1} \tag{3.17}$$

Un simple calcul algébrique conduit au résultat suivant :

$$y_{n+1} = (-g)^{n+1} y_0 + \sum_{i=0}^{n} (-g)^i P_{n-i+1} \tag{3.18}$$

Sous forme matricielle, ceci peut s'écrire

$$y_{n+1} = (-g)^{n+1} y_0 + \mathbf{K}^T \mathbf{P}$$

$$\text{où } \mathbf{K} = \begin{bmatrix} 1 \\ -g \\ (-g)^2 \\ \cdots \\ (-g)^n \end{bmatrix} \text{ et } \mathbf{P} = \begin{bmatrix} P_{n+1} \\ P_n \\ P_{n-1} \\ \cdots \\ P_1 \end{bmatrix}$$

En se basant sur la formule de la variance donnée par Dougherty [Dougherty (1990)], nous déduisons l'expression de la variance de la sortie en fonction de **K** et de la matrice de covariance du vecteur colonne **P** [Hardt et Siu (2002)].

$$\sigma^2_{y_{n+1}} = \mathbf{K}^T \mathbf{\Sigma}_{n+1} \mathbf{K} \tag{3.19}$$

Où

$$\mathbf{\Sigma}_n = \begin{bmatrix} Cov(P_1, P_1) & Cov(P_1, P_2) & \dots & Cov(P_1, P_n) \\ Cov(P_2, P_1) & Cov(P_2, P_2) & \dots & Cov(P_2, P_n) \\ \dots & \dots & \dots & \dots \\ Cov(P_n, P_1) & Cov(P_n, P_2) & \dots & Cov(P_n, P_n) \end{bmatrix}$$

Perturbation en Bruit Blanc Par définition du bruit blanc, les perturbations P_k sont des réalisations indépendantes et dont la distribution est supposée être normale $\mathcal{N}(0; \sigma^2)$. La matrice $\mathbf{\Sigma}$ est réduite dans ce cas particulier à la matrice identité multipliée par σ^2. Nous déduisons la variance de la sortie y_{n+1} par la formule suivante [Hardt et Siu (2002)] :

$$\sigma^2_{y_{n+1}} = \sigma^2 \sum_{i=0}^{n} ((-g)^{2i}) = \sigma^2 [\frac{1 - g^{2(n+1)}}{1 - g^2}] \tag{3.20}$$

La variance de la sortie en boucle fermée est ainsi toujours strictement supérieure à la variance de la perturbation, sauf si k_P est nul, auquel cas la commande proportionnelle est mise hors service. Intuitivement, les réalisations y_k étant perturbées par un bruit blanc, aucune information sur le cycle k ne serait utile (ne contiendrait de l'information utile) pour le ou les cycles futurs. La variance de la sortie est alors dégradée davantage en boucle fermée et tend d'une manière asymptotique vers une certaine valeur, exprimée ci dessous en fonction du gain k_P (équation 3.21). **AVar** est d'autant plus amplifiée que le gain k_P est grand.

$$\sigma^2_y = \sigma^2 [\frac{1}{1 - g^2}] \tag{3.21}$$

Marche Aléatoire Pour la perturbation ainsi modélisée, $P_k = \frac{1}{1-q^{-1}} \varepsilon_k$, le développement théorique de la variance de la sortie à chaque cycle k, de la même manière que le paragraphe précédent, serait une tâche fastidieuse. Nous allons nous contenter d'estimer la variance asymptotique **AVar**, une tâche plus aisée. Rappelons tout d'abord l'équation aux différences du système en boucle fermée.

$$y_{n+1} = (1 - g)\, y_n + g\, y_{n-1} + \varepsilon_{n+1} \tag{3.22}$$

Il s'agit d'un processus AR(2) non stationnaire [Castillo (2002), Box et al. (1994)] dont la variance asymptotique de la sortie **AVar** tend théoriquement vers l'infini (voir figure3.4). En clair, la commande proportionnelle est incapable de ramener à 0 la déviation y_k. Suite à ce constat et dans la mesure où la marche aléatoire est un modèle de base utile pour représenter les perturbations réelles, il serait inutile de continuer cette démarche avec d'autres perturbations plus complexes et plus générales (**IMA**$(1,1)$, **ARIMA**$(1,1,1)$). La commande proportionnelle sera toujours dans l'incapacité de réduire les fluctuations de y_k, sous l'effet de perturbations non stationnaires, notamment **IMA**$(1,1)$ et **ARIMA**$(1,1,1)$.

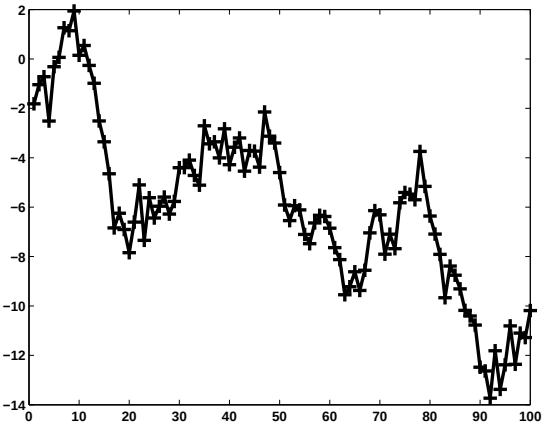

FIGURE 3.4 – *Illustration d'un processus **AR**(2) non-stationnaire pour $k_P = 0.2$.*

Conclusion

Au vu des différentes caractéristiques de la commande proportionnelle, il est manifeste que, implémentée toute seule, elle est inadaptée dans un contexte où les procédés sont sujets à des perturbations stochastiques non stationnaires. Une première limitation de cette commande est son incapacité à annuler notamment l'erreur statique ζ_∞, qui apparaît suite à l'application d'une perturbation en échelon constant. Certes, l'erreur statique diminue lorsqu'on augmente le gain, l'unique degré de liberté dont dispose cette commande. Mais, son augmentation au delà de certaines limites risque d'induire une réponse énergétique qui n'est pas sans danger : l'instabilité du système. L'étude de l'évolution de la variance de la variable de sortie nous a fourni une seconde limitation majeure : l'action **P** est impuissante face à une perturbation de nature marche aléatoire et à toute perturbation non stationnaire. L'indice **AVar** est alors théoriquement infinie, indépendamment du gain k_P.

3.2.2 La commande intégrale

Approche automaticienne

L'action intégrale est obtenue par la loi suivante :

$$u_{k+1} = u_k - k_I y_k \tag{3.23}$$

où k_I est le gain intégral. La fonction de transfert $\mathbf{T}_{BF}$ relative aux perturbations se déduit aisément de $\mathbf{T}_{BO} = \frac{bk_I q^{-1}}{1-q^{-1}}$. Nous avons :

$$\mathbf{T}_{BF} = \frac{1}{1+\mathbf{T}_{BO}} = \frac{1-q^{-1}}{1-(1-bk_I)q^{-1}} \tag{3.24}$$

Il s'agit d'un système du premier ordre. L'expression du pôle est $p_1 = 1 - b\,k_I$. Nous posons $f = b\,k_I$. Afin de garantir la stabilité du système en boucle fermée, f doit être strictement compris entre 0 et 2. Le graphe 3.5 présente l'évolution du temps de réponse à 2% en fonction de f. Un gain f proche de l'unité favorise une réponse rapide du système qui ralentit d'une façon exponentielle dès que f s'écarte de 1. A titre d'exemple, pour $f = 0.8$ (ou d'une façon symétrique 1.2), t_{rep} est de l'ordre de 6 cycles.

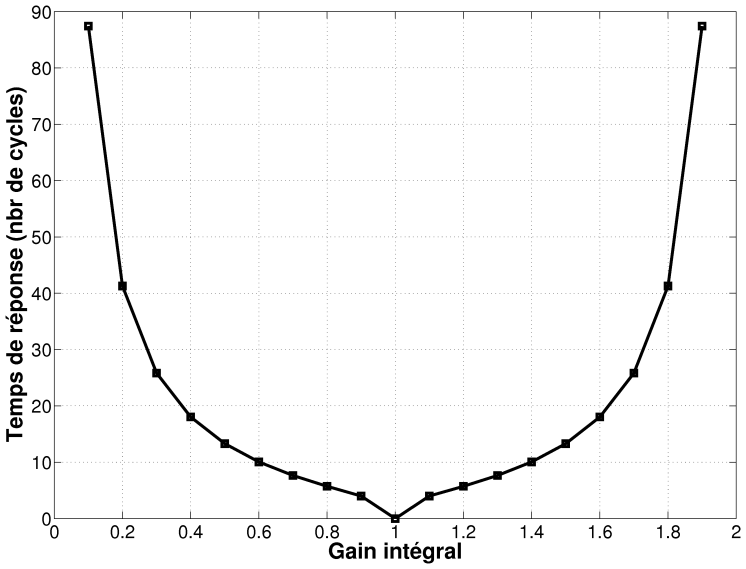

FIGURE 3.5 – *Evolution du temps de réponse du système asservi en fonction du gain intégral*

Enfin, en ce qui concerne l'écart ou le biais ζ_k, il est formulé à chaque cycle k par l'équation 3.25. Le calcul de ζ_∞ pour une perturbation en échelon donne une valeur nulle. Ceci garantit un rejet de ces perturbations, désignées dans l'industrie sous le terme anglo-saxon de *shift*.

$$\zeta_k = -\frac{1-q^{-1}}{1-(1-f)q^{-1}}P_k \tag{3.25}$$

Approche statisticienne

D'une manière équivalente à celle adoptée avec la commande proportionnelle, nous nous proposons de résoudre l'équation aux différences qui régit le système bouclé. Une solution élégante est exposée dans l'ouvrage de tsZ-Sin Siu [Hardt et Siu (2002)]. L'équation du système bouclé est donnée par la relation suivante :

$$y_{n+1} - (1-f)y_n = P_{n+1} - P_n \tag{3.26}$$

La solution de cette équation est la somme de la solution homogène et de la solution particulière. La solution homogène, $y_n = (1-f)^n y_0$, est identique à celle de la commande proportionnelle. La solution particulière est une combinaison linéaire de solutions particulières [6] de l'équation 3.17. La solution globale est alors :

$$y_{n+1} = (1-f)^{n+1}y_0 + P_{n+1} - (1-f)^n P_0 + (-f)\sum_{i=0}^{n-1}((1-f)^i P_{n-i}) \tag{3.27}$$

Sous forme matricielle, ceci peut s'écrire

$$y_{n+1} = (-f)^{n+1}y_0 + \mathbf{Q}^T\mathbf{P} \tag{3.28}$$

6. Il s'agit simplement d'une soustraction à faire.

$$\text{où } \mathbf{Q} = \begin{bmatrix} 1 \\ -f \\ -f(1-f) \\ \dots \\ -f(1-f)^{n-1} \\ -(1-f)^n \end{bmatrix} \text{ et } \mathbf{P} = \begin{bmatrix} P_{n+1} \\ P_n \\ \dots \\ \dots \\ \dots \\ P_0 \end{bmatrix}$$

Nous déduisons l'expression de la variance de la sortie en fonction de **Q** et de la matrice de covariance du vecteur colonne **P**.

$$\sigma^2_{y_{n+1}} = \mathbf{Q}^T \mathbf{\Gamma}_{n+1} \mathbf{Q} \tag{3.29}$$

$$\text{où } \mathbf{\Gamma}_n = \begin{bmatrix} Cov(P_0, P_0) & Cov(P_0, P_1) & \dots & Cov(P_0, P_n) \\ Cov(P_1, P_0) & Cov(P_1, P_1) & \dots & Cov(P_1, P_n) \\ \dots & \dots & \dots & \dots \\ Cov(P_n, P_0) & Cov(P_n, P_1) & \dots & Cov(P_n, P_n) \end{bmatrix}$$

Perturbation en Bruit Blanc La matrice $\mathbf{\Gamma}_n$ est réduite dans ce cas à la matrice identité multipliée par σ^2. Nous déduisons la variance de la sortie donnée par la formule suivante :

$$\sigma^2_{y_{n+1}} = \sigma^2(1 + (1-f)^{2n} + f[\frac{1-(1-f)^{2n}}{2-f}]) \tag{3.30}$$

Dans les travaux de Siu, le terme $(1-f)^{2n}$ est absent. Pour n suffisamment grand, nous constatons que la variance du système en boucle ouverte, σ^2, est amplifiée par le terme ρ (voir equation 3.31), appelé **rendement absolu**[7] [Box et Kramer (1992), Tsung et al. (1998)].

$$\rho = \frac{1}{1 - \frac{f}{2}} \tag{3.31}$$

Pour des valeurs de f égales à 0.5 et 1.5, ρ est égal respectivement à 1.33 et 4. Dans le cas de perturbations de nature bruit blanc, l'action intégrale est aussi peu performante que l'action proportionnelle. Ceci n'est pas surprenant comme nous l'avons déjà expliqué au paragraphe 3.1 : un procédé sujet au bruit blanc est un procédé sous contrôle statistique. Il n'est pas recommandé de réaliser un bouclage [Deming (2000), MacGregor (1990)].

Marche Aléatoire L'équation aux différences du système, dans ce cas, est un processus $\mathbf{AR}(1)$, comme le montre l'équation 3.32. Le rendement absolu ρ est formulé dans le tableau 3.1. La variance asymptotique de la sortie **AVar** est minimale autour d'un gain optimum f égal à l'unité. Dans la mesure où la variance de la déviation y_k est égale, dans ce cas, à celle des ε_k, la commande intégrale est confondue avec la commande **MMSE** (**M**inimum **M**ean **S**quare **E**rror). Nous y reviendrons plus loin dans ce chapitre.

$$y_{n+1} = (1-f)y_n + \varepsilon_{n+1} \tag{3.32}$$

7. Le terme anglais utilisé par Tsung dans [Tsung et al. (1998)] est *absolute efficiency*

Perturbation	θ	ω	$\rho = \frac{\mathbf{AVar}}{\sigma^2}$
Marche aléatoire	0	1	$\frac{1}{f(2-f)}$
IMA(1,1)	$\lvert\theta\rvert < 1$	0	$1 + \frac{(1-f-\theta)^2}{f(2-f)}$
ARIMA(1,1)	$\lvert\theta\rvert < 1$	$\lvert\omega\rvert < 1$	$\frac{\left(1+\theta^2\right)(1+\omega-f\,\omega)-2\,\theta\,(1+\omega-f)}{f\,(1+\omega)\,(2-f)\,(1-\omega+f\,\omega)\,(1-\omega)}$

TABLE 3.1 – *Résultats relatifs à la commande intégrale. Tableau récapitulatif des rendements absolus ρ en fonction de la perturbation* $P_k = \frac{1-\theta\,q^{-1}}{(1-q^{-1})(1-\omega\,q^{-1})}$.

Perturbation Intégrée à Moyenne Mobile IMA$(1,1)$ L'équation aux différences du système est transcrite ci-dessous (équation 3.33). Il s'agit d'un processus **IMA**(1,1). A partir de ses coéfficents et en se basant sur les expressions des fonctions d'autocovariance [Castillo (2002)], nous avons pu exprimer la variance de la sortie, donnée dans le tableau 3.1 . Nous constatons que pour un gain f égal à $(1-\theta)$, **AVar** est minimale. La commande I est encore une fois équivalente à la commande **MMSE**, dans le cas $f = 1-\theta$.

$$y_{n+1} = (1-f)y_n + \varepsilon_{n+1} - \theta\varepsilon_n \tag{3.33}$$

où $|\theta| < 1$ est le paramètre de la perturbation **IMA**(1,1), appliquée au procédé.

Perturbation Autorégressive Intégrée à Moyenne Mobile ARIMA$(1,1,1)$ Dans le cas d'une perturbation **ARIMA**$(1,1,1)$, caractérisée par le couple (θ,ω) [8], le rendement absolu ρ est fonction du gain f et de (θ,ω). Il est donné dans le tableau 3.1. A partir de la dérivée de la fonction ρ [9], nous pouvons toujours déduire une expression formelle du gain optimum f minimisant la variance de la variable de sortie. L'expression étant très complexe et sans grand intérêt pour l'interprétation des résultats, nous avons opté pour une minimisation numérique en utilisant un algorithme de type Simplex.

La perturbation **ARIMA** est un processus non-stationnaire. Alors appliquer une commande constante (boucle ouverte) serait synonyme à terme d'une variance théoriquement infinie. Réaliser au contraire une commande intégrale aura toujours un effet bénéfique, vis à vis du rendement ρ, comme le montre les figures 3.6 et 3.7. Le gain f optimum est en effet toujours non nul quelque soit le couple (θ,ω). Un rendement unitaire n'est en revanche pas toujours réalisable. Ceci est le cas en particulier des couples (θ,ω) proches de $(-1,1)$ et aussi de $(1,-1)$. Enfin, nous retrouvons une particularité relevée par Tsung de la figure 3.7 : une symétrie $\rho(\theta,\omega) = \rho(-\theta,-\omega)$ [Tsung et al. (1998)].

Conclusion

Pour clôturer cette section où nous nous sommes intéressés à la commande intégrale pure, il est essentiel de souligner le potentiel de cette commande. D'un point de vue automatique, nous avons vu que son domaine de stabilité est plus grand que la commande proportionnelle. Face à des perturbations en échelon déterministe, elle assure une élimination progressive de l'erreur statique, ainsi qu'un temps de réponse

8. L'ensemble de variation du couple (θ,ω) est $]-1,1[\times]-1,1[$, Voir références [Castillo (2002), Box et al. (1994)]

9. Il est important de vérifier que la dérivée seconde à l'optimum est positive

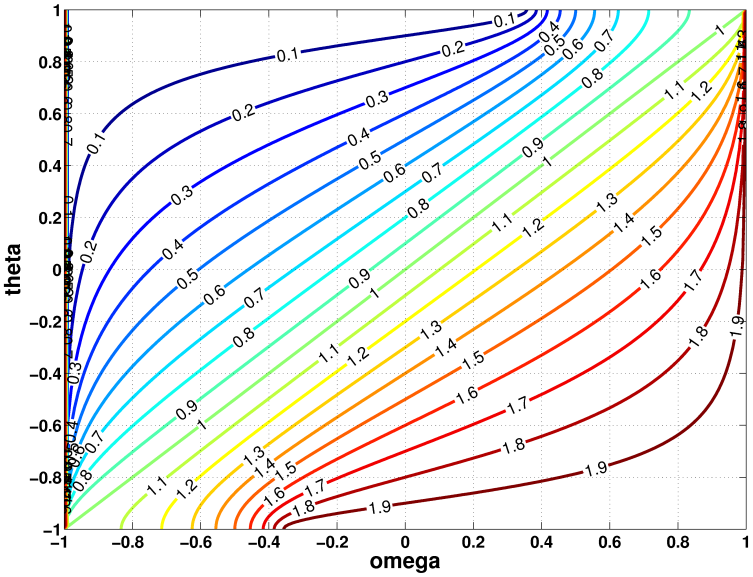

FIGURE 3.6 – *Résultats relatifs à la commande intégrale. Les courbes de niveau du gain f optimum en fonction du couple (θ,ω).*

faible pour un gain f proche de 1. D'un point de vue statistique, nous avons mis en exergue son impact positif sur la variance σ_y^2, en présence de perturbations non-stationnaires (**ARIMA**$(1,1,1)$, **IMA**$(1,1)$, marche aléatoire). C'est pour l'ensemble de ces qualités, que la commande intégrale a toujours été la loi de commande la plus présente dans la fabrication des composants microélectroniques. En effet, nous expliquerons plus loin que la loi de commande EWMA est équivalente à une commande intégrale pure.

3.2.3 La Commande Proportionnelle Dérivée

Approche automaticienne

La loi de commande proportionnelle dérivée se traduit par l'équation suivante :

$$u_{k+1} = -k_P(y_k + T_D(y_k - y_{k-1})) \tag{3.34}$$

où k_P est la gain proportionnel et $k_D = k_P\, T_D$ est la constante de dérivation. Peu d'ouvrages de statistique appliquée se sont intéressés à cette commande. DelCastillo en décrit l'esprit très brièvement dans [Castillo (2002)]. Il souligne le fait que la commande **PD** emploie une action proportionnelle à une prévision de l'erreur y_k à l'horizon T_D.

$$y_{k+T_D} \approx y_k + T_D\, \nabla\, y_k \tag{3.35}$$

Intuitivement, nous pouvons alors interpréter cette commande à travers les deux cas extrêmes :

1. Si T_D est trop grand, la prévision est médiocre et l'asservissement l'est aussi,
2. au contraire, si T_D tend vers 0, nous avons simplement une commande **P** pure.

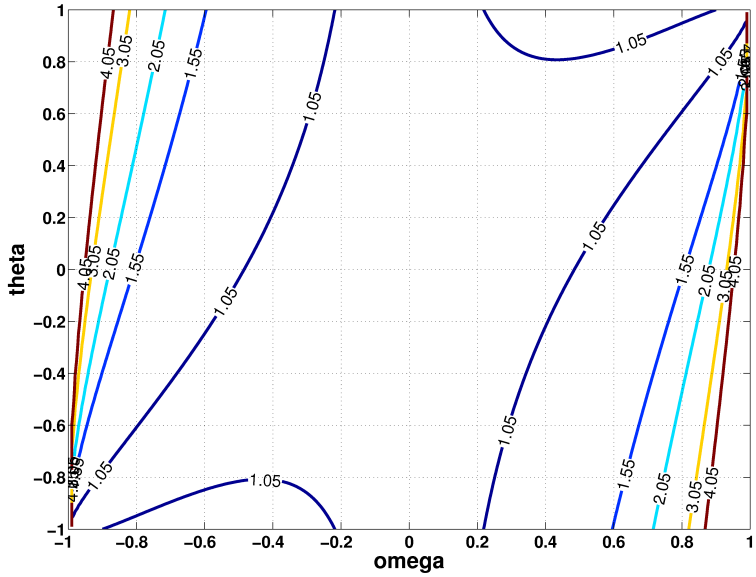

FIGURE 3.7 – *Résultats relatifs à la commande intégrale. Les courbes de niveau de ρ minimal en fonction du couple (θ,ω). Le ρ minimal correspond bien-entendu au gain f optimum.*

DelCastillo rajoute aussi que la composante dérivée d'une commande **PID** est souvent réservée aux systèmes dynamiques de second ordre. Dans ce paragraphe, nous allons chercher à comprendre les raisons d'un tel désintérêt pour cette commande dans le cas des procédés stochastiques. Reprenons la démarche que nous avons déployée jusqu'alors. La fonction de transfert en boucle ouverte s'écrit :

$$\mathbf{T}_{BO} = b\,k_P\,q^{-1}(1 + T_D\,(1 - q^{-1}))$$

Les conditions de stabilité du système en boucle fermée sont

$$b\,k_P > -1 \tag{3.36}$$
$$b\,k_P\,(1 + 2\,T_D) < 1 \tag{3.37}$$
$$-1 < b\,k_P\,T_D < 1 \tag{3.38}$$

Sous l'effet d'une perturbation en échelon déterministe d'amplitude κ, le biais en régime permanent ζ_∞ est non nul (voir équation 3.39). L'action dérivée n'a effectivement aucun effet sur le régime permanent, défini par la seule composante proportionnelle.

$$\zeta_\infty = \lim(1 - q^{-1})\zeta(q^{-1}) = -\frac{\kappa}{1 + bk_P} \tag{3.39}$$

Approche statisticienne

Perturbation en Bruit Blanc La représentation du rendement absolu ρ, exprimé par la relation 3.40, est réalisé sous forme de courbes de niveau 3.8. Le constat est encore une fois identique aux sections précédentes : le rendement absolu ρ est strictement supérieur à 1, sauf lorsque le correcteur est mis hors service ((k_P,T_D)=(0,0)). **Aussi, ρ croît plus rapidement en fonction de k_P (axe vertical) qu'en fonction de T_D (axe horizontal).**

$$\rho = \frac{\sigma_y^2}{\sigma_\varepsilon^2} = \frac{1 - b\,k_P\,T_D}{(1 + b\,k_P\,T_D)(1 + b\,k_P)(1 - b\,k_P\,(1 + 2\,T_D))} \tag{3.40}$$

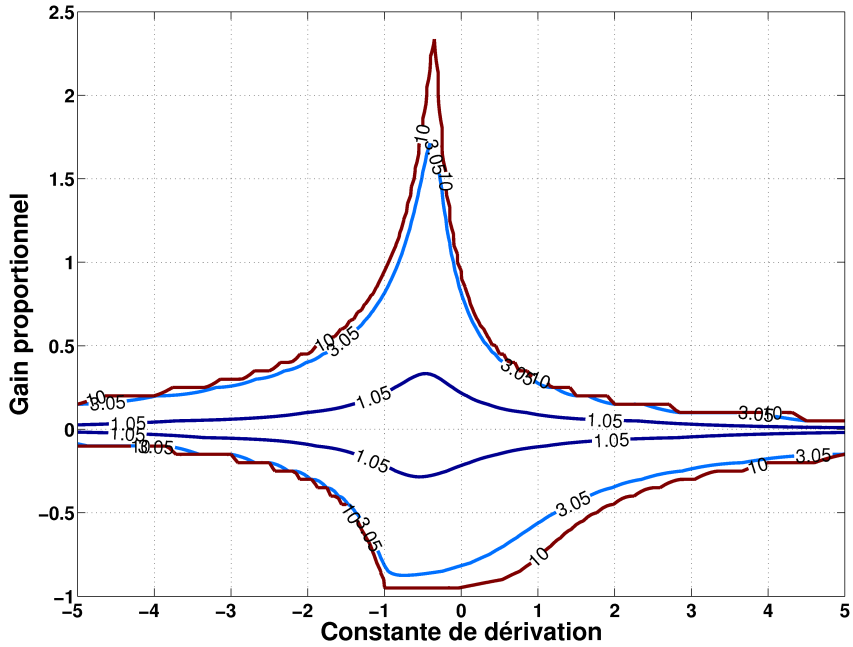

FIGURE 3.8 – *Résultats relatifs à la commande proportionnelle dérivée. Les courbes de niveau de ρ en fonction du couple (T_D,k_P). ρ est minimal et égal à 1 pour un k_P nul, autrement lorsque le contrôleur est off. Le rendement absolu augmente d'une manière trop importante aux limites du domaine de stabilité des paramètres T_D etk_P.*

Marche Aléatoire Dans ce cas, la variance asymptotique **AVar** est théoriquement infinie. D'une manière identique à la commande proportionnelle pure, la commande **PD** est incapable de stabiliser le système sous une perturbation non stationnaire. Il est alors inutile d'étudier les cas des processus **IMA**$(1,1)$ et **ARIMA**$(1,1,1)$.

3.2.4 La commande proportionnelle intégrale

La commande **PI** a beaucoup intéressé Box et luceno [Box et Luceño (1997b), Box et Luceño (1997a)], qui ont élaboré une excellente description de sa structure et de certaines applications en contrôle des procédés industriels. D'un autre côté, la robustesse de cette commande a été particulièrement bien détaillée par Tsung et al [Tsung et al. (1998)].

Approche automaticienne

La commande **PI** est une association d'une composante proportionnelle et une composante intégrale. Elle est donnée par la relation suivante [10] :

$$u_{k+1} = -k_P y_k - k_I \sum_{i=0}^{k} y_i \tag{3.41}$$

10. D'autres formulations équivalentes existent

Les fonctions de transfert en boucle ouverte et en boucle fermée en sont déduites et exprimées ci-dessous :

$$\mathbf{T}_{BO} = K\frac{1-\alpha\, q^{-1}}{1-q^{-1}}q^{-1} \tag{3.42}$$

$$\mathbf{T}_{BF} = \frac{1-\ q^{-1}}{1-\ (1-K)\, q^{-1} -\ \alpha\, K\, q^{-2}} \tag{3.43}$$

où $K = b\,(k_P + k_I)$, et $\alpha = \frac{k_P}{k_P+k_I}$.

En s'appuyant sur le critère de Jury, qui étudie la position des pôles dans le plan complexe $\boldsymbol{z}$, les conditions nécessaires et suffisantes pour garantir la stabilité du système en **BF** sont établies.

$$(1-\alpha)\, K = b\, k_I > 0 \tag{3.44}$$

$$(1+\alpha)\, K = b\,(2\, k_P + k_I) < 2 \tag{3.45}$$

$$-1 < \alpha\, K = b\, k_P < 1 \tag{3.46}$$

Si nous reprenons la notation utilisée précédemment, nous avons :

$$f > 0 \tag{3.47}$$

$$(2g + f) < 2 \tag{3.48}$$

$$-1 < g < 1 \tag{3.49}$$

Une caractéristique importante du système bouclé est le temps de réponse t_{rep}. Il est estimé à partir d'une expression empirique qui fait appel à la notion de pôle dominant. Dans les limites du domaine de stabilité, nous observons en figure 3.9 que t_{rep} dépend exclusivement du gain g, sur l'essentiel de la plage de variation du gain f. t_{rep} **est minimal pour un correcteur intégral pur à un gain** f **unitaire, et se dégrade lorsque** g **s'écarte de** 0**. A ce stade, le seul apport de la composante proportionnelle est un domaine de stabilité élargie. Elle conduit néanmoins à une dégradation de la dynamique de la réponse du système aux perturbations**. Enfin, la commande **PI** hérite de la commande **I** sa capacité à rejeter les perturbations de type échelon : $\zeta_\infty = 0$.

Approche statisticienne

Considérons l'équation aux différences du système bouclé 3.50. D'une manière identique aux commandes précédentes, nous allons estimer la variance asymptotique de la sortie pour diverses perturbations.

$$y_{n+1} = (1-K)y_n + \alpha\, K\, y_{n-1} + P_{n+1} - P_n \tag{3.50}$$

Perturbation en Bruit Blanc L'équation 3.50 est alors un processus **ARMA**$(2,1)$. Après manipulation des fonctions d'autocovariance, nous avons obtenu le rendement absolu ρ, formulé dans le tableau 3.2. Pour vérifier la validité de cette expression, nous pouvons toujours remplacer g et f par 0, ce qui correspond respectivement à une commande intégrale pure et une commande proportionnelle pure. Nous retrouvons bien

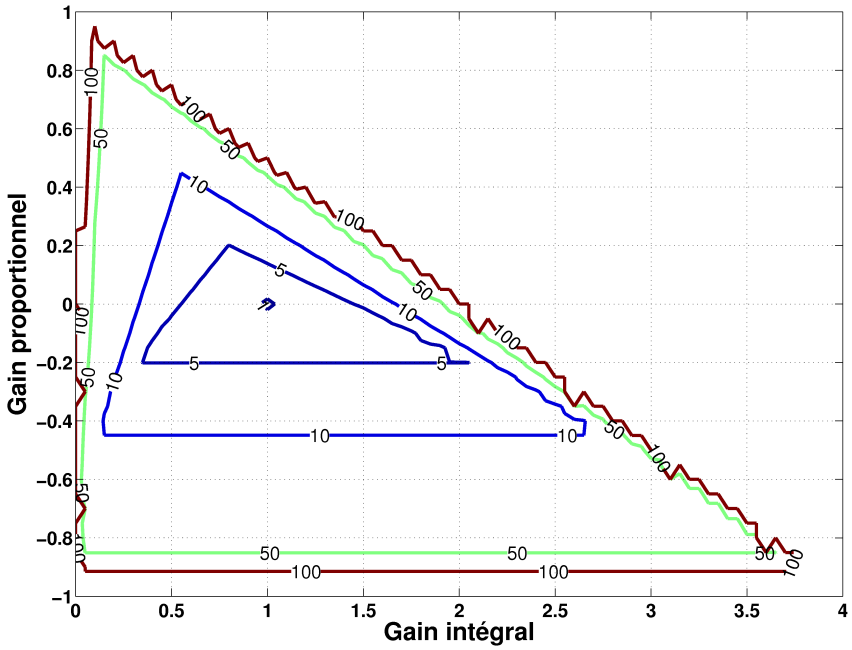

FIGURE 3.9 – *Résultats relatifs à la commande proportionnelle intégrale. Les courbes de niveau de t_{rep} en fonction du couple (g,f). t_{rep} est minimal pour une commande intégrale pure avec $f = 1$. Nous constatons que t_{rep} ne dépend quasiment pas de f. A noter que le domaine de stabilité du couple (f,g) est $]0,4[\times]-1,1[$.*

heureusement les expressions 3.21 et 3.31. En s'appuyant sur le graphe 3.10, nous retrouvons aussi le fait que la commande **PI**, comme toute sorte de bouclage (feedback), engendre l'inflation de la variance d'un procédé sous contrôle statistique. Choisir des gains g et f faibles proches de 0 permet de limiter ρ à des valeurs modérées ($\simeq 1$).

$$\rho = \frac{\sigma_y^2}{\sigma_\varepsilon^2} = \frac{2}{(1+g)(2-2g-f)} \tag{3.51}$$

Marche Aléatoire Dans le cas d'une perturbation de type marche aléatoire, la différence $P_{n+1} - P_n$ se réduit au seul terme ε_{n+1}. Le calcul du ratio ρ fournit la relation indiquée dans le tableau 3.2. Le tracé des courbes de niveau (figure 3.11) indique que la commande **PI** serait capable de réduire à 1 le rapport ρ, pourvu que le gain f soit proche de 1 et que le gain g soit faible. Autrement, dans le cas d'une commande intégrale pure.

Perturbation Intégrée à Moyenne Mobile IMA$(1,1)$ Nous reproduisons toujours la même démarche utilisée jusqu'à lors. Après identification de l'expression du rendement absolu ρ (tableau 3.2), nous avons tracé les courbes de niveau pour un θ quelconque (figure 3.12, $\theta = 0.5$). Comme nous l'avons démontré sous le paragraphe relatif à la commande intégrale, la loi de commande capable de réduire la variance de la sortie à celle des résidus ε_k est simplement une commande intégrale pure dont le gain est de $f = 1 - \theta$.

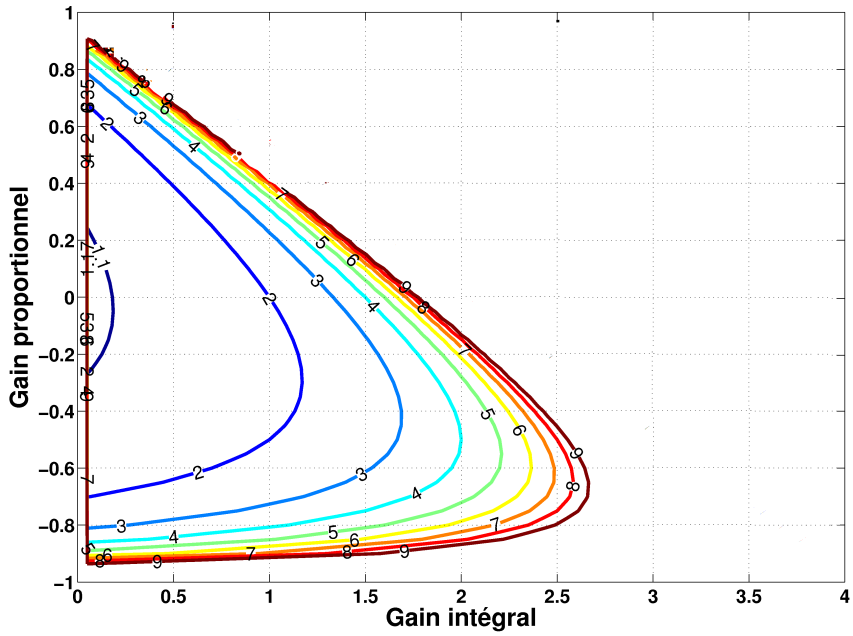

FIGURE 3.10 – *Résultats relatifs à la commande proportionnelle intégrale en présence d'une perturbation de nature bruit blanc. Les courbes de niveaux du rendement ρ en fonction des gains g et f. Le système étant sous contrôle statistique, ρ est naturellement minimal dans le cas $f = g = 0$, ie : boucle ouverte. A noter que le domaine de stabilité du couple (f,g) est $]0,4[\times]-1,1[$.*

Perturbation Autorégressive Intégrée à Moyenne Mobile ARIMA$(1,1,1)$ La minimisation du rendement ρ, dont l'expression est donné dans le tableau 3.2, pointe du doigt le rôle bienfaisant de l'action proportionnelle, qui, associée à l'action intégrale, repousse les fortes valeurs de ρ (> 2) dans des zones à la limite de l'ensemble de variation du couple (θ, $|\omega| > 0.9$), comme en témoignent les figures 3.13, 3.14 et 3.15. A titre d'exemple, $\rho(\omega = -0.8, \theta = 0.4)$ est de 2.05 sous une commande intégrale pure. Elle est de 1.4 si nous rajoutons une action proportionnelle optimale. En clair, la commande **PI** est une commande **MMSE** sur une grande partie (centrale) de l'espace de variation du couple (θ, ω), où ρ est égal à 1. A noter que dans les cas particuliers des perturbations de nature **IMA**$(1,1)$ (la droite $\omega = 0$) et marche aléatoire (la droite $\omega = \theta$), le gain proportionnel g est nul, ce qui rejoint les résultats du paragraphe 3.2.2. Le gain intégral f est égal à 1 pour une perturbation de nature marche aléatoire.

Perturbation	θ	ω	$\rho = \frac{\mathbf{AVar}}{\sigma^2}$
Marche aléatoire	0	1	$\frac{1-g}{f(1+g)(2-2g-f)}$
IMA(1,1)	$\lvert\theta\rvert < 1$	0	$\frac{(1-g)(1+\theta^2)-2\theta(1-g-f)}{f(1+g)(2-2g-f)}$
ARIMA(1,1)	$\lvert\theta\rvert < 1$	$\lvert\omega\rvert < 1$	$\frac{\left(1+\theta^2\right)\left(1-\phi_2-\phi_3(\phi_1+\phi_3)\right)-2\,\theta\,(\phi_1+\phi_2\,\phi_3)}{(1+\phi_3-\phi_2+\phi_1)\,(1-\phi_3-\phi_2-\phi_1)\left(1-\phi_3{}^2+\phi_1\,\phi_3+\phi_2\right)}$

TABLE 3.2 – *Résultats relatifs à la commande proportionnelle intégrale. Tableau récapitulatif des rendements absolus ρ en fonction de la perturbation $P_k = \frac{1-\theta\,q^{-1}}{(1-q^{-1})(1-\omega\,q^{-1})}$ et des constantes $\phi_1 = 1-g-f+\omega$, $\phi_2 = g-\omega\,(1-g-f)$ et $\phi_3 = -\omega\,g$.*

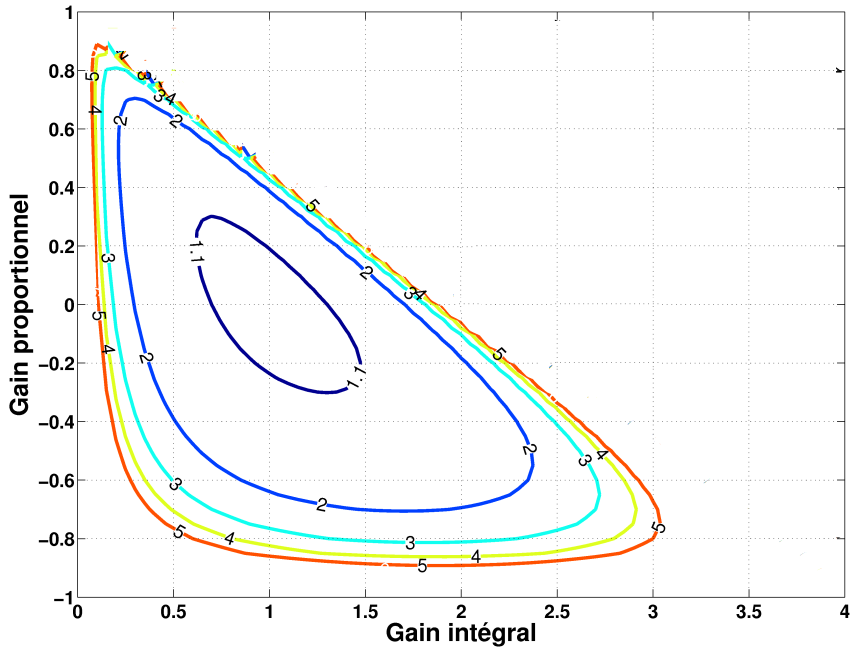

FIGURE 3.11 – *Résultats relatifs à la commande intégrale proportionnelle en présence d'une perturbation de nature marche aléatoire. Les courbes de niveaux du rendement ρ en fonction de g et f. ρ est minimal pour une commande intégrale pure avec un gain $f = 1$.*

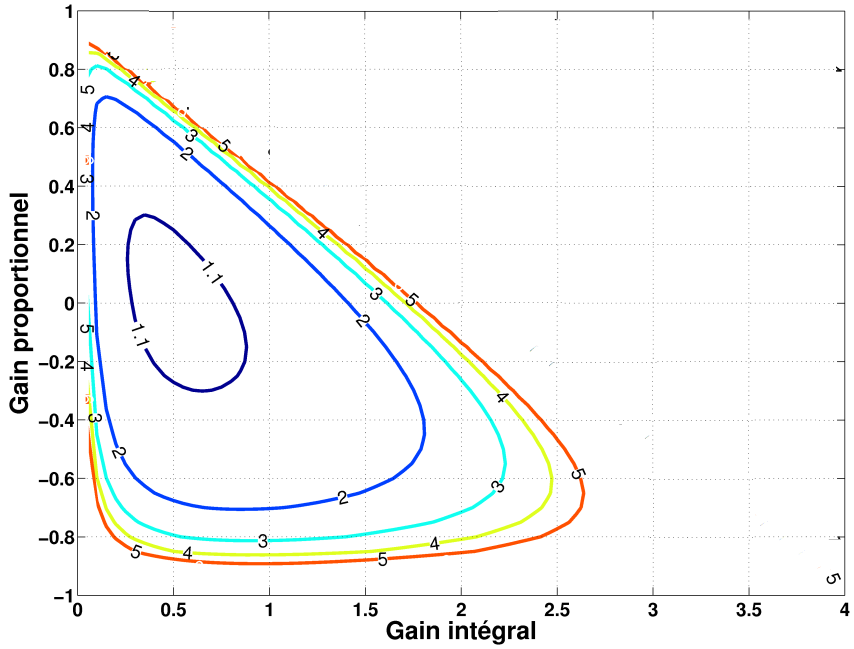

FIGURE 3.12 – *Résultats relatifs à la commande intégrale proportionnelle en présence d'une perturbation de nature* ***IMA***$(1,1)$*, avec θ=0.5. ρ est minimal pour une commande intégrale pure dont le gain f est égal à $1-\theta$=0.5.*

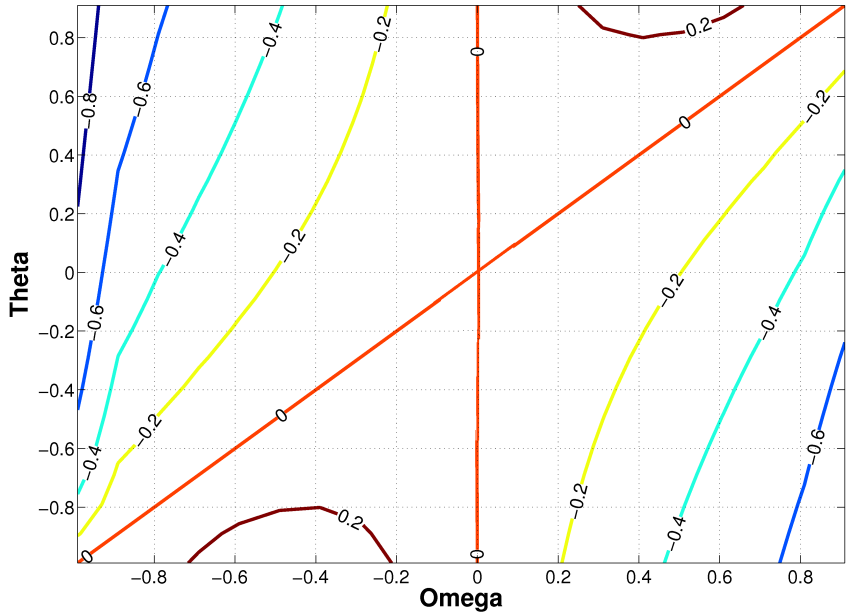

FIGURE 3.13 – *Résultats relatifs à la commande intégrale proportionnelle en présence d'une perturbation de nature* ***ARIMA***$(1,1,1)$. *Les courbes de niveaux du gain proportionnel* g *optimal en fonction de* θ *et* ω. g *est nul dans le cas des perturbations de nature* ***IMA***$(1,1)$ *(la droite* $\omega = 0$*) et marche aléatoire (la droite* $\omega = \theta$*).*

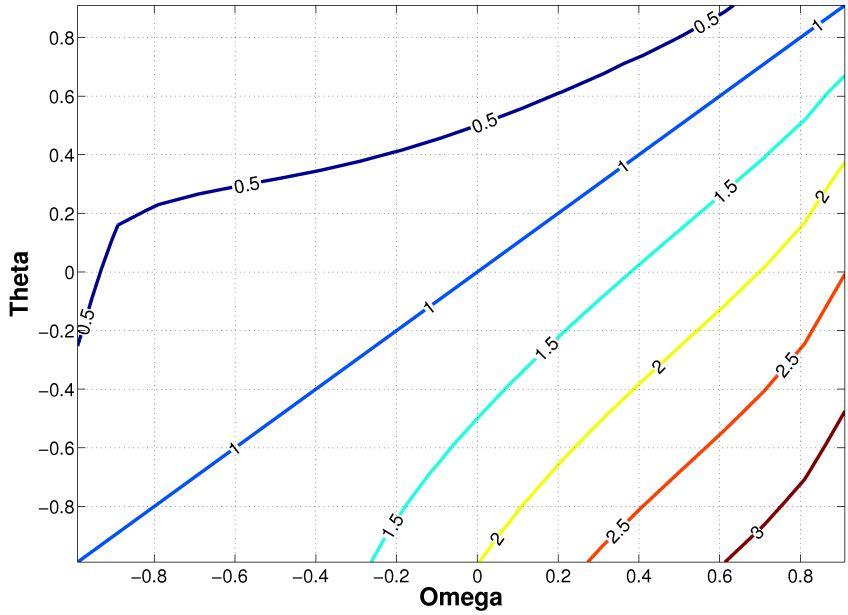

FIGURE 3.14 – *Résultats relatifs à la commande intégrale proportionnelle en présence d'une perturbation de nature* ***ARIMA***$(1,1,1)$. *Les courbes de niveaux du gain intégral* f *optimal en fonction de* θ *et* ω. f *est égal à l'unité dans le cas des perturbations de nature marche aléatoire (la droite* $\omega = \theta$*).*

Conclusion

L'action intégrale réalise gloablement des bonnes performances : une réponse rapide pour un gain f proche de l'unité, un domaine de stabilité large, un rendement absolu ρ proche de 1 sur une bonne partie de l'espace de variation des paramètres stochastiques des perturbations non-stationnaires (marche aléatoire, **IMA**$(1,1)$, **ARIMA**(1,1,1)). L'objet de ce paragraphe était alors d'évaluer l'apport de l'action proportionnelle. Nous avons démontré que l'ajout d'une action proportionnelle élargit davantage le domaine de stabilité. Associée à la commande **I** et dans le cas d'une perturbation **ARIMA**(1,1,1), elle engendre un rendement absolu égal à 1 sur une plus grande partie de l'espace de variation du couple (θ, ω) que l'action intégrale seule. La commande **PI** est alors confondue avec une commande **MMSE**. L'action **P** présente néanmoins peu d'intérêt dans le cas de perturbations de nature marche aléatoire et **IMA**$(1,1)$. **Elle dégrade, par ailleurs, le temps de réponse** t_{rep}.

3.2.5 La Commande Proportionnelle Intégrale Dérivée

Approche automaticienne

La commande **PID** est l'association des trois actions précédentes. Une formulation mathématique courante est :

$$u_k = k_P\, y_k \,+\, k_I \sum y_k + k_D\,(y_k - y_{k-1}) \iff \nabla\, u_k = c_1\, y_k \,+\, c_2\, y_{k-1} + c_3\, y_{k-2} \quad (3.52)$$

où y_k est l'écart entre la variable de sortie et la consigne, $c_1 = k_P + k_I + k_D$, $c_2 = -(k_P + 2\,k_D)$ et $c_3 = k_D$.

La fonction de transfert en boucle fermée relative à la perturbation est du troisième ordre, elle s'écrit :

$$\mathbf{T}_{BF} = \frac{1 - q^{-1}}{1 + \,(b\,c_1 - 1)\,q^{-1} + \,b\,c_2\,q^{-2} + \,b\,c_3\,q^{-3}} \quad (3.53)$$

Afin de répondre aux exigences de la stabilité en **BF**, les paramètres du **PID** doivent satisfaire les conditions suivantes :

$$|b\,c_3| < 1 \quad (3.54)$$
$$b\,(c_1 + c_2 + c_3) > 0 \quad (3.55)$$
$$b\,(c_1 - c_2 + c_3) < 2 \quad (3.56)$$
$$b\,c_3\,(b\,c_3 - b\,c_1 + 1) + b\,c_2 < 1 \quad (3.57)$$

Outre les spécifications en stabilité, des bonnes performances en précision et en temps de réponse sont essentielles. Avec la commande **PID**, une erreur statique nulle ($\zeta_\infty = 0$), suite à l'application d'un échelon, est garantie grâce à son action intégrale. Quant à la rapidité du système, nous proposons d'étudier l'effet des différentes composantes de la commande sur l'amplitude du pôle dominant. Prenons l'exemple de la figure 3.16. Pour chaque valeur du gain k_D, $|p_m|$ est rendu minimal en faisant varier le couple (k_I, k_P) dans la région de stabilité spécifiée plus haut. Les tracés 3.17 et 3.18 sont construits d'une manière identique.

A la vue de ces graphes, nous constatons plusieurs points :

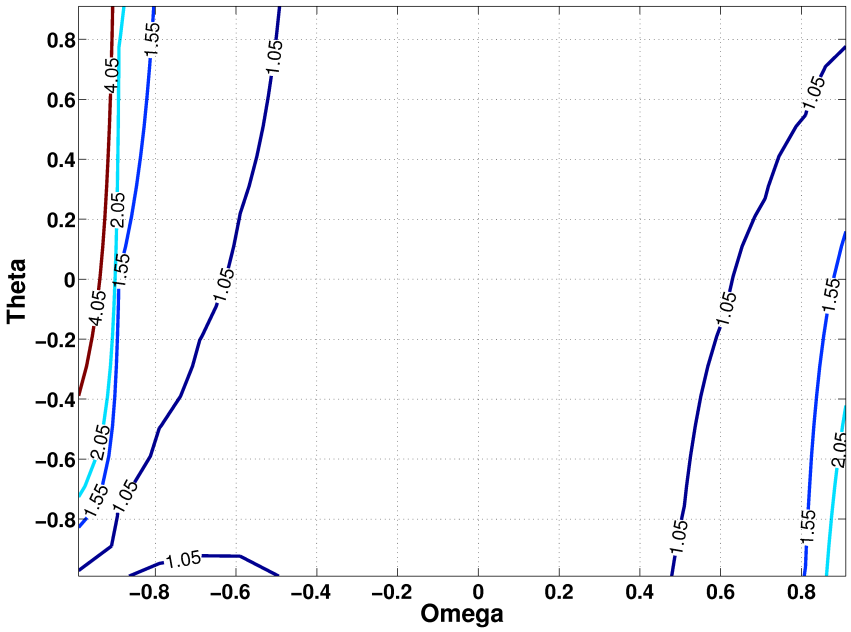

FIGURE 3.15 – *Résultats relatifs à la commande intégrale proportionnelle en présence d'une perturbation de nature* $\boldsymbol{ARIMA}(1,1,1)$. *Les courbes de niveaux du rendement absolu* ρ. *Grâce à une action* ***PI*** *optimale,* ρ *demeure proche de* 1 *sur l'essentiel de l'espace de variation du couple* (θ, ω).

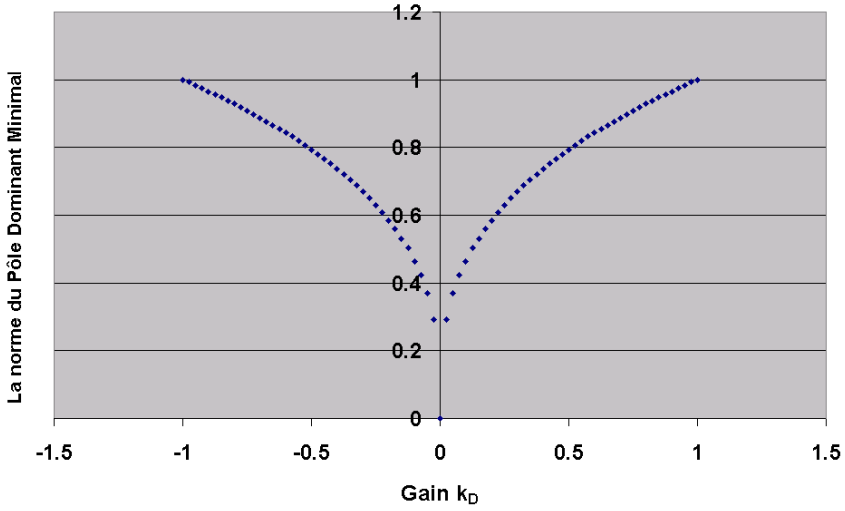

FIGURE 3.16 – *Evolution de l'amplitude du pôle dominant minimal en fonction du gain de la commande dérivée* k_D.

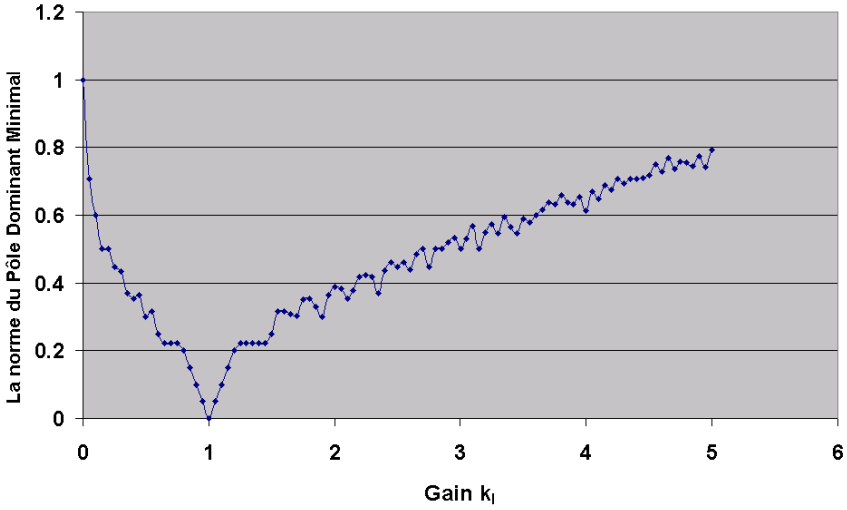

FIGURE 3.17 – *Evolution de l'amplitude du pôle dominant minimal en fonction du gain de la commande intégrale k_I.*

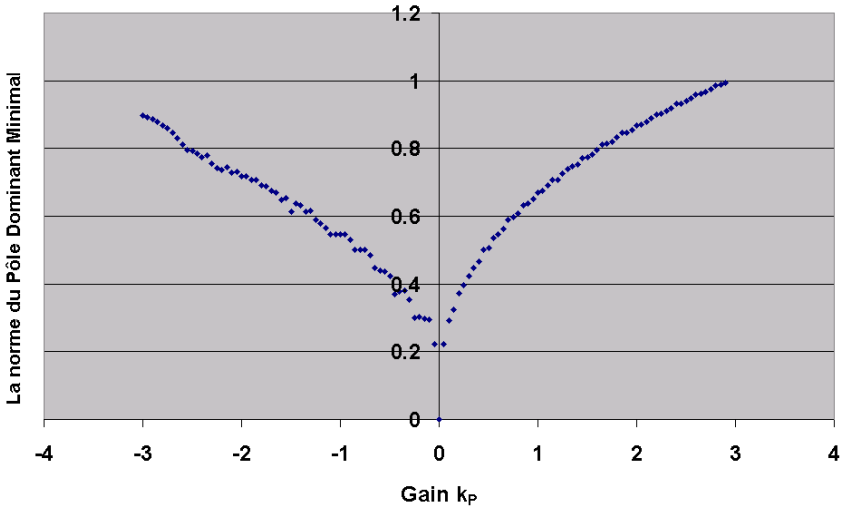

FIGURE 3.18 – *Evolution de l'amplitude du pôle dominant minimal en fonction du gain de la commande proportionnelle k_P.*

⋊ Le temps de réponse est minimal pour une commande **PID** où $k_I = 1$ et $k_D = k_P = 0$, autrement une commande intégrale pure à gain unitaire,

⋊ L'action dérivée et l'action proportionnelle (pour k_P positif) ont tendance à dégrader d'une manière exponentielle le temps de réponse t_{rep}. Pour un $k_D = 0.25$, le temps de réponse est au mieux égal à 18 cycles ($|p_m| = 0.6$),

⋊ L'évolution de $|p_m|$ en fonction du gain de l'action intégrale est linéaire et surtout moins explosive, comme c'est le cas des commandes **P** et **D**.

Approche statisticienne

A présent nous allons nous intéresser à l'inflation de la variance sous la commande **PID**. Formuler, à l'instar de la commande **PI**, le rendement absolu ρ en fonction des différents paramètres caractérisant le système (c_1, c_2, c_3, b, θ et ω) est une tâche très fastidieuse. Ici, nous avons utilisé le code fourni par Minh Vu afin de calculer le triplet des gains optimaux $(b\,k_i)_{i\in\{P,I,D\}}$ dans le cas d'une perturbation de nature **ARIMA**(1,1,1) [Vu (1992)]. Les gains sont optimaux dans le sens où ils minimisent la variance de y_k. L'étude des perturbations de nature marche aléatoire et **IMA**(1,1) ne présente pas un grand intérêt dans la mesure où nous avons d'ores et déjà démontré qu'un contrôleur intégrateur pur est optimal au sens **MMSE** dans les deux cas. L'analyse des figures 3.19, 3.20 et 3.21 montre que :

⋊ L'action intégrale est non nulle dans tout le domaine de variation du couple (θ, ω), et notamment dans le domaine ($\omega \to 1$) où la perturbation serait équivalente à une rampe déterministe [MacGregor et al. (1984)]. f tend néanmoins vers zero lorsque la perturbation tend vers un processus stationnaire ($\theta \to 1$) de type **AR**(1),

⋊ l'action dérivée demeure modérée, voire nulle dans l'essentiel du domaine $]-1,1[\times]-1,1[$,

⋊ enfin, concernant la composante proportionnelle, elle est bénéfique surtout dans le cas des $\omega \to 1$.

3.2.6 Choix des coefficients k_I, k_P et k_D

Bien que le correcteur **PID** soit une commande particulièrement répandue au sein de l'industrie, son emploi dans l'industrie du semiconducteur demeure souvent réduit à sa seule composante intégrale, notamment la fameuse commande EWMA. L'étude de la structure discrète de la commande **PID** et le choix des ses différents paramètres optimaux ont souvent été orientés vers des applications où le procédé est à flux continu (Chaudière à gaz, Climatiseur, etc). Les variables d'entrée et de sortie sont alors échantillonnées et la commande est numérique, embarquée dans un microcontrôleur. MacGregor, en s'appuyant sur deux exemples de l'industrie chimique, à savoir un réacteur chauffé à la vapeur et un échangeur thermique, a développé une méthodologie de conception d'une commande **PID** optimale [MacGregor et al. (1975)]. Elle est optimale dans le sens où elle minimise la fonction de coût formulée ci-après.

$$\text{Minimiser } E\{e_t^2 + \lambda\,[f(u_t)]^2\} \tag{3.58}$$

où E est l'opérateur de prévision, e_t est l'écart de la variable de sortie (la température) par rapport à la consigne, $f(u_t)$ est égale à u_t ou à $\nabla\, u_t$ selon qu'on veut contraindre la variance de la variable d'entrée ou la variance des changements (variations) de u_t

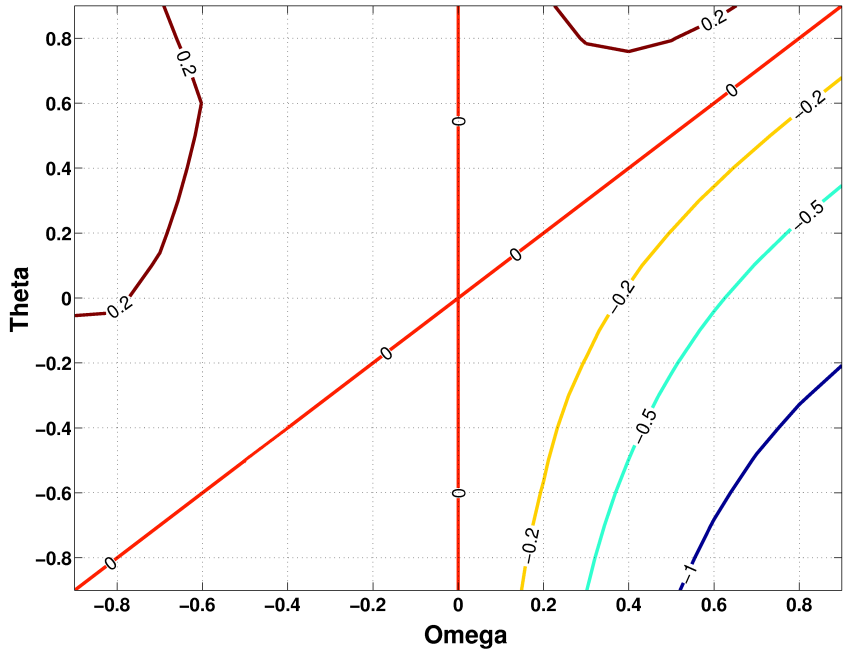

FIGURE 3.19 – *Résultats relatifs à la commande proportionnelle intégrale dérivée en présence d'une perturbation de nature* $\boldsymbol{ARIMA}(1,1,1)$. *Les courbes de niveaux du gain proportionnel* g *optimal en fonction de* θ *et* ω. g *est égal à* 0 *dans le cas des perturbations de nature marche aléatoire (la droite* $\omega = \theta$*) et* ***IMA****(1,1) (la droite* $\omega = 0$*). Le gain* g *optimal est souvent nul, sauf dans le cas* ($\omega \to 1$, $\theta \to -1$).

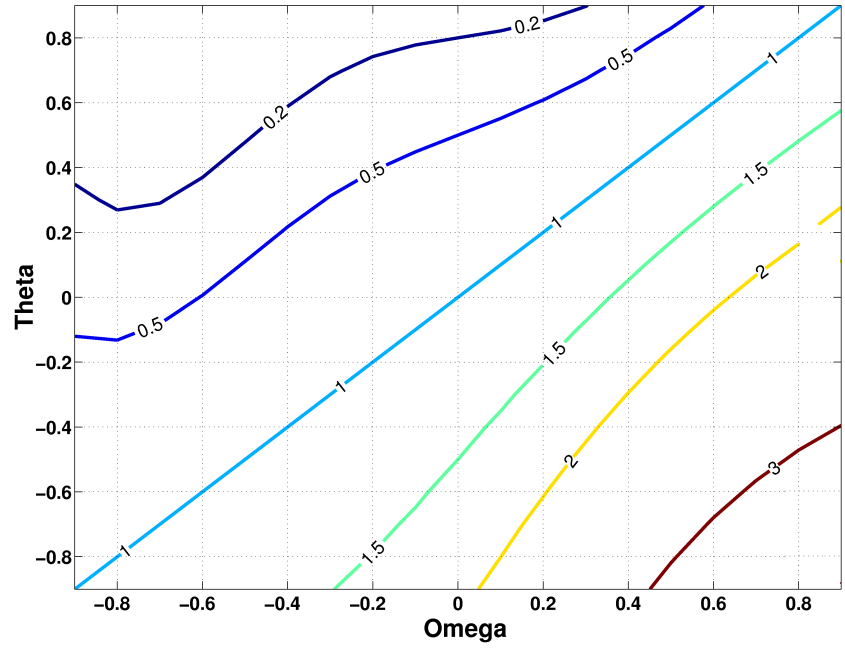

FIGURE 3.20 – *Résultats relatifs à la commande proportionnelle intégrale dérivée en présence d'une perturbation de nature* $\boldsymbol{ARIMA}(1,1,1)$. *Les courbes de niveaux du gain intégral* f *optimal en fonction de* θ *et* ω. f *est égal à l'unité dans le cas des perturbations de nature marche aléatoire (la droite* $\omega = \theta$*). Le gain* f *optimal croit d'une façon importante lorsque* $\omega \to 1$ *et* $\theta \to -1$.

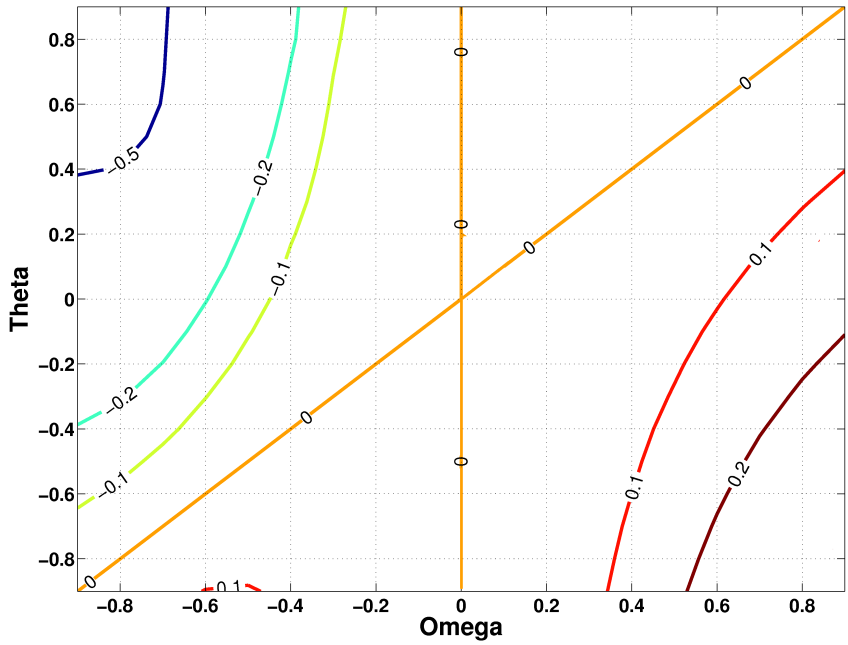

FIGURE 3.21 – *Résultats relatifs à la commande proportionnelle intégrale dérivée en présence d'une perturbation de nature* ***ARIMA***$(1,1,1)$. *Les courbes de niveaux de la constante de dérivation* k_D *optimale en fonction de* θ *et* ω. k_D *est égal à* 0 *dans le cas des perturbations de nature marche aléatoire (la droite* $\omega = \theta$*) et* ***IMA***$(1,1)$ *(la droite* $\omega = 0$*). La constante* k_D *optimale est proche de* 0 *sur l'essentiel du domaine de variation de* (θ, ω).

et λ est une constante qui détermine le poids que doit avoir $f(u_t)$.

Suite à l'identification du modèle Box-Jenkins du procédé et de la perturbation, Macgregor propose d'identifier les paramètres optimaux en se basant sur un tracé des courbes de niveaux de la variance de e_t et de $f(u_t)$, fonctions uniquement des paramètres **PID** [11]. Minh Vu a repris la même démarche [Vu (1992)], sauf qu'il s'est appliqué à exprimer la fonction de coût (Equation 3.58) en fonction de l'ensemble des paramètres caractérisant un système bouclé d'ordre quelconque [12]. Les gains optimaux sont alors obtenus par une simple minimisation codée dans le language C. Enfin Tsung [Tsung et Shi (1999)] a opté pour l'erreur quadratique moyenne de l'écart e_t comme fonction de coût, dont il a donné l'expression en fonction des paramètres stochastiques et des différents gains d'un **PID**. Il s'est limité en revanche au cas où la perturbation serait un processus $\mathbf{ARMA}(1,1)$ et le procédé à gain statique unitaire.

3.3 La commande MMSE (Minimum Mean Square Error)

Pour un exposé théorique complet sur cette commande, ponctué d'exemples, j'invite le lecteur à consulter l'ouvrage [Castillo (2002)]. Ici, rappelons le, il s'agit d'appliquer des concepts de régulation génériques au procédé à gain statique modélisé précédemment (figure3.2). Tout d'abord, rappelons l'équation de Box et Jenkins du système.

$$Y_k = b\,u_k + \frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})}\varepsilon_k \tag{3.59}$$

Y étant la déviation de la sortie, nous serons tentés de déduire

$$b\,u_k = -\frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})}\varepsilon_k \tag{3.60}$$

Cette loi de commande est celle d'un correcteur idéal. Elle est en effet impossible à implémenter car, ainsi formulée, pour calculer la commande u_k, nous aurions besoin de connaître la réalisation future de variable aléatoire ε_k. Cet écueil est encore plus vrai dans le cas où il y aurait un retard (d'ordre non nul) entre l'application de la commande et la sortie [Castillo (2002)]. Plusieurs $(\varepsilon_j, j \geq k)$ futures seraient alors à connaître.

A partir de l'équation 3.59, nous déduisons :

$$\varepsilon_k = \frac{\mathbf{D}(q^{-1})}{\mathbf{C}(q^{-1})}(Y_k - b\,u_k) \tag{3.61}$$

Nous introduisons, par ailleurs, le polynôme **G** tel que :

$$Y_k = bu_k + \frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})}\varepsilon_k = bu_k + \varepsilon_k + \frac{\mathbf{G}(q^{-1})}{\mathbf{D}(q^{-1})}\varepsilon_{k-1} \tag{3.62}$$

En substituant ε_{k-1} par son expression 3.61 en fonction de Y_{k-1} et u_{k-1} et en utilisant la relation qui relie les polynômes **C**, **D** et **G**, il vient :

$$Y_k = \varepsilon_k + \frac{\mathbf{G}(q^{-1})}{\mathbf{C}(q^{-1})}Y_{k-1} + b\frac{\mathbf{D}(q^{-1})}{\mathbf{C}(q^{-1})}\,u_k \tag{3.63}$$

11. L'expression des variances est obtenue à partir de l'équation aux différences du système en boucle fermée

12. Il a traité uniquement le cas $\lambda = 0$.

Soit $E_k[]$ l'opérateur de prévision E_k au cycle k. Autrement, toute variable aléatoire indexée par $j > k$ est considérée dans le futur par rapport au cycle présent k. Effectuer une prévision de la variable Y_k à l'horizon 1 (à partir du cycle $k-1$) revient à supposer nul le terme futur ε_k.

$$E_k[Y_k] = \frac{\mathbf{G}(q^{-1})}{\mathbf{C}(q^{-1})} Y_{k-1} + b\frac{\mathbf{D}(q^{-1})}{\mathbf{C}(q^{-1})} u_k \tag{3.64}$$

Par définition, la commande **MMSE** consiste à minimiser l'erreur quadratique moyenne (MSE) de Y_k (voir paragraphe 3.1.2). Soit T la valeur cible de Y_k.

$$\begin{aligned} MSE(Y_k) &= E_k[(Y_k - T)^2] && (3.65)\\ &= E_k[(Y_k - E_k[Y_k])^2] + (T - E_k[Y_k])^2 && (3.66)\\ &= E_k[(\varepsilon_k)^2] + (T - E_k[Y_k])^2 \geq \sigma_\varepsilon^2 && (3.67) \end{aligned}$$

Y_k étant une déviation, T est égale à 0. Minimiser $MSE(Y_k)$ équivaut à choisir $E_k[Y_k] = 0$ et la loi de commande **MMSE** relative au système 3.59 est alors déduite de l'équation 3.64 par :

$$\begin{aligned} &\min\{MSE(Y_k)\} = \min\{Var(Y_k)\} = Var(\varepsilon_k)\\ &\qquad \Longleftrightarrow E_k[Y_k] = 0 \Longleftrightarrow u_k = -\frac{\mathbf{G}(q^{-1})}{b\,\mathbf{D}(q^{-1})} Y_{k-1} \end{aligned} \tag{3.68}$$

Le diagramme du système bouclé est illustrée par la figure 3.22. Au delà des équations, il est important de saisir la double tâche qu'exécute la loi de commande 3.68 à tout instant :

⋈ Effectuer en premier lieu une prévision de la déviation à l'horizon 1 comme si le procédé était en boucle ouverte [13], c'est à dire de prévoir la réalisation de la variable aléatoire $E_k[Y_{k+1}|u_{k+1} = 0] = \frac{\mathbf{G}}{\mathbf{C}} Y_k$.

⋈ Trouver une commande u_{k+1} qui annule la prévision réalisée, $E_k[Y_{k+1}|u_{k+1} = 0]$, ce qui revient à appliquer u_{k+1} tel que

$$\frac{\mathbf{G}}{\mathbf{C}} Y_k = -b\,\frac{\mathbf{D}(q^{-1})}{\mathbf{C}(q^{-1})} u_{k+1} \tag{3.69}$$

En remplaçant la commande u_k par son expression 3.68 dans la description initiale du système 3.59, nous obtenons l'équation du système bouclé : $Y_k = \varepsilon_k$. La déviation Y_k en **BF** est alors un bruit blanc [14]. L'erreur quadratique moyenne est minimale et s'écrit :

$$MMSE(Y_k) = Var(Y_k) = \sigma_\varepsilon^2 \tag{3.70}$$

Si nous prenons l'exemple de perturbations de nature $\mathbf{IMA}(1,1)$, la loi de commande **MMSE** est donnée par la relation 3.71. Le contrôleur **MMSE** n'est autre qu'un contrôleur intégral pur. Le gain intégral est $k_I = \frac{1-\theta}{b}$. Nous pouvons en déduire la commande **MMSE** dans le cas d'une perturbation de nature bruit blanc. Il suffit de remplacer θ par 1. Le procédé étant sous contrôle statistique, toute commande de

13. Dans le cas des procédés à retard h non nul de plusieurs cycles (Temps mort), la prévision est effectuée naturellement à l'horizon h

14. Dans le cas des procédés à retard h non nul de plusieurs cycles (Temps mort), la déviation Y_k est un processus MA(h), $Var(Y_k) = \sigma_\varepsilon^2 \sum_{j=0}^{h} f_j^2$

retour engendrera l'inflation de la variance. La commande **MMSE**, ayant comme objectif de minimiser la variance de la déviation Y_k, consiste simplement à appliquer une commande constante.

$$u_k = -\frac{1}{b}\frac{1-\theta}{1-q^{-1}}Y_{k-1} \tag{3.71}$$

$$= u_{k-1} - \frac{1}{b}(1-\theta)Y_{k-1} \tag{3.72}$$

3.4 Extension de la commande MMSE

La méthode **MMSE** suppose la minimisation d'un critère de la forme $E[Y_k^2]$. Une extension naturelle de ce critère est de minimiser

$$J_1 = E[Y_k^2 + \mu\,(\nabla u_k)^2] \tag{3.73}$$

où μ pourrait être interprété comme **le multiplicateur de Lagrange** relatif à une contrainte d'inégalité 3.74 et $\nabla = 1 - q^{-1}$. L'objectif est alors d'avoir un compromis entre une variance minimale de la déviation de la sortie, $Var(Y_k)$, et une variance aussi faible des ajustements, $Var(\nabla\, u_k)$. Cette extension prend tout son intérêt dans certains cas particuliers où la variance des ajustements d'une commande **MMSE** est excessive [15].

$$E[(\nabla\, u_k)^2] \leq\ c^2 \tag{3.74}$$

Après une simple manipulation de la définition du critère J_1, il a été démontré [Castillo (2002)] que minimiser J_1 revient à minimiser la fonction J_2 donnée par l'équation suivante :

$$J_2 = E[\hat{Y_k}^2 + \mu\,(\nabla\, u_k)^2] \tag{3.75}$$

où $\hat{Y}_k$ est la prévision optimale au sens MMSE de Y_k sachant $\{Y_i\}_{i<k}$ (Voir Equation 3.64). La solution à ce problème de commande a été donnée par Clarke et **G**awthrop. Nous proposons de reprendre sa démarche toujours pour le procédé particulier décrit par l'équation 3.59. Selon l'équation 3.64, la fonction de coût J_2 s'écrit :

$$J_2 = (\frac{G}{\mathbf{C}}Y_{k-1} + b\frac{\mathbf{D}}{\mathbf{C}}\,u_k)^2 + \mu\,(\nabla\, u_k)^2 \tag{3.76}$$

La loi de commande qui permet de minimiser la fonction J_2 est solution de l'équation $\frac{dJ_2}{du_k} = 0$. Par ailleurs, remarquer que seuls les premiers termes des polynômes $\mathbf{C}$ et $\mathbf{D}$, à savoir $\mathbf{C}(0)$ et $\mathbf{D}(0)$, affectent la variable u_k à l'instant k conduit à la relation suivante

$$\frac{d\,\hat{Y}_k}{d\,u_k} = b\frac{\mathbf{D}(0)}{\mathbf{C}(0)} \tag{3.77}$$

Enfin, en exprimant $\frac{dJ_2}{du_k}$ et en faisant l'hypothèse que $\frac{\mathbf{D}(0)}{\mathbf{C}(0)} = 1$, nous obtenons la loi de commande Clark-Gawthrop énoncée ci-dessous. Naturellement, pour $\mu = 0$, nous retrouvons la commande MMSE.

$$u_k = -\frac{\mathbf{G}}{b\,\mathbf{D} + (\mu/b)\,\nabla\,\mathbf{C}}Y_{k-1} \tag{3.78}$$

15. C'est le cas notamment des systèmes dits à phase non-minimale

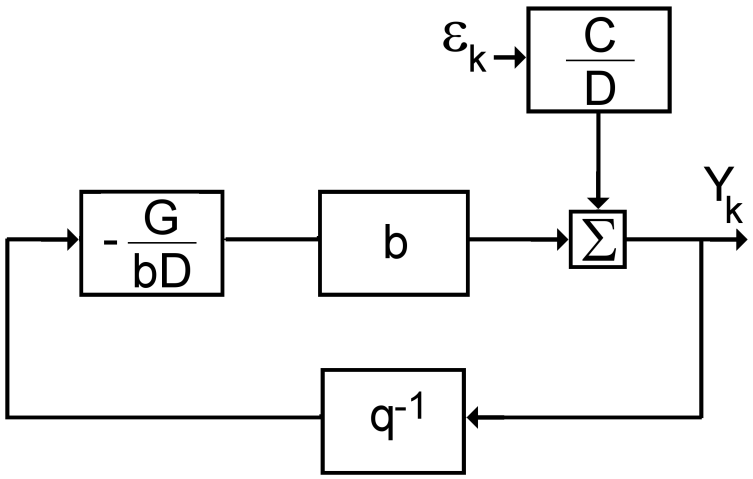

FIGURE 3.22 – *Mise en oeuvre d'une commande **MMSE**, Y_k est la déviation de la variable de sortie (consigne nulle), b est la gain du procédé, $P_k = \frac{C}{D}\varepsilon_k$ est la perturbation, $u_k = -\frac{G}{bD}Y_k$ est la loi de commande **MMSE** et q^{-1} est l'opérateur de retard.*

Prenons l'exemple d'une perturbation de nature marche aléatoire. La loi de commande Clark-Gawthrop s'écrit :

$$u_k = u_{k-1} - \frac{1}{b + \mu/b} Y_{k-1} \tag{3.79}$$

Il s'agit simplement d'une commande intégrale au gain variable en fonction de μ. En augmentant la valeur de μ, la variance des ajustements , $Var(\nabla u_k)$, est réduite fortement aux dépens d'une augmentation de la variance $Var(Y_k)$, qui demeure toutefois relativement faible (Voir Figure 3.23).

Le choix d'une valeur du facteur de pondération μ peut s'avérer difficile pour le concepteur d'une boucle de régulation. Il est souvent plus aisé de se fixer une limite supérieure c^2 aux variations des ajustements [16] $Var(\nabla u_k)$, comme le montre l'équation 3.74. Alors une manière d'obtenir un algorithme qui requiert seulement une valeur de c^2 est d'adopter au lieu d'un multiplicateur de lagrange μ constant, un μ qui s'auto-ajuste à chaque nouvelle observation selon un schéma approprié [Toivonen (1983), Castillo (2000)]. Le schéma adopté par Toivonen ainsi que DelCastillo est celui de Robbins-Monro formulé ci dessous.

$$\mu_{k+1} = \mu_k + \pi_{k+1}\, \mu_k((\nabla u_k)^2 - c^2) \tag{3.80}$$

où π_k est une suite de gains positifs. DelCastillo a rappelé les conditions nécessaires et suffisantes que doit satisfaire π_k afin de garantir la convergence de μ_k dans [Castillo (2000)].

16. Nous pouvons aussi restreindre la variance des déviations de u_k, ie : $Var(u_k)$ où u_k serait centré normalisé

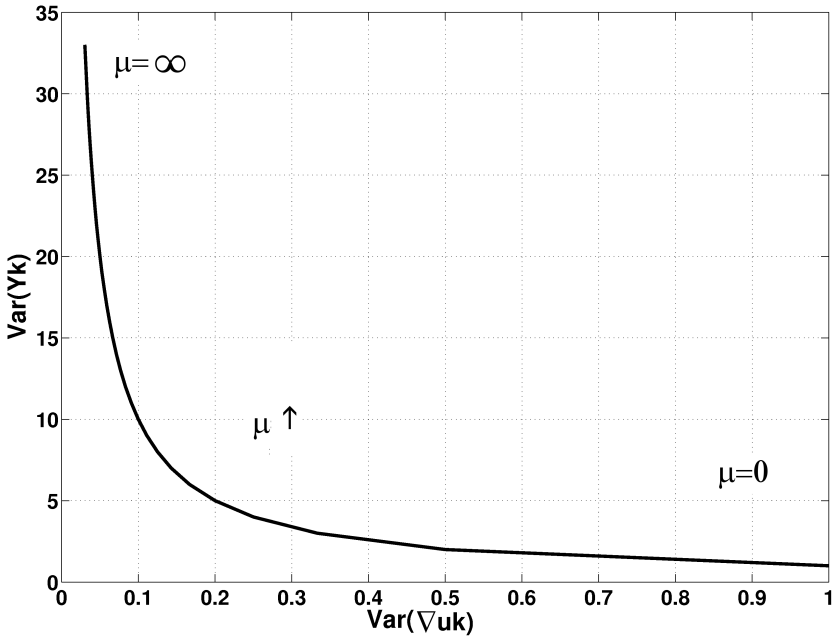

FIGURE 3.23 – *Comportement de la Commande Clark-Gawthrop dans le cas d'une marche aléatoire,* $b = 1$.

3.5 Le contrôleur Exponentially Weighted Moving Average (EWMA)

Rappelons le modèle simplifié du procédé.

$$y_k = b\, u_k + P_k$$

La loi de commande **EWMA**, est basée sur la statistique de prévision **EWMA**. Étant à l'instant k, l'idée est de prédire la perturbation P_{k+1} à l'aide de la statistique **EWMA** (voir équation 3.81) et d'ajuster la commande u_{k+1} en conséquence de manière à annuler la déviation de la sortie Y_{k+1}.

$$\hat{P}_{k+1} = \lambda P_k + (1-\lambda)\hat{P}_k \tag{3.81}$$

où λ est compris entre 0 et 1 et $\hat{P}_k$ est la prévision **EWMA** de la perturbation P_k .

Il s'en découle la loi de commande 3.82. A partir de la définition de cette commande, nous constatons d'ores et déjà que l'esprit de l'algorithme **EWMA** nous rappelle étrangement celui de la commande **MMSE**. L'unique différence est que la prévision au sein de la commande **MMSE** est une prévision optimale au sens de l'erreur quadratique moyenne, ce qui n'est pas forcément le cas pour une prévision basée sur la statistique **EWMA**. Néanmoins, il est vrai que cette statistique est optimale au sens **MSE**, lorsque la perturbation est de nature **IMA**$(1,1)$ [Box et Kramer (1992)]. Dans ce cas, il a été démontré que les deux commandes sont équivalentes si λ est pris égal à $1-\theta$.

$$u_{k+1} = \frac{-\hat{P}_{k+1}}{b} \tag{3.82}$$

Selon les équations 3.81 et 3.82,

$$\begin{aligned}\hat{P}_{k+1} &= \lambda P_k + (1-\lambda)\hat{P}_k \\ &= \lambda(y_k - bu_k) + (1-\lambda)\hat{P}_k \\ &= \lambda(y_k + \hat{P}_k) + (1-\lambda)\hat{P}_k \\ &= \lambda y_k + \hat{P}_k\end{aligned}$$

et puis

$$u_{k+1} = u_k - \frac{\lambda}{b} y_k$$

Le contrôleur **EWMA** est simplement un contrôleur intégral pure dont le gain intégral équivalent est $k_I = \frac{\lambda}{b}$. Toutes les propriétés vues précédemment sous le paragraphe 3.2.2 lui sont alors appliquées.

3.5.1 Stabilité

La commande EWMA est une commande intégrale dont le gain k_I est fonction du gain du système b. Le gain b, étant inaccessible par nature, est estimé par $\hat{b}$. Soit $\xi = \frac{b}{\hat{b}}$. Selon les résultats du paragraphe 3.2.2, la stabilité du système est assurée si l'inégalité suivante est valide .

$$0 < \lambda\xi < 2$$

Cette condition implique que b et β aient le même signe et que $\hat{b}$ soit supérieur à la moitié de λb. Ce résultat est bien connu et il a été démontré par divers façons dans la littérature [Sachs et Ingolfsson (1993), Castillo (2002), Chemali (2002)].

3.6 Le Contrôleur Adaptatif

Les méthodes de régulation classiques vues brièvement dans les paragraphes précédents, à travers la commande **PID**, **EWMA**, **MMSE**, sont conçus pour régler des systèmes linéaires et invariants dans le temps. Si nous venons à adopter cette hypothèse dans le cas du procédé lithographique, elle impliquerait deux éléments :

⋈ La sensibilité de la longueur de ligne de résine à la dose d'énergie, autrement le gain du procédé considéré comme statique, est invariante (constante).

⋈ Le procédé est sujet à une perturbation dont les paramètres stochastiques sont aussi constants.

L'hypothèse d'un gain constant est plausible. En revanche, une perturbation invariante, dans un environnement stochastique, est peu crédible. Ses caractéristiques varient en effet dans le temps et peuvent dégrader les performances du système bouclé d'une façon notable. Un exemple est celui de la commande **MMSE**. Un contrôleur **MMSE**, conçu pour rejeter des perturbations stationnaires de nature **ARMA**(1,1), échoue face à une perturbation non stationnaire **ARIMA**(1,1,1) [Tsung et al. (1998)]. Le système est alors instable et la variance de la variable de sortie est théoriquement infinie. Dans ces situations où les paramètres dynamiques et stochastiques de l'ensemble du système sont variables dans le temps, ou simplement inconnus, un remède peut être d'utiliser le contrôle adaptatif.

Par définition, un régulateur adaptatif est un régulateur muni de coefficients ajustables dont l'ajustement permet de maintenir un certain niveau de performances. La

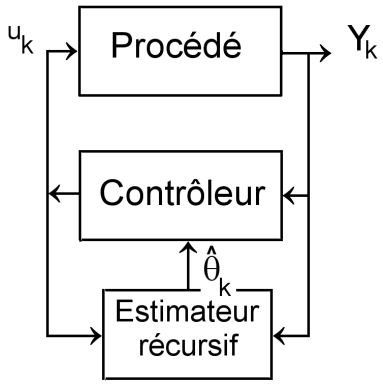

FIGURE 3.24 – *Vue Schématique d'une commande auto-ajustable.*

commande adaptative peut s'appliquer en boucle ouverte, ou pré-programmée. Une application de ce type de commande pourrait être envisagée dans le cas de la boucle de compensation (FF) entre la gravure et l'implantation ionique des poches (voir chapitre suivant). Il a été en effet question d'ajuster le centrage du modèle (il s'agit d'une constante définie pour chaque produit) en se basant sur le biais iso-dense de la grille (poly-Si). Par la suite, nous nous intéresserons à la commande adaptative en boucle fermée, et plus particulièrement à la commande auto-ajustable ou *self-tuning* (**ST**).

L'idée au centre de la commande auto-ajustable est de séparer l'estimation des paramètres, réalisée souvent via un estimateur récursif, de la loi de commande. La figure 3.24 montre la structure générale d'un contrôleur ST. Nous y voyons un bloc d'identification ainsi qu'un contrôleur aux coefficients ajustables. L'algorithme d'estimation fournit des valeurs estimées en ligne des paramètres, qui sont utilisées par la loi de commande comme s'ils étaient les vrais paramètres. Les automaticiens nomment cette démarche ***principe de l'équivalence certaine*** [Castillo (2002)].

3.6.1 Contrôleurs Auto-Ajustables Directs et Indirects

La technique indirecte de la commande ST fait usage d'un bloc d'identification pour estimer les paramètres caractéristiques du procédé et de la perturbation. Les valeurs estimées de ces paramètres serviront par la suite au contrôleur pour calculer la commande. Une seconde technique de contrôle adaptatif est la technique directe ou implicite où les paramètres de la loi de commande seront estimés directement par le bloc d'identification. Par souci de clarté, voici deux exemples qui illustrent la différence entre le schéma direct et indirect [Castillo (2002)].

Nous considérons un système de premier ordre avec une perturbation de nature **ARMA**(1,1) caractérisée par (θ, ω).

$$y_k = \frac{g\, q^{-1}}{1 - \phi\, q^{-1}}\, u_k + \frac{1 - \theta\, q^{-1}}{1 - \phi\, q^{-1}}\, \varepsilon_k \tag{3.83}$$

Si nous avons à réaliser une commande **MMSE**, elle serait formulée ainsi :

$$u_k = -\frac{\phi - \theta}{g}\, y_k \tag{3.84}$$

Le contrôleur auto-ajustable indirect correspondant est donné par l'expression 3.85. A chaque nouvelle observation de la variable de sortie, l'algorithme d'identification fournira à la loi de commande **MMSE** une estimation du vecteur $\begin{bmatrix}\hat{\phi} & \hat{\theta} & \hat{g}\end{bmatrix}'$.

$$u_k = -\frac{\hat{\phi} - \hat{\theta}}{\hat{g}}\, y_k \tag{3.85}$$

Pour le contrôleur auto-ajustable direct, la commande **MMSE** est réduite à la loi 3.86. Un seul paramètre L est alors à estimer dans le cas d'une configuration directe du contrôleur **ST**.

$$u_k = -\frac{\phi - \theta}{g}\, y_k = -L\, y_k \tag{3.86}$$

3.6.2 La Commande MMSE Auto-Ajustable

Le point de départ pour concevoir cette commande est l'équation 3.63, retranscrite ci-dessous.

$$\mathbf{C}\,(Y_{k+1} - \varepsilon_{k+1}) = \mathbf{G}\,Y_k + b\,\mathbf{D}\,u_{k+1} \tag{3.87}$$

où

$$G(q^{-1}) = g_0 + g_1 q^{-1} + \dots\, g_m q^{-m}$$
$$D(q^{-1}) = d_0 + d_1 q^{-1} + \dots\, d_l q^{-l}$$

et m et l sont respectivement les degrés des polynômes **G** et **D**. Cette équation définit un modèle du système au même titre que 3.59. L'intérêt est que ce modèle est paramétré directement par les polynômes G et D de la commande optimale **MMSE** (équation 3.68). Tout d'abord, nous introduisons l'hypothèse simplificatrice.

$$\mathbf{C}(q^{-1}) = 1 \tag{3.88}$$

Une commande MMSE auto-ajustable consisterait alors à implémanter un estimateur récursif en ligne, associé à la loi de commande

$$u_{k+1} = -\frac{\hat{\mathbf{G}}}{\hat{b}\,\hat{\mathbf{D}}} y_k \tag{3.89}$$

où $\hat{G}$ et $\hat{D}$ sont les estimées en ligne des polynômes G et D. L'estimation en ligne est effectuée selon l'équation 3.90. Comme la variable aléatoire ε_{k+1} est non corrélée au vecteur des régrésseurs $\{Y_{k-i}\}_{i\geq 0}$ et $\{u_{k+1-i}\}_{i\geq 0}$, un estimateur récursif de base tel les moindres carrés récursifs serait à même de fournir des estimés sans biais des coefficients $\{g_i\}_{i\geq 0}$ et $\{d_i\}_{i\geq 0}$ [Hunt (1986), Shah et Cluett (1991)]. Au bout de quelques itérations, l'algorithme convergera vers une commande à variance minimale.

$$Y_{k+1} = T_{k+1}\Theta_{k+1} + \varepsilon_{k+1} \tag{3.90}$$

où

$$\Theta_{k+1} = \begin{bmatrix} g_0, & g_1, & \cdots & g_m; & bd_0, & bd_1, & \cdots & bd_l \end{bmatrix}'$$
$$T_{k+1} = \begin{bmatrix} Y_k, & Y_{k-1}, & \cdots & Y_{k-m}; & u_{k+1}, & u_k, & \cdots & u_{k+1-l} \end{bmatrix}$$

Aussi surprenant que celà puisse paraître, cet algorithme convergera aussi vers une commande optimale (**MMSE**), même si l'hypothèse 3.88 n'est pas vérifiée

[Åström et Wittenmark (1973),Clarke et Gawthrop (1975)]. Un seul préalable à cette propriété fondamentale est la convergence de l'algorithme. Cette propriété est inattendue. La surprise tient à deux choses : d'une part, l'équation n'est plus fonction des seuls paramètres de la commande optimale et d'autre part, l'estimateur des moindres carrés récursifs est utilisé en dehors de ses conditions d'application ($C \neq 1$). Une explication intuitive est donnée dans [Harris et al. (1980)]

Reprenons l'exemple d'un procédé sujet à une perturbation **IMA**(1,1). Après l'obtention des expressions des polynômes **G** et **D**, l'équation 3.87 pourrait s'écrire de la façon suivante

$$(1 - \theta\, q^{-1})\,(Y_{k+1} - \varepsilon_{k+1}) = (1 - \theta)\,Y_k + b\,(1 - q^{-1})\,u_{k+1} \tag{3.91}$$

Selon les travaux d'Astrom et Wittenmark [Åström et Wittenmark (1973)], nous pouvons toujours substituer le polynôme $\mathbf{C} = 1 - \theta\, q^{-1}$ par 1. Ceci n'aurait aucun effet sut les propriétés asymptotiques de l'algorithme. Soit $\Theta' = \begin{pmatrix} 1-\theta, & b \end{pmatrix}$ et $z'_k = \begin{pmatrix} y_{k-1}, & u_k - u_{k-1} \end{pmatrix}$. L'estimation en ligne sera basée sur le modèle suivant :

$$Y_k - \varepsilon_k = \Theta' z_k \tag{3.92}$$

Le vecteur $\hat{\Theta}_k$ sera utilisé à chaque cycle pour appliquer une commande dont l'expression est la suivante :

$$u_k = u_{k-1} - \frac{1}{\hat{b}}(1 - \hat{\theta}) y_{k-1} \tag{3.93}$$

3.7 Etat de l'art : Régulation du procédé lithographique

Nous allons nous intéresser aux travaux de recherche en régulation appliqués à la lithographie et en particulier au contrôle de la variabilité lot à lot de la dimension critique de la résine L_{resist}. Hormis quelques rares travaux, l'algorithme de base mis en application a toujours été l'algorithme **EWMA** [Sachs et Ingolfsson (1993)], un simple intégrateur comme nous l'avons mentionné dans les paragraphes précédents. La variable d'entrée manipulée dans la majorité des cas industriels est la dose d'énergie du laser. Elle est en effet simple à ajuster d'un point de vue équipement et a en outre un effet linéaire sur la dimension critique de la résine. Afin de réduire davantage la variabilité du L_{resist}, d'autres variables ont été envisagées, notamment le focus, la température du recuit après insolation (Post Exposure Bake **PEB**), la vitesse de la centrifugeuse (dépôt de la résine), etc. Ceci n'est pas sans incidence bien entendu sur la complexité du modèle.

Geary et al ont implémenté un couple de structures dans les lignes de découpe qui présente une particularité très intéressante : les deux structures sont réalisées de sorte que les swings curves correspondantes soient en opposition de phase [Geary et Barry (2003)]. Le but est de séparer l'effet d'une éventuelle variation de l'épaisseur de résine de celui de toute autre source de variation (**PEB**, épaisseur de la couche sous-jacente, etc). La moyenne des L_{resist} de ce couple de structures serait en effet insensible aux variations de l'épaisseur de la résine, mais demeure linéairement dépendante de la dose d'énergie (Voir figure 3.25). Au contraire, la différence des L_{resist} dépend très peu de la variation de la dose mais est très sensible à la variation de l'épaisseur de la résine. Cette démarche prend tout son intérêt dans le cas où la mesure de

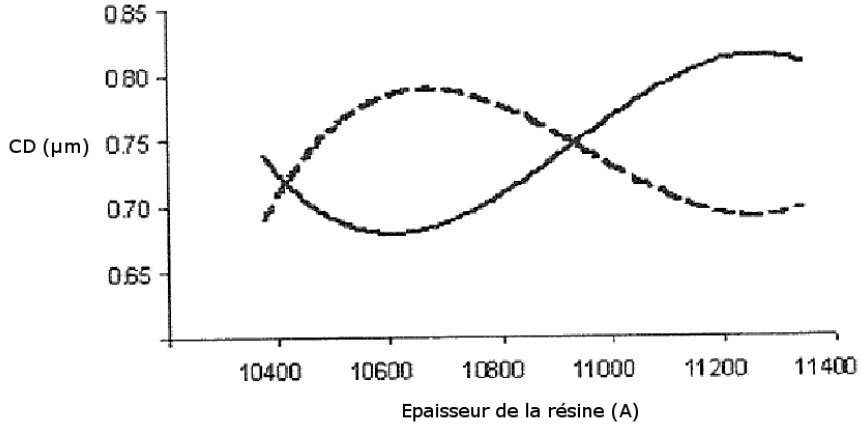

FIGURE 3.25 – *Les structures sont réalisées de telle façon que leurs 'swings curves' soient en opposition de phase.*

l'épaisseur de la résine serait impossible dans un contexte industriel [17]. Cet argument est à relativiser avec la venue de la scattérométrie, une métrologie capable de fournir les épaisseurs des différentes couches sous-jacentes ainsi que la dimension critique courant la même mesure.

Emir Gurer *et al.* ont aussi cherché à réduire la variabilité de l'épaisseur de résine [Gurer et al. (1998)]. Ils ont fait référence au rôle déterminant de l'humidité relative et la pression barométrique dans l'évaporation et la convection, phénomènes à l'origine de la formation de la couche de résine par centrifugation, avant d'introduire une boucle anticipative (FeedForward), qui ajuste la vitesse de la tournette en fonction de ces deux paramètres prélevés au préalable au sein de l'enceinte de la tournette [18].

Un travail remarquable est celui de Palmer et al [Palmer et al. (1996)], qui ont conçu un contrôleur qui ajuste à la fois la vitesse de la tournette et la température du recuit de la résine [19]. L'objectif était de réduire la variabilité de l'épaisseur de la résine et aussi celle de la concentration en composants photo-actifs PACs. Il s'agit d'un travail de référence dans la mesure où c'est la première fois que l'on utilise le filtre de Kalman pour une application en semi-conducteur. Son rôle était d'estimer à chaque nouvelle observation les coefficients du modèle linéaire qui relie les variables d'entrée (la température et la vitesse de la tournette) aux variables de sortie (l'épaisseur de la résine et la concentration en PAC).

John Stubber et al [Stuber et al. (2000)] ont souligné l'effet négatif de toute correction d'une déviation du focus par une compensation en dose d'énergie, notamment sur le profil des lignes de résine, et aussi le biais iso-dense. L'idée est alors de réaliser un régulateur qui ajuste à la fois le focus et la dose sous réserve de trouver une variable de sortie sensible au focus, et le rend donc observable. Ils ont opté pour le rapport entre le **B**=L_{resist}[Bottom], mesuré au pied de la ligne, et le **T**=L_{resist}[Top], mesuré au sommet de la ligne. Ce rapport est bien entendu équivalent à une mesure

17. Bien entendu, il faudrait commencer par démontrer la relative importance de la variabilité de l'épaisseur de la résine

18. Ils ont mis en place des capteurs pour la mesure de l'humidité et de la pression

19. Il s'agit d'une ancienne technologie où BARC et résine à amplification chimique n'existaient toujours pas

du profil, certainement inexistante dans les usines de fabrication à l'époque de ces travaux. Le modèle $2 * 2$ est formulé ci dessous. Le degré d'interaction entre les deux boucles est mesuré par le coéfficient $\Omega = \frac{1}{1-\frac{K_{12}K_{21}}{K_{11}K_{22}}}$. Un Ω proche de l'unité serait équivalent à une faible interaction entre la boucle $L_{resist} \circlearrowleft Dose$ et $\mathbf{T/B} \circlearrowleft Focus$.

$$\begin{bmatrix} L_{resist} \\ \mathbf{T/B} \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} Dose \\ Focus \end{bmatrix} \tag{3.94}$$

D'une façon totalement différente, Bao a proposé de générer une bibliothèque de profils à partir de simulations physiques [20]et cela pour différentes combinaisons de dose et de focus [Bao (2003)]. Le profil mesuré par scattérométrie, suite au processing de la plaque [21], sera alors comparé à ceux de la bibliothèque, bâtie lors d'une phase préparatoire. La déviation en dose et en focus par rapport aux valeurs appliquées (Recette) sont ainsi extraites et puis simplement soustraites pour le cycle suivant. En ce point, cette façon de réguler le procédé lithographique rappelle plusieurs travaux récents où l'on a cherché à extraire la dose et le focus effectifs, souvent via la mesure de structures spécifiques dans les lignes de découpe. Par effectif, nous entendons des valeurs du focus et de la dose équivalentes, résultats des déviations des autres paramètres (Température **PEB**, épaisseur de la résine, etc).

Dans la continuité des recherches menées par Bao, El Chemali a mis en relief un important avantage de manipuler le focus, celui de réduire la variabilité du biais iso-dense [Chemali et al. (2004)]. Dans son manuscrit de thèse [Chemali (2002)], il a réalisé une étude comparative entre 3 différentes configurations de régulation : chacune vise à minimiser les variations de certaines variables bien définies. La première manipule la dose et le focus afin d'ajuster le couple L_{resist}[i]/SWA des lignes isolées. Nous rappelons que SWA (Side Wall Angle) correspond au profil de la ligne de résine. De même, La seconde vise à réduire les variations de la dimension critique des lignes denses et des lignes isolées L_{resist}[d]/L_{resist}[i]. La dernière est conçue de façon à trouver le meilleur compromis pour le triplet L_{resist}[d]/L_{resist}[i]/SWA. En guise de conclusion de cette analyse comparative, El Chemali indique que les performances des deux dernières configurations (L_{resist}[d]/L_{resist}[i] et L_{resist}[d]/L_{resist}[i]/SWA)sont presque identiques. Elles sont en outre meilleures que le contrôleur L_{resist}[i]/SWA [22] dans la mesure où elle réduisent davantage les perturbations qui affectent le L_{resist}[i], le L_{resist}[d] et le SWA. Il est important de noter que ces résultats, ceux de Bao inclus, sont tirés de simulations effectuées à l'aide de PROLITH et n'ont pas été déployés dans un contexte industriel.

Les procédés lithographiques les plus avancés utilisent une résine à amplification chimique, qui rend le dessin des motifs relativement plus sensible que les résines novolaques au recuit **PEB**. Ce constat a conduit à la conception de plusieurs boucles de régulation qui agissent d'une façon ou d'une autre sur ce recuit. Musacchio a notamment réalisé un contrôleur qui manipule la dose d'énergie ainsi que la durée du recuit **PEB** [Musacchio (1998)]. L'objectif est de réduire la variabilité plaque à plaque de L_{resist}. Pour cela, il s'est appuyé sur les travaux de Jakatdar [Jakatdar et al. (1998)] qui montrent que la perte d'épaisseur de résine (Thickness loss) produite suite au recuit PEB est fortement corrélée au phénomène de déprotection de la

20. PROLITH, un simulateur célébre réalisé par Chris Mark

21. Bao a dévéloppé ce régulateur pour compenser les perturbations de plaque à plaque

22. Il a aussi montré qu'elles présentent une meilleure performance par rapport au simple contrôleur qui manipule seulement la dose d'énergie

matrice polymère. Une mesure de ce phénomène est très utile car elle contient des informations sur les effets cumulés du **PEB** et de l'insolation.

Sturtevant et al ont proposé une autre façon de monitorer en temps réel le phénomène de déprotection [Sturtevant et al. (1993)]. La technique est basée sur le potentiel diffractant de l'image latente qui se forme lors du recuit **PEB** [23]. Un capteur CCD placé au dessus de la plaque permet de collecter les premiers ordres de diffraction et d'observer les variations de l'intensité du signal diffracté en cours du recuit. A partir de là, Sturtevant et al ont défini un paramètre appelé T_c qui serait linéairement corrélé à la mesure SEM de la dimension critique (après développement). Ils proposent d'ajuster la dose d'énergie ou la température **PEB** de plaque à plaque afin de réduire la variabilité du paramètre T_c, calibré préalablement avec la mesure SEM.

3.8 Conclusion

Dans la première partie de ce chapitre, nous avons modélisé les procédés de fabrication en semi-conducteur, notamment la lithographie, dans un format bien spécifique. Il s'agit d'un modèle additif, inspiré de la littérature, somme d'une perturbation de type **ARIMA**(1,1,1) et d'un transfert entrée-sortie à gain statique. Nous avons revu, par la suite, un échantillon des méthodes de régulation utilisables potentiellement pour asservir les processus industriels sujets à des perturbations stochastiques.

La commande **PID** figure en tête de liste. Nous avons vu que cette commande est souvent réduite à une action combinée de la composante proportionnelle et intégrale. En effet, l'action dérivée est délaissée car elle dégrade le temps de réponse et n'a aucun effet sur le régime permanent. La composante intégrale est capable, au contraire, d'annuler les erreurs statiques produites suite à l'application des perturbations, dits *shifts* (ou échelon). Selon le choix du gain correspondant, le temps de réponse pourrait être aussi minimisé. Nous avons constaté que la commande **EWMA**, très bien implémentée dans l'industrie semi-conducteur, n'est qu'un contrôleur intégral pur.

La composante proportionnelle, associée à une action intégrale, est bénéfique dans la mesure où elle permet de garantir une commande optimale **MMSE** sur une plus grande partie du domaine de variation des paramètres stochastiques des perturbations que l'action intégrale seule, et ceci notamment lorsque la perturbation est assimilée à une rampe déterministe. L'action **P** a un impact négatif néanmoins sur le temps de réponse. A travers la dernière partie de la section qui traite de la commande **PID**, nous avons rappelé quelques démarches qui ont été adoptées pour choisir concrètement les différents gains.

Nous nous sommes aussi intéressés à la commande **MMSE**, appelée aussi commande à variance minimale (**MV**). Le développement mathématique du contrôleur a été réalisé dans le cas particulier des procédés stochastiques à gain statique. L'exemple d'une perturbation **IMA**$(1,1)$ montre que la commande **MMSE** peut être confondue avec une commande **PID**. Les gains sont dans ce cas fonctions des

23. Lors du recuit **PEB**, l'indice de réfraction des zones exposées change. Il en faut pas plus pour diffracter la lumière

paramètres stochastiques de la perturbation. Une extension de la commande **MV** est celle qui intègre une minimisation double, celle de la variance de la variable de sortie et aussi celle des ajustements. La commande Clarke-Gawthrop prend tout son intérêt dans le cas des systèmes à phase non minimale, où les ajustements pourraient être excessives. Ceci n'est pas le cas des procédés à gain statique.

Le contrôle adaptatif a été aussi introduit d'une façon succincte. Un contrôleur adaptatif, dans se version directe ou indirecte, est muni de coefficients qui sont ajustés à chaque nouvelle observation. Le but est alors de toujours avoir une commande optimale, en dépit de la variation éventuelle des caractéristiques de la perturbation. Bien que les perturbations dans l'industrie semi-conducteur sont variantes dans le temps, il n'est pas envisageable de mettre en production un contrôleur auto-ajustable. Les risques d'instabilité, d'inflation de la variance suite à des estimations en ligne biaisées (en phase transitoire ou en régime permanent) sont trop importants.

Après cet état des lieux des régulations potentiellement utilisables pour le procédé lithographique, nous ne pourrions émettre de recommendations, en absence d'une identification expérimentale (campagnes d'essais). Néanmoins, si nous nous appuyons sur les travaux de Macgregor [MacGregor et al. (1984)], où il établit une dualité entre les perturbations déterministes les plus fréquentes (rampe et échelon) et les processus **ARIMA**, une commande **PI** pourrait donner des bonnes performances.

Ce chapitre n'est qu'initiation et ouverture à une étude plus rigoureuse d'un point de vue automatique. Notre objectif étant de montrer l'interet de l'approche automaticienne dans la conception des régulations *run-to-run*, nous avons retenu plusieurs hypothèses simplificatrices, et notamment celle d'un retard unitaire dans la chaine de retour. Comme expliqué précédemment, ce retard est très souvent supérieur à une unité de temps, et il est par ailleurs variable. Nous avons aussi supposé d'une façon implicite que l'unité de temps est constante, ce qui est contraire à la réalité en production semiconducteur. Une unité de temps pourrait aussi bien être de l'ordre de quelques minutes que de quelques jours. Le recours à la théorie des sytèmes à retard et aux systèmes échantillonnés à pas variable permettrait d'aller plus loin dans la compréhension des régulations *run-to-run* dans l'industrie microélectronique.

Régulation Gravure-Implantation des Poches

4

Sommaire

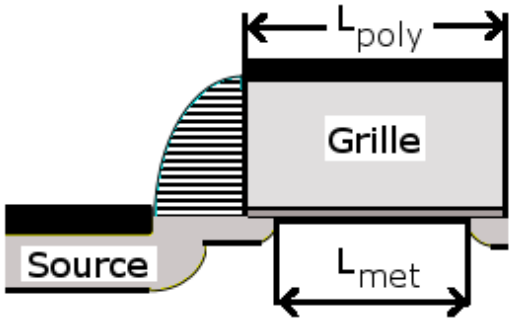

Figure 4.1 – *Illustration de la différence entre la longueur métallurgique L_{met} et la longueur de la grille en polysilicium L_{poly}*

Le chapitre 2 révèle une forte variabilité lot à lot de la longueur de grille gravée L_{poly}. La longueur effective du canal L_{eff} est définie comme celle qui conditionne le comportement du transistor. Elle est plus proche de la longueur métallurgique [1] (voir figure 4.1) et intervient dans les relations qui servent à décrire le comportement du transistor, les principaux sont la tension de seuil V_{th}, le courant de saturation I_{on}, et le courant de fuite I_{off}. Les variations de L_{poly} affectent naturellement L_{eff}, et entraînent par conséquent une fluctuation des caractéristiques électriques du transistor.

1. appelée aussi longueur de jonction à jonction

Dans ce chapitre, nous proposons d'étudier une régulation de compensation (*Feed-forward Controler*) entre la gravure du polysilicum et l'implantation des poches. Afin d'aligner la longueur L_{eff} de l'ensemble des lots et de compenser la variabilité de L_{poly}, le contrôleur ajuste la dose d'implantation des poches de chaque lot. En pratique, les lots qui ont une longueur de grille sur-dimensionnée auront une dose plus faible. Les lots qui ont une longueur de grille sous-dimensionnée auront en revanche une dose plus forte.

4.1 Principe physique de fonctionnement

Avant d'étudier les effets associés au fonctionnement du contrôleur, il est nécessaire de faire quelques rappels sur le fonctionnement du transistor MOS (*Metal-Oxyde-Semiconductor*) à effet de champ.

4.1.1 Principe d'un transistor MOS

La figure 4.2 montre le schéma d'un transistor à effet de champ (MOSFET) à canal N. Le principe de base du transistor MOS consiste à moduler la conductivité d'une couche de semi-conducteur (silicium) par le biais d'un champ électrique qui lui est perpendiculaire. Ce champ est appliqué par la grille en poly-silicium, à travers une couche diélectrique en oxyde de silicium, et permet de créer ce que nous appellons un canal de conduction entre la source et le drain. Le canal est constitué de charges électriques mobiles à l'interface Diélectrique-Semiconducteur. Le transistor (MOSFET), dans une représentation schématique, peut être considéré comme une résistance modulable, entre la source et le drain, commandée en tension par la grille. Pour une tension de drain donnée V_D, le courant de drain I_D n'augmente de façon notable que lorsque la tension de grille V_G est supérieure à une tension de seuil V_{th}. A partir de là, deux régimes sont définis : un régime à faible inversion lorsque $V_G < V_{th}$ et un régime à forte inversion où $V_G \geq V_{th}$ (voir figure 4.3).

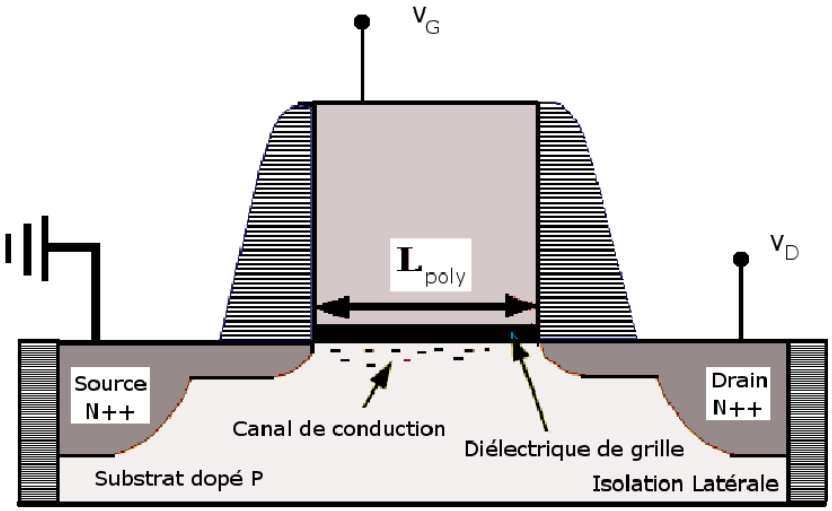

Figure 4.2 – *Schéma simplifié d'un transistor NMOS avec les électrodes source, drain et grille. La source est mise à la masse*

Pour les applications pratiques, il est souhaitable que I_D soit parfaitement nul lorsque le transistor est bloqué ($V_G = 0$ et $V_D \neq 0$), fonctionnant alors en régime de faible inversion, notamment pour réduire la puissance consommée par le dispositif. En réalité, ce n'est pas le cas, et il reste un courant de fuite résiduel que l'on note I_{off}. En régime de forte inversion ($V_G \geq V_{th}$), le transistor est dans l'état ouvert. Nous y distinguons deux régimes. Le premier correspond au régime linéaire où V_D est suffisamment faible, inférieure à $(V_G - V_{th})$. Dans ce cas, le transistor se comporte comme une résistance variable dont la valeur est contrôlée par la tension V_G. A

l'opposé, lorsque $V_D \geq (V_G - V_{th})$, le transistor est dit en régime de saturation. La résistance du canal est faible, varie peu avec la tension V_D et le courant de drain, dit courant de saturation I_{on}, est constant (voir équation 4.1).

$$I_{on} = \frac{1}{2}\mu C_{ox}\frac{W}{L_{eff}}(V_G - V_{th})^2, \tag{4.1}$$

Avec μ : la mobilité des charges électriques,
C_{ox} : la capacité de la couche diélectrique par unité de surface,
W : la largeur du transistor (zone active),
L_{eff} : la longueur effective de la grille.

A partir de cette relation simple, et sachant qu'en forte inversion, l'idéal est d'avoir un courant de drain I_{on} le plus élevé possible, notamment afin d'augmenter la vitesse de fonctionnement du circuit, nous voyons qu'il faut chercher à augmenter la capacité de l'oxyde et diminuer la longueur du canal, d'où la course à la miniaturisation dans laquelle l'industrie du semi-conducteur s'est engagée depuis les années 1970 [Schaller (1997)].

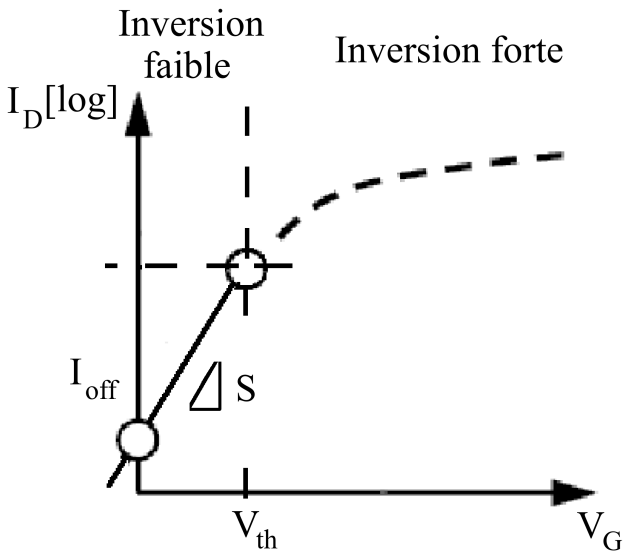

FIGURE 4.3 – *Caractéristique I_D=f(V_G) en échelle semi-logarithmique*

4.1.2 Les effets de la miniaturisation

Nous considérons le cas d'une structure MOS à canal N. Lorsque la tension V_G est suffisament grande, mais inférieure à V_{th}, le canal est dans un état de déplétion. Le champ électrique appliqué par la grille repousse les charges majoritaires (les trous) loin de l'interface entre la couche diélectrique et le substrat. Dans le cas d'un transistor long ($L_{eff} > 1\mu m$), la charge de déplétion dans le canal est majoritairement contrôlée par la grille. La jonction drain-substrat, zone, elle aussi, dépeuplée en porteurs majoritaires (zone de charge d'espace), s'étend en effet sur une partie du canal dont la longueur ΔL_{eff} demeure négligeable par rapport à celle du canal L_{eff}. Lorsque la dimension de la grille L_{poly} diminue, l'influence de la zone de charge d'espace (ZCE) induite par la jonction drain-substrat se trouve accrue. ΔL_{eff} est

alors comparable à L_{eff} et par conséquent la tension de seuil V_{th} chute. C'est l'effet géométrique du canal court (**SCE**, ***S**hort **C**hannel **E**ffect*).

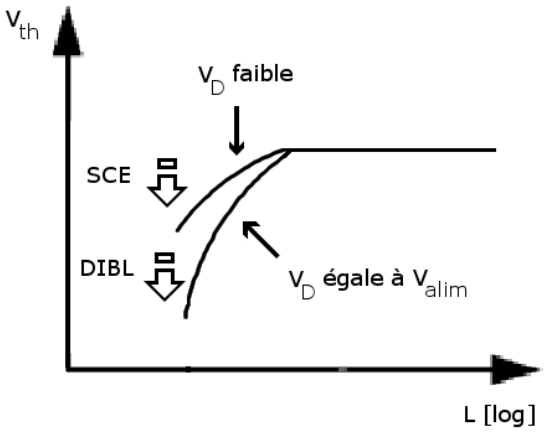

FIGURE 4.4 – *Schéma de l'évolution de V_{th} avec L_{poly}, en fonction de la tension appliquée sur le drain V_D.*

Par ailleurs, toute augmentation de la tension appliquée au drain V_D se répartit exclusivement sur la ZCE, entraînant l'extension de la charge de déplétion dans le canal. Ce phénomène contribue aussi à faire chuter la tension de seuil : c'est l'effet électrique DIBL (***D**rain **I**nduced **B**arrier **L**owering*). Comme le montre la figure 4.4, l'effet principal lié à l'augmentation relative des profondeurs de déplétion de la ZCE est l'abaissement de la tension de seuil avec la diminution de L_{poly} et sa dépendance au potentiel du drain V_D. Les effets canaux courts constituent une limitation importante pour la miniaturisation des technologies, car ils engendrent, via une réduction de la tension de seuil V_{th}, une augmentation incontrôlée du courant de fuite I_{off} (voir équation 4.2).

$$I_{off} \propto \exp(-ln(10)\frac{V_{th}}{S}), \tag{4.2}$$

où S est la pente sous le seuil de la caractéristique $I_D = f(V_G)$, illustrée en figure 4.3.

4.1.3 Implantation des poches [Gwoziecki (1999)]

La zone de charge d'espace (ZCE) des jonctions s'étend principalement dans le substrat à cause de son faible dopage en comparaison avec la source et le drain. Pour réduire la profondeur de la zone déplétée et compenser ainsi la chute de V_{th} due aux effets du canal court, une étape supplémentaire d'implantation a été intégrée aux différentes étapes de fabrication d'une technologie CMOS (voir figure 4.5). Grâce à cette étape d'implantation dite des poches, tiltée et auto-alignée avec la grille, le dopage du canal est augmenté localement autour des extensions LDD (***L**ow **D**oped **D**rain*) (voir figure 4.6). L'influence des poches sur le contrôle des effets de canal court est nettement mis en évidence sur la figure 4.7 : la décroissance de la tension de seuil avec la longueur de grille ou *roll-off* est réduite de manière significative. Les transistors longs ($> 1\mu m$) ne sont pas impactés.

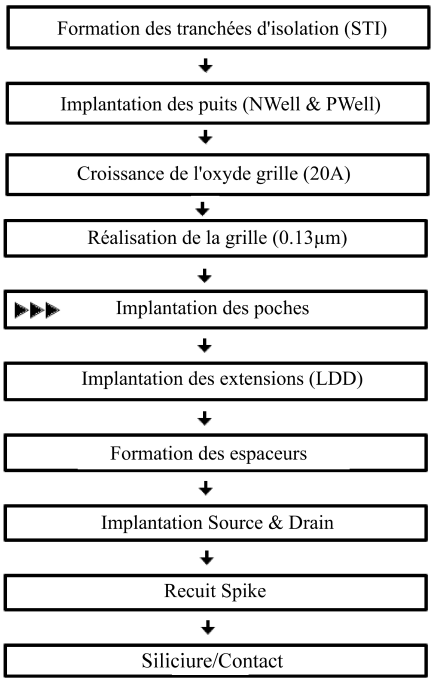

FIGURE 4.5 – *Les étapes clés du procédé de fabrication d'une technologie CMOS* 0.13*µm. Elles intégrent une étape d'implantation des poches, réalisée juste après la définition de la grille en poly-Si.*

Dans le cadre de ce travail, la dose d'implantation des poches a été choisie comme levier pour compenser les déviations de la dimension du poly-silicium. Cependant, nous devons signaler que d'autres leviers [Conchieri et al.], tel le tilt ou l'angle d'implantation des poches, semblent satisfaire le cahier des charges du contrôleur, mais elles n'ont pas été explorées. Bien-entendu, notre objectif n'était pas de réaliser un état des lieux complet des paramètres qui pourraient être envisagés, mais d'en choisir un, qui soit corrélé à la longueur effective des transistors courts, et dont l'ajustement n'affecte pas les autres dispositifs (transitors longs, ..). Il est aussi essentiel que ce paramètre soit facilement maniable. A titre d'exemple, il nous a été communiqué, au démarrage de nos travaux, de façon explicite de ne pas envisager l'énergie de l'implantation des poches comme variable d'ajustement. Toute manipulation de l'énergie d'implantation nominale, celle codée dans la recette du processus, est en effet impossible, contrairement à la dose. La dose d'implantation des poches satisfait ces spécifications. Il faut avouer aussi que notre décision de se focaliser exclusivement sur ce paramètre est due en partie à l'utilisation certaine de cette boucle par certains compétiteurs, même si nous n'en trouvons pas des publications.

4.2 Etat de l'art

Steven Merrill Ruegsegger *et al.* font partie des pionniers qui se sont intéressés à la régulation de compensation [Ruegsegger et al. (1998), Ruegsegger et al. (1999)]. Dans son manuscrit de thèse [Ruegsegger (1998)], Ruegsegger identifie deux éléments majeurs qui pourraient nuire à la bonne performance du contrôleur :

⋈ Tout d'abord, le bruit de la mesure de la variable d'entrée. Un contrôleur qui s'appuie sur une mesure de métrologie trop bruitée, au delà d'un certain seuil, pourrait en effet amplifier la variance de la variable de sortie. Le risque est de sur-ajuster.

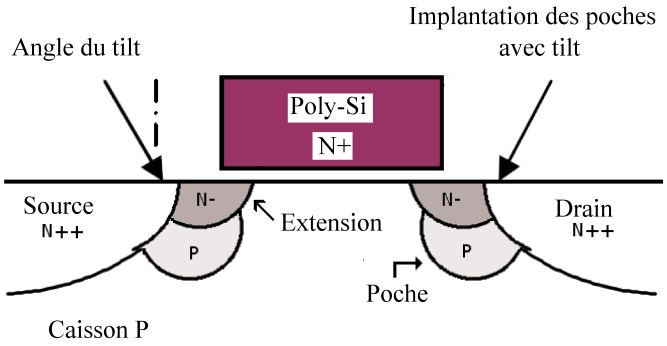

FIGURE 4.6 – *Illustration de l'implantation des poches pour un transistor NMOS : une implantation tiltée dont le dopage est de même nature que celui du substrat.*

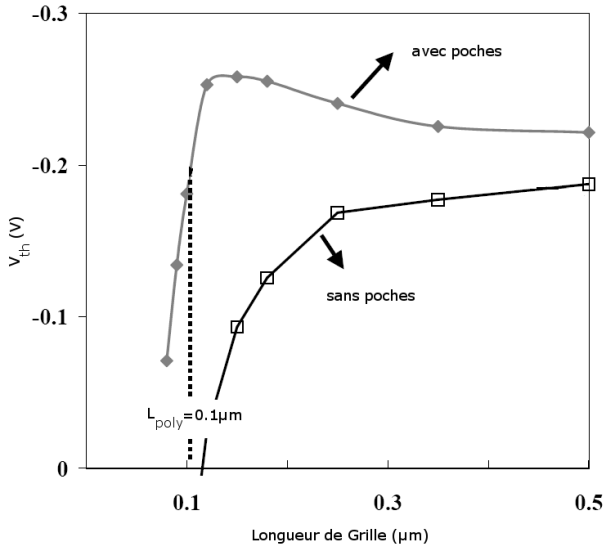

FIGURE 4.7 – *Caractéristiques $V_{th} = f(L_{poly})$ ($V_D = 1.5V$), obtenues par simulation pour deux architectures d'un transistor PMOS : sans et avec poches. L'implantation des poches permet de minimiser le phénomène de* ***roll-off****. Source : Gwoziecki (1999).*

En réponse à ce problème, Ruegsegger et Wagner ont développé un filtre statistique qui minimise l'erreur quadratique moyenne $E[(X - \hat{X})^2]$ entre X, la valeur réelle de la variable d'entrée et $\hat{X}$, son estimation [Ruegsegger et al. (1998)]. Sous certaines hypothèses statistiques, l'estimateur est formulé en étant simplement le produit de la mesure en ligne et d'un facteur S, appelé *Detuning factor*. Le facteur S ($0 < S < 1$) correspond à un degré de confiance dans la métrologie. Pour $S = 1$, la métrologie serait parfaite et la déviation mesurée serait réelle et due au procédé de fabrication. Au contraire, pour $S = 0$, la déviation mesurée est considérée comme étant du bruit.

⋈ En seconde position figure l'interaction de la variable de commande manipulée u_k avec des paramètres caractéristiques du produit, autres que le paramètre de sortie régulé y_k. Prenons l'exemple de la régulation FF entre les étapes de photolithographie et de gravure de grille. Son principe de fonctionnement est simple ; il s'agit d'ajuster le temps de *trimming* ou le flux d'oxygène en gravure selon la déviation de la longueur

de la résine mesurée en photolithographie. Cette méthode, bien que bénéfique en ce qui concerne la variabilité de L_{poly} [Marquet (2008)], pourrait dégrader le profil de grille (SWA). Afin de remédier à celà, Ruegsegger et Wagner proposent de sélectionner une recette parmi n recettes prédéfinies et qualifiées [Ruegsegger et al. (1999)]. Le choix du nombre de recettes n, de la valeur de la commande u_k pour chaque recette ainsi que des frontières de partage de l'intervalle de variation de u_k est optimisée.

4.3 Architecture du contrôleur Gravure-Implantation des Poches

Suite à la gravure du poly-silicium, une mesure de la longueur de la grille est réalisée par scattérométrie. Le niveau étant critique, la mesure des lots n'est pas échantillonnée et les mesures individuelles par site sont enregistrées dans une base de données. Comme le montre la figure 4.8, le contrôleur Gravure-Implantation des Poches comporte deux parties, un filtre et un module de prédiction.

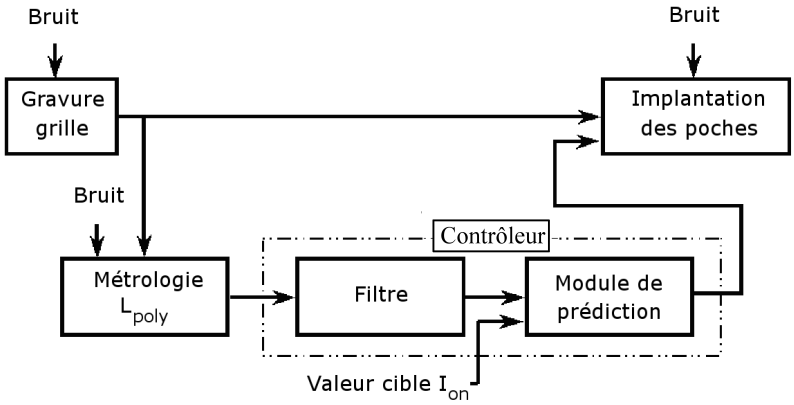

Figure 4.8 – *Vue schématique de la régulation Gravure-Implantation des Poches*

4.3.1 Le filtre

Le filtre écarte simplement les mesures individuelles dont la qualité d'ajustement **GOF**[2] (***G**oodness **O**f **F**it*) est inférieure à un seuil donné. Ces mesures ne seront pas incluses dans le calcul de la moyenne du lot. Nous n'avons pas intégré les travaux de Wagner et Ruegsegger sur le filtre **MMSE** (***M**inimum **M**ean **S**quare **E**stimator*) car le rapport entre la variance de l'outil de métrologie $\simeq (0.2)^2$, à savoir la scattérométrie, et la variance du procédé de gravure de grille $\simeq 2^2$ demeure assez faible de l'ordre de $(0.2)^2/2^2 \simeq 0.01$ [Ruegsegger et al. (1998)]. Avec un rapport aussi bas, la probabilité de sur-ajustement est minime.

Facteurs	Min	Max
L_{poly} (nm)	113	130
Dose I2 des Poches NMOS (at/cm^2)	2.5e13	3.1e13
Dose I2 des Poches PMOS (at/cm^2)	1.35e13	1.65e13

TABLE 4.1 – *Tableau récapitulatif des intervalles de variation des facteurs :1/ La longueur de grille 2/La dose d'implantation des poches pour les transistors NMOS et PMOS.*

4.3.2 Le module de prédiction

Le module de prédiction est construit autour d'un modèle prédictif polynomial qui lie la valeur du paramètre de sortie à la dose d'implantation et la dimension critique de grille L_{poly}. Le choix d'un modèle non-linéaire de type réseaux de neurones, comme c'est le cas dans [Cardarelli et al. (1996), Park et al. (2005)], n'est pas envisageable pour différentes raisons :

⋈ La version actuelle de l'application logicielle chargée d'intégrer l'algorithme de régulation, à savoir ProcessWorks de la société Adventa, ne peut être hôte de réseaux de neurones.

⋈ L'utilisation de modèles non-linéaires ne peut être justifiée sans passer par la case des modèles linéaires. Il faudrait en effet démontrer l'inefficacité de ces derniers avant d'envisager de complexifier la modélisation.

Nous avons modélisé deux paramètres critiques, à savoir, le courant de saturation I_{on} et le courant de fuite I_{off}. Dans ce manuscrit, nous avons choisi d'exposer les résultats du seul courant de saturation (I_{on}). La méthodologie employée est transposable d'une façon identique au courant de fuite I_{off}. Pour concevoir le modèle, nous avons opté pour un plan d'expériences de type plan factoriel complet à 3 niveaux 3^2 [Ozil (2002)]. Le choix des intervalles de variation des facteurs est résumé dans le tableau 4.1. Suite à une analyse de variance, seuls les termes $Dose$, $Dose^2$ et L_{poly} s'avèrent significatifs (voir figure 4.9). Le modèle peut s'écrire ainsi :

$$I_{on} = \phi + \alpha\, L_{poly} + \beta\, Dose + \gamma Dose^2 + \omega(0, \sigma_p^2), \tag{4.3}$$

Avec ϕ : Une constante appelée **centrage du produit**,
α : La sensibilité de I_{on} aux variations de L_{poly},
β, γ : Les coefficients du 1^{er} et du 2^{nd} ordre de la dose d'implantation,
ω : Les résidus du modèle. Par hypothèse, ω suit une loi normale $\mathcal{N}(0; \sigma_p^2)$.

Validation du modèle

Une fois le modèle établi par la méthode des moindres carrés, nous avons cherché à établir sa validité dans un environnement de production. En effet, la longueur de la grille et la dose des poches ne sont pas les seules sources de variabilité du courant de saturation. L'écart type des résidus σ_p estimé à partir du plan d'expérience n'est pas à l'image de l'erreur de prédiction réelle σ du modèle. L'objectif de cette étape

2. Le GOF est une valeur comprise entre 0 et 1, calculée pour chaque mesure individuelle en scattérométrie. Elle rend compte de la qualité de l'ajustement du modèle à la réponse scattérométrique mesurée par l'ellipsomètre : généralement, un GOF inférieur à 0.95 signifie que la mesure n'est pas à prendre en considération.

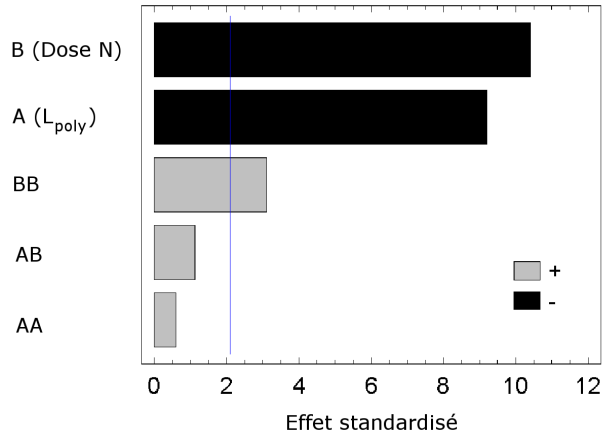

FIGURE 4.9 – *Diagramme de Pareto pour le paramètre I_{on} des transistors NMOS. Seuls les termes L_{poly}, Dose, $Dose^2$ sont significatifs à 95%*

est alors d'estimer l'erreur de prédiction du modèle $\hat{\sigma}$ et par la suite de statuer sur le potentiel gain du contrôleur en comparant $\hat{\sigma}$ et la variabilité lot à lot originale du courant de saturation. Pour cela, nous calculons simplement les résidus r de lots de production.

$$r = \hat{\alpha}\, L_{poly} + \hat{\beta}\, Dose + \hat{\gamma}\, Dose^2 - I_{on}, \tag{4.4}$$

$\hat{\sigma}$ est de l'ordre de $9\mu A$ (voir Figure 4.10), autrement dit, l'écart type asymptotique auquel tendra la variabilité de I_{on}, après la mise en production de la régulation, est de $9\mu A$. Sachant que le sigma actuel est de l'ordre de $12\mu A$, un gain de $3\mu A$ est potentiellement réalisable.

Calcul de la dose

Le modèle étant validé, il est implémenté au sein du module de prédiction. A chaque lot qui se présente au niveau de l'implantation des poches des transistors NMOS (ou PMOS), le contrôleur doit fournir la dose qui permettrait de ramener le courant I_{on} sur sa cible. Pour y parvenir, le module de prédiction ira chercher deux paramètres : la valeur cible $Target_{I_{on}}$ et L_{poly} moyen du lot. A ce stade, il s'agit d'une équation du second degré à une inconnue $Dose$.

Soit $K = \hat{\phi} + \hat{\alpha}\, L_{poly} - Target_{I_{on}}$, où L_{poly} est la moyenne du lot[3]. L'équation s'écrit alors :

$$g(Dose) = K + \hat{\beta}\, Dose + \hat{\gamma}\, Dose^2 = 0, \tag{4.5}$$

Le discriminant s'écrit $\Delta = \hat{\beta}^2 - 4K\hat{\gamma}$. Raisonnons par l'absurde. Si Δ était négative, ceci voudrait dire qu'aucune dose ne permettrait d'avoir un courant I_{on} égal à $Target_{I_{on}}$. A partir du fait que L_{poly} appartient au domaine expérimental de validité du modèle, ce constat est aberrant. Δ est alors positif ou nul. Mathématiquement,

3. L_{poly} doit appartenir au domaine de validité du modèle

deux solutions existent, et pourraient être confondues au cas où Δ serait nul.

$$Dose = (-\hat{\beta} \pm \sqrt{\Delta})/2\hat{\gamma}, \tag{4.6}$$

Pour choisir la dose appropriée, nous considérons deux cas :

1. $\hat{\gamma} > 0$, d'un point de vue physique, la fonction g est forcément décroissante [4] : la solution est par conséquent inférieure à $-\hat{\beta}/2\hat{\gamma}$ comme le montre le tableau de variation de la fonction g (Tableau 4.2),
2. $\hat{\gamma} < 0$, pour un raisonnement identique au précédent, la solution serait supérieure à $-\hat{\beta}/2\hat{\gamma}$ (Tableau 4.3).

Dans les deux cas, la solution retenue est $Dose = (-\hat{\beta} - \sqrt{\Delta})/2\hat{\gamma}$.

x	$-\infty$	0	$-\frac{\hat{\beta}}{2\hat{\gamma}}$	$+\infty$
$g'(x)$		$-$ $-$		$+$
g	$+\infty$	$\searrow$ K	$\searrow$ $K - \frac{\hat{\beta}^2}{4\hat{\gamma}}$	$\nearrow$ $+\infty$

TABLE 4.2 – *Tableau de variation de la fonction g dans le cas d'un $\hat{\gamma}$ positif.*

x	$-\infty$	$-\frac{\hat{\beta}}{2\hat{\gamma}}$	0	$+\infty$
$g'(x)$	$+$		$-$ $-$	
g	$-\infty$	$\nearrow$ $K - \frac{\hat{\beta}^2}{4\hat{\gamma}}$	$\searrow$ K	$\searrow$ $-\infty$

TABLE 4.3 – *Tableau de variation de la fonction g dans le cas d'un $\hat{\gamma}$ négatif.*

4.4 Simulation du contrôleur

L'objectif ultime de la régulation (FF) entre la gravure de grille et l'implantation des poches est d'aligner la longueur effective du canal L_{eff} de lot à lot et garantir ainsi une fenêtre vitesse-consommation du circuit dans les spécifications. Par ailleurs, il a été démontré d'une façon empirique que la vitesse du circuit et sa consommation sont fortement corrélées au couple I_{on} & I_{off} relatifs aux transistors courts. La question est alors de savoir lequel de ces deux paramètres choisir pour satisfaire l'objectif mentionné ci-dessus ? Quel impact aurait le choix de l'un ou l'autre sur le test IDDQ ?

Afin de donner des éléments de réponse à ces dernières questions, nous nous sommes basés sur le plan d'expérience réalisé précédemment pour modéliser le courant de saturation I_{on}, le courant de fuite I_{off} et la tension de seuil V_{th} des transistors

4. Physiquement, plus nous augmentons la dose d'implantation des poches déposée autour de la source et du drain, plus le courant de saturation des transistors NMOS diminue : $\hat{\beta} < 0$.

courts de type NMOS. D'une manière similaire à celle décrite dans le paragraphe précédent, nous avons estimé la variance des résidus de chacun des modèles. La variance de L_{poly} mesuré en production, a été aussi estimée. La figure 4.10 montre que l'erreur de prédiction du modèle de I_{on} est de l'ordre de $9\mu A$(I_{off} et V_{th} sont respectivement de l'ordre de 0.087 (log(A)) et 0.005 (mV)). Elle est engendrée par plusieurs facteurs dont :

⋊ La variabilité de caractéristiques géométriques et physiques non prises en compte par la modélisation, typiquement l'épaisseur de l'oxyde de grille,

⋊ L'incertitude des coefficients du modèle,

⋊ Le bruit de la mesure.

La déviation standard lot-à-lot de la longueur du poly-silicium est de l'ordre de 1.8 nm.

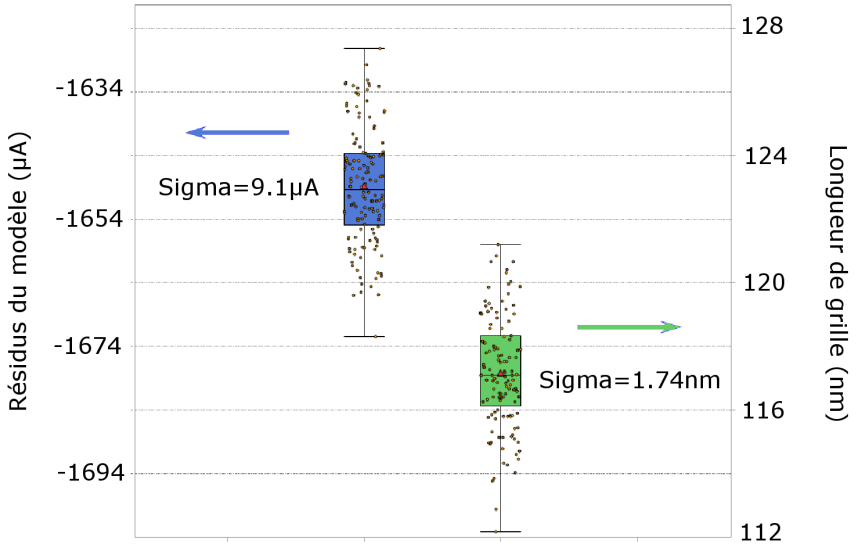

FIGURE 4.10 – *Estimation à partir d'une population de lots de production de 1/ l'erreur de prédiction du modèle $I_{on} = f(L_{poly}, Dose)$ relatif aux transistors NMOS 2/ et de l'écart-type de L_{poly}*

La simulation consiste simplement à générer un vecteur X_i de p nombres aléatoires, appartenant à une distribution normale de moyenne 117 nm et d'écart type 1.8 nm (voir Figure 4.10). Le vecteur X_i correspond à une série de p lots virtuels qui se présente à l'étape d'implantation des poches après l'achèvement de la gravure de grille. En se basant sur la modélisation réalisée au préalable, nous estimons la dose qui permettrait d'annuler l'écart entre la valeur de la variable de sortie, à savoir le courant de saturation ou le courant de fuite, et sa valeur cible. La moyenne ainsi que l'écart-type des paramètres électriques prédits des p lots sont par la suite calculés. Ce processus est répété n fois et les résultats obtenus sont comparés à ceux obtenus avec une dose d'implantation gardée constante.

4.4.1 Simulation avec I_{on} comme variable de sortie

Le courant cible $Target_{I_{on}}$ est de $650\mu A$. Comme le montre la figure 4.11, les tendances des moyennes ($I-A$ & $I-B$) et de l'écart-type ($II-A$ & $II-B$) des

caractéristiques électriques reflètent un effet positif de l'ajustement de la dose. La régulation permet à la fois un meilleur centrage du produit ainsi qu'une réduction de l'écart-type du courant I_{on} avec un gain potentiel de l'ordre de 30%. Le constat est identique pour le courant I_{off}, avec un gain de l'ordre de 15%. A noter que cette configuration ne permet pas de centrer le courant I_{off} sur sa valeur cible -7.8 log(A).

De surcroît, nous constatons une dégradation accrue de la variabilité de la tension de seuil dont le sigma passe de $5.10^{-3}\,mV$ en moyenne à $10^{-2}\,mV$, soit 100 % d'augmentation (voir figure 4.12). Ceci s'explique par un poids relativement plus important du dopage moyen du canal que du L_{poly} dans la variation de la tension V_{th} . La modélisation de V_{th} révèle en effet une sensibilité à L_{poly} presque nulle.

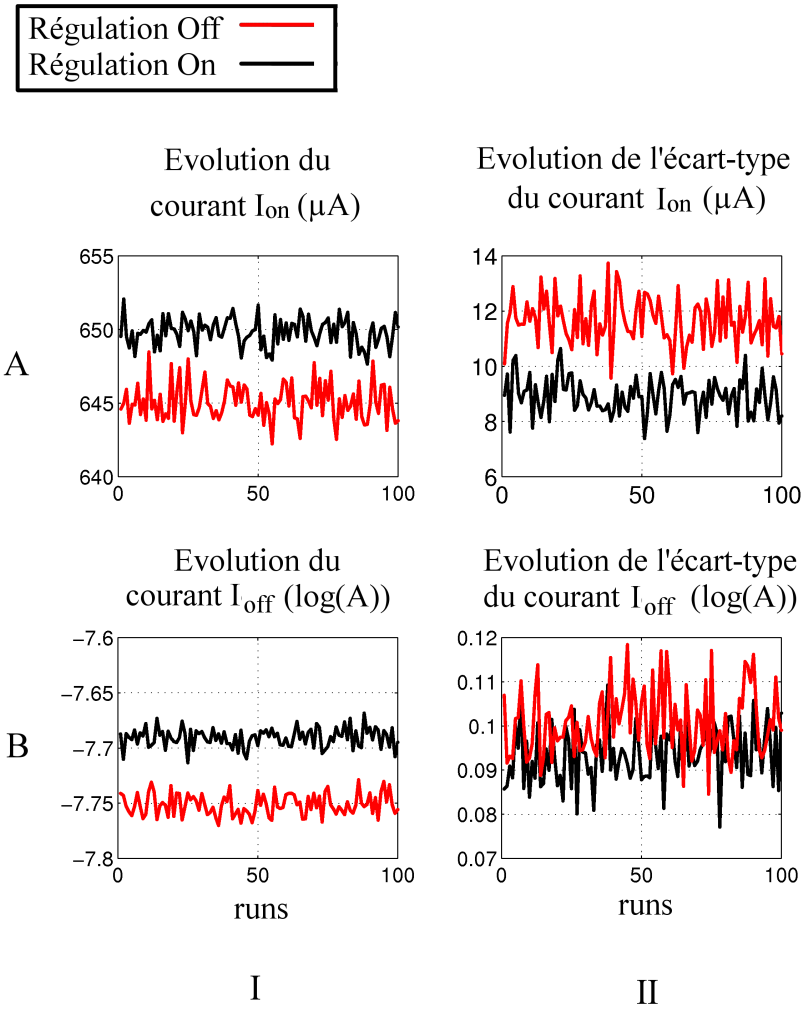

FIGURE 4.11 – *Résultats des simulations pour transistors NMOS, où la régulation vise à réduire la variabilité du courant $I_{on}(p = 100)$. Simulation de l'évolution des paramètres I_{on} et I_{off} 1/ à dose constante, 2/ et à dose ajustée (régulation en fonctionnement).*

Cette dernière remarque concernant l'impact de la régulation sur les variation de V_{th} nous a conduit lors de la mise en production à fixer une borne minimale $Dose_{min}$ et une borne maximale $Dose_{max}$ que la dose d'implantation ne pourra franchir. Au delà

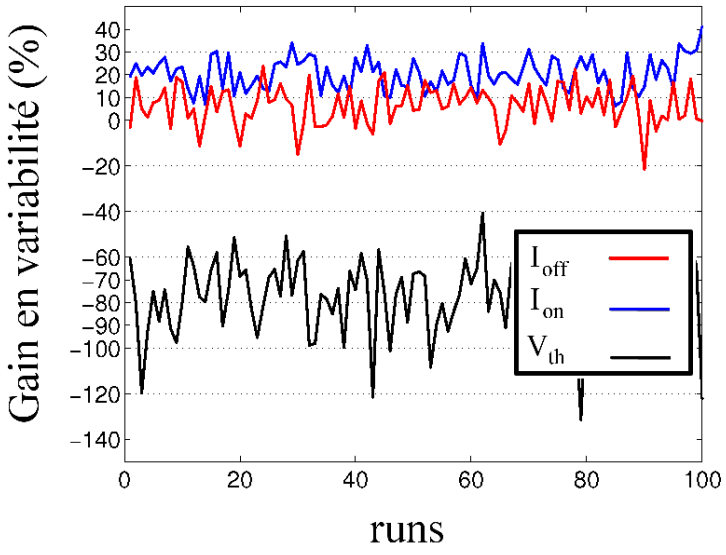

FIGURE 4.12 – *Résultats des simulations pour transistors NMOS, où la régulation vise à réduire la variabilité du courant $I_{on}(p = 100)$. Simulation de l'évolution du gain en écart-type relatif aux paramètres I_{on}, I_{off} et V_{th}. Le ratio relatif à V_{th} est négatif, synonyme d'une augmentation de sa variance.*

d'une déviation donnée de L_{poly}, par rapport à la longueur nominale, l'ajustement en dose est gardé inchangé (voir figure 4.13). Cette action nous permettra d'éviter des valeurs de V_{th} trop excentrées de la valeur cible.

4.4.2 Simulation avec I_{off} comme variable de sortie

Le courant cible $Target_{I_{off}}$ est de $-7.8log(A)$. Les résultats des simulations de cette alternative, illustrés en figure 4.14, vont dans le même sens que ceux de la première configuration : un gain de 23% pour le courant I_{on} $(II - A)$ et de 20% pour le courant I_{off} $(II - B)$. L'effet négatif de la régulation sur la variabilité de la tension de seuil V_{th} est bien moindre, de l'ordre de 50% (voir figure 4.15).

4.5 Implémentation et Résultats en production

Suite aux résultats encourageants prédits par la simulation, la première configuration du contrôleur avec le courant I_{on} comme paramètre de sortie a été implémentée en production. Le motif mesuré après gravure de grille est un réseau dense, où le pas h est égal à 310 nm (voir figure 1.11). Un seul produit à fort volume sera concerné par cette campagne de pré-production. L'objectif est de quantifier l'impact réel de l'ajustement de la dose d'implantation sur les caractéristiques du transistor ainsi que le gain de la régulation, en vue d'une mise en production générale pour tous les produits.

D'un point de vue pratique, l'algorithme (équations, limites, filtre) a été implémenté dans une application dédiée aux régulations **run to run**, Process Works de la société Adventa [5]. Cette application est configurée pour communiquer avec les équi-

5. www.adventact.com

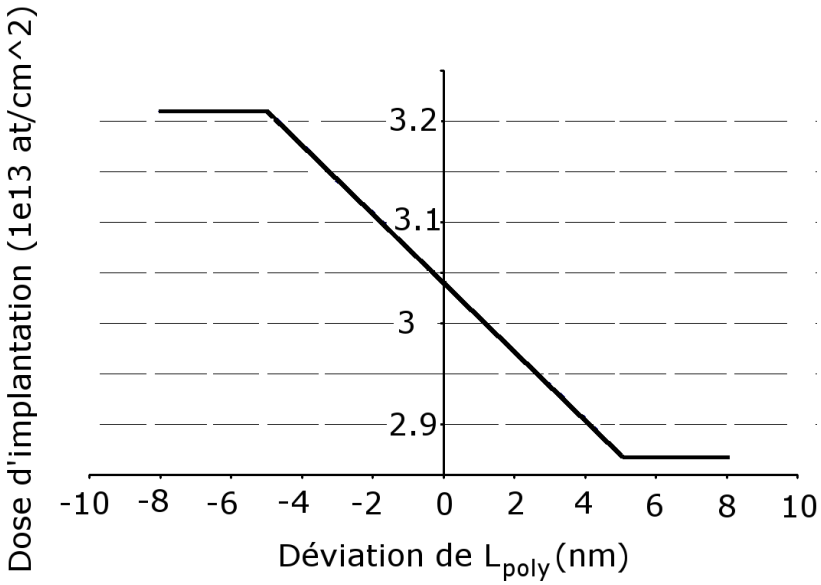

FIGURE 4.13 – *Vue schématique du profil de la dose d'implantation appliquée en fonction de la déviation de L_{poly} par rapport à la longueur nominale. Au delà d'une déviation de 5nm en valeur absolue, la correction en dose est constante.*

pements de gravure, les éllipsomètres et l'implanteur. Le diagramme générique, en figure 4.16, fait apparaître les différentes voies de communications entre l'application, la base de données de production (APC [6]), et le parc d'équipements. Pour chaque lot posé sur un équipement de process, auquel ProcessWorks est associé, l'application ira chercher différents éléments de son contexte de fabrication (route, technologie, opération, etc) à partir de la base APC. Seuls les lots concernés par les stratégies existantes de run-ro-run seront considérés et donneront effet à un échange d'information entre l'équipement de métrologie et de process d'un côté et l'application ProcessWorks de l'autre côté. Cette dernière stockera un certain nombre de variables dont les ajustements de la recette dans une base de données, appelée Gemstone. A noter que la régulation est complètement automatisée et transparente aux yeux des opérateurs. Les utilisateurs ont toutefois accès à Gemstone via une interface dédiée.

La figure 4.17 montre l'évolution du I_{on} des transistors NMOS avant et après la mise en production de la régulation. Nous constatons une nette réduction de la variabilité et un recentrage du produit sur la cible. Pour appuyer ce constat, la tendance du sigma cumulatif de la population a été tracée (voir figure 4.18). Chaque point correspond à l'écart type de la population des lots qui le précède. A l'instant où le contrôleur est mis en marche, le sigma est remis à 0. L'écart-type passe de $12\mu A$ à $9\mu A$ et la variabilité lot à lot est ainsi réduite de 25%. Le gain dépasse 40% pour le courant de saturation I_{on} des PMOS.

L'impact du fonctionnement du contrôleur sur le courant de fuite est aussi positif. L'évolution du sigma cumulatif du I_{off} des transistors PMOS, illustrée en figure 4.19, montre un gain de l'ordre de 25%. Conformément aux simulations, la variabilité de V_{th} a augmenté d'une manière significative de l'ordre de 100% (voir figure 4.20). L'écart type passe de $5mV$ à $10mV$. Cette augmentation, relativement importante, n'a

6. La base de données APC est mise à jour toutes les huit heures

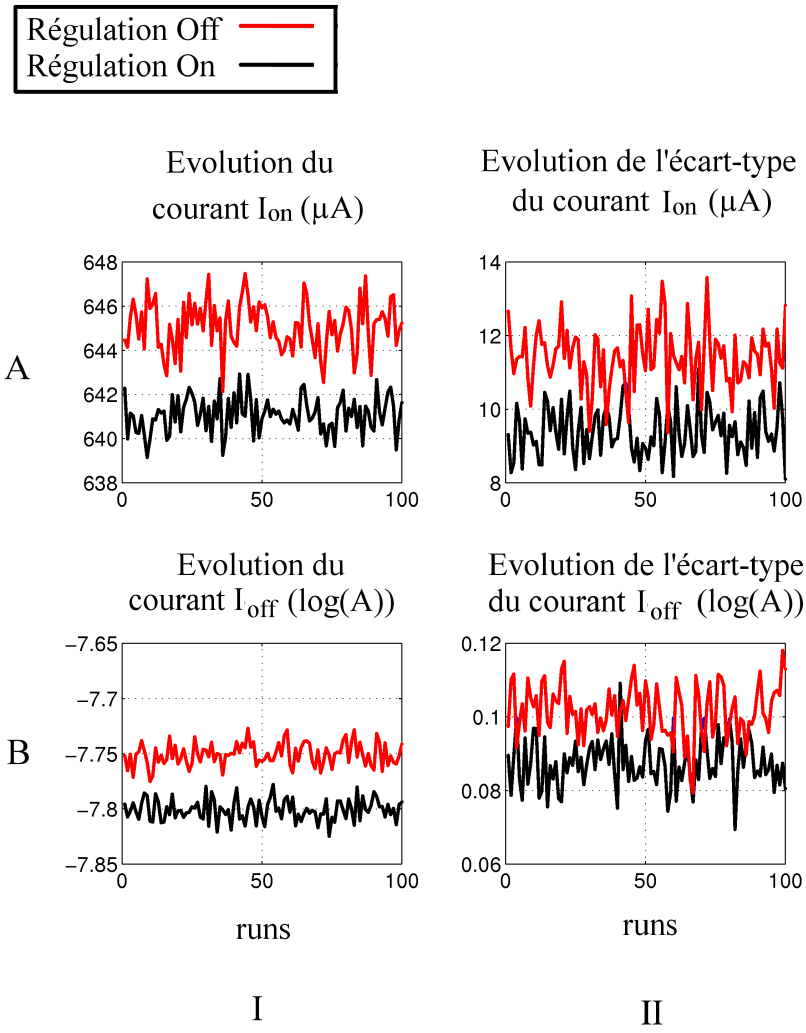

FIGURE 4.14 – *Résultats des simulations pour transistors NMOS ($p = 100$), où la régulation vise à réduire la variabilité du courant $I_{off}(p = 100)$. Simulation de l'évolution des paramètres I_{on} et I_{off}1/ à dose constante, 2/ et à dose ajustée (régulation en fonctionnement).*

qu'un impact modéré sur les fonctionnalités du produit, contrairement aux courants de saturation et de fuite.

Au delà de l'augmentation des Cpk paramétriques des produits, il est important de souligner l'amélioration du rendement sur certains produits sensibles, conséquence directe de la réduction des plaques rejetées pour I_{on} hors spécifications (2 à 5%). En complément à d'autres actions menées au sein des ateliers de photolithographie (remplacement des fours PEB RHP par des fours S-HRP) et de gravure, la régulation (FF) Gravure/Implantation des poches a permis d'une façon effective d'atteindre des performances technologiques, alignées avec les celles des usines les plus avancées, notamment un sigma du courant I_{on} de l'ordre de 17 μA.

Le service contrôle de production, chargé de la création des routes[7] des nouveaux

7. Une route est une succession dans l'ordre chronologique d'étapes élémentaires de process et de

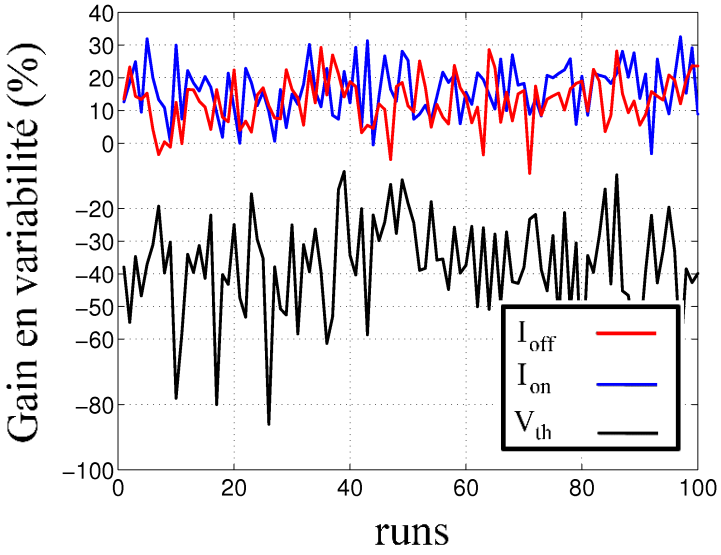

FIGURE 4.15 – *Résultats des simulations pour transistors NMOS, où la régulation vise à réduire la variabilité du courant $I_{off}(p = 100)$. Simulation de l'évolution du gain en écart-type relatif aux paramètres I_{on}, I_{off} et V_{th}. Le ratio relatif à V_{th} est négatif, synonyme d'une augmentation de sa variance.*

produits, a aussi bénéficié des bienfaits de cette régulation. Les produits ont souvent des routes différentes à cause simplement d'un courant **Target**$_{I_{on}}$ différent. La capacité du contrôleur à gérer plusieurs produits avec des cibles de courant différentes a permis de réduire le nombre de routes à créer.

Depuis l'implementation du contrôleur, les limites mises en place pour réduire l'excursion de la dose d'implantation des poches ont d'ores et déjà montré leur utilité. L'introduction d'un nouveau paramètre associé à la mesure du CD d'une nouvelle structure scattérométrique, a suffi pour induire un calcul erroné du CD au sein de l'application ProcessWorks. 45 lots ont été ainsi implantés avec une dose qui correspond à la dose limite maximale. Si ces limites n'existaient pas, nous aurions eu 45 lots de rebut.

4.6 La criticité du biais iso-dense Δ

On appellera le **biais iso-dense** la différence entre la dimension critique d'une ligne de poly-silicium dense et une ligne de poly-silicium isolée. Pour comprendre l'impact potentiel qu'a le biais iso-dense sur la performance du contrôleur, il faut rappeler deux points : la mesure en scattérométrie est réalisée sur un réseau dense ($h = 0.310\mu m$), alors que la mesure des caractéristiques paramétriques, notamment les courants I_{on} et I_{off}, est faite sur un transistor isolé. De ce fait, ajuster le courant d'un motif isolé en se basant sur la mesure d'un motif dense sous-entend l'existence d'une fonction g qui lie la dimension critique d'une ligne isolée à celle d'une ligne dense. Même si cette fonction g existe, comme le montre la figure 4.21 (g est simplement une

métrologie. A chaque étape, sont spécifiés la recette, les équipements qualifiés, etc. La création d'une route est un processus long qui demande vérification d'un nombre important d'ingénieurs.

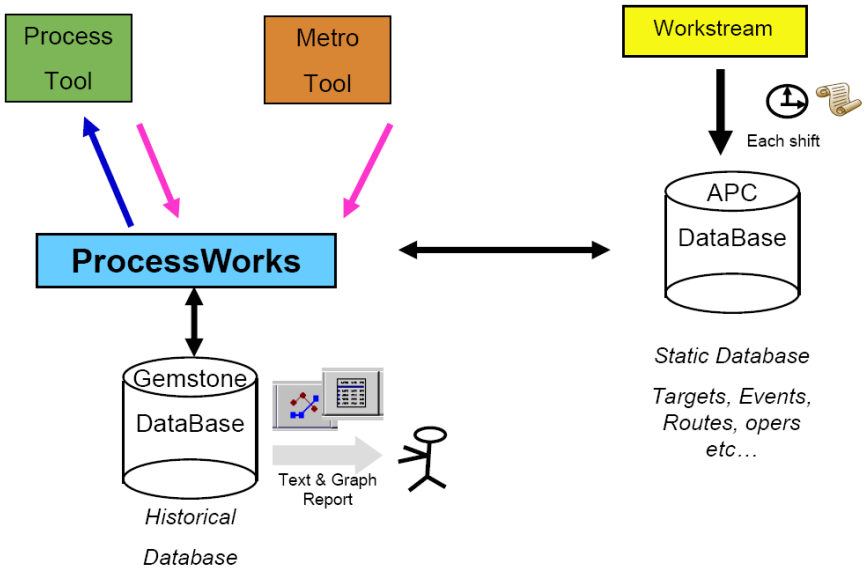

FIGURE 4.16 – *Vue schématique de l'environnement de l'application ProcessWorks. Source : Adventa ProcessWorks Basic Training.*

régression linéaire), elle est sujette à des variations dans le temps. En pratique, nous parlons plutôt de biais iso-dense Δ.

4.6.1 Mise en évidence de la criticité du biais iso-dense

Afin de valider ce raisonnement, nous avons réalisé un plan d'expérience où nous faisons varier le paramètre Δ. L'insolation de la grille est réalisée à travers une illumination annulaire. Cette technique d'exposition qui fait partie des techniques d'amélioration de la résolution **RETs** (***R**esolution **E**nhancement **T**echniques*), est caractérisée par le rapport $\sigma_{inner}/\sigma_{outer}$. La figure 4.22 présente une vue schématique d'une illumination annulaire faisant apparaître les différentes dimensions. Sachant que ce rapport affecte le biais iso-dense [Levinson (2005)], nous avons fait le choix de faire varier le σ_{inner} tout en gardant le σ_{outer} constant et égal à 0.86. Nous avons aussi inclus la dose d'énergie (mJ/cm^2) dans ce plan d'expérience, et ceci afin de couvrir toute la plage de variation de la longueur des lignes denses et isolées. Le plan retenu est encore une fois un plan factoriel à 3 niveaux de type 3^2. Les limites de variation des 2 facteurs sont résumées dans le tableau 4.4.

Suite à la réalisation des différentes expériences définies par le plan d'expérience, le lot est gravé et mesuré en ligne en scattérométrie et en microscopie électronique à balayage. Il est important de mentionner que la microscopie électronique à balayage nous sert à mesurer les lignes de poly-silicium isolées. La figure 4.23 présente un nuage de points qui témoigne, comme attendu, de l'absence de corrélation entre L_{poly} d'une ligne dense et L_{poly} d'une ligne isolée : la fonction g n'est plus valide.

Une fois achevé, le lot est testé au niveau du test paramétrique. Nous constatons que le courant I_{on} relatif aux transistors NMOS courts n'est plus corrélé à la mesure en scattérométrie du motif dense ($R^2 < 60\%$, voir figure 4.24), contrairement au motif isolé qui garde une corrélation forte avec le courant I_{on} ($R^2 \simeq 90\%$). En résumé, la variation du biais iso-dense Δ est très critique vis-à-vis de la performance du contrô-

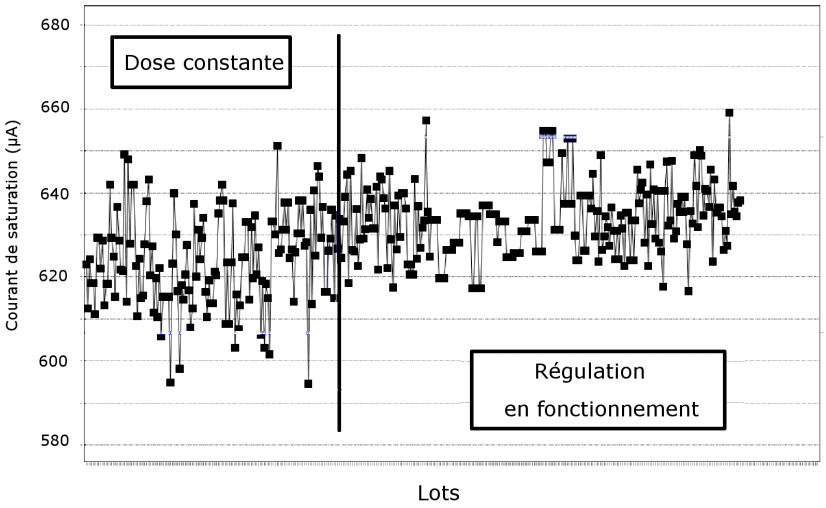

FIGURE 4.17 – *Evolution du courant de saturation I_{on} des transistors NMOS courts (0.13μm) avant et après la mise en production de la régulation. Meilleur centrage et réduction de la dispersion lot à lot du courant de saturation.*

Facteurs	**Minimum**	**Maximum**
Dose d'énergie (mJ/cm^2)	21	22.2
σ_{inner}	0.29	0.43

TABLE 4.4 – *Tableau récapitulatif des intervalles de variation des facteurs :1/ La dose d'énergie du scanner 2/Sigma inner.*

leur. Dans le cas d'une dérive ou d'un décalage de Δ, le contrôleur peut décentrer le produit, voire induire une perte de rendement pour les produits les plus sensibles.

4.6.2 Solution : une nouvelle structure scattérométrique

Afin d'augmenter la robustesse du contrôleur face aux variations du biais iso-dense, nous nous proposons de qualifier une nouvelle structure de scattérométrie à pas h plus large ($h = 0.780\mu m$) et de comparer son comportement à celui d'un motif isolé et d'un motif dense. Pour cela, nous avons repris le plan d'expériences décrit dans le paragraphe précédent, et nous y avons ajouté une mesure complémentaire, celle de la nouvelle structure. La figure 4.25 montre que la mesure en ligne du nouveau motif s'apparente à une mesure d'une ligne isolée en microscopie électronique à balayage. L'évolution de ces deux structures, sous l'effet des variations de la dose d'énergie et du σ_{inner} est équivalente.

Ce constat est conforté avec les résultats du test paramétrique. Le courant I_{on} demeure corrélé à la mesure de la nouvelle structure, contrairement au cas du motif dense (voir figure 4.26). Par ailleurs, la sensibilité du courant de saturation aux variations de la dimension critique de la grille est identique indépendamment du motif mesuré.

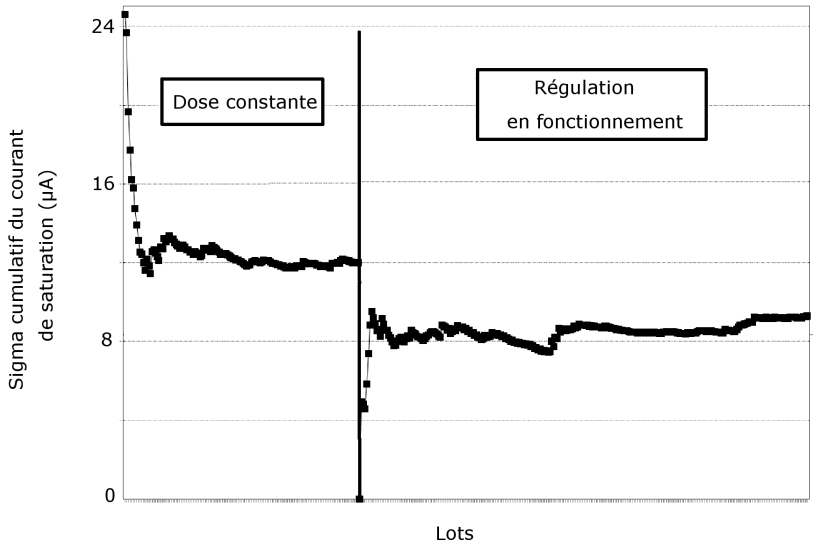

FIGURE 4.18 – *Evolution de l'écart-type cumulatif du courant de saturation I_{on} des transistors NMOS courts (0.13μm) avant et après la mise en production de la régulation. Le gain est de l'ordre de* 25%.

4.7 Définition d'agrégats (Threads)

Suite à une première expérience aux résultats positifs, il a été décidé de mettre en place la régulation Gravure-Implantation pour différents produits. Le modèle est-il valide pour tous les produits de la même technologie ? La réponse est non. Le modèle comporte plusieurs paramètres estimés au préalable à partir d'un plan d'expérience. Ils sont valables pour tout le domaine expérimental, mais *à priori* pour le seul produit (jeu de réticules) qui a servi pour la réalisation du plan d'expériences. L'objet de ce paragraphe est de spécifier les différents éléments (équipements, réticules, etc) qui pourraient faire varier ces paramètres.

4.7.1 La sensibilité α du paramètre de sortie à la longueur de grille

α est la sensibilité du courant de saturation aux variations de la dimension critique de grille. Le courant de saturation correspond à celui d'un transistor isolé, que l'on note **A**, tandis que la mesure en ligne en scattérométrie est celle d'un réseau de lignes, que l'on note **B**. Les deux motifs résident dans les lignes de découpe. La performance du contrôleur est intimement liée d'abord à la qualité du modèle, mais aussi à la fonction qui lie les dimensions des deux motifs **A** et **B**. Réduire cette fonction à la corrélation entre lignes denses et isolées g, voire au biais iso-dense Δ, dans le cas d'une structure B dense, est un premier indicateur du fonctionnement de la régulation. Il permet d'alerter l'ingénieur en cas de dérive, avant l'étape d'implantation des poches.

Par ailleurs, g est une régression linéaire dont la pente dépend du réticule et de l'équipement de gravure (voir figure 4.27). α est alors nécessairement dépendant du couple (Réticule/Equipement de gravure).

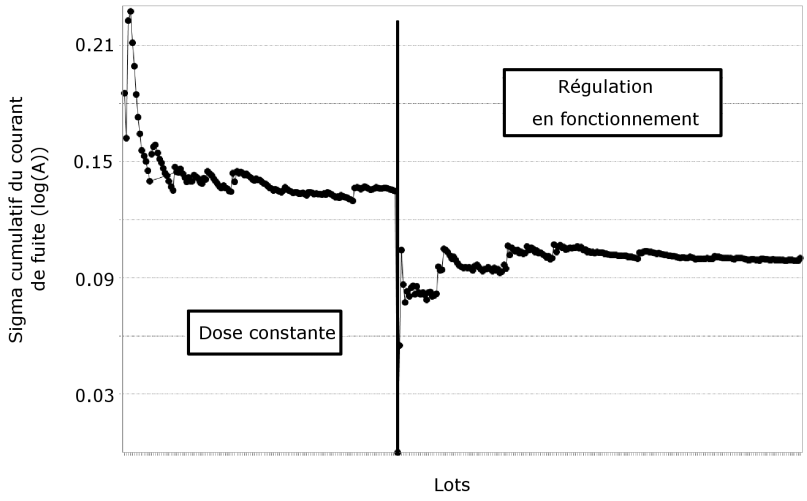

FIGURE 4.19 – *Evolution de l'écart-type cumulatif du courant de fuite I_{off} des transistors PMOS courts (0.13μm) avant et après la mise en production de la régulation. Le gain est de l'ordre de* 25%.

4.7.2 Les coefficients β et γ

Il s'agit des coefficients de premier et second ordre de la dose d'implantation dans le modèle polynomial adopté. Au moment de la mise en production de la régulation, un seul implanteur à moyen courant est qualifié et capable d'ajuster la dose en temps réel. De ce fait, nous n'avons pas à investiguer la dépendance du couple (β,γ) en fonction de plusieurs équipements, comme c'est le cas avec le coefficient α. Nous admettons enfin que ces coefficients sont indépendants du produit, et cela peut être justifié par les deux éléments qui suivent :

⋈ Au sein d'une même technologie, les produits diffèrent[8] uniquement par le jeu de réticules utilisé, qui correspond naturellement à un circuit et une application différente. Le processus de fabrication (implantation + recuit) est toutefois identique.

⋈ Pour les technologies avancées qui sont concernées par le déploiement de la régulation FF entre la gravure de la grille et l'implantation des poches, le nombre des produits s'élève à plusieurs dizaines. Réaliser un plan d'expérience pour chaque produit afin d'estimer les coefficients (β,γ) ne peut être envisageable par l'industriel, étant donné le coût trop élevé d'une telle initiative.

4.7.3 Le centrage du produit

Le centrage du modèle est un paramètre qui prend en compte le profil de la grille, le biais iso-dense et toute particularité (épaisseur d'oxyde, etc) du produit au sein d'une même famille technologique qui affecte le courant de saturation I_{on}. Il est fonction du jeu de réticules et de l'équipement de gravure.

En résumé et dans le cadre d'une première version de la régulation de compensation (FF) entre la gravure de grille et l'implantation des poches, un modèle sera

8. C'est le cas le plus souvent. D'une façon exceptionnelle, certains produits requièrent la modification ou le rajout d'un module.

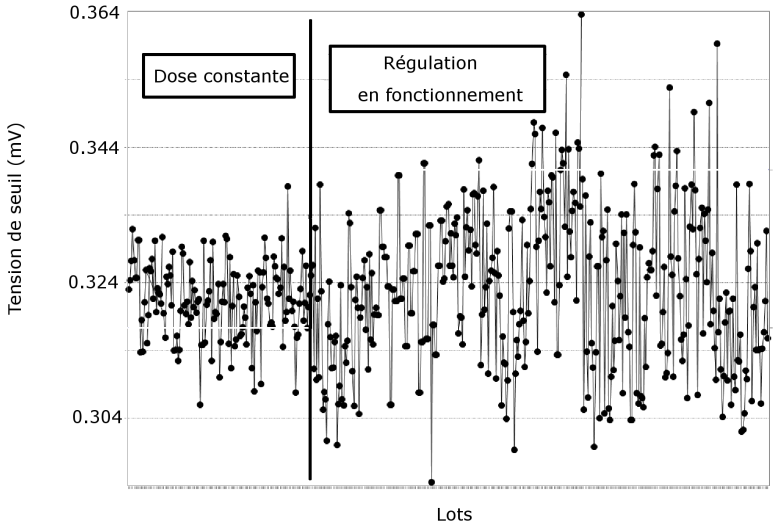

FIGURE 4.20 – *Evolution de la tension de seuil V_{th} des transistors NMOS courts (0.13μm) avant et après la mise en place de la régulation. Cette dégradation liée au contrôleur n'est pas critique.*

défini pour chaque couple (Réticule du niveau Grille, Equipement de gravure Grille). Le déploiement du contrôleur pour tous les produits critiques nécessitait ainsi une phase de modélisation, qui visait à spécifier le coefficient α et le centrage du produit ϕ. Le besoin en résultats de test paramétrique (lots de production) rend cette phase plus ou moins longue selon le produit. Dans le cas de produits à très faible volume ou très récents, cette phase pourrait s'étaler sur quelques mois, tandis que dans le cas de produits à fort volume, pour lesquels nous disposons d'un certain nombre de lots finis, et testés au test paramétrique (PT), elle est de l'ordre de quelques minutes.

4.8 Intégration de l'épaisseur de l'oxyde de grille

L'épaisseur de l'oxyde de grille t_{ox} est la deuxième source de variabilité des courants I_{on} et I_{off} des transistors courts (voir Chapitre 2). Souhaitant améliorer davantage la performance de la régulation, nous avons envisagé de prendre en compte le t_{ox} lors de l'ajustement de la dose des poches. La mesure t_{ox} sera réalisée en ellipsométrie et envoyée au même titre que la mesure de la longueur de grille L_{poly}, au module de prédiction pour l'intégrer dans le calcul de la dose. Signalons que cette nouvelle version de la régulation ne peut être envisagée pour cibler le courant de saturation (configuration retenue pour la régulation originale). En effet, si une telle configuration permet la réduction de variabilité de I_{on}, elle engendre une augmentation certaine de celle du I_{off}. Un raisonnement à L_{eff} constant permet d'illustrer ce constat.

⋈ La formulation de la tension de seuil par l'approche dite Transformation Tension-Dopage (VDT, *Voltage Dopage Transformation*) permet d'obtenir une relation analytique simple que j'ai choisi de rappeler ci dessous [Gwoziecki (1999)]. Ainsi, la tension de seuil d'un transistor de longueur effective L_{eff} s'obtient à partir de la relation canal long auquel se retranche un terme ΔV_{th} (Equation 4.8). A partir de cette équation 4.8, nous pouvons voir que la diminution de V_{th} avec L_{eff} est d'autant plus importante

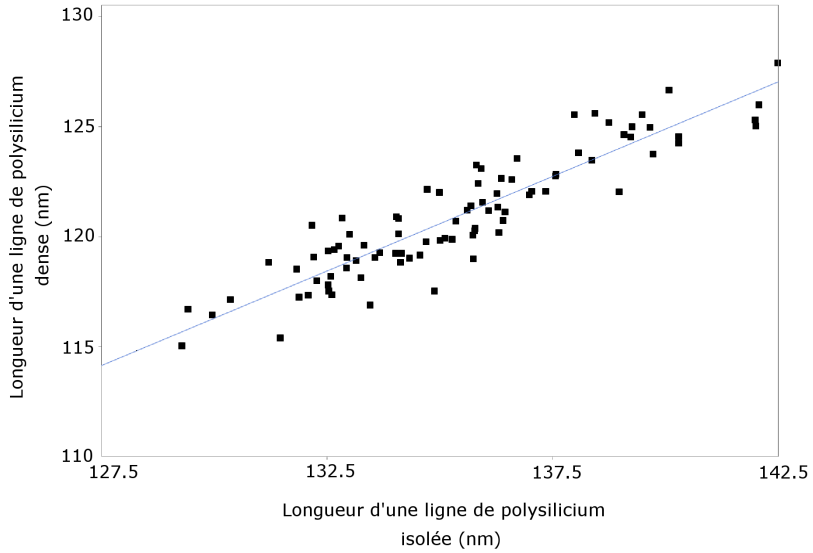

FIGURE 4.21 – *Illustration de la fonction g pour un produit A ;* $L_{poly}^{dense} = g(L_{poly}^{isole})$ *est une regression linéaire*

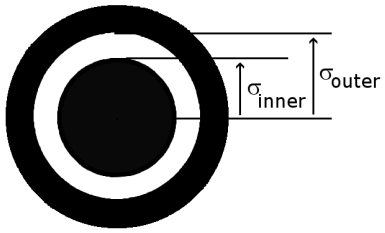

FIGURE 4.22 – *Avec une illumination annulaire, la lumière traverse un anneau centré sur l'axe optique, avant d'insoler le réticule. Cet anneau est caractérisé par un diamètre intérieur* σ_{inner} *et un second extérieur* σ_{outer}.

que l'épaisseur d'oxyde est élevée. Clairement, lorsque l'épaisseur de l'oxyde t_{ox} est supérieure à l'épaisseur nominale t_{nom}, l'effet $DIBL$ est exacerbé et la grille perd davantage le contrôle du canal [Skotnicki (2000)], ie : la tension de seuil V_{th} diminue. Les expressions du courant de fuite 4.2 et du courant de saturation 4.1 nous montrent de la même façon une tendance à la hausse du I_{off} et a contrario une tendance à la baisse du I_{on}. A partir de là, ajuster la dose d'implantation des poches pour viser une valeur cible $Target_{I_{on}}$ conduirait à une réduction du dopage moyen effectif du canal et par conséquent à une augmentation supplémentaire du courant I_{off} ;

$$V_{th} = V_{th_{long}} - \Delta V_{th} \tag{4.7}$$

$$= V_{th_{long}} - \kappa \frac{t_{ox}}{L_{eff}^2} \tag{4.8}$$

Où κ est une constante positive indépendante de L_{eff} et t_{ox}.

⋊ Dans le cas d'une épaisseur d'oxyde inférieure à l'épaisseur nominale, le courant de saturation aura tendance à augmenter, tandis que le courant de fuite tendra à diminuer. Pour compenser cette déviation et réduire son effet sur le courant I_{on}, le

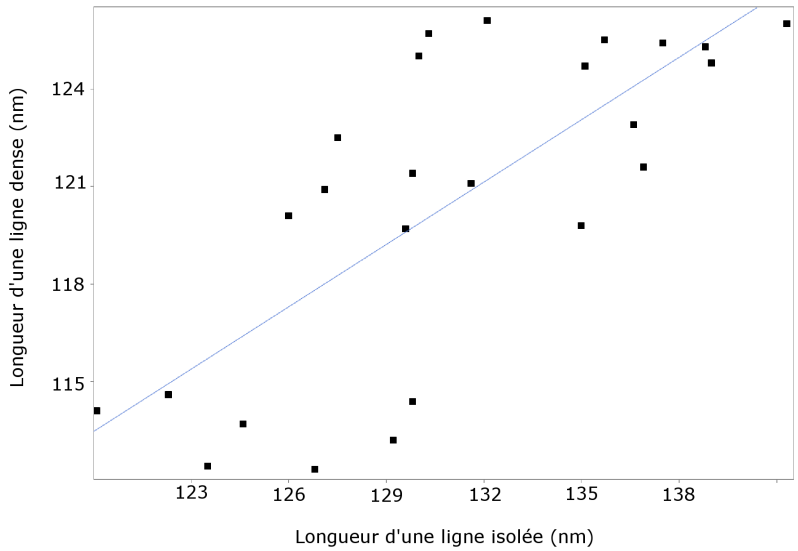

FIGURE 4.23 – *Suite au plan d'expérience réalisé (altération du biais iso-dense), la corrélation entre les dimensions d'une structure isolée et une structure dense n'existe plus.*

contrôleur commande une dose d'implantation plus élevée que la dose nominale. En conséquence, I_{off} est diminué davantage et sa variance est encore une fois diminuée.

4.8.1 Modélisation

Nous avons opté pour un plan d'expérience central composite avec trois facteurs : l'épaisseur d'oxyde t_{ox}, la longueur de grille L_{poly} et la dose d'implantation des poches. Le choix des intervalles de variations des paramètres est résumé dans le tableau 4.5.

Facteurs	**Min**	**Max**
L_{poly} (nm)	114	127
Épaisseur de l'oxyde de grille (Å)	19.7	24.9
Dose I2 des poches NMOS/PMOS ($1e13\,at/cm^2$)	$1.86 - 1$	$3.7 - 2$

TABLE 4.5 – *Tableau récapitulatif des intervalles de variation des facteurs :1/ La longueur de grille 2/ L'épaisseur de l'oxyde de grille et 3/ La dose d'implantation des poches pour les transistors NMOS et PMOS.*

A partir des mesures électriques du courant I_{off}, un modèle polynômial f est construit. Seuls les termes influents sur chaque réponse sont évidemment retenus. Le courant I_{off} est fonction des trois facteurs principaux (L_{poly}, $Dose$, t_{ox}), et des carrés de deux facteurs ($Dose^2$, t_{ox}^2). D'une manière identique à la méthodologie suivie dans la section 4.4, nous avons estimé l'erreur de prédiction de ce modèle f à partir des données de production. Dans la figure 4.28, nous comparons la variabilité de deux types de résiduels r_i à celle du courant I_{off}, et ceci pour trois produits différents **A**, **B**

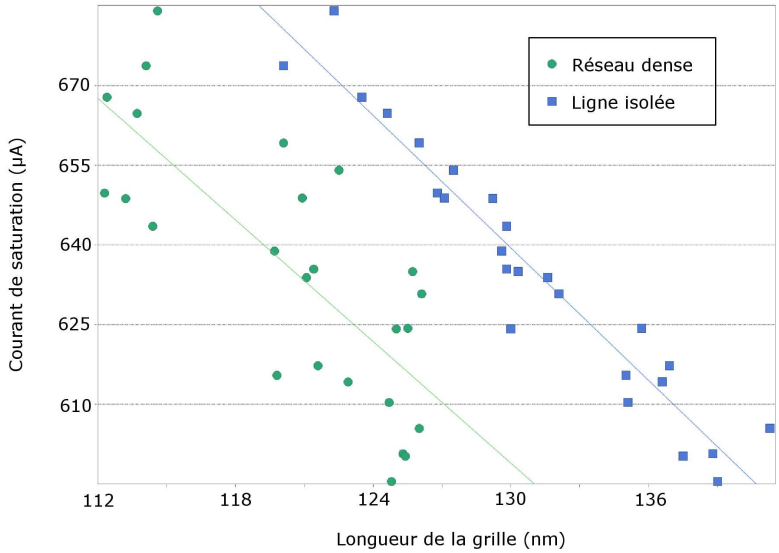

FIGURE 4.24 – *Caractéristique $I_{on} = f(L_{poly})$ pour les transistors NMOS courts (0.13µm). La variation du biais iso-dense Δ implique la perte de la corrélation entre la longueur des lignes denses et les paramètres électriques à l'instar de I_{on} : $R^2 = 59\%$. La longueur des lignes isolées demeure corrélée au courant I_{on} ($R^2 = 92\%$).*

et **C**[9].

⋊ Les résidus calculés à l'aide de l'expression $r_1 = I_{off} - f(L_{poly}, Dose, t_{ox})$;

⋊ Les résidus calculés à l'aide de l'expression $r_2 = I_{off} - f(L_{poly}, Dose, t_{ox} = 0)$, ie : seuls la longueur de grille L_{poly} et la dose $Dose$ sont prises en compte dans la prédiction du courant de fuite. L'épaisseur de l'oxyde est écartée.

Le modèle $f(L_{poly}, Dose)$ explique naturellement une partie de la variabilité du courant de fuite, qui varie selon le produit de 0.01 (10 %) à 0.05 log(A) (40%). Néanmoins, le rajout de l'épaisseur d'oxyde apporte peu, voire dégrade la capacité de prédiction du modèle. Une raison plausible que nous pouvons avancer pour expliquer cela est qu'il existe une ou plusieurs étapes de fabrication (espaceurs, recuit final,$\cdots$) qui auraient une influence telle sur les variations du courant I_{off} que les variations de t_{ox} passent au second plan et n'ont aucun poids. L'épaisseur d'oxyde t_{ox} présente de surcroît une faible variation lot à lot. Enfin, ces sources de variation majeures seront à identifier parmi des modules de process critiques qui n'ont pas été considérés dans le chapitre 2, notamment l'espaceur, les tranchés d'isolation (STI) et le recuit final.

4.9 CONCLUSION

Dans ce chapitre, nous avons conçu une régulation de compensation (FF) entre la gravure de la grille et l'implantation des poches. Le contrôleur s'appuie sur une mesure scattérométrique de la dimension critique L_{poly} pour ajuster la dose d'implantation des poches. L'objectif du contrôleur est en effet de compenser les déviations

9. Les produits diffèrent par l'architecture, mais le *process* est identique

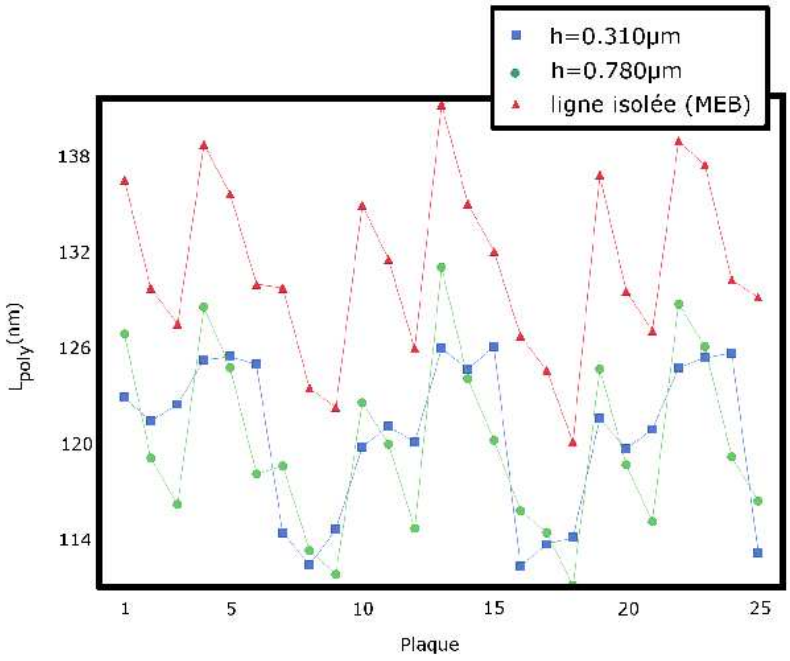

FIGURE 4.25 – *Résultats du plan d'expérience au niveau de la mesure en ligne de la longueur du polysilicium : Dans l'espace de variation de la dose d'énergie et du σ_{inner}, le motif isolé et celui dont le pas h est égal à 0.780μm évoluent d'une manière équivalente. De ce point de vue, le motif dense ($h = 0.310\mu m$) est un intrus.*

de L_{poly} et réduire les variations des caractéristiques électriques critiques autour de leurs valeurs cibles.

La simulation de cette régulation a permis de comparer deux configurations, où le paramètre de sortie utilisé est soit le courant de saturation (1) soit le courant de fuite (2). Bien que la variabilité de la tension V_{th} augmente dans le deux cas, les résultats des simulations mettent en exergue un effet positif certain sur la variation des courants I_{on} et I_{off}. La configuration (1) permet en effet de réduire l'écart-type du I_{off} et du I_{on}, respectivement de 15% et de 30%. L'extension de cette boucle avec la prise en compte de l'épaisseur d'oxyde a aussi été envisagée mais abandonnée à cause d'une capacité prédictive médiocre du modèle dans l'environnement de production.

Suite à l'implémentation de la régulation en production, les gains réels en variations lot à lot du I_{on} et du I_{off} sont venus confirmer ceux prédits par la simulation. Nous avons de surcroît réussi à atteindre les standards mondiaux en terme de variabilité globale du courant de saturation des transistors NMOS courts (0.13 μm), à savoir un sigma de $17\mu A$. Notons aussi un autre avantage de cette régulation : sa capacité à gérer plusieurs familles de produits avec des cibles de courants différentes.

Le contrôleur est maintenant en production depuis bientôt 3 ans et est devenu un outil clé dans le contrôle des procédés de la technologie 0.13 μm. A noter que le même contrôleur est en cours d'évaluation actuellement sur une technologie 90 nm. La performance du contrôleur peut cependant être compromise dans le cas d'une

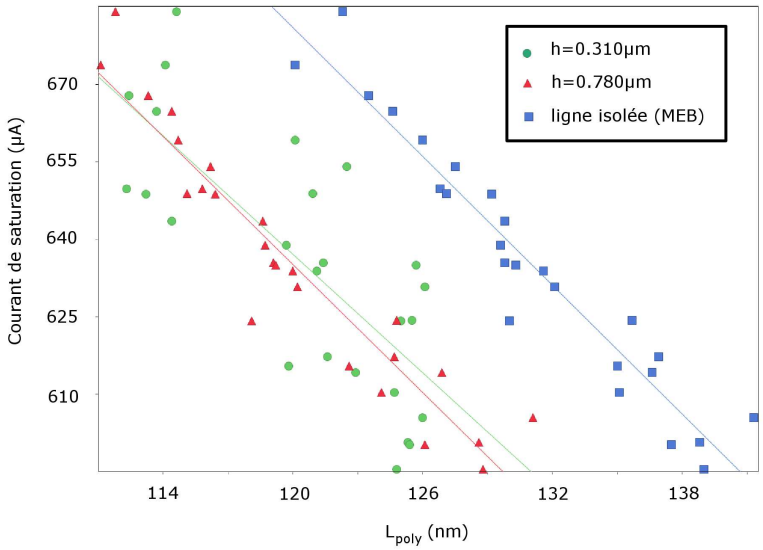

FIGURE 4.26 – *Résultats du plan d'expérience au niveau du test paramétrique : Dans l'espace de variation de la dose d'énergie et du σ_{inner}, le courant de saturation des transistors NMOS demeure fortement corrélé à la longueur des lignes isolées ($R^2 = 93\%$) et à celle des réseaux scattérométriques dont le pas h est égal à $0.780\mu m$ ($R^2 = 90\%$). Ceci n'est pas le cas des motifs denses ($R^2 = 59\%$).*

dérive ou d'un décalage du biais iso-dense Δ. Pour remédier à cela, nous envisageons, pour la deuxième version de cette régulation, de changer le motif mesuré en scattérométrie et adopter un réseau à pas plus large ($0.780\mu m$). Cette mesure augmentera la robustesse de la régulation face aux fluctuations de Δ.

Une autre évolution plus poussée de cette régulation consisterait à ajuster la dose d'implantation de plaque à plaque. Cette démarche ne peut être envisagée qu'en cas de progrès significatif en métrologie virtuelle [Chang et al. (2006), Khan et al. (2008)]. Dans ce contexte, la métrologie virtuelle aurait comme tâche de fournir à partir des quelques mesures individuelles de L_{poly}, et des informations en temps réel envoyées par les équipements de photolithographie et de gravure de grille, une estimation fiable de L_{poly} pour chaque plaque. Les travaux de Sendelbach montrent que l'intégration d'un scattéromètre à l'équipement de gravure est possible [Sendelbach et al. (2006)]. Elle permettrait d'une façon alternative et potentiellement moins complexe d'avoir une mesure moyenne de la longueur de grille pour chaque plaque, et ceci sans altérer le temps de gravure du lot.

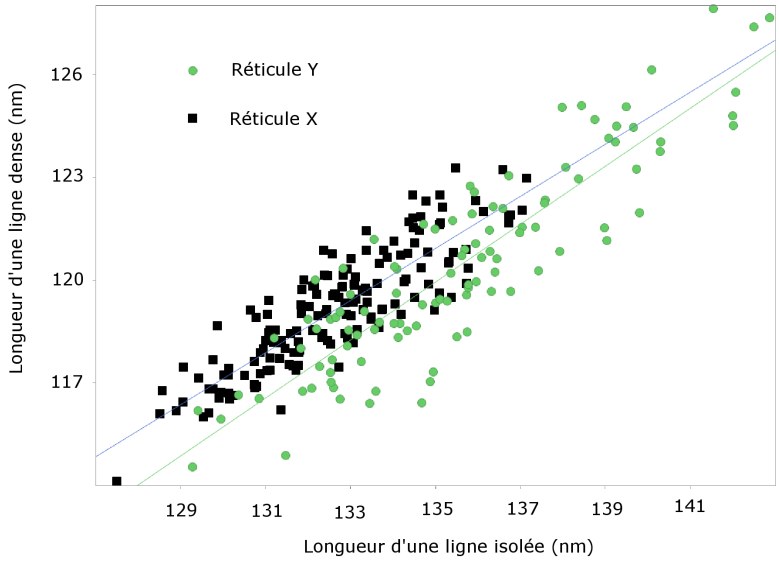

FIGURE 4.27 – *Illustration de la fonction g pour deux réticules X & Y ; $L_{poly}^{dense} = g(L_{poly}^{isole})$ est une regression, les pentes relatives aux deux réticules sont différentes de 10 %*

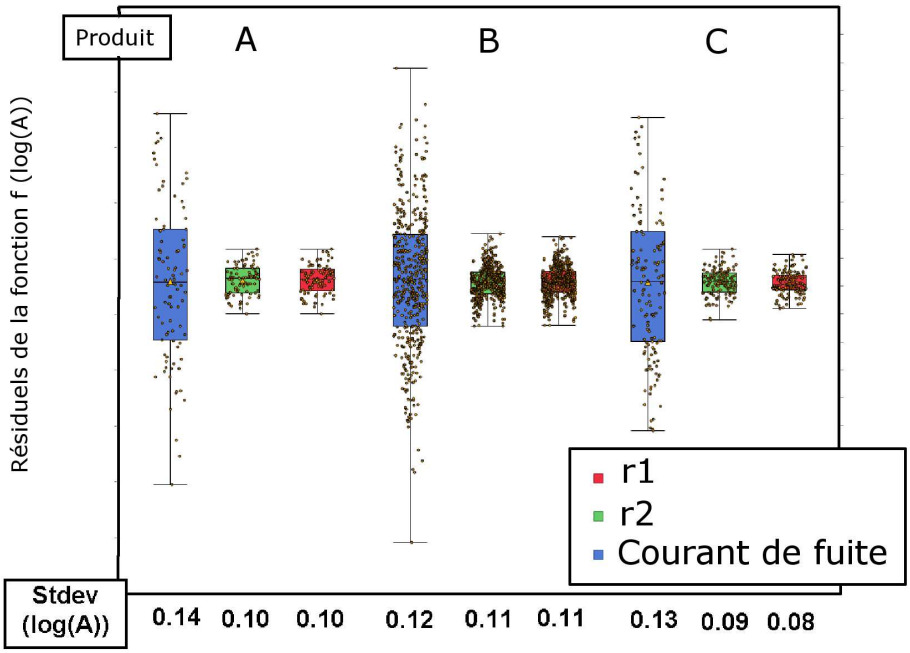

FIGURE 4.28 – *Estimation à partir d'une population de lots de production de 3 produits A, B & C de 1/l'erreur de prédiction du modèle $I_{off} = f(L_{poly}, Dose)$ en vert, 2/l'erreur de prédiction du modèle $I_{off} = f(L_{poly}, Dose, t_{ox})$ en rouge 3/et l'écart-type de I_{off} en bleu. Les mesures PT sont relatives aux transistors NMOS courts.*

Identification Du Procédé Lithographique

5

Sommaire

Dans un contexte de survie, du fait notamment d'un marché devenu très concurrentiel, les fabricants de composants sur le territoire européen, et notamment STMicroelectronics, ont entrepris un ensemble d'actions qui devrait permettre de réduire les coûts de fabrication et restaurer à la même occasion leur compétitivité face aux géants asiatiques. Parmi les actions menées dans ce cadre, j'en citerai deux principales : réduire les investissements et notamment l'achat d'équipements, et diversifier l'offre produit aux applications multiples (Imprimante, Portable, automobile, etc). Concrètement, le nombre de produits d'une même technologie a explosé avec des volumes plus au moins importants. Parallèlement, les équipements sont de moins en moins dédiés à des opérations ou des familles de produits spécifiques, et il est désormais courant de multiplexer différentes recettes sur la même machine. Ainsi décrit, cet environnement industriel est souvent nommé **high-mix fab** dans la littérature scientifique.

Face à cette tendance, les techniques de régulation avancée *run-to-run* (**R2R**), bien répandues dans les usines de fabrication de composants semi-conducteurs pour leur effet positif sur le rendement et les coûts de fabrication, doivent intégrer cette complexité et prendre en compte de nouvelles sources de variation : équipement de procédé et de métrologie, produit, niveau d'exposition, etc. Comment faire pour estimer le poids (ou la part) de chacune des sources de variation dans la déviation de la variable de sortie à un instant t quelconque ? La réponse à ce besoin est contenue dans la façon de réaliser le bloc **Estimateur**, ou encore **observateur** de la structure classique d'une boucle de régulation (Figure 5.1).

Suite à un état de l'art des principaux travaux qui ont traité cette question, nous allons nous intéresser à une architecture particulière de l'observateur, qui intègre

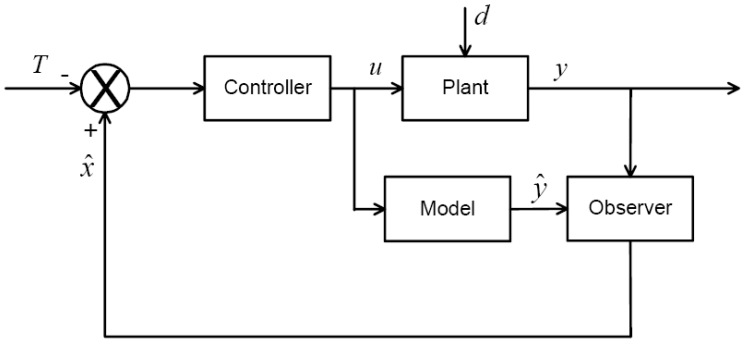

FIGURE 5.1 – *Architecture classique d'une boucle de régulation (R2R) composée d'un observateur, un modèle du procédé régulé et une loi de commande. Source : Campbell et al. (2002)*

un algorithme d'identification récursive (ou en ligne). A travers des simulations des données de production, nous allons chercher à caractériser les performances de deux catégories d'estimateurs récursifs, à savoir le filtre de Kalman ainsi qu'une variante des moindres carrés.

5.1 Prise en compte du contexte de fabrication : Etat de l'art

Michael Miller *et al.* ont été les premiers à s'intéresser à l'impact de la multiplicité des produits et des procédés [1] sur les boucles de régulation (R2R) dans l'industrie du semi-conducteur. Dans [Miller et al. (1997)], ils identifient quatre stratégies de contrôle différentes, à savoir les contrôleurs indépendants, groupés, coopératifs et enfin le contrôle composite.

L'indépendance des contrôleurs est une notion qui a été reprise par Bode dans ses travaux de doctorat [Bode (2001)]. Il s'agit d'identifier les sources de variation de la variable étudiée, et en extraire les plus pertinentes afin de définir ce qu'on appelle le *manufacturing context*. Face à chaque réalisation de ce contexte, est implantée une boucle de régulation avec un paramétrage optimal propre, fonctionnant d'une façon indépendante des autres contrôleurs. Pour davantage de clarté dans cet exposé, prenons l'exemple de la régulation de la dimension critique du polysilicium (Grille). Supposons avoir à réaliser trois produits différents de la même famille technologique. Une analyse off-line nous révèle que les équipements d'exposition en amont de la gravure, au nombre de deux, ainsi que ceux de gravure, au nombre de trois, forment des sources de variation non négligeables. Cette stratégie consisterait alors à concevoir $3 \times 2 \times 3 \equiv 18$ boucles, relative chacune à une combinaison du triplet (produit, équipement d'exposition, équipement de gravure).

Au delà d'un nombre potentiellement très grand de boucles à mettre en place et une réelle incapacité à gérer les passages d'une boucle à une autre, cette stratégie présente un défaut majeur, celui de la boucle *morte*. Dans le cas de l'exemple précédent, si nous avons la même fréquence de réalisation pour les différents combinaisons, chacune des boucles serait alimentées en données 5% du temps en moyenne : Un pourcentage trop bas pour garantir de bonnes performances.

Pour palier à cet inconvénient [2], Miller *et al.* suggèrent le contrôle groupé comme alternative [Miller et al. (1997)]. Il s'agit de regrouper les produits ou les équipements qui auraient un comportement *similaire* vis à vis de la variable régulée et réduire ainsi le nombre de boucles à déployer. Comme le souligne Toprac dans sa présentation orale [Toprac (2004)], cette stratégie n'est qu'une solution conceptuelle, et non une méthode automatisée. De surcroît, elle nécessite un travail d'engineering manuel lourd pour définir les nouveaux groupes ou agrégats.

Quant au contrôle composite, il requiert une modélisation physique globale du procédé en fonction des caractéristiques des plaquettes (produit, topographie, dimensions physiques). Un seul contrôleur est alors réalisé et associé à l'équipement du procédé. Encore une fois, il s'agit d'une solution très difficile à accomplir, du fait de la complexité des phénomènes physiques et chimiques mis en oeuvre dans les procédés de fabrication, et de la difficulté à mesurer certaines caractéristiques des plaquettes (topographie, etc).

La dernière solution, les contrôleurs coopératifs, est celle qui suscitera notre intérêt dans le reste de ce chapitre. L'idée centrale de ce mode de contrôle est de pouvoir

1. Si nous désirons faire un parallèle entre ces travaux pionniers de Miller *et al.* et les nôtres, les *procédés* seraient à mettre en face des niveaux d'exposition.

2. Toprac [Toprac (2004)] parle de *data starving*

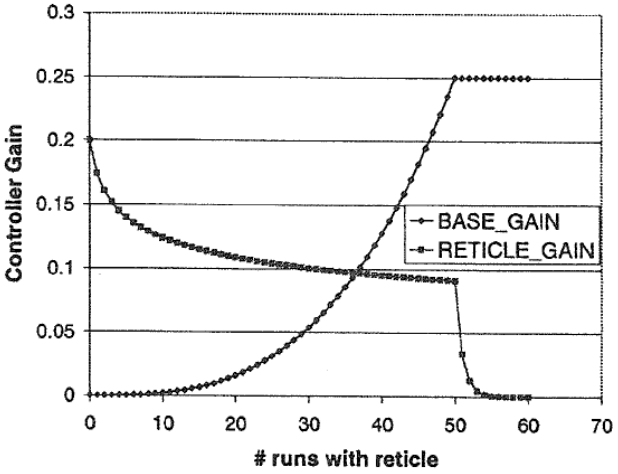

FIGURE 5.2 – *L'évolution des gains du contrôleur en fonction de l'usage du réticule. Le gain relatif au réticule a une valeur initiale non nulle et décroît au fur et à mesure que le réticule correspondant est utilisé. Au bout de* 50 *runs, gain_Ret* $= 0$*. D'une façon opposée, gain_Base croît d'une façon monotone avant d'atteindre un palier au* $50^{ème}$ *run. Source : Stuber (2003)*

partager entre les différentes boucles toute *information* concernant la perturbation. Les sources des perturbations sont alors découplées et la déviation de la variable étudiée est attribuée aux différentes composantes du *manufacturing context*. Dans le cas de l'exemple précédent, cette stratégie est traduite par l'équation 5.1.

$$\varepsilon_{Total} = \varepsilon_{Product} + \varepsilon_{Scanner} + \varepsilon_{Etcher} \tag{5.1}$$

où ε_{Total} est la déviation de la variable de réponse (variable contrôlée), $\varepsilon_{Product}$ est la déviation induite par le produit, $\varepsilon_{Scanner}$ est celle induite par l'équipement d'exposition et enfin ε_{Etcher} celle induite par l'équipement de gravure.

Cette démarche est aussi cohérente avec une politique d'échantillonnage où il est essentiel de maximiser l'utilisation des mesures dont nous disposons, et de minimiser le nombre de plaquettes test (*send-ahead wafers*). Elle permettrait aussi d'identifier les sources de toute déviation assez rapidement. Une première application industrielle de ce mode de contrôle est recensée chez la compagnie *Texas Instruments* (**TI**) que Stuber dévoile dans son article [Stuber (2003)]. Le modèle de base est la suivant.

$$\mathbf{CD}_k = M\,\mathbf{Dose}_k + Ret + Base_k \tag{5.2}$$

Où **CD** est la dimension critique de la ligne de résine,
Dose est la dose d'énergie,
M est ce qu'on appelle communément la pente,
Base est un offset caractéristique du couple Résine/Machine d'exposition,
Ret est l'Offset du réticule.

Les équations qui permettent la mise à jour des offsets *Base* et *Ret* sont formulées ci dessous.

$$Ret_{k+1} = Ret_k + gain\_Ret\,(\mathbf{CD}_k - \mathbf{CD}_{Target}) \tag{5.3}$$
$$Base_{k+1} = Base_k + gain\_Base\,(\mathbf{CD}_k - \mathbf{CD}_{Target}) \tag{5.4}$$

Où *gain_Ret* et *gain_Base* sont des gains qui évoluent dans le temps et $\mathbf{CD}_{Target}$ est la valeur cible du **CD**. Nous pouvons remarquer au passage qu'il s'agit simplement d'un correcteur EWMA à gain variable. Pour les cinquante premiers *runs* d'un réticule, la déviation du **CD**, $\mathbf{CD}_k - \mathbf{CD}_{Target}$, est totalement ou partiellement attribuée au réticule, comme le montre le graphe 5.2. Dans le cas d'un ancien réticule (nombre de *runs* supérieur à 50), le terme *Ret* demeure constant et seul l'offset *Base* est mis à jour.

Cette façon de faire est intuitive et ne repose sur aucun développement mathématique qui garantit l'estimation non-biaisée de *Ret* et *Base*. L'objectif de ce chapitre est de concevoir un contrôleur coopératif qui repose sur une théorie plus formelle. Ce contrôleur sera générique, transposable à plusieurs ateliers. Via des simulations de données virtuelles, nous chercherons à établir les performances d'une telle régulation en terme de biais, de fluctuations statistiques et de capacité de poursuite.

5.2 Prise en compte du contexte de fabrication : Modélisation

L'objet de ce paragraphe est de décrire la problématique d'un point de vue mathématique. Nous allons d'ores et déjà évoquer des mots clés de l'atelier de photolithographie dans la definition des sources de variabilité prises en compte. Ceci n'altère en aucun cas le caractère générique de ce récit qui pourrait s'appliquer à d'autres ateliers. Soit CD_k la mesure de la dimension critique réalisée suite au run k et z_k la valeur de l'état global définie par :

$$\begin{cases} CD_k &= \beta\, u_k + z_k \\ z_k &= \mu + s_{sc,k} + s_{l,k} + s_{r,k} + s_{c,k} \end{cases} \tag{5.5}$$

Où s_{sc}, s_l, s_r et s_c sont les biais relatifs respectivement à l'équipement d'exposition, au niveau d'exposition, au réticule (ou masque) et à l'équipement de métrologie[3]. β est la sensibilité du CD aux variations de la dose d'énergie u_k. Enfin, μ est une constante inconnue qui est aussi à estimer.

Soit n_i le nombre d'éléments pour chaque catégorie $i \in \{sc, l, r, c\}$. Dans un format matriciel, la variable de sortie du procédé CD_k à l'instant k s'écrit à partir de l'équation 5.5 :

$$z_k = \begin{bmatrix} 1 & \mathbf{T}_k \end{bmatrix} \mathbf{\Theta}_k = \mathbf{H}_k \mathbf{\Theta}_k \tag{5.6}$$

où $\mathbf{T}_k$ est un vecteur ligne de 1 et de 0, de dimension $\sum n_i$. Il définit le contexte de fabrication du run k. $\mathbf{\Theta}_k = \begin{bmatrix} \mu & s_{l_1} & s_{l_2} & \dots & s_{c_{n_c}} \end{bmatrix}'$ est le vecteur des paramètres inconnus de dimension $(1 + \sum n_i)$. Considérons le cas idéal où nous disposons de $(1 + \sum n_i)$ réalisations, chacune correspond à une configuration différente. L'objectif est alors de déterminer le vecteur $\mathbf{\Theta}$ de telle sorte que :

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1} & \mathbf{T} \end{bmatrix} \Theta = \mathbf{H}\,\Theta \tag{5.7}$$

où $\mathbf{Z}$ est le vecteur colonne des $(1 + \sum n_i)$ mesures de z_k, $\mathbf{1} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}'$, et $\mathbf{T}$ est une matrice de dimension $(1 + \sum n_i) \times \sum n_i$. Chacune de ses lignes correspond à un

3. Il s'agit des initiales de **l**ayer, **sc**anner, **r**eticle et **c**dsem.

vecteur $\mathbf{T}_k$ représentant un run particulier.

Les travaux de Ming-Da Ma *et al.* [Ma et al. (2008)], de Hanish [Hanish (2006)], de Firth *et al.* [Firth et al. (2006)] ainsi que ceux de Wang *et al.* [Wang et al. (2007), Wang et al. (2008)] nous révèlent une réalité inhérente au contrôle coopératif ainsi défini. Il est établi que même dans le cas particulier le plus favorable, la matrice $\mathbf{H}$ est de rang non complet. Pour appuyer ce constat, Wang *et al.* ont émis le théorème suivant :

Théorème 1 *Soit un processus dont le contexte comprend m catégories, chacune possède n_i éléments $(i = 1, 2, \ldots, m)$, alors le rang de matrice* $\mathbf{T}$ *satisfait l'inégalité suivante :*

$$rang(\mathbf{H}) = rang(\mathbf{T}) \leq \sum_{i=1}^{m} n_i - m + 1 \tag{5.8}$$

En clair, la matrice $\mathbf{H}$ -définie par l'équation 5.7- a un rang déficitaire au mieux de m. Par conséquent, le système n'est pas observable : l'observation de la variable de sortie ne permet pas de retrouver le vecteur d'état $\mathbf{\Theta}$ initial. Il n'est alors pas garanti d'avoir une estimation non-biaisée du vecteur $\mathbf{\Theta}$, et ceci, quelle que soit la méthode d'estimation mise en oeuvre (Moindres carrés, Kalman, Maximum de vraisemblance, etc). Pour remédier à cela, plusieurs stratégies ont été adoptées au sein de la communauté scientifique. La plus simple, comme l'ont suggéré Pasadyn et Edgar [Pasadyn et Edgar (2005)], est de rajouter aux runs de production, des mesures complémentaires ou des contraintes afin d'estimer certains offsets ou d'en réduire le nombre. Dans ce cas, l'industriel a habituellement recours aux tâches de qualité[4]. Alternativement, Firth et al. [Firth et al. (2006)] ont proposé d'altérer l'équation 5.7 de façon à tenir compte de l'estimation du vecteur $\mathbf{\Theta}$ à l'instant précédant.

$$\begin{bmatrix} \mathbf{Z}_q \\ \hat{\mathbf{\Theta}}_k \end{bmatrix} = \begin{bmatrix} \mathbf{H}_q \\ \mathbf{I} \end{bmatrix} \hat{\mathbf{\Theta}}_{k+1} \tag{5.9}$$

où $\mathbf{Z}_q$ et $\mathbf{H}_q$ sont des matrices qui correspondent respectivement aux matrices $\mathbf{Z}$ et $\mathbf{H}$, tronquées et de taille q[5]. $\mathbf{I}$ est la matrice identité. L'idée derrière ce formalisme est de mettre à jour uniquement les biais ou offsets intervenus dans ces q derniers runs. Les autres biais sont gardés constants.

Ming-Da Ma *et al.* ont développé une toute autre méthode qui s'appuie sur la technique d'ANOVA [Ma et al. (2008)]. Les valeurs absolues des états individuels étant inaccessibles, il suggère d'estimer des valeurs relatives par rapport à une moyenne globale. Si nous reprenons la notation jusqu'alors utilisée dans ce paragraphe, cela revient simplement à augmenter le système 5.5 de 4 équations.

$$\begin{cases} z_k & = \mu + s_{sc,k} + s_{l,k} + s_{r,k} + s_{c,k} \\ \sum_{i=1}^{n_{sc}} s_{sc_i} & = 0 \\ \sum_{i=1}^{n_l} s_{l_i} & = 0 \\ \sum_{i=1}^{n_r} s_{r_i} & = 0 \\ \sum_{i=1}^{n_c} s_{c_i} & = 0 \end{cases} \tag{5.10}$$

4. appelées aussi *send-ahead* wafers
5. Les lignes correspondantes aux q runs les plus récents sont gardées.

D'une manière équivalente, ce système peut s'écrire sous forme matricielle :

$$\mathbf{Z}^{\dagger} = \begin{bmatrix} \mathbf{Z} & 0 & 0 & 0 & 0 \end{bmatrix}' = \mathbf{H}^{\dagger}\mathbf{\Theta} \tag{5.11}$$

Avec

$$\mathbf{H}^{\dagger} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n \text{ runs.}} \quad \mathbf{T} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{4 \text{ cond.}} \quad \underbrace{\begin{matrix} 1 & \dots & 1 & 0 \\ 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \end{matrix}}_{n_s \text{ col.}} \quad \underbrace{\begin{matrix} 0 & \dots & 0 & 0 \\ 1 & \dots & 0 & 1 \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \end{matrix}}_{n_l \text{ col.}} \quad \underbrace{\begin{matrix} 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \\ 1 & \dots & 0 & 0 \\ 0 & \dots & 1 & 0 \end{matrix}}_{n_r \text{ col.}} \quad \underbrace{\begin{matrix} 0 & \dots & 0 \\ 0 & \dots & 0 \\ 0 & \dots & 0 \\ 1 & \dots & 1 \end{matrix}}_{n_c \text{ col.}}$$

Suite à des essais de simulation, Ming-Da Ma *et al.* rapportent que l'application de cet algorithme au contrôle de la gravure des tranchées d'isolation (*Shallow Trench Isolation*) aurait une valeur ajoutée certaine par rapport aux traditionnels contrôleurs indépendants, surtout pour les produits à faible volume. Ils ont aussi mis en exergue le rôle positif potentiel de cet algorithme dans le monitoring du conditionnement des équipements.

Ming-Da Ma *et al.* mettent toutefois en garde contre la **non-persistance de l'excitation**[6] du système, une seconde caractéristique inhérente à la fabrication de composants microélectroniques multi-produits [Wang et al. (2008)]. Du fait du nombre important des produits, il arrive que le volume d'un produit quelconque s'estompe puis s'arrête. Aussi, un équipement peut être à l'arrêt pendant une longue période (maintenance, déménagement, etc). Dans ces cas, il n'est pas garanti d'avoir une matrice $\mathbf{H}^{\dagger}$ de rang complet. L'estimation du vecteur $\mathbf{\Theta}$ pourrait être par conséquent biaisée. Dans les travaux axés sur l'identification récursive, la non-persistance de l'excitation a souvent été l'origine du développement mathématique de nouveaux algorithmes d'estimation récursive. Nous allons nous y intéresser dans les prochains paragraphes.

Partant du constat que seul l'état global z_k intervient dans le calcul de la recette par le contrôleur (R2R), Hanish [Hanish (2006)] et Wang *et al.* [Wang et al. (2008)] ont choisi d'estimer, non pas les paramètres originaux du contexte, mais des combinaisons linéaires de ceux-ci. Le nombre de combinaisons est bien entendu égal au rang de la matrice $\mathbf{H}$ et la dimension du vecteur inconnu $\mathbf{\Theta}$ est alors réduit de m (le nombre des catégories). La transformation suggérée par Wang *et al.* et exposée ci dessous sera reprise dans nos essais de simulations.

$$\begin{cases} y_{reference} = \mu + s_{sc_1} + s_{l_1} + s_{c_1} + s_{r_1} & \\ y_{s_i} = s_{sc_i} - s_{sc_1} & i = 2 \cdots n_s \\ y_{l_i} = s_{l_i} - s_{l_1} & i = 2 \cdots n_l \\ y_{r_i} = s_{r_i} - s_{r_1} & i = 2 \cdots n_r \\ y_{c_i} = s_{c_i} - s_{c_1} & i = 2 \cdots n_c \end{cases} \tag{5.12}$$

6. Voir paragraphe 5.3.2.

Où s_{l_j} est le biais relatif au $j^{ème}$ élément de la catégorie des *layers*, idem pour les autres catégories. Le nombre de paramètres inconnus est ainsi réduit de $m = 4$: une estimation non biaisée des variables y est désormais possible. Quelle que soit la stratégie mise en oeuvre pour aller au delà de la déficience du rang de la matrice des régresseurs **H**, elle est nécessairement couplée à un estimateur récursif.

5.3 L'identification récursive

L'identification récursive permet d'estimer le système en ligne au fur et à mesure que les mesures sont réalisées. Plusieurs algorithmes de base existent [Hunt (1986)].

5.3.1 Le filtre de Kalman

Le filtre de Kalman est une méthode d'estimation récursive classique dévéloppée par Kalman [Kalman (1960)] et utilisée d'ores et déjà dans une gamme étendue d'applications, notamment les régulations run-to-run [Palmer et al. (1996), Chemali (2002)]. Une condition *sine qua non* à sa mise en place est une représentation d'état (équation 5.13).

$$\begin{cases} \mathbf{\Theta}_k &= \mathbf{G}\,\mathbf{\Theta}_{k-1} + \mathbf{\Omega}_k \\ z_k &= \mathbf{F}\,\mathbf{\Theta}_k + \varepsilon_k \end{cases} \tag{5.13}$$

où **G** est la matrice de dynamique ou de transition. **F** est la matrice d'observation. ε_k et $\mathbf{\Omega}_k$ représentent respectivement le bruit de la mesure et les incertitudes liées à la modélisation de l'évolution du vecteur d'état $\mathbf{\Theta}$ (non-linéarité négligée, paramètres influents non-pris en compte). De tous les estimateurs qui sont fonctions linéaires des observations, le filtre de Kalman est le plus optimal au sens de l'erreur quadratique moyenne (*Minimum Mean Square Error Estimator*). Si ε_k et $\mathbf{\Omega}_k$ suivent des lois normales, il est de surcroît le plus optimal au sens MSE de tous les estimateurs (linéaires ou pas)[Castillo (2002)].

Dans le cadre de ces travaux de thèse, le filtre de Kalman sera simulé et nous allons devoir nous soumettre à ce formalisme mathématique. Néanmoins, il n'est pas question de développer un modèle physique de la perturbation [Crisalle et al. (1992)], du fait de la grande complexité de la tâche et du manque de moyens (instrumentation, équipements, etc). Il n'est pas non plus question de développer un modèle de la perturbation de type ARIMA[7]. Pour le lecteur qui serait intéressé par cette dernière démarche, je citerai les travaux de Vanli *et al.* [Vanli et al. (2007), Vanli (2007)]. Partant du constat que les mesures sont autocorrélées, l'idée de Vanli *et al.* est alors de partager l'ensemble des catégories, sources qualitatives de variation, en deux sous-ensembles : un premier qui contribue davantage à l'autocorrélation des données et un second qui contribue plutôt au décalage (shift, offset) de la moyenne. Pour modéliser les variables du premier groupe, Vanli *et al.* utilisent les processus autorégressifs **AR**. L'algorithme ainsi conçu, est basé sur une approche itérative qui sélectionne les quelques variables et interactions à modéliser par un processus **AR** et

7. Il s'agit d'un formalisme de séries chronologiques, mis en place par Box et Jenkins [Box et al. (1994)]

en optimise les paramètres.

Par la suite, nous allons opté pour un modèle assez simple pour décrire l'évolution de $\mathbf{\Theta}$: il s'agit de la marche aléatoire (voir Equation 5.14), adoptée par ailleurs dans plusieurs réalisations de boucles *run-to run* [Bode (2001), Mullins et al. (1997)].

$$\Theta_{k+1} = \Theta_k + \Omega_{k+1} \tag{5.14}$$

La marche aléatoire est à l'image de certaines perturbations courantes dans l'industrie de la microélectronique, notamment les décalages ou *shifts* : il suffit d'avoir une perturbation Ω dont la distribution est en forme de Dirac [MacGregor et al. (1984)]. Elle présente aussi un atout non négligeable : elle ne nécessite aucune démarche d'identification. Le modèle qui sera à la base de nos simulations dans les prochains paragraphes est finalement le suivant, sous une forme de representation d'état.

$$\begin{cases} \mathbf{\Theta}_{k+1} & = \mathbf{\Theta}_k + \mathbf{\Omega}_{k+1} \\ z_k & = \mathbf{H}_k \mathbf{\Theta}_k + \varepsilon_k \end{cases} \tag{5.15}$$

avec $\mathbf{H}_k = \begin{bmatrix} 1 & \mathbf{T}_k \end{bmatrix}$. Nous supposons que ε et $\mathbf{\Omega}$ suivent des lois normales : $\varepsilon \sim \mathcal{N}(0; R)$ et $\mathbf{\Omega} \sim \mathcal{N}(0; \mathbf{Q})$. L'algorithme d'estimation est alors décrit par les formules suivantes :

$$\begin{cases} \mathbf{K}_k & = \frac{\mathbf{P}_{k-1}\mathbf{H}_k'}{R + \mathbf{H}_k \mathbf{P}_{k-1} \mathbf{H}_k'} \\ \hat{\mathbf{\Theta}}_k & = \hat{\mathbf{\Theta}}_{k-1} + \mathbf{K}_k \left(z_k - \mathbf{H}_k \hat{\mathbf{\Theta}}_{k-1} \right) \\ \mathbf{P}_k & = \mathbf{P}_{k-1} - \mathbf{K}_k \mathbf{H}_k \mathbf{P}_{k-1} + \mathbf{Q} \end{cases} \tag{5.16}$$

$\mathbf{P}_k$ est la matrice de variance-covariance de $\hat{\mathbf{\Theta}}_k$ à un facteur multiplicateur près (voir le paragraphe 5.3.2). Nous constatons que l'application du filtre de Kalman sous-entend connaître R et $\mathbf{Q}$. Ceci n'est pas le cas du procédé lithographique, en raison simplement de sa **non-stationnarité** (décalages, dérives, etc). Le filtre de Kalman sera en effet sous-optimal, et nous nous devons de tester plusieurs couples (R,$\mathbf{Q}$) pour optimiser au mieux sa performance.

5.3.2 Les moindres carrés récursifs [Castillo (2002)]

Nous nous proposons de rappeler brièvement les étapes de développement de cette méthode, si répandue et célèbre. Nous reprenons l'équation du modèle 5.17 :

$$z_k = \mathbf{H}_k \mathbf{\Theta}_k + \varepsilon_k \tag{5.17}$$

avec $\varepsilon \sim \mathcal{N}(0; R)$. L'essentiel de cette technique est de déterminer une estimation du vecteur $\mathbf{\Theta}_k$ à l'instant k de sorte à minimiser la somme des carrés des écarts entre les z_i et leurs prévisions $\hat{z}_i = \mathbf{H}_i \hat{\mathbf{\Theta}}_i$, ceci sur un horizon de k mesures.

$$\min_{\mathbf{\Theta}} SS(\mathbf{\Theta}) = \min_{\mathbf{\Theta}} (\mathbf{Z} - \mathbf{H}\,\mathbf{\Theta}_k)' (\mathbf{Z} - \mathbf{H}\,\mathbf{\Theta}_k) \tag{5.18}$$

$$= \min_{\mathbf{\Theta}} \frac{1}{k} \sum_{j=1}^{k} (z_j - \mathbf{H}_j\,\mathbf{\Theta}_j)^2 \tag{5.19}$$

Le critère étant quadratique en Θ, le vecteur $\hat{\Theta}_t$ qui le minimise peut être exprimé de manière analytique.

$$\hat{\Theta}_k = [\sum_{j=1}^{k} \mathbf{H}_j' \mathbf{H}_j]^{-1} \sum_{j=1}^{k} \mathbf{H}_j' z_j \tag{5.20}$$

En introduisant la matrice de précision $\mathbf{P}_k$, égale à un facteur multiplicateur près à la matrice de variance-covariance de $\hat{\Theta}_k$

$$\mathbf{P}_k = [\sum_{j=1}^{k} \mathbf{H}_j' \mathbf{H}_j]^{-1} = \frac{Var(\hat{\Theta}_k)}{R} \tag{5.21}$$

et après quelques manipulations de l'équation 5.20, la forme récursive de l'algorithme en est déduite [Castillo (2002), Hunt (1986)]. Nous l'avons retranscrite ci-dessous [Castillo et Hurwitz (1997)]. Dans la littérature scientifique, on parle d'algorithme **RLS** (*Recursive Least Square*).

$$\begin{cases} \mathbf{K}_k &= \frac{\mathbf{P}_{k-1}\mathbf{H}_k'}{1+\mathbf{H}_k\mathbf{P}_{k-1}\mathbf{H}_k'} \\ \hat{\Theta}_k &= \hat{\Theta}_{k-1} + \mathbf{K}_k\,(z_k - \mathbf{H}_k\,\hat{\Theta}_{k-1}) \\ \mathbf{P}_k &= \mathbf{P}_{k-1} - \mathbf{K}_k\,\mathbf{H}_k\,\mathbf{P}_{k-1} \end{cases} \tag{5.22}$$

Une condition nécessaire pour garantir la convergence (absence de biais) de l'algorithme, valable d'ailleurs pour tout estimateur récursif, est **la persistance de l'excitation**. Le vecteur $\mathbf{H}_k$ doit varier suffisamment pour que des informations relatives aux paramètres du système puissent en être déduites. L'exemple d'une commande constante illustre bien cette notion. Dans ce cas, le gain ou la pente β est dit non-identifiable. Ce phénomène est défini de la façon suivante [Castillo (2002)] :

Définition 1 *Un vecteur $\mathbf{H}_k$ de taille n est dit **persistently exciting** d'ordre n si la matrice*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{H}_k' \mathbf{H}_k = \lim_{N\to\infty} \frac{1}{N} \mathbf{P}_N^{-1} \tag{5.23}$$

est définie positive.

La persistance de l'excitation, ainsi définie, implique que pour $t \to \infty$, la matrice $\mathbf{P}_t$ existe et elle est inversible. En s'appuyant sur la relation 5.21, nous pouvons alors conclure que la persistance de l'excitation est une condition nécessaire pour avoir une estimation fiable, ie : une variance majorée.

Par ailleurs, l'examen de l'algorithme des moindres carrés récursifs nous conduit au constat qu'il s'agit simplement d'un cas particulier du filtre de Kalman. Il suffirait de définir comme représentation d'état :

$$\begin{cases} \Theta_k &= \Theta_{k-1} \\ z_k &= \mathbf{H}_k\,\Theta_k + \varepsilon_k \end{cases} \tag{5.24}$$

L'algorithme RLS est en effet très approprié pour l'estimation de paramètres inconnus **constants** car la matrice $\mathbf{P}_k$, proportionnelle à la matrice de variance-covariance du

vecteur $\mathbf{\Theta}_k$, décroît dans le temps. L'étude de l'équation 5.22 où le terme correctif du second membre est toujours positif ou nul nous montre que $\mathbf{P}_k$ ne peut que décroître au fur et à mesure des réalisations. L'algorithme donne alors de moins en moins de poids aux nouvelles mesures et sa capacité à la poursuite des paramètres variables dans le temps se dégrade [Shah et Cluett (1991)].

Une variante de l'algorithme RLS, intitulée λ-RLS, permet de pallier à cette situation. Elle consiste à introduire une pondération de plus en plus faible sur les anciennes observations et un poids important aux plus récentes. Le critère des moindres carrés à facteur d'oubli consiste à minimiser :

$$\min_{\mathbf{\Theta}} SS(\mathbf{\Theta}) = \min_{\mathbf{\Theta}} \frac{1}{k} \sum_{j=1}^{k} \lambda^{k-j} \left(z_j - \mathbf{H}_j \mathbf{\Theta}_j\right)^2 \tag{5.25}$$

où λ est compris entre 0.95 et 0.99. Le choix de λ résulte d'un compromis entre une poursuite fidèle des paramètres (bonne précision) et une faible fluctuation statistique. Pour un λ proche de 1, la fenêtre d'observation ou encore l'horizon équivalent [Parkum et al. (1990)], défini par l'équation 5.26, est long et l'algorithme aura des difficultés à s'adapter fidèlement aux variations des paramètres. La variance de $\hat{\mathbf{\Theta}}$, fonction décroissante de λ, demeure en revanche faible. A l'inverse, pour un λ très inférieur à 1, la fenêtre d'observation est courte et l'algorithme poursuivra plus facilement l'évolution des paramètres. En contrepartie, la sensibilité au bruit de l'estimateur augmente et sa variance se dégrade. Sous une forme récursive, l'algorithme λ-RLS est retranscrit ci dessous (voir équations 5.27).

$$N = \sum_{i=0}^{\infty} \lambda^i = \frac{1}{1-\lambda} \tag{5.26}$$

Parmi les lacunes de cet algorithme, figure en premier ce qu'on appelle l'inflation de la covariance ou *covariance windup*. Elle a lieu lorsque les données fournies à l'estimateur ne sont pas suffisamment riches[8] : $\mathbf{P}_{k-1}\mathbf{H}_t' \approx 0$. Oublier les anciennes observations en faveur des nouvelles engendre dans ce cas une augmentation exponentielle de la matrice de variance $\mathbf{P}_t \approx \mathbf{P}_{t-1}/\lambda$; la moindre déviation du vecteur z_k sera alors suivi d'un fort ajustement des paramètres et l'algorithme est désormais très vulnérable au bruit de mesure. A partir de là, plusieurs variantes de l'algorithme λ-RLS ont été proposées et étudiées. Nous pouvons citer la technique des moindres carrés récursifs au facteur d'oubli variable, ou encore celle où la trace de la matrice $\mathbf{P}_k$ est gardée constante. Del Castillo *et al.* se sont servi de ce dernier pour concevoir un contrôleur auto-ajustable [Castillo et Hurwitz (1997)]. Wang *et al.* [Wang et al. (2008)] recommandent plutôt la technique développée par Salgado *et al.* [Salgado et al. (1988)]. Un aperçu de divers algorithmes issus des moindres carrés est donné par [Shah et Cluett (1991)].

$$\begin{cases} \mathbf{K}_k &= \frac{\mathbf{P}_{k-1}\mathbf{H}_k'}{\lambda + \mathbf{H}_k \mathbf{P}_{k-1} \mathbf{H}_k'} \\ \hat{\mathbf{\Theta}}_k &= \hat{\mathbf{\Theta}}_{k-1} + \mathbf{K}_k \left(z_k - \mathbf{H}_k \hat{\mathbf{\Theta}}_{k-1}\right) \\ \mathbf{P}_k &= \frac{1}{\lambda}\left(\mathbf{P}_{k-1} - \mathbf{K}_k \mathbf{H}_k \mathbf{P}_{k-1}\right) \end{cases} \tag{5.27}$$

Une autre cause de fragilité du λ-RLS est l'utilisation d'un unique facteur d'oubli pour l'ensemble des variables θ_i[9]. Ceci est préoccupant dans des situations où

8. Dans le sens da la définition 1
9. Un oubli uniforme dans l'espace

certains θ_i varient plus rapidement que d'autres. Dans ce cas, si nous choisissons λ en fonction des paramètres les plus rapides, ceci induirait une mauvaise précision pour les plus lents. A l'inverse, si λ est ajusté pour mieux identifier les plus lents, les variations rapides des autres θ_i ne pourront pas être suivies. L'atelier de photolithographie nous offre un exemple concret : Si les biais relatifs aux réticules sont constants, du fait que les lignes de chrome demeurent inchangées, l'équipement de fabrication, lui, présente un biais variable dans le temps (maintenance, dérive de température, vieillissement de la chambre de laser, etc). Des modifications de l'algorithme λ-RLS sont alors nées sous le nom de techniques d'oubli directionnel ou sélectif [Saelid et Foss (1983),Parkum et al. (1990)] : elles attribuent un facteur d'oubli différent à chaque paramètre. La plupart des pionniers à l'origine de ces nouvelles techniques ont utilisé des facteurs d'oubli variables dans le temps, et ceci afin de limiter l'augmentation de la variance lorsque le système n'est pas suffisamment excité dans une ou plusieurs directions.

L'algorithme de cette nouvelle technique qui intègre des facteurs d'oubli multiples est définie par le système d'équations suivant :

$$\begin{cases} \mathbf{K}_k &= \frac{\mathbf{P}_{k-1}\,\mathbf{H}'_k}{1+\mathbf{H}_k\mathbf{P}_{k-1}\mathbf{H}'_k} \\ \hat{\mathbf{\Theta}}_k &= \hat{\mathbf{\Theta}}_{k-1} + \mathbf{K}_k\,(z_k - \mathbf{H}_k\,\hat{\mathbf{\Theta}}_{k-1}) \\ \mathbf{P}_k &= \mathbf{\Lambda}^{-1}\,(\mathbf{P}_{k-1} - \mathbf{K}_k\,\mathbf{H}_k\,\mathbf{P}_{k-1})\,\mathbf{\Lambda}^{-1} \end{cases} \tag{5.28}$$

où $\mathbf{\Lambda} = diag[\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_n}]$. Les travaux de Yoshitani [Yoshitani et Hasegawa (1998)] et Oda [Oda K. et M. (1991)] en sont des exemples d'application.

Plus récemment, d'autres algorithmes ont été conçus pour répondre à ce même besoin : celui de gérer la différence en vitesse de changement de plusieurs paramètres inconnus [Cao et Schwartz (1999),Vahidi et al. (2003)]. Ces algorithmes possèdent une structure différente et requièrent à partir de là des modifications importantes du simulateur (voir section 5.4). En outre, les premières simulations que nous avons réalisées ont montré des performances équivalentes à celles de l'algorithme 5.28. A partir de là, nous avons choisi de nous concentrer sur l'emploi de l'algorithme 5.28 seulement, testant son potentiel pour une éventuelle mise en place dans un environnement de production de puces micro-électroniques. Il est important de signaler que tous les choix que nous faisons répondent aussi à une volonté de simplifier au mieux la conception de l'observateur. En outre des contraintes de temps, nous avons tenu à quantifier le potentiel d'une première version avant d'engager un travail d'amélioration à travers une nouvelle thèse par exemple.

5.3.3 JADE (Just-in-time Adapative Disturbance Estimation)

Dans les travaux de Firth *et al.* [Firth et al. (2006)], **JADE** est conçu tel qu'on suppose toujours connu et fixe un échantillon de données. Wang *et al.* ont repris cet algorithme et en ont développé une version récursive [Wang et al. (2008)].

$$\begin{cases} \mathbf{K} &= \frac{\mathbf{Q}_4^{-1}\,\mathbf{H}'_q}{\mathbf{Q}_1^{-1}+\mathbf{H}_q\mathbf{Q}_4^{-1}\mathbf{H}'_q} \\ \hat{\mathbf{\Theta}}_k &= \hat{\mathbf{\Theta}}_{k-1} + \mathbf{K}\,(Z_q - \mathbf{H}_q\,\hat{\mathbf{\Theta}}_{k-1}) \end{cases} \tag{5.29}$$

Les matrices $\mathbf{H}_q$, $\mathbf{Z}_q$ sont définies dans le paragraphe 5.2. $\mathbf{Q}_1$ est une matrice diagonale

de dimension $q \times q$ contenant les facteurs d'oubli pour les q dernières réalisations. $\mathbf{Q}_4$ est une matrice diagonale de dimension $n \times n$ ($n = dim(\mathbf{\Theta})$ est le nombre des paramètres à estimer). En comparant l'expression du gain (équation 5.29) à celle du filtre de Kalman [Christopher A. Bode et Edgar (2007)], Wang *et al.* ont constaté une équivalence entre les deux algorithmes lorsque seule la dernière mesure est prise en compte et

$$Q_1^{-1} = R \text{ et } Q_4^{-1} = \mathbf{P}$$

Du fait que la matrice de variance-covariance P est constante, la vitesse d'estimation ne décroît pas et l'algorithme **JADE** serait à même de détecter toute variation des variables θ_i (dérive d'outils de fabrication, décalage d'un équipement de métrologie, etc). Remettre la matrice P à sa valeur initiale à chaque nouvelle mesure la dépossède néanmoins de toute information concernant les propriétés statistiques de la perturbation, contenue dans les observations précédentes. De surcroît, une matrice P constante rendrait **JADE** plus vulnérable au bruit de la mesure. Par la suite, nous allons restreindre les simulations aux seuls cas des moindres carrés et du filtre de Kalman.

5.4 Simulation

Nous allons simuler le fonctionnement de différents algorithmes d'estimation récursive et caractériser leurs performances en rapport avec leur capacité à une poursuite fidèle (convergence, biais), leur fluctuation statistique (variance) et leur rapidité à converger. A partir de ces simulations, nous allons pouvoir désigner l'estimateur qui serait le plus à même à bien fonctionner dans un environnement de production de puces microélectroniques. A travers les hypothèses de départ, nous veillerons à décrire d'une façon la plus fidèle possible l'environnement de fabrication du niveau d'exposition des Vias 1 en photolithographie. Le choix de ce niveau est lié à un fort taux de recyclage [10], et un grand nombre d'équipements qualifiés. Les hypothèses de départ sont les suivantes :

⋈ Les différentes catégories impliquées sont : 5 équipements d'exposition $\mathbf{sc}_{k,\ k=1,2,3,4,5}$, 5 équipements de métrologie $\mathbf{c}_{i,\ i=1,2,3,4,5}$ et 14 réticules $\mathbf{r}_{j,\ j=1,\cdots,14}$.

⋈ Encore une fois, par souci de représentation de la production, les volumes des différents produits varieront du petit (1.5% ie : 13 lots) au grand (51%, ie : 553 lots).

⋈ Nous disposions en production d'une valeur établie et fiable de la pente β (ou le gain), souvent estimé à partir d'un plan d'expérience. La qualité de cette estimation sera prise en compte à travers l'indice $\varrho = \hat{\beta}/\beta$. Nous allons nous intéresser dans un premier temps au cas idéal où ϱ est égal à 1.

Comme le montre la figure 5.3, le simulateur comprend trois blocs principaux

1. Un bloc définissant les différents biais s et leurs profils (évolution dans le temps). Par hypothèse, seuls les équipements, de métrologie compris, pourraient subir des dérives et des décalages. Quant aux réticules, ils sont supposés avoir des biais constants.

10. Dans le milieu de la fabrication de semiconducteurs, nous utilisons le terme anglais *rework*. Cette opération, qui consiste à retravailler le lot, est possible dans la mesure où les matériaux déposés en lithographie sont organiques et solubles dans des solvants. Leur élimination est alors aisée.

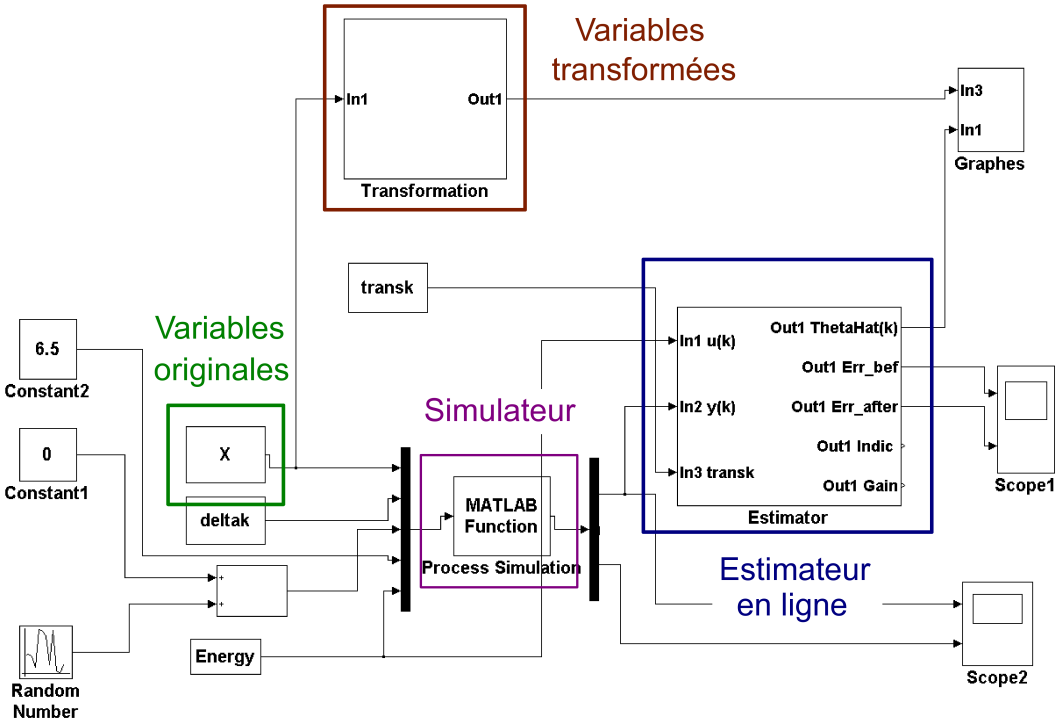

FIGURE 5.3 – *Schéma Simulink du simulateur de l'atelier de photolithographie. Le simulateur est en charge de fournir les mesures CD_k, en se basant sur les profils temporels des biais originaux, une séquence d'énergie et un contexte de fabrication qui lui sont fournis. L'estimateur en ligne a pour mission d'identifier les profils des variables transformées, déduites des variables originales grâce au bloc transformation.*

2. Un **bloc Simulation** basé sur le modèle 5.5. Il est en charge de fournir une séquence de mesures CD_k à partir d'une séquence d'entrées multiples : une variable aléatoire $\varepsilon_k \sim \mathcal{N}(0; \sigma_\varepsilon^2)$ représentative du bruit de mesure , la commande u_i, le gain β, la constante μ et les différents biais s relatifs aux états des différents éléments.
3. Un **bloc Estimation** qui aura comme tâche de retrouver à partir des entrées (u_i, CD_i) les différentes variables transformées définies ci-dessous.

$$\begin{cases} y_{reference} = \mu + s_{sc_{ref}} + s_{c_{ref}} + s_{r_{ref}} & \\ y_{sc}^k = s_{sc_k} - s_{sc_{ref}} & k \neq ref \text{ and } k \in \{1,2,3,4,5\} \\ y_c^i = s_{c_i} - s_{c_{ref}} & i \neq ref \text{ and } i \in \{1,2,3,4,5\} \\ y_r^j = s_{r_j} - s_{r_{ref}} & j \neq ref \text{ and } j \in \{1,2,3,4,5\} \end{cases} \tag{5.30}$$

$y_{reference}$ est une variable de référence. Elle intègre les effets d'éléments choisis au sein de leurs catégories comme éléments de référence. Plus loin dans ce chapitre, dans le cadre des simulations des données de production, ces éléments seront ceux les plus utilisés de toute leur catégorie. Une bonne précision de l'estimation de ces éléments de référence est en effet primordiale afin d'espérer des estimations correctes des variables transformées restantes y_{sc}, y_c et y_r.

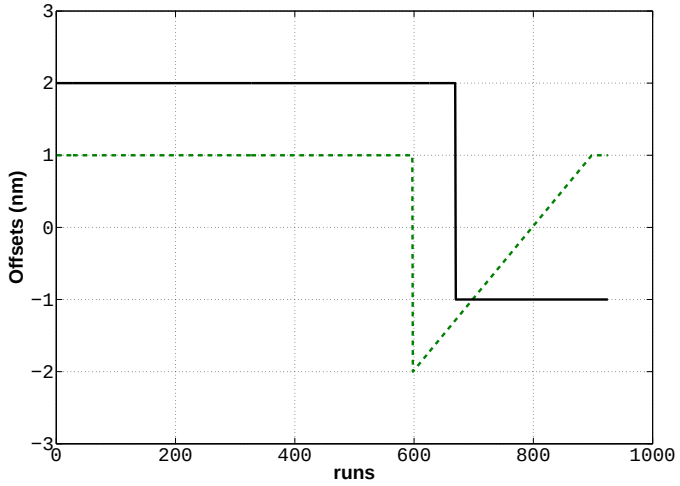

FIGURE 5.4 – *Profils de l'évolution des biais relatifs à un équipement d'exposition (vert) et à un équipement de métrologie (noir).*

Choix de $P(0)$ et $\Theta(0)$ [Castillo (2002), Falinower (2008)] La matrice P_k est inversible à partir de l'instant $k = \dim(\Theta) = n$. Alors, rigoureusement parlant, l'algorithme d'estimation récursive devrait être démarré à cet instant, pour que cette version récursive soit équivalente à l'algorithme non-récursif. En pratique, et en absence de toute information à priori, on démarre à $k = 0$ en posant :

$$\begin{cases} P(0) & = \delta^{-1}\,\mathbf{I},\ 0 < \delta \ll 1 \\ \Theta(0) & = \mathbf{0} \end{cases} \tag{5.31}$$

$\Theta(0)$ est alors interprété comme le vecteur des estimations à priori des paramètres inconnus, et $P(0)$ leur matrice de variance-covariance associée. Si nous avons davantage confiance à un autre Θ, δ pourrait être accru en conséquence.

Choix des états individuels Θ_i Comme nous l'avons indiqué en début de ce paragraphe, seuls les équipements pourraient subir des perturbations non stationnaires de type dérive ou décalage (voir figure 5.4). Les réticules auront des biais constants.

Choix de la commande $\mathbf{u}_k$ Nous allons utiliser une séquence de u_k, enregistrée en production. Ce choix va s'avérer judicieux dans le cas d'un indice ρ différent de l'unité. Nous y reviendrons dans la section 5.4.4.

5.4.1 Les Moindres Carrés Récursifs : résultats

Dans ce qui suit, l'algorithme **SF** désignera l'algorithme des moindres carrés à oubli sélectif qui a la capacité d'affecter un facteur d'oubli λ_i différent à chaque variable y_i. Pour davantage de simplicité, nous avons choisi d'avoir un facteur λ_c commun à la catégorie des équipements de métrologie et un second λ_{sc} commun à la catégorie des équipements d'exposition. D'une façon intuitive, les λ_i relatifs aux réticules sont choisis égaux à l'unité, tandis que celui associé à la variable $y_{reference}$ sera égal au $\min\{\lambda_c, \lambda_{sc}\}$. La variable $y_{reference}$ intègre en effet les biais d'un équipement

de métrologie de référence et un équipement d'exposition de référence. L'élément qui évolue le plus rapidement parmi ces deux équipements impose la vitesse de ses variations à la variable de référence.

La figure 5.5 montre l'estimation de $y_{reference}$ par trois variantes des moindres carrés, à savoir l'algorithme **RLS** classique, le λ-**RLS** ($\lambda = 0.99$) et le **SF** ($\lambda_{sc} = \lambda_c = 0.99$). L'objectif de ces premières simulations est en premier lieu une appréciation approximative des performances des différents algorithmes (convergence, rapidité, fluctuations). Par conséquent, le choix des valeurs numériques des λ_i n'a pas été optimisé. Tandis que la version originale des moindres carrés (**RLS**) a des difficultés à converger, comme attendu, les deux autres variantes convergent vers une estimation non biaisée de $y_{reference}$ et montrent des performances équivalentes en terme de capacité de poursuite et de fluctuation statistique.

La figure 5.6 décèle cependant une performance du λ-**RLS** beaucoup moins satisfaisante, dès qu'il s'agit de variable constante, telle les biais relatifs aux réticules. La variance de l'estimateur est alors importante. Deux facteurs en sont responsables. Tout d'abord, avec un facteur d'oubli égal à 0.99, l'estimateur n'incorpore pas l'information *à priori* sur l'invariabilité du paramètre à estimer. Il pondère les observations et il est de ce fait plus sensible au bruit de la mesure. D'autre part, la figure 5.6 illustre le cas d'un produit à très faible volume (12 lots), où la condition de la persistance de l'excitation (voir définition 1) n'est plus satisfaite. Nous savons que l'algorithme λ-**RLS** souffrirait dans ce cas de l'inflation de covariance. La figure 5.7 en donne une illustration. Nous constatons que la variance de la variable y_r croît d'une façon exponentielle entre deux utilisations du réticule correspondant, ce qui explique la réaction trop impulsive de l'estimateur. Au contraire, l'estimateur **RLS** et l'algorithme **SF** où les facteurs d'oubli relatifs aux réticules sont unitaires présentent une variance faible qui ne cesse de décroître (figure 5.7, courbes en vert et en bleue) au fur et à mesure des réalisations.

Il faudrait souligner toutefois une contre-performance de l'ensemble de ces variantes, qui font évoluer l'estimation $\hat{y}_r$, bien que le réticule ne soit pas utilisé. Ces algorithmes dérivés des moindres carrés corrigent l'estimation $\hat{y}_r$ à chaque nouvelle observation, et ceci d'une façon d'autant plus importante que l'erreur à priori $|z_k - H_k \hat{Theta}_{k-1}|$ est importante et que le nombre de runs passés, où le réticule est utilisé, est petit.

La figure 5.8 présente les résultats de la simulation dans le cas d'un produit à fort volume (131 lots), où l'estimateur serait préservé de la non-persistance de l'excitation. Nous y constatons une variance très importante du λ-**RLS**, en rapport avec une sensibilité accrue au bruit. Le **SF** se distingue encore une fois par une faible fluctuation, nettement moins importante que celle de l'algorithme **RLS** notamment. Les figures (en fin de chapitre) 5.32, 5.33, 5.34 et 5.35 sont une sélection aléatoire de résultats relatifs à l'estimation de certaines variables (équipements et réticules). Les remarques précédentes en ce qui concerne les différents algorithmes dérivés des moindres carrés demeurent valables :

⋈ L'algorithme **RLS** est clairement inadapté à un environnement de production sujet à des perturbations.

⋈ L'algorithme λ-**RLS**, avec un facteur d'oubli modéré (pas très agressif) de 0.99,

souffre de grandes fluctuations dans le cas de variables constantes, d'autant plus importantes qu'il s'agit de produits à faible volume. Augmenter davantage le facteur d'oubli compromettra sa performance dans le cas de variables sujettes à des dérives et à des décalages, notamment les équipements d'exposition et de métrologie.

$\rtimes$ L'algorithme **SF** est l'algorithme du compromis. A partir de ces premières simulations, il serait le plus adapté à notre contexte.

5.4.2 Le Filtre de Kalman : Résultats

Le filtre de Kalman possède deux paramètres que nous devons évaluer au préalable, R et **Q**. Dans notre cas, R est la variance du bruit de mesure et **Q** correspond à la matrice de covariance du bruit de process. **Q**, par définition, renseigne l'estimateur sur les variations du vecteur d'état $\boldsymbol{\Theta}$. Elle est naturellement plus difficile à déterminer car l'état du système est non-stationnaire[11]. A l'inverse, une estimation du scalaire R pourrait être obtenue à partir d'une étude de reproductibilité et de répétabilité[12] (**R&R**). Nous nous plaçons dans un contexte favorable idyllique et nous supposons connaître la valeur exacte de la variance R, ie : R sera égal à σ_ε^2. Afin d'approcher une performance satisfaisante du filtre de Kalman, nous avons réalisé des simulations avec différentes valeurs de la matrice **Q**. Généralement, cette matrice est choisie proportionnelle à la matrice identité **I** : $\mathbf{Q} \propto \mathbf{I}$. Pour mieux renseigner l'algorithme sur l'invariabilité de certaines variables y_i, nous avons opté pour un choix plus judicieux : $\mathbf{Q} \propto \mathbf{M}$, où **M** est exprimé ci-dessous (voir expression 5.32).

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & & & & \\ & 1 & & & \\ & & 0 & & 0 \\ & & & & 0 \\ 0 & \cdots & & 0 & 0 \end{pmatrix} \tag{5.32}$$

Le nombre des 1 sur la diagonale de la matrice est de 9. Rappelons que 9 correspond au nombre de variables y_i qui potentiellement varient dans le temps.

La figure 5.9 montre que la variance de l'estimateur est une fonction croissante du coefficient de proportionnalité entre la matrice **Q** et la matrice **M**, qu'on désignera par ϱ. Une forte valeur de ϱ rend compte en effet d'une forte variabilité du vecteur d'état que l'estimateur mettra à profit pour réduire l'erreur du modèle. L'erreur à posteriori est effectivement d'autant plus faible (figure 5.10), et les estimations d'autant plus bruitées que ϱ a des fortes valeurs. Une faible valeur de ϱ impliquerait par contre une faible fluctuation des paramètres. Elle pourrait dégrader néanmoins sa capacité de poursuite et être à l'origine d'un biais (voir figure 5.11, $\mathbf{Q} = \frac{R}{1000}\mathbf{M}$). Le graphe 5.12 qui correspond à l'estimation du biais d'un équipement d'exposition confirme

11. A priori, la matrice de covariance est alors infinie.

12. Il s'agit d'une analyse de variance dont l'objectif est de quantifier la contribution de plusieurs facteurs pouvant contribuer à la variabilité d'une mesure, et notamment la variabilité répétition (la précision de l'appareil de mesure), la variabilité temporelle, la variabilité humaine (l'effet opérateur) et la variabilité du processus de fabrication.

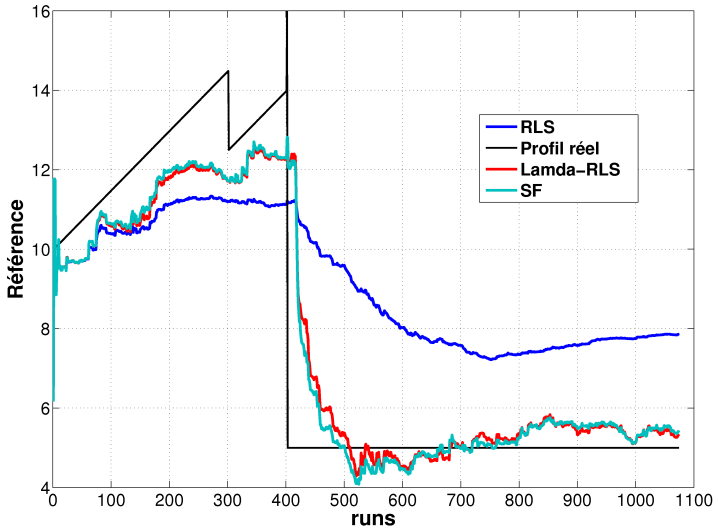

FIGURE 5.5 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'estimation de $y_{reference}$ pour trois variantes des moindres carrés :1/ Les moindres carrés classiques **RLS** en bleu, 2/l'algorithme λ-**RLS** avec $\lambda = 0.99$ en rouge et 3/ le **SF** ($\lambda_{sc} = \lambda_c = 0.99$) en vert.*

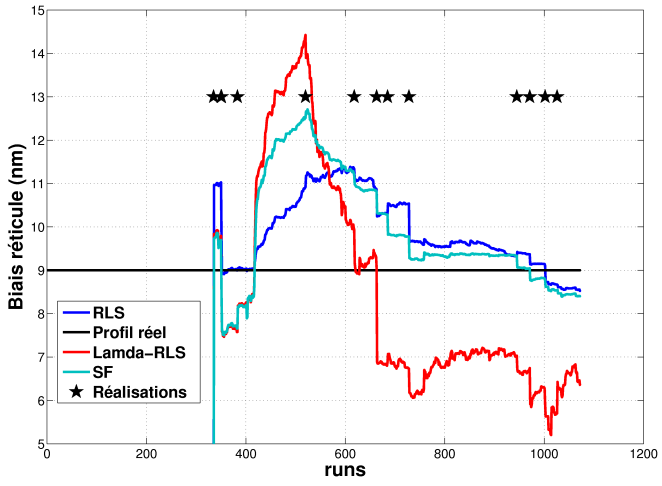

FIGURE 5.6 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'estimation du biais y_r, un réticule à faible volume, par trois variantes des moindres carrés :1/ Les moindres carrés classiques **RLS** en bleu, 2/l'algorithme λ-**RLS** avec $\lambda = 0.99$ en rouge et 3/ le **SF** ($\lambda_{sc} = \lambda_c = 0.99$) en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

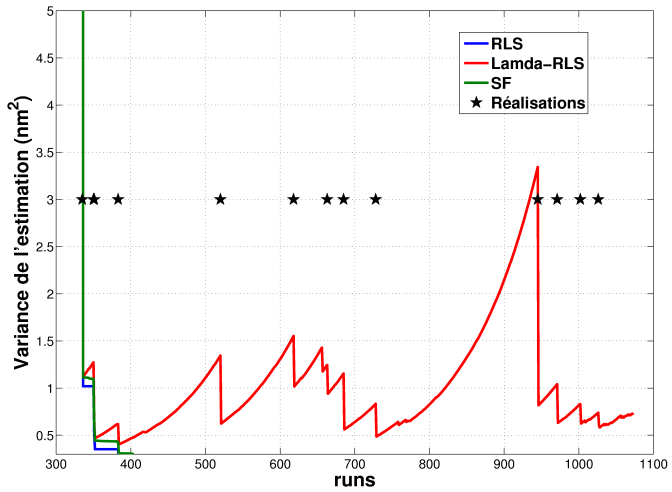

FIGURE 5.7 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de la variance de l'estimation du biais y_r, un réticule à faible volume, pour trois variantes des moindres carrés :1/ Les moindres carrés classiques **RLS** en bleu, 2/l'algorithme λ-**RLS** avec $\lambda = 0.99$ en rouge et 3/ le **SF** ($\lambda_{sc} = \lambda_c = 0.99$) en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

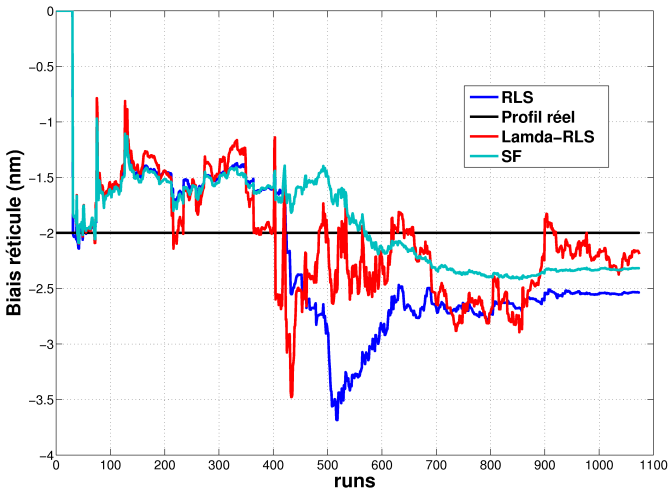

FIGURE 5.8 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'estimation du biais y_r, un réticule à fort volume, par trois variantes des moindres carrés :1/ Les moindres carrés classiques **RLS** en bleu, 2/l'algorithme λ-**RLS** avec $\lambda = 0.99$ en rouge et 3/ le **SF** ($\lambda_{sc} = \lambda_c = 0.99$) en vert.*

ces dernières remarques.

Les figures 5.13, 5.36 et 5.37 montrent l'évolution des estimations de variables constantes y_r, relatives à des réticules au volume important. Nous constatons une bonne précision du filtre de Kalman et ceci indépendamment de ϱ. Ceci n'est pas toujours le cas et notamment quand il s'agit de produits à faible volume et des réalisations bien étalées dans le temps. Les figures 5.14 et 5.38 illustrent effectivement ce cas. A noter que dans le cadre de ces simulations (et uniquement dans ce cadre), une valeur intermédiaire entre $\varrho = \frac{R}{100}$ et $\varrho = \frac{R}{1000}$ semble être un bon compromis entre une faible variance de l'estimateur et une bonne capacité de poursuite.

En guise de conclusion, Il apparaît d'une façon claire que les algorithmes **RLS** et λ-**RLS** sont inadaptés à l'environnement de production et notamment l'atelier de photolithographie. Le premier ne détient aucune capacité de poursuite et il est donc incapable d'estimer des variables qui évoluent dans le temps. Le second utilise un unique facteur d'oubli comme paramètre de réglage pour ajuster le compromis entre le biais et la variance. Ce compromis, nous l'avons vu, n'est pas satisfaisant. Nous allons nous intéresser de ce fait au couple Kalman/**SF**.

5.4.3 Comparaison des Moindres Carrés (SF) au filtre de Kalman

La figure 5.15 pointe du doigt un comportement propre aux moindres carrés : en présence de perturbations, l'algorithme corrige les estimations de variables y_i bien qu'elles soient associées à des éléments non utilisés. La figure 5.15 correspond au cas d'un équipement d'exposition, qui a été à l'arrêt pendant une longue période. Ceci n' a pas empêché l'estimateur **SF** de faire dériver $\hat{y}_{sc}$ de -27 nm vers -10 nm. Nous observons que ce comportement est d'autant plus accentué que :

- ⋊ Le nombre de réalisations passées, où l'équipement est utilisé, est petit,
- ⋊ l'erreur à priori ($z_k - H_k \hat{Theta}_{k-1}$) est importante (perturbations, bruit).

Les figures 5.16 (courbe en rouge) et 5.17 (courbe en vert) illustrent davantage cette propriété. Quant au filtre de Kalman, il fige l'estimation à l'instant k jusqu'à avoir une nouvelle réalisation où l'élément correspondant sera utilisé, d'où cette impression d'une estimation constante par morceaux (figure 5.38).

Cette différence entre les deux algorithmes donne un avantage au filtre de Kalman, notamment dans le cas d'un nombre faible de réalisations (produit à faible volume, équipement peu utilisé). En ce qui concerne le cas des variables avec un nombre important de réalisations, les performances des deux algorithmes en terme de capacité de poursuite et de fluctuation statistique sont globalement équivalentes (figures 5.18, 5.19 et 5.39). L'optimisation des facteurs d'oubli et de ϱ pourrait bien entendu améliorer davantage les résultats, mais ceci est valable pour l'un et pour l'autre.

En ce qui concerne les paramètres variables dans le temps tel le biais relatif à un équipement de métrologie (figure 5.20), nous notons toutefois une contre-performance valable aussi bien pour le filtre de Kalman que pour le **SF**. Il s'agit d'une convergence à cinétique très lente (de l'ordre de quelques dizaines de runs) au niveau des dérives et des décalages. Toute tentative d'améliorer la capabilité des estimateurs à coller aux profils réels, par le biais des facteurs d'oubli pour l'un et le choix de **Q** pour l'autre, serait aux dépens d'une sensibilité accrue au bruit de

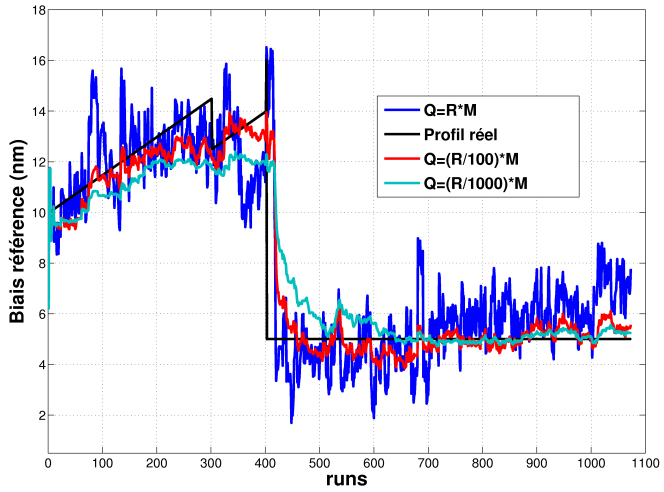

FIGURE 5.9 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_{ref} relatif à la référence pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en rouge 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en vert.*

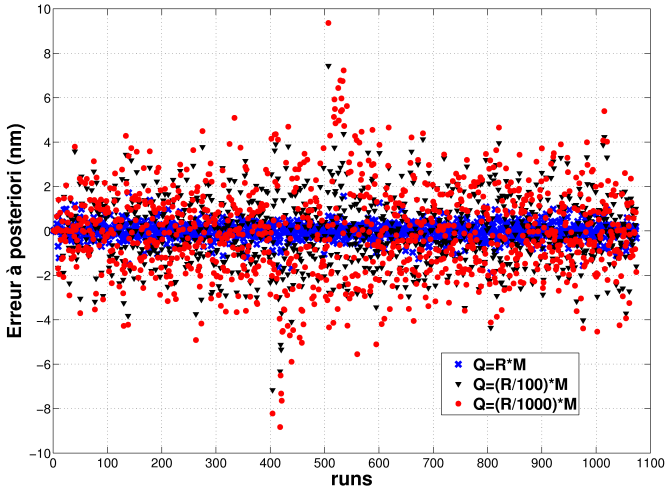

FIGURE 5.10 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'erreur à posteriori du modèle pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en noir 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en rouge.*

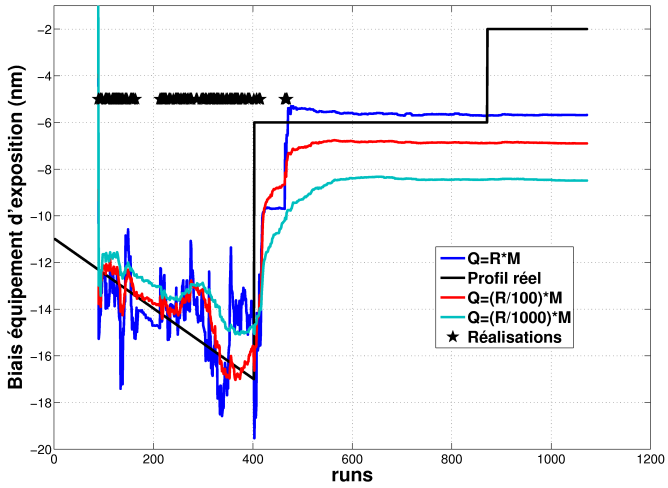

FIGURE 5.11 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_{sc} relatif à un équipement d'exposition donné pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu, 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en rouge, et 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

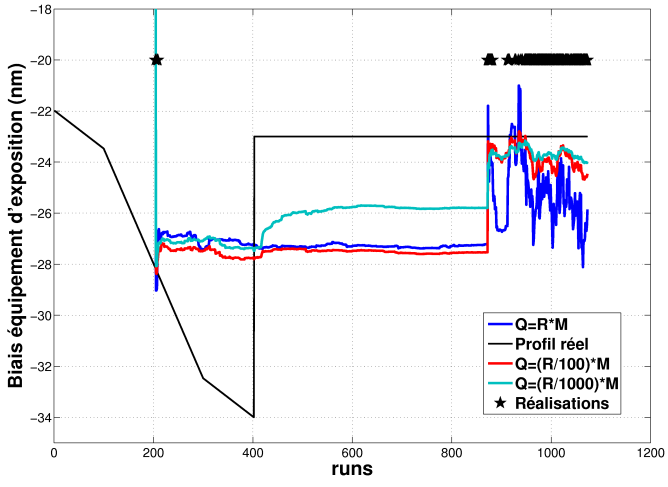

FIGURE 5.12 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_{sc} relatif à un équipement d'exposition donné pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu, 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en rouge, et 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

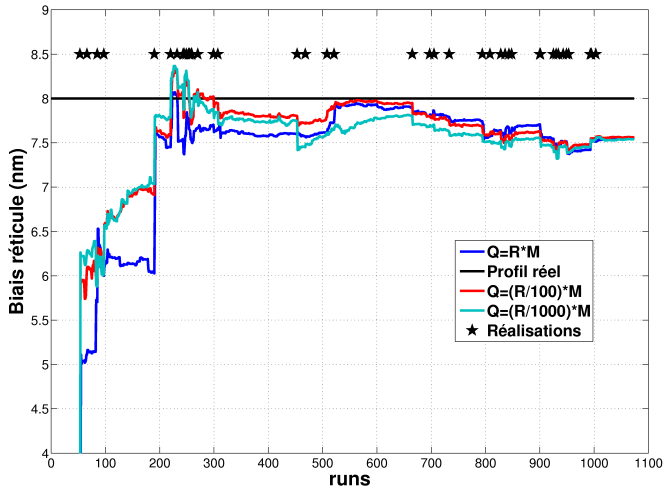

FIGURE 5.13 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_r relatif à un réticule donné pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu, 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en rouge, et 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

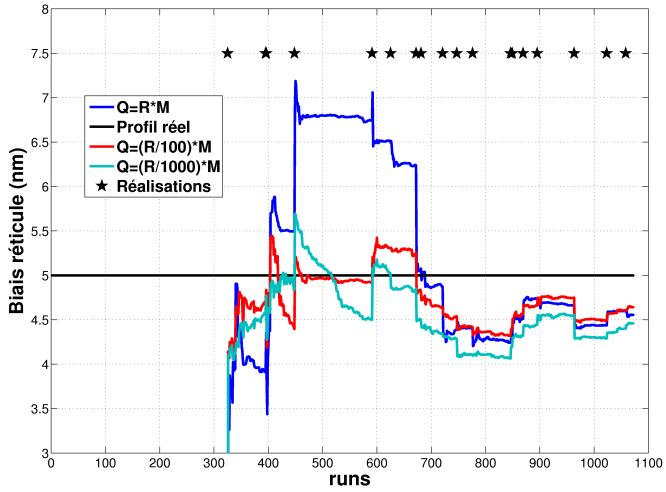

FIGURE 5.14 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_r relatif à un réticule donné pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu, 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en rouge, et 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

mesure et d'une mauvaise précision.

Pour pallier à cet inconvénient, Wang *et al.* recommandent d'utiliser un algorithme de détection de changements [Wang et al. (2008), Wang et He (2007)]. L'idée est en effet de préserver la précision de l'estimation et sa fonction de filtre en laissant évoluer la matrice de variance **P** comme dans le cas d'un processus aux paramètres constants. La matrice **P** ne pourra alors que décroître au fur et à mesure des observations garantissant une bonne précision. En parallèle, **un détecteur de changements** pourrait être mis en place pour repérer tout changement significatif des paramètres et réinitialiser à la même occasion la matrice **P** à une forte valeur. Cette opération permet aux estimations dans ce cas de réagir et converger rapidement au bout de quelques runs. Dans le cadre de nos travaux, étant donné la tâche très importante que ça représente [Basseville et Nikiforov (1993)], nous n'avons pas eu le temps de faire un état de l'art des différents algorithmes de détection et en incorporer un dans l'architecture de l'observateur.

5.4.4 Impact de ϱ

L'hypothèse simplificatrice qui stipule que ϱ soit égal à l'unité ne correspond pas à la réalité. Une longue liste de paramètres tels le bruit de la mesure, les conditions expérimentales (température, pression).. rendent l'obtention de la valeur exacte de la pente β impossible. Afin de saisir l'impact de notre méconnaissance de la pente sur l'estimation du vecteur d'état $\hat{\mathbf{\Theta}}$, nous avons réalisé des simulations avec un ϱ différent de 1. Comme le montre la figure 5.21, une valeur de ϱ aussi petite que 0.985 engendre une estimation biaisée du biais de référence. Les estimations des variables restantes demeurent, en revanche, inchangées. Ce résultat est valable aussi bien pour le filtre de Kalman que pour les moindres carrés.

Nous avons alors entrepris d'estimer la pente en ligne, au même titre que les biais y_i. L'utilisation du filtre de Kalman ($\mathbf{Q}=\frac{R}{100}*\mathbf{M}$) dans ce cas donne une estimation assez précise de la pente (voir figure 5.22). Le biais $\Delta\beta = \beta - \hat{\beta}$ est de l'ordre de 1% à 3% seulement. Même dans ce cas assez favorable, l'estimation $\hat{y}_{ref}$ demeure biaisée comme illustré en figure 5.23. Le graphe 5.24 révèle en effet des résidus $\Delta z_k = z_k - \hat{z}_k$ non centrés sur 0, tandis que la distribution de l'erreur globale ($\Delta CD = CD_k - \hat{CD}_k$) l'est.

Pour comprendre les origines de ce comportement double des résidus, nous avons transcrit ci dessous la relation qui relie les deux erreurs. Toute déviation de la pente $\Delta\beta$ est amplifiée par la commande u_k, dont les valeurs varient entre 20 et 30 mJ/cm^2. Au delà des premiers instants où $\hat{\beta}$ oscille autour de la valeur réelle, la déviation $\Delta\beta$ garde un signe constant (voir figure 5.22) et induit ainsi une distribution du terme $(\Delta\beta_k)\,u_k$ non centrée sur 0. Dans sa course à la minimisation de l'erreur globale, l'algorithme compense ce biais et engendre naturellement une estimation de z_k à biais non nul.

$$\Delta CD_k = (\Delta\beta_k)\,u_k + \Delta z_k \tag{5.33}$$

Par ailleurs, un examen attentif fait ressortir une influence importante de la séquence des commandes sur le biais d'estimation de la pente $\Delta\beta$. La figure 5.25 illustre ce constat. Nous y observons un biais de l'ordre de 1 à 3% dans le cas d'une séquence $\{u_k\}$ enregistrée en production. Le simple fait de changer l'ordre de réalisations de

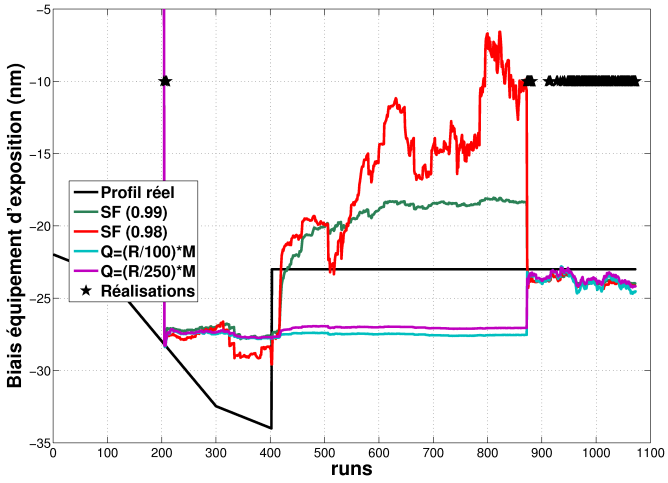

FIGURE 5.15 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_s relatif à un équipement d'exposition donné pour plusieurs algorithmes 1/ SF ($\lambda_{sc} = \lambda_c = 0.99$) en vert, 2/ SF ($\lambda_{sc} = \lambda_c = 0.98$) en rouge, 3/ Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$) en bleu et 4/Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{250}\boldsymbol{M}$) en violet.*

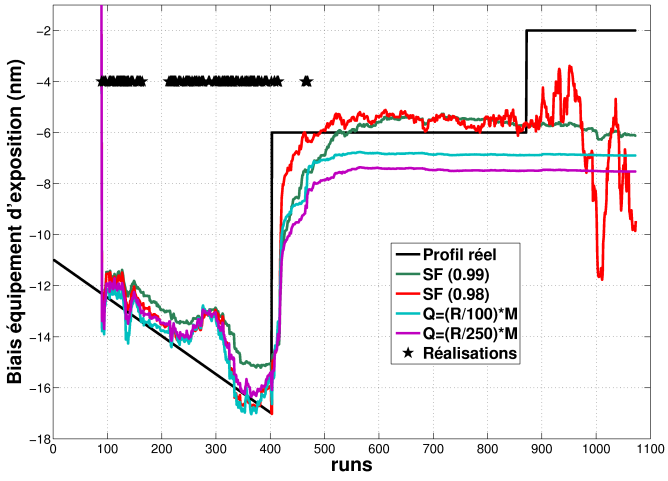

FIGURE 5.16 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_s relatif à un équipement d'exposition donné pour plusieurs algorithmes 1/ SF ($\lambda_{sc} = \lambda_c = 0.99$) en vert, 2/ SF ($\lambda_{sc} = \lambda_c = 0.98$) en rouge, 3/ Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$) en bleu et 4/Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{250}\boldsymbol{M}$) en violet.*

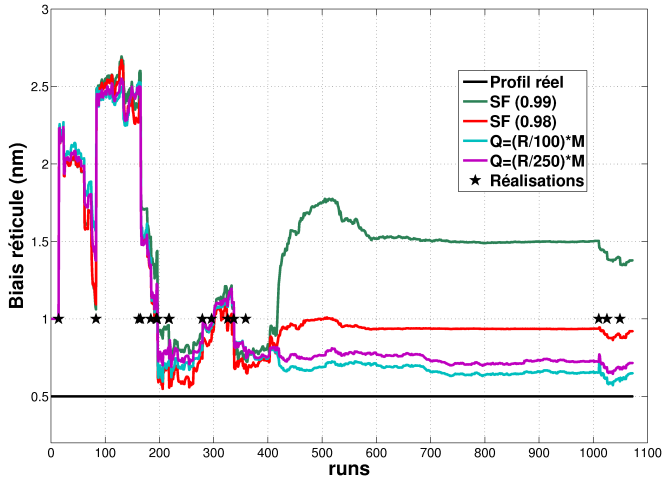

FIGURE 5.17 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_r relatif à un réticule donné pour plusieurs algorithmes 1/ SF ($\lambda_{sc} = \lambda_c = 0.99$) en vert, 2/ SF ($\lambda_{sc} = \lambda_c = 0.98$) en rouge, 3/ Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$) en bleu et 4/Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{250}\boldsymbol{M}$) en violet.*

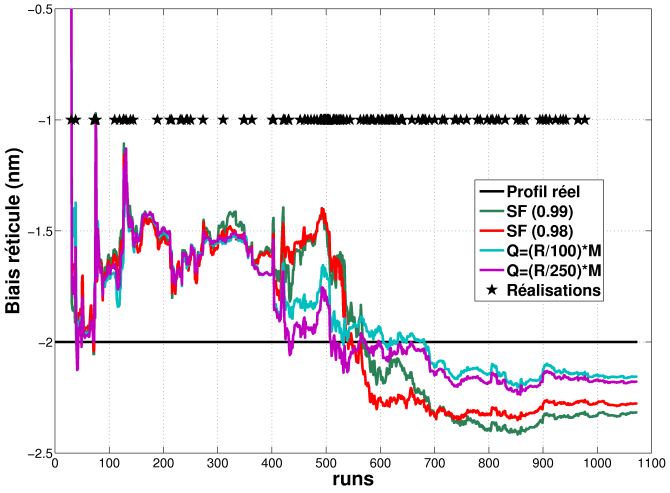

FIGURE 5.18 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_r relatif à un réticule donné pour plusieurs algorithmes 1/ SF ($\lambda_{sc} = \lambda_c = 0.99$) en vert, 2/ SF ($\lambda_{sc} = \lambda_c = 0.98$) en rouge, 3/ Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$) en bleu et 4/Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{250}\boldsymbol{M}$) en violet.*

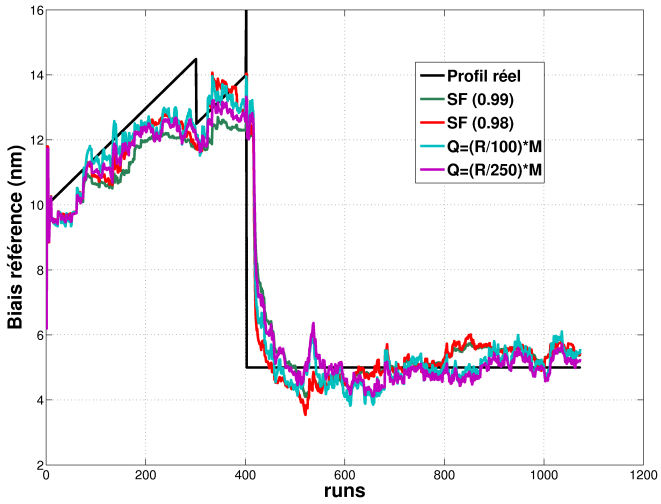

FIGURE 5.19 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_{ref} relatif à la variable référence pour plusieurs algorithmes 1/ SF ($\lambda_{sc} = \lambda_c = 0.99$) en vert, 2/ SF ($\lambda_{sc} = \lambda_c = 0.98$) en rouge, 3/ Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$) en bleu et 4/Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{250}\boldsymbol{M}$) en violet.*

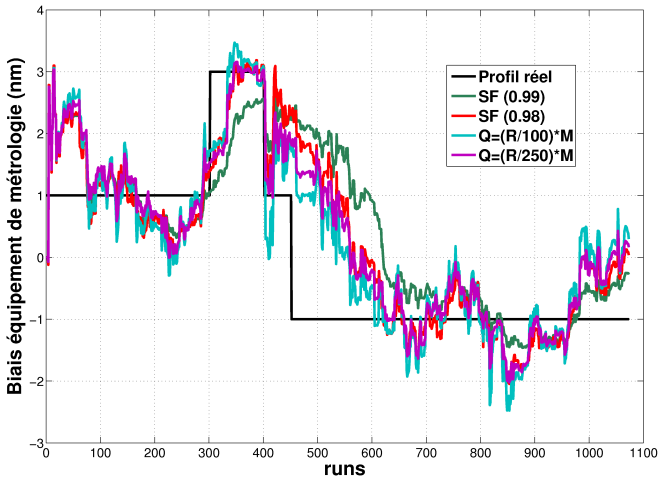

FIGURE 5.20 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_c relatif à un équipement de métrologie donné pour plusieurs algorithmes 1/ SF ($\lambda_{sc} = \lambda_c = 0.99$) en vert, 2/ SF ($\lambda_{sc} = \lambda_c = 0.98$) en rouge, 3/ Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$) en bleu et 4/Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{250}\boldsymbol{M}$) en violet.*

la séquence pourrait dégrader ou au contraire réduire davantage le biais. Cet état des choses est intimement lié à la capacité de la séquence $\{u_k\}$ à exciter le système. Il est évident que dans le cas d'une commande constante, l'algorithme est incapable d'estimer la pente. De la même façon, dans le cas de séquences $\{u_k\}$ constantes par morceaux [13], le biais est non nul.

5.4.5 Synthèse

La performance du filtre de Kalman et des moindres carrés récursifs à oubli sélectif (**SF**) a été analysée dans cette section. Certes, les deux techniques sont capables de poursuivre les trajectoires des perturbations, mais ceci est souvent aux dépens d'une forte sensibilité au bruit. Alors pour préserver la variance des estimateurs, leur capacité de poursuite est souvent bridée.

Que ce soit le filtre de Kalman ou l'algorithme **SF**, l'estimateur peine à coller au profil d'un état, variant dans le temps, notamment aux points d'application d'une dérive ou un décalage. Selon les simulations réalisées dans cette section, la transition s'étalerait en effet sur une plage d'une cinquantaine de réalisations. Ceci serait inacceptable dans un environnement de production de composants micro-électroniques. Wang *et al.* proposent de coupler à l'algorithme d'estimation récursive un algorithme pour détecter les changements (impulsion, échelon, rampe, etc). Les résultats des simulations réalisées dans [Wang et al. (2008)] montrent une amélioration significative de la variance des résidus.

Les résultats rappelés précédemment ont été illustrés dans le cas d'un gain connu ($\varrho = 1$). Un ϱ différent de 1 engendre une estimation biaisée du vecteur Θ. Nous avons alors envisagé d'intégrer le gain dans l'ensemble des paramètres à identifier. Les résultats des simulations rendent compte de l'importance du choix de la séquence des commandes u_k. Une séquence de valeurs aléatoires (bruit blanc) garantit une estimation à biais nul du gain β, ainsi que des différentes variables y_i. A l'inverse, certaines séquences enregistrées en production (longs paliers à dose constante) donnent lieu à un gain et un vecteur Θ biaisés.

5.5 Application aux données de production : Atelier de photolithographie

Par le biais de simulations, nous avons mis en évidence le potentiel du filtre de Kalman ainsi qu'une variante des moindres carrés, appelée **selective forgetting** (**SF**) à identifier en ligne les contributions des différents éléments du contexte de fabrication. Par la suite, nous allons appliquer ces méthodes aux données de production, relatives à l'exposition des vias (via1) d'une technologie particulière. Le choix de ce niveau est lié à l'importance du parc d'équipements impliqué (5 équipements d'exposition et 5 équipements de métrologie) et le grand taux de recyclage au moment des travaux. L'objectif de ce paragraphe est d'estimer le gain attendu d'une éventuelle mise en

13. La séquence est alors comprise de paliers où chaque palier correspond à une campagne d'insolation d'un produit sur un équipement donné. L'amplitude des variations au niveau des paliers est de ce fait faible.

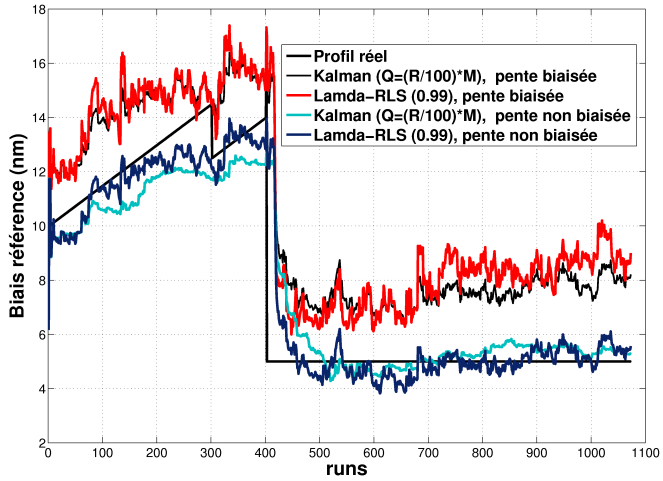

FIGURE 5.21 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$). L'évolution de l'estimation du biais y_{ref} relatif à la variable référence pour deux valeurs de ϱ 1/ $\varrho = 1$, Le filtre de Kalman et les moindres carrés avec facteur d'oubli (0.99) donnent des estimations non biaisées de y_{ref}, et 2/ $\varrho = 0.985$, les mêmes algorithmes donnent des estimations biaisées de y_{ref}.*

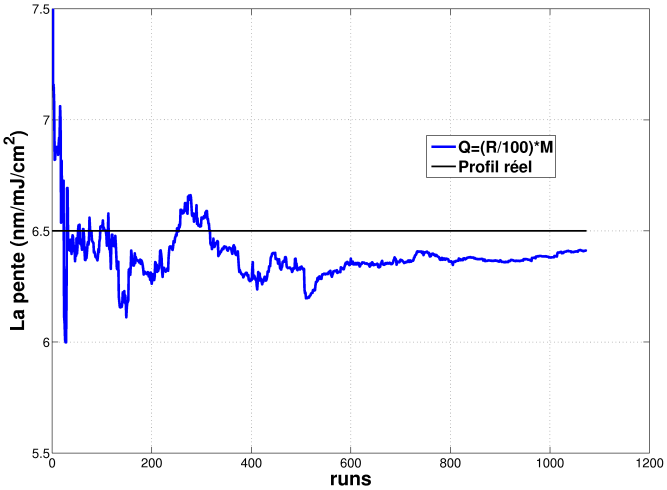

FIGURE 5.22 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$). L'évolution de l'estimation de la pente $\hat{\beta}$ par le filtre de Kalman (**Q**=(R/100)***M**).*

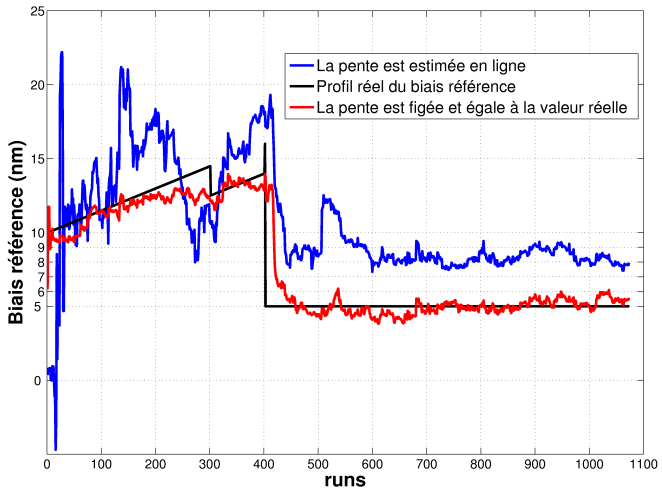

FIGURE 5.23 – *Résultats de la simulation ($\sigma_\epsilon^2 = 3$). L'évolution de l'estimation du biais y_{ref} per le filtre de Kalman (**Q**=$\frac{R}{100}$***M**) dans deux cas 1/ La pente est figée et $\varrho = 1$, $\hat{y}_{ref}$ est alors non biaisé, et 2/ La pente est estimée en ligne, et $\hat{y}_{ref}$ est alors biaisée.*

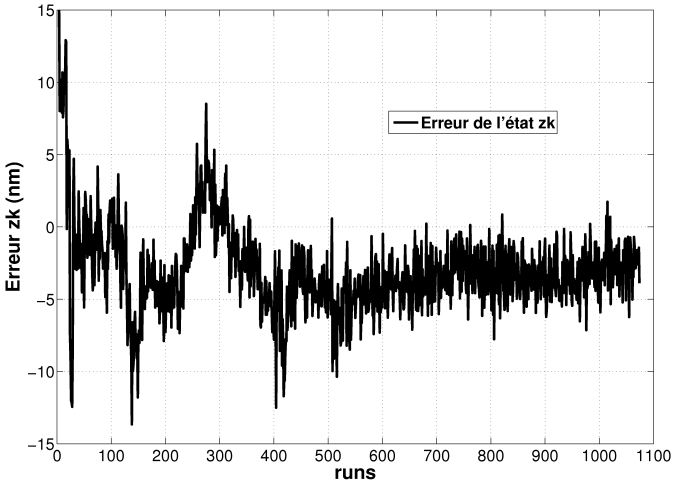

FIGURE 5.24 – *Résultats de la simulation ($\sigma_\epsilon^2 = 3$). L'évolution des résidus $\Delta z_k = z_k - \hat{z}_k$ suite à l'utilisation du filtre de Kalman. La pente est estimée en ligne, et **Q**=$\frac{R}{100}$***M**. A Noter le centrage non nul des résidus.*

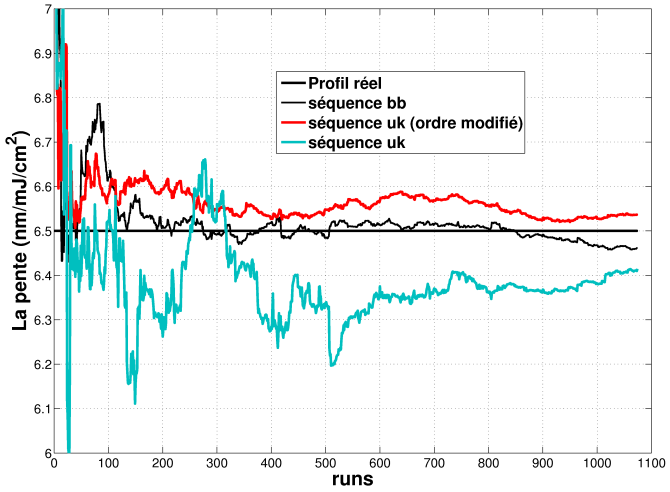

FIGURE 5.25 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$) pour le filtre de Kalman, l'évolution du gain $\hat{\beta}$ pour 3 séquences de u_k 1/bruit blanc en noir, 2/ séquence utilisée en production en vert 3/ La même séquence de production mais l'ordre de réalisation a été modifié d'une façon aléatoire, la courbe est en rouge.*

place d'un observateur, tel celui simulé dans la section précédente, par rapport à la régulation existante.

5.5.1 Description

Les données fournissent un renseignent précieux sur un ensemble d'éléments : la date de passage du lot, la commande en dose d'énergie, la mesure de la dimension critique de la résine ainsi que l'équipement de métrologie utilisé. Elles couvrent la période allant du premier avril 2008 au 9 février 2009. Nous aurions pu inclure d'autres niveaux d'insolation qui font appel à une même chimie [Stuber (2003)], à l'instar des vias des couches ultérieures (via2, via3, via4). Le choix du seul niveau d'exposition répond uniquement à une volonté de simplifier le problème. Inclure d'autres niveaux pourrait être envisagé après avoir démontré le gain d'une telle architecture de l'observateur.

Dans cette optique, nous avons extrait des données relatives à l'exposition des via1 de plusieurs produits d'une même technologie. Ces étapes font appel en effet à la même résine et naturellement à la même recette de couchage. Nous avons recensé 14 réticules avec un nombre minimal de lots exposés, à savoir 10. La figure 5.26 illustre la fréquence de passage des différents produits, un cas typique d'une fabrication high-mix en semi-conducteur.

Afin de veiller à l'alignement ou *matching* des équipements de photolithographie d'une même catégorie (DUV 193 nm, DUV 248 nm, I-line), une tâche de qualité (TQ) est régulièrement effectuée. Elle consiste à appliquer une dose d'énergie commune à l'ensemble du parc, à travers un unique réticule, dédié à cette tâche. Le personnel de production vérifie par la suite que la dimension critique des lignes de résine exposée reste dans les spécifications. Dans le cas d'un résultat où un fort écart est constaté, et confirmé par une seconde mesure, nous devons alors recalibrer l'intensité du laser à travers les facteurs de conversion d'un couple de capteurs intégrés à l'outil d'inso-

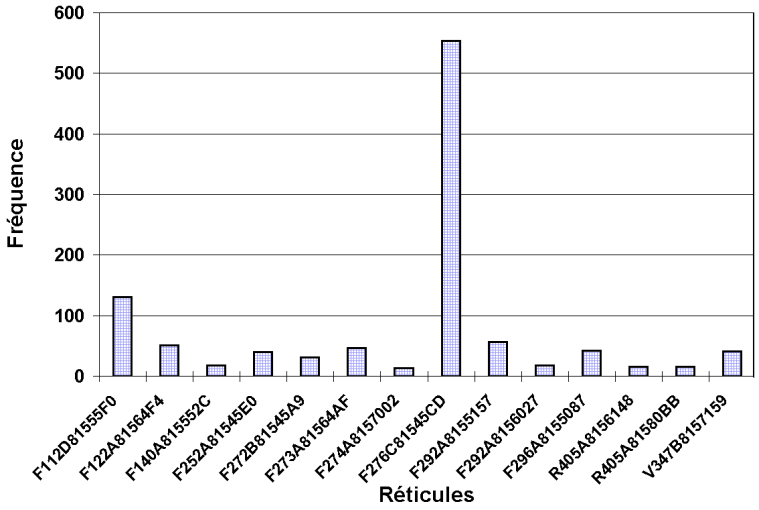

FIGURE 5.26 – *Fréquence de passage des différents réticules.*

lation : Spot Sensor (SS) & Energy Sensor (ES). Nous allons tenter d'en expliquer le principe d'une façon sommaire.

Le capteur ES est situé en amont de la lentille de projection et du réticule. Il est en charge de mesurer l'intensité du laser durant l'insolation de la plaque. Grâce à ces mesures d'intensité, la dose d'énergie délivrée par le laser est asservie à la valeur souhaitée par une boucle de retour, comme le montre la vue schématique 5.27. C'est pour calibrer ce capteur, dont les mesures ne tiennent pas compte des phénomènes d'absorption qui ont lieu dans la lentille que le capteur SS prend tout son intérêt. Ce dernier est en effet positionné sur le même support que la plaquette. La calibration est réalisée off-line, et revient à ajuster le facteur de conversion du capteur ES de sorte que l'énergie déposée dans la résine corresponde à celle mesurée par le capteur SS.

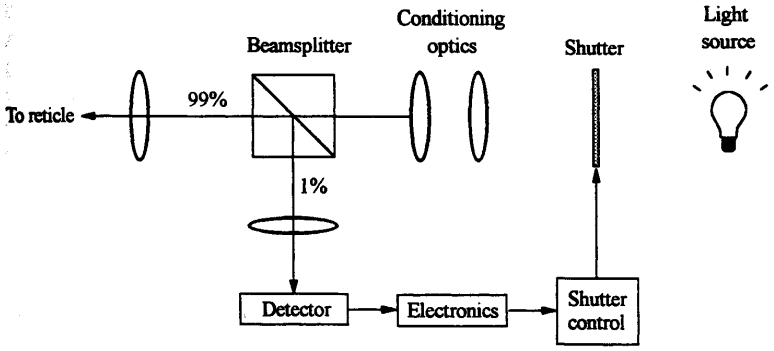

FIGURE 5.27 – *Asservissement pour réguler la dose d'énergie délivrée par le stepper. Source : Levinson (2005).*

Les deux capteurs sont identiques. Grâce à une association de photodiodes, de matériaux spécifiques (filtre, matériaux pour convertir l'ultraviolet en lumière visible) et de circuiterie électronique, le capteur génère un courant électrique, qui sera amplifié puis converti en une tension analogique. Au moyen d'un convertisseur Analogique Numérique (AN), cette différence de potentiel est convertie à son tour en une tension numérique. Comme son nom pourrait le laisser supposer, le facteur de conversion

f_{conv} relatif à chacun des capteurs correspond au rapport entre la dose d'énergie mesurée et la tension délivrée.

$$f_{conv} = \frac{E}{V} \tag{5.34}$$

où E est la dose d'énergie (nJ/cm^2) et V est la tension délivrée par le capteur (bit). Dans le cas d'une tâche de qualité (TQ) où la mesure serait hors spécifications, les facteurs f_{conv} des capteurs ES et SS sont modifiés de sorte à retrouver le point de fonctionnement de référence. Pendant la période que couvrent les données, cette intervention a été réalisée à plusieurs reprises. Alors, pour préserver nos travaux d'identification de toute source d'erreur supplémentaire, nous allons nous intéresser à la fraction des données la plus longue, relative à un facteur de conversion constant par équipement. Cette séquence comporte 1073 réalisations.

5.5.2 Identification en Boucle fermée

Dans le cadre de ces travaux d'identification, la collecte des observations a été réalisée en boucle fermée : une commande linéaire de retour est en place pour asservir le procédé de lithographie, notamment au niveau de l'insolation des vias 1. Ceci entraîne une corrélation entre la commande u_{k+1} et ε_k, ie : $E(u_{k+1}\,\varepsilon_k)|_{\hat{\beta}=\beta} \neq 0$. Par suite de cette corrélation, il a été démontré que même si l'algorithme démarre avec $\hat{\beta}(0) = \beta$, un biais proportionnel à $E(u_{k+1}\,\varepsilon_k)$ sera introduit [Falinower (2008)].

L'identification en boucle fermée est une thématique qui a intéressé les chercheurs depuis les années soixante-dix. La solution préconisée par Box et MacGregor [Box et MacGregor (1976)] est de superposer au signal d'entrée u commandé par le contrôleur un signal D (dither signal) de nature aléatoire. Cette technique permet de rompre la dépendance linéaire et d'augmenter sensiblement, de ce fait, la précision de l'estimation du gain. Il est aussi intéressant de remarquer que l'identifiabilité est aussi restaurée dans le cas d'un contrôleur d'ordre supérieur. D'une manière plus générale, le gain peut être estimé si la commande est retardée, altérée ou encore si les paramètres du correcteur varient avec le temps [Ellingsen (1976)]. Cette dernière alternative a été étudiée notamment par Luceno [Luceño (1997)]. Certes, les techniques proposées par les pionniers Box et MacGregor et par Luceno plus récemment sont effectives et garantissent l'identification des caractéristiques stochastiques et dynamiques du procédé. Mais ces approches requièrent une intervention sur le système, soit par l'injection d'un signal D, soit par la modification des paramètres du correcteur. Ces interventions ne sont pas sans coût pour l'industriel, et ne peuvent être justifiées dans le cas de nos travaux.

Au sein de l'atelier de photolithographie, la mesure de métrologie (exemple : dimension critique de la résine) n'est pas réalisée d'une manière systématique à la suite de l'étape de photolithographie. Selon le degré de priorité du lot, la charge des équipements de métrologie, la disponibilité des opérateurs, la mesure pourrait être réalisée plus au moins rapidement. Naturellement, d'autres lots de productions sont 'processés' pendant cet intervalle de temps, ceci avec une dose d'énergie u_k indépendante de l'erreur toujours pas mesurée. Si nous rajoutons aussi le fait qu'il existe des contrôleurs indépendants pour chacun des équipements, nous pouvons espérer que la condition $E(u_{k+1}\,\varepsilon_k)|_{\hat{\beta}=\beta} = 0$ soit satisfaite.

Toutefois, les simulations réalisées dans la section précédente ont montré que

l'estimation en ligne de la pente β pourrait engendrer une erreur ou un biais, qui est fonction de la séquence des commandes u_k. Cet exercice d'estimer la pente au même titre que les variables y_i à partir des données de production a montré, de surcroît, une grande dépendance de $\hat{\beta}$ aux paramètres des algorithmes utilisés (facteurs d'oublis, matrices de covariances d'un filtre de Kalman, etc) [14].

Pour tout cela, nous avons opté pour une valeur constante égale à $6.5\,nm/mJ.cm^{-2}$. Il s'agit de la même valeur utilisée par la boucle de régulation déployée en production. Ce choix est compréhensible dans la mesure où notre objectif est de quantifier la valeur ajoutée de ce nouvel observateur par rapport à la boucle existante. Il pourrait néanmoins être avantageux d'étudier la performance de l'observateur (la variance de la variable de réponse) en fonction de la valeur choisie et d'en optimiser ainsi le choix. Faute de temps, nous laissons cette initiative aux lecteurs intéressés.

5.5.3 Résultats

Nous allons examiner les résultats du processus d'identification appliqué aux données de production. Notre démarche consiste à optimiser en premier lieu les performances du filtre de Kalman et de l'algorithme **SF**. L'optimisation s'appuiera sur le choix des paramètres (R,**Q**) pour le premier et le couple (λ_{sc}, λ_c) pour le second, afin de minimiser la variance des erreurs de prédiction à priori relatifs à chaque produit [15]. Par définition, l'erreur de prédiction à priori correspond à la différence entre la valeur réelle de CD_{k+1} (à l'instant $k+1$) et sa valeur prédite $\hat{CD}_{k+1}$. $\hat{CD}_{k+1}$ est calculé en se basant sur l'estimation de l'état $\hat{\Theta}_k$ réalisée à l'instant k. La variance des erreurs à priori Σ est une mesure du potentiel prédictif de l'observateur. Une variance Σ plus petite que la variance de la dimension critique (données de production) est le signe d'une possible amélioration de la variabilité du CD grâce à cette nouvelle architecture de régulation.

Le filtre de Kalman

Le filtre de Kalman possède deux paramètres R et $\mathbf{Q} = \varrho\mathbf{M}$. Afin d'en optimiser les performances comme décrit ci-dessus, nous avons fait varier le rapport $r = \frac{R}{\varrho}$. C'est cette fraction qui ajuste le compromis entre la capacité de poursuite et la sensibilité au bruit de mesure. L'évolution de l'écart-type $\sqrt{\sum}$ en fonction de r est présenté en figure 5.28. Le graphique montre que les valeurs extrêmes du rapport r, notamment au delà de 10^4 et en dessous de 10^1, dégradent la capacité de prédiction de l'observateur.

Dans cette étude, nous définissons le gain comme étant le pourcentage d'inflation ou de diminution de l'écart-type de l'erreur à priori ($\sqrt{\sum}$) par rapport à celui des mesures du **CD** (σ_{CD}). Par ailleurs, le σ_{CD} relatif à la population entière des lots est de 5.7 nm, ce qui laisse espérer un gain global supérieur à 8% (voir figure 5.28). Après avoir identifié un domaine de variation plus restreint de r, à savoir $[10; 10^3]$, nous nous proposons de refaire le même calcul pour chaque produit. Certains produits ont en effet des cibles différentes et de ce fait, le gain associé à chaque produit pourrait

14. Il n'y a pas de grand intérêt à appuyer nos dires avec des figures.

15. Un produit est associé à un couple (réticule, valeur cible). Autrement, deux produits différents pourraient avoir le même réticule mais pas la même valeur cible.

Produit	Nombre de lots	$\mathbf{r} = 10^p$			
		$p = 1$	$p = 1.5$	$p = 2$	$p = 2.5$
FF274ACG-02	**12**	-17	-13	-12	-16
F405XXXZ-03	**14**	18	25	30	33
FF292ADG-03	**14**	19	16	8	-2
F405XXXY-01	**15**	15	12	11	11
FF140ABG-03	**17**	-32	-19	-8	1
FF272BAG-05	**30**	-44	-43	-45	-47
FV347BCG-02	**38**	-12	-9	-2	5
FF296ACG-02	**40**	22	22	20	18
FF273ABG-02	**45**	-17	-15	-12	-9
FF122ACG-02	**48**	5	9	11	12
F403XXXY-03	**54**	-1	3	1	-4
FF276CEG-14	**78**	-16	-6	0	1
F401XXXX-02	**115**	10	13	12	9
FF276CEG-13	**459**	-4	-1	0	-1
Total	979	4	7	8	7

TABLE 5.1 – *Tableau récapitulatif des gains enregistrés en fonction du produit et du coefficient* ***r****. Un gain négatif correspond à une dégradation de l'écart-type.*

être sensiblement sous-estimé, ou du moins, différent si on se limite au calcul du gain global.

Le tableau 5.1 donne un apercu détaillé du gain estimé pour chaque produit et pour plusieurs valeurs de **r** (**r** appartenant à l'intervalle $[10; 10^3]$). La valeur $\mathbf{r} = 10^2$ semble être la valeur de **r** la plus poche de l'optimum, garantissant une réduction de la variabilité globale de 8%. Parmi 14 produits, il y en 9 où l'implémentation de ce nouvel observateur permettrait de réduire la variance de la réponse (CD). Le gain oscillerait dans ce cas entre 0 et 30%. D'un autre côté, nous constatons une dégradation de la variance dans le cas des produits **FF27**, **FV347** et **FF140**. L'inflation reste modérée, sauf dans le cas du **FF272BAG-05**, où l'inflation est de l'ordre de 40%.

L'évolution des biais des réticules qui correspondents aux produits mis en cause (figures 5.29, 5.30, et 5.31) souligne effectivement une précision médiocre des estimations $\hat{y}_r$. L'étendue de leurs variations varie en effet entre 2.5 à 7nm, selon le réticule.

L'algorithme SF

L'application de l'algorithme **SF** requiert deux paramètres λ_c et λ_{sc}. Nous avons choisi λ_c égale à λ_{sc} (Soit $\lambda = \lambda_{sc} = \lambda_c$) afin d'en simplifier l'optimisation. Les résultats montrent une dégradation de la variabilité de la réponse, pour λ variant entre 1 et 0.97. D'une façon globale, l'inflation est de 10%.

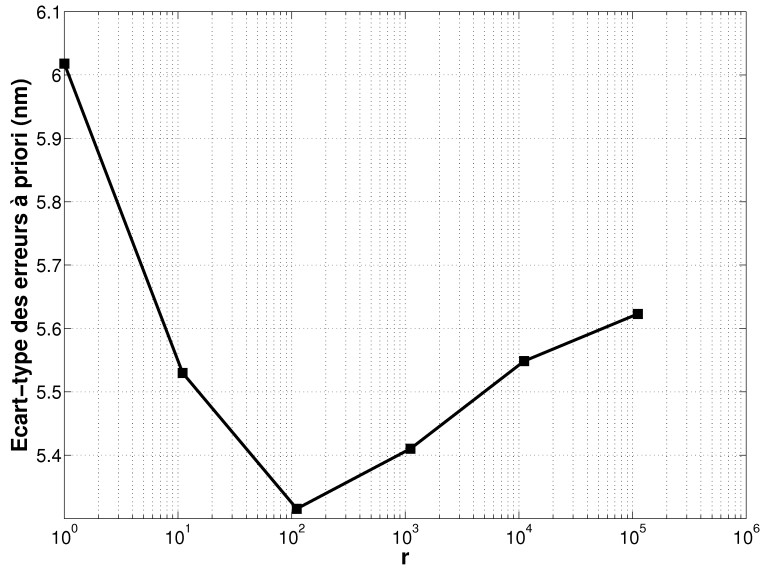

FIGURE 5.28 – *Résultats de l'application du processus d'identification (filtre de Kalman) aux données de production. Evolution de l'écart-type des erreurs à priori en fonction du rapport* $r = \frac{R}{\varrho}$.

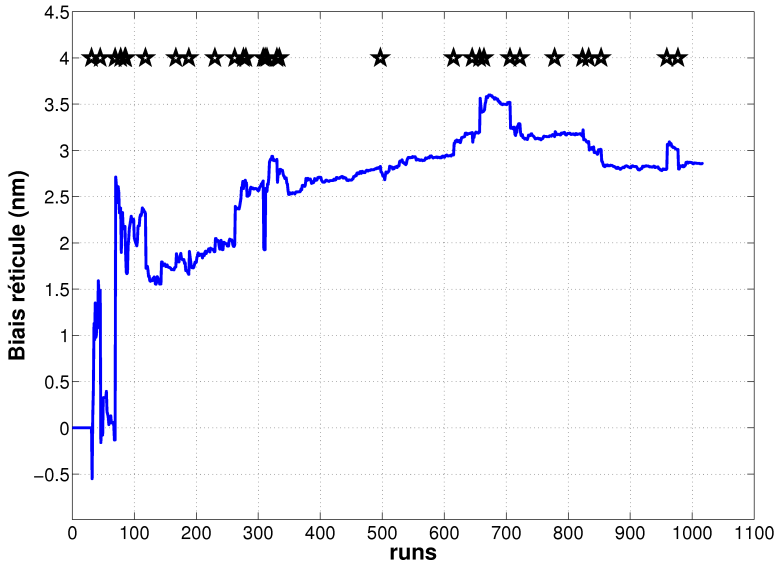

FIGURE 5.29 – *Résultats de l'application du processus d'identification (filtre de Kalman) aux données de production. Evolution de l'estimation du biais* y_r *relatif à un réticule* r_1.

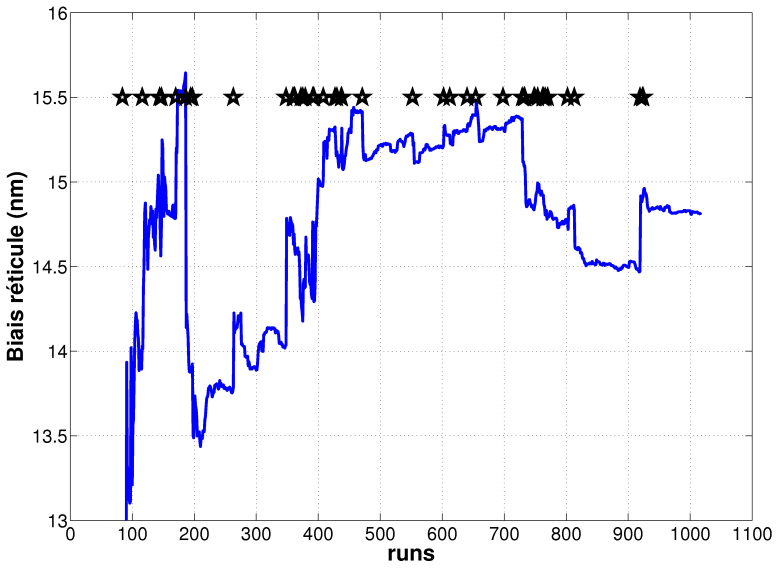

FIGURE 5.30 – *Résultats de l'application du processus d'identification (filtre de Kalman) aux données de production. Evolution de l'estimation du biais y_r relatif à un réticule r_2.*

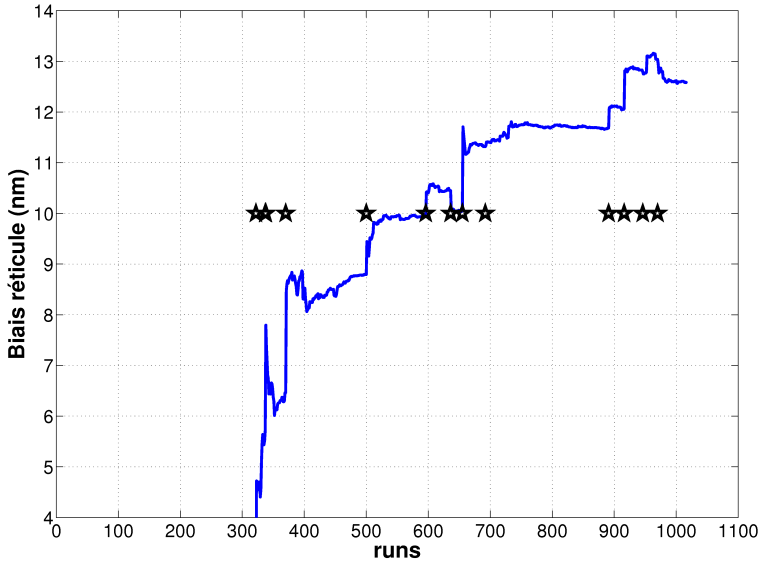

FIGURE 5.31 – *Résultats de l'application du processus d'identification (filtre de Kalman) aux données de production. Evolution de l'estimation du biais y_r relatif à un réticule r_3.*

5.6 Conclusion

L'objet de ce chapitre était de mettre en application des algorithmes d'identification récursive afin d'améliorer l'efficacité des algorithmes de régulation run-to-run. La nouvelle architecture intègre différents éléments du contexte de fabrication (équipements, recettes, etc) et présente ainsi un second avantage, celui de réduire le nombre de boucles à déployer en production. Nous nous sommes penchés sur la performance du filtre de Kalman ainsi que celle d'une variante des moindres carrés, appelée *selective forgetting algorithm* (SF).

Nous avons réalisé des simulations qui ont mis en relief l'existence de compromis entre la capacité de poursuite et la fluctuation statistique des estimateurs. Ce compromis est atteint en ajustant les facteurs d'oubli dans le cas de l'algorithme **SF** et le couple (R,**Q**) pour le filtre de Kalman. Nous avons montré en revanche que l'estimation en ligne de la pente β dépend de la séquence des commandes. Une séquence u_k dont les réalisations sont autocorrélées pourrait conduire à un $\hat{\beta}$ erroné et ainsi à un vecteur Θ biaisé, tandis qu'une séquence de commande de type bruit blanc garantit une estimation non biaisée de β.

Par la suite, nous avons considéré un cas pratique, celui de l'étape de photolithographie des VIA1. Certes cette étape est d'ores et déjà régulée, mais la variabilité du CD demeure forte et le taux de recyclage élevé. Les simulations réalisées à partir des données de production ont démontré que le filtre de Kalman pourrait réduire davantage la variabilité de la réponse CD. Globalement, le gain correspondant serait de l'ordre de 10%. Nous avons constaté néanmoins une dégradation relativement à l'existant de la variabilité de quelques produits. Quant à l'algorithme **SF**, les simulations prédisent une inflation de l'écart-type de l'ordre de 10%.

Je tiens à signaler que nous aurions pu aller plus loin tant au niveau de l'élaboration de l'architecture de l'observateur, qu'au niveau de la simulation des données de production et l'interprétation des résultats. Faute de temps [16], j'ai dû me limiter à l'usage d'un simple estimateur récursif, sans l'associer à un détecteur de changements. Ce dernier aurait, selon Wang *et al.* [Wang et al. (2007)] amélioré de manière significative la performance de l'estimateur. J'ai aussi adopté plusieurs hypothèses simplificatrices qui ont peut être limité la capacité de l'observateur, notamment le choix de la pente β égal à celui de l'application existante ($6.5\,nm/mJ.cm^{-2}$) et le choix d'un facteur d'oubli commun à tous les équipements de fabrication et de métrologie.

En outre du facteur temps, défavorable à une étude plus approfondie, nous n'avons pas bénéficié d'une volonté plus marquée de la part du partenaire industriel. Dès le début de mes travaux sur l'identification récursive et les nouvelles architectures des boucles de régulation, il a été décidé de déployer sur le site de Rousset une nouvelle application run-to-run, appelée Symphonie. Symphonie est une application conçue et développée sur le site de Crolles. Dédiée à l'atelier de photolithographie, elle est lourdement paramétrée pour prendre en compte les effets des réticules, des équipements d'expositions, des équipements de métrologie, et les recettes (ou niveaux d'exposition). Symphonie est en clair un contrôleur coopératif et suit ainsi la tendance générale de l'évolution des régulations run-to-run [Miller et al. (1997)].

16. l'absence de la volonté de l'industriel a aussi joué un rôle

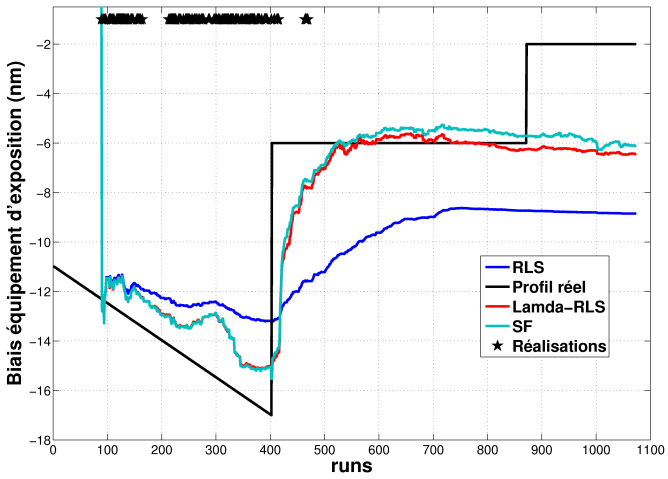

FIGURE 5.32 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'estimation de y_{sc} pour trois variantes des moindres carrés :1/ Les moindres carrés classiques **RLS** en bleu, 2/l'algorithme λ-**RLS** avec $\lambda = 0.99$ en rouge et 3/ le **SF** ($\lambda_{sc} = \lambda_c = 0.99$) en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

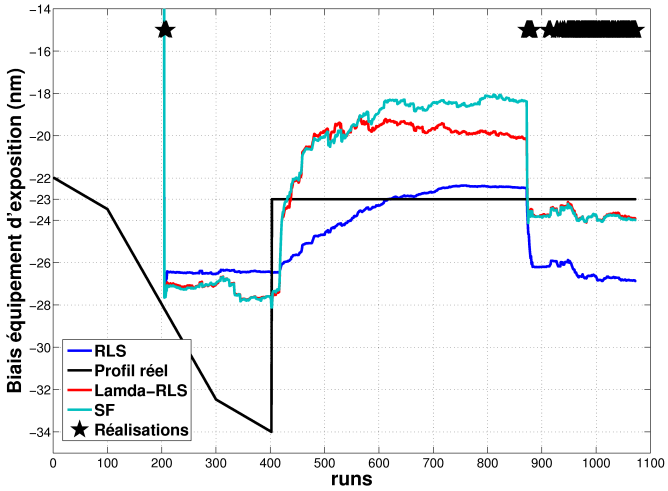

FIGURE 5.33 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'estimation de y_{sc} pour trois variantes des moindres carrés :1/ Les moindres carrés classiques **RLS** en bleu, 2/l'algorithme λ-**RLS** avec $\lambda = 0.99$ en rouge et 3/ le **SF** ($\lambda_{sc} = \lambda_c = 0.99$) en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

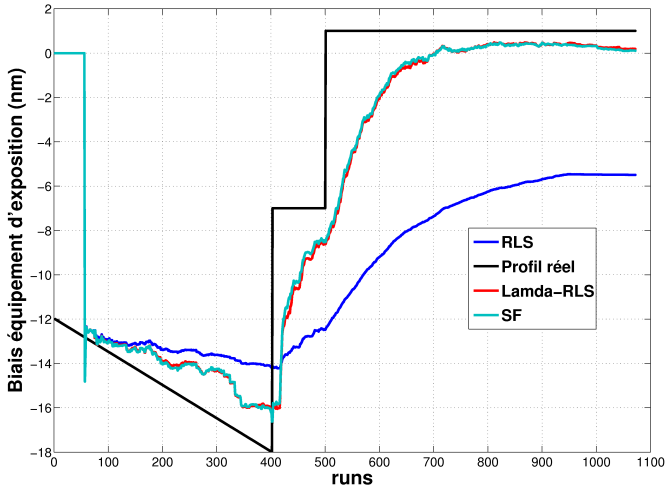

FIGURE 5.34 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'estimation de y_{sc} pour trois variantes des moindres carrés :1/ Les moindres carrés classiques **RLS** en bleu, 2/l'algorithme λ-**RLS** avec $\lambda = 0.99$ en rouge et 3/ le **SF** ($\lambda_{sc} = \lambda_c = 0.99$) en vert.*

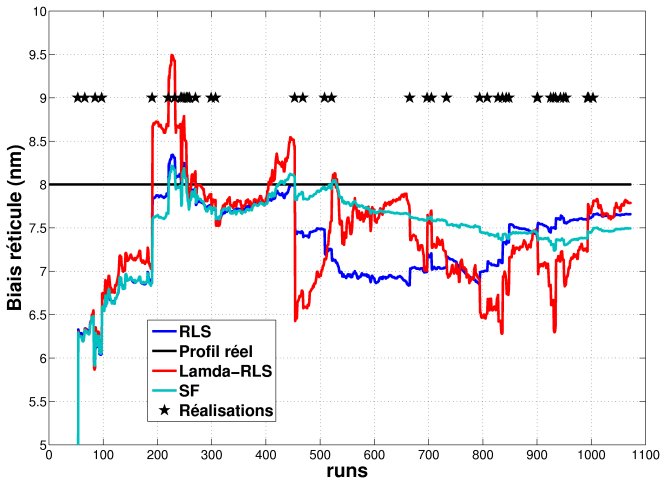

FIGURE 5.35 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'estimation du biais y_r, un réticule à faible volume, par trois variantes des moindres carrés :1/ Les moindres carrés classiques **RLS** en bleu, 2/l'algorithme λ-**RLS** avec $\lambda = 0.99$ en rouge et 3/ le **SF** ($\lambda_{sc} = \lambda_c = 0.99$) en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

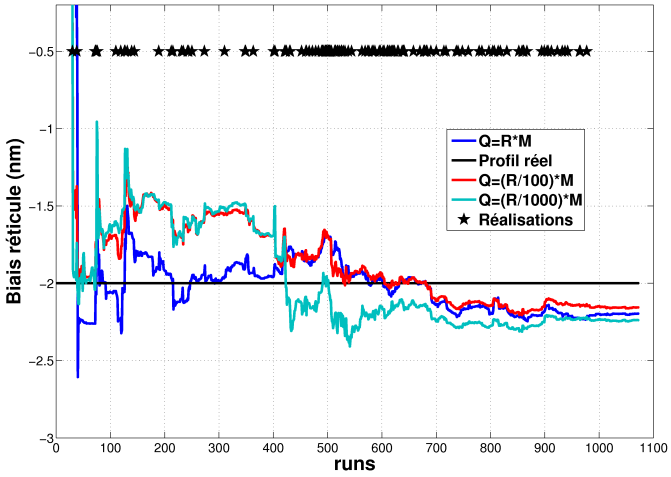

FIGURE 5.36 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_r relatif à un réticule donné pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu, 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en rouge, et 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

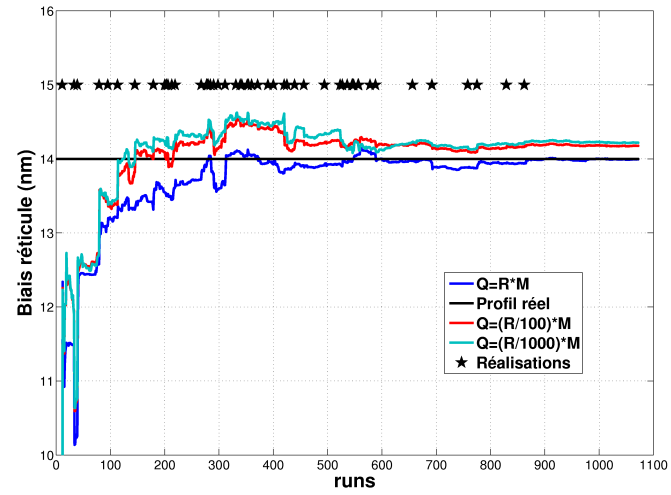

FIGURE 5.37 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_r relatif à un réticule donné pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu, 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en rouge, et 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

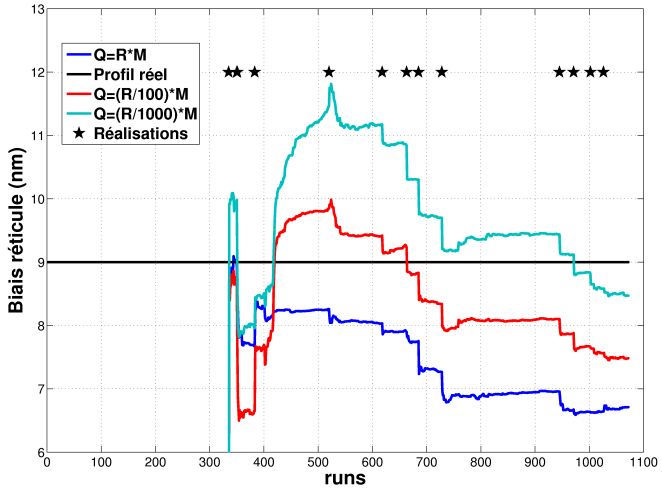

FIGURE 5.38 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_r relatif à un réticule donné pour plusieurs configurations du filtre de Kalman : 1/ $R = 3$ & $Q = R\boldsymbol{M}$ en bleu, 2/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$ en rouge, et 3/ $R = 3$ & $\boldsymbol{Q} = \frac{R}{1000}\boldsymbol{M}$ en vert. Les étoiles noires indiquent les runs où le réticule est utilisé.*

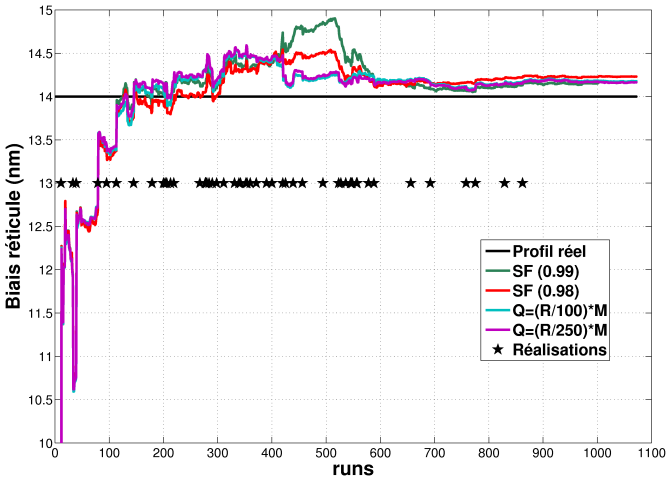

FIGURE 5.39 – *Résultats de la simulation ($\sigma_\varepsilon^2 = 3$, $\varrho = 1$). L'évolution de l'estimation du biais y_r relatif à un réticule donné pour plusieurs algorithmes 1/ SF ($\lambda_{sc} = \lambda_c = 0.99$) en vert, 2/ SF ($\lambda_{sc} = \lambda_c = 0.98$) en rouge, 3/ Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{100}\boldsymbol{M}$) en bleu et 4/Filtre de Kalman ($R = 3$ & $\boldsymbol{Q} = \frac{R}{250}\boldsymbol{M}$) en violet.*

Conclusion Générale

La variabilité paramétrique de technologies **CMOS** avancées, constatée en bout de chaîne de fabrication, requiert la mise en place d'actions appropriées, afin d'en réduire l'ampleur. Ces actions sont d'autant plus essentielles que la variabilité paramétrique affecte les rendements et produit une image défavorable de la fabrication des circuits auprès des clients internes (les divisions) et externes. Les travaux présentés dans le cadre de cette thèse portent sur l'application de techniques de contrôle avancé des procédés à la fabrication des semi-conducteurs. Nous nous sommes intéressés aux méthodes de régulation **run to run** et particulièrement aux boucles de compensation en boucle ouverte et au contrôle coopératif.

Dans la première partie de cette thèse, nous avons conduit une campagne de mesures et avons identifié la longueur de grille L_{poly} comme source majeure de la **variabilité paramétrique spatiale** (variations intra-plaque) et **temporelle** (variations lot à lot et plaque à plaque). Nous avons mis en évidence le poids important des unités de recuit après exposition (**PEB**) dans les variations intra-plaque de L_{poly}. Ces résultats ont été décisifs pour remplacer les fours **RHP** (Rapid Hot Plate) existants par des fours plus récents, les **SRHP** (Super Rapid Hot Plate). Cette action a permis de réduire l'étendue de la dimension critique des lignes de résine de 50%.

Afin de réduire la variabilité paramétrique temporelle et notamment lot à lot, nous avons investi le domaine du contrôle statistique avancé des procédés, en l'occurrence les boucles de régulation **run to run**. Nous avons réalisé un état de l'art de régulations, potentiellement envisageables dans le contexte d'une fabrication de semi-conducteurs. Nous avons considéré le cas des procédés à gain statique, perturbés par des processus de type **ARIMA**(1,1,1). A partir de ce choix, nous avons déroulé une double approche automaticienne/statisticienne afin d'étudier la première des commandes, la commande **PID**.

Nous avons démontré que la commande **I** est essentielle dans un contexte où les procédés sont sujets à des dérives et des décalages. C'est moins le cas de la composante proportionnelle qui dégrade le temps de réponse du système. En revanche, associée à l'action **I**, l'action proportionnelle garantit une variance minimale dans un espace de variation des paramètres stochastiques plus large. Elle est dans ce sens bénéfique dans le cas des perturbations de type rampe ou dérive. Nous avons vu par ailleurs que la commande dérivée présente peu d'intérêt quant il s'agit de procédés à gain statique.

Dans ce même état de l'art, nous avons rappelé quelques aspects généraux de certaines lois de commandes, à savoir la commande **MMSE** (*Minimum Mean Square Error*), la commande **EWMA** (*Exponentially Weighted Moving Average*) et la commande adaptative. Nous avons aussi réalisé un état des lieux des applications des régulations **run to run** dans les usines de fabrication de composants microélectroniques, en

l'occurrence en photolithographie.

Cette thèse a débouché concrètement sur le développement d'une boucle de régulation de compensation (en boucle ouverte). Le contrôleur compense la déviation de L_{poly}, mesurée par scattérométrie, en ajustant la dose d'implantation des poches. Son implémentation en production a conduit à une réduction de la variabilité de caractéristiques électriques critiques, notamment le courant de saturation (30%) et le courant de fuite (15%). Le déploiement de cette boucle est venu poursuivre et achever l'oeuvre dont le remplacement des fours **PEB** était la première brique. Nous avons pu ainsi atteindre les standards mondiaux en terme de variabilité paramétrique de la technologie CMOS 0.13μm.

Cet outil de régulation est vite devenu un outil clé dans le contrôle des procédés de la technologie 130nm, et il est d'ores et déjà en cours d'évaluation pour une technologie plus avancée (90nm). Nous avons toutefois identifié une faiblesse qui pourrait éventuellement altérer la capacité du contrôleur à centrer les produits. Il s'agit du biais iso-dense Δ. Alors pour remédier à cela et augmenter la robustesse de la régulation face aux fluctuations de Δ, nous avons proposé de remplacer la structure scattérométrique, utilisée en première version, par une seconde structure de test plus adaptée.

Dans le dernier chapitre, nous nous sommes intéressés au développement d'un contrôleur coopératif. L'idée du contrôle coopératif est de concevoir un observateur, capable d'attribuer en ligne la déviation de la variable de sortie aux différentes sources de variation qualitatives. Chacune de ces sources, en l'occurrence les équipements de métrologie, de process ainsi que les produits, possède un **offset** (ou biais) qui est mis à jour à chaque nouvelle mesure. L'observateur, ainsi défini, est simplement un estimateur récursif.

Nous avons simulé le comportement du filtre de Kalman et d'une variante des moindres carrés, dite à oubli sélectif (**SF**). Les résultats ont mis en exergue l'existence d'un compromis entre la capacité de poursuite de l'algorithme et la fluctuation statistique de l'estimateur. Nous avons aussi constaté la difficulté d'estimer la pente en ligne, dont la précision est subordonnée à la nature de la séquence des commandes. En seconde partie, nous avons considéré un cas pratique, celui de l'étape de photolithographie des VIA1 d'une technologie CMOS 0.18μm. Les simulations réalisées à partir des données de production ont montré l'incapacité de l'estimateur des moindres carrés **SF** à faire mieux que la régulation existante. Les simulations prédisent en effet une inflation de l'écart-type de 10%. A contrario, le filtre de Kalman paraît à même de réduire la variabilité de la variable de réponse, en l'occurrence le diamètre des VIA1, de l'ordre de 10%.

Annexes

A

Sommaire

A.1 Notions Élémentaires en Modélisation des Processus Stochastiques Discrets

Un processus stochastique est une suite de variables aléatoires indexées par le temps :

$$(X_t; t \in \mathbb{Z}) \tag{A.1}$$

Ici t appartient à un ensemble discret, ce qui définit un processus en temps discret. Le choix des processus discrets s'explique par la spécifité de l'industrie du semiconducteur qui déploie exclusivement des procédés de type discontinu ou encore par lots.

A.1.1 Modèles ARIMA

La modélisation de réalisations auto corrélées est souvent réalisée à l'aide des équations aux différences stochastiques. Les travaux fondateurs de Box et Jenkins Box et al. (1994) ont donné naissance à un formalisme simple de ces équations, appelées communément modèles ARIMA (*Autoregressive Integrated Moving Average*). Un modèle ARIMA est composé potentiellement de trois types de processus : les processus auto-régressifs AR, les processus de moyenne mobile MA et les processus d'intégration I. Pour permettre une expression formelle plus simple de ces différents processus, nous introduisons l'opérateur retard défini de la façon suivante :

On considère un processus stochastique $(x_t ; t \in \mathbb{Z})$, l'opérateur retard noté B [1], est défini tel que :

$$Bx_t = x_{t-1} \ \forall t \in \mathbb{Z} \tag{A.2}$$

Modèles auto-régressifs AR(p)

Un processus auto-régressif d'ordre p suppose que la valeur courante est déterminée par la somme pondérée des p valeurs précédentes, plus un terme aléatoire d'erreur. Ainsi, le processus $(x_t ;\ t \in \mathbb{Z})$ satisfait une représentation AR d'ordre p, notée AR(p), si et seulement si :

$$\Phi(B)x_t = c + \varepsilon_t \tag{A.3}$$

avec

$$c \in \mathbb{R},$$
$$\Phi(B) = \sum_{j=0}^{p} \phi_j B^j \,/\, \text{où } \forall j < p, \phi_j \in \mathbb{R}, \phi_0 = 1, \phi_p \in \mathbb{R}^*,$$
$$\varepsilon \equiv i.i.d(0, \sigma^2).$$

1. Suivant les ouvrages, il est aussi noté L pour *Lag*

Modèles à moyenne mobile MA(q)

Un processus de moyenne mobile d'ordre q suppose que la valeur actuelle est déterminée par la somme pondérée des erreurs ayant entaché les q valeurs précédentes, à laquelle s'ajoute un terme spécifique d'erreur. Un processus (x_t ; t $\in \mathbb{Z}$) satisfait une representation MA d'ordre q, notée MA(q), si et seulement si :

$$x_t = m + \Theta(B)\varepsilon_t \tag{A.4}$$

avec

$$\begin{aligned}
&m \in \mathbb{R} \\
&\Theta(B) = \sum_{j=0}^{q} \theta_j B^j \text{ où } \forall j < q, \theta_j \in \mathbb{R}, \theta_0 = 1, \theta_q \in \mathbb{R}^* \\
&\varepsilon \equiv i.i.d(0, \sigma^2)
\end{aligned}$$

Modèles ARMA(p,q)

Naturellement, les processus ARMA se définissent par l'adjonction d'une composante autorégressive AR et d'une composante moyenne mobile MA. D'une manière identique aux paragraphes précédents, en voilà la définition.

Le processus (x_t ; t $\in \mathbb{Z}$) satisfait une representation ARMA ; d'ordre p et q, notée ARMA(p ; q), si et seulement si :

$$\Phi(B)x_t = c + \Theta(B)\varepsilon_t \tag{A.5}$$

avec

$$\begin{aligned}
&c \in \mathbb{R}, \\
&\Theta(B) = \sum_{j=0}^{q} \theta_j B^j \text{ où } \forall j < q, \theta_j \in \mathbb{R}, \theta_0 = 1, \theta_q \in \mathbb{R}^*, \\
&\Phi(B) = \sum_{j=0}^{p} \phi_j B^j \text{ où } \forall j < p, \phi_j \in \mathbb{R}, \phi_0 = 1, \phi_p \in \mathbb{R}^*, \\
&\varepsilon \equiv i.i.d(0, \sigma^2).
\end{aligned}$$

Modèle d'intégration I(d)

Le modèle ARMA, tel que défini précédemment, s'applique à des processus stationnaires, ie oscillant autour d'une valeur moyenne constante et ayant une variance finie. Sauf qu'en pratique, il n'est pas toujours raisonnable de considérer les procédés industriels comme stationnaires. Dans ce cas, l'idée est alors de se ramener au cas stationnaire par utilisation de l'opérateur ∇ tel que :

$$\nabla x_t = (1 - B)x_t = x_t - x_{t-1} \forall t \in \mathbb{Z} \tag{A.6}$$

Un processus d'intégration ou intégré d'ordre d est un processus qui a besoin d'être différencié d fois, $\nabla^d x_t$, avant d'atteindre la stationnarité.

Modèles ARIMA(p,d,q)

Au vu du paragraphe précédent, un processus non stationnaire (x_t ; t$\in \mathbb{Z}$) sera dit ARIMA(p,d,q) s'il existe un entier positif d tel que $\nabla^d x_t$ soit un processus ARMA(p,q).

A.1.2 La prévision

Une fois un modèle ARIMA spécifié et estimé, nous pouvons l'utiliser pour effectuer des prévisions. Se situant à un instant t quelconque, nous nous proposons d'effectuer une prévision à l'horizon $h \in \mathbb{N}^*$, c'est à dire de prévoir la réalisation du processus à l'instant t+h, x_{t+h}. Cette prévision sera basée sur la connaissance d'un ensemble d'informations disponibles à cette date, notamment l'ensemble des observations du passé ($x_{t-j}/j \in \mathbb{N}$).

Soit $(x_t, t \in \mathbb{Z})$ un processus ARMA(p,q). Il se formule comme suit :

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} \ldots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \ldots + \theta_q \varepsilon_{t-q} \tag{A.7}$$

Soit $\widehat{x}_{t+h|t}$ la prévision de la variable aléatoire x_{t+h} calculée à l'instant t. La qualité de la prévision est mesurée à l'aide d'une distance à la valeur réelle de la réalisation x_{t+h}, appelée fonction de coût. S'étant donné une fonction mesurant le coût de la prévision, la minimisation du coût moyen, conditionnellement à l'information disponible, conduit à une prévision dite optimale. Si nous retenons l'erreur quadratique moyenne (*Mean Square Error*) comme fonction de coût, car elle se prête facilement aux calculs, la prévision optimale est définie par l'équation suivante :

$$\min_{\widehat{x}_{t+h|t}} MSE(\widehat{x}_{t+h|t}) = \min_{\widehat{x}_{t+h|t}} E\left[(x_{t+h} - \widehat{x}_{t+h|t})^2 / (x_{t-i}, i \in \mathbb{N})\right] \tag{A.8}$$

La résolution de cette équation montre que la prévision optimale est égale à l'espérance conditionnelle de x_{t+h} (à définir).

$$\widehat{x}_{t+h|t} = E_t[x_{t+h}] \tag{A.9}$$

Le calcul de l'espérance conditionnelle dans le cas d'un processus ARMA est très simple. Il suffit de reécrire le modèle sous forme d'équation aux différences(A.7) et d'appliquer les règles suivantes [Castillo (2002)] :

1. Remplacer $x_{t-j}(j \geq 0)$ par les valeurs observées.
2. Remplacer $x_{t+j}(j \geq 1)$ par $\widehat{x}_{t+j|t}$.
3. Remplacer $\varepsilon_{t-j}(j \geq 0)$ par $\widehat{\varepsilon}_t = x_{t-j} - \widehat{x}_{t-j|t-j-1}$.
4. Remplacer $\varepsilon_{t+j}(j \geq 1)$ par $E_t[\varepsilon_{t+j}] = 0$.

Par souci de clarté, voici un exemple d'un processus IMA $(y_t, t \in \mathbb{Z})$ à moyenne nulle. Le modèle s'écrit :

$$y_t = y_{t-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1}$$

La prévision à un pas, optimale au sens de l'erreur quadratique moyenne, est obtenue comme suit :

$$\widehat{y}_{t+1|t} = y_t - \theta_1 (y_t - \widehat{y}_{t|t-1})$$

Nous obtenons après arrangement :

$$\widehat{y}_{t+1|t} = (1 - \theta_1) y_t + \theta_1 \widehat{y}_{t|t-1}$$

Nous constatons que la prévision optimale (MMSE) à un pas (h=1) d'un processus IMA(1,1) est équivalente à *la Moyenne mobile pondérée de façon exponentielle* (EWMA) [2], avec $\lambda = 1 - \theta_1$.

2. La méthode EWMA sera détaillée davantage par la suite

A.2 Outils d'analyse des systèmes discrets en automatique

Pour caractériser un système bouclé, l'automaticien a recours à une boîte d'outils lui permettant d'analyser trois éléments importants : la stabilité du système, son comportement en régime transitoire et son comportement en régime permanent suite à l'application d'une perturbation. S'agissant de systèmes discrets, un outil mathématique privilégié est la transformation en z, dont nous allons donner un bref aperçu par la suite. Rappelons que cette approche traite exclusivement les perturbations déterministes, et notamment la perturbation en échelon.

A.2.1 Transformation en z

La **transformée en** z $X(z)$ d'un signal discret $x(k)$ est donnée par :

$$X(z) = \sum_{k=0}^{+\infty} x(k) \cdot z^{-k} \quad |z| > r_0 \tag{A.10}$$

où z est une variable complexe.

Parmi les propriétés de cette transformation (linéarité, transformée d'un produit de convolution, etc) [3], je mettrai l'accent sur deux propriétés essentielles à la compréhension de la suite de ce chapitre.

Translation avant (signal retardé)

La transformée en z d'un signal discret **retardé** de d périodes est donnée par la transformée en z du signal non-retardé multipliée par z^{-d} :

$$Z\{x(k-d)\} = z^{-d} \cdot Z\{x(k)\} = z^{-d} \cdot X(z) \tag{A.11}$$

Théorème de la valeur finale

La valeur finale d'un signal discret x(k) peut se calculer par :

$$x_\infty = x(\infty) = \lim_{k\to\infty} x(k) = \lim_{z\to 1} ((z-1) \cdot X(z)) \tag{A.12}$$

Cette formule s'avérera très utile ultérieurement pour calculer le gain permanent de systèmes dynamiques linéaires discrets.

A.2.2 Fonction de transfert discrète

Soit un système discret décrit par son équation aux différences, ou y est le signal de sortie et u le signal d'entrée.

$$\begin{aligned} &y(k) + a_1 \cdot y(k-1) + ... + a_{n-1} \cdot y(k-n+1) + a_n \cdot y(k-n) \\ &\quad = b_0 \cdot u(k-d) + b_1 \cdot u(k-d-1) + ... + b_{m-1} \cdot u(k-n+1) + b_m \cdot u(k-n) \end{aligned} \tag{A.13}$$

La transformée en z des deux membres de l'équation aux différences donne :

$$\begin{aligned} &\left(1 + a_1 \cdot z^{-1} + ... + a_{n-1} \cdot z^{1-n} + a_n \cdot z^{-n}\right) \cdot Y(z) \\ &\qquad = z^{-d} \cdot \left(b_0 + b_1 \cdot z^{-1} + ... + b_{m-1} \cdot z^{1-m} + b_m \cdot z^{-m}\right) \cdot U(z) \end{aligned} \tag{A.14}$$

3. Pour le lecteur intéressé par une compréhension plus poussée de la transformation en z, nous le renvoyons à l'ouvrage Godoy et Ostertag (2003)

d'où :

$$G(z) = \frac{Y(z)}{U(z)} = z^{-d} \cdot \frac{b_0 + b_1 \cdot z^{-1} + ... + b_{m-1} \cdot z^{1-m} + b_m \cdot z^{-m}}{1 + a_1 \cdot z^{-1} + ... + a_{n-1} \cdot z^{1-n} + a_n \cdot z^{-n}} \tag{A.15}$$

$G(z)$ est appelée fonction de transfert du système . Elle apparaît ici comme une fraction rationnelle en z, mise sous forme de puissances de z négatives.

Pôles et zéros et ordre du système

Les valeurs de z qui annulent le numérateur de $G(z)$ en sont les **zéros**. $G(z)$ compte donc m zéros, réels ou complexes. Les zéros sont ainsi $z_1, z_2, ..., z_m$ valeurs dépendant des coefficients b_0 à b_m. Quant aux valeurs de z annulant le dénominateur de $G(z)$, elles portent le nom de **pôles**, au nombre de n. Ceux-ci sont également réels ou complexes. Les pôles sont $p_1, p_2, ..., p_n$ valeurs dépendant des coefficients a_1 à a_n. L'*ordre* d'un système dynamique linéaire est égal à son nombre de pôles n.

A.2.3 Stabilité

La stabilité est l'un des plus points les plus importants dans l'étude des systèmes asservis. Au sens de la stabilité EB-SB, elle garantie que la sortie du système sera bornée, du moment que les entrées (commande et perturbations) le sont aussi.

Théorème

Un système dynamique linéaire discret est stable, si et seulement si, tous les pôles de sa fonction de transfert sont situés à l'intérieur du disque-unité :

$$|p_i| < 1 \tag{A.16}$$

La stabilité est donc une propriété intrinsèque, dépendant exclusivement des paramètres et de la structure du système, mais aucunement des signaux d'entrée. Lorsqu'un ou plusieurs pôles sont à l'extérieur du cercle-unité, le système est instable. Pour des pôles situés exactement sur le cercle-unité, le système est à *stabilité marginale*.

Éléments de synthèse de régulateurs : Lieu des pôles ou lieu d'Evans

Le lieu des pôles représente dans le plan complexe l'évolution des pôles de la fonction de transfert en boucle **fermée** lorsque le gain de la boucle k_o varie de 0 à l'infini. Si la fonction de transfert en boucle ouverte d'un système de régulation a pour expression générale

$$G_o(z) = \frac{Y(z)}{E(z)} = k_o \cdot \frac{n_o(z)}{d_o(z)} \tag{A.17}$$

La fonction de transfert en boucle fermée s'écrit, lorsque le retour est unitaire (figure A.2) :

$$G_w(z) = \frac{Y(z)}{W(z)} = \frac{G_o(z)}{1 + G_o(z)} = \frac{k_o \cdot \frac{n_o(z)}{d_o(z)}}{1 + k_o \cdot \frac{n_o(z)}{d_o(z)}} = \frac{k_o \cdot n_o(z)}{d_o(z) + k_o \cdot n_o(z)} \tag{A.18}$$

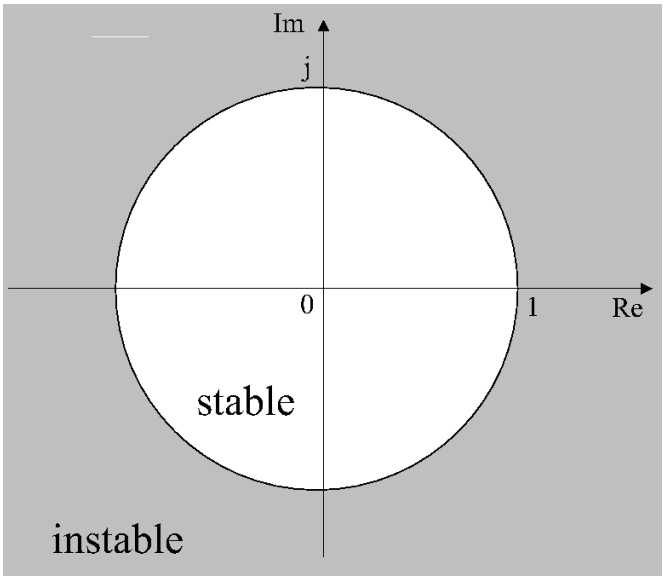

FIGURE A.1 – *Le partage du plan complexe z en zone de stabilité (intérieur du cerle unitaire) et en zone d'instabilité*

Les pôles de la fonction de transfert en boucle fermée sont les valeurs de z annulant le dénominateur de $G_w(z)$. Ils sont donc solutions de l'équation de l'équation caractéristique

$$d_c(z) = d_o(z) + k_o \cdot n_o(z) = 0 \tag{A.19}$$

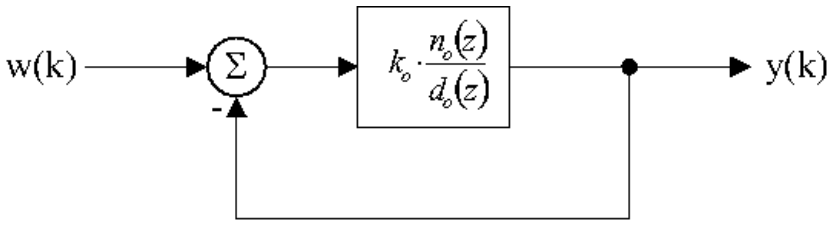

FIGURE A.2 – *Schéma fonctionnel d'un système de régulation discret, présenté de façon à ce que le retour soit unitaire et mettant en évidence la fonction de transfert en boucle ouverte* $G_o(z)$.

Partant d'un système asservi ayant pour fonction de transfert de boucle G_o, il s'agit de déterminer la valeur k du facteur d'Evans k_o telle que les pôles dominants en boucle fermée répondent à des spécifications particulières (taux d'amortissement, nature du régime transitoire, etc).

A.2.4 Précision en régime permanent

La précision d'un système asservi est obtenue en chiffrant la valeur de l'erreur $\zeta(t)$ entre la sortie et la valeur cible $\zeta(k) = y_C(k) - y(k)$. On se limite ici à la précision en régime permanent. Dans le cas d'une valeur cible constante et égale à 0, et en appliquant le théorème de la valeur finale, l'erreur statique s'écrit :

$$\zeta_\infty = \lim_{k\to\infty} -y(k) = \lim_{z\to 1} -(z-1)Y(z) \tag{A.20}$$

A.2.5 La rapidité du système : Durée de réponse

La **durée de réponse** $T_{\text{rég}}$ est la durée mesurée entre l'instant d'application d'une perturbation déterministe $P(t)$ et l'instant où la grandeur réglée $y(t)$ ne s'écarte plus d'une bande de tolérance de $\pm 2\%$ tracée autour de sa valeur finale y_∞. Elle nous permet d'évaluer la réponse du système en boucle fermée. Nous l'approximons par la formule suivante.

$$t_{rep} = -\frac{4}{\ln(|p_m|)} \tag{A.21}$$

Où p_m est le pôle dominant du système bouclé, $p_m = \{p, |p| = \max(|p_i|)\}$. En s'appuyant sur cette relation, nous constatons que plus les pôles se rapprochent de 0 (dans le plan complexe z), plus le système est rapide. Si nous prenons l'exemple d'un pôle dominant égal à 1. Dans ce cas, le système asservi a un pôle situé sur le cercle-unité. Il est alors à stabilité marginale et n'atteindra jamais le régime permanent : $t_{rep} = \infty$.

Bibliographie

Peter J. Apostolakis. Statistical approach to optimizing advanced low-voltage sem operation. volume 1464, pages 406–412. SPIE, 1991. (Cité page 16.)

Junwei Bao. *An Optical Metrology System for Lithography Process Monitoring and Control*. PhD thesis, EECS Department, University of California, Berkeley, 2003. (Cité pages 16, 18, 19, 20 et 81.)

Michèle Basseville et Igor V. Nikiforov. *Detection of abrupt changes : theory and application*. Prentice-Hall, Inc., 1993. ISBN 0-13-126780-9. (Cité page 136.)

Mortini Benedicte. *Etude des resines photolithographiques positives 193 nm a amplification chimique et mise au point de leurs conditions de procede = study of 193 nm positive chemically amplified photoresists and optimization of their process conditions*. PhD thesis, Université Joseph Fourier De Grenoble, 2001. (Cité page 6.)

Christopher A. Bode. *Run-To-Run Control Of Overlay And Linewidth In Semiconductor Manufacturing*. PhD thesis, University of Texas, Austin, 2001. (Cité pages 115 et 121.)

Pierre Borne et Jean-Pierre Richard. *Analyse et régulation des processus industriels*. Editions TECHNIP, 1993. ISBN 2710806436, 9782710806431. (Cité pages 47 et 49.)

George Box, Gwilym M. Jenkins, et Gregory Reinsel. *Time Series Analysis : Forecasting & Control*. Prentice Hall, 3rd édition, Février 1994. ISBN 0130607746. (Cité pages 45, 52, 56, 120 et 158.)

George Box et Tim Kramer. Statistical process monitoring and feedback adjustment : a discussion. *Technometrics*, 34 :251–267, 1992. (Cité pages 46, 55 et 75.)

George E. P. Box et Alberto Luceño. Discrete proportional-integral adjustment and statistical process control. *Journal of Quality Technology*, 29 :248–260, 1997a. (Cité page 59.)

George E. P. Box et Alberto Luceño. *Statistical Control : By Monitoring and Feedback Adjustment*. Wiley-Interscience, 1 édition, Septembre 1997b. ISBN 0471190462. (Cité pages 46 et 59.)

George E. P. Box et John F. MacGregor. Parameter estimation with closed-loop operating data. *Technometrics*, 18 :371–380, Novembre 1976. ISSN 00401706. (Cité page 145.)

W. Jarrett Campbell, Stacy K. Firth, Anthony J. Toprac, et Thomas F. Edgar. A comparison of run-to-run control algorithms. Dans *American Control Conference, 2002. Proceedings of the 2002*, volume 3, pages 2150–2155 vol.3, 2002. ISBN 0743-1619. (Cité page 114.)

Liyu Cao et Howard M. Schwartz. A novel recursive algorithm for directional forgetting. Dans *American Control Conference, 1999. Proceedings of the 1999*, volume 2, pages 1334–1338 vol.2, 1999. (Cité page 124.)

Gino Cardarelli, Mario Palumbo, et Pacifico M. Pelagagge. Use of neural networks in modeling relations between exposure energy and pattern dimension in photolithography process [mos ics]. *Components, Packaging, and Manufacturing Technology, Part C, IEEE Transactions on*, 19 :290–299, 1996. ISSN 1083-4400. (Cité page 93.)

Alain Casali, Christian Ernst, Franck Gasnier, et Jamel Stephan. Extracting correlated sets using the chi-squared measurement within n-ary relations : an implementation. Dans *Proceedings of the European AEC/APC Conference*, 2007. (Cité page 28.)

Enrique Del Castillo. A variance-constrained proportional-integral feedback controller that tunes itself. *IIE Transactions*, 32 :479–491, 2000. (Cité page 74.)

Enrique Del Castillo. *Statistical Process Adjustment for Quality Control*. Wiley-Interscience, 1 édition, Avril 2002. ISBN 0471435740. (Cité pages 45, 46, 52, 56, 57, 71, 73, 76, 77, 120, 121, 122, 127 et 160.)

Enrique Del Castillo et Arnon M Hurwitz. Run-to-run process control : Literature review and extensions. *Journal of Quality Technology*, 29(2) :184–196, 1997. (Cité pages 47, 122 et 123.)

Yaw-Jen Chang, Yuan Kang, Chin-Liang Hsu, Chi-Tim Chang, et Tat Yan Chan. Virtual metrology technique for semiconductor manufacturing. Dans *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, pages 5289–5293, 2006. (Cité page 111.)

Chadi El Chemali, Jim Freudenberg, Matt Hankinson, et Joseph J. Bendik. Run-to-run critical dimension and sidewall angle lithography control using the prolith simulator. *Semiconductor Manufacturing, IEEE Transactions on*, 17 :388–401, 2004. ISSN 0894-6507. (Cité page 81.)

Chadi Elias El Chemali. *Run-to-Run Control for Wafer Patterning in Semiconductor Manufacturing*. PhD thesis, University of Michigan, 2002. (Cité pages 76, 81 et 120.)

Q. Peter He Christopher A. Bode, Jin Wang et Thomas F. Edgar. Run-to-run control and state estimation in high-mix semiconductor manufacturing. *Annual Reviews in Control*, 31 :241–253, 2007. (Cité page 125.)

David W. Clarke et Peter J. Gawthrop. A self-tuning controller. volume 122, pages 929–934, Mai 1975. (Cité page 79.)

Brian P. Conchieri, Steven M. Ruegsegger, et John J. Ellis-Monaghan. Effective channel length control using ion implant feed forward. (Cité page 90.)

Oscar D. Crisalle, Robert A. Soper, Duncan A. Mellichamp, et Dale E. Seborg. Adaptive control of photolithography. *AIChE Journal*, 38 :1–14, 1992. (Cité page 120.)

Veronica Czitrom et Patrick D. Spagon. *Statistical Case Studies for Industrial Process Improvement (ASA-SIAM Series on Statistics & Applied Probability)*. Society for Industrial Mathematics, pap/cas édition, 1987. ISBN 0898713943. (Cité page 30.)

Philippe de Larminat. *Automatique : commande des systèmes linéaires*. Editions HERMES, 1996. ISBN 2-86601-515-0. (Cité pages 47 et 49.)

W. Edwards Deming. *Out of the Crisis*. The MIT Press, Août 2000. ISBN 0262541157. (Cité pages 45 et 55.)

Edward R. Dougherty. *Probability and statistics for the engineering, computing, and physical sciences*. Prentice-Hall, Inc., 1990. (Cité page 52.)

David Drain. *Statistical Methods for Industrial Process Control*. Springer, 1st édition, Février 1997. ISBN 0412085119. (Cité page 30.)

Walter R. Ellingsen. Discussion of : Parameter estimation with closed-loop operating data. *Technometrics*, 18 :381–384, Novembre 1976. (Cité page 145.)

Clément-Marc Falinower. Introduction à la commande adaptative. Polycopié, Supélec, 4 2008. (Cité pages 127 et 145.)

Stacy K. Firth, W. Jarrett Campbell, Anthony J. Toprac, et Thomas F. Edgar. Just-in-time adaptive disturbance estimation for run-to-run control of semiconductor processes. *Semiconductor Manufacturing, IEEE Transactions on*, 19 :298–315, 2006. ISSN 0894-6507. (Cité pages 118 et 124.)

Shane Geary et Ronan Barry. Neural network-based run-to-run controller using exposure and resist thickness adjustment. volume 5044, pages 150–160, 2003. (Cité page 79.)

Emmanuel Godoy et Eric Ostertag. *Commande numérique des systèmes*. Ellipses Marketing, Juillet 2003. ISBN 2729817247. (Cité page 162.)

Emir Gurer, Tom Zhong, John Lewellen, et Reese Reynolds. Model-based adaptive process control : A cd-control example. *Solid state technology*, 41 :205–212, 1998. (Cité page 80.)

Romain Gwoziecki. *Etude de nouveaux concepts d'architectures drain-sources pour les technologies CMOS sub-0.18 microns*. PhD thesis, Institut national polytechnique de Grenoble, France, 1999. (Cité pages 89, 91 et 106.)

Andrew Habermas, Dongsung Hong, Matthew F. Ross, et William R. Livesay. 193-nm cd shrinkage under sem : modeling the mechanism. volume 4689, pages 92–101, Santa Clara, CA, USA, Juillet 2002. SPIE. (Cité page 16.)

Courtney K. Hanish. Run-to-run state estimation in systems with unobservable states. Dans *Proceedings of the European AEC/APC Conference*, 2006. (Cité pages 118 et 119.)

David E. Hardt et Tsz-Sin Siu. Cycle to cycle manufacturing process control, 2002. (Cité pages 47, 52 et 54.)

Thomas J. Harris, John F. MacGregor, et J. D. Wright. Self-tuning and adaptive controllers : An application to catalytic reactor control. *Technometrics*, 22 :153–164, Mai 1980. ISSN 00401706. (Cité page 79.)

Joost Van Herk, Jean De Caunes, Francois Pasqualini, et Stephane Hubac. Guidline to start a r2r control loop. Rapport technique, Alliance Crolles 2 : STMicroelectronics, Freescale and Philips, 2005. (Cité pages 26 et 27.)

Kenneth J. Hunt. A survey of recursive identification algorithms. *Transactions of the Institute of Measurement and Control*, 8 :273–278, 1986. (Cité pages 78, 120 et 122.)

Armann Ingolfsson. Run by run process control. Master's thesis, Massachusetts Institute of Technology, 1991. (Cité page 3.)

Nickhil H. Jakatdar, Xinhui Niu, John T. Musacchio, et Costas J. Spanos. In-situ metrology for deep-ultraviolet lithography process control. Dans *Proc. SPIE Metrology, Inspection, and Process Control for Microlithography XII, Bhanwar Singh ; Ed.*, volume 3332, pages 262–270, 1998. (Cité page 81.)

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D) :35–45, 1960. (Cité page 120.)

Chih-Ming Ke, Tsai-Sheng Gau, Pei-Hung Chen, Anthony Yen, Burn J. Lin, Tadashi Otaka, Takashi Iizumi, Katsuhiro Sasada, et Kazuo Ueda. Effect of various arf resist shrinkage amplitudes on cd bias. volume 4689, pages 997–1006, Santa Clara, CA, USA, 2002. SPIE. (Cité page 16.)

Robin De Keyser et James Donald. Model based predictive control in rtp semiconductor manufacturing. Dans *Control Applications, 1999. Proceedings of the 1999 IEEE International Conference on*, volume 2, pages 1636–1641 vol. 2, 1999. (Cité page 45.)

Robin De Keyser et Clara Mihaela Ionescu. The disturbance model in model based predictive control. volume 1, pages 446–451 vol.1, 2003. ISBN 1085-1992. (Cité page 45.)

Aftab A. Khan, James R. Moyne, et Dawn M. Tilbury. Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares. *Journal of Process Control*, 18 :961–974, Décembre 2008. (Cité page 111.)

Harry J. Levinson. *Principles of Lithography, Second Edition*. SPIE Publications, 2 édition, Février 2005. ISBN 0819456608. (Cité pages 6, 8, 9, 10, 102 et 144.)

Ramon C. Littell, George A. Milliken, Walter W. Stroup, et Russell Wolfinger. *SAS System for Mixed Models*. SAS Publishing, Juillet 1996. ISBN 1555447791. (Cité page 30.)

Alberto Luceño. Parameter estimation with closed-loop operating data under time varying discrete proportional-integral control. *Communications in Statistics - Simulation and Computation*, 26 :215, 1997. ISSN 0361-0918. (Cité page 145.)

Ming-Da Ma, Chun-Cheng Chang, David Shan-Hill Wong, et Shi-Shang Jang. Identification of tool and product effects in a mixed product and parallel tool environment. *Journal of Process Control In Press*, 2008. (Cité page 118.)

John MacGregor. A different view of the funnel experiment. *Journal of Quality Technology*, 22 :255–259, 1990. (Cité pages 45 et 55.)

John F. MacGregor, Thomas J. Harris, et J. D. Wright. Duality between the control of processes subject to randomly occurring deterministic disturbances and arima stochastic disturbances. *Technometrics*, 26 :389–397, 1984. (Cité pages 46, 68, 83 et 121.)

John F. MacGregor, J. D. Wright, et Huynh Man Hong. Optimal tuning of digital pid controllers using dynamic-stochastic models. *Industrial & Engineering Chemistry Process Design and Development*, 14 :398–402, Octobre 1975. (Cité page 68.)

Chris A. Mack. *Field Guide to Optical Lithography*. SPIE Publications, 2006. ISBN 0819462071. (Cité page 6.)

Séverine Marquet. *Maîtrise de la variabilité des procédés de fabrication par le développement de modèles de régulation*. PhD thesis, Université Joseph Fourier De Grenoble, 2008. (Cité page 92.)

Henry Mathieu. *Physique des semiconducteurs et des composants électroniques*. Dunod, 5e éd. édition, Avril 2001. ISBN 2100056549. (Cité page 21.)

Michael L. Miller, Abe Ghanbari, et Anthony J. Toprac. Impact of multi-product and -process manufacturing on run-to-run control. volume 3213, pages 138–146, Austin, TX, USA, 1997. SPIE. (Cité pages 3, 115 et 150.)

Kevin M. Monahan et Sadri Khalessi. Application of statistical models to decomposition of systematic and random error in low-voltage sem metrology. volume 1673, pages 36–41, San Jose, CA, USA, 1992. SPIE. (Cité page 16.)

James Moyne, Enrique Del Castillo, et Arnon M. Hurwitz. *Run-to-Run Control in Semiconductor Manufacturing*. CRC, 1 édition, Novembre 2000. (Cité page 45.)

James A. Mullins, W. Jarrett Campbell, Allen D. Stock, Abe Ghanbari, et Anthony J. Toprac. Evaluation of model predictive control in run-to-run processing in semiconductor manufacturing. Dans *Process, Equipment, and Materials Control in Integrated Circuit Manufacturing III*, volume 3213, pages 182–189, Austin, TX, USA, 1997. SPIE. (Cité page 121.)

John Musacchio. Run to run control in semiconductor manufacturing. Rapport technique, EECS Department, University of California, Berkeley, 1998. (Cité page 81.)

Kenneth R. Muske et James B. Rawlings. Model predictive control with linear models. *AIChE Journal*, 39 :262–287, 1993. (Cité page 44.)

Tsujii M. Oda K., Takeuchi H. et Ohba M. Practical estimator for self-tuning automotive cruise control. pages 2066–2071, 1991. (Cité page 124.)

Patrick Ozil. Plans d'expériences. Polycopié, INPG, 2 2002. (Cité page 93.)

Evan Palmer, Wei Pen, et Costas. J. Spanos. Control of photoresist properties : a kalman filter based approach. *Semiconductor Manufacturing, IEEE Transactions on*, 9 :208–214, 1996. ISSN 0894-6507. (Cité pages 80 et 120.)

Erwine Pargon. *Analyse des mécanismes mis en jeu lors de l'élaboration par gravure plasma de structures de dimensions déca-nanométriques : Application au transistor CMOS ultime*. PhD thesis, Université Joseph Fourier De Grenoble, 2004. (Cité page 10.)

Seong-Jin Park, Moon-Sang Lee, Sung-Young Shin, Kwang-Hyun Cho, Jong-Tae Lim, Bong-Su Cho, Young-Ho Jei, Myung-Kil Kim, et Chan-Hoon Park. Run-to-run overlay control of steppers in semiconductor manufacturing systems based on history data analysis and neural network modeling. *Semiconductor Manufacturing, IEEE Transactions on*, 18 :605–613, 2005. ISSN 0894-6507. (Cité page 93.)

Jens Parkum, Niels K. Poulsen, et Jan Holst. Selective forgetting in adaptive procedures. Dans *The 11th IFAC World Congress in Tallinn*, volume 3, pages 180–185, 1990. (Cité pages 123 et 124.)

Alexander J. Pasadyn et Thomas F. Edgar. Observability and state estimation for multiple product control in semiconductor manufacturing. *Semiconductor Manufacturing, IEEE Transactions on*, 18 :592–604, 2005. ISSN 0894-6507. (Cité page 118.)

Michael Quirk et Julian Serda. *Semiconductor Manufacturing Technology*. Prentice Hall, united states ed édition, Novembre 2000. ISBN 0130815209. (Cité page 14.)

Christopher J. Raymond, S. Sohail H. Naqvi, et John R. McNeil. Scatterometry for cd measurements of etched structures. volume 2725, pages 720–728, Santa Clara, CA, USA, Mai 1996. SPIE. (Cité page 17.)

Steve M. Ruegsegger, Aaron Wagner, Jim Freudenberg, et Dennis Grimard. Optimal feedforward recipe adjustment for cd control in semiconductor patterning. volume 449, pages 573–577, Gaithersburg, Maryland (USA), Novembre 1998. (Cité pages 90, 91 et 92.)

Steve M. Ruegsegger, Aaron Wagner, Jim Freudenberg, et Dennis Grimard. Feedforward control for reduced run-to-run variation in microelectronics manufacturing. *Semiconductor Manufacturing, IEEE Transactions on*, 12 :493–502, 1999. ISSN 0894-6507. (Cité pages 90 et 92.)

Steven M. Ruegsegger. *Feedforward Control for Reduced Run-to-Run Variation in Microelectronics Manufacturing*. PhD thesis, University of Michigan, 1998. (Cité pages 2, 15 et 90.)

Emanuel Sachs, Albert Hu, et Armann Ingolfsson. Run by run process control : combining spc and feedback control. *Semiconductor Manufacturing, IEEE Transactions on*, 8 :26–43, 1995. ISSN 0894-6507. (Cité page 3.)

Emanuel Sachs et Armann Ingolfsson. Stability and sensitivity of an ewma controller. *Journal of Quality Technology*, 25 :271–287, 1993. (Cité pages 3, 76 et 79.)

Steinar Saelid et Bjarne Foss. Adaptive controllers with a vector variable forgetting factor. Dans *Decision and Control, 1983. The 22nd IEEE Conference on*, volume 22, pages 1488–1494, 1983. (Cité page 124.)

Anne-Claire Salaun. Initiation au microscope electronique à balayage (meb). Notes de cours, Institut d'Electronique et de Télécommunications de Rennes, 2004. (Cité pages 15 et 16.)

Mario E. Salgado, Graham C. Goodwin, et Richard H. Middleton. Modified least squares algorithm incorporating exponential resetting and forgetting. *International Journal of Control*, 47(2) :477–491, 1988. (Cité page 123.)

Robert R. Schaller. Moore's law : past, present and future. *Spectrum, IEEE*, 34 :52–59, 1997. ISSN 0018-9235. (Cité pages 1 et 88.)

Matthew Sendelbach, Charles N. Archie, Bill Banke, Jason Mayer, Hideaki Nii, Pedro Herrera, et Matt Hankinson. Correlating scatterometry to cd-sem and electrical gate measurements at the 90-nm node using tmu analysis. volume 5375, pages 550–563, Mai 2004. (Cité page 19.)

Matthew Sendelbach, Andres Munoz, Kenneth A. Bandy, Dan Prager, et Merritt Funk. Integrated scatterometry in high-volume manufacturing for polysilicon gate etch control. Dans *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6152, pages 134–150, Avril 2006. (Cité pages 19 et 111.)

Sirish L. Shah et William R. Cluett. Recursive least square based estimation schemes for self-tuning controller. *Canadian Journal of Chemical Engineering*, 69 :89–96, Février 1991. (Cité pages 78 et 123.)

Thomas Skotnicki. Transistor mos et sa technologie de fabrication. *Techniques de l'Ingénieur*, E2 430, 2000. (Cité page 107.)

Brian E Stine, Student Member, Duane S Boning, et James E Chung. Analysis and decomposition of spatial variation in integrated circuit processes and devices. *IEEE Transactions on Semiconductor Manufacturing*, 10 :24–41, 1997. (Cité page 2.)

Karl Johan Åström et Björn Wittenmark. On self-tuning regulators. volume 9, pages 185–199. 1973. (Cité page 79.)

John D. Stuber. Multiloop photolithography control using hierarchical context information for apc models. Dans *Advanced Process Control and Automation. Edited by Hankinson, Matt ; Ausschnitt, Christopher P. Proceedings of the SPIE*, volume 5044, pages 75–82, 2003. (Cité pages 116 et 143.)

John D. Stuber, Francois Pagette, et Susan Tang. Device dependant run-to-run control of transistor critical dimension by manipulating photolithography exposure settings. Dans *Proceedings of the AEC/APC Symposium XII*, pages 1–21, 2000. (Cité page 80.)

John L. Sturtevant, Steven J. Holmes, Theodore G. Van Kessel, Philip C. Hobbs, Jerry C. Shaw, et Robert R. Jackson. Postexposure bake as a process-control parameter for chemically amplified photoresist. Dans *Proc. SPIE Integrated Circuit Metrology, Inspection, and Process Control VII, Michael T. Postek ; Ed.*, volume 1926, pages 106–114, Août 1993. (Cité page 82.)

Genichi Taguchi, Elsayed A. Elsayed, et Thomas C. Hsiang. *Quality Engineering in Production Systems*. Mcgraw-Hill College, Septembre 1988. (Cité page 48.)

Michel Tenenhaus. *La régression PLS : Théorie et pratique*. Technip, Août 1998. ISBN 2710807351. (Cité page 27.)

Hannu T. Toivonen. Variance constrained self-tuning control. *Automatica*, 19 :415–418, 1983. (Cité page 74.)

Anthony J. Toprac. Solving the high-mix control problem. Dans *Proceedings of the European AEC/APC Conference*, 2004. (Cité page 115.)

Fugee Tsung et Jianjun Shi. Integrated design of run-to-run pid controller and spc monitoring for process disturbance rejection. *IIE Transactions*, 31 :517–527, Mai 1999. (Cité page 71.)

Fugee Tsung, Huaiqing Wu, et Vijayan N. Nair. On the efficiency and robustness of discrete proportional-integral control schemes. *Technometrics*, 40 :214–222, 1998. (Cité pages 48, 55, 56, 59 et 76.)

Stéphane Tufféry. *Data Mining et statistique décisionnelle : l'intelligence dans les bases de données*. Editions Technip, Août 2005. (Cité page 28.)

Ardalan Vahidi, Maria Druzhinina, Anna Stefanopoulou, et Huei Peng. Simultaneous mass and time-varying grade estimation for heavy-duty vehicles. Dans *American Control Conference, 2003. Proceedings of the 2003*, volume 6, pages 4951–4956 vol.6, 2003. ISBN 0743-1619. (Cité page 124.)

Omer Arda Vanli. *Statistical Process Adjustment Problems in Short-Run Manufacturing*. PhD thesis, Pennsylvania State University, 2007. (Cité page 120.)

Omer Arda Vanli, Nital S. Patel, Mani Janakiram, et Enrique Del Castillo. Model context selection for run-to-run control. *Semiconductor Manufacturing, IEEE Transactions on*, 20 :506–516, 2007. ISSN 0894-6507. (Cité page 120.)

K. Minh Vu. Optimal setting for discrete pid controllers. *Control Theory and Applications, IEE Proceedings D*, 139 :31–40, 1992. (Cité pages 68 et 71.)

Jin Wang et Q. Peter He. A bayesian approach for disturbance detection and classification and its application to state estimation in run-to-run control. *Semiconductor Manufacturing, IEEE Transactions on*, 20 :126–136, 2007. (Cité page 136.)

Jin Wang, Q. Peter He, et Thomas F. Edgar. A general framework for state estimation in high-mix semiconductor manufacturing. Dans *American Control Conference, 2007. ACC '07*, pages 3636–3641, 2007. ISBN 0743-1619. (Cité pages 118 et 150.)

Jin Wang, Q. Peter He, et Thomas F. Edgar. State estimation in high-mix semiconductor manufacturingy. *Journal of Process Control In Press*, 2008. (Cité pages 118, 119, 123, 124, 136 et 140.)

N. Yoshitani et A. Hasegawa. Model-based control of strip temperature for the heating furnace in continuous annealing. *Control Systems Technology, IEEE Transactions on*, 6 :146–156, 1998. ISSN 1063-6536. (Cité page 124.)

École Nationale Supérieure des Mines
de Saint-Étienne

N° d'ordre: **535 M**

Nader JEDIDI

Applications of Advanced Process Control Techniques within the Semiconductor Manufacturing Industry

Speciality: **Microelectronics**

Abstract: With pattern dimensions decreasing in CMOS technology, chip performance is becoming increasingly sensitive to process variation, especially for the 130 nm node and below. Due to fluctuations in the fabrication process, variability in device parametric characteristics such as transistor drive currents is increased resulting in device yield degradation. To address this variability, we focused on the development of advanced process control tools, and in particular the design of run to run loops.

First a statistical analysis of the parametric variations, in both saturation and subthreshold regimes has been carried out, in order to identify the main process contributors. The biggest cause of parametric transistor variability and hence Yield loss is large gate line-width (CD) variation. Since the lot-to-lot component of the gate CD variations is significant, we considered a novel strategy in run-to-run control which is more adapted to the high-mix environment, namely the cooperative control. We developed an in-line estimator to estimate the states of each product and tool. We investigated two recursive identification algorithms: Kalman filter and selective forgetting least square. With reference to the simulations we have achieved using production data, we highlighted a real potential of using the Kalman filter scheme.

This thesis deals also with compensating for the residual lot-to-lot CD variations. A full-automated run-to-run feed-forward control scheme from Gate Etching to Pocket Implantation (FFE-I2) has been considered. Using feed-forward control, the measured deviation of the post-etch gate line-width is automatically compensated by adjusting the pocket dose. The implementation of the FFE-I2 in production allowed a 40% reduction of the lot-to-lot main parametric characteristics variations.

Keywords: **Advanced Process Control**, **Run-to-Run**, **Semiconductor Manufacturing**

École Nationale Supérieure des Mines
de Saint-Étienne

N° d'ordre : **535 M**

Nader JEDIDI

Titre de la thèse : **Applications de Techniques Avancées de Contrôle des Procédés en Industrie du Semi-conducteur**

Spécialité : **Microélectronique**

Résumé : Cette thèse porte sur le développement d'outils de contrôle avancé des procédés et leurs applications à l'industrie de fabrication des composants microélectroniques. L'accent est mis en particulier sur les boucles de régulation, connues sous le nom de boucles run-to-run dans la littérature anglo-saxonne.

Dans la première partie de ces travaux, l'objectif était d'identifier, à travers une analyse statistique multi-variée, les étapes de fabrication critiques, principalement responsables des variabilités temporelle et spatiale des performances électriques du dispositif, à savoir les courants de saturation, de fuite, et la tension de seuil des transistors courts (L=0.13µm). Les résultats obtenus montrent une forte variabilité lot à lot du courant de saturation et du courant de fuite, due, au premier ordre à la variabilité de la longueur de grille en poly-silicium.

Ce constat a été le point de départ pour le développement d'une nouvelle stratégie de régulation run-to-run mieux adaptée à un environnement de production *high-mix*, à savoir le contrôle coopératif. Un contrôleur coopératif s'appuie sur un algorithme d'identification récursif afin d'estimer en ligne les états des sources de variations qualitatives prises en compte, notamment le produit et l'équipement de fabrication. Dans ce cadre, nous avons simulé les performances de plusieurs estimateurs en ligne, en particulier le filtre de Kalman et les moindres carrés récursifs à oubli sélectif. Ces simulations montrent un avantage significatif en faveur du filtre de Kalman.

D'une façon complémentaire au développement du contrôleur coopératif, nous nous sommes intéressés aux régulations de compensation et, en particulier, de celle entre la gravure de la grille et l'implantation des poches. Elle vise à compenser la déviation de la longueur de la grille en ajustant la dose d'implantation des poches. L'objectif est alors de réduire la dispersion lot à lot de la longueur effective du canal. Le déploiement de cette boucle de compensation en production a permis de réduire de 40% les variations lot à lot des principales caractéristiques paramétriques.

Mots clefs : **Fabrication des composants microélectroniques, contrôle avancé des procédés, run-to- run**