Statistical Learning on Circular Domains For Advanced Process Control in Microelectronics

Espéran Padonou

NNT : 2016LYSEM009

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de
L'Ecole des Mines de Saint-Etienne

Ecole Doctorale N° 488
Sciences, Ingénierie, Santé

Spécialité de doctorat : **Mathématiques appliquées**

Soutenue publiquement le 13/05/2016 par :
Espéran Padonou

Apprentissage Statistique en Domaine Circulaire Pour la Planification de Contrôles en Microélectronique

---

Devant le jury composé de :

Gamboa Fabrice, Professeur, Institut de Mathématiques de Toulouse, Président

Iooss Bertrand, Chercheur Sénior HDR, Electricité de France, Rapporteur
Vicario Grazia, Professeur, Ecole polytechnique de Turin, Rapporteur
Reis Marco, Maitre de Conférences, Université de Coimbra Examinateur

Roustant Olivier, Professeur, Mines Saint – Étienne, Directeur de thèse
Blue Jakey, Maitre de Conférences, Mines Saint – Étienne, Co-encadrant
Duverneuil Hugues, Ingénieur – Manager, STMicroelectronics, Co-encadrant

Spécialités doctorales	Responsables :	Spécialités doctorales	Responsables
SCIENCES ET GENIE DES MATERIAUX	K. Wolski Directeur de recherche	MATHEMATIQUES APPLIQUEES	O. Roustant, Maître-assistant
MECANIQUE ET INGENIERIE	S. Drapier, professeur	INFORMATIQUE	O. Boissier, Professeur
GENIE DES PROCEDES	F. Gruy, Maître de recherche	IMAGE, VISION, SIGNAL	JC. Pinoli, Professeur
SCIENCES DE LA TERRE	B. Guy, Directeur de recherche	GENIE INDUSTRIEL	X. Delorme, Maître assistant
SCIENCES ET GENIE DE L'ENVIRONNEMENT	D. Graillot, Directeur de recherche	MICROELECTRONIQUE	Ph. Lalevée, Professeur

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

ABSI	Nabil	CR	Génie industriel	CMP
AUGUSTO	Vincent	CR	Image, Vision, Signal	CIS
AVRIL	Stéphane	PR2	Mécanique et ingénierie	CIS
BADEL	Pierre	MA(MDC)	Mécanique et ingénierie	CIS
BALBO	Flavien	PR2	Informatique	FAYOL
BASSEREAU	Jean-François	PR	Sciences et génie des matériaux	SMS
BATTON-HUBERT	Mireille	PR2	Sciences et génie de l'environnement	FAYOL
BEIGBEDER	Michel	MA(MDC)	Informatique	FAYOL
BLAYAC	Sylvain	MA(MDC)	Microélectronique	CMP
BOISSIER	Olivier	PR1	Informatique	FAYOL
BONNEFOY	Olivier	MA(MDC)	Génie des Procédés	SPIN
BORBELY	Andras	MR(DR2)	Sciences et génie des matériaux	SMS
BOUCHER	Xavier	PR2	Génie Industriel	FAYOL
BRODHAG	Christian	DR	Sciences et génie de l'environnement	FAYOL
BRUCHON	Julien	MA(MDC)	Mécanique et ingénierie	SMS
BURLAT	Patrick	PR1	Génie Industriel	FAYOL
CHRISTIEN	Frédéric	PR	Science et génie des matériaux	SMS
DAUZERE-PERES	Stéphane	PR1	Génie Industriel	CMP
DEBAYLE	Johan	CR	Image Vision Signal	CIS
DELAFOSSE	David	PR0	Sciences et génie des matériaux	SMS
DELORME	Xavier	MA(MDC)	Génie industriel	FAYOL
DESRAYAUD	Christophe	PR1	Mécanique et ingénierie	SMS
DJENIZIAN	Thierry	PR	Science et génie des matériaux	CMP
DOUCE	Sandrine	PR2	Sciences de gestion	FAYOL
DRAPIER	Sylvain	PR1	Mécanique et ingénierie	SMS
FAVERGEON	Loïc	CR	Génie des Procédés	SPIN
FEILLET	Dominique	PR1	Génie Industriel	CMP
FOREST	Valérie	MA(MDC)		CIS
FOURNIER	Jacques	Ingénieur chercheur CEA		CMP
FRACZKIEWICZ	Anna	DR	Sciences et génie des matériaux	SMS
GARCIA	Daniel	MR(DR2)	Génie des Procédés	SPIN
GAVET	Yann	MA(MDC)	Image Vision Signal	CIS
GERINGER	Jean	MA(MDC)	Sciences et génie des matériaux	CIS
GOEURIOT	Dominique	DR	Sciences et génie des matériaux	SMS
GONDRAN	Natacha	MA(MDC)	Sciences et génie de l'environnement	FAYOL
GRAILLOT	Didier	DR	Sciences et génie de l'environnement	SPIN
GROSSEAU	Philippe	DR	Génie des Procédés	SPIN
GRUY	Frédéric	PR1	Génie des Procédés	SPIN
GUY	Bernard	DR	Sciences de la Terre	SPIN
HAN	Woo-Suck	MR	Mécanique et ingénierie	SMS
HERRI	Jean Michel	PR1	Génie des Procédés	SPIN
KERMOUCHE	Guillaume	PR2	Mécanique et Ingénierie	SMS
KLOCKER	Helmut	DR	Sciences et génie des matériaux	SMS
LAFOREST	Valérie	MR(DR2)	Sciences et génie de l'environnement	FAYOL
LERICHE	Rodolphe	CR	Mécanique et ingénierie	FAYOL
MALLIARAS	Georges	PR1	Microélectronique	CMP
MOLIMARD	Jérôme	PR2	Mécanique et ingénierie	CIS
MOUTTE	Jacques	CR	Génie des Procédés	SPIN
NIKOLOVSKI	Jean-Pierre	Ingénieur de recherche	Mécanique et ingénierie	CMP
NORTIER	Patrice	PR1		SPIN
OWENS	Rosin	MA(MDC)	Microélectronique	CMP
PERES	Véronique	MR	Génie des Procédés	SPIN
PICARD	Gauthier	MA(MDC)	Informatique	FAYOL
PIJOLAT	Christophe	PR0	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR1	Génie des Procédés	SPIN
PINOLI	Jean Charles	PR0	Image Vision Signal	CIS
POURCHEZ	Jérémy	MR	Génie des Procédés	CIS
ROBISSON	Bruno	Ingénieur de recherche	Microélectronique	CMP
ROUSSY	Agnès	MA(MDC)	Génie industriel	CMP
ROUSTANT	Olivier	MA(MDC)	Mathématiques appliquées	FAYOL
STOLARZ	Jacques	CR	Sciences et génie des matériaux	SMS
TRIA	Assia	Ingénieur de recherche	Microélectronique	CMP
VALDIVIESO	François	PR2	Sciences et génie des matériaux	SMS
VIRICELLE	Jean Paul	DR	Génie des Procédés	SPIN
WOLSKI	Krzystof	DR	Sciences et génie des matériaux	SMS
XIE	Xiaolan	PR1	Génie industriel	CIS
YUGMA	Gallian	CR	Génie industriel	CMP

Mise à jour : 01/02/2016

Remerciements

Lorsqu'enfant, je mettais pieds à l'école primaire de Malanville au Bénin, il était peu probable que mes études s'achèvent par une thèse de doctorat à Saint-Étienne. A l'heure où j'écris ces lignes, je revois le film de ce parcours sinueux, fait de travail, de joie et de passion, de doutes et d'émotions. J'exprime ma profonde gratitude à tous ceux qui m'ont accompagné tout au long du chemin.

Je suis tout d'abord reconnaissant envers l'École des Mines de Saint-Étienne et STMicroelectronics qui m'ont donné l'opportunité de réaliser ce projet. En particulier, je remercie Olivier Roustant, mon directeur de thèse qui, par sa rigueur, son expérience et son humanité a été le véritable GPS de mon itiniraire de chercheur, depuis l'affaire *Holt-Winters* avec Michel Lutz jusqu'aux *Polars GPs*. Je dis également merci à mon manager Hugues Duverneuil, cet inlassable homme d'action et d'idées qui, par son optimisme et sa détermination a permis le déploiement industriel de ce travail. Je ne saurais non plus oublier les apports de Jakey Blue qui, dans un style didactique et ludique, m'a été d'un soutien inouï pour construire et aviver le pont entre les mathématiques et l'industrie microélectronique. Un grand merci à ceux qui ont contribué au développement de l'application *zeus*, dédiée à la mise en oeuvre de mes travaux chez STMicroelectronics. Il s'agit d'une part des super-stagiaires Igor Chidlovskii (mon petit noir) et Samiath Amadou (la khôlleuse) qui en ont développé le premier prototype, et d'autre part de Pierre Belin, Marc Mikolajczak et Aurelie Tallandier-ede qui en assurent l'implémentation et la perennité. J'attribue une mention spéciale à David Ginsbourger, Nicolas Durrande et aux autres membres du consortium ReDice, un groupe au sein duquel s'allient somptueusement rire et science.

J'ai une pensée spéciale pour ceux que j'ai eu le plaisir de rencontrer à Saint-Étienne, à Crolles et à Gardanne dans le cadre de mes travaux. Il me plait à ce stade d'emprunter quelques mots à Oscar Wilde pour peindre le tableau de l'immense bonheur que m'a procuré le temps passé avec eux: grâce à vous, je n'ai pas simplement rajouté trois années de thèse à ma vie, mais de la vie à mes années de thèse. Je regrette dejà les gâteaux de Mickael Binois, mon grand camarade de classe et de vie, friand de bissap et de manioc. De même, je conserve des souvenirs vivants des anecdotes d'Isabelle, des restaus avec JC, du "sourire Colgate" d'Hossein, du stress-relaxant d'Hassan, sans oublier les discussions passionnantes, passionnées et probablement passionelles avec Afafe. Je me souviendrai toujours de certains caractères d'exception: le style savant et taquin de Xavier^D, les bons plans et l'extraordinaire ouverture d'esprit de Rodolphe, la science et la conscience de Xavier^B, les khôlles d'Éric, les chocolats de Marianne et surtout la patiente-générosité de Christine. Je fais un clin d'œil à Didier, mon mutualiste des repas de midi, et un gros bisou à Mireille pour ses conseils et son assistance dont l'apparition des premiers fruits remonte à ma deuxième année d'études à Saint-Étienne. Par ailleurs, c'est avec un pincement au cœur que je quitte mes collègues de Crolles: le service ICT et le club des footballeurs. Puisqu'il m'est impossible de mentionner ici le nom de chacun, je fais un zoom sur Rémi Poinas mon frère Haoussa, Jérôme Henot alias Professeur Silla, Patrick Féraud et Ian Smith. J'accorde la grâce présidentielle à Anne-Sophie qui n'a pas voulu m'épouser et à Patrick Farouche dont les opinions footballistiques demeurent consternantes. Merci à Kadi (optimal update), JP, Eli, Paolo, Gabrielle, Alain, Claude, Dominique et Stéphane pour nos agréables et enrichissantes discussions.

A présent, je pense à mes soeurs Éméline, Larissa, Doris, Orelle et Gloria. Je vous ai volé du temps au profit des études et de l'affection au profit d'amis et amantes. La réalité est que je vous aime et que vous me manquez... A l'image de mes soeurs, je reste redevable à ma grande famille, large au sens africain, et au sein de laquelle j'ai eu les modèles comme Doris et Christelle qui m'ont fortement émulé. Par la même occasion, je remercie mes camarades de classe et d'associations. Spécial bisou à Arélyss, Tatiana et Barriath qui m'ont souteu pendant les moments les plus difficiles, aux garçons de la première compagnie, à Sinath, Roland, Brian, Martial et Lorens. Je n'oublie pas Orianne et Pierre-Channel pour leur forte implication dans l'organisation de ma soutenance et je leur souhaite, quand le moment sera venu, de porter mieux que moi le flambeau du Quartier Latin de l'Afrique. J'embrasse chaleureusement les familles Idohou, Taffin, Guélen, Forget, Marcot, Margarit et Quinto.

Pour finir, je voudrais honorer Odon Vallet et sa Fondation dont l'action m'a permis de briser bien de barrières sociales aux moments où la seule force de mes parents ne suffisait plus pour financer mes études supérieures. Je remercie aussi Benjamin Royannez pour son écoute, sa disponibilité durant notre période stéphanoise et son action au sein de la Fondation. Dans la même perspective, je suis reconnaissant au complexe scolaire Sainte Félicité où mes études secondaires ont entièrement été prises en charge: un modèle d'une extrême rareté sur le continent africain.

Merci à tous !

À mon père et à ma mère qui,
loin des turpitudes de ce monde,
ont tout misé sur mon éducation.

Contents

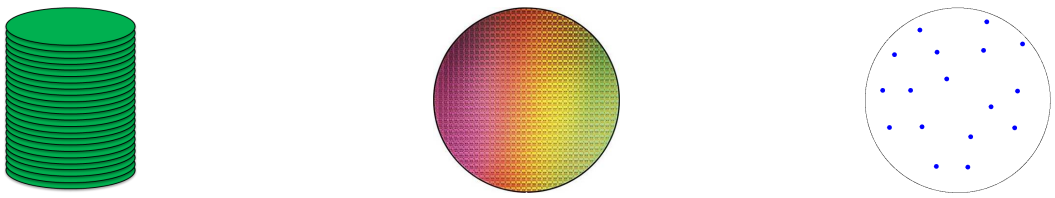
Chapter 1

From industrial needs to contributions in applied mathematics

In the era of the Internet of Things (IoT), electronic devices revolutionise our daily lives and play a prominent role in all economic sectors such as telecommunications, healthcare, automotive and military applications. The driven factors of this boosting change are technological advances in microelectronics, governed by an exponential speed known as Moore's law since 1965. Microelectronics is related to the study and production of very small electronic components, called integrated circuits (IC), and embedded in modern appliances. Behind the technological achievements in this sector, important economic issues are arisen. For instance, as transistors can shrunk to a size around few units of nanometers thanks to FinFETs and FD-SOI technologies [1], their production involves expensive investments and costly controls.

This thesis is part of the overall framework of cost reduction in quality control. Driven by the needs of our industrial partner, STMicroelectronics, our research project consists in developing original probabilistic models for spatial and temporal datasets. In the following sections, we introduce these contributions through the industrial framework.

1.1 Industrial needs in spatial and temporal modelling



(a) A lot (25 wafers). (b) A silicon wafer. (c) 17 sites on a wafer.

Figure 1.1 – The three levels of quality control in semiconductor industry.

The IC production consists in building functional modules, layer-by-layer, on a circular slice of a semiconductor material called wafer (Figure 1.1b). The quality of a wafer is controlled

[1] FinFETs (FIN Field Effect Transistor) and FD-SOI (Fully Depleted Silicon On Insulator) are the two latest transistor architectures, expected to shrunk transistors size around few nanometers.

through measurements at a limited set of locations, which are commonly called sites in microelectronics and design points in statistics (Figure 1.1c). Due to logistical constraints, wafers are processed by batches called lots (Figure 1.1a). Production and control strategies are then designed and executed with respect to lots, and several quality indicators are aggregated to the lot level. However, as illustrated in Figure 1.2, there are several factors of non-homogeneity within one lot. Therefore, modern quality control systems are based on

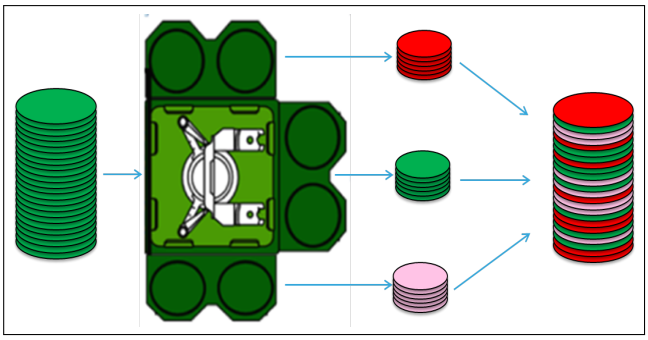


Figure 1.2 – A source of variability within one lot.

wafers since they represent the central elements in manufacturing. The issue is then to assess and regulate a production system over time, based on the measurements on wafers. For this purpose, Statistical Process Control (SPC) methods are intensively used. A common practice is to monitor the time-series of the successive mean values, estimated wafer by wafer. In practice, such procedures are often applied under the assumption that the measurements are independent and identically distributed.

In practical operating conditions, the assumption of independent and identically distributed measurements is proved to be unrealistic (Figure 1.3). Indeed, there exists a spatial de-

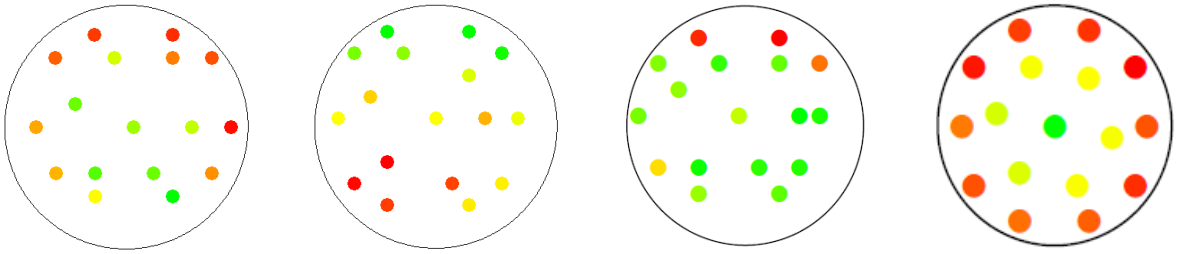
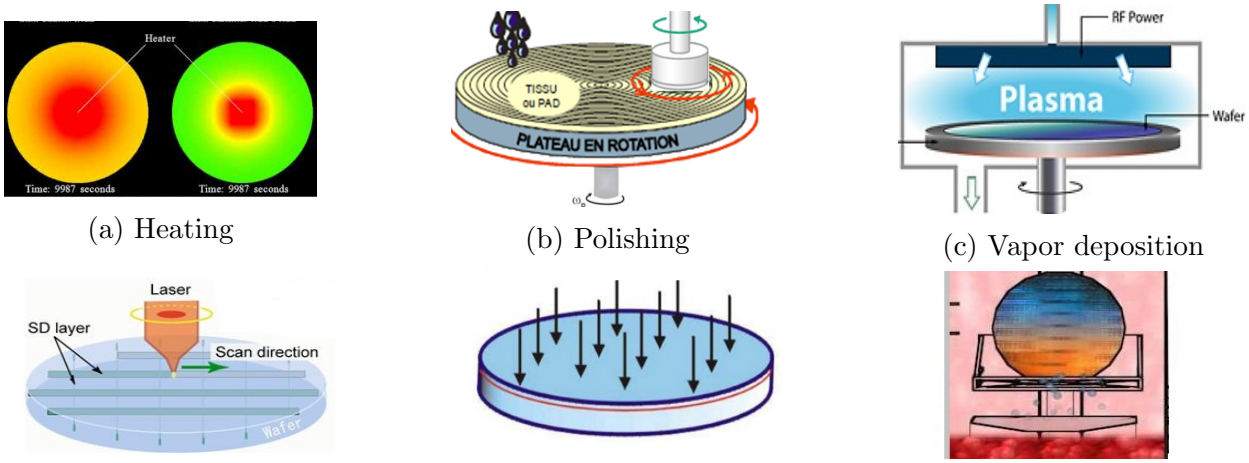


Figure 1.3 – Different examples of measurements over wafers in semiconductor industry.

pendence among measurements. Not only should quality be controlled over time, but also it must be assessed spatially due to this dependence. Given this, we address the issue of quality control as a profile monitoring problem [111, 80]. The procedure allows to deal with spatial and temporal aspects separately. The spatial problem, consisting in modelling the relationship between design points and measurements is then tackled with response surface methods, including statistical models and designs of experiments. The temporal component is treated with SPC tools, including control charts and time-series models. In each part, visualization and interpretation tools, which are essential for practitioners, are provided.

1.2 From physical processes to spatial patterns

In semiconductor industry, production involves various manufacturing processes. Taking a common heating process as an example, when a heating source is placed in the center of a wafer (Figure 1.4a), thermal conduction is governed by heat equation, resulting in a radial variation of temperature. As a consequence, the physical characteristics of the material are different from the center to the boundary of the wafer. A radial pattern is also observed in the vapor deposition process shown in Figure 1.4c. The result is consistent with the underlying physical principle, namely Fick's law of diffusion. Conversely, the laser processing shown in Figure 1.4d generates the pattern which varies from top to bottom and from left to right. This is due to horizontal and vertical movements of the scanner, while its lens gets warm over time. Through these examples, we understand why quality characteristics vary spatially over wafers. From there comes the need to model wafers spatial patterns for the purpose of quality control and regulation.



(a) Heating (b) Polishing (c) Vapor deposition

(d) Laser processing (e) Doping (f) Drying

Figure 1.4 – Examples of manufacturing processes in microelectronics, and resulting patterns. *Source: STMicroelectronics (b, d and e); YouTube (a and f); SlideShare (c).*

To model wafers spatial patterns, an ideal solution should be to use the physical principle underlying the manufacturing process. However, in complex systems like semiconductor fabs, there are hundreds of production steps. Therefore, statistical approaches are employed to recover the wafer pattern from a limited set of points, sampled for quality check. In other words, the idea is to fit a response surface model over a circular domain.

In the general context of response surface models, a wide variety of tools is available: polynomials, splines, radial basis functions, neural networks, Kriging, etc. In the particular case of circular domains, Zernike polynomials [118, 89] and harmonic basis functions [102] are the most famous and widely applied. Furthermore, seen as a subset of the complex plane, the disk is also endowed with algebra inherited from complex analysis. As an example, Fourier expansions using the Poisson kernel[2] are useful to model datasets over circular domains (see e.g. [66], Chapter 7). Among the available methods, Zernike polynomials are attractive due to their interpretability, and Kriging models arouse our interest for their capacity to quantify

[2]Kernel in the sense of Fourier series. The Poisson kernel extends functions defined on the circle to harmonic functions on the disk.

uncertainty. This research is then motivated to develop novel methods in the same fashion.

When considered individually, the physics of each manufacturing process may be modelled in Kriging. For instance, using harmonic covariance functions is proven to be optimal for heating problems [40]. Rather than specificities, our work is focused on more general methods. Therefore, in parallel to investigations on statistical models, a review of manufacturing processes in microelectronics revealed two main families of technologies. The first group $\mathcal{G}_1$ leads to variations aligned with Euclidean directions (Figures 1.4d, 1.4e and 1.4f), and the second group $\mathcal{G}_2$ results in radial and angular patterns (Figures 1.4a, 1.4b and 1.4c). Different datasets, that seemed to correspond to these two groups were studied in [89], based on Kriging. Traditional Kriging models are formulated with the Euclidean distance and correspond to $\mathcal{G}_1$. They may lead to poor results for datasets from $\mathcal{G}_2$. Furthermore, Kriging models are usually formulated with respect to a Cartesian basis. A priori, the choice of such coordinate system is arbitrary when the input domain is circular. Potentially better solutions would consist in selecting a basis according to the dataset.

Contributions to spatial models

Regarding the spatial pattern modelling, our contributions are four-fold. In Chapter 2, the main results on Zernike polynomials are reviewed. Their orthogonality property, formulated with the uniform measure over the disk, is extended to the uniform measure in the space of polar coordinates. The resulting functions are more suitable for patterns of type $\mathcal{G}_2$. We also propose a data-driven solution to select a Cartesian basis over the disk. Based on the geostatistical concept of geometric anisotropy [4, 48], it allows to detect the main direction of variations. In Chapter 3, we introduce polar Gaussian processes. Defined in the space of polar coordinates to include radial and angular correlations in Kriging predictions, they lead to a significant improvement when the manufacturing process is of type $\mathcal{G}_2$. Polar Gaussian processes are also generalized to hyperballs, corresponding in computer experiments to a directional input in higher dimension. Finally in Chapter 4, we investigate the Sobol decomposition of Kriging models, with a focus on polar Gaussian processes. Derived from the properties of centred kernels, the Sobol decomposition provides a framework to interpret and visualize Kriging based response surfaces. In particular, it allows to quantify the importance of radial and angular effects. Furthermore, they outperform standard Kriging models.

Contributions in designs of experiments

In the area of designs of experiments, our research was focused on optimal strategies for Kriging models over the disk. In Chapter 5, we introduce maximin Latin cylinders in order to reproduce the properties of Latin hypercubes in the space of polar coordinates. This family of designs is suitable to learn polar Gaussian processes and can be adapted to fill the disk for general situations. In Chapter 6, IMSE-optimal designs are investigated in a static setting and the key properties are numerically studied. Then, we develop in Chapter 7 a sequential procedure to relocate design points. The resulting dynamical design of experiments is proved to be convergent.

1.3 Statistical Process Control for spatial and temporal data

Two main kinds of temporal datasets are under focus in this thesis. The first group originates from the successive spatial models that are developed for process control. The stability of the resulting multivariate time-series is checked over time, based on SPC tools. The second group involves hundreds of indicators, quantifying industrial performance. These indicators consist of daily, weekly or monthly variables, collected in production and IT databases. Because they describe the manufacturing and IT activities, the resulting time-series depend on external factors such as sales growth, outstanding, seasonality, etc. The challenge is to find a general procedure to monitor such diversified time-series.

SPC for spatial data

This part of our thesis falls within the general framework of profile monitoring. In the special case of semiconductor industry, the issue was addressed by [37] who used thin-plate splines to detect abnormal profiles whereas [10] monitored the spatial variance over wafers via the so-called spatial variance spectrum. Since spatial patterns are modelled with Zernike polynomials and Kriging in our thesis, we focus on profile monitoring based on these models. Although several studies were dedicated to profile monitoring based on regression models, Zernike polynomials received few attention. To deepen the descriptive analysis conducted by [89], Chapter 8 presents a control chart for spatial patterns represented in terms of Zernike polynomials. We also address profile monitoring based on Gaussian process parameters. This issue seems not to be addressed in the literature. The difficulty comes from non-Gaussianity of Kriging parameters, which are also harder to interpret than regression coefficients. The approaches found in the literature are focused on the monitoring of deviations from a target profile [19], or machine learning methods such as decision trees and variable selection [16]. We first consider Kriging parameters themselves. Control charts are then proposed for the different Gaussian process models implemented in the thesis. A monitoring procedure is also developed for spatial variance, based on Sobol indices. For the sake of interpretation in industry, spatial patterns are predicted and diagnosed in Chapter 9, based on continuous and categorical predictors, representing manufacturing parameters.

SPC for temporal data

The main requirement of the monitoring of performance indicators is to be applicable to a wide variety of observed time-series. To distinguish normal operations from exceptional events, a standard procedure is to model time series, and then monitor the model residuals [76]. A Holt-Winters smoothing was previously implemented since this algorithm is easy to manage by non-statisticians [70, 69]. Despite a good detection rate, the procedure generates too many false alarms, due to outliers and structural changes. In order to address both kinds of problems, we develop a robust and adaptive control chart in Chapter 10. The method embeds a detection of structural breaks in the robust smoothing introduced by [39, 21].

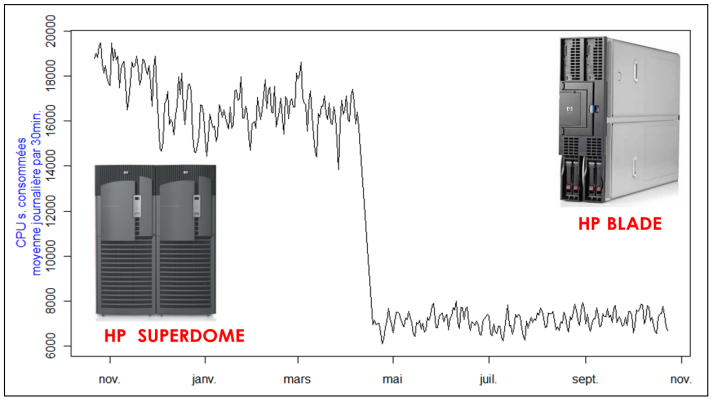


Figure 1.5 – An example of structural change (adapted from [70]).

Résumé en Français

Motivés par des besoins en industrie microélectronique, nos travaux apportent des contributions en modélisation probabiliste de données spatiales, et en maîtrise statistique de procédés.

Le problème spatial a pour spécificité d'être posé sur un domaine circulaire. Il se représente par un modèle de krigeage dont la partie déterministe est constituée de polynômes orthogonaux et la partie stochastique d'un processus gaussien. Traditionnellement définis avec la norme euclidienne et la mesure uniforme sur le disque, ces choix n'exploitent pas les informations a priori sur les procédés d'usinage. Pour tenir compte des mécanismes de rotation ou de diffusion à partir du centre, nous formalisons les processus gaussiens polaires sur le disque. Ces processus intègrent les corrélations radiales et angulaires dans le modèle de krigeage, et en améliorent les performances dans les situations considérées. Ils sont ensuite interprétés par décomposition de Sobol et généralisés en dimension supérieure. Des plans d'expériences sont proposés dans le cadre de leur utilisation. Au premier rang figurent les cylindres latins qui reproduisent en coordonnées polaires les caractéristiques des hypercubes latins.

Pour intégrer à la fois les aspects spatiaux et temporels du problème industriel, la maîtrise statistique de procédé est abordée en termes d'application de cartes de contrôle aux paramètres des modèles spatiaux. De cartes adaptées au suivi temporel des paramètres de Krigeage sont proposées. Pour finir, les séries temporelles contrôlées comportent parfois des données atypiques et des changements structurels, sources de biais en prévision, et de fausses alarmes en maitrise de risques. Ce problème est traité par lissage robuste et adaptatif.

PART I

Response Surface Models on Circular Domains

Chapter 2

Zernike polynomials and Kriging

In this chapter, we consider a physical or a computer experiment on the unit disk $\mathcal{D}$, represented with Cartesian or polar coordinates: $\mathcal{D} = \{(x,y) \in \mathbb{R}^2, x^2+y^2 \leq 1\} = \{(\rho\cos\theta, \rho\sin\theta), \rho \in [0,1], \theta \in \mathbb{S}\}$, where $\mathbb{S} = \mathbb{R}/2\pi\mathbb{Z}$ is the unit circle. We call $\mathbf{X} = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right)$ the design points, and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ the response values. Two regression methods are considered to model $\mathbf{Y}$ given $\mathbf{X}$. The first one uses Zernike polynomials and the second one, known as Kriging, is based on Gaussian processes.

2.1 Zernike polynomials

2.1.1 Historical context and key properties

In 1934, the physics community was limited by several computational issues. Particularly in optics, people had to model complex images such as superimposed fringes, displayed on a circular support area, representing a lens or an eye's pupil. This involved the computation of dozens of polynomial's coefficients whereas modern computers did not exist. Moreover, among the estimated coefficients, only few were needed. In this context, the challenge was to find a set of orthogonal functions over the disk to allow a computation of coefficients, independently from one another. Within the framework of regression models, an interesting question is then to find $(P_k)_k$, a basis of orthogonal polynomials in x and y over $\mathcal{D}$.

Denote $L^2(\mathcal{D})$ the space of square-integrable complex-valued functions over $\mathcal{D}$, with the scalar product:

$$\langle f, g\rangle = \int_{\mathcal{D}} f(x,y)\,\overline{g(x,y)}dxdy = \int_{\mathcal{D}} f\big(\rho\cos(\theta), \rho\sin(\theta)\big)\overline{g\big(\rho\cos(\theta), \rho\sin(\theta)\big)}\rho d\rho d\theta \quad (2.1)$$

When the $(P_k)_k$'s are unit vectors, the orthogonality condition is written:

$$\int_{\mathcal{D}} P_k(x,y)\,\overline{P_{k'}(x,y)}dxdy = \delta_{k,k'} \quad (2.2)$$

where δ denotes the kronecker symbol (1 if $k = k'$ and 0 otherwise). There exist infinite sets of polynomial basis which satisfy condition 2.2. Indeed, applying the Gram–Schmidt orthogonalization process to $(1, x, y, xy, x^2, y^2, \ldots)$ will lead to different orthogonal sets, depending on the initialization of the algorithm. However, in practice, a natural choice is ascending and lexicographic degrees: $1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3, \ldots$ When modelling

over the disk, another interesting property is independence with respect to the coordinates system. Relevant polynomials over the disk should have the same form after rotating the Cartesian basis, i.e.

$$P_k(x,y) = G(\theta) P_k(x',y') \tag{2.3}$$

where G is a 2π-periodic function in θ satisfying $G(0) = 1$ and (x', y') denotes the transform of (x, y) by rotation with angle θ:

$$\begin{cases} x' = x\cos(\theta) - y\sin(\theta) \\ y' = x\sin(\theta) + y\cos(\theta) \end{cases}$$

Proposition 2.1.1. *If the complex-valued polynomial basis $(P_k(x,y))_k$ satisfies 2.2 and 2.3, then*

$$P_k(x,y) = r_k(\rho) g_k(\theta)$$

where $g_k(\theta) = e^{im\theta}$ with $m \in \mathbb{Z}$, and $r_k(\rho)$ is of the form:

$$r_{n,m}(\rho) = a_m \rho^m + a_{m+2}\rho^{m+2} + \cdots + a_n \rho^n \tag{2.4}$$

where n has the same parity as m. In addition, the r_k's are orthogonal polynomials over $[0,1]$ with respect to the inner product $\langle f, g\rangle = \int_0^1 f(\rho) g(\rho) \rho d\rho$.

Proof. (Adapted from [106])
A necessary condition on G is obtained from Equation 2.3, by considering a rotation with angle $\theta_1 + \theta_2$, which a equivalent to two successive rotations with angles θ_1 and θ_2. By noting (x', y') the image of (x, y) by the rotation with angle θ_1, we have:

$$\begin{aligned} G(\theta_1 + \theta_2) P_k(x,y) &= G(\theta_2) P_k(x',y') \\ &= G(\theta_2) G(\theta_1) P_k(x,y) \end{aligned}$$

Therefore, $G(\theta_1 + \theta_2) = G(\theta_1) G(\theta_2)$. After a differentiation with respect to θ_1, setting $\theta_1 = 0$ shows that $\theta_2 \mapsto G(\theta_2)$ is a solution of the differential equation:

$$f'(u) = \alpha f(u)$$

Since G is 2π-periodic and satisfies $G(0) = 1$, it is of the form

$$G(\theta) = e^{im\theta}, \quad m \in \mathbb{Z}$$

From there, setting $y' = 0$ in Equations 2.3 leads to $P_k(x,y) = e^{im\theta} P_k(\rho, 0)$. It follows that $P_k(x,y) = r_k(\rho) e^{im\theta}$ with $m \in \mathbb{Z}$ and $r_k(\rho) = P_k(\rho, 0)$. Let n be the degree of $r_k(\rho)$. Therefore, there exist $a_0, \ldots, a_n \in \mathbb{C}$ such that $r_k(\rho) = \sum_{l=0}^n a_l \rho^l$. Then

$$\begin{aligned} P_k(x,y) &= \left(\sum_{l=0}^n a_l \rho^l\right) e^{im\theta} = \left(\sum_{l=0}^n a_l \rho^l\right) \left(\frac{x+iy}{\rho}\right)^m \\ &= \sum_{l=0}^n a_l \rho^{l-m} (x+iy)^m \\ &= \sum_{l=0}^n a_l \left(x^2+y^2\right)^{\frac{l-m}{2}} (x+iy)^m \end{aligned}$$

Thus $P_k(x,y)$ is a polynomial if and only if $l-m \in 2\mathbb{N}$ for $l = 0, 1 \dots n$. As a consequence, $n-m \in 2\mathbb{N}$ and $r_k(\rho)$ is of the form $r_{n,m}(\rho) = a_m\rho^m + a_{m+2}\rho^{m+2} + \cdots + a_n\rho^n$.

The orthogonality of the $r_{n,m}$'s is derived by separability of integrals over $[0,1]\times[0,2\pi]$. Based on this property, the $r_{n,m}$'s are uniquely defined (up to a constant) by applying the Gram–Schmidt process to the stair-step polynomials $\rho^m, \rho^{m+2}, \rho^{m+4}, \dots$ □

Back to the framework of real valued functions, we use the conventional notation $g_m(\theta) = \cos(m\theta)$ for real parts, and $g_{-m}(\theta) = \sin(m\theta)$ for imaginary parts, with $m \in \mathbb{N}$. The resulting polynomials are then of the form $P_n^m(x,y) = r_{n,m}(\rho)\cos(m\theta)$ or $P_n^{-m}(x,y) = r_{n,m}(\rho)\sin(m\theta)$.

2.1.2 Definition

Zernike polynomials are a sequence of orthogonal functions of $L^2(\mathcal{D})$. Indexed by two integers n and m such that $n-m \in 2\mathbb{N}$, each Zernike polynomial is either odd or even, and conventionally noted Z_n^m or Z_n^{-m} according to this parity. The orthogonality condition is:

$$\int_{\mathcal{D}} Z_n^m(\rho,\theta)\, Z_{n'}^{m'}(\rho,\theta)\,\rho d\rho d\theta = \frac{\epsilon_m}{2n+2}\delta_{n,n'}\delta_{m,m'} \tag{2.5}$$

where ϵ_m is the Neumann factor ($\epsilon_0 = 2$ and $\epsilon_m = 1$ if $m \geq 1$), and δ denotes the kronecker symbol ($\delta_{k,k'} = 1$ if $k = k'$ and 0 otherwise). In polar coordinates, Zernike polynomials are the product of a radial polynomial $r_n^m(\rho)$ by an angular function $g_m(\theta)$:

$$P_n^m\Big(\rho\cos(\theta), \rho\sin(\theta)\Big) = r_{n,m}(\rho)\, g_m(\theta) \tag{2.6}$$

with:

$$r_{n,m}(\rho) = \sum_{k=0}^{\frac{n-|m|}{2}} (-1)^k \binom{n-k}{k}\binom{n-2k}{\frac{n-|m|}{2}-k}\rho^{n-2k} \tag{2.7}$$

$$g_m(\theta) = \begin{cases}\cos(m\theta) & \text{if } m \geq 0,\\ \sin(m\theta) & \text{if } m < 0\end{cases} \tag{2.8}$$

$m=0$	$m=1$	$m=2$
$r_0^0 = 1$	$r_1^1 = \rho$	$r_2^2 = \rho^2$
$r_2^0 = 2\rho^2 - 1$	$r_3^1 = 3\rho^3 - 2\rho$	$r_4^2 = 4\rho^4 - 3\rho^2$
$r_4^0 = 6\rho^4 - 6\rho^2 + 1$	$r_5^1 = 10\rho^5 - 12\rho^3 + 3\rho$	$r_6^2 = 15\rho^6 - 20\rho^4 + 6\rho^2$

Table 2.1 – Analytical expressions of Zernike radial polynomials.

The orthogonality of Zernike polynomials is due to the orthogonality of the $r_{n,m}$'s over $[0,1]$ and the orthogonality of the g_m's over $[0,2\pi]$.

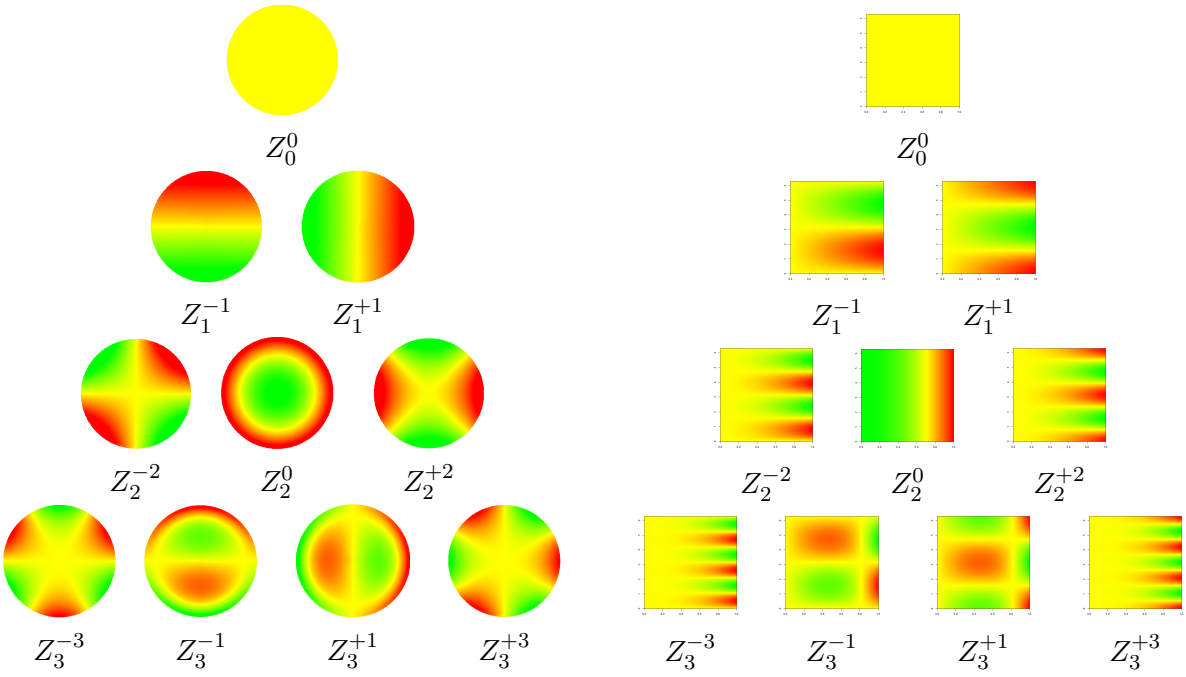


Figure 2.1 – Color representation of ten Zernike polynomials: y vs x (left), and θ vs ρ (right).

Initially used for image analysis in the conception of the phase contrast microscope, Zernike polynomials meet a wide variety of applications. They serve as shape descriptors for strain maps in mechanics, and for atmospheric turbulence in fluid dynamics. They also allow to classify breast cancers, and are used in microelectronics to model spatial patterns of wafers [89]. In Figure 2.1 are provided color representations for the first ten Zernike polynomials. They exhibit the following symmetry and rotation invariance properties:

$$Z_n^m(\rho, -\theta) = \text{sign}(m) Z_n^m(\rho, \theta) \tag{2.9}$$

$$Z_n^m(\rho, \theta + \pi) = (-1)^m Z_n^m(\rho, \theta) \tag{2.10}$$

$$Z_n^m\left(\rho, \theta + \frac{2k\pi}{m}\right) = Z_n^m(\rho, \theta), k \in \mathbb{Z} \tag{2.11}$$

From now on, we will use the normalized version of Zernike polynomials to benefit from orthonormality properties: $\frac{Z_n^m}{\|Z_n^m\|}$, with $\| Z_n^m \|^2 = \frac{\epsilon_m}{2n+2}$.

2.1.3 Measure modification

Throughout this thesis, we will deal with some responses whose interpretation are related to polar coordinates. Keeping this in mind, we aim at finding a set of orthogonal functions with respect to the uniform measure over the space $[0,1] \times [0, 2\pi]$ of polar coordinates:

$$\int_{\mathcal{D}} P_k(\rho\cos(\theta), \rho\sin(\theta)) \overline{P_{k'}(\rho\cos(\theta), \rho\sin(\theta))} d\rho d\theta = \int_{\mathcal{D}} P_k(x,y) \overline{P_{k'}(x,y)} \frac{dxdy}{\sqrt{x^2+y^2}} = \delta_{k,k'} \tag{2.12}$$

There exist an infinite sets of basis which satisfy Condition 2.12. As in the case of Zernike polynomials, we focus on functions which meet simultaneously Conditions 2.12 and 2.3 to take into account the geometry of the disk.

Proposition 2.1.2. *If the polynomial basis $(P_k(x,y))_k$ satisfies 2.12 and 2.3, then*

$$P_k(x,y) = \tilde{r}_k(\rho) g_k(\theta)$$

where $(\tilde{r}_k(x,y))_k$ is a polynomial basis over $[0,1]$ and $g_k(\theta) = e^{im\theta}$ with $m \in \mathbb{Z}$. Furthermore, the r_k's are of the form 2.4 and orthogonal with respect to the inner product $\langle f, g\rangle = \int_0^1 f(\rho)\, g(\rho)\, d\rho$.

Proof. By noting that the difference between Propositions 2.1.1 and 2.1.2 comes from the integration measure, the proof is straightforward. It remains to find a basis $(\tilde{r}_k(x,y))_k$ of the form $\tilde{r}_{n,m}(\rho) = a_m\rho^m + a_{m+2}\rho^{m+2} + \cdots + a_n\rho^n$ with $n - m \in 2\mathbb{N}$, and that are orthogonal with respect to $\langle f, g\rangle = \int_0^1 f(\rho)\, g(\rho)\, d\rho$. □

Construction with Legendre polynomials

Legendre polynomials are orthogonal functions over $[-1,1]$ with respect to the inner product $\langle f, g\rangle = \int_{-1}^1 f(x)\, g(x)\, dx$. They are defined as:

$$L_n(x) = \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k}^2 (x-1)^{n-k} (x+1)^k \tag{2.13}$$

So, the shifted Legendre polynomials $\tilde{L}_n(x) = L_n(2x-1)$ represent an orthogonal basis over $[0,1]$ with respect to $\langle f, g\rangle = \int_0^1 f(x)\, g(x)\, dx$. Then, the family $\left(\tilde{L}_n(\rho) g_m(\theta)\right)_{n-m\in 2\mathbb{N}}$ meets Conditions 2.12 and 2.3. However, these functions are not polynomials in x and y because Legendre polynomials are neither odd, nor even, such that $\tilde{L}_n(\rho)$ is not of the form 2.4. Among drawbacks, $\tilde{L}_n(\rho) g_m(\theta)$ is not always differentiable at the center of the disk.

Construction by orthogonalization

To obtain orthogonal polynomials in the sense of Conditions 2.12 and 2.3, the radial terms corresponding to a given $m \in \mathbb{N}$ must be of form 2.4: $\tilde{r}_{n,m}(\rho) = a_m\rho^m + a_{m+2}\rho^{m+2} + \cdots + a_n\rho^n$, with $n - m \in 2\mathbb{N}$. Then, $\tilde{r}_{n,m}$ is recursively defined by Gram-Schmidt orthogonalization:

$$\begin{aligned}
\tilde{r}^m_m(\rho) &= \rho^m \\
\tilde{r}^m_{m+2}(\rho) &= \rho^{m+2} - \frac{\langle \rho^{m+2}, \tilde{r}^m_m\rangle}{\langle \tilde{r}^m_m, \tilde{r}^m_m\rangle} \tilde{r}^m_m(\rho) \\
\tilde{r}^m_{m+4}(\rho) &= \rho^{m+4} - \frac{\langle \rho^{m+4}, \tilde{r}^{m+2}_m\rangle}{\langle \tilde{r}^{m+2}_m, \tilde{r}^{m+2}_m\rangle} \tilde{r}^{m+2}_m(\rho) - \frac{\langle \rho^{m+4}, \tilde{r}^m_m\rangle}{\langle \tilde{r}^m_m, \tilde{r}^m_m\rangle} \tilde{r}^m_m(\rho) \\
&\cdots
\end{aligned}$$

Based on this recurrence relation, we obtain the family P^m_n of orthogonal polynomials $\tilde{r}^m_n(\rho) g_m(\theta)$, represented in Figure 2.2. Remark that given $n \in \mathbb{N}$, we have $P^n_n = Z^n_n$ and $P^{-n}_n = Z^{-n}_n$ by construction: the initialization of the Gram-Schmidt process is the same as for Zernike.

Proposition 2.1.3. *Up to a constant, there exists a unique polynomial basis, with increasing degrees in x and y, that satisfies 2.2 and 2.3. The same goes for Conditions 2.12 and 2.3.*

$m=0$	$m=1$	$m=2$
$\tilde{r}_0^0 = 1$	$\tilde{r}_1^1 = \rho$	$\tilde{r}_2^2 = \rho^2$
$\tilde{r}_2^0 = \rho^2 - \frac{1}{3}$	$\tilde{r}_3^1 = \rho^3 - \frac{3}{5}\rho$	$\tilde{r}_4^2 = \rho^4 - \frac{5}{7}\rho^2$
$\tilde{r}_4^0 = \rho^4 - \frac{6}{7}\rho^2 + \frac{3}{35}$	$\tilde{r}_5^1 = \rho^5 - \frac{70}{63}\rho^3 + \frac{5}{21}\rho$	$\tilde{r}_6^2 = \rho^6 - \frac{882}{693}\rho^4 + \frac{35}{99}\rho^2$

Table 2.2 – Analytical expressions of the modified radial polynomials $\tilde{r}_n^m$'s.

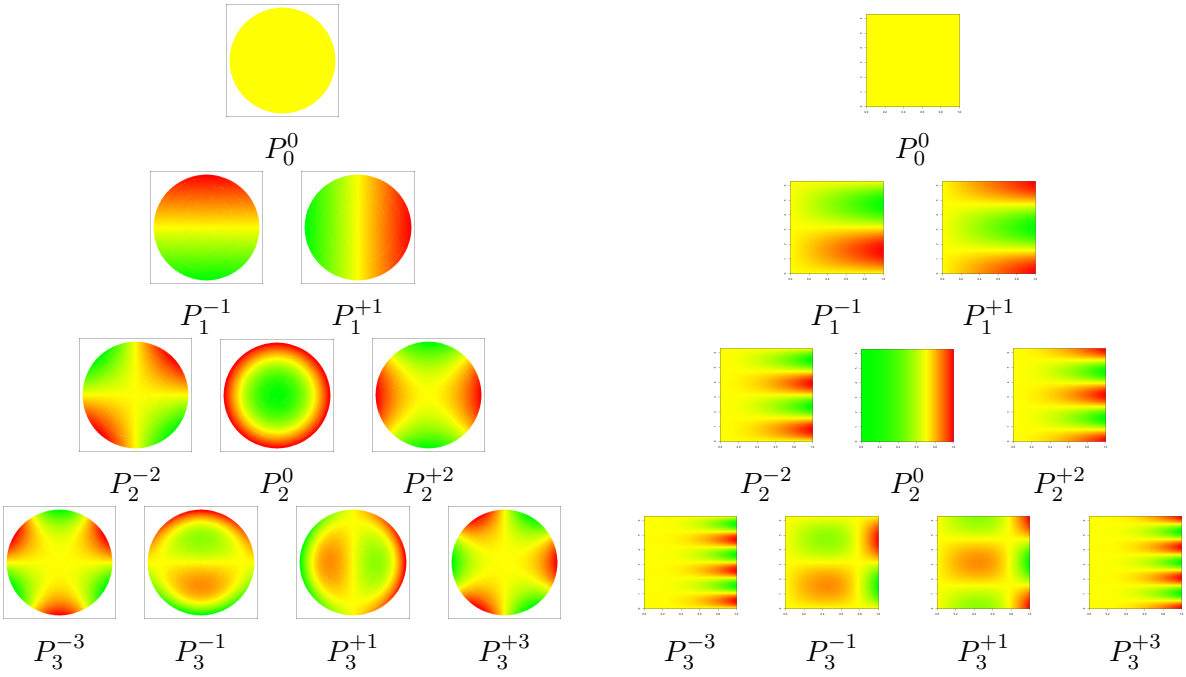


Figure 2.2 – Orthogonal polynomials over the disk with respect to the uniform measure over the space of polar coordinates: y vs x (left), and θ vs ρ (right).

Proof. Due to the necessary condition resulting from 2.3, such polynomials are of the form $Q_n^m(\rho)\cos(m\theta)$ or $Q_n^m(\rho)\sin(m\theta)$. Different solutions would come from the radial polynomials $Q_n(\rho)$. The additional condition of increasing degrees in x and y, and therefore in ρ, ensures uniqueness (up to a constant). □

Notice that other kinds of measure modifications could be investigated. In 1D, the question is widely addressed, especially when the new measure is the product (or quotient) of the old one by a polynomial (see e.g.[58], Section 2.7).

2.1.4 Estimation

Estimation with Fourier coefficients

In the framework of large datasets, interpolation using Zernike polynomials is done in a deterministic way, based on truncated Fourier series (see e.g. [26]):

$$Y(\mathbf{x}) = \sum_{n=0}^{d} \sum_{\substack{m=-n \\ n-|m| \in 2\mathbb{N}}}^{n} \beta_n^m Z_n^m(\mathbf{x}) \tag{2.14}$$

where $(\beta_n^m)_{m,n} = \boldsymbol{\beta}^\top$ is the vector of coefficients to estimate. Equation 2.14 defines the coordinates of Y onto $\text{vec}(Z_0^0, Z_1^{-1}, ..., Z_d^d)$, which is an orthonormal system. The β_n^m's are then uniquely defined by orthogonal projections via the scalar product:

$$\beta_n^m = \langle Y, Z_n^m \rangle = \int_{\mathcal{D}} Y(\rho,\theta) Z_n^m(\rho,\theta)\, \rho d\rho d\theta \tag{2.15}$$

Given a finite number of observations $\left(\mathbf{x^{(i)}}, Y_i\right)$ with $1 \leq i \leq N$, the β_n^m's are estimated as Fourier coefficients:

$$\hat{\beta}_n^m = \sum_{i=1}^{N} Y_i Z_n^m\left(\mathbf{x^{(i)}}\right) \tag{2.16}$$

By exploiting the orthogonality property, the procedure has the advantage of computing separately the β_n^m's. In particular, only a subset of coefficients can be estimated. However, its validity is subject to constraints on the design $\mathbf{X}$. First, the sample size must be large enough to ensure convergence when approximating integral terms. Second, the sample density must be uniform on $\mathcal{D}$ to be consistent with the integration measure in Equation 2.1. Such conditions, especially uniformity with respect to polar angles, are hard to meet in practice.

Estimation based on Ordinary Least Squares

The second method to estimate $\boldsymbol{\beta}$ is based on Ordinary Least Squares (OLS), corresponding to the maximum likelihood estimator when ε is a Gaussian noise (see e.g. [49], Chapter 3). The d-order linear regression model is:

$$Y_i = \sum_{n=0}^{d} \sum_{\substack{m=-n \\ n-|m| \in 2\mathbb{N}}}^{n} \beta_n^m Z_n^m\left(\mathbf{x^{(i)}}\right) + \varepsilon_i, \quad 1 \leq i \leq N \tag{2.17}$$

where the ε_i's are independent error terms, modelled as:

$$\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right) \tag{2.18}$$

Under this assumption, the vector $\hat{\boldsymbol{\beta}}$ follows a multivariate normal distribution:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{F}^\top \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{F}^\top \mathbf{F}\right)^{-1}\right) \tag{2.19}$$

where $\mathbf{F}$ is the experimental matrix defined as $\left(\mathbf{Z_n^m}\left(\mathbf{x}^{(i)}\right)\right)$, with $i = 1 \ldots N$, $n = 0 \ldots d$, and $n - |m| \in 2\mathbb{N}$. The estimation procedure thus defined is the Best Linear Unbiased Estimator (BLUE) for the regression model 2.17 and will be used in this thesis.

2.2 Kriging

Given the responses $Y_1, \ldots, Y_n$ at $\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}$ and a new point $\mathbf{x^{(0)}} \in \mathcal{D}$, an interesting question is to predict $Y_0 = Y\left(\mathbf{x^{(0)}}\right)$ as an affine combination of $Y_1, \ldots, Y_n$:

$$Y_0 = \mu + \sum_{i=1}^{n} \lambda_i \left(Y_i - \mu\right),$$

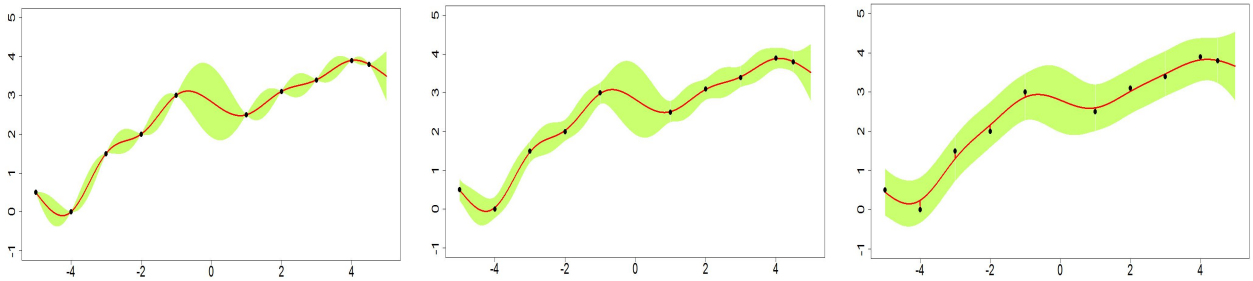
where μ represents the response mean, and $\lambda_1, \ldots, \lambda_n$ are the weights of the $\mathbf{x}^{(i)}$'s in the prediction at $\mathbf{x^{(0)}}$. Kriging consists in choosing the λ_i's to minimize the quadratic risk $\mathbb{E}\left[\left(Y_\mathbf{0} - \mu - \sum_{i=1}^{n} \lambda_i \left(Y_i - \mu\right)\right)^2\right]$, based on the assumption that $(Y_0, Y_1, \ldots, Y_n)$ originate from a random realization of a Gaussian random field $(Z_\mathbf{x})_{\mathbf{x}\in\mathcal{D}}$. Under this assumption that will be detailed hereafter, the estimation is the Best Linear Unbiased Predictor (BLUP). The resulting model is called Kriging in honour of Daniel Krige for his pionneering works in geostatistics [73].

2.2.1 Definition

From a probabilistic point of view, the observations $Y_1, \ldots, Y_n$ can be modelled as the sum of a deterministic trend function μ, and a realization of the Gaussian random field $(Z_\mathbf{x})_{\mathbf{x}\in\mathcal{D}}$. In this sense, Kriging is also called Gaussian Process Regression [90], since it infers the expected value of Z, conditionally on observations. The model is:

$$Y_i = \mu(\mathbf{x^{(i)}}) + Z(\mathbf{x^{(i)}}) + \eta_i \tag{2.20}$$

where, $\eta_1, \ldots, \eta_n$ are Gaussian random variables with law $\mathcal{N}(0, \tau^2)$. τ^2 is an homogeneous variance term called "nugget" or "jitter" such that the model is an interpolator if $\tau = 0$, and becomes a smoother when $\tau > 0$. The trend function μ describes deterministic variations,



(a) Interpolation ($\tau = 0$) (b) Smoothing ($\tau = 0.01$) (c) Smoothing ($\tau = 0.05$)

Figure 2.3 – Kriging predictions with different values of τ. The red line represents estimated values, black points are observations, and the green area is the prediction interval (95%)

and is usually specified in terms of basis functions such as polynomials. Z is a centred Gaussian process (GP), completely defined by its covariance function or kernel k. In the examples in Figure 2.3, the trend μ is constant. At each point, Kriging provides a prediction interval, based on a Gaussian conditional distribution (see Paragraph 2.2.3 for more details).

2.2.2 Covariance functions or kernels

The kernel k is defined as $k(\mathbf{x}, \mathbf{x}') = \mathrm{cov}\Big(Z(\mathbf{x}), Z(\mathbf{x}')\Big)$. Unlike the mean function μ, k is more difficult to parametrize, due to requirements of positive definiteness:

$$\sum_{i,j=1}^{m} a_i a_j k(\mathbf{x_i}, \mathbf{x_j}) \geq 0, \quad \forall m \in \mathbb{N},\ \forall\, \mathbf{x_1}, \ldots, \mathbf{x_m} \in \mathcal{D},\ \forall\, a_1, \ldots, a_m \in \mathbb{R} \tag{2.21}$$

Condition 2.21 ensures that the covariance matrix of $(Z_{\mathbf{x^{(1)}}}, \ldots, Z_{\mathbf{x^{(n)}}})$ is positive semidefinite. Parametric families of kernels are proposed in the literature, based on particular assumptions. Stationarity and isotropy are especially two important concepts in Kriging. When Z is stationary, $\mathrm{cov}\big(Z(\mathbf{x}), Z(\mathbf{x}')\big)$ depends only on the difference $\mathbf{x} - \mathbf{x}'$, and all observations have the same variance. In addition, if $\mathrm{cov}\big(Z(\mathbf{x}), Z(\mathbf{x}')\big)$ depends only on the Euclidean distance $\| \mathbf{x} - \mathbf{x}' \|$, then Z is said to be isotropic.

In Table 2.3 are presented some commonly used kernels in dimension 1 and information about their differentiability. Since the GP sample paths smoothness is linked to its kernel smoothness[90], the sample paths generated by the Gaussian kernel will be infinitely smooth whereas those originating from k_{Brown} will be very harsh. The range parameter ℓ indicates the characteristic distance between two points that leads to significant changes in the process. Finally σ^2 allows to control the overall variance of the GP.

Kernel	Formula	Parameter	Smoothness (quadratic sense)
Brownian motion	$k_{Brown}(u,v) = \min(u,v)$	–	C^0
Gaussian	$k_G(x,v) = \sigma^2 \exp\left(-\left(\frac{u-v}{\ell}\right)^2\right)$	$\ell > 0$	C^∞
Power-exponential	$k_{\exp}(u,v) = \sigma^2 \exp\left(-\lvert\frac{u-v}{\ell}\rvert^\alpha\right)$	$0 < \alpha < 2, \quad l > 0$	C^0
Matérn$_{\frac{3}{2}}$	$k_{m\frac{3}{2}}(u,v) = \sigma^2\left(1 + \frac{\sqrt{3}\lvert u-v\rvert}{\ell}\right)\exp\left(-\frac{\sqrt{3}\lvert u-v\rvert}{\ell}\right)$	$\ell > 0$	C^1
Matérn$_{\frac{5}{2}}$	$k_{m\frac{5}{2}}(u,v) = \sigma^2\left(1 + \frac{\sqrt{5}\lvert u-v\rvert}{\ell} + \frac{5(\lvert u-v\rvert)^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}\lvert u-v\rvert}{\ell}\right)$	$\ell > 0$	C^2
Matérn$_\nu$	$k_\nu(u,v) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\lvert u-v\rvert}{\ell}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{\lvert u-v\rvert}{\ell}\right)$	$\ell > 0, \quad \nu \in \frac{1}{2}\mathbb{N} \setminus \{0\}$	$C^{(\nu-1)}$

Table 2.3 – Some common kernels in 1D, with K_ν the modified Bessel function.

Higher dimension kernels can be designed by combining 1D kernels, as we will see in Chapter 3. In dimension 2 for instance, the tensor-product Gaussian kernel of Equation 2.22 is a common choice.

$$\mathrm{cov}\big(Z(\mathbf{x}), Z(\mathbf{x}')\big) = k_G(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\left(\frac{x_1 - x_1'}{\theta_1}\right)^2 - \left(\frac{x_2 - x_2'}{\theta_2}\right)^2\right) \tag{2.22}$$

This kernel belongs to $C^\infty(\mathcal{D} \times \mathcal{D})$. Then, it produces infinitely differentiable response surfaces. However, to get a well-conditioned covariance matrix, less smooth kernels are recommended. The Matérn family in Table 2.3 allows to parametrize smoothness via the parameter ν. As a centred Gaussian vector is fully defined by its covariance matrix, Kriging predictions will be accurate if a correct kernel is set. The choice is governed by prior knowledge on Y, such as smoothness and symmetries.

2.2.3 Prediction

When all the parameters of a Kriging model are known, prediction with Equation 2.20 at a new point $\mathbf{x} \in \mathcal{D}$ is given by a Gaussian conditional distribution, knowing the observations $\mathbf{y} = (Y_1, \ldots, Y_n)$. The moments of this distribution, called Simple Kriging mean and Simple Kriging variance, are provided in Equations 2.23 and 2.24.

$$m_{\mathrm{SK}}(\mathbf{x}) = \mu(\mathbf{x}) + \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\big(\mathbf{y} - \mu(\mathbf{x})\big) \tag{2.23}$$

$$s^2_{\mathrm{SK}}(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\mathbf{k}(\mathbf{x}) \tag{2.24}$$

where $\mathbf{K} = \Big(k\left(\mathbf{x^{(i)}}, \mathbf{x^{(j)}}\right)\Big)_{1 \le i,j \le n}$ is the covariance matrix at the design points, $\mathbf{k}(\mathbf{x}) = \Big(k\left(\mathbf{x}, \mathbf{x^{(i)}}\right)\Big)_{1 \le i \le n}$ is called covariance vector at $\mathbf{x}$.

In practice, the trend function μ is often unknown and estimated by maximum likelihood (ML). The model is called Universal Kriging. Universal Kriging mean and variance are given in Equations 2.25 and 2.26.

$$m_{\mathrm{UK}}(\mathbf{x}) = \hat{\mu}(\mathbf{x}) + \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\big(\mathbf{y} - \hat{\mu}(\mathbf{x})\big) \tag{2.25}$$

$$s^2_{\mathrm{UK}}(\mathbf{x}) = s^2_{\mathrm{SK}}(\mathbf{x}) + \big(\mathbf{f}(\mathbf{x}) - \mathbf{F}^\top\mathbf{K}^{-1}\mathbf{k}(\mathbf{x})\big)^\top \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right)\big(\mathbf{f}(\mathbf{x}) - \mathbf{F}^\top\mathbf{K}^{-1}\mathbf{k}(\mathbf{x})\big) \tag{2.26}$$

where $\mathbf{F}$ is the experimental matrix used to estimate the trend as defined in Section 2.1 and $\hat{\mu}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \hat{\boldsymbol{\beta}}$. The ML estimation of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \big(\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F}\big)^{-1}\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{z}$, with covariance matrix $\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \big(\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F}\big)^{-1}$. The paramaters of the kernel are usually estimated by maximum likelihood. The model is then called "Plug-in-Kriging", which means Kriging based on an estimated kernel. Cross-validation strategies such as leave-one-out can also be used for noise-free responses [90, 97].

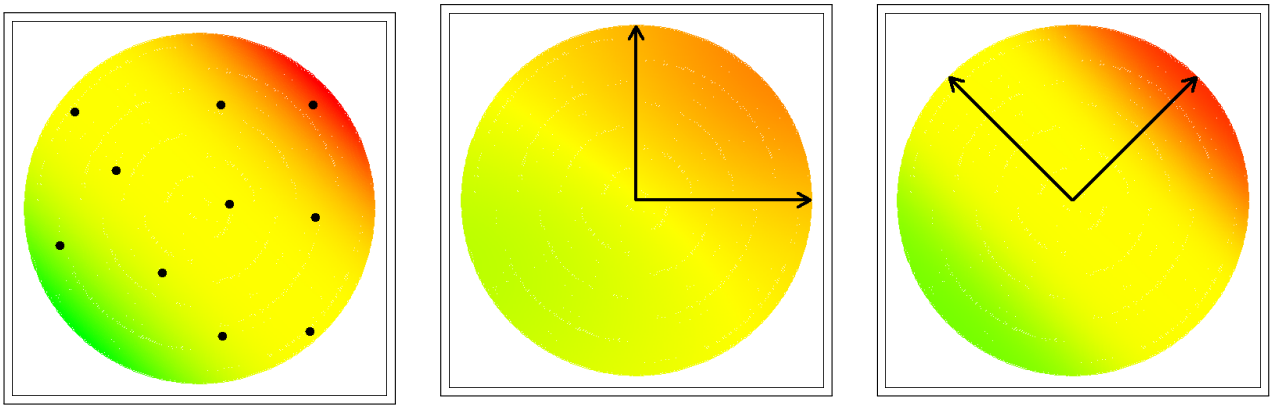
2.3 Geometric anisotropy in Kriging

2.3.1 An introducing example

We consider the tensor-product Gaussian kernel in Equation 2.22. Given two points $\mathbf{x} = (x_1, x_2)$ and $\mathbf{x}' = (x'_1, x'_2)$ on the disk, the covariance between $Z(\mathbf{x})$ and $Z(\mathbf{x}')$ is written:

$$\mathrm{cov}\Big(Z(\mathbf{x}), Z(\mathbf{x}')\Big) = \sigma^2 \exp\left(-\left(\frac{x_1 - x'_1}{\theta_1}\right)^2\right) \exp\left(-\left(\frac{x_2 - x'_2}{\theta_2}\right)^2\right)$$

Therefore, $\mathrm{cov}\big(Z(\mathbf{x}), Z(\mathbf{x}')\big)$ decreases in function of two terms. The first one depends on the horizontal distance $\left|x_1 - x'_1\right|$, and the second one depends on the vertical distance $\left|x_2 - x'_2\right|$. For the sake of clarity, we call these two terms *horizontal and vertical correlations*. The ratio $\frac{\theta_1}{\theta_2}$ quantifies the relative importance of *horizontal and vertical correlations*. In particular, $\theta_1 = \theta_2$ leads to an isotropic kernel, whereas high differences between θ_1 and θ_2 correspond to more important variations according to x_1 or x_2. This is an example of anisotropy, a property of being dependent on directions. When the input domain is a disk, there is no natural choice for x_1 and x_2 axes. Instead of arbitrary settings, we study a data-driven solution. Let us consider the dataset in Figure 2.4 where the response depends only on $x_1 + x_2$. Based on the



(a) Response values and DoE. (b) $\left(\frac{\hat{\theta}_2}{\hat{\theta}_1} = 1.2, \text{RMSE} = 0.45\right)$ (c) $\left(\frac{\hat{\theta}_2}{\hat{\theta}_1} = 13, \text{RMSE} = 0.15\right)$

Figure 2.4 – Kriging models over $\mathcal{D}$, based on $f(x_1, x_2) = (x_1 + x_2)^3$

10-points design in Figure 2.4a, two simple Kriging models are estimated. The first one uses the Cartesian basis shown in Figure 2.4b, and the second uses the basis in Figure 2.4c. We observe that the result can significantly change according to the basis. In this example, the basis represented in Figure 2.4c allows to better describe spatial correlations. The estimated range parameters $\left(\hat{\theta}_1 = 0.7, \hat{\theta}_2 = 10\right)$ are consistent with the orientation. Notice that similar results are obtained with exponential and Matérn kernels, with and without nugget. The property of varying mainly in one direction is known in geostatistics as geometric anisotropy, and the degenerated case where Z is constant in one direction is called zonal anisotropy [4].

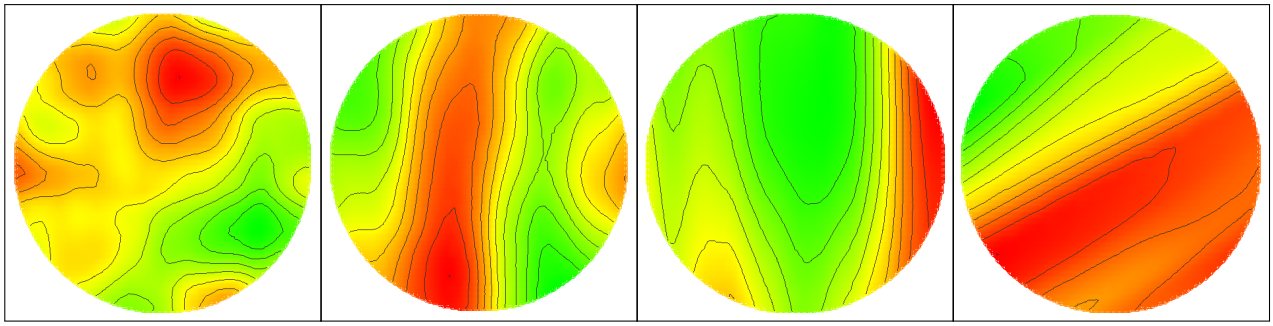
2.3.2 Definition and key properties

Definition and GP simulations on the unit disk

A simple way to include geometric anisotropy is Kriging models consists in applying a linear transformation to the coordinate system (see e.g. [4, 48]). We consider kernels of the form:

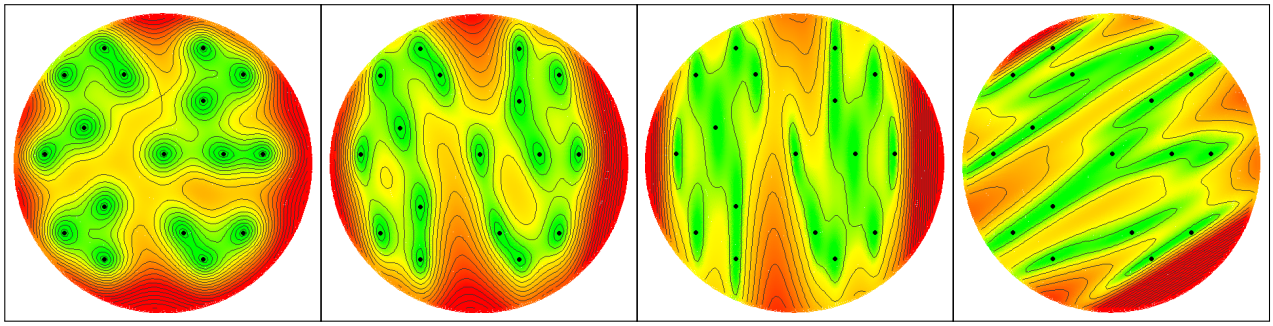
$$k_\phi(\mathbf{x}, \mathbf{x}') = k\Big(R_\phi(\mathbf{x}), R_\phi(\mathbf{x}')\Big), \phi \in \left[0, \frac{\pi}{2}\right] \tag{2.27}$$

where k is a 2D kernel, Matérn$\frac{5}{2}$ for instance, and R_ϕ the rotation with angle ϕ. To describe the effects of geometric anisotropy in Kriging models, we draw simulated surfaces and Kriging standard deviations, based on 17 observations and varying $(\theta_1, \theta_2, \phi)$. The simulated surfaces vary more in the direction with the lower range (Figures 2.5). Kriging standard deviations in Figures 2.6 allow to describe spatial uncertainty over $\mathcal{D}$. The neighborhoods of design points look like ellipses, oriented in the directions ϕ and $\phi + \frac{\pi}{2}$, one of which corresponds to largest correlations. Following the terminology of *horizontal and vertical correlations*, the kernel in Equation 2.27 allows to model *oblique correlations*.



$(\theta_1 = \frac{1}{2}, \theta_2 = \frac{1}{2}, \phi = 0)$ $(\theta_1 = \frac{1}{2}, \theta_2 = 1, \phi = 0)$ $(\theta_1 = \frac{1}{2}, \theta_2 = 3, \phi = 0)$ $(\theta_1 = \frac{1}{2}, \theta_2 = 3, \phi = \frac{\pi}{3})$

Figure 2.5 – Simulations of Gaussian processes over $\mathcal{D}$ with different scenarios of anisotropy.



$(\theta_1 = \frac{1}{2}, \theta_2 = \frac{1}{2}, \phi = 0)$ $(\theta_1 = \frac{1}{2}, \theta_2 = 1, \phi = 0)$ $(\theta_1 = \frac{1}{2}, \theta_2 = 3, \phi = 0)$ $(\theta_1 = \frac{1}{2}, \theta_2 = 3, \phi = \frac{\pi}{3})$

Figure 2.6 – Kriging standard deviations under different scenarios of anisotropy.

Model estimation: key points for Matérn kernels

Haskard [48] carried out an intensive study, including simulations, on Kriging models with geometric anisotropy. Focused on Matérn kernels, it resulted in the following properties:

1. Geometric anisotropy has no negative influence on parameter estimation.
2. Omitting an existing anisotropy can substantially worsen Kriging predictions.
3. Over-fitting by including a non-existing anisotropy is less damaging.

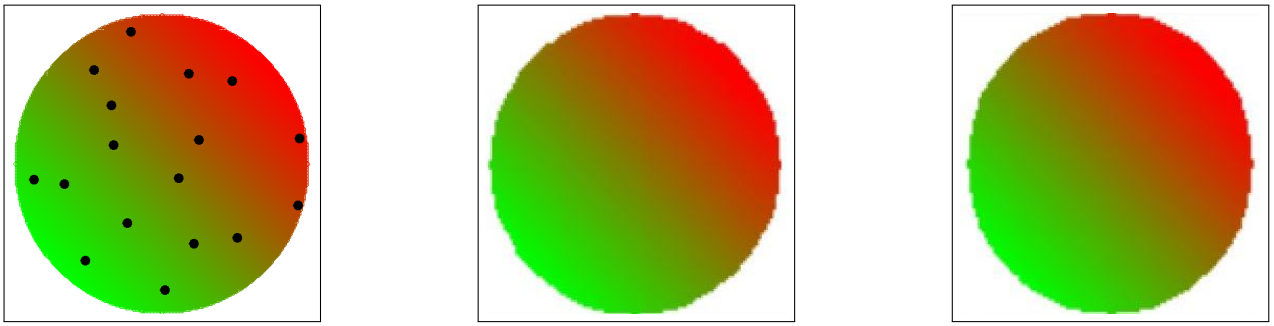
Link with single index models

Geometric anisotropy in 2D means that the response mainly depends on a linear combination of x_1 and x_2. More generally in machine learning, a common issue is related to the large numbers of predictors $(x_1, \ldots, x_p) = \mathbf{X_p}$. An alternative to selecting a subset of variables is to use a dimension reduction method, such as projection pursuit regression (PPR). In PPR, the response is assumed to depend on M $(M \leq p)$ linear combinations of $x_1, \ldots, x_p$ [49]:

$$f(\mathbf{X_p}) = \sum_{m=1}^{M} L_m \left(w_m^T \mathbf{X_p} \right) \tag{2.28}$$

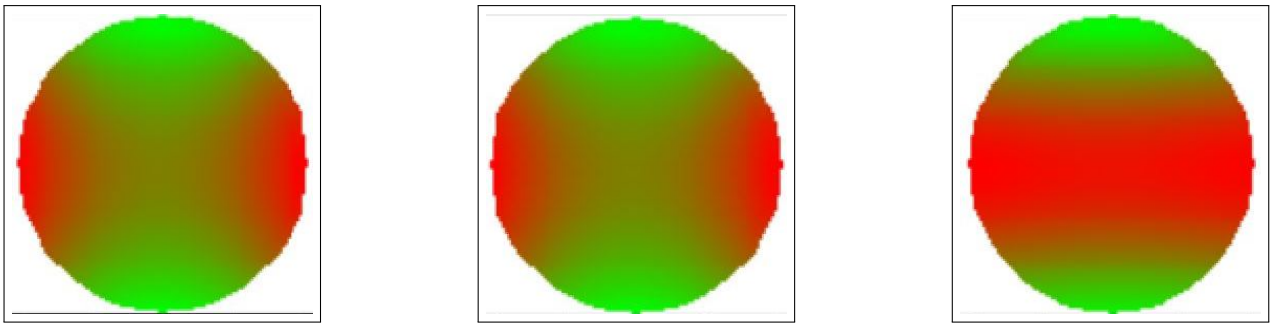
where $w_m \in \mathbb{R}^p$ such as $\|w_m\| = 1$, and L_m is a link function. $M = 1$ means that the response depends on a single linear combination of the input variables. This case is refereed to as Single-Index Model, and Gaussian process Single-Index Model (GP-SIM) if the

link function L_m is a GP [46]. Therefore, Kriging with geometric anisotropy is similar to a projection pursuit regression. In particular, a GP-SIM in dimension 2 is a Kriging model with zonal anisotropy. To understand the difference between Single-Index Models and Kriging with geometric anisotropy, we consider 3 analytical functions to be recovered over $\mathcal{D}$, based on the design points of Figure 2.7a. The predicted values with a GP-SIM (Figures 2.7c, 2.8c and 2.9c) are compared to those produced by Kriging with geometric anisotropy (Figures 2.7b, 2.8b and 2.9b).



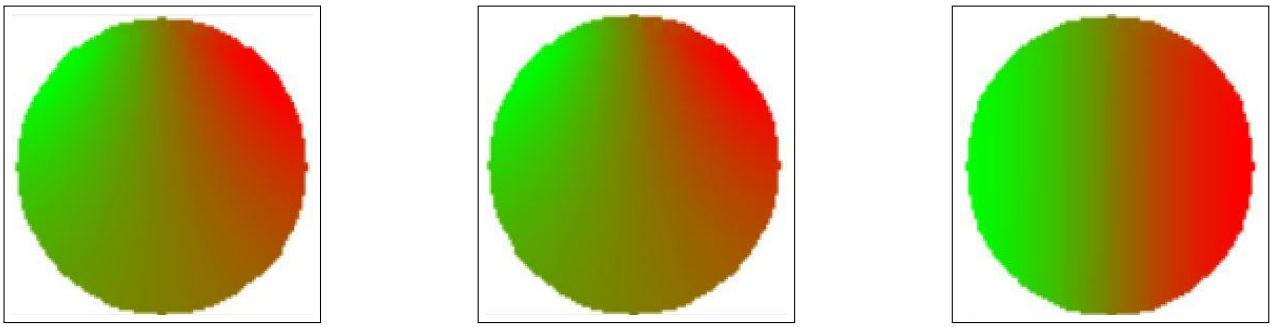
(a) Real values (b) Kriging (c) GP-SIM

Figure 2.7 – Kriging and GP-SIM predictions for the analytical response $\mathrm{sh}\Big(5(x_1+x_2)\Big)$



(a) Real values (b) Kriging (c) GP-SIM

Figure 2.8 – Kriging and GP-SIM predictions for the response $\mathrm{sh}\Big(5(x_1+x_2)\Big)\mathrm{sh}\Big(5(x_1-x_2)\Big)$



(a) Real values (b) Kriging (c) GP-SIM

Figure 2.9 – Kriging and GP-SIM predictions for the analytical response $\sin(x)e^y$

Through these examples, we see that whether modelling geometric anisotropy or using a GP-SIM allows to select the main direction of variation 2.7c. However, the GP-SIM is

virtually in dimension 1. It will always privilege one direction, even not relevant (Figures 2.8c and 2.9c). Remark that GP-SIMs meet more applications in higher dimension [46].

2.3.3 Assessment with analytical functions

The purpose of this paragraph is to assess the benefits from modelling geometric anisotropy, based on toy functions, representing different scenarios in Figure 2.10, with S the sigmoid $\mathrm{S}(t) = \frac{1}{1+e^{-t}}$ and B the Branin function $B(s,t) = \left(15t - \frac{5(15s-5)^2}{4\pi^2} + \frac{5}{\pi}(15s-5) - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(15s-5) + 10$. For each test function, 20 points filling the disk are used as

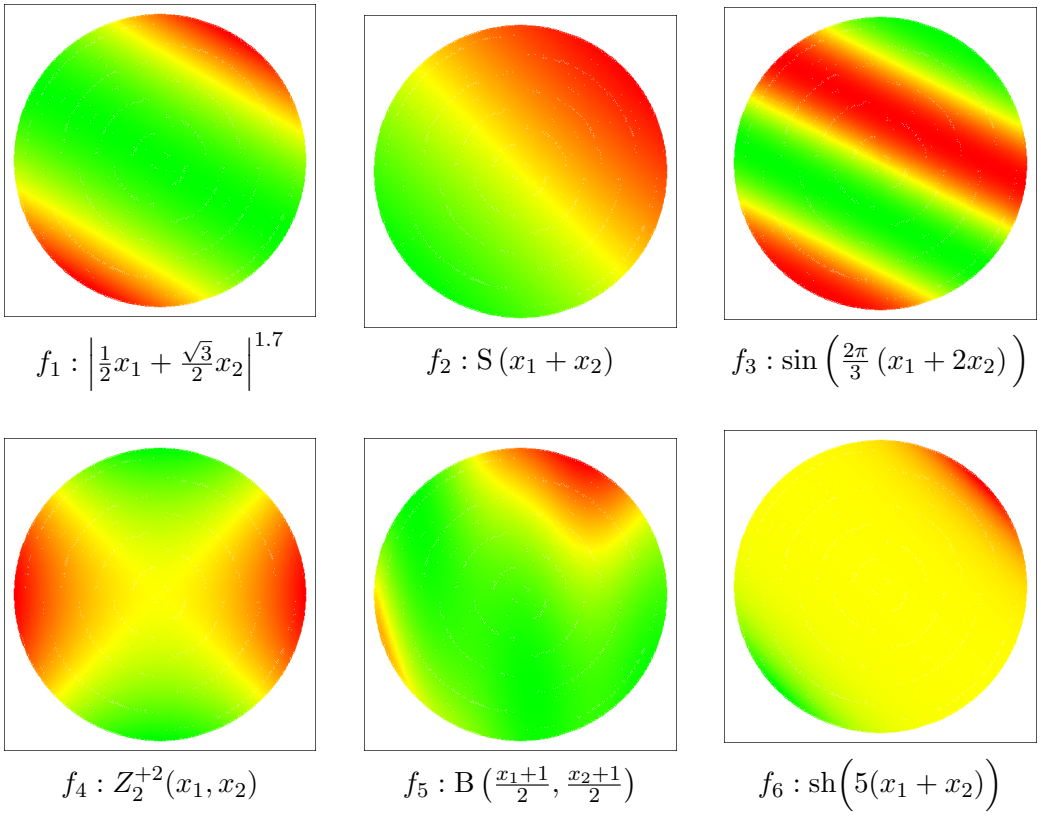


Figure 2.10 – Representation of the test functions.

DoE. Two GP models are estimated. The first one is the usual anisotropic tensor-product Gaussian kernel. The second one embeds geometric anisotropy (Equation 2.27). The RMSE of each model is computed, based on a regular grid with 1000 points over $\mathcal{D}$. The gain due to modelling geometric anisotropy is defined as:

$$Gain = 100\left(\frac{RMSE_{Gauss} - RMSE_{k_\phi}}{\sigma(Y)}\right)$$

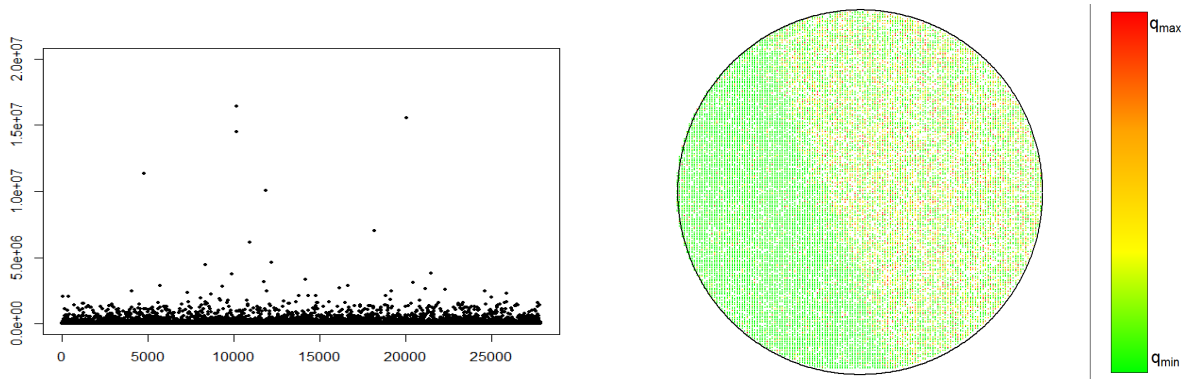
In Table 2.4 are shown the gains for the different test functions. The estimated anisotropy angle is also provided and compared with the angle resulting from a PLS regression (Partial Least Squares [49]). In this table, KLM refers to Kriging with a two-order Zernike polynomial as a trend, and KM correspond to a constant trend. Based on these scenarios, we can extend the conclusions provided in [48] for the Gaussian kernel. In particular, modelling geometric anisotropy leads to significant improvements if the response is strongly non-linear according to one direction (functions f_1, f_2 and especially f_3). In addition, setting the PLS angle as initial value for likelihood maximization would be a relevant choice.

	Gain (%)		Estimation $\hat{\phi}$ (degrees)			
Functions	KM	KLM	KM	KLM	$Real$	PLS
f_1	2.5	0	60	65	60	71
f_2	1.5	0	44	53	45	44
f_3	86	81	64	62	60	61
f_4	0	0	90	28	-	49
f_5	0	0	9	89	-	71
f_6	3	0	40	48	45	49

Table 2.4 – Gains due to modelling geometric anisotropy in different Kriging models.

2.4 Application in microelectronics

To show the importance of geometric anisotropy in microelectronics, we consider an electrical variable Y, measured at 27824 points of a wafer. The data are rescaled to $[0, 1]$ such that higher values correspond to good electrical performances, and lower values represent bad integrated circuits. As shown in Figure 2.11a, the dataset contains some outliers, resulting from unsuccessful tests. Therefore, the largest 10% of observations are discarded from the color representation in Figure 2.11b to provide a relevant visualization. There are obviously



(a) Response values Y.

(b) Variations of Y over $\mathcal{D}$.

Figure 2.11 – Outputs of an electrical test over a wafer.

two subareas of the wafer, corresponding to good and bad products respectively. In addition, the response varies mainly in an oblique direction, orthogonal to the border line between the two groups.

2.4.1 Zernike regression

As a first option to describe this pattern, a Zernike regression of order 2 is performed. The regression coefficients, related to normalized polynomials, are shown in Figure 2.12. They allow to describe the shape of Y. In particular, the ratio $\frac{\hat{\beta}_1^{-1}}{\hat{\beta}_1^{+1}} = 2.8$ suggests an angle ϕ around 20 degrees to model geometric anisotropy.

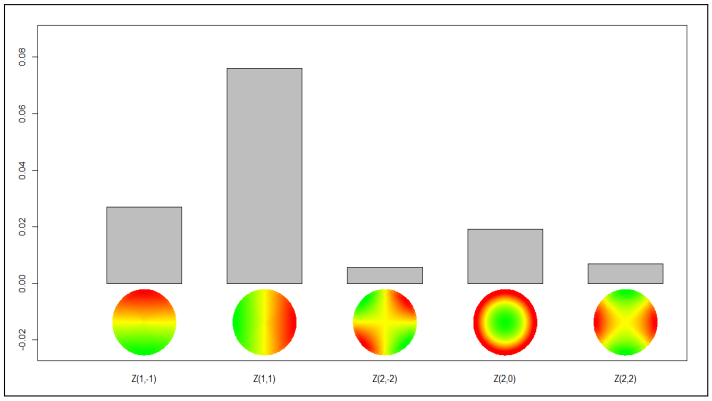
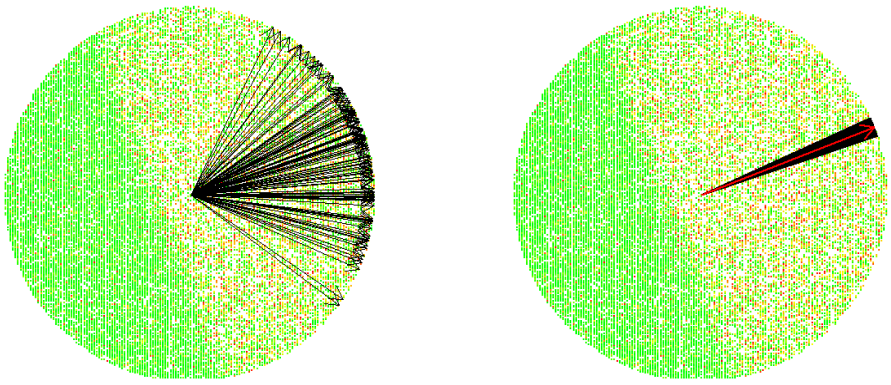


Figure 2.12 – Absolute values of 6 regression coefficients, representing the influence of different Zernike polynomials when modelling Y with Equation (2.17) and $d = 2$.

2.4.2 Estimation of the anisotropy angle

In order to confirm this intuition, a robust estimation is performed, based on PLS regressions. PLS provides a direct estimation of ϕ because it models Y as a linear combination of x_1 and x_2. 100 estimations of ϕ are then obtained, based on 100 designs of 50 random observations. They are displayed in Figure 2.13a. The population of simulated angles can be modelled with the von-Mises distribution, the circular analogue of the normal distribution. This leads to $\phi \in [18°, 25°]$ as 95% confidence interval, which is consistent with Zernike coefficients.

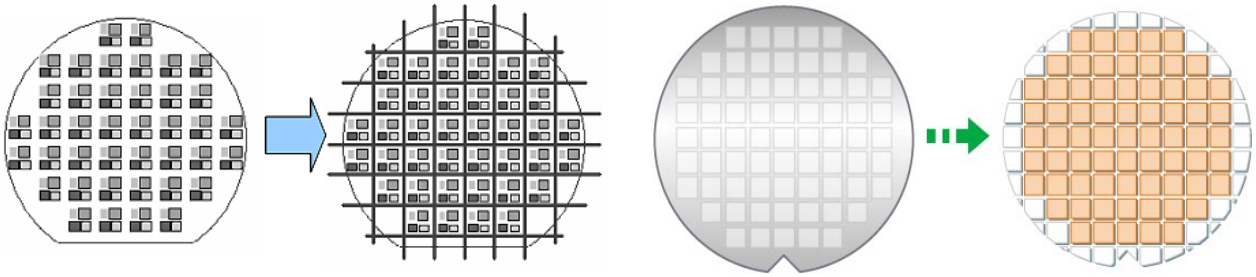


(a) Estimated angles. (b) Confidence interval for ϕ.

Figure 2.13 – Estimation of ϕ with PLS regressions and von Mises distribution.

2.4.3 Wafer notch orientation

Geometric anisotropy often occurs in microelectronics, and markers of directions are needed to describe spatial patterns. In practice, there is a flat cut on wafers, oriented according to the horizontal axis or the first bisector. This cut, called notch, allows also to indicate the crystallographic direction 2.14a. On the most recent wafers, the notch is shrunken for economic reasons (Figure 2.14b).



(a) Notch on a wafer. (b) A modern notch.

Figure 2.14 – Examples of wafer notch orientation.

Résumé en Français

Deux modèles de surface de réponse en domaine circulaire sont considérés. Le premier est une régression utilisant les polynômes de Zernike et le second est une technique d'interpolation ou de lissage à base de processus gaussiens.

Les polynômes de Zernike sont une base de fonctions orthogonales par rapport à la mesure uniforme sur le disque unité. Ils se distinguent des autres polynômes par certaines propriétés d'invariance par rotation qui sont particulièrement utiles en optique. Leurs caractéristiques sont aussi exploitées en microélectronique pour la modélisation spatiale de variables physiques et électriques. Une analyse des propriétés des polynômes de Zernike nous a permis de déduire une deuxième famille de polynômes dotés des mêmes propriétés de symétrie et de rotation, mais orthogonaux par rapport à la mesure uniforme dans l'espace des coordonnées polaires.

En ce qui concerne les modèles de krigeage, on remarque que leur fonction de covariance (noyau) se définit usuellement avec les coordonnées cartésiennes. Or, tout choix de repère cartésien sur le disque est arbitraire a priori et nous remarquons que certains repères sont plus adaptées que d'autres selon les données. Dans la perspective d'une sélection judicieuse et automatique du système de coordonnées sur le disque, nous modélisons l'anisotropie géométrique dans les noyaux de covariance. Cette notion, initialement introduite en géostatistique, répond à notre problème spatial en retrouvant les axes correspondant aux principales directions de variation.

Chapter 3

Polar Gaussian processes

The results of this chapter are based on the contribution "Polar Gaussian Processes for Predicting on Circular Domains" [81], by Padonou and Roustant, in revision for SIAM/ASA *Journal on Uncertainty Quantification.*

3.1 Introduction

This research aims at analyzing costly computer or physical experiments on a disk. The question was motivated by two industrial problems. The first one comes from semiconductor industry where integrated circuits are produced on disks called wafers. Several technological processes such as lithography, heating or polishing, exploit the geometry of the disk through rotations or diffusions from the center. A common issue is to reconstruct a quantity of interest over the whole disk, from few measurements at specific locations. The second problem is related to air pollution modelling for environmental impact assessment. Greenhouse gas concentrations are simulated by a computer code. Among the input variables, the pair (speed, direction) of wind characteristics can be represented on a disk, the radius of which corresponds to the maximal speed. Here also, the goal is to predict the gas concentration from some simulated experiments.

Approximation problems on the disk have been considered since the works of Zernike [118] in optics. Zernike polynomials are orthogonal with respect to the usual scalar product on the unit disk, a useful property for linear models. For such models, it is shown that optimal design of experiments are included in concentric circles [26]. More recently, a stochastic model consisting of a Gaussian process (GP), also called Kriging, has been proposed for microelectronics applications [89]. Among the existing interpolation and approximation methods, Kriging models are famous for their ability to provide both accurate prediction and uncertainty quantification, as pointed out in [45]. However their performance relies on the choice of a covariance kernel, often simply called kernel hereafter. Traditional kernels do not take into account the geometry of the disk. This may be a drawback, at least for technological or physical processes involving a diffusion from the center of the disk, or a rotation.

The main aim of the paper is to propose GP models that incorporate the geometry of the disk in their covariance kernel. For that purpose, we consider the parametrization of the unit disk in polar coordinates: $\mathcal{D} = \{(\rho\cos\theta, \rho\sin\theta), \rho \in [0,1], \theta \in \mathbb{S}\}$ where $\mathbb{S}$ represents the unit circle $\mathbb{R}/2\pi\mathbb{Z}$. The idea is to define a GP on the parametrization space $\mathcal{C} = (0,1]\times\mathbb{S}$ defined by (ρ,θ). This implies constructing a kernel on a product of the Euclidean space $(0,1]$ and of the circle $\mathbb{S}$, which can be done by algebraically combining kernels on these

two spaces with sum, product or ANOVA operations for instance. The corresponding GPs will be called here *polar* GPs, and the usual ones based on Cartesian coordinates, *Cartesian* GPs.

The construction of kernels on $\mathbb{S}$ can be achieved in several ways, and is connected to the literature of directional data (see e.g. [72, 35]) and periodic functions (see e.g. [90]). A first option in [107], based on Bochner's theorem, consists in using the spectral representation of 2π-periodic functions with positive Fourrier transforms. In other words, it uses the Schoenberg's representation of correlation functions on spheres (see e.g.[42], Equation 13). Another possibility is to use so-called wrapped GP, obtained by transforming a multinormal density to a periodic one by applying an operator written as an infinite sum [60]. Both require to truncate infinite series for practical implementation. Here we focus on simpler approaches that provide explicit kernel expressions, either by considering restriction to $\mathbb{S}$ of a 2-dimensional GP [90], involving the chordal distance on $\mathbb{S}$, or by using the recent results of Gneiting [42], involving the geodesic distance on $\mathbb{S}$. The geodesic distance on a general manifold was recently used in the context of free-form monitoring, with so-called geodesic GPs [23]. However, the goal and the approach are quite different here, where the form is fixed (the unit disk) and the geodesic distance known analytically. Furthermore, here the geodesic is relative to the manifold $\mathbb{S}$ which is only an algebraic portion of the mapped space $\mathcal{C}$.

Second, we address the issue of defining an initial design of experiments (DoE) for circular domains. Considering the space $\mathcal{C}$ of polar coordinates is natural, but standard designs cannot be used directly due to its non-Euclidean structure. By considering a valid distance, we obtain maximin Latin hypercube designs (LHD, [74]) on $\mathcal{C}$. That class of designs is recommended when the process has a physical interpretation in polar coordinates. In order to deal with more general situations, we also propose a modified version, which still has the LHD structure with respect to ρ and θ, and is well filling the disk $\mathcal{D}$.

The paper is organized as follows. Section §3.2 presents the background and defines notations. Section §3.3 introduces so-called polar GPs. Section §3.4 shows the strength of the approach on two real applications, in microelectronics and environments. Section §3.5 investigates an extension to higher dimensions, where the disk is replaced by a hyperball. Section §3.6 discusses the range of applicability of polar GPs and gives perspectives for future research.

3.2 Background and notations

Let $\mathcal{D}$ denote the unit disk represented either in Cartesian or polar coordinates:

$$\mathcal{D} = \{(x, y) \in \mathbb{R}^2, x^2 + y^2 \leq 1\} = \{(\rho\cos\theta, \rho\sin\theta), \rho \in [0, 1], \theta \in \mathbb{S}\}$$

where $\mathbb{S} = \mathbb{R}/2\pi\mathbb{Z}$ is the unit circle. In various situations, one has to predict a variable of interest which is measured at a limited number of locations in $\mathcal{D}$. For that purpose, we will consider the framework of Gaussian Process Regression [90] also called Kriging in reference to its origins in geostatistics (see e.g. [73]). The measurement locations, also called design points, will be denoted by $X = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right)$. In Gaussian Process Regression, the observed values at X are modelled by:

$$Y_i = \mu(\mathbf{x^{(i)}}) + Z(\mathbf{x^{(i)}}) + \eta_i \tag{3.1}$$

where μ is a trend function, $Z \sim GP(0, k)$ is a centered Gaussian process (GP) with covariance function – or *kernel* – k, and $\eta_1, \ldots, \eta_n$ are Gaussian random variables representing noise. We now briefly detail the three parts of the model.

The trend function μ is deterministic and often modeled as a linear combination of basis functions. Here, Zernike polynomials [118] are good candidates since they constitute an orthogonal basis for the usual scalar product on $\mathcal{D}$. Their shape including regular patterns are suited to describe symmetries or rotations. They were recently used for predicting on a disk [89]. The first Zernike polynomials, up to order 2, are shown in Fig. 3.1. The reader is

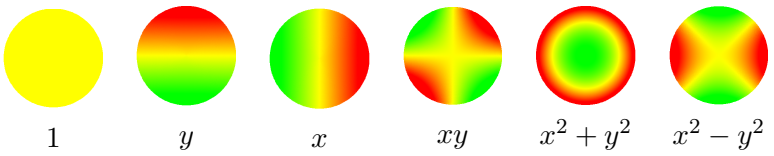


Figure 3.1 – The six first Zernike polynomials.

referred to [118] for more details.

The stochastic part of model 3.1 comprises a GP and a noise. The GP Z takes into account the spatial dependence, which thus entirely depends on its kernel k. The choice of k is crucial for applications, and may be done in order to include knowledge such as smoothness, periodicity, symmetries, etc. There are many ways to construct a kernel, and a comprehensive presentation is found in [90], Section 4. A key idea is that multidimensional kernels can be obtained by algebraic operations, such as sum or products, of 1-dimensional kernels.

Finally the noise part represented by the η_i's may have different purposes: Modelling a measurement noise or potential discrepancies between the dataset and the kernel, and addressing numerical issues such as ill-conditioning ([6, 44]). The η_i's are modeled as independent $N(0, \tau^2)$, where τ^2 is an unknown homogeneous variance term often called "nugget" or "jitter". When conditioning on the observed values, the model is an interpolator if $\tau = 0$. It is a smoother when $\tau > 0$, which gives more flexibility.

When all parameters are known, prediction with Equation (3.1) is given in a closed form by a Gaussian conditional distribution knowing the observations Y_i, $i = 1, \ldots, n$. Its two moments are known as Kriging mean and Kriging variance. Analytical expressions are also available when the parameters are estimated, known as universal Kriging formulas that we use here (see e.g. [90]). An important fact is that the Kriging mean at a new site $\mathbf{x}$ is obtained as an affine combination of the observed values Y_i that are correlated to $\mathrm{Z}(\mathbf{x})$. Though all the locations may be involved in the prediction, the neighboring locations, corresponding to high correlations, typically play a key role.

3.3 Polar Gaussian processes

One way to define a GP on the unit disk $\mathcal{D}$ is to use the restriction of a GP on the square $[0, 1]^2$, defined in Cartesian coordinates. In this paper, we will call them *Cartesian* GPs. In our work, we propose to further exploit the geometry of the disk by using the polar coordinates. The associated GPs will be called *polar* GPs.

When using the polar coordinates, the unit disk $\mathcal{D}$ is connected to the cylinder $\mathcal{C} = (0, 1] \times \mathbb{S}$, where $\mathbb{S}$ denotes the unit circle:

$$\Psi : (\rho, \theta) \in \mathcal{C} \mapsto (\rho \cos\theta, \rho \sin\theta) \in \mathcal{D} \setminus \{\mathbf{0}\} \tag{3.2}$$

It is a one-to-one correspondence from $\mathcal{C}$ to the unit disk without its center. The fact that the center is lost in the mapping may be a problem in theory. In practice a design point located at the center of the disk can be replaced by a set of design points placed on a closed concentric circle. A GP on $\mathcal{D}$ can then be obtained by using Ψ^{-1}, resulting in kernels on $\mathcal{D} \times \mathcal{D}$ of the form:

$$k(\mathbf{x}, \mathbf{x}') = k_{\mathcal{C}}\left(\Psi^{-1}(\mathbf{x}), \Psi^{-1}(\mathbf{x}')\right) \tag{3.3}$$

where $k_{\mathcal{C}}$ is a kernel on $\mathcal{C} \times \mathcal{C}$. Such transformations are referred to as "warping" in the context of GP modeling (see e.g. [90], Section 4.2.3.).

Kernels on the cylinder can be defined by exploiting its product structure. This can be done by combining a kernel k_{r} on $(0, 1] \times (0, 1]$ and a kernel k_{a} on $\mathbb{S} \times \mathbb{S}$. A first way is by using the tensor product:

$$k_{\mathrm{prod}}(\mathbf{u}, \mathbf{u}') = k_{\mathrm{r}}\big(\rho, \rho'\big) k_{\mathrm{a}}(\theta, \theta') \tag{3.4}$$

where $\mathbf{u} = (\rho, \theta)$ and $\mathbf{u}' = (\rho', \theta')$ are in $\mathcal{C}$. This formulation implicitly assumes that the GP Z is the product of two independent components: a radial process R_ρ and an angular process A_θ ($Z_{\mathbf{u}} = R_\rho A_\theta$). It corresponds to a simple form of interaction. For processes that do not have interactions between these components ($Z_{\mathbf{u}} = R_\rho + A_\theta$), an additive kernel should be more appropriate:

$$k_{\mathrm{add}}(\mathbf{u}, \mathbf{u}') = k_{\mathrm{r}}\big(\rho, \rho'\big) + k_{\mathrm{a}}(\theta, \theta') \tag{3.5}$$

A trade-off between these two extreme approaches is the ANOVA kernel defined as:

$$k_{\mathrm{ANOVA}}(\mathbf{u}, \mathbf{u}') = \Big(1 + k_{\mathrm{r}}\big(\rho, \rho'\big)\Big)\Big(1 + k_{\mathrm{a}}(\theta, \theta')\Big) \tag{3.6}$$

The expanded form of Equation (3.6) shows that a process $Z_{\mathbf{u}}$ with ANOVA kernel can be viewed as a sum of four independent GPs: a constant process Z^0, a radial process R_ρ with kernel k_{r}, an angular process A_θ with kernel k_{a}, and a process Z^{inter} on $\mathcal{C}$ with kernel $k_{\mathrm{r}} k_{\mathrm{a}}$. From the ANOVA point of view, these processes are similar to constant term, main effects, and second-order interaction [31], but without respecting the unicity constraints such as centering. For more details on how to make new kernels from old, we refer the reader to [90].

Let us now define the kernels k_{r} on $(0, 1] \times (0, 1]$ and k_{a} on $\mathbb{S} \times \mathbb{S}$. We recall that valid kernels must be positive definite. The domain $(0, 1]$ is a segment of a 1-dimensional Euclidean space. As a consequence, traditional kernels are suitable for k_{r}. In particular, Matérn kernels are attractive for their ability to control the smoothness of the process and to ensure numerical stability. In dimension 1, the Matérn $\frac{5}{2}$ kernel is given by:

$$k_{\mathrm{m}}(x, x') = \left(1 + \frac{\sqrt{5} \mid x - x' \mid}{\ell} + \frac{5(x - x')^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5} \mid x - x' \mid}{\ell}\right) \tag{3.7}$$

A simple way of defining kernels on $\mathbb{S} \times \mathbb{S}$ is mentioned in [42]. They are based on the chordal distance $d_1(\theta, \theta') = 2 \sin\left(\frac{\theta - \theta'}{2}\right)$ and the geodesic distance $d_2(\theta, \theta') = \operatorname{acos}\big(\cos(\theta - \theta')\big)$ illustrated in Figure 3.2.

To define a kernel on $\mathbb{S} \times \mathbb{S}$, one could be tempted to compose usual kernels with d_1 or d_2. Unfortunately, positive definiteness is not guaranteed for the resulting functions when d_2 is used. As a counter-example, if the Gaussian kernel is chosen for k_{a}, then $k_{\mathrm{a}} \circ d_2$ is not positive definite ([42], Th. 8). Alternatively, two sufficient conditions of positive definiteness over $\mathbb{S} \times \mathbb{S}$ are provided by Gneiting [42]. Define F_d the class of continuous functions $\varphi : [0, \infty) \to \mathbb{R}$, with $\varphi(0) = 1$ and such that the function $(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \varphi(\|\mathbf{x} - \mathbf{x}'\|)$ is positive definite. Then:

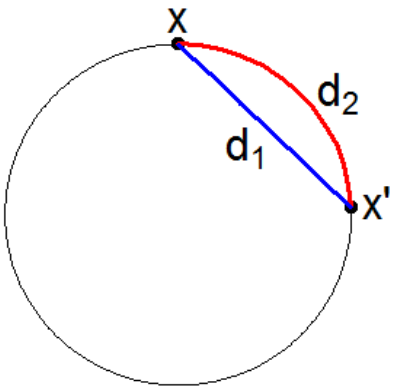


Figure 3.2 – Chordal (d_1) and geodesic (d_2) distances on $\mathbb{S}$.

(i). If $\varphi \in F_2$, then $\varphi \circ d_1$ is a kernel on $\mathbb{S} \times \mathbb{S}$.

(ii). If $\varphi \in F_1$ is such that $\varphi(t) = 0$ for $t \geq \pi$, then $\varphi \circ d_2$ is a kernel on $\mathbb{S} \times \mathbb{S}$.

Kernels satisfying (i) were initially proposed by Yadrenko in 1983 and are often used to describe periodic functions (see e.g. [90]). They correspond to restrictions of 2-dimensional isotropic GPs on $\mathbb{R}^2$ to $\mathbb{S}$. The second result is due to Lévy in 1961. Kernels satisfying (ii) can be constructed from compactly supported functions on $\mathbb{R}$ such as the C^2-Wendland function defined for $0 \leq t \leq \pi$:

$$W_c(t) = \left(1 + \tau \frac{t}{c}\right)\left(1 - \frac{t}{c}\right)_+^{\tau}, \qquad c \in (0, \pi]; \tau \geq 4 \tag{3.8}$$

For the geodesic distance, we use $c = \pi$, which is the largest possible value due to condition (ii) above. With this choice, the covariance between two angles θ, θ' is zero when $d_2(\theta, \theta') = \pi$, and strictly positive for $d_2(\theta, \theta') < \pi$. The same interpretation is possible for the chordal distance with $c = 2$, though it is not necessary to use a compactly supported function in that case. From now on, we will use the Wendland function in both cases, resulting in the two following kernels on $\mathbb{S} \times \mathbb{S}$:

$$k_{\text{chord}}(\theta, \theta') = W_2(d_1(\theta, \theta')), \tag{3.9}$$
$$k_{\text{geo}}(\theta, \theta') = W_\pi(d_2(\theta, \theta')), \tag{3.10}$$

and the corresponding GPs will be denoted *polar GP (chordal)* and *polar GP (geodesic)*.

GP simulations on the unit disk

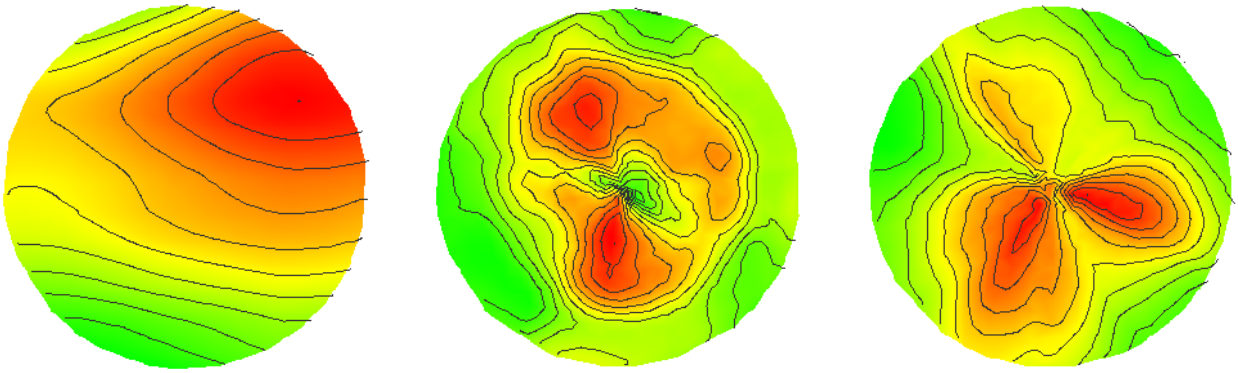
In order to better understand the specificities of polar GPs, it is useful to draw simulated surfaces. For the sake of simplicity, we propose to focus on the ANOVA combinations. We consider a Cartesian GP and the two polar GPs (chordal, geodesic) defined in Equations (3.9), (3.10). Their expressions are written below, including variance factors s^2, α_1^2, α_2^2:

(a) $k(\mathbf{x}, \mathbf{x}') = s^2\Big(1 + \alpha_1^2\, k_{\text{m}}\big(x, x'\big)\Big)\Big(1 + \alpha_2^2\, k_{\text{m}}\left(y, y'\right)\Big)$

(b) $k(\mathbf{x}, \mathbf{x}') = s^2\Big(1 + \alpha_1^2\, k_{\text{m}}\big(\rho, \rho'\big)\Big)\Big(1 + \alpha_2^2\, k_{\text{chord}}\left(\theta, \theta'\right)\Big)$

(c) $k(\mathbf{x}, \mathbf{x}') = s^2\Big(1 + \alpha_1^2\, k_{\mathrm{m}}(\rho, \rho')\Big)\Big(1 + \alpha_2^2\, k_{\mathrm{geo}}(\theta, \theta')\Big)$

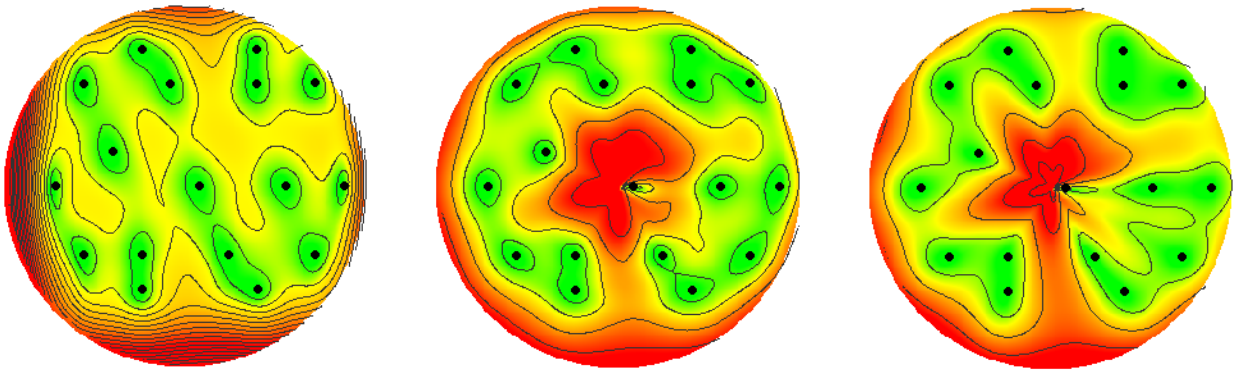
Simulation results are displayed in Figure 3.3. We can see that with polar GPs, the simulated surface exhibits radial and angular patterns around the center of the disk. Such kernels may thus be suitable to describe physical phenomena involving such effects. Figure 3.4 shows



Cartesian GP | Polar GP (chordal) | Polar GP (geodesic)

Figure 3.3 – Simulations of Cartesian and polar GPs with kernels (a)-(c).

via Kriging standard deviation how model uncertainty varies over $\mathcal{D}$, given a design of 17 points. Two striking differences are visible, especially between the Cartesian GP and the polar GP (geodesic), about uncertainty at the center of the disk, and uncertainty regions at the vicinity of design points. On one hand, the neighborhoods produced by the Cartesian GP look like elliptical regions at any location of the circular domain. On the other hand, those produced by the polar GP (geodesic) look like pie chart sectors, oriented towards the center of the disk, which plays a particular role. This is also true for the polar GP (chordal), though less pronounced.



Cartesian GP | Polar GP (chordal) | Polar GP (geodesic)

Figure 3.4 – Kriging standard deviations for Cartesian and polar GPs (kernels (a)-(c))

3.4 Applications

3.4.1 Quality control in microelectronics

In microelectronics, integrated circuits are produced on circular slices of semiconductor materials called wafers. For quality monitoring, physical and electrical variables are collected on a set of locations of these wafers. In this example, the characteristic of interest is thickness, a key parameter affecting performance of integrated circuits. In our industrial background, only 17 predefined points are measured for economic reasons. The statistical challenge consists in predicting non-measured locations in order to assess the spatial risk of default from this dataset. For the purpose of this study, thickness is further measured at 64 new locations to serve as a test grid. For the sake of confidentiality, the technological process is not detailed and the thickness values are rescaled. It produces here data with a pronounced radial pattern. However, we will not assume that the model is purely radial, which is a too strong assumption in practice, due to the numerous successive operations on a wafer, and the possible slacks in processing. The aim of this section is to compare the Cartesian and po-

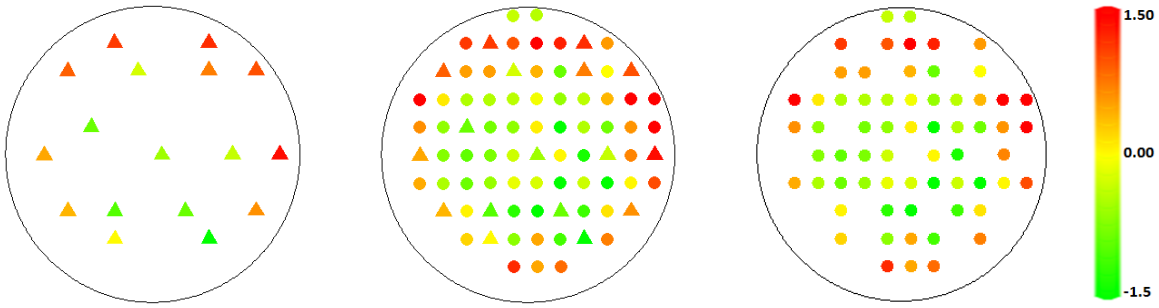


Figure 3.5 – Rescaled thickness values. The 81 measurement locations are shown in the middle, including 17 design points (triangles, left) and 64 test points (bullets, right).

lar GPs (chordal, geodesic), obtained with 3 types of algebraic combination (product, sum, ANOVA). The Cartesian GPs considered here are obtained by tensor product, tensor sum or ANOVA product of the 1-dimensional Matérn kernel of Equation (3.7). For the polar GPs, we use the same combinations for a kernel k_{r} on $(0, 1]$ and a kernel k_{a} on $\mathbb{S} \times \mathbb{S}$, accordingly to Equations (3.4), (3.5), (3.6). For k_{r}, we use again the Matérn kernel, whereas for k_{a} we choose k_{chord} or k_{geo} (see Equations (3.9), (3.10)). The range parameters τ and θ, as well as the variance factors s^2, α_1^2, α_2^2 are estimated by Maximum Likelihood (ML) with R package kergp [27]. The optimizer used is the method `L-BFGS-B` proposed in the `optim` function in R: an adaptation of the quasi-Newton method BFGS for boundary constraints. To improve its performances, we added a multistart step: 10 initial points were sampled at random, and for each of them a separate optimization was performed. The best result among the ten was finally chosen. The model accuracy is computed on the 64 test points, with the root mean squared error (RMSE) criterion. The results are summarized in Table 3.1 when μ is constant in Equation 3.1. They show that the smallest prediction errors are obtained with the polar GPs, corresponding to gains around 20% compared to the Cartesian GP. Adding Zernike polynomials as a trend slightly improves the result for the Cartesian GP, but the untrended polar GPs still outperform with a gain of 15%. Actually the trend captures the main part of the phenomenon and the GP part has then a minor effect: results are the same as for a pure linear model based on Zernike polynomials of order 2.

GP type	Cartesian			Polar (chordal)			Polar (geodesic)		
Kernel type	k_{prod}	k_{add}	k_{ANOVA}	k_{prod}	k_{add}	k_{ANOVA}	k_{prod}	k_{add}	k_{ANOVA}
RMSE	0.75 *	0.77	0.76	0.69	0.60 *	0.62	0.68	0.61 *	0.65

Table 3.1 – RMSE computed on 64 test points for several GPs with a constant trend. For each GP type, the combination resulting in the smallest RMSE is marked by an asterisk. When a Zernike trend is added, the best RMSE is equal to 0.71 for all GP types, corresponding to the score of the trend only.

In order to further analyze the results, we select for each GP type the kernels corresponding to the best combination, indicated by an asterisk in Table 3.1. The prediction surfaces obtained with these 3 kernels are shown on Figure 3.6. All the GPs succeed in recovering the radial pattern of the dataset, visible on Figure 3.5, middle. However, it is less faithfully identified by the Cartesian GP. The differences on the predicted values can be explained by thinking at the space in which the kernel is defined. For polar GPs, prediction at one location will particularly involve the locations corresponding to a high correlation according to ρ or θ. Typically, the resulting neighborhoods in $\mathcal{D}$ may look like pie chart sectors (high radial correlation) or ring portions (high angular correlation). Here, a closer look at estimated parameters reveals that there is a high angular correlation. Therefore, prediction at the bottom of the disk involves the other points that are close to the boundary. On the other hand, for the Cartesian GP, the predicted thickness has a low value, since the measurement points around, in the (x, y) space, have a low value. Finally notice that the predicted value at the extreme boundary of the disk should be considered with care, since no test points are defined on this region due to technical constraints.

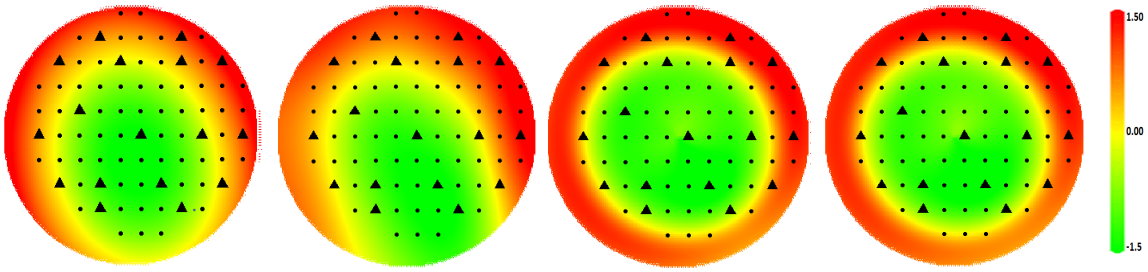
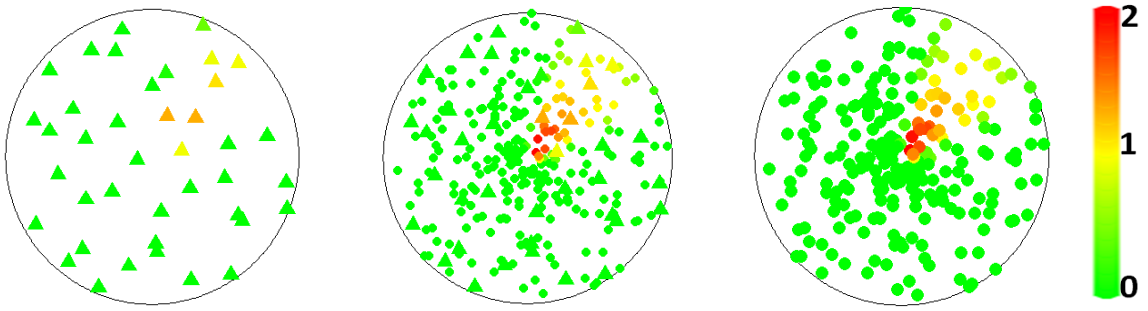


Figure 3.6 – Prediction surface for the best untrended GP models of Table 3.1. When adding a Zernike trend, the prediction surface is approximately the same as for a pure Zernike regression represented on the left. Black bullets correspond to test points, triangles to design points.

3.4.2 Air pollution modelling with a directional input

The problem tackled here is an environmental question. A greenhouse gas emitted by a known source, usually an industrial plant, is measured at a given location for air quality monitoring. In the absence of sensors, gas concentration must be predicted. For simple landscapes, analytical expressions are available based on transport and diffusion equations. However, for complex landscapes, gas concentration is simulated by numerical codes [7]. The

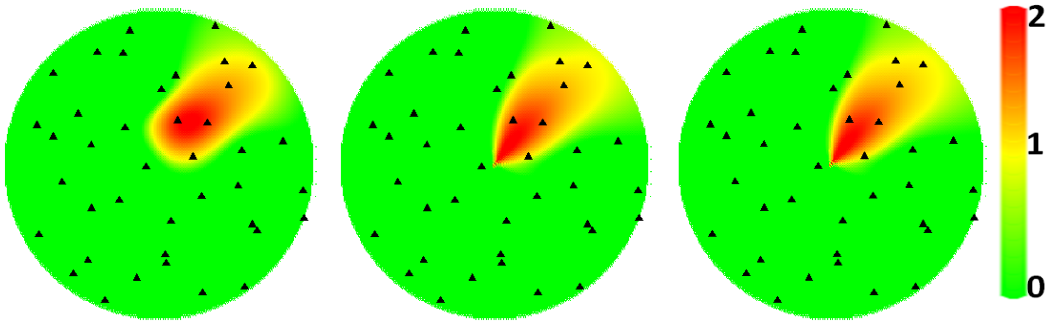
input variables include the emitted flow, landscape characteristics and meteorological variables. Here we focus on wind speed and wind direction. In this short study, 242 simulations were carried out, 30 of which serve as design points and the other ones are used for tests, as illustrated in Figure 3.7. The wind speed, initially given on the range $[0; 12]$ $(m.s^{-1})$, is rescaled to $[0, 1]$. With this transformation, the domain of the variables (speed, direction) is the unit disk. The aim of this study is simply to compare the prediction accuracy of Carte-



Design points. Design and test points. Test points.

Figure 3.7 – Rescaled gas concentrations. The 242 simulation locations are shown in the middle, including 30 design points (triangles, left) and 212 test points (bullets, right).

sian and polar GPs, without using a priori information. In particular, we do not specify the constraints of positivity or nullity of the gas concentration on a known subregion. We use the same kernels as in the first application, corresponding to 3 algebraic combinations (product, sum, ANOVA). Here, the best model is obtained for the tensor-product combination for all kinds of GPs. This claims in favor of an interaction speed-direction for the wind on gas concentration. Notice that adding a Zernike polynomial trend does not improve the results here, since the angular shape is restricted to a region of the disk, which is hard to capture with Zernike polynomials. The results are displayed in Figure 3.8. In terms of prediction accuracy (measured by the RMSE criterion) the polar GPs are clearly outperforming, corresponding to gains around 40% compared to the standard tensor-product Matérn kernel. Furthermore, for the polar GPs the influence of wind direction on gas concentration has an angular shape, which is intuitive, and corresponds to the true shape visible in Figure 3.7 (middle). On the other hand, this shape is rectangular for the Cartesian GP.



Cartesian GP
RMSE = 0.61

Polar GP (chordal)
RMSE = 0.38

Polar GP (geodesic)
RMSE = 0.37

Figure 3.8 – Estimated gas concentrations according to wind speed (ρ) and direction (θ), for untrended Cartesian and polar GPs. Adding a Zernike polynomial trend does not improve the results. Triangles correspond to design points.

3.5 Generalization to hyperballs

In computer experiments, the problem dimension is often higher than in spatial statistics, and the aim of this section is to investigate an extension of polar Gaussian processes in higher dimensions. More precisely, we investigate situations where the angular part of the inputs is in higher dimension. As an example, a force vector may be represented as a pair (magnitude, direction), where direction is a point on a sphere.

3.5.1 Polar Gaussian processes on hyperballs

Let us consider that the input domain is the unit d-dimensional ball $\mathbb{B}^d$ ($d > 1$), represented either in Cartesian coordinates by $\mathbb{B}^d = \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| \leq 1\}$, where $\|.\|$ is the Euclidian norm, or in spherical coordinates $(\rho, \theta_1, \ldots, \theta_{d-1})$. As in Section 3.3, we call Cartesian GP any restriction to $\mathbb{B}^d$ of usual GPs on $\mathbb{R}^d$. Polar GPs are generalized to $\mathbb{B}^d$ by using the product structure $\mathbb{B}^d = [0,1] \times \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| = 1\}$ denotes the $(d-1)$-sphere. Their kernels are obtained by combining the kernel k_{r} on $[0,1] \times [0,1]$ and a kernel k_{a} on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$.

A simple way to construct kernels on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ is to restrict a kernel on $\mathbb{R}^d \times \mathbb{R}^d$, remarking that positive definitiveness is preserved by restriction. This gives for instance the kernels defined with the chordal distance, $d_1(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$, i.e. $k_{\mathrm{a}}(\mathbf{u}, \mathbf{v}) = \phi(d_1(\mathbf{u}, \mathbf{v}))$ where $(\mathbf{x}, \mathbf{x}') \mapsto \phi(d_1(\mathbf{x}, \mathbf{x}'))$ is a kernel on $\mathbb{R}^d \times \mathbb{R}^d$. This also includes restriction of anisotropic kernels. For example, $(\mathbf{u}, \mathbf{v}) \mapsto \sigma^2 \exp\left(-\sum_{j=1}^d \left(\frac{u_j - v_j}{\ell_j}\right)^2\right)$ defines a kernel on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. The drawback of this construction is that it does not involve the geometry of the sphere: When distances define correlations, they lie on the Euclidian space $\mathbb{R}^d$ and not on the sphere.

A second way is to define a kernel on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ from a distance on the sphere. The theory is well developed for *isotropic* kernels, meaning that the covariance function depends only on the geodesic distance $d_2(\mathbf{u}, \mathbf{v}) = \mathrm{acos}(\cos\langle \mathbf{u}, \mathbf{v} \rangle)$. In this context, positive-definiteness is harder to meet. Thus, the approach used in Section 3.3 for $d = 2$ consisting in plugging d_2 in a compactly supported correlation function, is only valid for $d \leq 3$ [42]. For $d \geq 4$, conditions for positive-definiteness are provided in [42]. A first option is to plug the geodesic distance d_2 in a completely monotonic function, i.e. a function f admitting derivatives at any order and with alternate derivative signs: $(-1)^m f^{(m)} \geq 0$ for all integer m. As an example, $(\mathbf{u}, \mathbf{v}) \to \exp\left(-\frac{d_2(\mathbf{u}, \mathbf{v})}{\tau}\right), \tau > 0$ is a kernel on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. Another option is to use a correlation function which admits a representation as an infinite sum of cosine powers, called Gegenbauer expansion, with strictly positive coefficients (see [42] for more details). As an example, $\varphi_{\sin} \circ d_2$ is a kernel over $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ for $d \geq 2$, with $\varphi_{\sin}$ defined as:

$$\varphi_{\sin}(t) = 1 - \left(\sin\left(\frac{t}{2}\right)\right)^{\alpha}, \quad \alpha \in (0, 2) \tag{3.11}$$

3.5.2 Space-filling designs on hyperballs

We now aim at extending the space-filling designs considered in Section 5.2 to hyperballs. Let us first remark that there are two difficulties in extending the space-filling Latin cylinder designs. Indeed, when $d \geq 3$ the geometry of the hypersphere $\mathbb{S}^{d-1}$ is more complex and the mapping to an hypercube with boundary constraints (of the kind $2\pi = 0$) is not clear.

Furthermore, although the Φ_p maximin criterion can be generalized, its optimization in dimension nd, where n is the design size, seems much harder when d increases. For instance, when $d = 10$ and with the rule of thumb $n = 10d$, the optimization problem is in dimension 1000.
On the other hand, it is easy to simulate uniform designs on hyperspheres. A simple procedure described in [98] consists in remarking that if $\mathbf{X} \sim \mathcal{N}(0, I_d)$ then $\mathbf{T} = \frac{X}{\|\mathbf{X}\|}$ is uniform on $\mathbb{S}^{d-1}$. Furthermore if R is a random variable drawn independently and uniformly on $[0, 1]$ then $R^{1/d}\,\mathbf{T}$ is uniform on $\mathbb{B}^{d-1}$ (see e.g. [29], Theorem 2.2.1.) This extends the case of the disk (Section 5.2). Notice however that uniform designs on hyperballs may not be the best designs when radial or angular patterns are present: In the 2-dimensional case, we obtained better results when the radius R was sampled uniformly (Section **??**). This suggests two strategies:

- Common part: Simulate independently $R \sim \mathcal{U}[0, 1]$, $\mathbf{T} \sim \mathcal{U}(\mathbb{S}^{d-1})$
- Strategy 1 "$\mathcal{U}_{\mathbb{B}}$" (Uniform sampling on hyperballs): Compute $R^{1/d}\,\mathbf{T}$.
- Strategy 2 "$\mathcal{U}_{\mathrm{r}} \times \mathcal{U}_{\mathbb{S}}$" (Uniform sampling of radial and angular parts): Compute $R\,\mathbf{T}$.

3.5.3 Case study on toy functions

In order to investigate the behavior of polar GPs in a dimension higher than 2, we consider the following test functions:

- $f_1 : (x_1, \ldots, x_d) \mapsto \|\mathbf{x}\|^2$
- $f_2 : (\rho, \theta_1, \ldots, \theta_{d-1}) \mapsto \sum_{i=1}^{d-1} \cos(3\theta_i)$
- $f_3 : (x_1, \ldots, x_d) \mapsto \left(\sum_{i=1}^{d} x_i\right)^2$

The function f_1 is purely radial, and f_2 purely angular. On the other hand, f_3 does not exhibit any radial or angular pattern.
We perform numerical tests with $d = 10$. For each test function, three GP models with a constant trend are tested. Recall that k_{m} denotes the Matérn $\frac{5}{2}$ kernel (see Eq. 3.7). Then we consider:

- A Cartesian GP with a tensor-product kernel with a common characteristic length l: $\prod_{j=1}^{d} k_{\mathrm{m}}(x_j, x'_j; \ell)$.
- A polar GP based on chordal distance, with kernel $k = k_{\mathrm{r}} + k_{\mathrm{a}}$ where $k_{\mathrm{r}} = k_{\mathrm{m}}$ for the radius, and $k_{\mathrm{a}} = k_{\mathrm{m}} \circ d_1$ (restricted to $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$) for the angular part.
- A polar GP based on the geodesic distance, with kernel $k = k_{\mathrm{r}} + k_{\mathrm{a}}$, where $k_{\mathrm{r}} = k_{\mathrm{m}}$ for the radius, and isotropic sine power kernel (Eq. 3.11) for k_{a}.

Their kernels are denoted by $\mathrm{k}_{\mathrm{Cart}}$, $\mathrm{k}_{\mathrm{chord}}$ and $\mathrm{k}_{\mathrm{geo}}$. Notice that the other algebraic combinations among sum, product and ANOVA have been tried for all kernels, without modifying the conclusions. Moreover, the proposed kernels take into account the symmetry of the problem in their definition (isotropy, common parameter value per dimension) and thus depend on a very small number of parameters. The numerical likelihood maximization is then highly reliable, and was carefully done using ten different initial values.

Finally, the two design strategies presented in the previous section are applied. The design size is fixed to $n = 10d$. In order to assess model accuracy, the RMSE criterion is computed over a test set of size 1000, sampled uniformly in $\mathbb{B}^d$. For the sake of interpretability, the RMSE is shown as a percentage of the standard deviation of the output values on the test set.
Finally, the whole study is repeated $N = 100$ times, and the boxplot characteristics of the RMSE values over the N repetitions are shown in Table 3.2: median and interquartile values.

Function	$f_1(\mathbf{x}) = \|\mathbf{x}\|^2$			$f_2(\mathbf{x}) = \sum_{i=1}^{d-1} \cos(3\theta_i)$			$f_3(\mathbf{x}) = \left(\sum_{i=1}^{d} x_i\right)^2$		
Kernel	k_{Cart}	k_{chord}	k_{geo}	k_{Cart}	k_{chord}	k_{geo}	k_{Cart}	k_{chord}	k_{geo}
"$\mathcal{U}_{\mathbb{B}}$"	28.9 (6.2)	**0.0** (0.0)	**0.0** (0.0)	15.1 (10.0)	**8.1** (0.6)	**8.1** (0.6)	23.6 (5.3)	91.6 (8.2)	97.9 (8.6)
"$\mathcal{U}_{\mathrm{r}} \times \mathcal{U}_{\mathbb{S}}$"	14.2 (3.1)	**0.1** (0.2)	**0.2** (0.5)	11.1 (1.1)	**8.2** (0.8)	**8.1** (0.7)	**17.4** (6.6)	34.4 (9.9)	65.4 (14)

Table 3.2 – Model accuracy of three GP models and two design strategies on toy functions. Each experiment is repeated 100 times, and the median of the normalized RMSE (i.e. divided by the output standard deviation) is reported as well as the interquartile interval (into brackets).

We observe that polar GPs give better results for the two functions that exhibit a radial and angular pattern, and a worse result for the other one. In particular, predicting a radial function is done much more accurately with a polar GP. This may be explained by the reconstruction process and geometry considerations in high dimension. Indeed, as in the 2-dimensional case, polar GPs reconstruct a radial function by using the points located on closed concentric hypersphere (high angular correlation) whereas Cartesian GPs use the neighbors (in the sense of Euclidian norm) which are very few in high dimension (see e.g. [49], 2.5.). Notice that even if we double the number of experiments ($n = 20d$) to learn this radial function, the performance of the Cartesian GP with the best design strategy has a median RMSE equal to 5% (not shown in Table 3.2), which is still worse than polar GPs.
The results about design strategy on these toy functions are in favor of sampling uniformly the radial part, rather than sampling uniformly on the hyperball. Finally polar GP construction with chordal distance d_1 perform better than for geodesic distance d_2. In addition to the sine power kernel, we also tested the exponential kernel $\exp\left(-\frac{d_2(\mathbf{u},\mathbf{v})}{\tau}\right)$, but it gave worst results. However, other kernels could have been tried with a possible different conclusion, and a deeper investigation should be done in the future.

3.6 Discussion

We addressed the issue of analyzing costly computer or physical experiments on a disk. Such problems are encountered in various industrial applications, where the geometry of the disk is exploited for several technological processes involving rotations or diffusions from the center. For prediction purpose, we introduced so-called polar GP models that take into account the geometry of the disk both in their mean and covariance kernel. The new kernels are defined in polar coordinates. They are obtained as a combination of a kernel for the radius using an Euclidean distance, and a kernel for the angle, based on either chordal or geodesic distances on the unit circle. It was shown in two industrial examples where radial and angular patterns are visible that the approach significantly improves prediction. The

best algebraic combination was found to be either a tensor product or a tensor sum, which claims in favor of using a kernel mimicking the more general ANOVA decomposition [41]. Furthemore, in these applications there were only few differences in the results obtained with the polar GPs based on chordal or geodesic distances. This can be explained by the strong monotonic relationship between the chordal and geodesic distance. However, in theory the geodesic distance does not distort distances on the circle, and should be preferred. Finally, though not reported here, similar results were obtained with other kernel choices such as Matérn $\frac{3}{2}$ or exponential kernels for the Cartesian GP.

It is important to precise when polar GPs, based on distances on the unit circle, are relevant. One main difference between polar GPs and the usual ones, called here Cartesian GPs, is about the neighborhoods used for prediction. Since kernels of polar GPs are mapped to the polar space (ρ, θ), the prediction at one location will particularly involve the locations corresponding to a high radial or angular correlation with respect to ρ or θ. Typically, the resulting neighborhoods in the disk may look like pie chart sectors (high radial correlation) or ring portions (high angular correlation). This explains why polar GPs give more accurate predictions when there are radial or angular patterns, as may happen for technological processes that involve a rotation or a diffusion from the center. In other situations, involving for instance translations, Cartesian GPs may give better results. These two cases might correspond to the "two clusters of profiles over a circular grid" mentioned in [89] without any additional information about their origin. A knowledge of the process or historical data may help to choose which kernel is appropriate. In any case, there remains a lot of degrees of freedom about a GP model definition, concerning at least the trend shape or the different kernels corresponding to a given distance. To address this problem, aggregation techniques may be a solution.

Finally, we investigated an extension of the whole methodology to higher dimensions, replacing the disk by a hyperball. We performed empirical tests on several toy functions in 10-dimensions. Similar general conclusions hold, i.e. that polar GPs give better results for the functions that exhibit radial or angular patterns. In particular radial functions are much better reconstructed with polar GPs. This may be explained by the fact that in high dimensions points are located on the boundaries. Now, reconstruction with a polar GP involves points located on closed concentric hyperspheres (high angular correlation) while reconstruction with a Cartesian GP involves the neighbors, which are very few. Among other conclusions, kernels based on geodesic distance are here more difficult to handle, and on our first trials they performed worse than kernels based on chordal distance.

Résumé en Français

Traditionnellement définis avec la norme euclidienne, les noyaux de covariance n'exploitent pas les informations sur les procédés de fabrication. En particulier, les mécanismes de rotation et de diffusion à partir du centre génèrent d'importants effets radiaux et angulaires dans les données spatiales. Les processus gaussiens polaires que nous formalisons sur le disque intègrent les corrélations radiales et angulaires dans les modèles de krigeage et en améliorent les performances dans de telles situations. Ils permettent aussi de modéliser des simulateurs à entrées directionnelles à l'instar des modèles de dispersion atmosphérique.

En dimensions quelconque, le noyau d'un processus gaussien polaire se construit par combinaison algébrique d'un noyau sur l'espace du rayon polaire et d'un noyau sur l'espace des

angles, représenté par une sphère et donc muni de la distance cordale ou de la distance géodésique. Les avantages provenant de cette construction sont illustrés à travers des fonctions tests en dimensions 2 et 10.

Chapter 4

Linear models based on Gaussian processes

The Gaussian process models presented in the previous chapter rely on kernels formulated by algebraically combining one dimensional functions. Three kinds of combinations were tested for each GP type: product, addition and ANOVA. It turned out that the best choice was either the product or the addition in a case by case basis. The ANOVA kernel is a trade-off between the product and the addition. However, it cannot recover purely product or additive kernels. The purpose here is to formulate a more general GP model to automatically recover additive or product structures. For the purpose of sensitivity analysis, its kernel is designed to be a zero-mean function: a useful property for a simple computation of sensitivity indices. In addition, it allows to visualize the radial and angular effects.

4.1 Sobol-Hoeffding decomposition

Let $x_1, \ldots, x_d$ be d independent random variables and $D = \prod_{i=1}^{d} D_i$ with $D_i = [a_i, b_i]$, a hypercube of $\mathbb{R}^d$. Given $f \in L^2(D)$, f is uniquely decomposed as (see e.g. [32, 105, 57]):

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j}^{d} f_{ij}(x_i, x_j) + \cdots + f_{1\ldots d}(x_1, \ldots, x_d), \tag{4.1}$$

with $\mathbf{x} = (x_1, \ldots, x_d)$, f_0 a constant, and $\forall I \subseteq \{1 \ldots d\}$, f_I fulfills the centring condition

$$\mathbb{E}\big(f_I(x_I)\big) = 0, \tag{4.2}$$

and the non-simplification conditions

$$\mathbb{E}\big(f_{i_1 i_2}(x_{i_1}, x_{i_2}) \,|x_{i_1}\big) = \mathbb{E}\big(f_{i_1 i_2 i_3}(x_{i_1}, x_{i_2}, x_{i_3}) \,|x_{i_1}, x_{i_2}\big) = \cdots = 0. \tag{4.3}$$

This unicity condition can be rewritten:

$$\int_{D_i} f_I(x_I) d\nu_i(x_i) = 0 \quad \forall i \in I \tag{4.4}$$

Equation 4.1 is also refereed to as Functional ANOVA decomposition. The f_I's are recursively obtained:

$$\begin{aligned}
&f_0 = \mathbb{E}\big(f(\mathbf{x})\big) \\
&f_i(x_i) = \mathbb{E}\big(f(\mathbf{x}) \,|x_i\big) - f_0, \quad i = 1 \ldots d \\
&f_{ij}(x_i, x_j) = \mathbb{E}\big(f(\mathbf{x}) \,|x_i x_j\big) - f_i(x_i) - f_j(x_j) - f_0, \quad i, j = 1 \ldots d \\
&\ldots
\end{aligned}$$

More generally, $f_I(x_I) = \mathbb{E}\big(f(\mathbf{x})\,|x_I\big) - \sum_{J\subsetneq I} f_J(\mathbf{x_J})$. The non-simplification conditions imply orthogonality, which allows a variance decomposition:

$$\mathrm{var}(\mathbf{x}) = \sum_{i=1}^{d} \mathrm{var}\big(f_i(x_i)\big) + \sum_{i<j}^{d} \mathrm{var}\big(f_{ij}(x_i,x_j)\big) + \cdots + \mathrm{var}\big(f_{1\ldots d}(x_1,\ldots,x_d)\big). \tag{4.5}$$

The so-called partial variances $V_I = \mathrm{var}\big(f_I(x_I)\big)$ quantify the importance of f_I in f in terms of variance. By denoting the total variance by V, the ratios $\frac{V_I}{V}$ are called Sobol indices.

4.2 The model

Denote $D = \prod_{i=1}^{d} D_i$ with $D_i = [a_i, b_i]$, a hypercube of $\mathbb{R}^d$, endowed with the probability measure $\nu = \otimes_i^d \nu_i$. Given $I \subseteq \{1\ldots d\}$ and $\mathbf{x} = (x_1,\ldots,x_d)$ an element of $\mathbb{R}^d$, we call $\mathbf{x_I}$ the sub-vector of $\mathbf{x}$ corresponding to I. Following the Sobol decomposition of Gaussian random field paths introduced in [41], we consider the linear model with interactions:

$$\begin{aligned} Z(\mathbf{x}) &= \mu + \sum_{\substack{I\subseteq\{1\ldots d\}\\ |I|\leq 2}} Z_I(x_I) \\ &= \mu + Z_1(x_1) + \cdots + Z_d(x_d) + \sum_{i<j} Z_{ij}(x_i,x_j) \end{aligned} \tag{4.6}$$

where μ is constant, $Z_i(x_i) \sim GP(0,k_i)$ and $Z_{ij}(x_i,x_j) \sim GP\left(0, k_i \bigotimes k_j\right)$ are $d(d+1)/2$ Gaussian processes with continuous sample paths such that:

(i). The Z_I's are independent

(ii). $\int_{D_i} k_i(u,v)d\nu_i(u) = 0 \quad \forall v \in D_i$

Remark 1. In model (4), only second order interactions are considered to have a tractable number of terms. However, the model can easily be extended to the case $|I| \geq 2$.

2. Condition (ii) means that the k_i's are centred. It allows to interpret each term of the model independently of the others as we see now.

3. Contrarily to [41], the aim is to build a structured GP and not to decompose an existing one.

Proposition 4.2.1. *Equation* (4.6) *defines the Sobol decomposition of* Z*:*

$$\big(Z(\mathbf{x})\big)_I = Z_I(x_I), \quad \forall I \subseteq \{1\ldots d\} \tag{4.7}$$

Proof. In [41], it is proven that centred k_i's are equivalent to centred sample paths:

$$\int_{D_i} Z_i(x_i)d\nu_i(x_i) = 0$$

Given $I \subseteq \{1 \dots d\}$ such that $|I| \geq 2$, and $i \in I$ we have:

$$\begin{aligned}
\operatorname{var}\left(\int_{D_i} Z_I\left(x_I\right) d\nu_i(x_i) \right) &= \operatorname{cov}\left(\int_{D_i} Z_I\left(x_I\right) d\nu_i(x_i), \int_{D_i} Z_I\left(y_I\right) d\nu_i(y_i) \right) \\
&= \iint_{D_i^2} \operatorname{cov}\Big(Z_I\left(x_I\right), Z_I\left(y_I\right) \Big) d\nu_i(x_i) d\nu_i(y_i) \\
&= \iint_{D_i^2} k_I\left(x_I, y_I\right) d\nu_i(x_i) d\nu_i(y_i) \\
&= \left(\prod_{j \in I \setminus i} k_j(x_j, y_j) \right) \iint_{D_i^2} k_i\left(x_i, y_i\right) d\nu_i(x_i) d\nu_i(y_i) = 0
\end{aligned}$$

In addition, Z_I is centred and thus $\mathbb{E}\left[\left(\int_{D_i} Z_I\left(x_I\right) d\nu_i(x_i) \right)^2 \right] = 0$. As a consequence, $\left(\int_{D_i} Z_I\left(x_I\right) d\nu_i(x_i) \right)^2 = 0$ since the sample paths of Z_I are continuous. Sobol's uniqueness condition is then fulfilled and (4.6) defines the functional ANOVA decomposition of Z. □

Proposition 4.2.2. *Sobol decomposition of the Kriging mean*
Let $m(\mathbf{x})$ be the Kriging mean at $\mathbf{x}$, based on the DoE $\left(\mathbf{x^{(1)}}, \dots, \mathbf{x^{(n)}}\right)$ and the response vector $\mathbf{y}$. Denote $\mathbf{K} = \left(k\left(\mathbf{x^{(i)}}, \mathbf{x^{(j)}}\right)\right)_{1 \leq i,j \leq n}$ the covariance matrix of the Kriging model, and $\mathbf{k_I}(\mathbf{x_I}) = \left(k_I\left(\mathbf{x_I}, \mathbf{x_I}^{(\mathbf{i})}\right)\right)_{i=1,\dots,n}$ the covariance vectors of the sub-models corresponding to Z_I's: $k_I(x_I) = \prod_{i \in I} k_i(x_i)$. Then, the Sobol decomposition of $m(\mathbf{x})$ is:

$$m(\mathbf{x}) = \mu + \sum_{\substack{I \subseteq \{1 \dots d\} \\ |I| \leq 2}} m_I(\mathbf{x_I}) \tag{4.8}$$

and

$$var\left(m_I\left(\mathbf{x}\right)\right) = \alpha^\top \mathbf{\Gamma}_I \, \alpha \tag{4.9}$$

where $\alpha = \mathbf{K}^{-1}\left(\mathbf{y} - \mu\right)$, $m_I(\mathbf{x_I}) = \alpha^\top \mathbf{k_I}(\mathbf{x_I})$ and $\mathbf{\Gamma}_I = \bigodot_{i \in I} \mathbf{\Gamma}_i$, with $\mathbf{\Gamma}_i = \int_{D_i} \mathbf{k_i}(x_i) \mathbf{k_i}(x_i)^\top d\nu_i(x_i)$.

The variances defined by Equation (4.9) represent the relative contributions of the sub-models Z_I's to the total variance of the output. They correspond to unscaled Sobol indices.

Proof. (Adapted from [31])
By independence of the Z_I's, $\mathbf{k} = \sum_{|I| \leq 2} \mathbf{k_I}(\mathbf{x_I})$. Thus, the Kriging mean at $\mathbf{x}$ is:

$$\begin{aligned}
m(\mathbf{x}) = \mu + \alpha^\top \mathbf{k}(\mathbf{x}) &= \mu + \alpha^\top \left(\sum_{I \subseteq \{1 \dots d\}, |I| \leq 2} \mathbf{k_I}(\mathbf{x_I}) \right) \\
&= \mu + \sum_{\substack{I \subseteq \{1 \dots d\} \\ |I| \leq 2}} m_I(\mathbf{x_I}
\end{aligned}$$

with $m_I(\mathbf{x_I}) = \alpha^\top \mathbf{k_I}(\mathbf{x_I}) = \mathbf{k_I}(\mathbf{x_I})^\top \alpha$. For the first point of the proposition, we conclude by uniqueness of Sobol decomposition since:

$$\begin{aligned}\forall i \in I, \quad \int_{D_i} m_I(x_I)\, d\nu_i(x_i) &= \alpha^\top \int_{D_i} \mathbf{k_I}(\mathbf{x_I}) d\nu_i(x_i) \\ &= \alpha^\top \left(\prod_{j \in I \setminus i} k_j(x_j) \right) \int_{D_i} k_i(x_i) d\nu_i(x_i) = 0\end{aligned}$$

For the second point, as $m_I(\mathbf{x})$ is centred, $\mathrm{var}(m_I(\mathbf{x})) = \mathrm{E}\left(m_I(\mathbf{x})^2\right)$. So,

$$\begin{aligned}\mathrm{var}(m_I(\mathbf{x})) &= \alpha^\top \left(\mathrm{E}\Big(\mathbf{k_I}(\mathbf{x_I}) \mathbf{k_I}(\mathbf{x_I})^\top \Big) \right) \alpha \\ &= \alpha^\top \left(\int_{D_I} \bigodot_{i \in I} \mathbf{k_i}(x_i) \mathbf{k_i}(x_i)^\top d\nu_I(x_I) \right) \alpha \\ &= \alpha^\top \bigodot_{i \in I} \left(\int_{D_i} \mathbf{k_i}(x_i) \mathbf{k_i}(x_i)^\top d\nu_i(x_i) \right) \alpha\end{aligned}$$

□

Practical aspects

In practice, we use GPs of the form $Z_I = \beta_I Y_I$, with $\beta_I \geq 0$ and $\int_{D_I} \mathrm{var}(Y_I(x_I))\, d\nu_I(x_I) = 1$. In this way, the Y_I's represent a set of elementary patterns with the same order of magnitude, and the β_I's represent their influence in the total variability. For stationary processes, the condition $\int_{D_I} \mathrm{var}(Y_I(x_I))\, d\nu_I(x_I) = 1$ is equivalent to $k_i(x_i, x_i) = 1$ for $i \in \{1 \dots d\}$.

Moreover, since the additive structure of the process is not always satisfied by the data, one should add an independent Gaussian noise with variance ε to the model via the "nugget" (see chapter 3). The model (4.6) becomes:

$$Y(\mathbf{x}) = \mu + \sum_{\substack{I \subseteq \{1 \dots d\} \\ |I| \leq 2}} \beta_I Y_I(x_I) + \eta, \quad \eta \sim N(0, \varepsilon^2), \beta_I \geq 0 \tag{4.10}$$

4.3 Making zero mean kernels from old

4.3.1 Centred kernels on segments

In this paragraph, we focus on the case of dimension 1. Denote $Z \sim GP(0, k)$ on $[a, b]$ and ν a probability measure on $[a, b]$. The purpose is to find a centred kernel k^0 on $[a, b]^2$.

Proposition 4.3.1. *If k is a kernel on $[a, b]^2$, then*

$$k^*(u, v) = k(u, v) - \int_{[a,b]} k(u, v) d\nu(u) - \int_{[a,b]} k(u, v) d\nu(v) + \iint_{[a,b]^2} k(u, v) d\nu(u) d\nu(v)$$

and

$$k^\dagger(u, v) = k(u, v) - \frac{\int_{[a,b]} k(u, v) d\nu(u) \int_{[a,b]} k(u, v) d\nu(v)}{\iint_{[a,b]^2} k(u, v) d\nu(u) d\nu(v)}$$

are centred kernels on $[a, b]^2$.

Proof. The proof is straightforward by direct calculations. Notice that these two functions are respectively the kernels of $Z(x) - \int_a^b Z(x)d\nu(x)$ and $Z(x) - \mathrm{E}\left(Z(x) \mid \int_a^b Z(x)d\nu(x)\right)$. The second one was introduced by [31]. □

Examples of centred kernels based on the uniform measure over [a, b]

Given a kernel k over $[a,b]^2$, Proposition 4.3.1 provides two procedures to center k. With respect to the uniform measure over [a, b], their use is only subject to computing

$$I_1(v) = \frac{1}{b-a}\int_{[a,b]} k(u,v)du \text{ and } I_2 = \frac{1}{(b-a)^2}\iint_{[a,b]^2} k(u,v)dudv.$$

These integrals are provided for some usual kernels by [41].

- Exponential kernel: $k(u,v) = \exp\left(-\frac{|u-v|}{\ell}\right)$
  - $I_1(v) = \omega\Big(2 - k(a,v) - k(v,b)\Big)$
  - $I_2 = 2\omega\left(1 - \omega + \omega\exp\left(-\frac{1}{\omega}\right)\right)$

  with $\omega = \frac{\ell}{b-a}$.
- Matérn$_{\frac{3}{2}}$: $k(u,v) = \left(1 + \frac{\sqrt{3}|u-v|}{\ell}\right)\exp\left(-\frac{\sqrt{3}|u-v|}{\ell}\right)$.
  - $I_1(v) = \omega_3\Big[4 - A\left(\frac{1}{\omega_3}\frac{y-a}{b-a}\right) - A\left(\frac{1}{\omega_3}\frac{b-y}{b-a}\right)\Big]$
  - $I_2 = 2\omega_3\Big[2 - 3\omega_3 + (1+3\omega_3)\exp\left(-\frac{1}{\omega_3}\right)\Big]$

  with $A(x) = (2+x)e^{-x}$ and $\omega_3 = \frac{\ell}{\sqrt{3}(b-a)}$.
- Matérn$_{\frac{5}{2}}$: $k(u,v) = \left(1 + \frac{\sqrt{5}|u-v|}{\ell} + \frac{5(|u-v|)^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}|u-v|}{\ell}\right)$
  - $I_1(v) = \frac{\omega_5}{3}\Big[4 - B\left(\frac{1}{\omega_5}\frac{y-a}{b-a}\right) - B\left(\frac{1}{\omega_5}\frac{b-y}{b-a}\right)\Big]$
  - $I_2 = \frac{1}{3}\omega_5\left(16 - 30\omega_5\right) + \frac{2}{3}\Big(1 + 7\omega_5 + 15\omega_5^2\Big)\exp\Big(-\frac{1}{\omega_5}\Big)$

  with $B(x) = \left(8 + 5x + x^2\right)e^{-x}$ and $\omega_5 = \frac{\ell}{\sqrt{5}(b-a)}$.

Examples of centred kernels based on the density function $f(u) = 2u$ over [0, 1]

Recall that if R, T are independent random variables with uniform distribution on $[0,1]$ and $[0,2\pi]$ respectively, then $(\sqrt{R}, T)$ is uniform on $\mathcal{D}$. Keeping in mind the need to use the uniform measure over the disk, we consider a second probability measure with density $f(u) = 2u$, which is the probability density function of $\sqrt{R}$ over $[0,1]$. The following two integrals are then needed to center kernels over $[0,1]^2$ with respect to this measure:

$$J_1(v) = 2\int_{[0,1]} k(u,v)udu \text{ and } J_2 = 4\iint_{[0,1]^2} k(u,v)uvdudv.$$

Now, we provide J_1 and J_2 for some usual kernels. The integrals have been computed with the help of the formal calculus software "Xcas", and numerically checked.

- Exponential:
  - $J_1(v) = 2\ell\Big(2y + \ell k(0,y)\Big) - 2\Big(\ell + \ell^2\Big)k(1,y)$
  - $J_2 = 8\ell^4\Big(1 - k(0,1)\Big) - 8\ell^3 k(1,0) - 4\ell^2 + \frac{8}{3}\ell$
- Matérn$_{\frac{3}{2}}$:
  - $J_1(v) = 8y\ell + 2\,(y + 3\ell)\,\ell e^{-\frac{y}{\ell}} + 2\big(\,(\ell+1)\,(y - 3\ell) - 1\big)e^{-\frac{1-y}{\ell}}$
  - $J_2 = 4\big(\frac{4}{3} - 3\ell + 10\ell^3\big)\ell - 8\ell^2\,(1 + 5\ell + 5\ell^2)\,e^{-\frac{1}{\ell}}$
- Matérn$_{\frac{5}{2}}$:
  - $J_1(v) = \frac{32}{3}\ell y + \frac{2}{3}\Big(15\ell^2 + 7\ell y + y^2\Big)e^{-\frac{y}{\ell}}$
    $- \frac{2}{3\ell}\Big(15\ell^3 + (1+\ell)y^2 - y(7\ell^2 + 7\ell + 2) + 6\ell + 1\Big)e^{-\frac{1-y}{\ell}}$
  - $J_2 = 4\ell\Big(\frac{16}{9} - 5\ell + \frac{70}{3}\ell^3 - \frac{2}{3}\big(1 + 10\ell + 35\ell^2\,(1+\ell)\,\big)e^{-\frac{1}{\ell}}\Big)$

4.3.2 Centred kernels on the circle

Proposition 4.3.2. *Case of isotropic periodic kernels*

Let $T > 0$, λ the uniform measure over $[0,T]$, and $\varphi : [0,\infty) \to \mathbb{R}$ a T-periodic function such that $k : (u,v) \mapsto \varphi(\,|\,u - v\,|\,)$ is a kernel on $[0,T]^2$. Then, $\int_0^T k(u,v)d\lambda(u)$ is constant. As a consequence, $k(u,v) - s$ is a centred kernel on $[0,T]^2$, and we have $k^*(u,v) = k^\dagger(u,v) = k(u,v) - s$ with $s = \int_0^T \varphi(t)d\lambda(t)$.

Proof. By T-periodicity,

$$\begin{aligned}\int_0^T \varphi\,(|\,u - v\,|\,)\,d\lambda(u) &= \int_v^{T+v} \varphi(u - v)d\lambda(u)\\ &= \int_0^T \varphi(t)d\lambda(t) = s, \text{ by uniformity of } \lambda\end{aligned}$$

Remark that s corresponds to the three integral terms of k^* and $k^\dagger$ of proposition 4.3.1, leading to the second part of the proposition. □

Examples

From proposition 4.3.2, centred kernels based on the geodesic distance on the circle can be obtained by centring compactly supported correlation functions.

4.4 Simulations and applications

In addition to the multiplicative, additive and ANOVA combinations used in chapter 3, we consider a fourth construction, mimicking model (4.6). The resulting kernel, denoted by FAD in reference to Functional ANOVA Decomposition, is given by:

$$k_{\text{FAD}}\,(\mathbf{u},\mathbf{u}') = \sigma_1^2 k_1^0\big(u_1,u_1'\big) + \sigma_2^2 k_2^0\,(u_2,u_2') + \sigma_{12}^2 k_1^0\big(u_1,u_1'\big)k_2^0\,(u_2,u_2') \tag{4.11}$$

Correlation functions	Analytic expressions	Parameters	$s = \bar{\varphi}$
Sine power	$\varphi(t) = 1 - \left(\sin(\frac{t}{2})\right)^{\alpha}$	$\alpha \in (0,2)$	$1 - \frac{\sqrt{\pi}}{\pi}\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}+1\right)}$
Askey	$\varphi(t) = \left(1 - \frac{t}{c}\right)_+^{\tau}$	$\tau \geq 2$	$\frac{c}{\pi(\tau+1)}$
C^2-Wendland	$\varphi(t) = \left(1 + \tau\frac{t}{c}\right)\left(1 - \frac{t}{c}\right)_+^{\tau}$	$c \in (0,\pi]; \tau \geq 4$	$\frac{2c}{\pi(\tau+2)}$
C^4-Wendland	$\varphi(t) = \left(1 + \tau\frac{t}{c} + \frac{\tau^2-1}{3}\frac{t^2}{c^2}\right)(1 - \frac{t}{c})_+^{\tau}$	$c \in (0,\pi]; \tau \geq 6$	$\frac{8c}{3\pi(\tau+3)}$

Table 4.1 – Some compactly supported correlation functions φ and their mean s such that $(\theta, \theta') \mapsto \varphi\big(\operatorname{acos}(\cos(\theta - \theta'))\big) - s$ is a centred kernel on $\mathbb{S} \times \mathbb{S}$.

where $\mathbf{u} = (\rho, \theta)$ for polar GPs and $\mathbf{u} = (x, y)$ for Cartesian GPs. The notations k_1^0 and k_2^0 mean that the kernels are centred. Polar GPs are centred in two ways. The first one uses the uniform measure over $[0,1]$ for polar radius, and the uniform measure over $\mathbb{S}$ for angles. The resulting kernel, corresponding to uniform weights in the space of polar coordinates, is denoted by $k_{\text{FAD:}[0,1]\times\mathbb{S}}$. The second one is obtained with the density function $f(u) = 2u$ over $[0,1]$ for polar radius, and the uniform measure over $\mathbb{S}$ for angles. It corresponds to uniform weights over the disk and the corresponding kernels are denoted by $k_{\text{FAD:}\mathcal{D}}$. Regarding Cartesian GPs, the disk does not have a product structure with respect to xy-axes. Therefore, we consider the restriction to the disk of centred Cartesian GPs over $[-1,1] \times [-1,1]$. Their kernels are denoted by $k_{\text{FAD:}[-1,1]^2}$.

We also assess a fifth kind of kernel, involving an additive and a multiplicative parts, but without centring conditions:

$$k_{\text{prod+add}}(\mathbf{u}, \mathbf{u}') = \sigma_1^2 k_1\big(u_1, u_1'\big) + \sigma_2^2 k_2\left(u_2, u_2'\right) + \sigma_{12}^2 k_1\big(u_1, u_1'\big) k_2\left(u_2, u_2'\right) \tag{4.12}$$

Actually, Equation (4.11) corresponds to model (4.10) with $\sigma_{\text{I}}^2 = \beta_I$ and Equation (4.12) represents its non-centred version. Both have the advantage of becoming additive if $\sigma_{12}^2 = 0$ or multiplicative if $\sigma_1^2 = \sigma_2^2 = 0$.

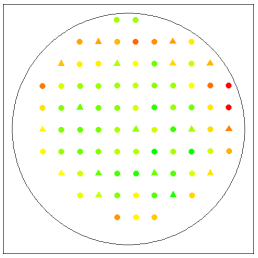
4.4.1 Simulations

In this section, different GP models are compared, based on 6 analytical toy functions. Among these models, the centred GPs are interpreted based on Sobol decomposition (Equations (4.9) and 4.8). The computation settings and results are detailed in Appendix 10.5.

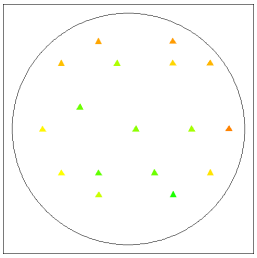
4.4.2 Applications

This section is dedicated to the industrial datasets in chapter 3, Sections 3.4.1 and 3.4.2. The matérn$\frac{5}{2}$ kernel is used over $[-1,1]^2$ and $[0,1]^2$, and the C^2-Wendland kernel over $\mathbb{S}$. Through these two datasets, we will see that Model (4) meets our expectations by improving Kriging models and allowing to visualize radial and angular effects. For instance, in the microelectronics example, $k_{\text{fad:}\mathcal{D}}$ leads to the best predictions whereas Cartesian GPs are the worst.

Application in microelectronics



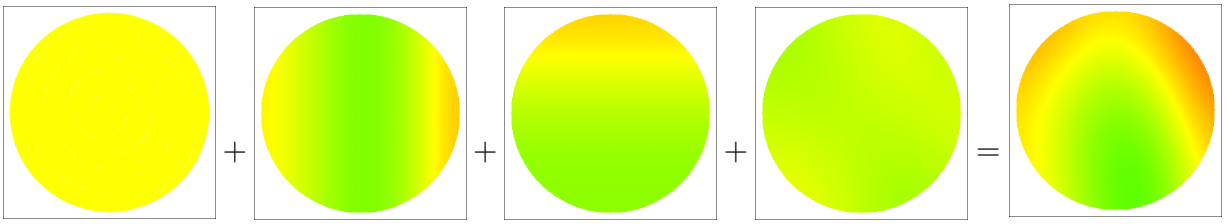
Dataset



DoE

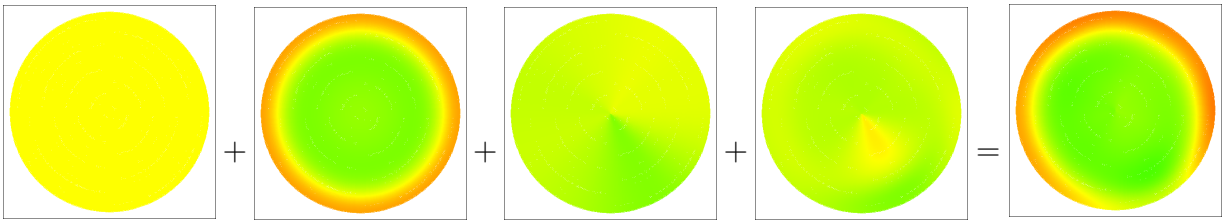
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	0.82	-20
Cartesian	$k_{\text{fad:}[-1,1]^2}$	0.84	-19
Polar	$k_{\text{prod+add}}$	0.66	-16
Polar	$k_{\text{fad:}[0,1]\times\mathbb{S}}$	0.71	-16
Polar	$k_{\text{fad:}\mathcal{D}}$	0.66	-15

Assessment of different GP models



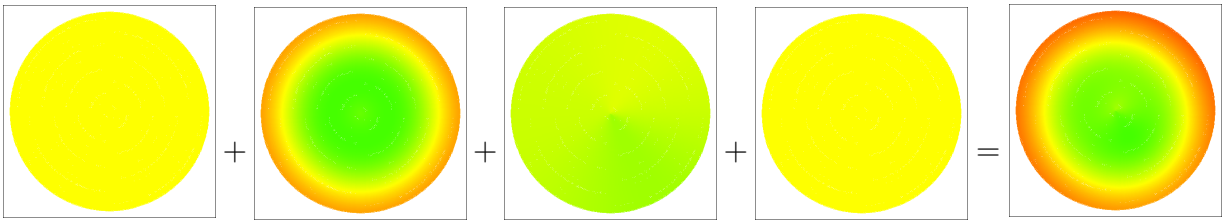
$\mu(x,y)$	$m_1(x)$	$m_2(y)$	$m_{12}(x,y)$	$m(x,y)$
–	$S = 46$	$S = 52$	$S = 2$	Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 85$	$S = 8$	$S = 7$	Kriging mean

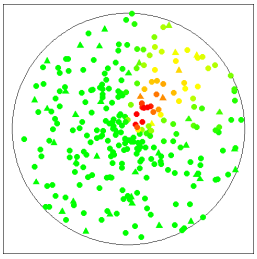
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



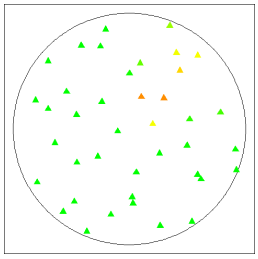
$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 74$	$S = 4$	$S = 22$	Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

By giving the same importance to $m(x,\theta)$ and $m(y,\theta)$, the Cartesian GP predictions are consistent with the pronounced radial effect which is observed in the data. However, $m(y,\theta)$ is not fully recovered because the y-axis is not fully filled by design points. Polar GPs overcome this problem by giving a high importance to the radial sub-model.



Dataset

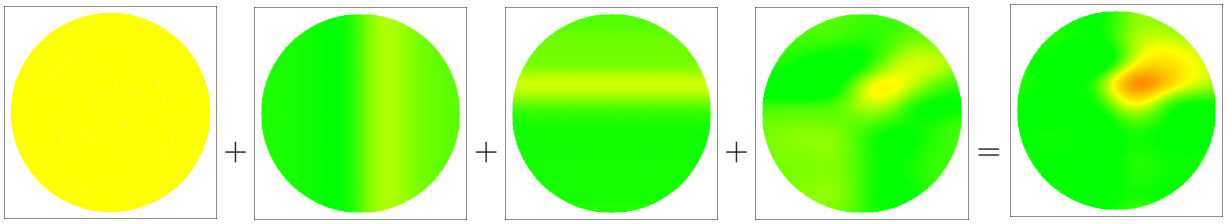


DoE

GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	0.58	-26
Cartesian	$k_{\text{fad:}[-1,1]^2}$	0.63	-29
Polar	$k_{\text{prod+add}}$	0.36	-11
Polar	$k_{\text{fad:}[0,1]\times\mathbb{S}}$	0.38	-16
Polar	$k_{\text{fad:}\mathcal{D}}$	0.32	-5

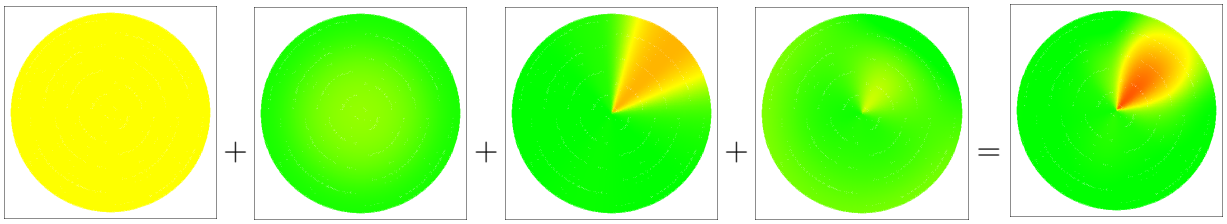
Assessment of different GP models

Application to air quality monitoring



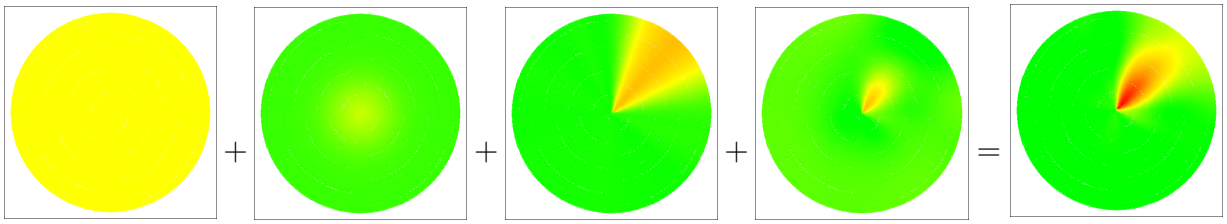
$\mu(x,y)$ — | $m_1(x)$ $S=26$ | $m_2(y)$ $S=29$ | $m_{12}(x,y)$ $S=45$ | $m(x,y)$ Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$ — | $m_1(\rho)$ $S=8$ | $m_2(\theta)$ $S=81$ | $m_{12}(\rho,\theta)$ $S=11$ | $m(\rho,\theta)$ Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1]\times\mathbb{S}$)



$\mu(x,y)$ — | $m_1(\rho)$ $S=6$ | $m_2(\theta)$ $S=80$ | $m_{12}(\rho,\theta)$ $S=14$ | $m(\rho,\theta)$ Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Here too, polar GPs and $k_{\text{fad:}\mathcal{D}}$ in particular give the best results.

Résumé en Français

L'objectif de ce chapitre est double: réaliser une analyse de sensibilité pour interpréter les modèles de krigeage et formaliser une famille générique de processus gaussiens qui englobe les noyaux de type additif, multiplicatif et ANOVA. Les modèles résultants sont définis sur un domaine hypercubique muni d'une mesure produit. Ils comprennent des composantes additives indépendantes et des termes d'interaction, tous centrés au sens de Sobol.

Les modèles de krigeage ainsi définis s'interprètent en termes de décomposition de Sobol. Les indices de sensibilité qui en sont déduites quantifient l'importance des effets radiaux et angulaires dans les processus gaussiens polaires. En plus de combler les attentes espérées, ces modèles donnent de meilleurs résultats que les processus gaussiens usuels, notamment lorsque la mesure d'intégration utilisée pour leur centrage est adaptée.

PART II

Design of experiments

Chapter 5

Maximin Latin Cylinders

In the first part of this thesis, different response surface models were proposed over the disk. Despite their relevancy, their performance can be worsen if a suitable design of experiments is not chosen. The question of designs of experiments is now addressed, and this first chapter focuses on static designs.

The results of this chapter are based on the contribution "Polar Gaussian Processes for Predicting on Circular Domains" [81], by Padonou and Roustant, in revision for SIAM/ASA *Journal on Uncertainty Quantification*.

5.1 Some usual designs on the disk

Among the DoEs that are specific to the disk, there are optimal designs for Zernike polynomials and spirals.

5.1.1 D-Optimal designs for Zernike polynomials

The D-optimal designs were investigated in [26] and were found to be contained in few concentric circles, as illustrated in Figure 5.1. D-optimal DoEs for regression models are not

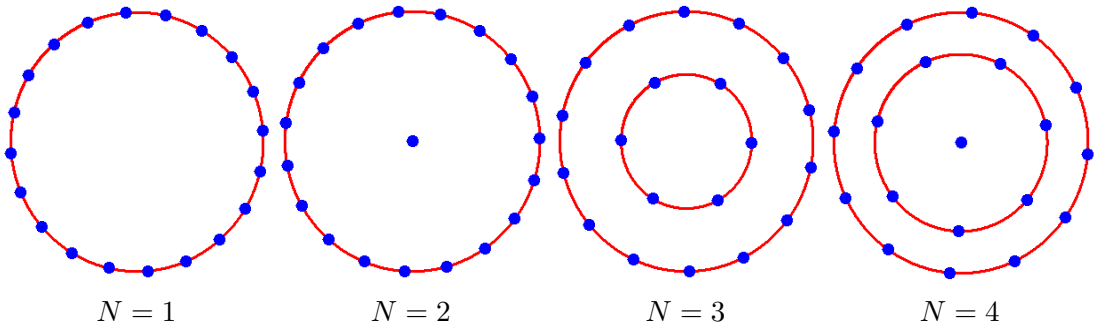
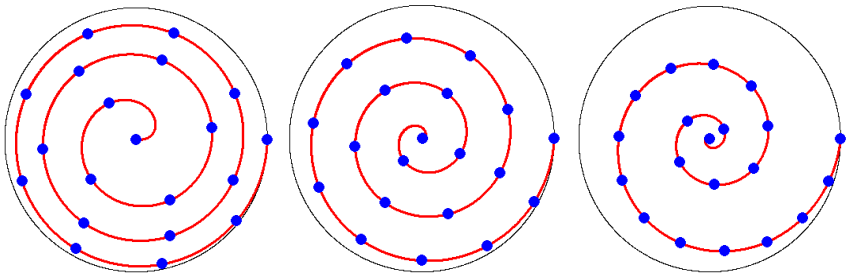


Figure 5.1 – 20-point D-optimal DoEs for Zernike polynomials of degree N.

robust to departures from the assumed shapes [53], and do not fill the space, a property usually required in the framework of GP modelling for capturing potential non-linearities.

5.1.2 Spirals

Spirals, hereafter denoted spiral DoEs, are used in various industrial settings: microelectronics, optics, microbiology, etc. They allow to control the density of the design (see e.g. [79]). Some of them are represented in Figure 5.2, corresponding to the equation $\rho = a\theta^p + b$.



Fermat: $s = \frac{1}{2}$ Archimede: $s = 1$ Galilee: $s = 2$

Figure 5.2 – 20-point DoEs defined from spirals of the form $\rho = a\theta^s + b$ with $\theta \in [0, 6\pi]$. The parameter s controls the speed with which the curve moves away from the center, and a, b are chosen such that the spirals start at the center and end at the boundary.

Poor space-filling properties are also visible for spirals in the space (ρ, θ) of polar coordinates, as shown in Figure 5.3, though they may correctly fill the disk.

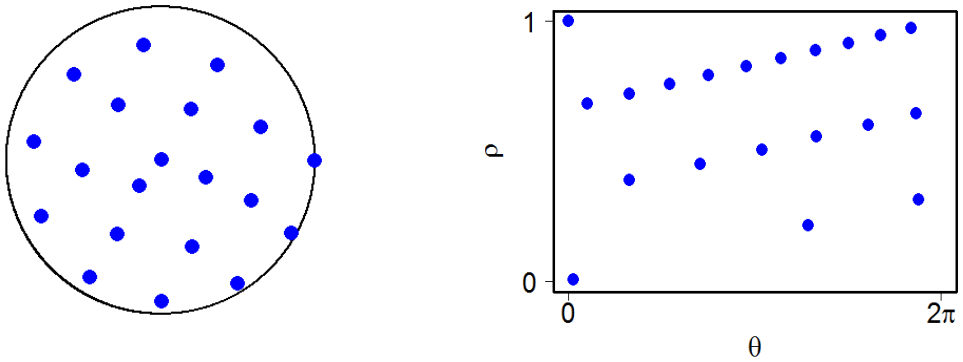


Figure 5.3 – Cartesian (left) and polar (right) representations of the Archimedean spiral DoE. This DoE is filling well the disk but not the cylinder of polar coordinates.

5.2 Maximin Latin hypercubes for polar coordinates

For metamodelling a potentially complex phenomenon, two main properties are expected from a good DoE: Space-filling, in order to capture non-linearities, and uniformity of the marginal distributions, to avoid redundancies in projection. Among the indicators used to assess space-fillingness, the maximin criterion [77] is a common choice. In addition, Latin hypercube designs (LHD, [74]) provide good projection properties onto marginal dimensions. Thus, maximin LHDs are often proposed as initial DoEs. However such designs cannot be directly used in polar coordinates, due to the non-Euclidean structure of $\mathcal{C}$. The aim of this section is to adapt their construction.

Let us first recall the construction of a maximin LHD over the hypercubic domain $[0, 1]^2$. Given a design $\mathbf{X} = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right)$ of elements of $[0, 1]^2$, we denote $\Phi_{\text{Mn}}(\mathbf{X})$ so-called

maximin criterion, giving the minimal distance among design points:

$$\Phi_{\text{Mn}}(\mathbf{X}) = \min_{i \neq j} \left(\| \mathbf{x}^{(\mathbf{i})} - \mathbf{x}^{(\mathbf{j})} \| \right) \tag{5.1}$$

A maximin DoE is a design that maximizes Φ_{Mn}. However, Φ_{Mn} is hard to optimize and a regularized version Φ_p, more suitable for optimization, was proposed in [78]:

$$\Phi_p(\mathbf{X}) = \left(\sum_{1 \leq i < j \leq n} \| \mathbf{x}^{(\mathbf{i})} - \mathbf{x}^{(\mathbf{j})} \|^{-p} \right)^{\frac{1}{p}} \tag{5.2}$$

For $p \to \infty$, maximizing Φ_{Mn} is equivalent to minimizing Φ_p. Following [78, 22], we will use $p = 50$. In software, the algorithms used for optimization are often based on simulated annealing or evolutionary strategies (see e.g. [36]). When the input variables are not provided in the same unit of measure, a maximin LHD is first designed over $[0,1]^2$, corresponding to dimensionless variables.

Now let us consider the cylinder $\mathcal{C}$ of polar coordinates. The construction of a Latin hypercube on $\mathcal{C}$ is identical for an hypercubic domain, by considering discretizations of $[0,1]$ and $\mathbb{S}$. For the sake of clarity, we propose to call *polar Latin cylinder design (polar LCD)* or simply LCD, a LHD defined in polar coordinates, referring to the geometry of the polar space. As for the maximin criterion, two modifications are needed for polar coordinates. First, a valid distance on $\mathcal{C}$ must fill the condition $\| \mathbf{u} - \mathbf{u}' \| = 0$ for $\mathbf{u} = (\rho, \theta)$ and $\mathbf{u}' = (\rho, \theta')$, with $\theta = \theta' \pmod{2\pi}$. In particular the Euclidean distance is no further valid since it does not see that the points $(\rho, 0)$ and $(\rho, 2\pi)$ are the same in $\mathcal{C}$. Second, the range of the polar angle θ is π, which is the maximum value of the geodesic distance over $\mathbb{S}$. Therefore, any distance over the dimensionless cylinder $[0,1] \times \left(\frac{1}{\pi}\mathbb{S}\right)$ applies to the polar space (ρ, θ). A natural choice is the geodesic distance given by:

$$\| \mathbf{x} - \mathbf{x}' \|_{\text{Polar}} = \sqrt{(\rho - \rho')^2 + \left(\frac{d_2(\theta, \theta')}{\pi} \right)^2} \tag{5.3}$$

Notice that the factor $\frac{1}{\pi}$ rescales d_2 to $[0,1]$ and weighs equivalently the radius and the angle.

From now on we will denote Φ_{Polar} (resp. $\Phi_{\text{Cartesian}}$) the Φ_p criteria computed with $\| \cdot \|_{\text{Polar}}$ (resp. $\| \cdot \|_2$). Minimizing Φ_{Polar} leads to a maximin LCD. A 20-point maximin LCD is displayed in Figure 5.4, where the cylinder is represented as a 2-dimensional map. As expected it is well filling the space of polar coordinates. Though it looks similar to a maximin LHD obtained in an hypercubic domain with the usual Euclidean distance, the difference is visible on the left and right boundaries which correspond to the same points in $\mathcal{C}$: the design points near the left and right boundaries are also spread out from each other. LCDs are recommended when the studied phenomenon has a physical interpretation with respect to polar coordinates. First, if the phenomenon is purely radial (resp. angular), the Latin structure ensures that all the design radius (resp. angles) values are different, so that no information is lost by projection. Furthermore, the maximin property helps in capturing non-linearities with respect to ρ and θ. However, when no a priori information about the phenomenon is known, the maximin LCD may be inappropriate, due to non-uniform filling that they produce on $\mathcal{D}$, as visible in Figure 5.4. Though it is not possible to optimize simultaneously maximin criteria based on distances in Cartesian and polar coordinates, a multi-criteria approach could be investigated. In this paper, as a first study, we focus on a simple transformation of a maximin LCD which helps improving space-fillingness on $\mathcal{D}$

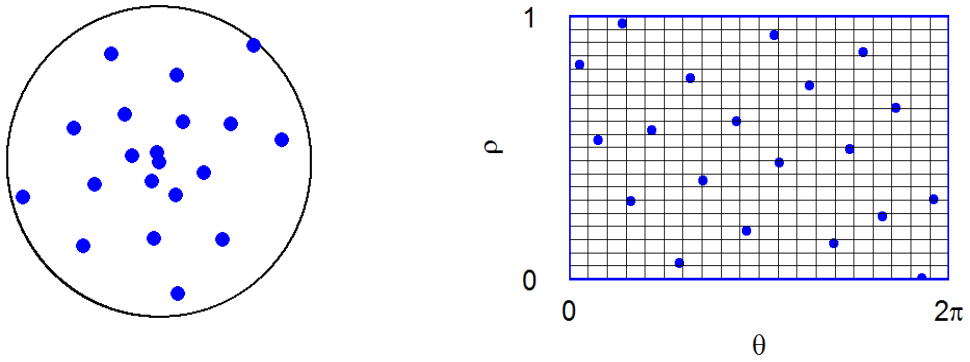


Figure 5.4 – Cartesian (left) and polar (right) representations of a 20-point maximin Latin cylinder design (LCD). The design is well-filling the cylinder $\mathcal{C}$ of polar coordinates, displayed as a 2-dimensional map: In particular, the design points near the left and right boundaries are also spread out from each other.

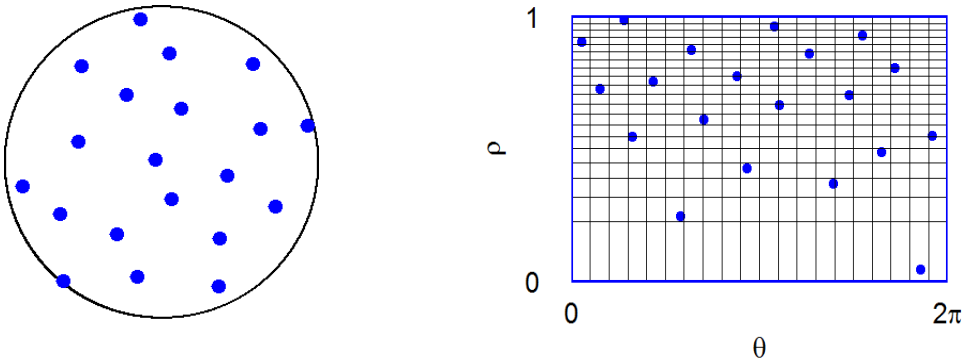


Figure 5.5 – Cartesian (left) and polar (right) representations of the LCD obtained by transforming the maximin LCD of Figure 5.4 with $\rho \mapsto \sqrt{\rho}$.

while preserving the Latin structure on $\mathcal{C}$. This is done by applying the transform $\rho \mapsto \sqrt{\rho}$, based on the well-known fact that if R, T are independent random variables with uniform distribution on $[0, 1]$ and $[0, 2\pi]$ respectively, then $(\sqrt{R}, T)$ is uniform on $\mathcal{D}$. This transformation was applied to the design of Figure 5.4, resulting in the design displayed in Figure 5.5.

5.3 Comparison

The aim of this section is to compare the DoEs presented above with respect to quality criteria, and to evaluate their performance on a set of toy functions. We will denote Dopt1, ..., Dopt4 the D-optimal DoEs for Zernike regression of order N ($1 \leq N \leq 4$) shown in Figure 5.1, and Spiral-F, Spiral-A, Spiral-G the spiral DoEs (Fermat, Archimede, Galilee) of Figure 5.2. We also denote maxLCD the maximin LCD of Figure 5.4 and maxLCD* its transformed version with $\rho \mapsto \sqrt{\rho}$ (Figure 5.5). All these 20-point DoEs are compared according to the following scheme:

(i). An assessment is made according to space-filling and D-optimality criteria. For space-filling, two indicators are used: the minimum Euclidean distance, and the minimum

geodesic distance (Equation 5.3) between design points. The D-optimality criterion for the N-order Zernike regression (see [26]) is given in log-scale.

(ii). A comparison in term of prediction accuracy. The RMSE over a test grid of 1.000 points is computed for the 6 analytical functions shown in Figure 5.6, illustrating various non-linear patterns. For each DoE, the best model is chosen among Zernike polynomials up to order 4, Cartesian GPs and polar GPs with kernels obtained by combination (sum, product, ANOVA) of 1-dimensional kernels as in Section §3.4.

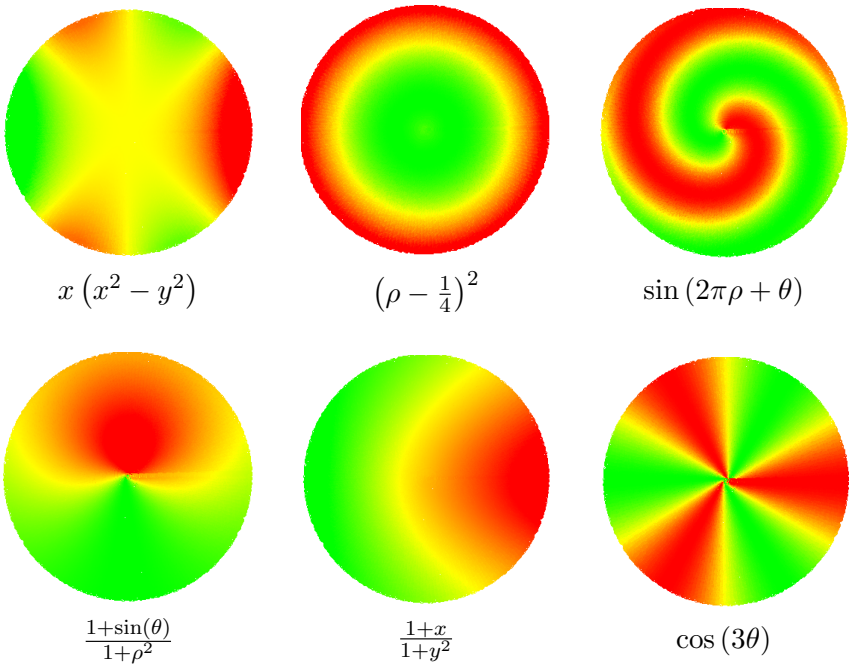


Figure 5.6 – Color representation of test functions.

	D-optimality			$\min_{i\neq j}\left(\| \mathbf{x}^{(i)}-\mathbf{x}^{(j)} \|\right)$	
	$D_{N=2}$	$D_{N=3}$	$D_{N=4}$	$\| \cdot \|_{\text{Polar}}$	$\| \cdot \|_2$
D-opt1	-159.9	-308.3	-448.3	0.01	0.31
D-opt2	**36.6**	-135.6	-353.1	0.01	0.33
D-opt3	35.4	**49.1**	-18.9	0.02	**0.45**
D-opt4	34.4	47.5	**63.5**	0.03	0.32
Spiral-F	29.7	37.2	44.1	0.04	0.20
Spiral-A	27.2	31.6	32.1	0.03	0.22
Spiral-G	23.3	19.3	-1.4	0.01	0.13
maxLCD	22.2	20.3	2.4	**0.06**	0.06
maxLCD*	27.5	32.8	33.0	0.04	0.28

Table 5.1 – Comparison of DoEs according to D-optimality and space-filling criteria.

The results of Table 5.1 are consistent with the theory of D-optimality and exhibit the lack of robustness of D-optimal designs in case of departure from their assumptions, especially when N is underestimated. The comparison also shows that spiral DoEs have rather good scores for all criteria. The best spirals for Zernike polynomials are the one that have the smaller p (Spiral-F), but the intermediate one (Spiral-A) has the best space-filling scores; It

seems to be the best trade-off among spirals. As expected, the maximin LCD is interesting for the polar GPs because it optimally fills the polar space, but has the worst space-filling score in Cartesian coordinates. This weakness is overcome by its modified version maxLCD*, which seems to accomplish the best trade-off for the different criteria among all the DoEs considered.

In Table 5.2, we see that D-optimal designs of low order $(1, 2)$ have in general poor scores

	Prediction RMSE (as percentage of the standard deviation)					
	$x\left(x^2-y^2\right)$	$\left(\rho-\frac{1}{4}\right)^2$	$\sin\left(2\pi\rho+\theta\right)$	$\frac{1+\sin(\theta)}{1+\rho^2}$	$\frac{1+x}{1+y^2}$	$\cos\left(3\theta\right)$
D-opt1	14.0	153.1	84.7	2.6	2.9	2.8
D-opt2	14.4	46.6	84.0	3.0	1.7	0.2
D-opt3	0.0	9.1	62.1	1.1	**0.4**	**0.2**
D-opt4	0.0	9.1	50.0	1.9	0.8	1.4
Spiral-F	0.0	0.4	35.2	4.0	0.7	4.5
Spiral-A	0.0	0.1	46.0	2.0	0.7	2.8
Spiral-G	0.0	**0.0**	49.1	2.9	1.3	0.8
maxLCD	0.0	0.0	31	1.0	1.0	0.8
maxLCD*	0.0	0.3	**23.3**	**1.0**	0.4	2.6

Table 5.2 – Comparison of DoEs in terms of predictive performance on toy functions.

in term of RMSE for the functions considered here, that present non-linearities. Spirals and maxLCD perform rather well. maxLCD met our expectations when radial and angular patterns are dominant (functions 1, 2, 3 and 6), and the modified maxLCD* seems to adapt well to the range of functions and models considered here, confirming its robustness among other DoEs. Finally, notice that the function $z = \sin(2\pi\rho + \theta)$ was poorly reconstructed by all models, whatever the DoE. An acceptable fit would require a model with geometric anisotropy in the space of polar coordinates, or more than 20 points.

Résumé en Français

Parmi les plans d'expériences spécifiques au disque, les plans D-optimaux pour les polynômes de Zernike sont adaptés aux modèles de régression. La planification d'expériences le long de spirales est aussi une option. En effet, en plus d'être des formes fréquentes en microbiologie, les spirales peuvent être facilement paramétrées de façon à remplir les domaines circulaires.

Etant donné le manque de robustesse des plans D-optimaux et les redondances produites par les spirales, nous avons introduit les cylindres latins. Ils permettent d'éviter les redondances dans l'espace polaire qu'ils remplissent par ailleurs uniformément. Nous montrons également qu'une simple transformation analytique permet de transformer les cylindres latins en plans remplissant le disque tout en limitant les redondances dans l'espace des coordonnées polaires. Les performances des cylindres latins sont ensuite évaluées via différents critères: D-optimalité, remplissage de l'espace et performance en prévision.

Chapter 6

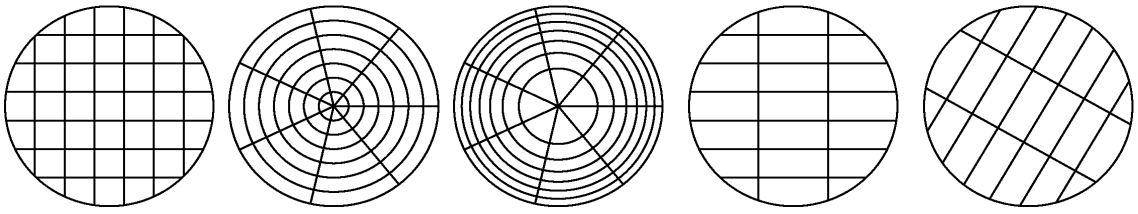
IMSE-optimal designs

We have seen in Chapter 5 that maximin Latin cylinders represent good tradeoffs when no information is available on the response surface. However in a dynamic setting, knowledge is gradually acquired through successive experiments. The purpose of this chapter is to include the available knowledge on the process, represented by its kernel, in the choice of design points. For this, the Integrate Mean Square (IMSE) criterion is used and the parameters of the kernel are supposed known.

6.1 Problematic and formulation

6.1.1 Motivation

The traditional approach to represent a spatial risk is to sample the input domain by square areas ([92], Chapter 6). In particular, Latin hypercubes belong to this class of designs. However, there exist many other discretization methods such as polygons, tessellations, triangulations, etc. A relevant choice should primarily be suited to the data [92]. As an example,



(a) (b) (c) (d) (e)

Figure 6.1 – Different discretization schemes for the disk

behind the discretization of Figure 6.1a, there is an implicit assumption of isotropy. Conversely, those of Figures 6.1b and 6.1c correspond to Latin Cylinders (section 5.2). They should better capture radial and angular variations. Recall that these DoEs were constructed without any knowledge on the response surface. In particular, the maximin criterion used for their optimization gives the same importance to radial and angular correlations. As a consequence, when the radial or the angular part is predominant, quantifying such information remains an open question. In the case of Cartesian GPs too, the discretization should be adapted for anisotropic processes, as illustrated in Figure 6.1d. However, the corresponding design is not a Latin hypercube, due to the non linear bound of the disk. In addition, using

the Euclidean distance raises actually the issue of geometric anisotropy. Since the choice of the x and y axes will influence the distribution of design points (Figure 6.1e), it deserves a careful assessment.

In this chapter, we aim at taking into account the available knowledge on the process in the DoE. We are also interested in a flexible formulation which, unlike grids, does not depend on the coordinates system. Our third aim is to find a procedure which is sequentially applicable, contrarily to Latin hypercubes. In Kriging, a key role of the kernel is to quantify spatial correlations such as those mentioned in the paragraph above. This knowledge can be exploited via the Integrated Mean Square Error (IMSE) criterion proposed in [100]. The IMSE criterion was successfully used in the framework of circular domains by [11] to reduce the size of a Design of Experiments (DoE)

6.1.2 Formulation

Given a design of experiments $\mathbf{X} = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right)$, and a Kriging model with stochastic part $Z \sim GP(0, k)$, the IMSE criterion related to $\mathbf{X}$ is [100]:

$$\text{IMSE}\left(\mathbf{X}\right) = \int_{\mathcal{D}} \mathbb{E}\left((Z_{\mathbf{x}} - \hat{Z}_{\mathbf{x}})^2\right) d\nu(\mathbf{x}) \tag{6.1}$$

Where $\hat{Z}_{\mathbf{x}}$ denotes the Kriging mean at $\mathbf{x}$. ν is an integration measure over $\mathcal{D}$. As the Kriging mean $\hat{Z}_{\mathbf{x}}$ is the orthogonal projection of $Z_{\mathbf{x}}$ onto the space spanned by $(Z_{\mathbf{x^{(1)}}}, \ldots, Z_{\mathbf{x^{(n)}}})$, we have by Pythagoras $\mathbb{E}\left({Z_{\mathbf{x}}}^2\right) = \mathbb{E}\left({\hat{Z}_{\mathbf{x}}}^2\right) + \mathbb{E}\left((Z_{\mathbf{x}} - \hat{Z}_{\mathbf{x}})^2\right)$. By noting $\mathbf{K}$ the covariance matrix and $\mathbf{k}(\mathbf{x})$ the covariance vector as introduced in chapter 2.2, we get $\mathbb{E}\left((Z_{\mathbf{x}} - \hat{Z}_{\mathbf{x}})^2\right) = \text{var}(Z_{\mathbf{x}}) - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{k}(\mathbf{x})$. Therefore,

$$\text{IMSE}\left(\mathbf{X}\right) = \int_{\mathcal{D}} \Big(\text{k}\left(\mathbf{x}, \mathbf{x}\right) - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{k}(\mathbf{x})\Big) d\nu(\mathbf{x}) \tag{6.2}$$

Remark The IMSE criterion depends only on the kernel and design points. Given a kernel, the IMSE-optimal design is not unique in general. In the isotropic case for instance, the IMSE criterion is invariant under symmetries and rotations. Notice that each numerical computation of the IMSE is time consuming and fast approaches are proposed in the literature [38].

6.1.3 The choice of an integration measure

The usual choice for ν (say ν^*) is the uniform measure over the disk:

$$\text{IMSE}^*\left(\mathbf{X}\right) = \int_0^1 \int_0^{2\pi} \Big(\text{k}\left(\mathbf{x}, \mathbf{x}\right) - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{k}(\mathbf{x})\Big) \rho d\rho \frac{d\theta}{2\pi}$$

As remarked for Latin cylinders, the resulting designs will not be uniform on the cylinder of polar coordinates. Indeed, the large values of ρ are actually overrepresented. To fill uniformly the space of polar coordinates, the uniform measure over the cylinder (say $\nu^\dagger$) should be preferred:

$$\text{IMSE}^\dagger\left(\mathbf{X}\right) = \int_0^1 \int_0^{2\pi} \Big(\text{k}\left(\mathbf{x}, \mathbf{x}\right) - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{k}(\mathbf{x})\Big) d\rho \frac{d\theta}{2\pi}$$

Either of these two measures can be used according to the interpretation of the response. When the the disk is physically a surface, IMSE^* would be appropriate. When the interpretation of the process is rather related to ρ and θ (wind speed and direction for instance), $\text{IMSE}^\dagger$ should be preferred.

6.2 Implementation and assessment

In this section, we compute IMSE-optimal designs with different settings in order to represent a wide variety of industrial scenarios. The considered kernels correspond to linear models with interactions presented in Chapter 4. For each simulation, the number of design points is 20, and the genetic algorithm using derivatives of [75] is used for optimization. This algorithm has the advantage to combine a global strategy, a randomized evolutionary search, with a local quasi-Newton optimization (L-BFGS-B). The integrals are numerically approximated over two uniform grids represented in Figure 6.4. The first one $G^\dagger$ is a uniform sample of $N = 1000$ points over $[0,1] \times [0, 2\pi]$. The second one G^* is obtained by transforming $G^\dagger$ with $\rho \mapsto \sqrt{\rho}$. To avoid the potential confusions between the integration measures ν^* and $\nu^\dagger$, corresponding to G^* and $G^\dagger$, we note: $\mathbf{X}^* = \text{argmin}\Big(\text{IMSE}^*\left(\mathbf{X}\right)\Big)$ and $\mathbf{X}^\dagger = \text{argmin}\Big(\text{IMSE}^\dagger\left(\mathbf{X}\right)\Big)$. In the following result tables, the designs are represented in the Euclidean space (y vs x) and in the polar space (ρ vs θ).

6.2.1 IMSE-optimal designs for polar GPs

In this section, the kernel is given by: $\mathrm{k}\left(\mathbf{x},\mathbf{x}'\right)=\mathrm{k}_1\left(\rho,\rho'\right)+\mathrm{k}_2\left(\theta,\theta'\right)+\mathrm{k}_1\left(\rho,\rho'\right)\mathrm{k}_2\left(\theta,\theta'\right)$. k_1 is the matérn$_{\frac{5}{2}}$ with range parameter ℓ_r and k_2 the C^2-Wendland with shape parameter τ and support $c=\pi$. The resulting designs are presented in Tables 6.1 and 6.2.

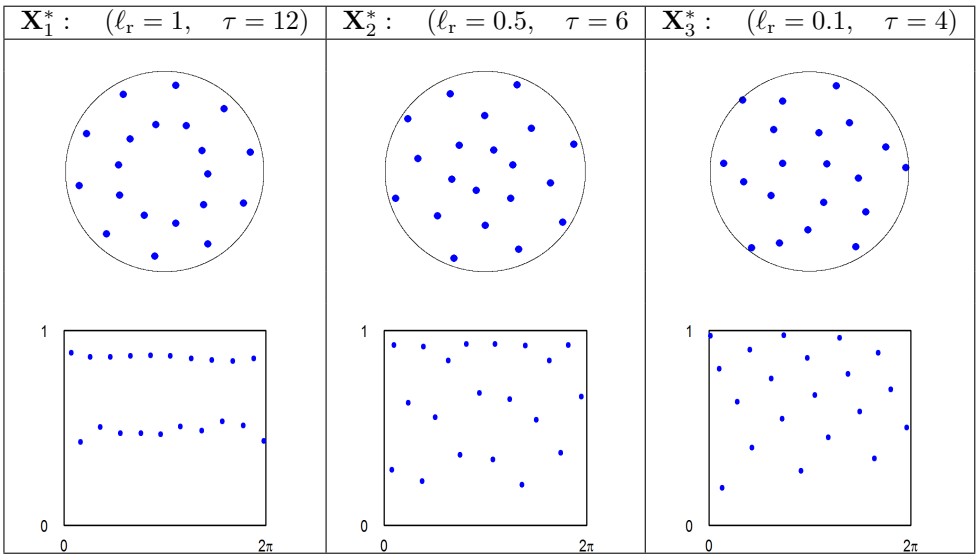


Table 6.1 – ν^*-IMSE-optimal designs for polar GPs with varying (ℓ_r,τ)

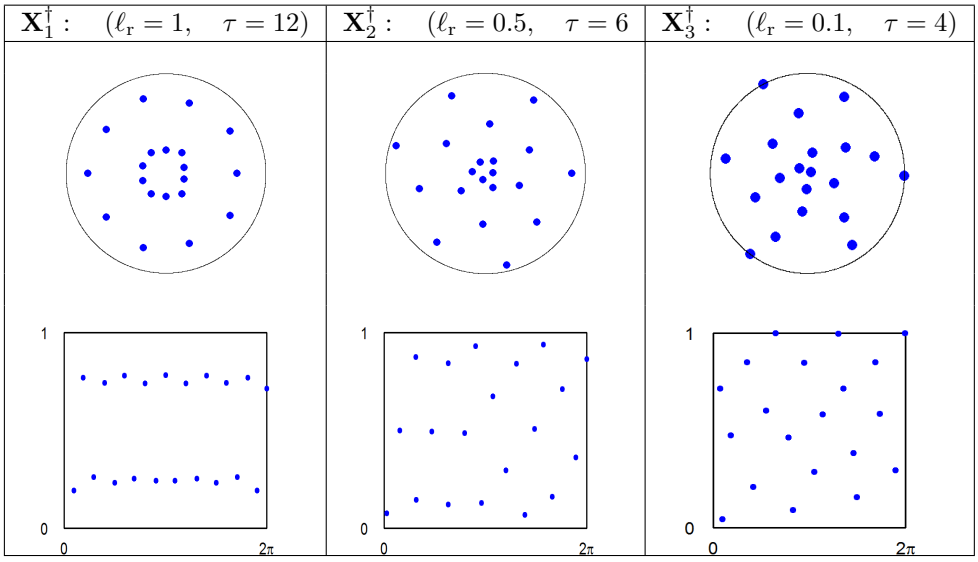


Table 6.2 – $\nu^\dagger$-IMSE-optimal designs for polar GPs with varying (ℓ_r,τ)

6.2.2 IMSE-optimal designs for Cartesian GPs

The kernel is defines as $\mathrm{k}\left(\mathbf{x},\mathbf{x}'\right)=\mathrm{k}_1\left(x,x'\right)+\mathrm{k}_2\left(y,y'\right)+\mathrm{k}_1\left(x,x'\right)\mathrm{k}_2\left(y,y'\right)$, with k_1 the matérn$_{\frac{5}{2}}$ with range parameter ℓ_1 and k_2 the k_2 the matérn$_{\frac{5}{2}}$ with range parameter ℓ_2. The results are presented in Tables 6.3 and 6.4.

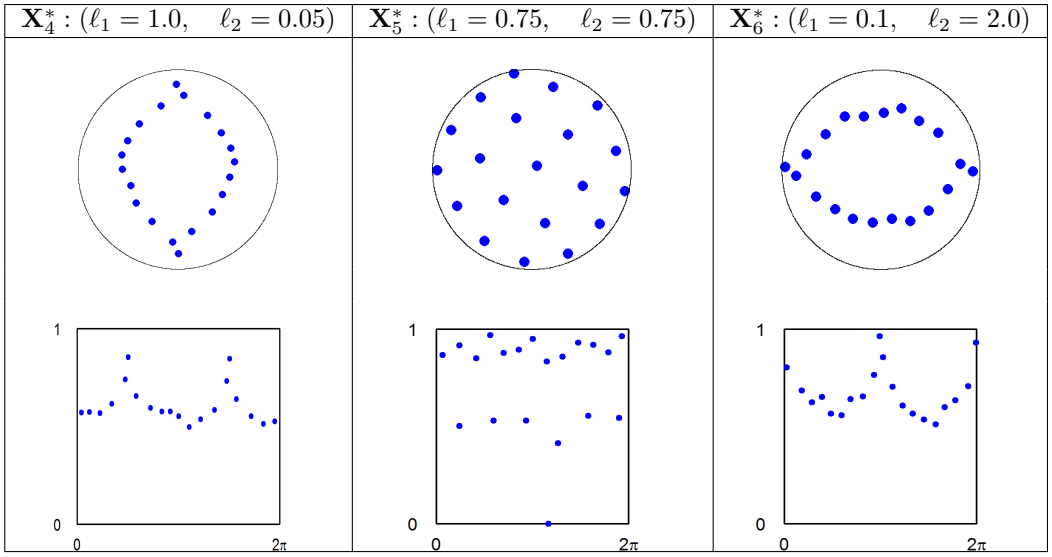


Table 6.3 – ν^*-IMSE-optimal designs for Cartesian GPs with varying (ℓ_1, ℓ_2)

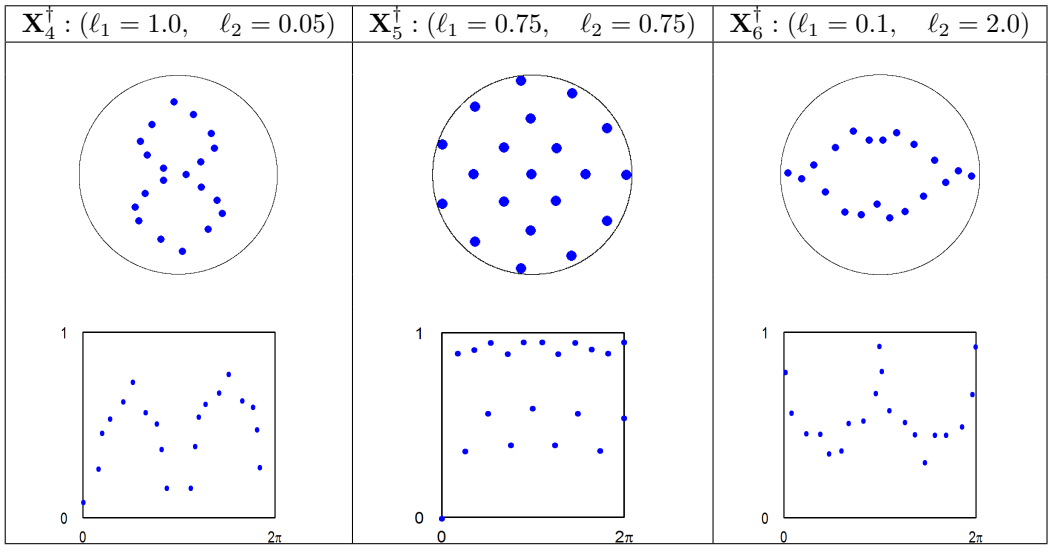


Table 6.4 – $\nu^\dagger$-IMSE-optimal designs for Cartesian GPs with varying (ℓ_1, ℓ_2)

6.3 Assessment through simulations

In the computed IMSE-optimal designs, various effects are visible, the first of which are symmetries: a central symmetry for polar GPs and axial reflections for Cartesian GPs. The integration measure too is influential. ν^* tends to fill the Euclidean disk uniformly, whereas $\nu^\dagger$ generates more points near the center to fill the polar space. Speaking of space-filling, the impact of correlations is drastic. A manifold with weak correlations is uniformly filled. Conversely, a manifold with strong correlations may present a poor coverage. Consequently, there are pronounced alignments in design points, corresponding to specific phenomena.

Remarks

1. In this study, symmetry properties were automatically recovered by the minimization of the IMSE. This is quite time consuming, due to combinatorial aspects and integral approximations. Setting symmetries a priori would allow to divide the dimension of the problem by 2 or 4.

2. Geometric anisotropy can also be included in IMSE-optimal designs for Cartesian GPs. Though not detailed here, it simply translates into a rotation of the DoE.

Validation through simulations

For the sake of conciseness, we only consider the 6 IMSE-optimal designs corresponding to the measure ν^* in this paragraph. They correspond to $\mathbf{X_1}^*, \ldots, \mathbf{X_6}^*$ in Tables 6.1 and 6.3. We call $\mathrm{k}_1, \ldots, \mathrm{k}_6$ the kernels under which these 6 DoEs are optimal. Recall that $\mathrm{k}_1, \ldots, \mathrm{k}_6$ are represented by their parameters (ℓ_r, τ), or (ℓ_1, ℓ_2) in Tables 6.1 and 6.3. We also consider the maximin Latin Cylinders maxLCD and maxLCD* to provide a reference level. The following simulation protocol is used:

1. Generate 260 points over the disk, including a regular grid of 100 points, and the $(6 + 2) \times 20$ points corresponding to the 6 IMSE optimal designs and the 2 Latin Cylinders.

2. (a) Simulate one realization of $Z \sim GP(0, \mathrm{k}_1)$ at the 260 points.
   (b) Fit 8 different Kriging models corresponding to the 8 DoEs, with this (known) kernel k_1.
   (c) Compute the Mean Square Error (MSE), based on the regular grid for each Kriging model.
   (d) Repeate 100 times the steps a, b and c, and average the MSEs by design.
   (e) Rank the 11 designs by increasing MSEs.

3. Repeat step 2 with $Z \sim GP(0, \mathrm{k}_2), \ldots, Z \sim GP(0, \mathrm{k}_6)$

4. Compute the median and worst ranks by design.

The results are summarized in Table 6.5. They are fully consistent with the theory of IMSE-optimality. For each kernel k_i, the best design is repaired by the rank 1 in red. Each IMSE-optimal design achieves the best result for the kernel under which it is computed. However, when the "wrong" kernel is specified, the IMSE-optimal designs become less competitive.

	k_1	k_2	k_3	k_4	k_5	k_6	Median	Worst
$\mathbf{X_1}^*$	**1**	3	5	6	3	6	4	6
$\mathbf{X_2}^*$	2	**1**	3	2	2	2	2	3
$\mathbf{X_3}^*$	8	4	**1**	5	5	5	5	8
$\mathbf{X_4}^*$	6	8	8	**1**	8	8	8	8
$\mathbf{X_5}^*$	4	5	6	4	**1**	3	4	6
$\mathbf{X_6}^*$	5	7	7	8	7	**1**	7	8
LCD*	3	2	2	3	4	4	3	4
LCD	7	6	4	7	6	7	6	7

Table 6.5 – Ranks of the 8 DoEs when estimating 6 different GPs with kernels $k_1, \ldots, k_6$, where each $\mathbf{X_i}^*$ is IMSE-optimal under k_i.

Among them, $\mathbf{X_2}^*$ is the most robust. The point is that $\mathbf{X_2}^*$ is a space-filling design without obvious alignment, so adapted to capture a wide variety of non-linearities. The uniform maximin Latin cylinder (LCD*) is confirmed to accomplish the best trade-off with regard to the studied scenarios.

6.4 The discrete case

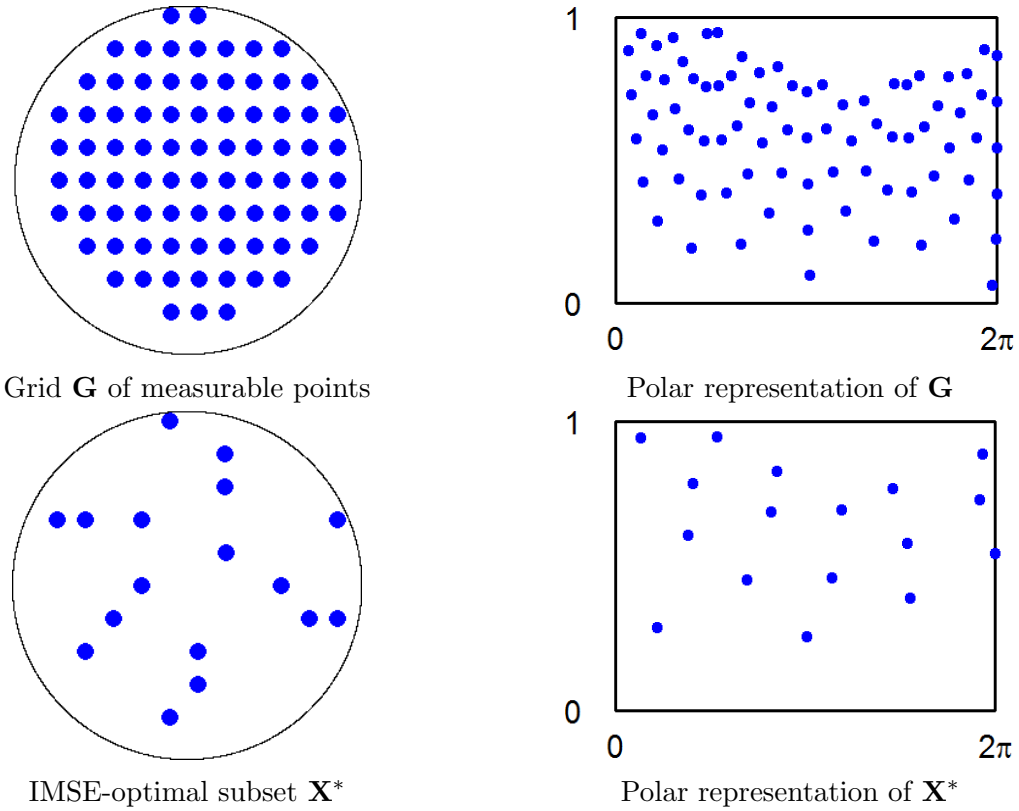
In many industrial settings, only a finite set of points can be measured. The DoE can only be a subset of these measurable points and IMSE-optimality becomes a discrete optimization problem. Denote $\mathbf{G} = \{\mathbf{g^{(1)}}, \ldots, \mathbf{g^{(N)}}\}$ this set of all measurable points of $\mathcal{D}$, and n the desired size of the DoE. The problem is to find $\mathbf{X} = \{\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\} \subseteq \mathbf{G}$ such that IMSE $(\mathbf{X})$ is minimal.

Study of an industrial case We consider the dataset of Paragraph 3.4.1, including 81 measurable points on a wafer, 17 of which served as DoE. Given the different Kriging models that were previously tested for this dataset, we retain the kernel estimated with Equation 4.12 in Section 4.4.2. This information is now used to select a new subset of 17 points, based on IMSE-optimality. From a combinatorial point of view, reaching the exact solution is difficult. Nevertheless, global strategies such as simulated annealing were showed suit this class of problem [11]. In the same perspective, we use the genetic research presented in [75]. The grid $\mathbf{G}$ and the estimated IMSE-optimal subset $\mathbf{X}^*$ are displayed in Figure 6.2. The estimated optimal subset is now compared to the usual industrial design presented in Figure 3.5. Though their Kriging means are similar, the IMSE-optimal subset lead to a smaller uncertainty: Figure 6.3. Notice that the points of $\mathbf{G}$ are shifted to the top right of the disk. This is due to technical constraints resulting from layout design rules [1].

Résumé en Français

Le critère IMSE est un indicateur d'incertitude globale sur la surface du disque. Puisqu'il se calcule directement lorsque les paramètres de krigeage sont connus, le minimiser aboutit aux

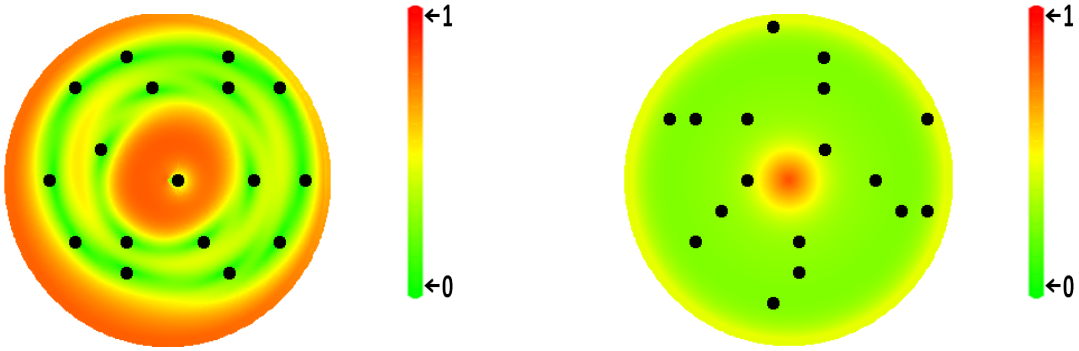
[1] Integrated circuit layout, also known as mask design, is the representation of an integrated circuit in terms of geometric shapes, corresponding to elementary components. The performance of integrated circuits depends on the positions and connections among these geometric shapes. Hundreds of rules govern the layout of modern circuits.



Grid **G** of measurable points | Polar representation of **G**

IMSE-optimal subset $\mathbf{X}^*$ | Polar representation of $\mathbf{X}^*$

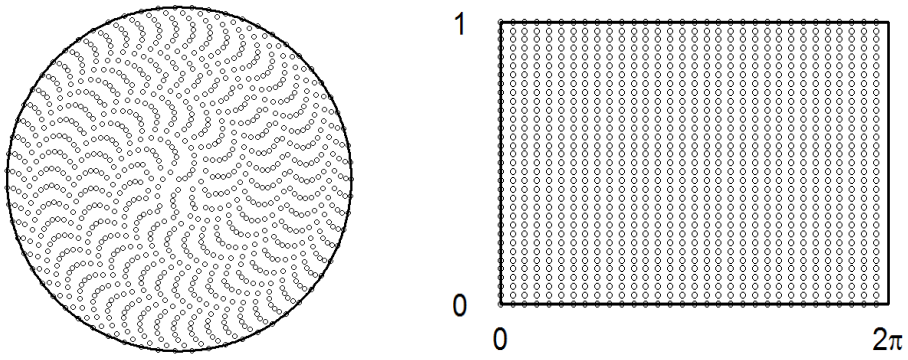
Figure 6.2 – Grid of the available points, and the selected subset of 17 points.



Mean integrated variance: 0.25 | Mean integrated variance: 0.10

Figure 6.3 – Kriging standard deviation, using the 17 designs points of Figure 3.5 (left) and the IMSE-optimal DoE presented in Figure 6.2 (right)

plans d'expériences optimaux en termes de risque quadratique. Dans ce chapitre, nous avons étudié les plans IMSE-optimaux pour les modèles de krigeage sur le disque avec différents scénarios, simulés par variation des paramètres des processus gaussiens. Les motifs obtenus, notamment leurs propriétés de symétrie et de rotation dépendent du type de processus gaussien. Les sous - espaces marginaux (x et y ou ρ et θ) ne sont par ailleurs pas remplis de la même façon selon l'importance relative des corrélations horizontales et verticales, ou des corrélations radiales ou angulaires, déterminées par les paramètres de krigeage. La pertinence des plans IMSE-optimaux est enfin montrée via simulations.



G^* (y vs x) $G^\dagger$ (ρ vs θ)

Figure 6.4 – Integration grids associated to ν^* and $\nu^\dagger$

Chapter 7

IMSE-optimal relocations

IMSE-optimal designs are recommended for Kriging models when the kernel is known. However, their implementation in industry should not cause abrupt changes in operating systems. In this chapter, a sequential procedure is investigated to gradually improve a design in terms of IMSE-optimality. The parameters of the kernel are supposed known.

7.1 Motivations and working hypothesis

7.1.1 Motivations

A key point in Statistical Process Control is repeatability, as we will see in Chapter III. A sudden change of design points may wrongly lead to suspect a drift in the process mean, as illustrated in Figure 7.1. To avoid such instabilities, a progressive strategy is investigated to optimize the DoE. This translates into two industrial constraints: do not increase the cost of an experiment, and change the original design as little as possible. From there comes the issue of optimal relocation of a design point, which is addressed via the IMSE.

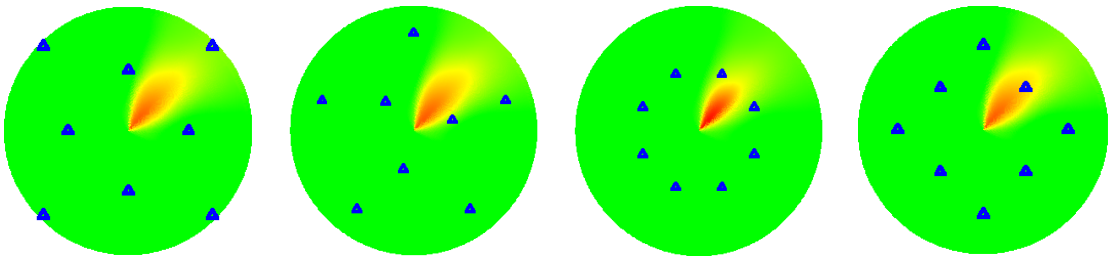


Figure 7.1 – Instability when monitoring air quality (see Section 3.4.2 and [7] for more details). $\bar{\mathbf{y}}$ is an estimation of the response mean, based on observations at blue points.

7.1.2 Assumptions

Given the DoE $\mathbf{X} = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right)$ over $\mathcal{D}$, the question of relocation is formulated as removing a design point $\mathbf{x^{(j)}}$ and adding a new one $\mathbf{x}^\star \in \mathcal{D}$. There is a duality between these two operations (adding and removing). Given two choices of point to remove, the best point to add may be different, and vice-versa. Consequently, for N potential points to add,

there are $N \times n$ possible relocations. Though n is usually of the order of a few dozens, N may reach very high levels and increase the solution time. This issue remains present in the continuous case, due to the increasing number of local optima. In our industrial context, the relocation procedure will be repeated in continuous, for thousand times. To reduce the computation time, we do not test the $N \times n$ possible relocations. Instead, the point to add is first proposed, regarding the initial DoE. Only then, is the removal suggested. Finally, the relocation is confirmed if it leads to a lower IMSE. Doing the addition first is motivated by a purpose of sequential computation. Its also allows to explore the areas of the disk which are not filled by the initial DoE.

7.2 Sequential relocation of a design point

7.2.1 Sequential addition

Given an initial design $\mathbf{X} = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right)$, and an integration measure ν over $\mathcal{D}$, the question is to find $\mathbf{x}^{(n+1)}$ which minimizes the criterion:

$$I_{n+1} = \int_{\mathcal{D}} \sigma^2_{n+1}(\mathbf{u}) d\nu = \int_{\mathcal{D}} \Big(\mathrm{k}\left(\mathbf{u}, \mathbf{u}\right) - \mathbf{k_{n+1}}(\mathbf{u})^\top (\mathbf{K_{n+1}})^{-1} \mathbf{k_{n+1}}(\mathbf{u}) \Big) d\nu$$

where $\mathbf{K_{n+1}} = \left(\mathrm{k}(\mathbf{x^{(i)}}, \mathbf{x^{(j)}})\right)_{1 \leq i,j \leq n+1}$ and $\mathbf{k_{n+1}}(\mathbf{u}) = \left(\mathrm{k}(\mathbf{u}, \mathbf{x^{(i)}})\right)_{1 \leq i \leq n+1}$ are the covariance matrix and the covariance vector at $\mathbf{u}$ for the Kriging model based on the design $\left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n+1)}}\right)$. I_{n+1} is the IMSE of the Simple Kriging model after adding $\mathbf{x}^{(n+1)}$ to the DoE. When the integral is approximated through a grid with $j = l^2$ points, N evaluations of I_{n+1} will require $N \times l^2$ different evaluations of the quantity $\mathbf{k_{n+1}}(\mathbf{u})^\top (\mathbf{K_{n+1}})^{-1} \mathbf{k_{n+1}}(\mathbf{u})$, which is computationally demanding. Therefore, we use a sequential formula to compute I_{n+1} when I_n is already known. The Kriging variance $\sigma^2_{n+1}(\mathbf{u})$ after adding $\mathbf{x}^{(n+1)}$, is obtained from the Kriging variance $\sigma^2_n(\mathbf{u})$ by:

$$\sigma^2_{n+1}(\mathbf{u}) = \sigma^2_n(\mathbf{u}) - \frac{\left(e - \mathbf{w}^\top \mathbf{d}\right)^2}{c - \mathbf{w}^\top \mathbf{b}} \tag{7.1}$$

with $e = \mathrm{k}(\mathbf{x^{(n+1)}}, \mathbf{u})$, $c = \mathrm{k}\left(\mathbf{x^{(n+1)}}, \mathbf{x^{(n+1)}}\right)$, $\mathbf{b} = \mathbf{k_n}\left(\mathbf{x^{(n+1)}}\right)$, $\mathbf{d} = \mathbf{k_n}(\mathbf{u})$, and $\mathbf{w} = \mathbf{K_n^{-1}} \mathbf{k_n}\left(\mathbf{x^{(n+1)}}\right)$. With this formula adapted from ([101], Section 5.2), solving a $n+1$ by $n+1$ equation system is transformed into $2n+2$ multiplications and 3 additions.

7.2.2 Sequential deletion

The traditional leave-one-out and its limits

Given the design $\mathbf{X} = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right)$, the purpose is to remove $\mathbf{x^{(j)}}, j \in \{1, \ldots, n\}$ while keeping the maximum amount of information. This can be addressed with a leave-one-out cross-validation. Leave-one-out consists in estimating the response at each design point when the corresponding observation is removed from the learning set. In the framework of Kriging models, there exist quick versions of leave-one-out, based on Dubrule's formula. Implementations are provided in two R packages, kergp [27] and DiceKriging [97]. Denote $\varepsilon_1, \ldots, \varepsilon_n$ the leave-one-out absolute errors corresponding to $\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}$, and ε_j their minimum. Then $\mathbf{x^{(j)}}$ is the point to remove. Indeed, $\mathbf{x^{(j)}}$ is the design point whose deletion

has the smallest consequence. However, leave-one-out remains questionable in the case of noisy observations. When the noise magnitude gets large, the procedure loses in reliability. Furthermore, leave-one-out cannot be applied when some response values are missing. This is the case when $\mathbf{x}^{(n+1)}$ is added to design points.

Leave-one-out based on Kriging variance

We consider a leave-one-out procedure, based on Kriging variance. Then, all the response values are no longer required, especially when the kernel is known. Given $Z \sim GP(0, k)$ and design points $\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n+1)}} \in \mathcal{D}$, we call leave-one-out Kriging variance the variance of Z at a design point, conditionally on observations at the others.

$$\sigma^2_{-i,i} = \mathbb{E}\left(Z^2_{\mathbf{x^{(i)}}} \middle| Z_{\mathbf{x^{(1)}}}, \ldots, Z_{\mathbf{x^{(i-1)}}}, Z_{\mathbf{x^{(i+1)}}}, \ldots, Z_{\mathbf{x^{(n+1)}}}\right), \quad i \in 1 \ldots n+1$$

If σ^2_j is the minimum leave-one-out variance, then $\mathbf{x^{(j)}}$ is the point to remove.

Leave-one-out based on IMSE-optimality

The leave-one-out procedure using Kriging variance is based on a local criterion since the expected variance is optimized at a single point. Following our initial purpose of a global reduction of uncertainty, we propose to use the IMSE criterion to select the best point to remove. Given $\mathbf{u} \in \mathcal{D}$, we define leave-one-out Kriging variances at $\mathbf{u}$ as:

$$\sigma^2_{-i}(\mathbf{u}) = \mathbb{E}\left(Z^2_{\mathbf{u}} \middle| Z_{\mathbf{x^{(1)}}}, \ldots, Z_{\mathbf{x^{(i-1)}}}, Z_{\mathbf{x^{(i+1)}}}, \ldots, Z_{\mathbf{x^{(n+1)}}}\right), \quad i \in 1 \ldots n+1$$

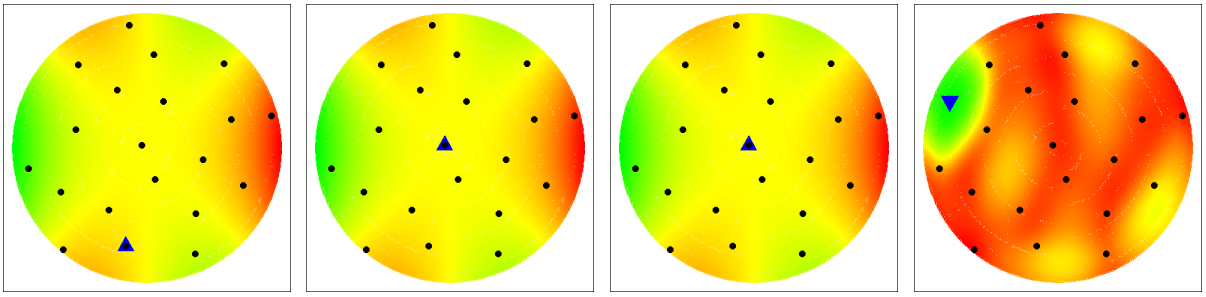
The leave-one-out IMSEs are given by: $I_{-i} = \int_{\mathcal{D}} \sigma^2_{-i}(\mathbf{u}) d\nu$. If I_{-j} is the minimum of these leave-one-out IMSEs, then the point to remove is $\mathbf{x^{(j)}}$.

7.3 Illustration and iteration of relocation

The relocation procedure is first illustrated with analytical functions to represent different scenarios. Secondly, it is iterated to simulate successive applications in a dynamic system.

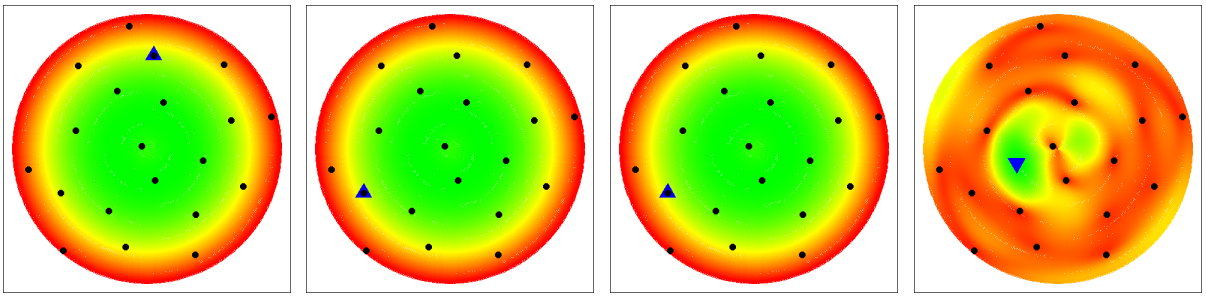
7.3.1 Illustration of relocation on toy functions

As initial DoE, we use the maximin Latin Cylinder LCD* (see section 5.2). For each toy function, two GP models are estimated. The first one is a linear model based on a Cartesian GP and the second one is a linear model based on a polar GP (Chapter 4). The model with the higher likelihood is used to propose the relocation. For each function, 4 graphs are displayed. On the first one, the point to remove is proposed by the traditional leave-one-out. On the second one, it is proposed by leave-one-out based on Kriging variance. On the third one, it is proposed by IMSE-optimality. The fourth graph displays the criterion I_{n+1}, where the point to add is marked with a triangle point-down. Further analytical functions are tested in Appendix 10.5. The suggested relocations, including those presented in Appendix 10.5, tend to fill the unexplored regions of the disk. In particular, the spatial indicator I_{n+1} exhibits different kinds of poor space-filling for the different patterns. The three deletion strategies may lead to different results. However, their are equivalent in Figure 7.2 (same abscissa) and in Figure 7.3 (same polar radius).



Leave-one-out 1 | Leave-one-out 2 | Leave-one-out 3 | $I_{n+1}(\mathbf{u})$

Figure 7.2 – 3 relocation strategies for the function $x^3 - xy^2$, based on a Cartesian GP. The triangles point-up are proposals for relocation, and the triangle point-down is the new location.



Leave-one-out 1 | Leave-one-out 2 | Leave-one-out 3 | $I_{n+1}(\mathbf{u})$

Figure 7.3 – 3 relocation strategies for the function $(\rho - \frac{1}{4})^2$, based on a polar GP. The triangles point-up are proposals for relocation, and the triangle point-down is the new location.

7.3.2 Iteration of the relocation procedure

In this section, we denote the initial design by $\mathbf{X}^0 = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right)$ and we call I_0 the corresponding IMSE. Let $\mathcal{R}$ be the procedure of relocation of one design point. Plainly, $\mathcal{R}\left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(n)}}\right) = \left(\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(j-1)}}, \mathbf{x}^\star, \mathbf{x^{(j+1)}}, \mathbf{x^{(n)}}\right)$, where $\mathbf{x}^\star$ and j are chosen by IMSE-optimality (Sections 7.2.1 and 7.2.2). We consider the sequence $\left(\mathbf{X}^k\right)_{k\in\mathbb{N}}$, with first element $\mathbf{X}^0$, and defined as $\mathbf{X}^{k+1} = \mathcal{R}\left(\mathbf{X}^k\right)$. Then, the sequence $I_k = \text{IMSE}\left(\mathbf{X}^k\right)$ is convergent. Indeed, this sequence is decreasing and bounded. As a consequence, iterating the procedure $\mathcal{R}$ will result in a solution to the problem of IMSE-optimality. Given the sequential nature of the algorithm, this solution is a local optimum.

An analytical example

In the example below, the maximin Latin cylinder LCD (see section 5.2) is used as initial design. We consider the Ridge function $f_c : (x, y) \to |x\cos\left(\frac{\pi}{3}\right) + y\sin\left(\frac{\pi}{3}\right)|^\beta$, with $\beta = 1 + \frac{\sqrt{2}}{2}$. Recall that LCD is designed for polar GPs. Therefore, it does not fill the disk uniformly. In particular, the sample density is very high at the center and a poor space filling is observed near the boundary. In contrast, because f_c corresponds to a Cartesian GP, a space-filling design should be chosen. The purpose is to explain the impact of successive relocations on this case of mismatch between the response and the design. The relocation procedure is iterated 10 times (Figure 7.5). After 7 iterations displayed in Figure 7.4, the sequence reaches a stationary state. Global strategies, involving simultaneous relocations of several

points, would ensure that a global optimum is obtained. But this is outside scope.

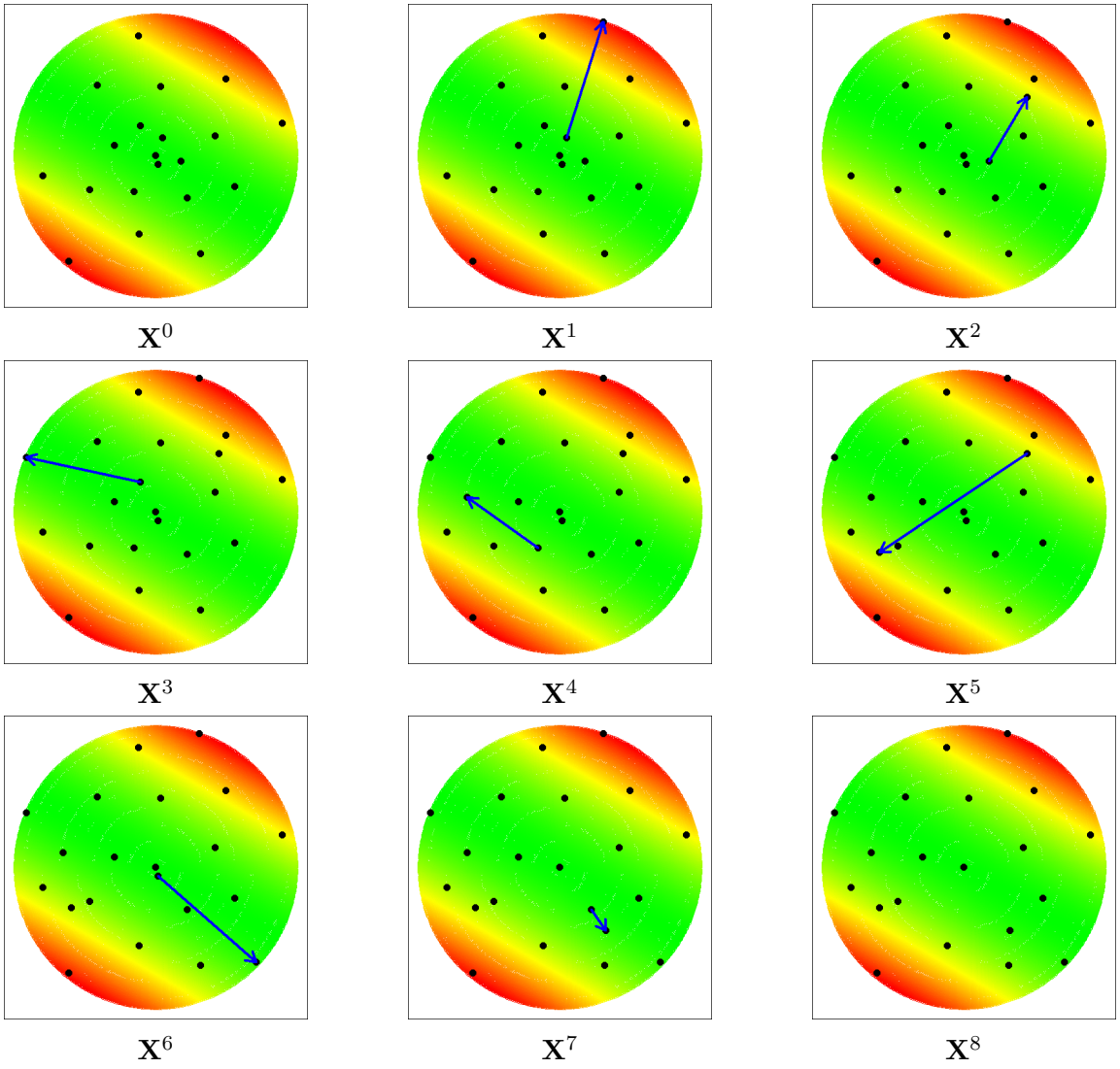


Figure 7.4 – Successive relocations with a GP estimated from a Ridge function.

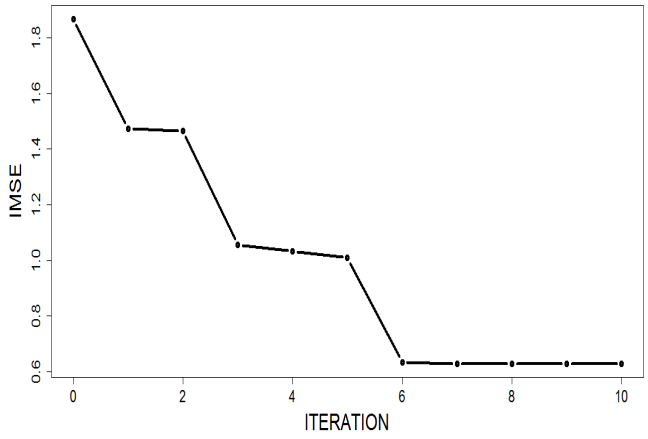


Figure 7.5 – Evolution of the IMSE for 10 iterations

Application: sequential design for air quality monitoring

The industrial DoE used is Section 3.4.2 is considered. The dataset is presented in Figure 3.7. It corresponds to the concentration of a greenhouse gas for different values of wind speed (polar radius) and direction (polar angles). The design, including 30 points filling the disk, was chosen without any knowledge on the process. It was shown in Section 3.4.2 that the response corresponds to a polar GP. The parameters of the kernel were also estimated. Now, they are used to optimize the design through successive relocations. In Figure 7.7 is displayed the evolution of the IMSE during 10 relocations displayed in Figure 7.6. The

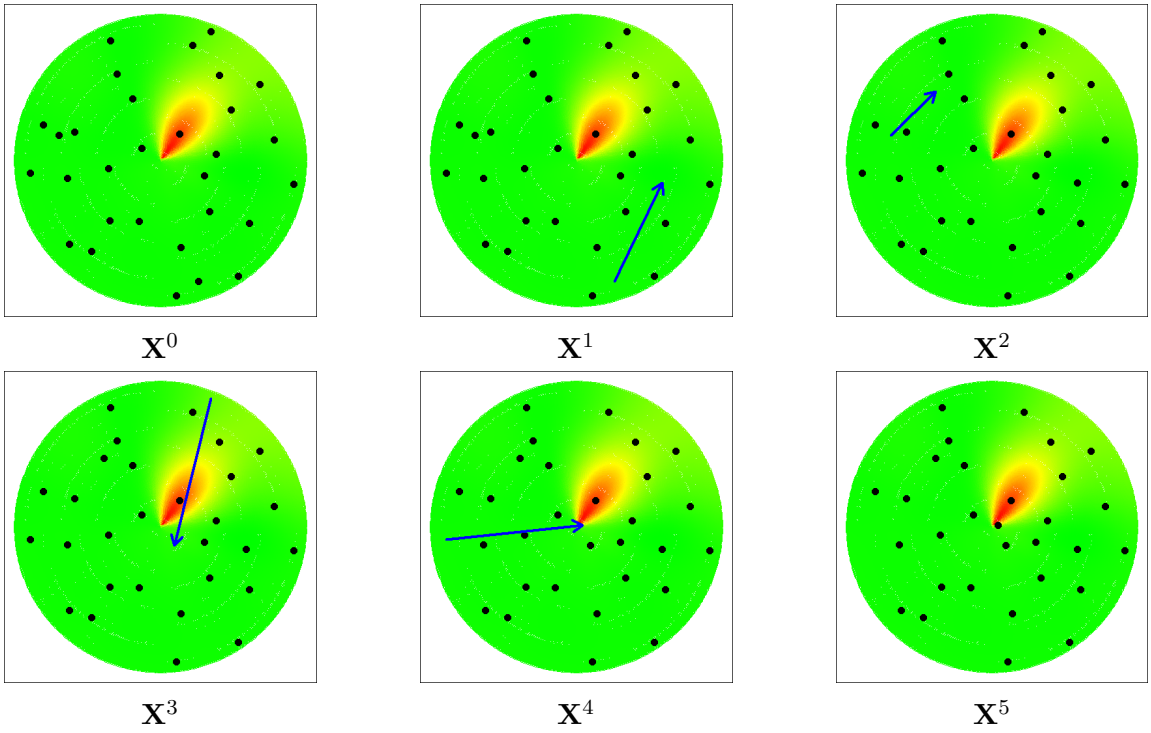


Figure 7.6 – Successive relocations with a polar GP.

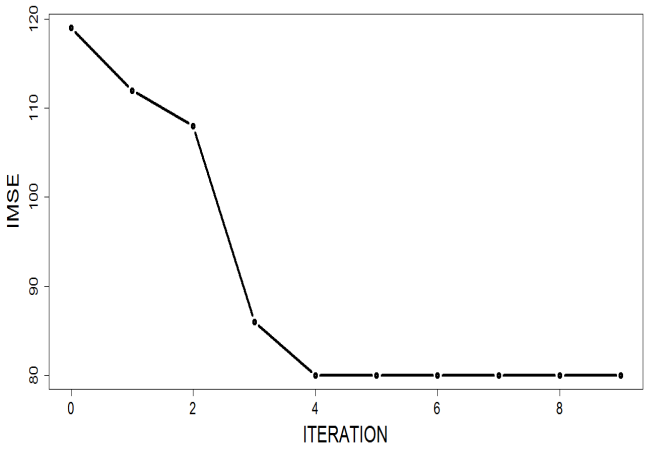


Figure 7.7 – Evolution of the IMSE for 10 iterations.

initial design $\mathbf{X}^0$ does not fill the space of polar coordinates. For a polar GP, this is an inadequacy which was partially offset by successive relocations.

Résumé en Français

La procédure présentée dans ce chapitre permet de déplacer un point d'un plan d'expériences de façon IMSE-optimale. Pour faire face à la forte complexité du problème d'optimisation sous-jacent, la question est découplée en deux parties: le retrait d'un point et l'ajout d'un autre sous contrainte d'IMSE-optimalité. Cette formulation permet de calculer le critère IMSE de façon séquentielle. La pertinence et la convergence de la procédure est enfin illustrée à travers des fonctions tests et des cas industriels.

PART III

Monitoring of spatial and temporal data

Chapter 8

Profile monitoring on the disk

In this Chapter, we focus on statistical quality control. After a brief review of conventional Statistical Process Control tools, we tackle the case where the quality indicator is the pattern defined over the disk, regarded as a profile. Control charts based on Zernike polynomials and Gaussian process models are investigated.

Some results of this Chapter are published in the proceedings [82], "Spatial risk assessment on circular domains: Application to wafer profile monitoring", by Padonou, Roustant, Blue and Duverneuil.

8.1 Statistical Process Control (SPC)

SPC refers to the use of statistical methods in order to monitor and improve a production process, or more generally, a business environment. The topic was introduced by Shewhart [103]. Ever since, it has received an increasing attention in industry. The key idea is that in any production process, there are two distinct origins of variation: common-causes and special-causes (see e.g. [76]). Common causes include permanent and quantifiable background noises. They are unavoidable and represent an inherent part of the process. As examples, machine vibrations, measurement errors and computers' response time are common-cause variations. Conversely, special causes are unusual and generally large when compared to the background noise. Computer crash, machinery failures, faulty controllers and human errors are examples of special-cause variations. They are usually arisen from dysfunctions. In SPC, control charts represent simple and powerful tools which allow to detect special-cause variations. Among them, Shewhart control charts are the most intuitive.

8.1.1 Shewhart control charts

Following the theory of variations aforementioned, quality can be statistically controlled by monitoring some key indicators over time. In the example of Figure 8.1, 100 successive observations of a quality characteristic x are displayed versus time. The center line corresponds to the average value of x. The other lines, called upper control limit UCL and lower control limit LCL, are chosen to wrap the majority of observations when the process is in-control. In this sense, the two last observations correspond to special causes and are called out-of-control signals. Such a graphical display is a Shewhart control chart, or simply control chart. It can be roughly interpreted as successive hypothesis testing. However, control limits are usually chosen to match quality requirements, and may not satisfy all the statistical assumptions

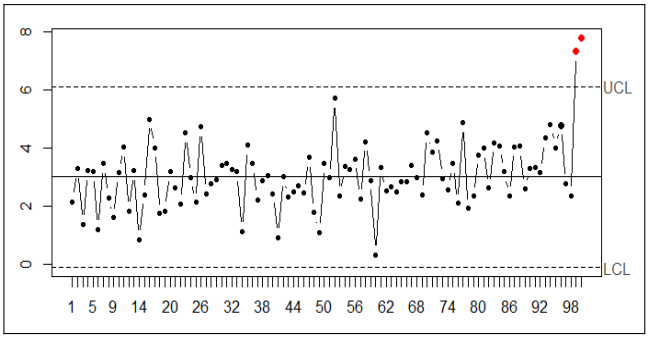


Figure 8.1 – An example of Shewhart control chart

[76]. Designing a control chart is commonly done in two steps. First, an historical dataset, collected under normal operating conditions, is used to estimate the control limits. This phase is qualified as retrospective. Second, a prospective phase consists in detecting out-of-controls, based on the estimated limits. When there is no historical record, a confidence interval is estimated from the recent dataset and used. In this case, robust methods are recommended to avoid bias.

8.1.2 CUSUM charts

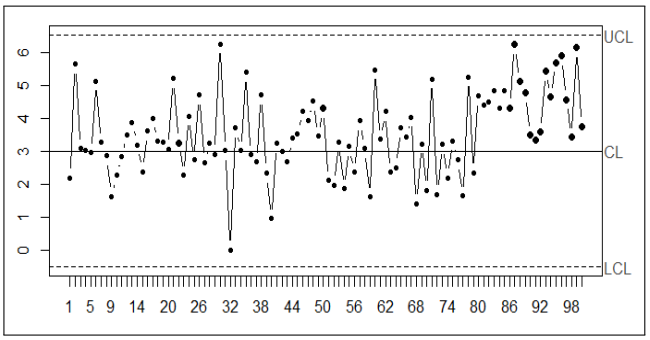


Figure 8.2 – Control chart with a small shift from the date 80.

In another example, $n = 100$ observations $x_1, \ldots, x_n$ at regular time periods $t_1, \ldots, t_n$ are plotted in the control chart of Figure 8.2. From date 80 to date 100, all observations are in-control, but above the center line. This is apparently caused by a non-random factor, indicating a small but persistent change in the process (Western Electric rules, Chapter 5 of [76]). To detect the small shifts, cumulative sum of observations (CUSUM) are more efficient. The first CUSUM charts were introduced in 1954 by E. S. Page [84]. Nowadays, the Tabular CUSUM chart is the mostly used. It consists in monitoring two statistics C_i^+ and C_i^-, respectively dedicated to detect increases or decreases in the x_is', and defined as:

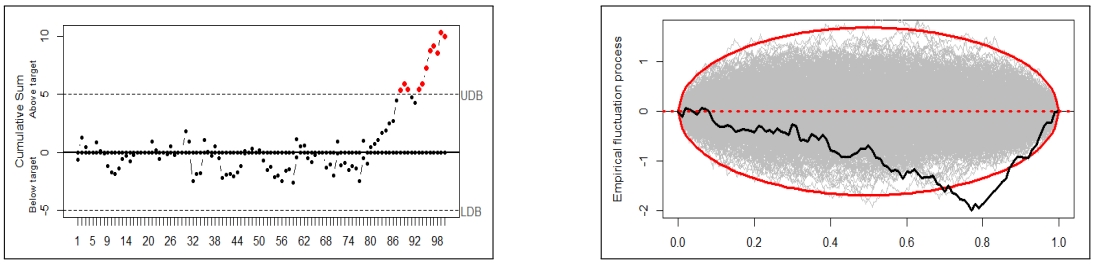
$$\begin{cases} C_i^+ = \max\left(0, x_i - k^+ + C_{i-1}^+\right) \text{ and} \\ C_i^- = \max\left(0, k^- - x_i + C_{i-1}^-\right), \end{cases} \tag{8.1}$$

where $C_0^+ = C_0^- = 0$. The parameters $k^+ = \bar{x} + \delta$ and $k^- = \bar{x} - \delta$ represent tolerable magnitudes for increases and decreases of the mean. In practice, δ is a small percentage of the estimated standard deviation $\hat{\sigma}$ and the control limits are set around $5\hat{\sigma}$. The Tabular

CUSUM applied to the dataset of Figure 8.2 is illustrated in Figure 8.3a. Although the procedure is designed to be efficient, it usually fails to detect early and late shifts. This drawback is due to the initialization and the time delay needed to accumulate deviations from the target. Therefore, [115] proposed alternative bounds for CUSUM charts, based on asymptotic properties. The observation periods are rescaled between 0 and 1 ($t_1 = 0, \ldots, t_n = 1$) and the CUSUM quantity at time $t \in [0, 1]$ is given by:

$$W_n^0(t) = \frac{1}{\hat{\sigma}\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} (x_i - \bar{x}) \tag{8.2}$$

where $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. If the x_is' are independent and identically distributed, then $W_n^0(t) \to B^0(t)$ when $n \to \infty$. $B^0(t)$ is Brownian bridge, with variance $t(1-t)$. As a consequence, time dependant control limits of form $UCL = \lambda\sqrt{t(1-t)}$ and $LCL = -\lambda\sqrt{t(1-t)}$ allow to detect early and late shifts. The resulting chart is displayed in Figure 8.3b. The peak, close to date 0.8, corresponds to the change date [115]. In addition to detecting the change, the chart is adapted to the data in terms of distribution. As illustrated in Figure 8.3b, 1000 random samples of $W_n^0(t)$ are simulated, and fall within UCL and UCL.

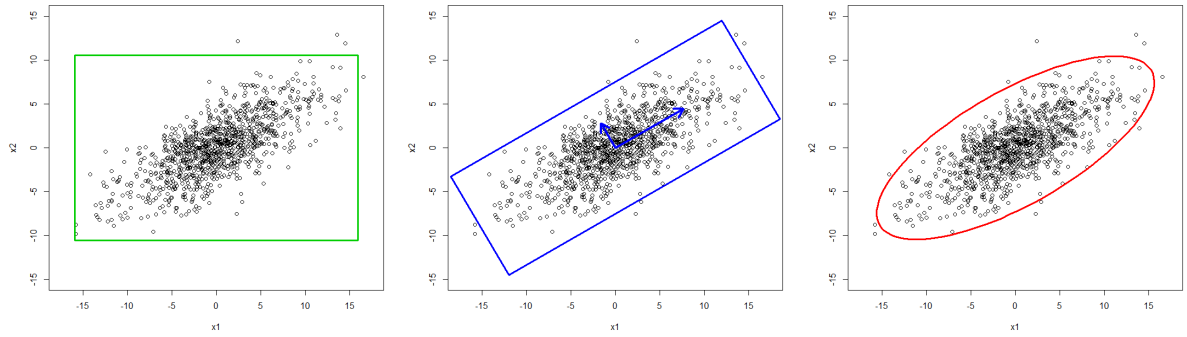


(a) Tabular CUSUM chart. (b) CUSUM chart using $W_n^0(t)$.

Figure 8.3 – Different CUSUM charts to detect a small drift.

8.1.3 Multivariate charts

With the advances of sensor technologies, there are many quality indicators nowadays. Monitoring them separately from one another may be misleading. In particular, when they are correlated, univariate control charts are no longer suitable. In the example in Figure 8.4a, x_1 and x_2 are correlated. In this case, univariate charts will surely fail to detect inter-variables deviations. Furthermore, when the number of dependent variables gets large, univariate control charts will generate too many false alarms (see e.g. [76], Chapter 11). Though Bonferroni corrections can help to overcome false alarms problems in low dimension, multivariate approaches are employed to consolidate the large number of dependent variables into some simple indicators. A common practice is to project multi-dimensional data onto orthogonal directions and monitor the resulting latent variables (see e.g. [65]). Among existing projection methods, Principal Components Analysis is mostly used. In the case of linear correlations, it significantly improves the detection of univariate control charts as shown in Figure 8.4b. Such procedures are very helpful, especially for very high dimensional problems wherein several variables are redundant and noisy. Only a significant subset of the latent variables will be selected by means of dimension reduction. Given the poor knowledge and



(a) Univariate charts (b) PCA chart (c) Hotelling's T^2 chart

Figure 8.4 – Different control charts for a 2-dimensional Gaussian vector.

assumptions on the distribution of the dataset, the Hotelling's T^2 chart is a standard choice. Its control bound appears to be an elliptical confidence region, in accordance with the data distribution. As illustrated in Figure 8.4c, the Hotelling's T^2 chart is suitable for monitoring Gaussian distributed variables. Comprehensive studies on control charts can be found in [76] and [18].

8.2 Profile monitoring

raditionally, quality is defined as the fitness of a product or service to meet the requirements, such as specifications and satisfaction, of the user. Montgomery[76] introduced a new definition of quality, which is inversely proportional to variability. Profile monitoring is absolutely a part of this modern vision. It tackles the issue of process control when the quality to monitor is described by a functional relationship between a response variable and one or more explanatory variables [80]. Considered as the most promising area of research in SPC [113], profile monitoring meets several needs in industry, such as quality control, and pattern recognition and classification. The problem is this: at each time-stamp, the object to control is a curve, a spatial profile, or characteristics in higher dimensions, represented by n observations of the input variables $\mathbf{X}$, and the n corresponding responses $\mathbf{y} = y_1(t), \ldots, y_n(t)$. Profile monitoring is addressed through a standard approach including two steps [80, 113]:

1. At each time stamp, the profile model $\mathbf{y} = f(\mathbf{X})$ is estimated;
2. Some parameters of f are monitored over time with a control chart.

Based on the concept above, several profile monitoring schemes have been developed. As an example in a healthcare application, [63] proposed T^2 control charts to monitor the parameters of linear models. In automotive industry, [5] used control charts to monitor the parameters of a linear mixed model, including a second-order polynomial and autocorrelated residuals. [59] developed a different approach for stamping processes. They decomposed 1D profiles into segments, where the levels are monitored, because the faults to detect affect only few portions of the profile. To deal with multiscale structures, [91] used wavelets to monitor paper surface through two separate Shewhart control charts. A comprehensive study of profile monitoring, with further applications is provided by [80]. In the specific case of wafers, that is a circular input domain, [37] used thin-plate splines to detect abnormal

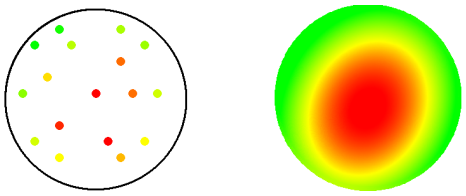
products. Rather than the profile itself, [10] investigated spatial variations of the variance. The resulting indicator, called spatial variance spectrum, is then monitored with control limits calculated based on a χ^2 distribution. Though [89] obtained promising results by modelling profiles with GP models and Zernike polynomials, we remark that there are no control charts based on GP parameters or Zernike coefficients in the literature. In the profile monitoring scheme of [19] using Gaussian processes, the monitored indicator is rather formulated through deviations from a target profile.

8.2.1 Profile monitoring based on Zernike polynomials

In this section, we implement the two-steps profile monitoring procedure, based on Zernike polynomials and Hotelling's T^2 control chart. The methodology is illustrated with an industrial dataset, involving $q = 506$ wafers processed one after the other. The response $\mathbf{y}$ is a physical characteristic measured at 17 locations on each wafer.

Step 1: profile model

Due to the properties presented in 2.1, Zernike polynomials represent the natural polynomial model for wafers profile monitoring in circular domains. This first step consists in fitting $\mathbf{y} = f(\mathbf{X})$ for each wafer, where f is given by Equation (2.19). Now, the details of the model are provided for a single wafer in Figure 8.5. Given the 17 measurements, a two-order Zernike regression is used to avoid over-fitting. The estimated shape shows a strong radial



Observed data Fitted profile

Figure 8.5 – Profile estimation with Zernike polynomials.

Table 8.1 – Zernike regression. Coefficients β_n^m and p-values

Polynomials	Z_0^0	Z_1^{-1}	Z_1^1	Z_2^{-2}	Z_2^0	Z_2^2
Coefficient $\hat{\beta}_n^m$	–	-0.010	0.002	0.006	-0.028	-0.005
p-values (%)	< 0.1	< 0.1	41	0.5	< 0.1	8.8

effect, confirmed by the regression coefficients in Table 8.1. Indeed, $\hat{\beta}_2^0$, corresponding to the radial polynomial Z_2^0 is the largest in absolute value. In addition, the model exhibits two main directions of deformation, obtained by rotating the x and y axes by 80 degrees approximately. The resulting model corresponds to an adjusted coefficient of determination $R^2_{\text{adj}} = 94\%$. It confirms the existence of a link between measurements and locations, and the second step of profile monitoring can be performed.

Step 2: control chart

The regression model presented above summarizes a wafer profile in $p = 6$ coefficients: $\hat{\beta}_0^0$, $\hat{\beta}_1^{-1}$, $\hat{\beta}_1^1$, $\hat{\beta}_2^{-2}$, $\hat{\beta}_2^0$, $\hat{\beta}_2^2$. Actually, the q wafers profiles are represented by a $q \times p$ matrix $\mathbf{M}$ where the columns and rows correspond to the regression coefficients and wafers indices respectively. This second step consists in monitoring $\mathbf{M}$ over time, based on a multivariate control chart (Section 8.1.3). Notice that the approach differs from traditional SPC which would only involves the monitoring of the constant term β_0^0. By assuming that $\mathbf{M}$ has Gaussian entries and that spatial patterns are independent over time, the monitoring can be done with the Hotelling's T^2 chart. The vector of Hotelling's T^2 statistics is given by:

$$T^2 = (\mathbf{M} - \hat{\mu})^\top \hat{\mathbf{\Sigma}}^{-1} (\mathbf{M} - \hat{\mu}) \tag{8.3}$$

where $\hat{\mathbf{\Sigma}}$ and $\hat{\mu}$ are the estimated mean and covariance matrix of $\mathbf{M}$. By denoting $F_\alpha(k, n-k)$ the α-percentile of the F-distribution with parameters k and $n-k$, the upper control limit of T^2 is:

$$T^2_{UCL} = \frac{k(n-1)(n+1)}{n(n-k)} F_\alpha(k, n-k) \tag{8.4}$$

As mentionned in Section 8.1.1, T^2 statistics is ideally estimated with an historical dataset, collected under normal operating conditions [76]. In this experimental study, such data are not available. We use robust estimations for $\mathbf{\Sigma}$ and μ in order to avoid different kinds of bias due to process instability or outliers [95, 83]. The resulting chart is displayed in Fig. 8.6.

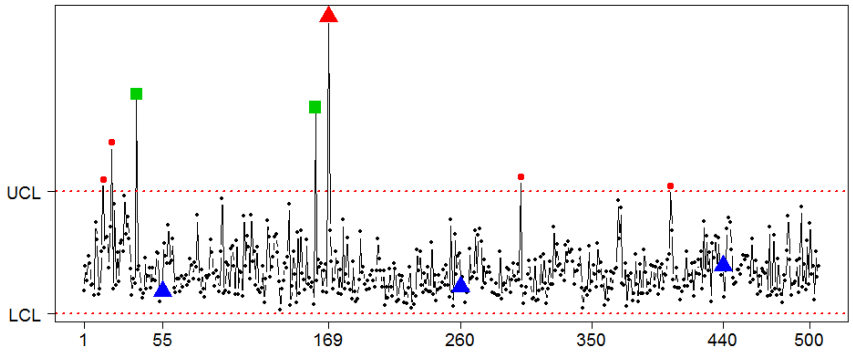


Figure 8.6 – Hotelling's T^2 control chart for the 506 wafers

Among the 7 abnormal signals, we focus on the most significant one, corresponding to wafer 169. For this wafer, we display in Figure 8.7 (see [76], Section 11.3 for more details) a decomposition of the T^2 statistic, indicating the relative contribution of each Zernike polynomial. These contributions reveal that coefficients β_2^{-2} and β_2^2 are the main causes of the problem. The underlying Zernike polynomials Z_2^{-2} and Z_2^2 suggest an angular drift.

To visually confirm the result, we display the four profiles, corresponding to wafers 55, 169, 260 and 440 which are marked with triangles in the control chart. The corresponding profiles in Figure 8.8 exhibit a pronounced dissimilarity of the $169th$ wafer compared to the others. Actually, it is obvious that monitoring the mean value β_0^0 would not allow to detect the problem. By detecting the dissimilarity in spatial patterns, the implemented profile monitoring scheme is then an improvement, compared to standard SPC. Notice that the time-series of the T^2 statistics in Figure 8.6 reveal a slight autocorrelation. This temporal effect is minor in the example and the analysis of time series will be addressed in Chapter 10.

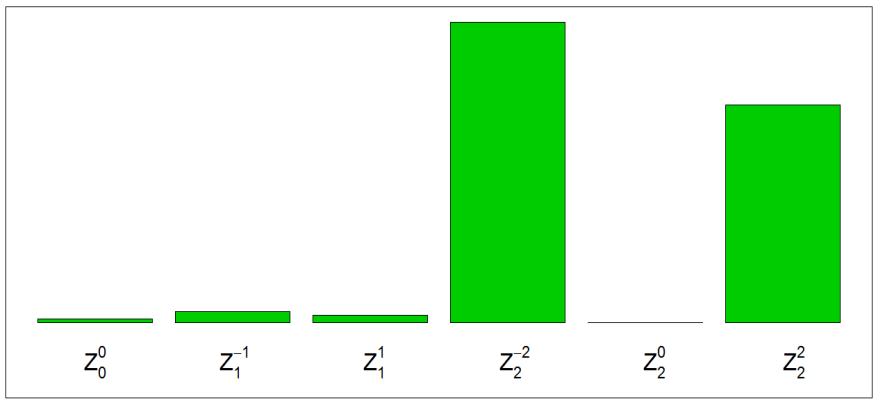
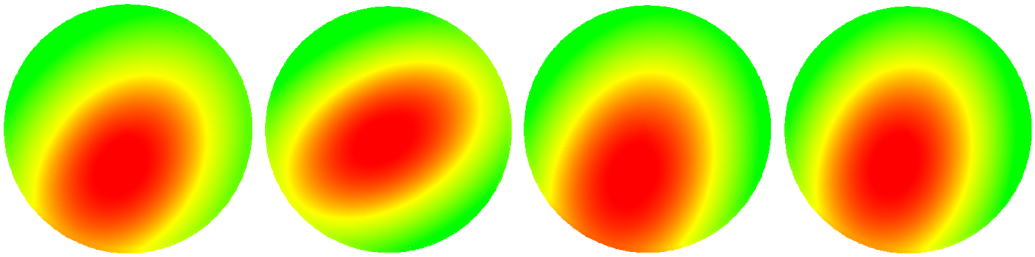


Figure 8.7 – Partial contributions to the T^2 statistics for wafer 169.



Wafer 55 Wafer 169 Wafer 260 Wafer 440

Figure 8.8 – Profiles of the 4 wafers marked with triangles in Fig. 8.6.

8.2.2 Profile monitoring based on Gaussian processes

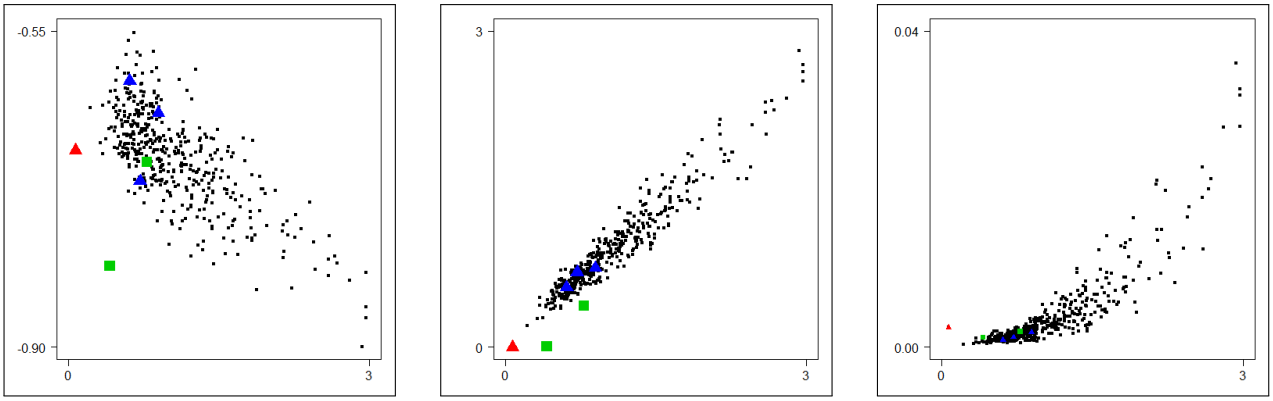
As shown in the first part of this thesis, Kriging allows to improve wafers spatial models. Therefore, we are wondering whether Kriging parameters, like Zernike coefficients, can be used in profile monitoring. For validation, we use the same dataset of 506 wafers and focus on wafers 169, 37 and 160 which represent the main out-of-control signals according to Zernike polynomials. The wafers 55, 260 and 440 are marked with blue triangles.

Profile monitoring based on a Cartesian GP

Profile monitoring is now applied to the sequence of 506 wafers. The polynomial regression is replaced by a GP model with constant trend and a tensor-product Matérn$_{\frac{3}{2}}$ kernel:

$$\mathrm{cov}\Big(Z(\mathbf{x}), Z(\mathbf{x}')\Big) = \sigma^2 k_{m\frac{3}{2}}(x_1, x_1'|\ell_1)\, k_{m\frac{3}{2}}(x_2, x_2'|\ell_2)$$

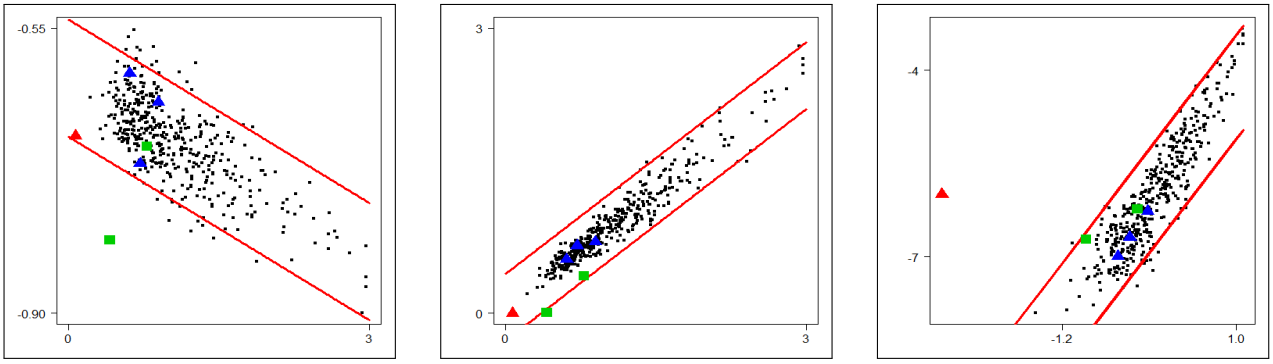
$\mathbf{M}$ is now a 506 by 4 matrix whose columns correspond to estimated values of $(\ell_1, \ell_2, \mu, \sigma)$. For the sake of readability, we omit the usual hat indicating that the parameters are estimated. We observe in Figure 8.9 that (σ^2, ℓ_1) is not drawn from a Gaussian distribution and the Hotelling's chart does not apply to $\mathbf{M}$. However, an important fact in Figure 8.9 is the existence of a relationship between the different pairs of estimated parameters. Given the scatter-plots, these relationships should be readily described by parametric models. A linear regression confirms that $\ell_2 \simeq 0.8\ell_1 + 0.07$ with coefficient of determination $R^2 = 92\%$. The heteroscedasticity observed in Figure 8.9c led us to investigate the link between ℓ_1 and σ^2 in a logarithmic scale. This resulted in $\log(\sigma^2) \simeq 1.8\log(\ell) - 6$ with $R^2 = 81\%$. Though less significant, the affine relation between μ and ℓ_1 is modelled as $\mu \simeq -0.08\ell_1 - 0.6$,



(a) μ vs ℓ_1 (b) ℓ_2 vs ℓ_1 (c) σ^2 vs ℓ_1

Figure 8.9 – Estimated Kriging parameters for 506 wafers.

corresponding to a coefficient of determination of $R^2 = 53\%$. In Figure 8.10 are represented the fitted models.



(a) $\mu = -0.08\ell_1 - 0.6$ (b) $\ell_2 = 0.8\ell_1 + 0.07$ (c) $2\log(\sigma) = 1.8\log(\ell_1) - 6$

Figure 8.10 – Monitoring of the residuals of 3 regressions models that link Kriging parameters.

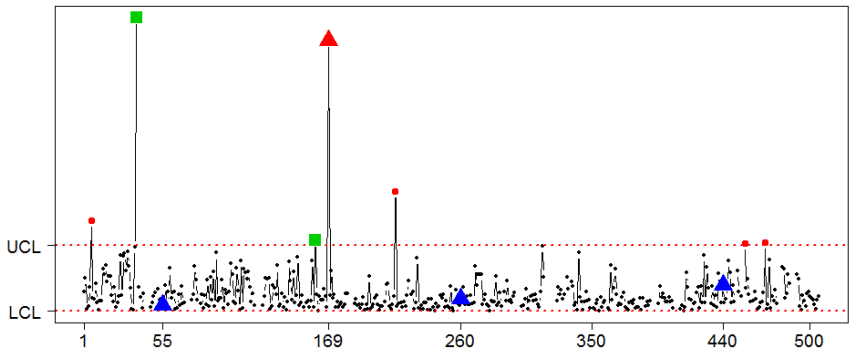
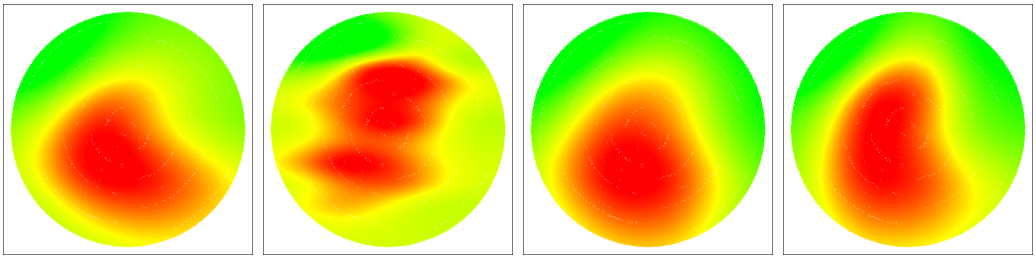


Figure 8.11 – Hotelling's T^2 control chart for $(\epsilon_1, \epsilon_2, \epsilon_3)$.

From now on, we denote the residuals of the regression models by $\epsilon_1, \epsilon_2, \epsilon_3$, displayed in Figures 8.10a, 8.10b and 8.10c. Based on the assumption that $\epsilon_1, \epsilon_2, \epsilon_3$ follow a Gaussian distribution, 95% control limits are set for each pair of Kriging parameters. This would result

in the 3 independent control charts displayed in Figures 8.10. In order to control the false alarm rate in higher dimension, we use the Hotelling's T^2 chart to monitor $(\epsilon_1, \epsilon_2, \epsilon_3)$. In Figure 8.11, we see that this multivariate control chart succeeds in detecting the $169th$ wafer in addition to the two other reference signals, namely wafers 37 and 160. As an example, the profile of wafer 169 is drawn in Figure 8.13. We observe that it has a different pattern.



Wafer 55 Wafer 169 Wafer 260 Wafer 440

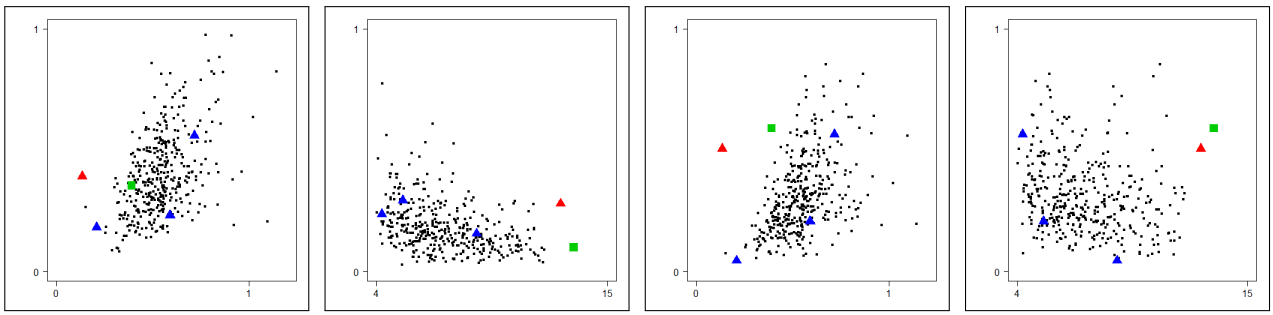
Figure 8.12 – Profiles of the 4 wafers marked with triangles in Fig. 8.11.

Profile monitoring based on a centred polar Gaussian process

The considered model involves 4 independent terms, corresponding to the functional ANOVA decomposition presented in Section 4. The kernel of the underlying centred GP is:

$$\mathrm{cov}\Big(Z(\mathbf{x}), Z(\mathbf{x}')\Big) = \sigma_1^2 k_1^0\big(\rho, \rho'|\ell\big) + \sigma_2^2 k_2^0\left(\theta, \theta'|\tau\right) + \sigma_{12}^2 k_1^0\big(\rho, \rho'|\ell\big) k_2^0\left(\theta, \theta'\tau\right)$$

where k_1^0 and k_2^0 are centred Matérn$_{\frac{5}{2}}$ and C^2-Wendland kernels over $[0,1]^2$ and $\mathbb{S}^2$. In Figure 8.13, we observe via the estimated parameters that wafers 169 and 160 are far from the other observations. The same goes for wafer 37 which is beyond the limits of the graphics to provide a clear visualization of the different scatter-plots. Based on the procedure presented for Cartesian GPs, we display in Figure 8.14 a Hotelling's T^2 control chart for $(\ell, \tau, \sigma_1^2, \sigma_2^2, \sigma_{12}^2)$. As observed, the abnormal wafers 169, 160 and 37 are detected. In particular, wafer 37 was removed before fitting the different regression models to avoid biasing the T^2 control limits.



(a) σ_1^2 vs ℓ (b) σ_2^2 vs τ (c) σ_{12}^2 vs ℓ (d) σ_{12}^2 vs τ

Figure 8.13 – Estimated parameters of a centred polar GP for 506 wafers.

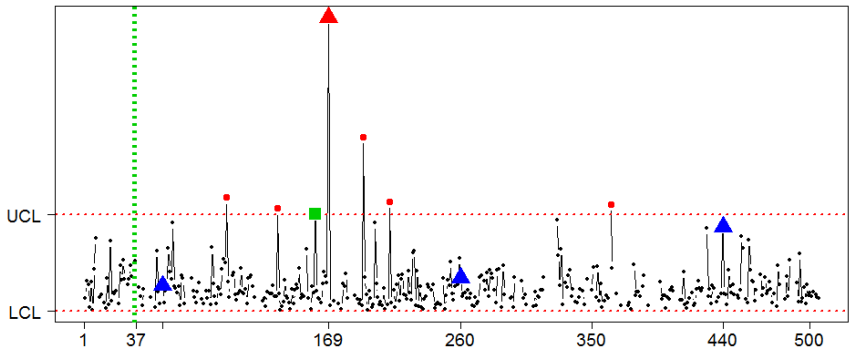
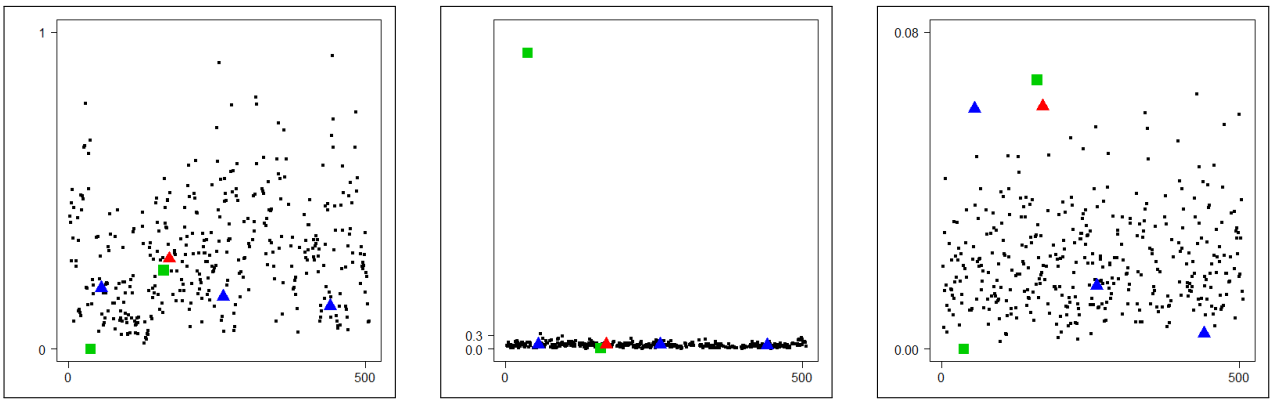


Figure 8.14 – Hotelling's T^2 control chart based on a polar GP, after transformation and whitening.

8.2.3 Profile monitoring based on Sobol decomposition

Compared to polynomial regression, Kriging is harder to interpret due to the non-parametric nature of GPs. However, as explained in Chapter 4, Sobol indices allow to interpret GPs in terms of variances. The resulting variances are furthermore related to independent processes. In Figure 8.15 are displayed the 506 wafers Sobol indices. One finding is immediate: monitoring the Sobol indices would allow to detect the abnormal profiles. In particular, wafer 37 differs too much from the population by its very large "angular variance". Wafers 160 and 169 too have large "interaction variances", when compared to the others.



(a) Radial. (b) Angular. (c) Interaction.

Figure 8.15 – Sobol indices for 506 wafers, based on a centred polar GP.

8.3 General methodology and practical issues

There are two important facts in this chapter. First, the control chart based on Zernike polynomials is easily implemented and interpreted. Conversely, it is harder to monitor GP parameters. Secondly, the GP parameters are mutually linked. As a consequence, when monitoring Kriging parameters, being far from the other observations in the sense of the Euclidean distance is not always a sign of dissimilarity (see an illustration in Figure 8.9a,

or [119] for asymptotic properties). To monitor such "structured" datasets, Hawkins [50] introduced control charts based on Regression Adjusted Variables. The procedure consists in monitoring the residuals from the regression of each variable on all others. Given the empirical results here-above, we propose a 3 steps monitoring scheme to adapt regression adjustment in the framework of Kriging:

1. At each time stamp, estimate the Kriging model $\mathbf{y} = f(\mathbf{X})$;
2. Perform a linear (or log-linear) regression for each pair of estimated parameters;
3. Monitor the regression residuals with standard control charts.

In Section 8.2.2, the regressions corresponding to step 2 were motivated by empirical considerations. However, the resulting models are coherent with the dependencies among Kriging parameters (see e.g. [20], Section 5.3.3). For a practical implementation, we provide some relevant quantities that can be monitored.

- The ratio $\frac{\ell_1}{\ell_2}$ of characteristic lengths for tensor-product Cartesian GPs which was used by [89] to describe the differences between clusters of profiles. It is furthermore consistent with the empirical finding $\ell_2 = 0.8\ell_1$ in Section 8.2.2.
- The angle ϕ of geometric anisotropy which is identifiable modulo $\frac{\pi}{2}$ unless the process is isotropic [48]. It should be monitored based on a suitable periodic distribution.
- The coefficients a and b in the linear model $\log(\sigma) = a\log(\ell) + b$ for Matérn kernels. For a large number of observations in a bounded domain, monitoring these coefficients would indicate if the successive profiles are equal [119].
- The Sobol indices based on centred GPs since they represent the variance ratios of independent processes as explained in Section 4. Monitoring them would allow to detect changes in the process variance as shown in this chapter.

Résumé en Français

La Maitrise Statistique de Procédés (MSP ou SPC en anglais) fait référence à l'emploi d'outils statistiques pour le suivi et l'amélioration de performances en entreprise. Les cartes de contrôle représentent l'outil le plus intuitif et le plus utilisé dans cette discipline. Dans ce premier chapitre, nous avons fait une revue des principes de base de la SPC, notamment les différents types de cartes de contrôle: univariées ou multivariées, adaptées à la détection de changements brusques ou de dérives lentes. Les cartes sont ensuite utilisées dans le but suivre dans le temps des données spatiales définies sur le disque. Ce problème, connu en anglais sous le nom de "Profile Monitoring", se résoud en deux étapes de façon standard. La première étape consiste à modéliser (surface de réponse) les données spatiales à chaque pas de temps. La deuxième utilise les outils de la SPC pour le suivi temporel des paramètres des modèles spatiaux. Notre contribution a été de proposer des cartes de contrôles pour les profils spatiaux, en utilisant les coefficients de Zernike d'une part et les paramètres de processus Gaussiens d'autre part. A l'issue de l'étude, nous avons remarqué que les polynômes de Zernike offrent un cadre simple et interprétable pour suivre dans le temps des données définies sur le disque. les paramètres de krigeage, qui ne sont pas gaussiens, sont plus difficiles à modéliser avec les cartes de contrôle standard. Pour faire face à ce problème,

nous avons proposé une démarche empirique consistant à modéliser les résidus de régressions entre différents paramètres de krigeage. Les aspects pratique de la méthodologie sont enfin évoqués.

Chapter 9

Spatial Pattern Prediction and Diagnosis

Up to now, wafers patterns are reconstructed from spatial observations, and control charts are developed in order to detect abnormal patterns and potential changes in the process. The aim in this chapter is to link these statistical models to manufacturing parameters. For this purpose, the focus is on Zernike polynomials which provide a simple and intuitive decomposition of spatial patterns.

9.1 Problem and motivation

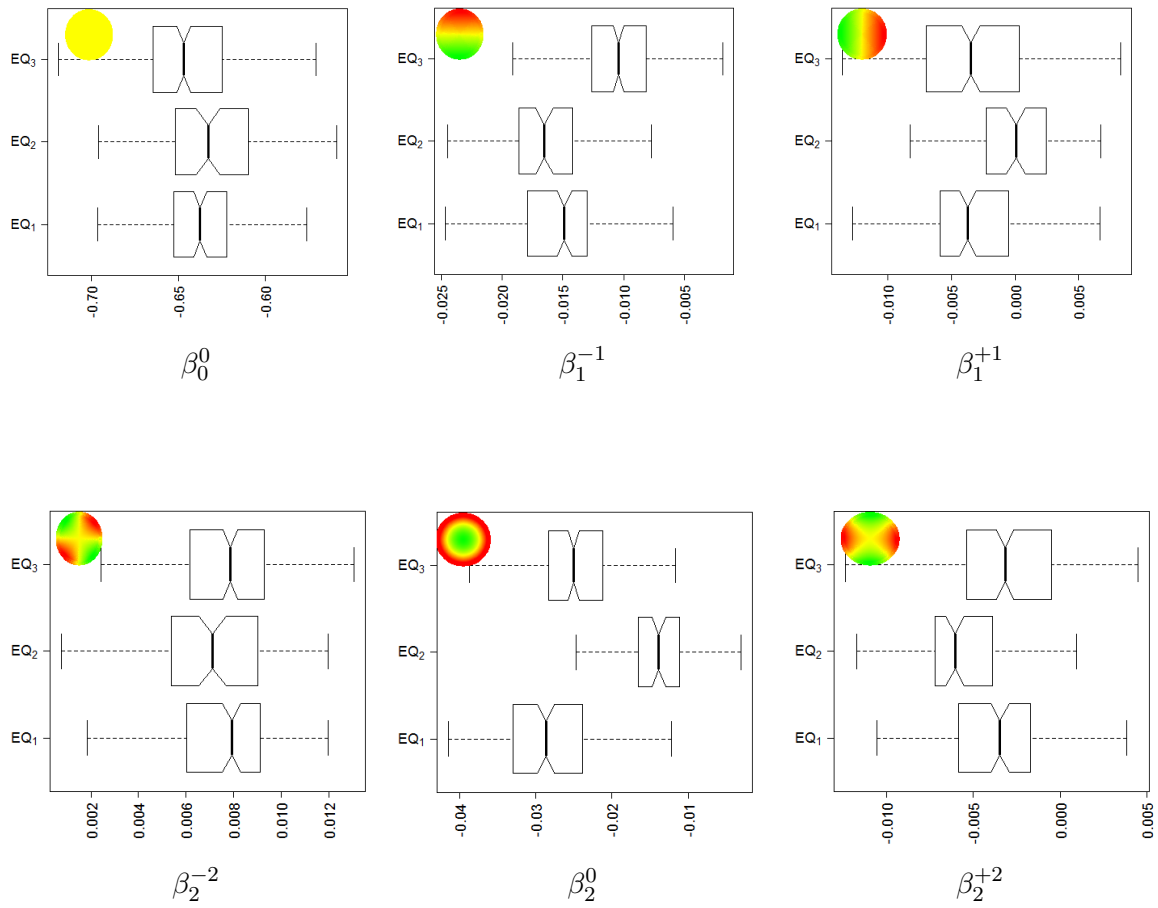


Figure 9.1 – Regression coefficients for the 506 wafers, grouped by three process machines.

To introduce the problem, we consider the dataset of the 506 wafers monitored in Section 8.2.1. The manufacturing process, which is not detailed here, was done by 3 different machines (EQ_1, EQ_2, EQ_3). In Section 8.2.1, each spatial pattern was modelled with 6 Zernike polynomials, resulting in a 506 by 6 matrix $\mathbf{M}$ where rows correspond to wafers and columns indicate the Zernike coefficients: β_0^0, β_1^{+1}, β_1^{-1}, β_2^{-2}, β_2^0 and β_2^{+2}. These coefficients are grouped by machines and displayed in Figure 9.1. Significant differences among the three machines for the same product can be observed intuitively. Similar studies were exploited by [37] who modelled spatial patterns with splines in order to detect and classify equipment faults.

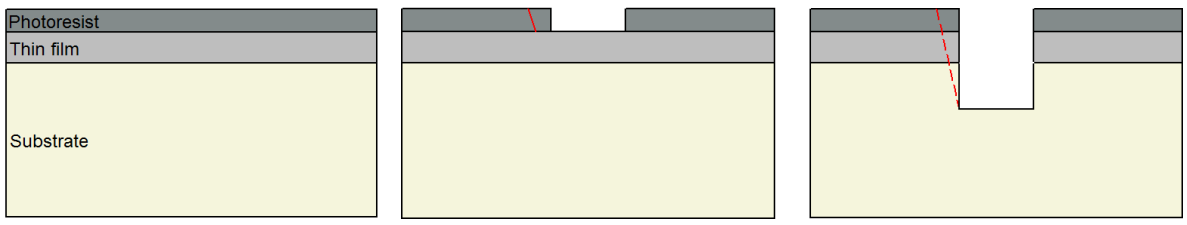
Other than the impact from machines, wafers spatial patterns can be linked with the process physics too. For instance, heating from the center (Figure 1.4a) is a process that surely generates a radial pattern, which corresponds intuitively to high values of β_2^0. Given the knowledge on the production executive records, spatial patterns can firstly be decomposed into significant Zernike polynomials and then explained further. In general, there are hundreds of manufacturing parameters that potentially influence the wafers profiles. Rather than reconstructing the spatial patterns from measurements at sampled locations, the purpose of the study in this chapter is to infer and explain them, based on the available processing characteristics.

Apart from the causal relationships among processing characteristics and spatial patterns, there is no clear understanding on the effects of each parameter. The predictors are furthermore dependant on one-another. Given that, causal models such as bayesian networks [87] would be a relevant choice to predict and interpret spatial patterns. However, these probabilistic models rely on a representation of conditional dependencies as a directed acyclic graph. This hypothesis is not fulfilled in the presence of feedback regulation, which is a very common control mechanism in semiconductor fabs. Therefore, we focus on linear regression models, with the essential purpose of providing interpretation and diagnosis tools in industry.

9.1.1 Characteristics of Process Variables

The influence of successive processing steps

Given that semiconductor manufacturing involves hundreds of processing steps, a relevant study of a single production step cannot be conducted regardless of the previous ones. For instance, it is well-known that wafers surfaces resulting from etching processes will depend on the previous operation (lithography). As can be seen in Figure 9.2, an imperfect lithography process may impact the product quality after etching.



Initial surface. After lithography. After etching.

Figure 9.2 – An example of dependency between two production steps in microelectronics.

Variables related to one process step

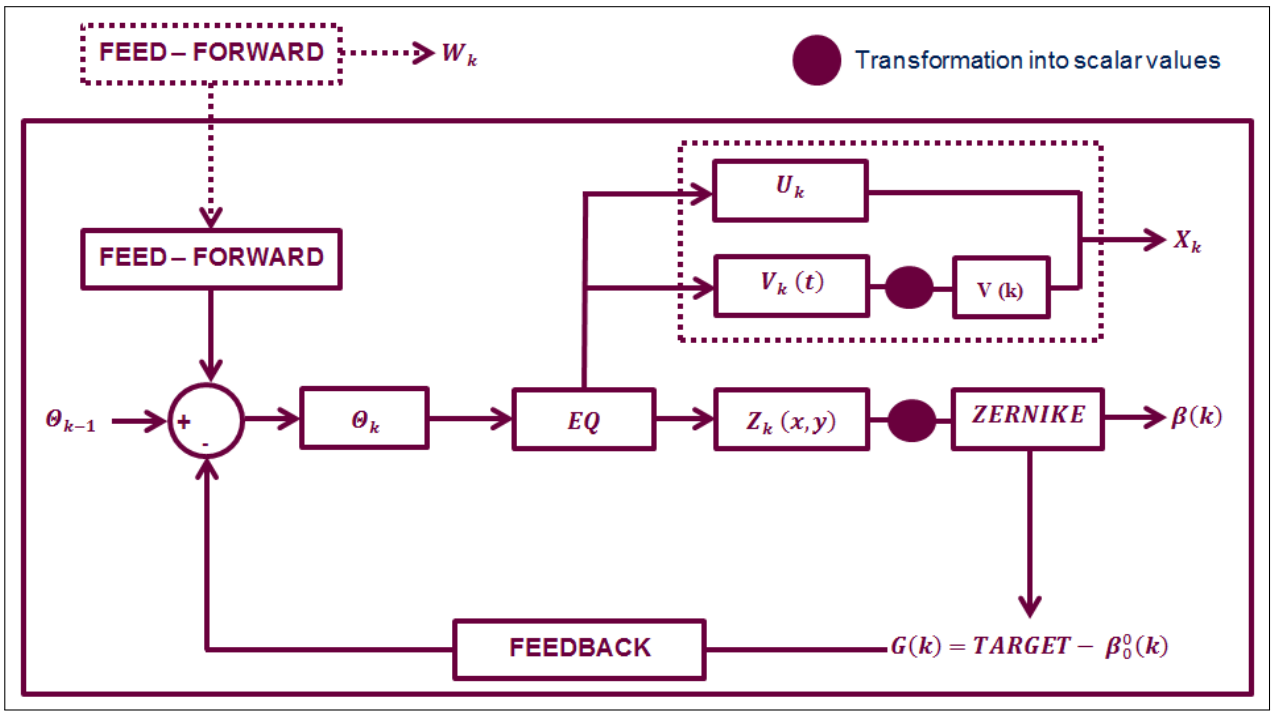


Figure 9.3 – Schematization of a regulated process control framework.

To introduce the different variables involved in the prediction of spatial patterns, we consider the framework of a production step in Figure 9.3. The successive wafers, processed by a machine EQ are numbered by their running orders, and k is the wafer currently processed by EQ. $\mathbf{w_k}$ denotes the measurements recorded over k after the previous production step. The operating parameters of EQ are expressed as Θ_k, which is provided by the regulation system, referred to Run-to-Run control (R2R) in semiconductor manufacturing [24]. R2R allows dynamic modifications of operating parameters between successive product runs in order to minimize the gap to the target and the variability of the process. Θ_k is then the model-based information used to compensate the deviation observed in the measurements of the precedent wafer.

When the wafer is processed in the machine, embedded sensors are turned-on to monitor the key physical parameters, such as temperature, gas flow, pressure. These data are usually collected continuously during the process and expressed as multivariate time series ($\mathbf{V_k}(t)$). They are conventionally referred to Fault Detection and Classification (FDC) data in semiconductor industries, practically summarized into scalar values (intercept, slope, max, min ...), and stored in the vector $\mathbf{V_k}$. In parallel, automation parameters such as the machine name, the lot name and counters related to maintenance are recorded and represented by the vector $\mathbf{U_k}$. Evidentally, $\mathbf{U_k}$ is made of discrete and categorical variables.

At the end of the product run, the processed wafer is assigned to quality control. Physical and electric tests are assessed at n different locations on the wafer. It results in a n by 3 matrix $\mathbf{Z_k} = (x_i, y_i, z_i)_{i=1...n}$ where $(x_i, y_i)_i$ are the coordinates of the measured points

and z_i's are the associated responses. At this stage, there are two ways to represent the spatial values, depending on whether the measurement locations are fixed or not. When the measurement locations are the same from wafer to wafer, $\mathbf{Z_k}$ is represented by the vector $\mathbf{z}_k = (z_1, \dots z_n)$. Otherwise, $\mathbf{Z_k}$ is represented by the coefficients of a d-order Zernike regression $\mathbf{y}_k = \left(\beta_0^0(k), \dots, \beta_d^{+d}(k)\right)$.

9.1.2 Link the Product Quality to Process Characteristics

With all the process parameters settled down, $\mathbf{Z_k}$ can be modelled, based on successive observations of $\left(\Theta_k, \mathbf{U_k}, \mathbf{V_k}, \mathbf{W_k}\right), k = 1 \dots N$. We recall that each spatial response $\mathbf{Z_k}$ is represented by the vector $\mathbf{y_k}$, corresponding to the spatial measurements, or Zernike coefficients. Switching from one to another is simply done by regression or computation of polynomials. For the sake of readability, we suppose that $\mathbf{y_k}$ corresponds to Zernike coefficients with $d = 2$, i.e.

$$\mathbf{y_k} = \left(\beta_0^0(k), \beta_1^{+1}(k), \beta_1^{-1}(k), \beta_2^{-2}(k), \beta_2^0(k), \beta_2^{+2}(k)\right)^\top,$$

where $\mathbf{y_k} \in \mathcal{M}_{1,q}(\mathbb{R})$, with $q = (d+1)(d+2)/2$ being the number of Zernike coefficients. Let $\mathbf{x_k} = \left(\Theta_k, \mathbf{U_k}, \mathbf{V_k}, \mathbf{W_k}\right)^\top$ be the inputs of successive observations where categorical predictors are transformed into binary variables by dummy coding. An essential point to keep in mind is the common identity of spatial measurements. Obviously, the n measured locations over one wafer share the same records through manufacturing operations, except the xy-coordinates. This implies that point-to-point predictors, except the xy-coordinates, are the same. Taking into account the xy-coordinates of the measurements, the q Zernike coefficients can be modelled with the same predictors. The regression model is then formulated as:

$$\mathbf{y_k} = \mathbf{b} + \mathbf{A}^\top \mathbf{x_k}, \tag{9.1}$$

where $\mathbf{b} = (b_1, \dots, b_q)^\top \in \mathbb{R}^q$ and $\mathbf{A} = \left(a_{ij}\right)$ is a p by q matrix, p being the number of predictors in $\mathbf{x_k}$, including dummy variables. Expanding Equation (9.1) leads to:

$$\begin{bmatrix} \beta_0^0(k) \\ \vdots \\ \beta_d^d(k) \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_q \end{bmatrix} + \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{q1} & \cdots & a_{qp} \end{bmatrix} \mathbf{x_k} \tag{9.2}$$

Given 9.2, the rows of the matrix $\mathbf{A}^\top$ are computed separately. The underlying assumption is that the outputs are mutually independent because of the orthogonality of Zernike polynomials. Therefore, q linear regression models are evaluated, corresponding to the q Zernike coefficients (based on the same predictor $\mathbf{x_k}$). While considering more than one process steps, there are many input variables (40 and 76 in the datasets below) and variables selection becomes a critical issue to obtain significant models.

9.1.3 Selection of Significant Process Variables

Given the large number of manufacturing parameters and the multivariate output, the influential variables are selected in two stages, in order to limit the negative effects of missing data. Firstly, each input-output pair is tested separately, based on the standard Pearson correlation for continuous inputs, the Kruskal-Wallis test for categorical ones, and the ANOVA (Analysis of variance) test for integers. Secondly, the q different models in 9.2 are estimated based on the subset of the selected pairs. The importance of each pair is now quantified by the *p-value*, resulting from these q regressions.

9.2 Case study

9.2.1 Process variables selection based on Zernike coefficients

In the first dataset for methodology validation, there are 40 predictors, involving 3 regulation parameters and a dozen of categorical variables, integers included. The modelled spatial patterns are represented by 6 Zernike coefficients: β_0^0, β_1^{-1}, β_1^{+1}, β_2^{-2}, β_2^0 and β_2^{+2}. In order to provide a relevant visualization, the variables selection results are summarized as a matrix $G = \left(G_{ij}\right), i = 1, \ldots 40, j = 1, \ldots 6$. G_{ij} being the importance of the ith input when modelling the jth output, G is displayed in Figure 9.4 where:

- black cells identify the non-selected input-output pairs,
- green cells indicate the selected pairs which are not significant,
- orange cells correspond to selected and significant pairs,
- red cells indicate selected and very significant pairs.

Regarding the ten outputs, the most influential process variables correspond to the subset $\{1, 4, 24, 36\}$. They include one regulation signal and three Zernike coefficients related to the previous manufacturing step. These influential variables stable enough from output to output. Remark that higher order polynomials seem to depend on fewer variables than the others. However, the variations observed in wafers profiles are due to non-constant terms, and therefore mainly explained by variables 24 and 36.

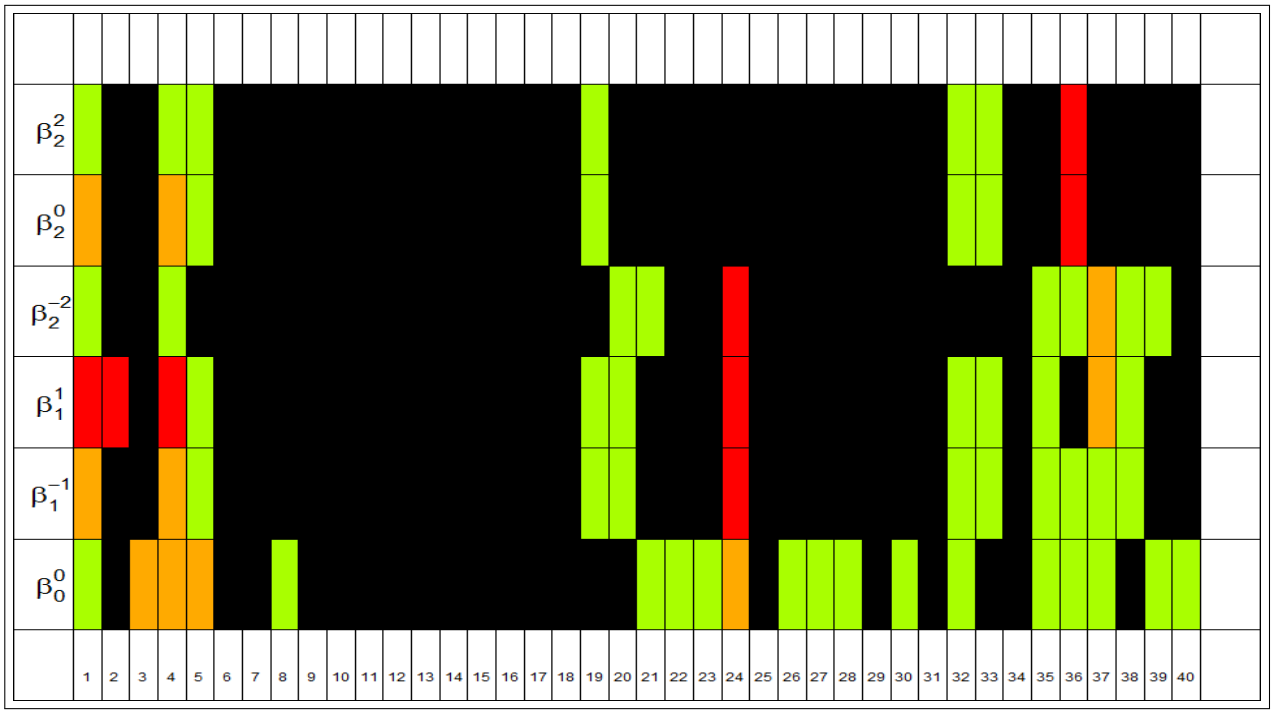


Figure 9.4 – Variable selection based on 40 inputs and 6 Zernike coefficients as output.

9.2.2 Process variables selection based on spatial measurements

As explained in Section 9.1.1, spatial data can also be represented by the n spatial values when the n measurement locations are fixed, which is the case for our second dataset. In Equation 9.1, the 6 Zernike coefficients are now replaced by $n = 17$ spatial measurements. The variables selection procedure is then conducted based on these 17 outputs. For this second dataset, the number of inputs is higher too (76), due to more variables from the previous process step. The matrix G, summarizing the variables selection, has now 17 rows and 76 columns. G is displayed in Figure 9.5.
In this second model, the three regulation signals are the most influential variables, followed by some measurements at the previous step. As an explanation, only the mean value is regulated over wafers. That makes the 17 spatial measurements vary together. Therefore, all the outputs depend on the same inputs. Remark that the variable 4 (machine name) is always selected but never significant in the final model. It becomes very important when regulation signals are removed from $\mathbf{X}$. Indeed, the machine name is already an input variable for the R2R feedback. Although the R2R signals are able to provide accurate predictions, they behave as black box systems and are more difficult to interpret than the concrete elements such as machines and measurements.

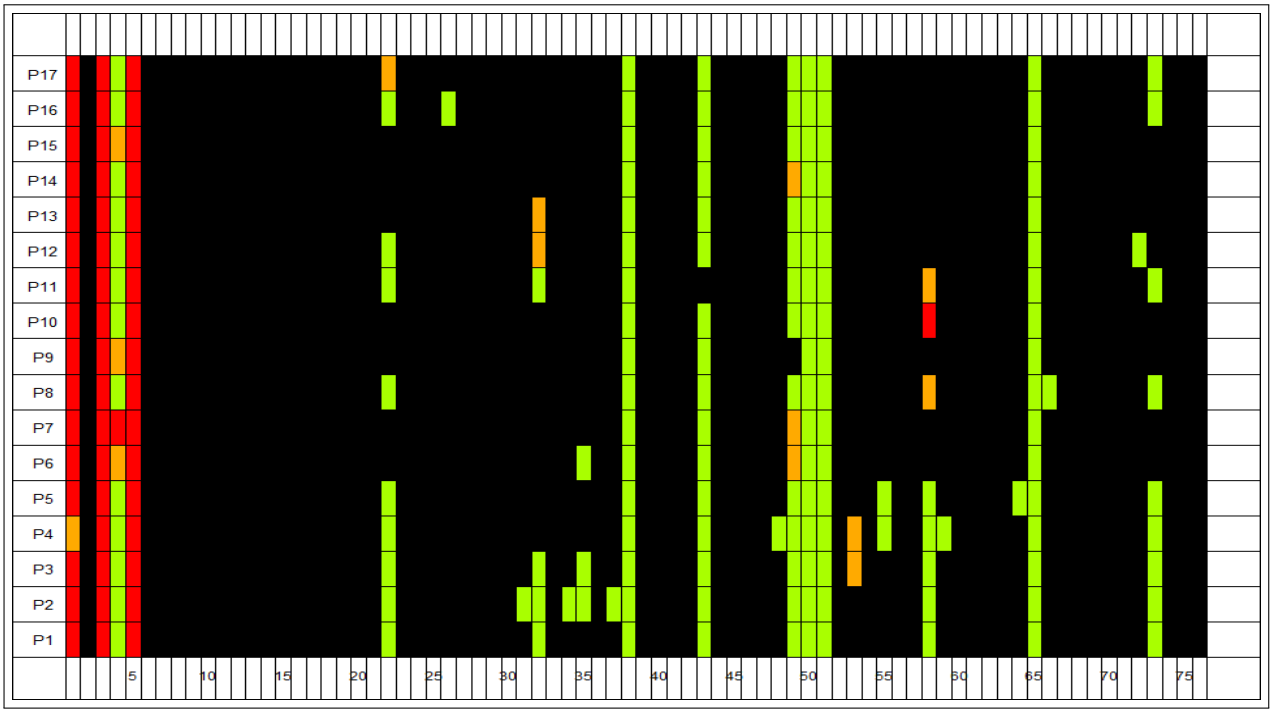


Figure 9.5 – Variable selection based on 76 inputs and 17 spatial samples as outputs.

9.2.3 Predictions

The model corresponding to the second dataset is now validated, and only the 5 most significant inputs are used. The predicted spatial measurements and the corresponding wafers profiles are displayed in Figure 9.6 and 9.7. As observed, wafers spatial patterns can be mainly explained by 5 key processing characteristics. The next challenge would be to embed such dependencies in response surface models after having consolidated the variable selection method.

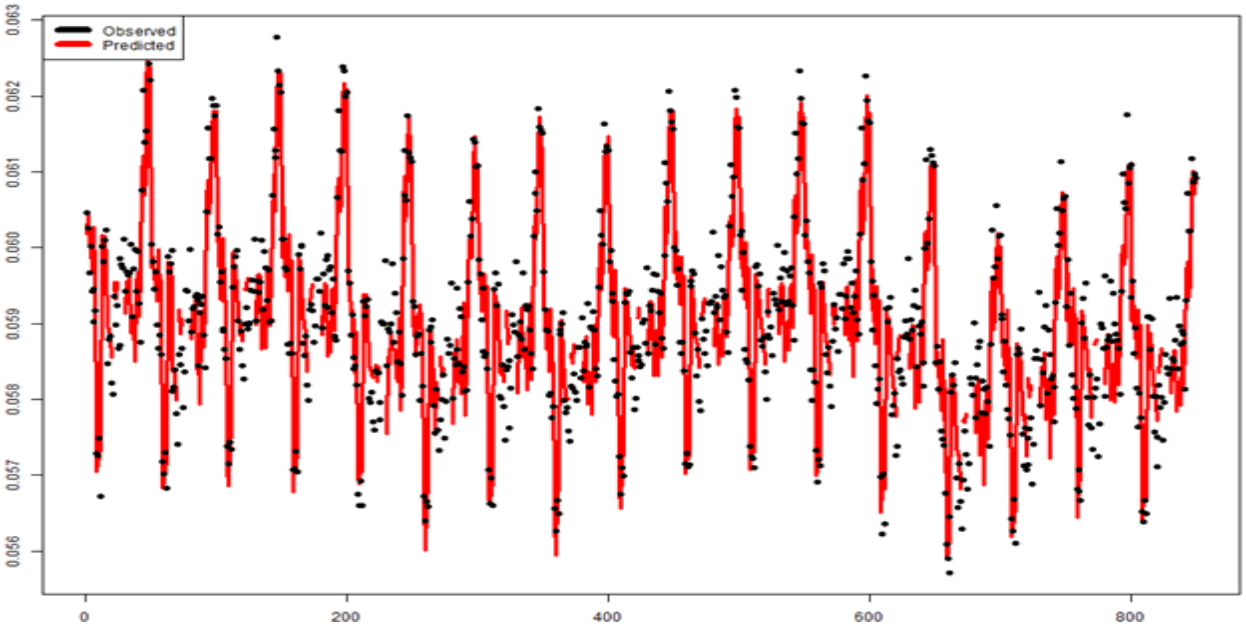
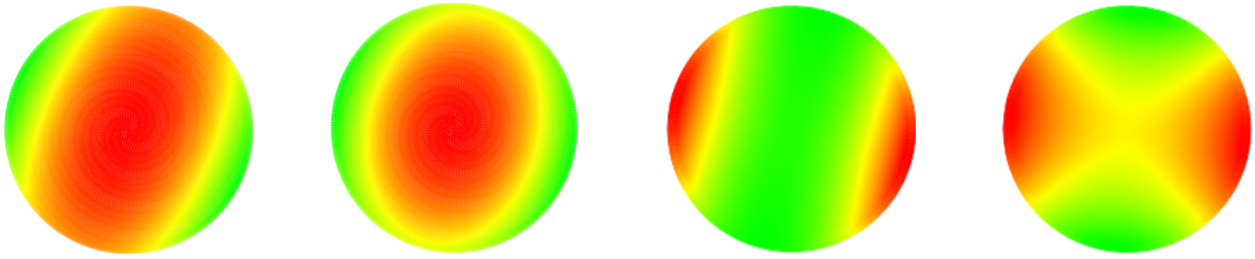
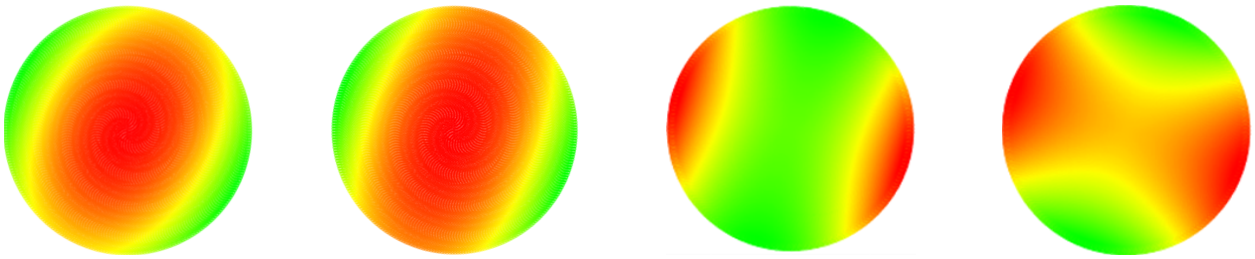


Figure 9.6 – Site-to-site predicted values.



(a) Prediction 1 (b) Prediction 2 (c) Prediction 3 (d) Prediction 4



(e) Observation 1 (f) Observation 2 (g) Observation 3 (h) Observation 4

Figure 9.7 – Fitted profiles, based on spatial measurements (observations), and predicted profiles based on processing characteristics (predictions).

Résumé en Français

Dans les chapitres précédents, les données spatiales sont modélisées à partir d'observations en un nombre fini de points sur le disque. L'objectif de cette étude est de prédire les données spatiales à partir de variables continues et catégorielles, représentant les paramètres de fabrication en microélectronique. La méthodologie que nous avons proposée consiste à représenter les courbes définies sur le disque par un nombre fini de coefficients de Zernike ou par les observations lorsque les points de mesure sont fixes. Ces représentants des courbes sont alors modélisés par régression en fonction des paramètres de production, puis reconver-

ties en données spatiales. La procédure ainsi définie permet décomposer les profils des wafers et de les lier aux paramètres de production, donc aux connaissances métier de l'industrie microélectronique. Dans le cas des données utilisées, on remarque notamment une forte influence des mécanismes de régulation et des équipements de production sur la qualité des produits. Par ailleurs, il est confirmé que les défauts spatiaux se transmettent d'une étape d'usinage à une autre.

Chapter 10

Robust monitoring of time series with structural changes

This paper presents original research initiated by the monitoring needs of a semiconductor production plant. The industrial operations rely on an Information Technology (IT) system, and several time series data are controlled statistically. Unfortunately, these variables often contain outliers, as well as structural changes due to external decisions in the IT activity. As a consequence, it has been observed that the monitoring results obtained with standard techniques could be severely biased.

This paper attempts to overcome such difficulties. A new monitoring method is proposed, based on robust Holt-Winters smoothing algorithm, and coupled with a relearning procedure for structural break detection. Such a method is flexible enough for a large-scale industrial application. We evaluate performance through simulation, and show its usefulness in real industrial applications for univariate and multivariate time series. The scope of application deals with IT activity monitoring, but the introduced statistical methods are generic enough for being used in other industrial fields.

The results of this chapter are based on the publication "Robust Monitoring of an Industrial IT System in the Presence of Structural Change" [83], by Padonou, Roustant and Lutz. The industrial outcome [70] was honored with the Greenfield Challenge prize in celebration of the International Year of Statistics.

10.1 Introduction

10.1.1 Industrial background and literature review

Manufacturing efficiency relies increasingly on Information Technologies (IT). This is clearly true for wafer fabrication plants, where this research is ongoing through a partnership with the company STMicroelectronics. Several activities are useful for proper IT management (see eg. Rudd *et al.* [99]). This paper is focused on monitoring. Careful monitoring is based on a cautious observation of the IT system. The objective is to have a close look at all data recorded and stored to track the activity of a plant IT system. As the size and complexity of IT systems strongly increase, the viability of monitoring can only be ensured by employing automated procedures as mentioned by Dugmore *et al.* [30]. In our background, methods intuitive enough to be reported on graphic charts were needed. Thus, threshold exceeding

detection mentioned by Rudd [99] is welcome. In addition, facing the enormous diversity and expanse of the components of modern IT systems, experts do not have always enough knowledge or time to determine a priori the critical thresholds.

There has also been a great interest in statistical methods of monitoring. Several publications have concerned the theoretical principles and the implementation of these methods (see e.g. Lowry *et al.* [68] , Montgomery [76] , Doganaksoy *et al.* [28]). Simultaneously several authors classified these monitoring tools according to the state of the monitored processes: independent vs dependent data, model-specific vs model generic methods, normal vs non normal distributions. In this way, they provided recommendations for the choice of the monitoring procedures (see e.g. Ben - Gal *et al.* [8] , Alfaro *et al.* [3] , Abbasi *et al.* [1])

Seeing the quantity and diversity of the data under study, automated procedures that do not require any preliminary hypothesis about the monitored variables, are required. In this case, monitoring requires the ability to distinguish normal system operations from exceptional events. To quantitatively detect such abnormal behavior in the fluctuation of a time series, one standard procedure is the following: 1) Model the time series to be monitored; 2) Monitor model residuals, through an appropriate control chart (Montgomery [76]). This statistical solution should be able to solve the problem of automatic monitoring of an IT information system.

A literature review presented in 2006 by De Gooijer and Hyndman [43] showed that many researches have been devoted to the first step of the procedure here-above. Among them, ARIMA models are famous for their ability to describe a large variety of industrial processes (Box and Jenkins [12, 13]). These models are expected to produce independent and normally distributed residuals, which allow the automatic setting of the critical thresholds in the second step.

Concurrently to ARIMA models, Holt-Winters smoothing is a flexible algorithm for time series that do not need any preliminary analysis. It is also a fairly good approximation for many kinds industrial variables (see Makridakis [71]), especially for those encountered in our research field. Since ARIMA models are equivalent to Holt-Winters smoothing under certain conditions (see e.g. Gardner Jr. [62] , Hyndman [56]), we decided to implement this latter, which is furthermore easier to manage by non-statisticians. By way of example, Hellerstein [51] developed an approach based on ARIMA modeling, whereas Brutlag [14] and Leikis [67] employed Holt-Winters smoothing to monitor IT systems.

However at STMicroelectronics, monitoring based on the usual Holt-Winters algorithm revealed 2 classes of limitations:

1. **Sensitivity to outliers:** outlying observations are used to causing severe bias in the two phases of the monitoring procedure. Many solutions were suggested to overcome these difficulties. First of all, robust estimators (see e.g. Rousseeuw *et al.* [93, 96, 94, 95] , Huber *et al.* [55, 53] , Hubert [54]) can be used to make control limits resistant to outliers. Along the same lines, different methods are offered to limit the influence of these outliers during the time series modelling phase. In case of deterministic trend, Gardner [61] presented a pre-cleaning method where outliers are detected via a 2-sigma detection rule and replaced by averages. Cipra *et al.*[17] developed robust Kalman filters that can also serve to perform Holt-Winters smoothing. More recently Gelper

et al. [39] and Croux *et al.* [21] introduced a robust version of the Holt-Winters smoothing with enough recommendation on its implementation.

2. **Structural changes due to external decisions:** They can lead to poor behavior of the robust Holt-Winters algorithm proposed by References [39, 21] . For instance, after a shift in the time series, this robust Holt-Winters algorithm may furnish worse results than its original version. Many authors have already focused on this question of change point analysis also known as structural change monitoring. The problem consists in testing at each stage if there is a change or not and when it really occurred (Woodall and Montgomery [112]). Zeileis[117] drew up the inventory of statistical tests to monitor changes in linear regressions models. But more generally, and beyond linear regressions models, the existing methods for change point detection can be grouped in two major families: frequentists and Bayesians.

   (a) The very first frequentist solutions have been proposed since the fifties by Page [84, 85, 86] and deepened over years up to more recent formulations using support vector machine (see e.g. Desobry[25])

   (b) Bayesian formulations are also endowed with a significant history that began with Chernoff and Zacks[15]. They keep growing and represent a dynamic and current field of statistics and probabilities (see Zacks and Kenett[114] , Colosimo[18] , Erdman and Emerson[33] for more details).

10.1.2 Contributions and contents of this paper

The originality of this paper consists of tackling simultaneously the two kinds of problems mentioned earlier that weaken the Holt-Winters algorithm. For that, the robust approach proposed by References [39, 21] is tested and we show in an industrial background how it allows solving the first class of issues. Afterwards, we develop a new methodology to adapt this robust smoothing so that its responds to the two classes of problems aforementioned. This methodology is dynamic detecting structural breaks in a time series and reinitializing the robust scheme of References [39, 21] accordingly. We prove that the new scheme improves the flexibility of References [39, 21] 's algorithm by solving the coupled problem of outliers and structural changes.

Finally, it should be stated that an industrial monitoring implies a simultaneous control of several dozens of variables. There are numerous approaches in multivariate process control (see e.g. Bersimis *et al.* [9] for a review), and a specific investigation of robust techniques in this context is beyond the scope of the paper. Alternatively, we show how the robust controlling procedure can be adapted to the framework of multivariate process control.

This paper is organized as follows. In Section 2, we introduce the main statistical methods that have been deployed to control the STMicroelectronics IT system: The robust Holt-Winters monitoring proposed by Gelper *et al.* [39] and Croux *et al.* [21] ; The dynamical contribution developed through our research, improving its flexibility in changing environments. In Section 3, some performance tests are introduced. They are grounded on quantitative simulations to: 1) Compare the usual and robust Holt-Winters monitoring procedures; 2) Evaluate the structural break detection capacity of our new adaptive procedure. Lastly,

Section 4 introduces some examples from real industrial cases. In addition, univariate and multivariate applications are considered.

10.2 Investigation of robust monitoring for trended time series with structural changes

The main principles of Holt-Winters based monitoring (HW), and its robust version (RHW) introduced by References [39, 21] will be presented. Then, a new methodology (RHW-SC) in presence of structural changes is furnished.

10.2.1 Monitoring based on Holt-Winters smoothing

The Holt-Winters (HW) algorithm is a popular technique used to provide short-term forecasts of a given time-series (see e.g. Makridakis *et al.* [71]). The predictions are built iteratively as a linear combination of the observed values and the prediction obtained at last step.

For illustration, let us consider a time series y, observed at dates $1, 2, \ldots, n-1$. For the sake of simplicity, we assume that y is non-seasonal, though the methodology is similar. The HW algorithm[52, 110] is based on the assumption that y is a sum of two time-series α and β corresponding respectively to a local level (order of magnitude) and a trend. These auxiliary time series are estimated iteratively as averages of the last observation and the last predictions, weighted by two parameters λ_1 and λ_2:

$$\hat{\alpha}_t = \lambda_1 y_t + (1-\lambda_1)\hat{y}_{t|t-1}, \qquad t = 1, \ldots, n-1 \tag{10.1}$$

$$\hat{\beta}_t = \lambda_2(\hat{\alpha}_t - \hat{\alpha}_{t-1}) + (1-\lambda_2)\hat{\beta}_{t-1}, \qquad t = 1, \ldots, n-1 \tag{10.2}$$

Logically, the one-step-ahead forecast done at date $t-1$ for date t is then given by:

$$\hat{y}_{t|t-1} = \hat{\alpha}_{t-1} + \hat{\beta}_{t-1}, \qquad t = 1, \ldots, n \tag{10.3}$$

which gives in particular the prediction at date n. In practice, λ_1 and λ_2 are estimated by minimizing a criterion (often the least-square criterion) based on the forecast errors:

$$E_t = y_t - \hat{y}_{t|t-1}, \qquad t = 1, \ldots, n-1 \tag{10.4}$$

and the algorithm is initialized by a linear regression on the first m values.

As a second step, monitoring can be performed (Montgomery[76]). While applying a control chart to y is not recommended, due to the violations of the usual assumption *identically and independently distributed data* especially if y is trended, the forecast errors E_t may be close to satisfy it (Box and Jenkins [12, 13]). Then, assuming furthermore that E_t are normally distributed $N(0, S^2)$ the upper and lower control limits for E_t are then given by:

$$UCL = +q_{\alpha/2} * \hat{S} \tag{10.5}$$

$$LCL = -q_{\alpha/2} * \hat{S} \tag{10.6}$$

where $q_{\alpha/2}$ is the quantile of a Student distribution at level $\alpha/2$, and $\hat{S}^2$ is the usual variance estimator:

$$\hat{S}^2 = \frac{1}{n-m-1} \sum_{t=m+1}^{n-1} E_t^2 \tag{10.7}$$

These limits intend to detect the dates that correspond to an anomaly: A value of E_t outside the interval [UCL, LCL] should be a strong indication of an abnormal behavior (for a given confidence level α). However, the limits themselves are sensitive to outliers, since the variance estimator overestimates the true variance in presence of outliers. Furthermore, the predicted value $\hat{y}_{t|t-1}$ depends linearly on past values that may contain outliers. These problems are solved by the robust version of the Holt-Winters monitoring.

10.2.2 Robust monitoring based on Holt-Winters smoothing

The Robust Holt-Winters algorithm (RHW) introduced by Gelper *et al.* [39] and Croux *et al.* [21] considers two additional auxiliary time series: y^*, representing a *cleaned* proxy of y after outliers treatment, and σ the expected prediction error, representing a robust estimate of the forecast error E_t. To obtain a robust algorithm, large values are truncated when larger than a given threshold. More precisely, the expected errors σ_t are computed recursively by:

$$\hat{\sigma}_t^2 = \lambda_\sigma \left[\psi_k \left(\frac{E_t}{\hat{\sigma}_{t-1}} \right) \right]^2 \hat{\sigma}_{t-1}^2 + (1-\lambda_\sigma)\hat{\sigma}_{t-1}^2 \tag{10.8}$$

where λ_σ is a given weight, and ψ_k is the Huber function with boundary value k:

$$\psi_k(x) = \begin{cases} x & \text{if } |x| \le k, \\ \text{sign}(x) \times k & \text{if } |x| > k \end{cases} \tag{10.9}$$

The error E_t is still given by $E_t = y_t - \hat{y}_{t|t-1}$, with $\hat{y}_{t|t-1} = \hat{\alpha}_{t-1} + \hat{\beta}_{t-1}$, but the local level and trend are now estimated by using the cleaned time series y^*:

$$\hat{\alpha}_t = \lambda_1 y_t^* + (1-\lambda_1)\hat{y}_{t|t-1} \tag{10.10}$$

$$\hat{\beta}_t = \lambda_2(\hat{\alpha}_t - \hat{\alpha}_{t-1}) + (1-\lambda_2)\hat{\beta}_{t-1} \tag{10.11}$$

where y_t^* is given by:

$$y_t^* = \psi_k \left(\frac{E_t}{\hat{\sigma}_t} \right) \times \hat{\sigma}_t + \hat{y}_{t|t-1} \tag{10.12}$$

Notice that the role of the Huber function ψ is to truncate the forecast errors E_t when larger than k times the expected prediction error $\hat{\sigma}_t$:

$$\psi_k \left(\frac{E_t}{\hat{\sigma}_t} \right) \times \hat{\sigma}_t = \begin{cases} E_t & \text{if } |E_t| < k\hat{\sigma}_t, \\ \text{sign}(E_t) \times k\hat{\sigma}_t & \text{if } |E_t| > k\hat{\sigma}_t \end{cases} \tag{10.13}$$

For instance, Equation (10.12) can be rewritten in a simpler way:

$$y_t^* = \begin{cases} y_t & \text{if } |E_t| < k\hat{\sigma}_t, \\ \text{sign}(E_t) \times k\hat{\sigma}_t + \hat{y}_{t|t-1} & \text{if } |E_t| > k\hat{\sigma}_t \end{cases} \tag{10.14}$$

Finally the parameters λ_1 and λ_2, and the standard deviation S of the forecast errors are computed with robust procedures (References [39, 21]):

$$(\lambda_1^{\text{opt}}, \lambda_2^{\text{opt}}) = \underset{\lambda_1,\lambda_2}{\text{argmin}} \left\{ S_0^2 \sum_{t=m+1}^{n-1} \left[\psi_k \left(\frac{E_t}{S_0} \right) \right]^2 \right\} \tag{10.15}$$

$$\hat{S}^2 = C_k \frac{S_0^2}{n-m-1} \sum_{t=m+1}^{n-1} \left[\psi_k \left(\frac{E_t}{S_0} \right) \right]^2 \tag{10.16}$$

Here, $S_0 = \text{med}_{m+1 \leq t \leq n} |E_t|$ and C_k is a consistency factor. For the common choice $k = 2$ and $\lambda_\sigma = 0.3$, we have $C_k \approx 1.404$ (see References [39, 21] for more details).

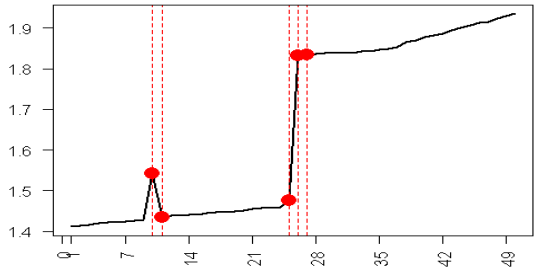
The algorithm initialization is also done robustly by repeated median regression. For that, we employed a period of length 7: This is short enough to assume a local linear trend and long enough to be resistant to 2 outliers.

Since the RHW smoothing is fully robust, the control charts based on the errors E_t (see Section 10.2.1) should now be resistant to outliers, which is a clear improvement to the (non-robust) HW-based monitoring. However, adaptation is necessary in presence of a structural change.

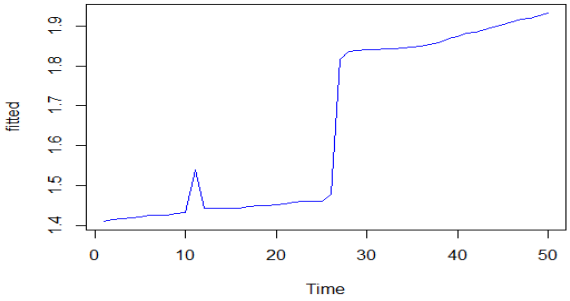
In this section, the focus is on time series that possibly contain outliers and structural changes. We consider frequentist approaches to detect the latter. To detect structural changes in parametric models, three main classes of methods exist (Zeileis [116]): F statistics, fluctuation tests and maximum likelihood scores. We choose to use a common and simple F statistics, the Chow test, since it can be easily adapted to a robust framework by using a robust regression and robust F statistics. The Chow test splits the sample into 2 groups: The first before the break date and the second after. The model parameters are estimated for both of them so that an F test be performed to judge whether they are equal or not. The Chow test is easy to use but restricted by 2 limitations. The first, mentioned by Hansen [47] , is that the break date must be known a priori. Moreover, the exact number of changes is unknown. The second problem is a question of robustness: A break may be missed or falsely detected because of outliers. In this section, we show how RHW smoothing can deal with these problems and deduce a strategy for structural change monitoring.

An introducing example Consider the time series in Figure 1 with an outlier at date 10 and a break at date 26. As expected, the robust algorithm is not sensitive to the outlier at date 10 contrarily to classic smoothing: This is its main advantage. But after that, the level changes suddenly at date 26. The robust algorithm does not admit this modification quickly and many false alarms are generated. Nevertheless, this specificity can be used to detect break dates. Indeed, when a structural change happens, there is a quite long period (here 26—34) of successive false alarms corresponding to successive large errors in the RHW smoothing. During this period, the predicted values $\hat{y}_{t|t-1}$ seem to exhibit a deterministic pattern (Figure 1, d): Actually, we show below (Figure 2 and Proposition 1) that they match exactly with an analytical function increasing exponentially. These two facts strongly suggest that the periods of successive large errors given by robust Holt-Winters smoothing are useful for the detection of break dates.

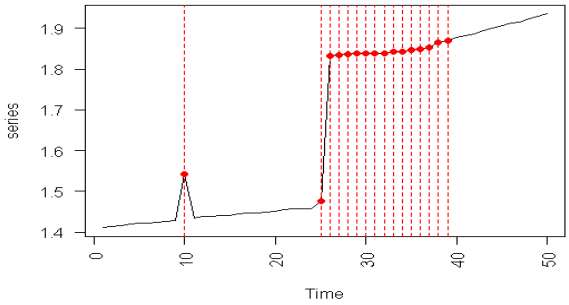
When a structural break occurs in practice, there is a sequence of consecutive false alarms due to a succession of large errors. RHW smoothing enables to quantify the importance of



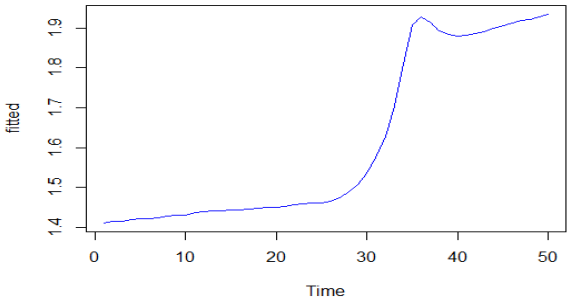
(a) Alarms given by the classical HW forecasting



(b) Predicted values by the classical HW.



(c) Alarms given by the robust HW forecasting



(d) Predicted values by the robust HW.

Figure 10.1 – A real time series with outlier and structural change

these errors by comparing them to their predicted values. Let us call *relative error* the ratio $\frac{E_t}{\hat{\sigma}_t}$. A succession of large values of this ratio is a forewarning sign of structural change. More formally, we call t_1 *a suspicious date* for structural change if there exists an integer $p \geq 3$ such that:

$$\psi_k\left(\frac{E_{t_1}}{\hat{\sigma}_{t_1}}\right) = \psi_k\left(\frac{E_{t_1+1}}{\hat{\sigma}_{t_1+1}}\right) = \cdots = \psi_k\left(\frac{E_{t_1+p-1}}{\hat{\sigma}_{t_1+p-1}}\right) = k \tag{10.17}$$

or

$$\psi_k\left(\frac{E_{t_1}}{\hat{\sigma}_{t_1}}\right) = \psi_k\left(\frac{E_{t_1+1}}{\hat{\sigma}_{t_1+1}}\right) = \cdots = \psi_k\left(\frac{E_{t_1+p-1}}{\hat{\sigma}_{t_1+p-1}}\right) = -k \tag{10.18}$$

The period $[t_1, \ldots, t_1 + p - 1]$ is a *suspicious period*: a period when forecasting errors remain k times higher than their expected values.

Proposition 1 During a suspicious period $[t_1, \ldots, t_1 + p - 1]$, the predictions $\hat{y}_{t|t-1}$ of the robust HW smoothing are given by a deterministic and monotonic function f which does not depend on any observation posterior to date t_1. Its form is:

$$f(t) = \frac{r^{(t+1)}}{(r-1)^2} + a.t + b \tag{10.19}$$

with

$$r = \sqrt{\lambda_\sigma(k^2 - 1) + 1} = 1.378405 \tag{10.20}$$

and where a and b are given by $\hat{y}_{t_1}$ and $\hat{y}_{t_1-1}$

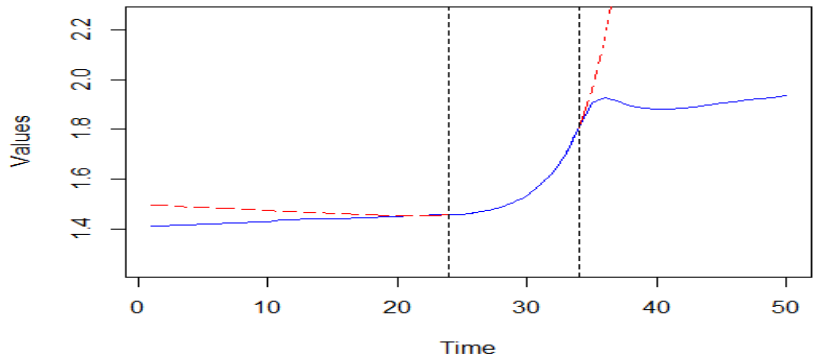


Figure 10.2 – Zoom on Figure 1 (d): The predicted values (solid line) coincide with a function f of the form $f(t) = ar^t + bt + c$ (dotted lines) during the period [26-34].

Proof See Appendix 1.

A new methodology for structural change monitoring

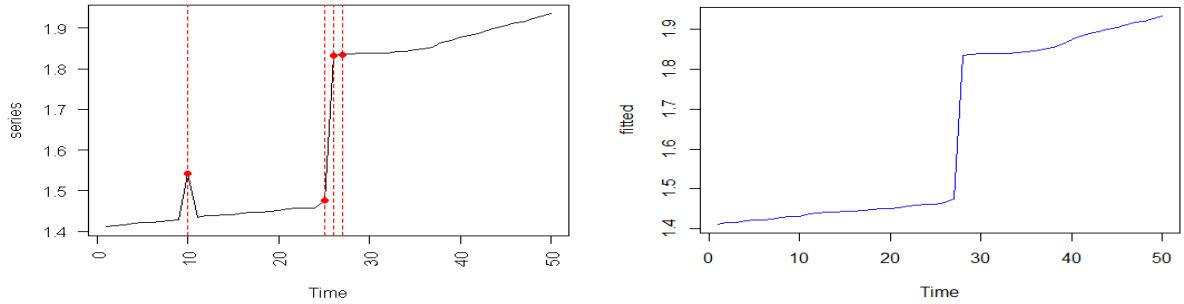
The previous observations and results about RHW smoothing in case of a structural change suggest the following methodology called RHW-SC:

1. Find the suspicious dates by looking at consecutive relative errors given by the robust HW smoothing (Equations 10.17 and 10.18)

2. Apply a robust version of the Chow test to the suspicious dates detected in 1. One robust version of the test consists in replacing the usual linear regression by the repeated median regression (Siegel [104] , Rousseeuw *et al.* [96]) and using a robust estimator of the residuals' variance (Croux *et al.* [21]) to compute the F statistic.

This strategy tackles the two main issues mentioned at the beginning of this section: All the possible break dates (number and locations) are automatically detected by the algorithm itself, and for a given date the statistical test for structural change is done in a robust way. In practice the methodology is applied dynamically, and when the robust Chow test is positive, the robust HW smoothing is reinitialized.

Performance of the RHW-SC methodology on the introducing example To solve the problem raised by the introducing example, the RHW-SC methodology is performed dynamically with $p = 3$. The methodology detected one suspicious date, namely day 26, which indeed corresponds to the structural change. Notice that only one Chow test has been used contrarily to exhaustive methods such as Quandt-Anderson that systematically test all dates. We reinitialized the robust HW smoothing at date 26, by using the repeated median regression coefficients. As a result, RHW-SC methodology remains resistant to the outlier at date 10 without generating a long sequence of false alarms after the break.



(a) Alarms given by RHW-SC.

(b) Predicted values by RHW-SC.

Figure 10.3 – Alarms and predicted values by the robust version with structural change detection

10.3 Global performance tests

Some ARIMA models correspond to exponential smoothing methods (see e.g. Hyndman *et al.* [56]). In particular, a Holt-Winters forecasting with smoothing parameters (λ_1 , λ_2) is equivalent to an ARIMA (0, 2, 2) model of parameters (θ_1 , θ_2) if: $-1 \leq \theta_2 \leq 1$, $\theta_2 - \theta_1 \leq 1$ and $\theta_2 + \theta_1 \leq 1$, with :

$$\lambda_1 = 1 + \theta_1 \tag{10.21}$$

$$\lambda_2 = \frac{1 - \theta_1 - \theta_2}{1 + \theta_2} \tag{10.22}$$

Thus, we first use ARIMA (0, 1, 1) time series with parameter $\theta = 0.5$. Indeed, they correspond to the special case where $\theta_2 = 0$. Next, we perform our simulations on ARIMA (0, 2, 2) models with (θ_1 , θ_2) = (1 , -0.25). These parameters have been estimated from an industrial time series by assuming that it comes from an ARIMA(0, 2, 2) model. In this section, we compare first the performances of the RHW and the HW smoothing statically and dynamically. Afterwards, we evaluate the ability of the RHW-SC method to detect a structural change.

10.3.1 Comparison tests in a static setting

We are interested in comparing 2 characteristics of the HW and the RHW methods:

1. The *power*: The probability that an outlying observation is detected
2. The *false detection rate*: The probability that a normal observation is detected. This risk is called *size* of the control chart when there is no outlier. Below, we use 0.95 as a confidence level, so the false detection rate is expected to be 0.05.

Simulation results for ARIMA (0, 1, 1) time series

We use the following strategy:

1. Generation of a time series: simulate an ARIMA (0, 1, 1) time series of length 160 with parameters $\theta = 0.5$. The first 60 values serve as a training sample (including $m = 7$ values for HW and RHW initializing).

2. Generation of outliers: For a fixed contamination rate R, choose randomly $160R$ dates among the 160 dates. Contaminate these observations by adding or subtracting to them a value e. Whether it is an addition or a subtraction is chosen at random. Two cases are considered for e:

   (a) e is a fixed value among 10, 20 or 30

   (b) "Mix": e is chosen at random uniformly among 10, 20, 30

3. Perform the HW and the RHW methods and estimate:

   (a) Their *power*: as the percentage of the contaminated observations that are really detected

   (b) Their *false detection rate*: as the percentage of non contaminated observations that are detected.

4. Repeat 100 times the steps 1 to 3.

The RHW and HW monitoring are equivalent when there is no outlier: same power and same false alarm rate (Reference [21]). Their differences become significant when the data gets contaminated (see below).

On the one hand, Table 10.1 and Table 10.2 below show that the HW control chart is subject to two effects. The first one is the widening of its control limits by outliers, which tends to reduce abnormally its false alarms rate. The second one concerns the dates following these outliers: the corresponding predictions are biased and this fact tends to raise the false alarms rate. Here, the first effect is predominant. This explains the too low false alarms risk and the poor detection rate of the HW method.

On the other hand, Table 10.3 and Table 10.4 show that the RHW methodology outperforms the HW smoothing especially when the number of outliers raises and when the magnitude of these outliers is not fixed: which is a realistic case. In fact, the RHW control chart does not lose either its power, or the stability of its false alarms rate that remains stable around the theoretical value 5%.

Table 10.1 – HW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	93.5	70.0	31.0
e = 20	94	74	35
e = 30	91	72	36
Mix	72	55	33

Table 10.2 – HW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	3.3	3.6	2.1
e = 20	1.7	3.0	7.5
e = 30	1.9	6.0	10
Mix	3	5.3	4

Table 10.3 – RHW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	100	100	99.2
e = 20	100	100	99.9
e = 30	100	100	100
Mix	100	100	99.8

Table 10.4 – RHW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	5.2	4.8	6.1
e = 20	5.2	5.2	6.4
e = 30	5.2	4.8	7.1
Mix	5.1	4.9	6.1

Simulation results for ARIMA (0, 2, 2) time series

The same strategy as in the previous section is used. Only the first step is modified to generate ARIMA (0, 2, 2) models with $(\theta_1 , \theta_2) = (1 , -0.25)$. The results are summarized below (Tables 10.5, 10.6, 10.7 and 10.8). They show that the RHW control chart remains better in term of detection. Nevertheless, its false risk becomes higher than expected even if it remains stable enough for a fixed contamination rate contrarily to HW smoothing.

Table 10.5 – HW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	97.6	91.3	67.0
e = 20	100	92.2	74.4
e = 30	97.1	91.0	73.3
Mix	92.7	77.3	57.5

Table 10.6 – HW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	3.3	2.7	3.7
e = 20	2.1	4.1	8.0
e = 30	5.2	1.5	4.0
Mix	2.2	3.9	6.4

Table 10.7 – RHW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	99.8	99.0	83.0
e = 20	100	99.9	93.5
e = 30	100	99.9	95.0
Mix	99.7	99.3	91.5

Table 10.8 – RHW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	6.4	7.2	9.0
e = 20	7.1	7.1	8.5
e = 30	6.2	7.2	9.7
Mix	7.0	7.4	9.0

10.3.2 Comparison tests in a dynamical setting for ARIMA (0, 2, 2) time series

In practice, industrial variables are tracked daily; the smoothing parameters and control limits are re-estimated every day. So, dynamic simulations were performed. They use the same procedure as for the static setting but with updating the smoothing parameters and controls limits at each iteration.

Tables 10.11 and 10.12 show how suitable HW monitoring is to update its smoothing parameters and control limits contrarily to the RHW method. Indeed, the HW detection rate is improved even if the power of the RHW smoothing remains better. Let us remark that the HW false alarm risk has raised and approaches better the theoretical value 5%. As for the RHW false alarms risk, it has decreased to approach this same theoretical value; but this later improvement is less obvious. However, the RHW false alarms risk has become very stable and non dependent on the magnitude of outliers, which is not the case for the HW

Table 10.9 – HW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	99.6	95.4	80
e = 20	100	93.6	80
e = 30	100	97	78.6
Mix	95	70	52

Table 10.10 – HW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	4.6	4.3	4.2
e = 20	3.3	2.2	2
e = 30	2.5	2.2	2.4
Mix	3.0	3.2	2.0

Table 10.11 – RHW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	100	100	99.7
e = 20	100	100	100
e = 30	100	100	100
Mix	100	100	100

Table 10.12 – RHW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	6.3	6.8	8.6
e = 20	6.0	7.1	9.1
e = 30	6.2	6.9	9.3
Mix	6.1	7.0	7.2

control chart. Finally, the RHW control chart outperforms when the scale of the outliers is more realistically (randomly) chosen.

Comparison using ROC analysis According to the analysis here-above, the RHW smoothing presents higher detection rates, but also higher false alarm rates. To have a clear view of the benefit of the robust procedure, we use a receiver operating characteristics (ROC) graph. Indeed, the final goal is to classify observations in two groups: normal or abnormal. So, for an ARIMA (0, 2, 2) time series randomly contaminated as previously ($e =$ "Mix" and $R = 5\%$), we perform the RHW and the HW smoothing. Instead of fixing the confidence level, it is varied with the aim of plotting the corresponding false alarm rates versus the detection rates for the two procedures (Fawcett[34]).

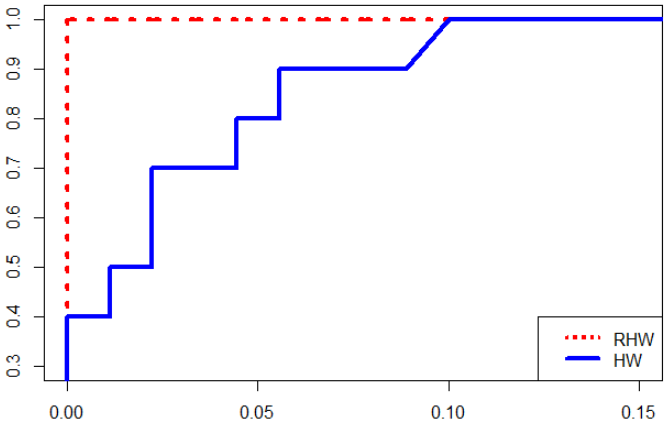


Figure 10.4 – ROC curves for one contaminated ARIMA(0, 2, 2) time series:
Detection rate (y-axis) vs. False alarm rate (x-axis)

These curves show clearly that the RHW smoothing is the best choice. For instance, a user who tolerates 5% of false alarms can still reach more than 99% of detection rate with

the RWH smoothing instead of only 80% for the classic version. Furthermore, the results remain similar when the simulation is repeated, but this is not the focus here. Let us remark that these results are consistent to the values in Tables 10.9, 10.10, 10.11 and 10.12.

10.3.3 Structural change detection performance tests

To evaluate the RHW-SC methodology, the following strategy was used:

1. Generation of a time series: simulate an ARIMA (0, 2, 2) time series of length 160 with parameters $(\theta_1\ ,\ \theta_2) = (1, -0.25)$. The first 60 values serve as a training sample (including $m = 7$ values for RHW initializing).

2. Generation of outliers: For a fixed contamination rate $R = 0.05$, choose randomly $160R$ dates among the 160 dates. Contaminate these observations by adding or subtracting to them a fixed value $e = 5$. Whether it is an addition or a subtraction is chosen is random.

3. Generation of structural changes: Choose randomly 1 date among the 100 last values. From this date to the last one, add a linear function: $A * DATE + B$ to the time series.

4. Perform the RHW-SC methods with 0.95 as confidence level for the robust Chow test and compute:

   (a) The detection rate for structural changes,

   (b) The false detection rate for structural changes, corresponding here to the probability of detecting a change at a wrong date.

5. Repeat 1000 times the steps 1 to 4.

Table 10.13 – Detection rates of changes

	B = 0	B = ±50
A = 0	-	95
A = ±50	86	95

Table 10.14 – False detection rates of changes

	B = 0	B = ±50
A = 0	-	0.6
A = ±50	0.4	0.5

The results show that the RHW-SC methodology leads to a very low risk of false alarms. Indeed, the probability of occurrence of a suspicious period with length $p \geq 3$ at another date than the break is extremely small. Since these suspicious periods represent a necessary condition for change detection, therefore Chow-tests are performed very rarely: ergo fewer alarms.

In addition, the RHW-SC methodology is efficient for structural change detection, especially when a shift happens. When there is no shift (only a slope change), the detection becomes less efficient. Nevertheless, in that case, RHW period of successive false alarms is short and there is no need to reinitialize.

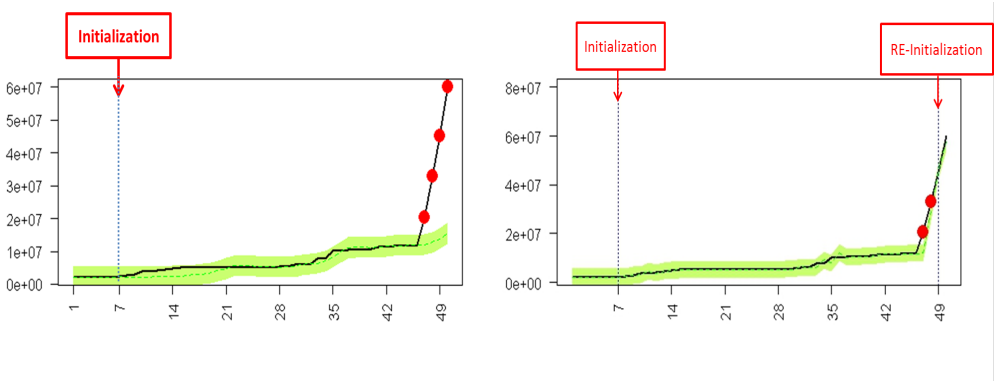
We are now interested in another performance characteristic: the Conditional Expected Delay (CED), see Kenett *et al.* [64] . The CED is the time between the first opportunity to detect a change and the first true alarm related to this change. Here, the RHW-SC detections usually occur at the third post-change observation. According to the simulations above, the Conditional Expected Delay is estimated to be $p = 3$.

10.4 Applications

The RHW and the RHW-SC methods are performed daily on many indicators of the information system of a company. In general, the results are satisfactory. Here, we present some of these examples and a multivariate case.

10.4.1 Examples of univariate time series

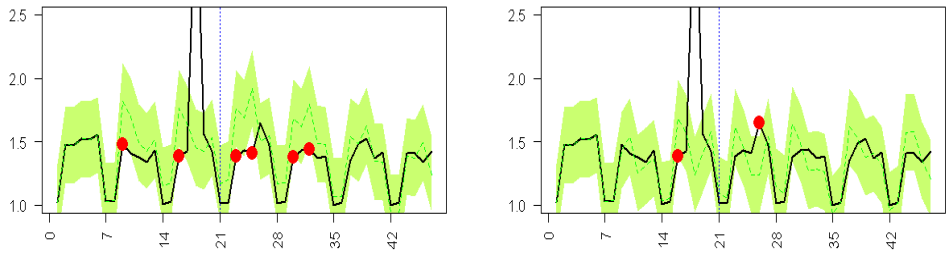
This first example is illustrative of structural changes detection by the RHW-SC methodology.



(a) RHW (b) RHW-SC

Figure 10.5 – An industrial time series with a structural change

The variable of Figure 10.5 is an indicator of CPU consumption. It is subject to one slope change at date 47. This leads to four false alarms when the RHW methodology is performed (10.5a). The change is detected 3 days later by the RHW-SC monitoring, resulting in a reinitialization and reduction of the number of false alarms.
The second example shows an extension of the RHW smoothing to seasonal time series (Reference [39]). This differentiates the normal operating conditions (5 weekdays activity and drop of activity on weekends) from exceptional events.



(a) Seasonal HW (b) Seasonal RHW

Figure 10.6 – An industrial and seasonal time series

Figure 10.6 points out the sensitivity of the HW method to outliers. The abnormal observation of the third week has deformed the seasonal component. Because of the over-estimation of this seasonality, the 26th observation that is really aberrant is not detected. Moreover, this leads to four false alarms the two following weeks. These problems no longer exist with the seasonal RHW smoothing.

10.4.2 The multivariate case

In our company STMicroelectronics, several indicators are tracked daily. So, monitoring them separately leads to false alarms every day. This is foreseeable given that the theoretical false alarms risk tends to 100% when the number of independent variables approaches 100: Hence, the necessity to perform a multivariate monitoring. Notice that there are several approaches in the multivariate framework (see e.g. Bersimis *et al.* [9]), and our aim is neither to do a comparative study nor to propose a best one. Rather, we show how the new robust methodology can be adapted to the multivariate case. Thus, as an example, we constitute groups of variables. For each group of p variables, the following strategy is used:

1. Perform the RHW smoothing for each variable of the group.
2. Use a robust Hotelling T2 control chart to analyze simultaneously the p vectors of residuals given by step 1.

Among existing robust Hotelling T2 control charts (see e.g. Rousseeuw [93], Alfaro *et al.* [2]) we consider here the computation of the confidence ellipsoid with the Minimum Covariance Determinant criterion introduced by Rousseuw *et al.* [93, 95] . There are two identified difficulties: A poor orientation of the confidence ellipsoid and an underestimation of its size. To face these problems, recent solutions found in the literature were employed. Firstly, the orientation is improved by choosing a subset of the 75 % *best points* for the MCD criterion instead of 50% (Huber *et al.* [55]). Secondly, the size estimation is improved by using two correcting factors: One asymptotic factor of consistency to the chi square distribution (Rousseeuw *et al.* [96]) and one empirical result for small samples (Pison *et al.* [88]).

Now, we present an example with a group of $p = 4$ variables. BIN1 is related to the user activity (software transactions activities) whereas BIN2, BIN3 and BIN4 concern the volume and activity of a data base (Oracle statistics as DBtime, Redo Size, Session logical Read: see the Oracle data base documentation for more details). Usually, the data base activity follows the user's behavior. This explains the positive correlation among the variables.

Each day, the Hotelling T2 statistic is computed for this group. The result is plotted in Figure 10.7. This control chart has successfully detected the 6 most important outliers but does not say where the problems come from. This is the well known problem of multidimensional *out of controls* interpretation. As an example, let us focus on the alarm on date 56. Its cause can be known by looking at each variable separately. Then, Figure 10.8b shows that the variable BIN2 is mainly responsible.

Nevertheless, this solution is not realistic for an industrial use because it produces too many graphs that need to be examined. Among the numerous existing criteria to interpret multivariate signals, the partial relative contributions mentioned by Montgomery [76] are very popular for their efficiency. In Figure 10.9, these relative contributions at date 56 confirm the influence of BIN2.

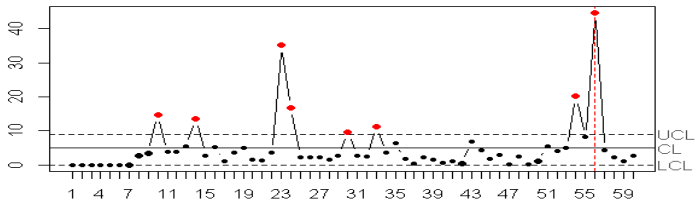
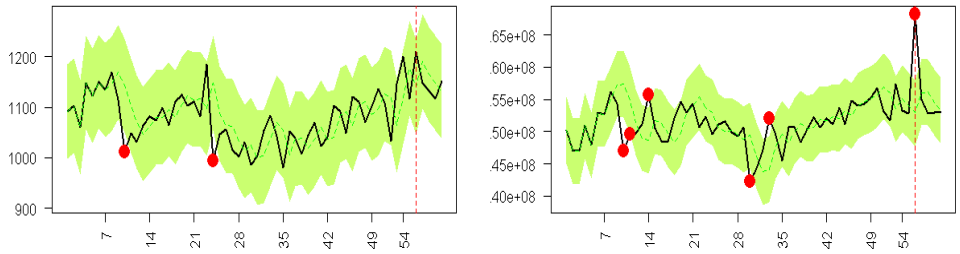
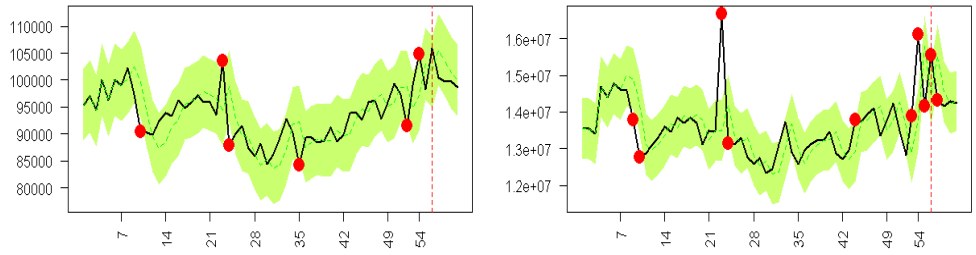


Figure 10.7 – Hotelling daily statistics for a group of four variables



(a) Variable BIN1 (b) Variable BIN2



(c) Variable BIN3 (d) Variable BIN4

Figure 10.8 – RHW smoothing for the four variables of the group

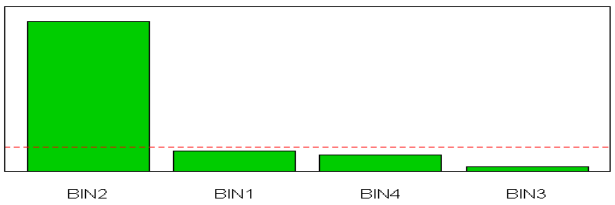


Figure 10.9 – Partial contributions to variability at date 56

10.5 Discussion

In this paper, a monitoring based on robust Holt-Winters smoothing, as proposed by Gelper *et al.* [39] and Croux *et al.* [21] was studied. Based on an industrial application of this

method at STMicroelectronics and simulation studies, its high robustness was confirmed. However, as poor results were noticed in case of structural changes, we have also proposed and evaluated an improved dynamical approach, for a better integration of changing environments. The efficiency of this robust and dynamical method has been demonstrated on real univariate and multivariate STMicroelectronics case studies. This contribution has been developed as an improvement for the STMicroelectronics IT system monitoring, but it could be applied to many other industrial applications, where time dependent variables has to be statistically controlled.

For further research, there are several interesting outstanding questions. Firstly, fine tuning could be investigated. Hence, the truncation parameter was set to $k = 2$ and the smoothing parameter λ_σ to 0.3 (still proposed by References [39, 21]). We also recommended to use $m = 7$ for initialization to be resistant to 2 outliers. These values could be further evaluated: Do they always provide optimal results or should they be contextually adapted? Secondly, when structural changes occur, we decided to reinitialize the Holt-Winters parameters. One alternative could be to use dynamical smoothing parameters that may change over time, as proposed by Williams [109] or Taylor [108] . Finally, in our methodology, we chose a Chow test to detect structural breakpoints, since it is easily adapted to the robust framework. Nonetheless, as mentioned earlier, other approaches do exist (Bayesian formulations, other F statistics, fluctuation tests, maximum likelihood scores). Their potential of application in our monitoring procedure and a comparative study or their performance may lead to interesting further insights.

Résumé en Français

L'évaluation de la performance industrielle passe par le suivi statistique d'indicateurs de qualité. Le travail présenté dans ce chapitre s'inscrit dans la continuité de ceux de Lutz [69], Gelper [39] et Croux [21]. Pour assurer la maitrise statistique des ressources en système d'information, Lutz avait couplé un lissage exponentiel de Holt-Winters aux cartes de contrôle usuelles. En contexte industriel, l'algorithme engendre de fausses alarmes, dues à des observations atypiques. Le problème est alors traité par Gelper et Croux qui ont développé une version robuste du lissage de Holt-Winters et de sa carte de contrôle. Le modèle robuste répond alors au problème des données atypiques mais détériore les performances initiales en cas de changements structurels. Pour traiter simultanément le problème de données atypiques et celui des changements structurels, nous avons proposé une version à la fois robuste et adaptative du lissage exponentiel de Holt-Winters. L'algorithme embarque une détection de changements brusques dans le lissage robuste qu'il réinitialise dans le cas échéant. Théoriquement justifiée par l'équivalence entre modèles ARIMA et lissage de Holt-Winters, cette méthodologie bénéficie des avantages des deux versions précédentes.

Conclusion and outlook

Driven by industrial needs in microelectronics, this thesis resulted in novel contributions in spatial statistics, designs of experiments, statistical process control and time-series modelling. The major outcomes in applied mathematics are generic and meet other applications in computer experiments, environmental engineering, IT system monitoring. We now suggest possible avenues of research related to our contributions.

Spatial statistics. The problem of spatial data had the specificity of being defined on circular domains. It was addressed through Universal Kriging, a model involving a deterministic trend and a stochastic term represented by a Gaussian process. In traditional Kriging, the geometry of the disk is included neither in the trend, nor in the Gaussian process part. In continuity with the work of [89] who used Zernike polynomials to model circular features in the trend, we introduced polar Gaussian processes to embed the geometry of the disk in the stochastic part. Defined in the space of polar coordinates by mapping the disk to a cylinder, polar Gaussian processes improve Kriging estimation when radial and angular components are predominant since the reconstruction is made on radial and angular correlations. Then, we extended polar Gaussian processes to hyperballs, corresponding to a higher dimension directional input in computer experiments. We also investigated the deterministic part, namely Zernike polynomials. Since these functions are orthogonal with respect to the uniform measure over the disk, their properties are not fully exploited when the model is defined in polar coordinates. Given this, we conducted a study on measure modification. As a result, we defined a new set of functions that are orthogonal with respect to the uniform measure over the space of polar coordinates, and that have the same symmetry and rotation properties than Zernike polynomials. As a visualization and interpretation tool, a sensitivity analysis based on centred polar Gaussian processes is conducted. The Sobol decomposition of Gaussian processes, proposed by [41] and [31] in hypercubic domains, was then extended to periodic functions. In addition to recovering additive and tensor-product kernels, the resulting Gaussian processes allow a direct comparison of Cartesian and polar Gaussian processes, based on their likelihoods.

The future works on polar Gaussian processes should put the focus on higher dimensions. In particular, anisotropic kernels over hyperspheres represent a relevant question. An idea would consist in using the hyperspherical coordinates system and estimating different covariance parameters for each angular coordinate. The challenge would then be to find meaningful distances and positive-definite functions under the geometric constraints resulting from hyperspherical coordinates. Regarding sensitivity analysis, the Sobol decomposition proposed for polar Gaussian processes relies on the separability of probability measures over the space of polar coordinates. This does clearly not apply to standard Kriging models on the disk,

due to the dependence between x and y coordinates at the boundary the disk. A formal study should include this dependence in integrals terms rather than extending the input domain to a hypercube as done at now.

Designs of Experiments. The contributions of this thesis to designs of experiments are twofold. First, we introduced Latin cylinders in order to reproduce the main properties of Latin hypercubes in the space of polar coordinates. In particular, space-filling is optimized based on the geodesic distance over the cylinder. The resulting class of designs is recommended when the process is inherent in polar coordinates. Moreover, we showed through a simple measure modification, how Latin cylinders can be adapted to fill the disk and suit more general situations. Secondly, when there is some information on the response, we carried out a simulation study of IMSE-optimal designs for standard and polar Gaussian processes. Their key properties such as symmetries and optimality in prediction are investigated in a static setting. For dynamic systems, we tackled the question of IMSE-optimal relocation of a design point. When iterated, the implemented relocation procedure is shown to gradually improve the initial design of experiments until convergence.

D-optimality was already investigated for Zernike polynomials by Dette [26]. For further research, the same criterion should be studied for orthogonal polynomials with respect to the uniform measure over the polar space. One could adapt the demonstration provided in [26] after remarking that the two families of polynomials have the same rotation properties and differ only with respect to their radial terms. Another issue is the computational time of IMSE-optimal designs in the discrete case. An approximation of the IMSE, based on spectral methods for instance [38], and combined with a suitable optimization algorithm may allow to overcome the difficulty. The symmetry properties should also be exploited in order to reduce the dimension of the problem.

Statistical Process Control. In SPC, our contributions lie in the framework of profile monitoring. The specificity of the problem was to monitor curves over time, instead of scalar values. Following the standard procedure presented in [111] and [80], we developed a control chart to monitor the coefficients of Zernike regression. Based on this reference model, we proposed two options for profile monitoring with Gaussian processes. The first one applies the Regression-Adjustment procedure of Hawkins [50] to transformed Kriging parameters. The second one consists in monitoring the process variance via Sobol indices.

Among the perspectives in process monitoring, control charts based on Gaussian processes represent a promising avenue of research. The empirical study that we presented needs to be strengthened through simulations or more theoretical derivations on the distribution of Kriging parameters. Seen as a classification problem, the topic may also be interesting for the machine learning community. In this context, a relevant metric should be investigated in the "non Euclidean space of Kriging parameters", as observed in this thesis.

Our last contribution in SPC consists in using control charts to monitor autocorrelated datasets in the presence of outliers and structural changes at the same time. The robust Holt-Winters smoothing proposed by Croux and Gelper [21] hardly adapts to structural changes by becoming temporarily deterministic. Based on the equivalence between Holt-Winters smoothing and ARIMA models, we provided a characterization of this deterministic behavior with the aim of detecting structural changes. An adaptive monitoring scheme was

then implemented to simultaneously tackle the problems of outliers and structural changes. Further research should extend the method to multivariate datasets in order to align with the latest trends in industry.

Applications in microelectronics. This thesis highlighted two main families of profile patterning: first, radial and angular wafers profiles, and second, variations according to Euclidean directions. To model variations according to Euclidean directions, we realized that an arbitrary choice of the Cartesian basis over the disk is not optimal. As a solution, we proposed a data-driven approach based on the concept of geometric anisotropy. By decomposing and predicting spatial patterns with processing parameters, we also provided a simple tool to interpret and classify wafers profiles as suggested by [89]. The proposed methods are implemented in the R package DiskLearn, in preparation and used to display most of the graphical outputs in this thesis. The collection of these implementations are successfully tested in industry with two applications. The first one is dedicated to a daily time-series monitoring, and the second one is used for profile monitoring and spatial uncertainty quantification in real-time. Automatic parametrizations and visualizations are also implemented. As future work, given the high complexity in semiconductor manufacturing, it will be worth investigating multivariate models.

List of Figures

List of Tables

Bibliography

[1] S.A. Abbasi and A. Miller. On proper choice of variability control chart for normal and non-normal processes. *Quality and Reliability Engineering International*, 28(3):279–296, 2012.

[2] J.L. Alfaro and J.F. Ortega. A robust alternative to Hotelling's T2 control chart using trimmed estimators. *Quality and Reliability Engineering International*, 24(5):601–611, 2008.

[3] J.L. Alfaro and J.F. Ortega. A new control chart in contaminated data of t-student distribution for individual observations. *Applied Stochastic Models in Business and Industry*, 29(1):79–91, 2013.

[4] D. Allard, R. Senoussi, and E. Porcu. Anisotropy models for spatial data. *Mathematical Geosciences*, pages 1–24, 2015.

[5] A. Amiri, W. A. Jensen, and R. B. Kazemzadeh. A case study on monitoring polynomial profiles in the automotive industry. *Quality and Reliability Engineering International*, 26(5):509–520, 2010.

[6] I. Andrianakis and P. G. Challenor. The effect of the nugget on gaussian process emulators of computer models. *Comput. Stat. Data Anal.*, 56(12):4215–4228, December 2012.

[7] M. Batton-Hubert, M. Binois, and E. Padonou. Inverse modeling to estimate methane surface emission with optimization and reduced models: application of waste landfill plants. In *13th Annual Conference of the European Network for Business and Industrial*, Ankara, Turkey, 2013.

[8] I. Ben-Gal, A. Shmilovici, and G. Morag. Context-based statistical process control: A monitoring procedure for state-dependent processes. *Technometrics*, 45:293–311, 2003.

[9] S. Bersimis, S. Psarakis, and J. Panaretos. Multivariate statistical process control charts: An overview. *Quality and Reliability Engineering International*, 23(5):517–543, 2007.

[10] J. Blue and A. Chen. Spatial Variance Spectrum Analysis and Its Application to Unsupervised Detection of Systematic Wafer Spatial Variations. *IEEE Transactions on Automation Science and Engineering*, 8(1):56–66, January 2011.

[11] R. Borgoni, L. Radaelli, V. Tritto, and D. Zappa. Optimal reduction of a spatial monitoring grid: Proposals and applications in process control. *Computational Statistics & Data Analysis*, 58:407–419, 2013.

[12] G.E.P. Box and G.M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1976.

[13] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley, 4th edition, 2008.

[14] J.D. Brutlag. Aberrant Behavior Detection in Time Series for Network Monitoring. *Proceedings of the 14th Systems Administration Conference (LISA 2000)*, 2000.

[15] H. Chernoff and S. Zacks. Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time. *The Annals of Mathematical Statistics*, 35:999–1018, 1964.

[16] P. Cicorella. *Surface Reconstruction and Monitoring via Gaussian processes*. PhD thesis, Politecnico di Milano, 2014.

[17] T. Cipra and R. Romera. Kalman filter with outliers and missing observations. *Test*, 6(2):379–395, 1997.

[18] B. M. Colosimo. *Bayesian Control Charts*, volume 1, pages 169–174. John Wiley & Sons, 2008.

[19] B. M. Colosimo, P. Cicorella, and M. Blaco. Monitoring: Spc for cylindrical surfaces via gaussian processes. *Journal of Quality Technology*, 46(2):95–113, 2014.

[20] N. A. C. Cressie. *Statistics for spatial data*. John Wiley and Sons. John Wiley & Sons, New York, 1993.

[21] C. Croux, S. Gelper, and K. Mahieu. Robust control chart for time series data. *Expert Systems with Applications*, 38(11):13810–13815, 2011.

[22] G. Damblin, M. Couplet, and B. Iooss. Numerical studies of space filling designs: optimization of Latin hypercube samples and subprojection properties. *Journal of Simulation*, 7(4):276–289, 2013.

[23] E. Del Castillo, B. M. Colosimo, and S. D. Tajbakhsh. Geodesic Gaussian processes for the parametric reconstruction of a free-form surface. *Technometrics*, 57(1):87–99, 2015.

[24] E. Del Castillo and A. M. Hurwitz. Run-to-Run Process Control: Literature Review and Extensions. *Journal of Quality Technology*, 29(2):184–196, 1997.

[25] F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *Trans. Sig. Proc.*, 53(8):2961–2974, 2005.

[26] H. Dette, V. B. Melas, and A. Pepelyshev. Optimal designs for statistical analysis with Zernike polynomials. *Statistics*, 41(6):453–470, 2007.

[27] Y. Deville, D. Ginsbourger, and N. Roustant, O. Contributors: Durrande. *kergp: Gaussian Process Laboratory*, 2015. R package version 0.1.0.

[28] N. Doganaksoy, F.W. Faltin, and W.T. Tucker. Identification of out of control quality characteristics in a multivariate manufacturing environment. *Communications in Statistics - Theory and Methods*, 20(9):2775–2790, 1991.

[29] V. Dubourg. *Adaptive surrogate models for reliability analysis and reliability-based design optimization*. PhD thesis, Université Blaise Pascal - Clermont II, 2011.

[30] J. Dugmore and S. Lacy. *Capacity management*. British Standards Institution, 2005.

[31] N. Durrande, D. Ginsbourger, O. Roustant, and L. Carraro. ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, 115:57–67, 2013.

[32] B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981.

[33] C. Erdman and J.W. Emerson. bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems. *Journal of Statistical Software*, 23(3):1–13, 2007.

[34] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[35] N.I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1995.

[36] J. Franco, D. Dupuy, O. Roustant, G. Damblin, and B. Iooss. *DiceDesign: Designs of Computer Experiments*, 2014. R package version 1.6.

[37] M. M. Gardner, J. C. Lu, R. S. Gyurcsik, J. J. Wortman, B. E. Hornung, H. H. Heinisch, E. A. Rying, S. Rao, J. C. Davis, and P. K. Mozumder. Equipment fault detection using spatial signatures. *IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part C*, 20:295–304, 1997.

[38] B. Gauthier and L. Pronzato. Spectral approximation of the imse criterion for optimal designs in kernel-based interpolation models. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):805–825, 2014.

[39] S. Gelper, R. Fried, and C. Croux. Robust forecasting with exponential and Holt-Winters smoothing. *Journal of Forecasting*, 29(3):285–300, 2010.

[40] D. Ginsbourger, O. Roustant, and N. Durrande. On degeneracy and invariances of random fields paths with applications in gaussian process modelling. *Journal of Statistical Planning and Inference*, 170:117 – 128, 2016.

[41] D. Ginsbourger, O. Roustant, D. Schuhmacher, N. Durrande, and N. Lenz. On ANOVA decompositions of kernels and Gaussian random field paths. *ArXiv e-prints*, September 2014.

[42] T. Gneiting. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4):1327–1349, 09 2013.

[43] J.G. De Gooijer and R.J. Hyndman. 25 years of time series forecasting. *International Journal of Forecasting*, 2006.

[44] R. B. Gramacy and H. K. H. Lee. Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722, 2012.

[45] R. B. Gramacy, J. Niemi, and R. M. Weiss. Massively parallel approximate gaussian process regression. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):564–584, 2014.

[46] R.B. Gramacy and H. Lian. Gaussian process single-index models as emulators for computer experiments. *Technometrics*, 54(1):30–41, 2012.

[47] B. Hansen. The new econometrics of structural change: Dating breaks in U.S. Labor Productivity. *Journal of Economic Perspectives*, 15:117–128, 2001.

[48] K. A. Haskard. *An anisotropic Matern spatial covariance model: REML estimation and properties*. PhD thesis, University of Adelaide, School of Agriculture, Food and Wine, 2007.

[49] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2009.

[50] D. M. Hawkins. Multivariate quality control based on regression-adjusted variables. *Technometrics*, 33(1):61–75, 1991.

[51] J.M. Hellerstein. Quantitative Data Cleaning for Large Databases. United Nations Economic Commission for Europe (UNECE), 2008.

[52] C.C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.

[53] P.J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1981.

[54] M. Hubert. *Theory and Applications of Recent Robust Methods*. Statistics for industry and technology. Birkhäuser, 2004.

[55] M. Hubert and K. Van Driessen. Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45:301–320, 2004.

[56] R.J. Hyndman, M.L. King, I. Pitrun, and B. Billah. Local Linear Forecasts Using Cubic Smoothing Splines. *Australian and New Zealand Journal of Statistics*, 47(1):87–99, 2005.

[57] B. Iooss. Revue sur l'analyse de sensibilité globale de modèles numériques. *Journal de la Société Française de Statistique*, 152(1):3–25, 2011.

[58] M. E. H. Ismail. *Classical and quantum orthogonal polynomials in one variable*, volume 98 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2005. With two chapters by Walter Van Assche, With a foreword by Richard A. Askey.

[59] J. Jin. Individual station monitoring using press tonnage sensors for multiple operation stamping processes. *Journal of Manufacturing Science and Engineering*, 126(5):83–90, 2004.

[60] G. Jona-Lasinio, A. Gelfand, and M. Jona-Lasinio. Spatial Analysis of Wave Direction Data using Wrapped Gaussian Processes. *The Annals of Applied Statistics*, 6(4):1478–1498, 2012.

[61] E. S. Gardner Jr. Rule-based forecasting vs. damped-trend exponential smoothing. *Management Science*, 45(8):1169–1176, 1999.

[62] E.S. Gardner Jr. Exponential smoothing: The state of the art–part ii. *International Journal of Forecasting*, 22(4):637–666, 2006.

[63] L. Kang and S.L. Albin. On-line monitoring when the process yields a linear profile. *JOURNAL OF QUALITY TECHNOLOGY*, 32(4):418–426, 2000.

[64] R.S. Kenett and M. Pollak. On assessing the performance of sequential procedures for detecting a change. *Quality and Reliability Engineering International*, 28(5):500–507, 2012.

[65] T. Kourti and J. F. MacGregor. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28:3–21, 1995.

[66] S.G. Krantz. *Handbook of Complex Variables*. Birkhäuser Boston, 2012.

[67] C.A. Leikis. Consolidated Capacity and Performance Reporting. *Int. CMG-CONFERENCE*, 2:527–534, 2007.

[68] C.A. Lowry and D.C. Montgomery. A review of multivariate control charts. *IIE Transactions*, 27:800–810, 1995.

[69] M. Lutz. *Industrial decision-aid socio-statistical methods : Applied to the capacity management of an IS in the microelectronics industry*. Theses, Ecole Nationale Supérieure des Mines de Saint-Etienne, May 2013.

[70] M. Lutz, E. Padonou, and O. Roustant. Implementing statistical methods to improve information system management in a semiconductor industry. In *13th Annual Conference of the European Network for Business and Industrial*, Ankara, Turkey, 2013. International Year of Statistics, Greenfield Challenge.

[71] S. Makridakis, S. Wheelwright, and R.J. Hyndman. *Forecasting, methods and applications*. Wiley, 3rd edition, 1998.

[72] K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 2000.

[73] G. Matheron. *Principles of geostatistics*, volume 58. Society of Economic Geologists, 1963.

[74] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[75] W. R. Mebane, Jr. and J. S. Sekhon. Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software*, 42(11):1–26, 2011.

[76] D.C. Montgomery. *Statistical Quality Control: A Modern Introduction*. John Wiley & Sons, 2012.

[77] M.D. Morris. Gaussian surrogates for computer models with time-varying inputs and outputs. *Technometrics*, 54(1):42–50, 2012.

[78] M.D. Morris and T.J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43:381–402, 1995.

[79] R. Navarro and J. Arines. *Complete Modal Representation with Discrete Zernike Polynomials - Critical Sampling in Non Redundant Grids*. INTECH Open Access Publisher, 2011.

[80] R. Noorossana, A. Saghaei, and A. Amiri. *Statistical Analysis of Profile Monitoring*. Wiley Series in Probability and Statistics. Wiley, 2011.

[81] E. Padonou and O. Roustant. Polar Gaussian Processes and Experimental Designs in Circular Domains. SIAM/ASA Journal on Uncertainty Quantification (forthcoming), March 2016.

[82] E. Padonou, O. Roustant, J. Blue, and H. Duverneuil. Spatial risk assessment on circular domains: Application to wafer profile monitoring. In *26th Advanced Semiconductor Manufacturing Conference (ASMC), 2015*, pages 223 – 225, Saratoga, New York, United States, May 2015.

[83] E. Padonou, O. Roustant, and M. Lutz. Robust Monitoring of an Industrial IT System in the Presence of Structural Change. *Quality and Reliability Engineering International*, 2014.

[84] E.S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1/2):100–115, 1954.

[85] E.S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3-4):523–527, 1955.

[86] E.S. Page. Cumulative sum schemes using gauging. *Technometrics*, 4(1):97–109, 1962.

[87] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[88] G. Pison, S. Van Aelst, and G. Willems. Small Sample Corrections for LTS and MCD. *Metrika*, 55:111–123, 2002.

[89] G. Pistone and G. Vicario. Kriging prediction from a circular grid: application to wafer diffusion. *Applied Stochastic Models in Business and Industry*, 29(4):350–361, 2013.

[90] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[91] M. S. Reis and P. M. Saraiva. Multiscale statistical process control of paper surface profiles. *Quality Technology & Quantitative Management*, 3(3):263–282, 2006.

[92] B.D. Ripley. *Spatial Statistics*. Wiley Series in Probability and Statistics. Wiley, 2005.

[93] P.J. Rousseeuw. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[94] P.J. Rousseeuw and C. Croux. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

[95] P.J. Rousseeuw and K. Van Driessen. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41:212–223, 1998.

[96] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

[97] O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1):1–55, 2012.

[98] R.Y. Rubinstein. Generating random vectors uniformly distributed inside and on the surface of different regions. *European Journal of Operational Research*, 10(2):205 – 209, 1982.

[99] C. Rudd and V. Lloyd. *Service Design, ITIL, Version 3*. Stationery Office Books, 2007.

[100] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 11 1989.

[101] R.L. Sandland and G.M. Laslett. Precision and accuracy of kriging estimators with inter-laboratory trial information. In M. Armstrong, editor, *Geostatistics*, volume 4 of *Quantitative Geology and Geostatistics*, pages 797–808. Springer Netherlands, 1989.

[102] T. Sheil-Small. Analytic and harmonic functions in the unit disc. In *Complex Polynomials*, pages 125–171. Cambridge University Press, 2002. Cambridge Books Online.

[103] W. A. Shewart. *Economic control of Quality of Manufactured Product*. Van Nostrand Reinhold Co., New York, 1931.

[104] A.F. Siegel. Robust regression using repeated medians. *Biometrika*, 69:242–244, 1982.

[105] I. Sobol. Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments*, 1:407–414, 1993.

[106] A. Soumelidis, Z. Fazekas, F. Schipp, and M. Pap. *Electronic Engineering and Computing Technology*, chapter Discrete Orthogonality of Zernike Functions and Its Application to Corneal Measurements, pages 455–469. Springer Netherlands, Dordrecht, 2010.

[107] E. T. Spiller, M. J. Bayarri, J. O. Berger, E. S. Calder, A. K. Patra, E. B. Pitman, and R. L. Wolpert. Automating emulator construction for geophysical hazard maps. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):126–152, 2014.

[108] J.W. Taylor. Smooth transition exponential smoothing. *Journal of Forecasting*, 23:385–394, 2004.

[109] T.M. Williams. Adaptive Holt-Winters forecasting. *Journal of Operational Research Society*, 38(6):553–560, 1987.

[110] P.R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960.

[111] W. H. Woodall. Current research on profile monitoring. *Production*, 17:420 – 425, 12 2007.

[112] W. H. Woodall and D.C. Montgomery. Research issues and ideas in statistical process control. *Journal of Quality Technology*, 31(4):376–386, 1999.

[113] W.H. Woodall, D.J. Spitzner, D.C. Montgomery, and S. Gupta. Using control charts to monitor process and product quality profiles. *Journal of Quality Technology*, 36(3):309–320, 2004.

[114] S. Zacks and R.S. Kenett. Process tracking of time series with change points. *Recent Advances in Statistics and Probability*, Proceedings of the 4th International Meeting of Statistics in the Basque Country:155–171, 1994.

[115] A. Zeileis. Alternative boundaries for cusum tests. *Statistical Papers / Statistische Hefte*, 2004.

[116] A. Zeileis. A Unified Approach to Structural Change Tests Based on ML Scores, *F* Statistics, and OLS Residuals. *Econometric Reviews*, 24:445–466, 2005.

[117] A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. Strucchange: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of statistical Software*, 7:1–38, 2002.

[118] F. Zernike. Diffraction theory of the cut procedure and its improved form, the phase contrast method. *Physica*, 1:689–704, 1934.

[119] H. Zhang. Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, March 2004.

Appendix 1

We prove the Proposition presented in Section 10.2.2: If $t_1, t_1+1, \ldots, t_1+p-1$ is a suspicious period, then the sequence $\hat{y}_{t_1}$, $\hat{y}_{t_1+1}$, $\ldots \hat{y}_{t_1+p-1}$ is given by a deterministic and monotonic function f which does not depend on any observation posterior to date t_1.

Proof

The period $t_1, t_1+1, \ldots, t_1+p-1$ is defined by either Equation 10.17 or Equation 10.18. Without lost of generality, consider Equation 10.17. By combining it and Equation 10.8, we obtain:

$$\hat{\sigma}_t^2 = (1 + \lambda_\sigma(k^2-1))\hat{\sigma}_{t-1}^2 \qquad t_1+1 \leq t \leq t_1+p \tag{10.23}$$

which shows that the predicted errors $\hat{\sigma}_t$ follow a geometric progression with common ratio,

$$r = \sqrt{1+\lambda_\sigma(k^2-1)} \tag{10.24}$$

Consequently, by denoting $t_0 := t_1 - 1$, the date before t_1, the predicted errors are given by:

$$\hat{\sigma}_t = r^{t-t_0}\hat{\sigma}_{t_0} \qquad t_1+1 \leq t \leq t_1+p \tag{10.25}$$

With the common choice $k = 2, \lambda_\sigma = 0.3$, we have $r \approx 1.378 > 1$. Thus, the expected error goes increasing exponentially. Furthermore, the cleaned time series y_t^* becomes:

$$y_t^* = k.r^{t-t_0}\hat{\sigma}_{t_0} + \hat{y}_t \tag{10.26}$$

Now, relying on the equivalence with ARIMA(0,2,2) model (see e.g. Hyndman [56]), the forecast values of the Holt-Winters smoothing with parameters (λ_1, λ_2) follow the recursive scheme:

$$\hat{y}_{t+1} = (2-\theta_1)\, y_t^* + \theta_1 \hat{y}_t - (1+\theta_2)\, y_{t-1}^* + \theta_2 \hat{y}_{t-1} \tag{10.27}$$

where θ_1 and θ_2 are the parameters of the corresponding ARIMA(0, 2, 2) model (See Equations 10.21 and 10.22). Given Equation 10.26, this scheme becomes:

$$\hat{y}_{t+1} - 2\hat{y}_t + \hat{y}_{t-1} = P(t) \tag{10.28}$$

with:

$$P\,(t) = k[(2-\theta_1)\, r - (1+\theta_2)]r^{(t-1-t_0)}\hat{\sigma}_{t_0} \tag{10.29}$$

This is a linear equation, whose solutions are given by the sum of the solutions of the homogeneous linear equation and a particular solution:

$$\hat{y}_t = f(t) = \frac{r^{t+1}}{(r-1)^2} + a.t + b \tag{10.30}$$

The values of the constants a and b are imposed by $\hat{y}_{t_1}$ and $\hat{y}_{t_1+1}$. This shows indeed that the predicted values are purely deterministic and increasing exponentially during the suspicious period.

Sensitivity analysis

6 analytical fonctions are used to further understand the meaning of the sensitivity indices presented in Chapter 4. For each function, the 5 following GP models are estimated:

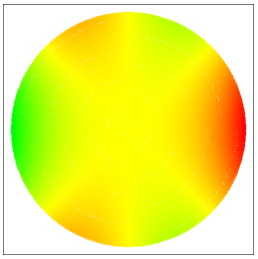
1. A Cartesian GP with kernel $k_{\text{prod+add}}$
2. A centred Cartesian GP with kernel $k_{\text{fad:}[-1,1]^2}$
3. A polar GP with kernel $k_{\text{prod+add}}$
4. A centred polar GP with kernel $k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$
5. A centred polar GP with kernel $k^{*}_{\text{fad:}\mathcal{D}}$

These functions are recovered over $\mathcal{D}$ based on the two maximin Latin cylinders LCD and LCD* presented in Section 5.2. We recall that LCD fills uniformly the space (ρ, θ) whereas LCD* fills the space (x, y). In the same logic, the $\mu^{\dagger}$ is uniform over (ρ, θ) whereas μ^{*} is uniform over the disk. For each toy function and each design, the 5 models are assessed, based on the following indicators:

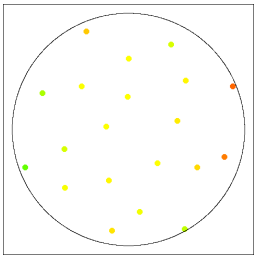
1. RMSE based on predictions at 1000 points filling the disk
2. Likelihood in a logarithmic scale
3. Sobol decomposition for $k_{\text{fad:}[-1,1]^2}$, $k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$ and $k^{*}_{\text{fad:}\mathcal{D}}$

As we are going to see, the centred GP models presented in Chapter 4 ouperform the standard additive and tensor products kernels in addition to providing relevant Sobol indices. A key point is that the probability measures $\mu^{\dagger}$ and μ^{*} may lead to different results in the presence of radial shapes. Indeed, the two measures correspond to the same marginal distribution of polar angles. In the study, the GP models are fitted with the matérn$\frac{5}{2}$ kernel over $[0, 1]^2$ and $[-1, 1]^2$ the C^2-Wendland function fover $\mathbb{S}^2$.

Test function 1



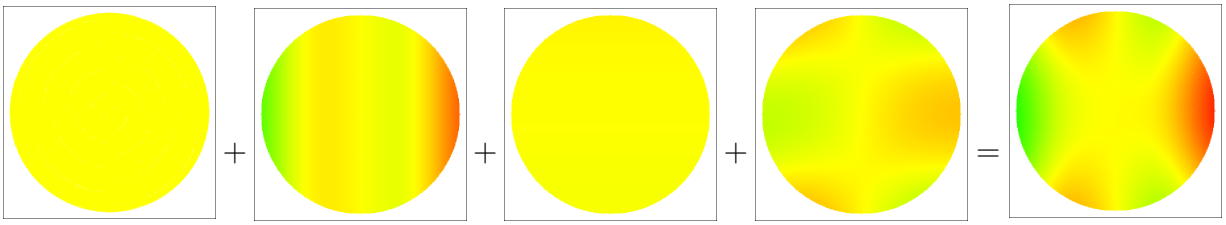
Dataset



DoE

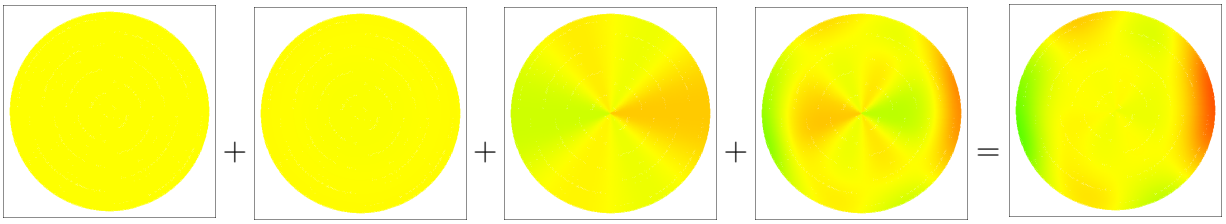
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	34.7	5.4
Cartesian	$k_{\text{fad:}[-1,1]^2}$	10.2	21.7
Polar	$k_{\text{prod+add}}$	32.9	0.4
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	27.9	3.3
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	31.6	7.4

Assessment of different GP models



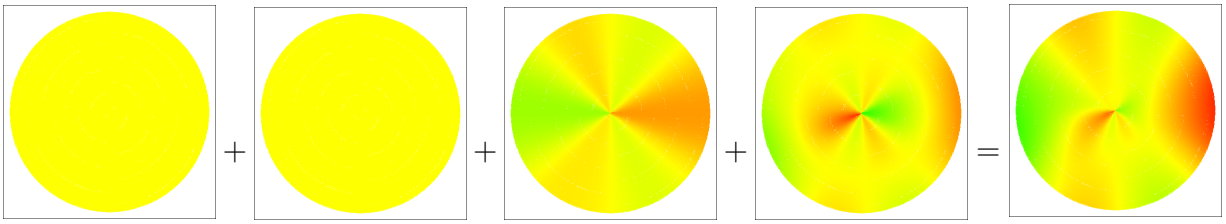
$\mu(x,y)$	$m_1(x)$	$m_2(y)$	$m_{12}(x,y)$	$m(x,y)$
–	$S = 74.7$	$S = 0.4$	$S = 24.8$	Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 0.3$	$S = 41.6$	$S = 58.1$	Kriging mean

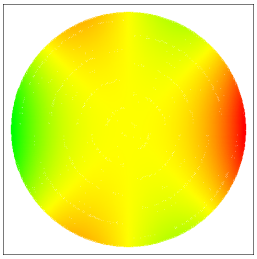
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



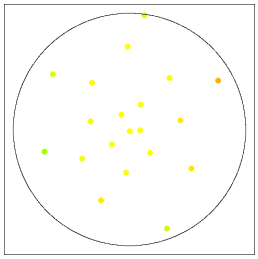
$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 0$	$S = 60.4$	$S = 39.6$	Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 1



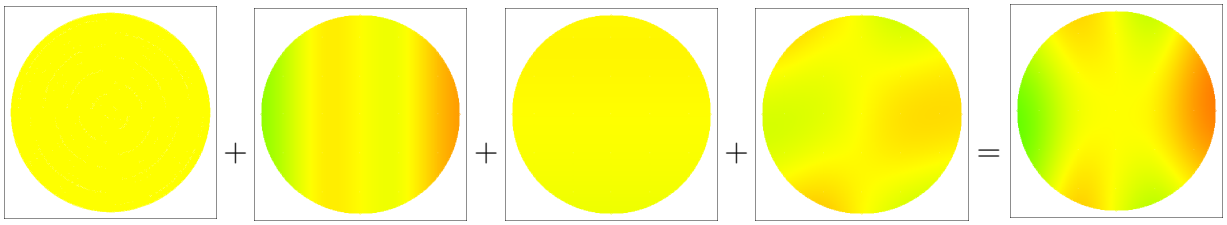
Dataset



DoE

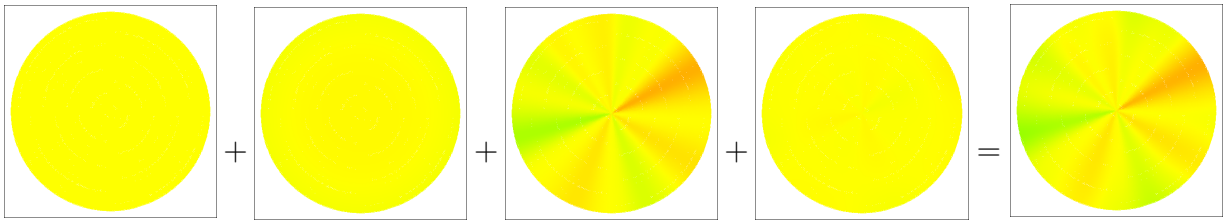
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	61	17.7
Cartesian	$k_{\text{fad:}[-1,1]^2}$	33	35.1
Polar	$k_{\text{prod+add}}$	94	13
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	81.3	14.8
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	83.6	14.7

Assessment of different GP models



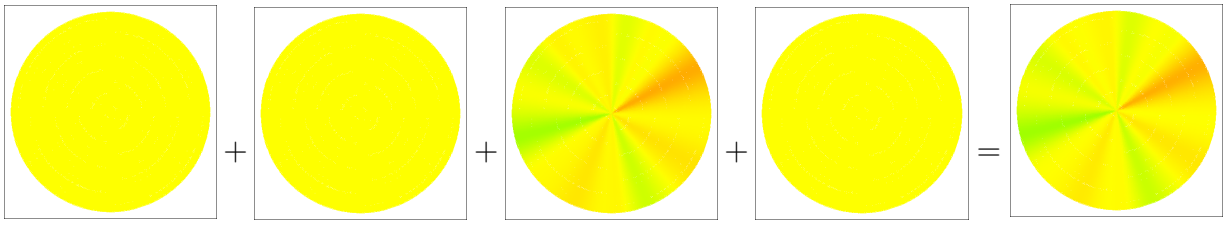
$\mu(x,y)$ — | $m_1(x)$ $S = 76.5$ | $m_2(y)$ $S = 2$ | $m_{12}(x,y)$ $S = 21.4$ | $m(x,y)$ Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$ — | $m_1(\rho)$ $S = 2.1$ | $m_2(\theta)$ $S = 97.1$ | $m_{12}(\rho,\theta)$ $S = 0.8$ | $m(\rho,\theta)$ Kriging mean

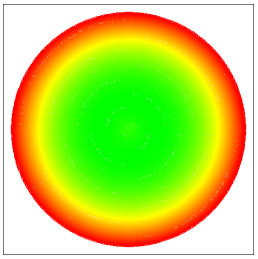
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



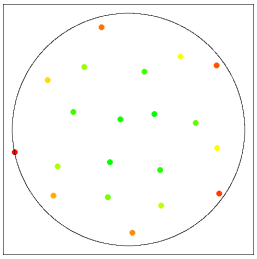
$\mu(x,y)$ — | $m_1(\rho)$ $S = 0$ | $m_2(\theta)$ $S = 100$ | $m_{12}(\rho,\theta)$ $S = 0$ | $m(\rho,\theta)$ Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 2



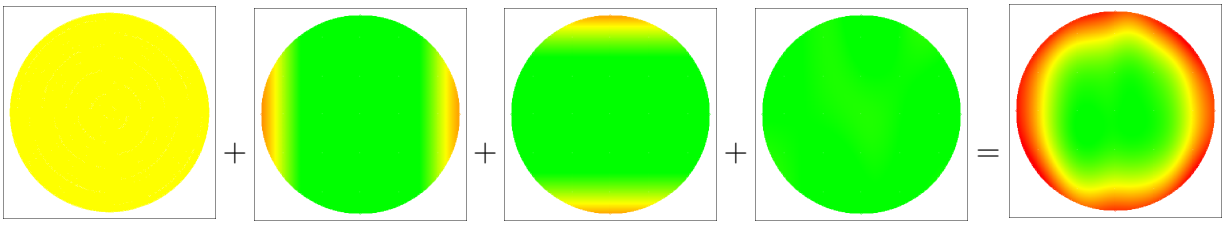
Dataset



DoE

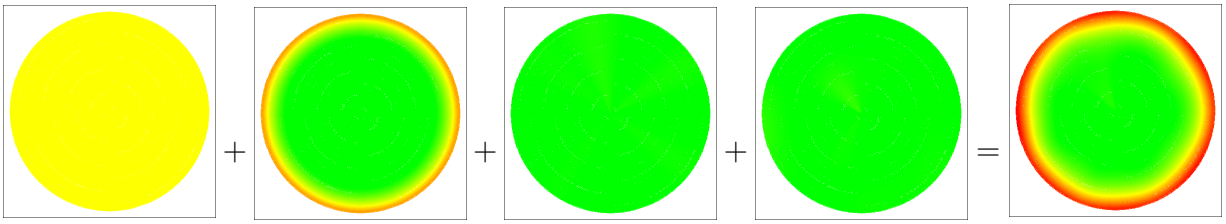
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	21.4	13.6
Cartesian	$k_{\text{fad:}[-1,1]^2}$	16.8	21.5
Polar	$k_{\text{prod+add}}$	0.3	67.4
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	13.8	21.8
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	0	140.4

Assessment of different GP models



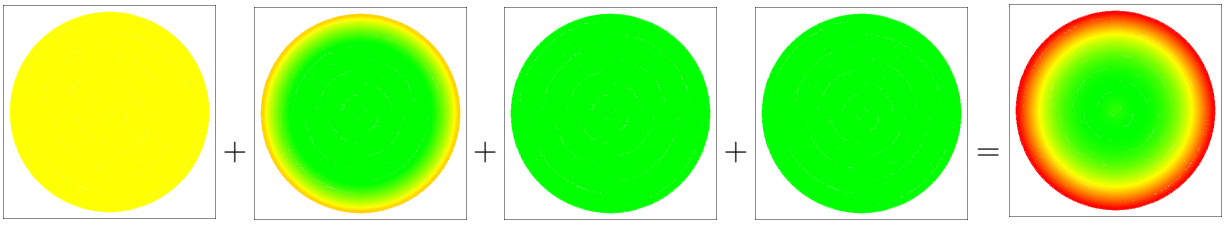
$\mu(x,y)$ — | $m_1(x)$ $S = 53.3$ | $m_2(y)$ $S = 45.7$ | $m_{12}(x,y)$ $S = 1$ | $m(x,y)$ Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$ — | $m_1(\rho)$ $S = 98.9$ | $m_2(\theta)$ $S = 0.6$ | $m_{12}(\rho,\theta)$ $S = 0.6$ | $m(\rho,\theta)$ Kriging mean

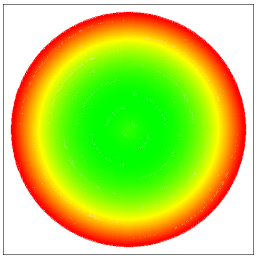
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



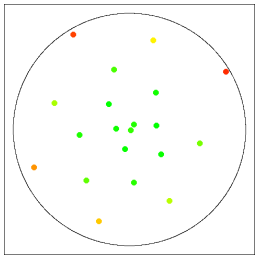
$\mu(x,y)$ — | $m_1(\rho)$ $S = 100$ | $m_2(\theta)$ $S = 0$ | $m_{12}(\rho,\theta)$ $S = 0$ | $m(\rho,\theta)$ Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 2



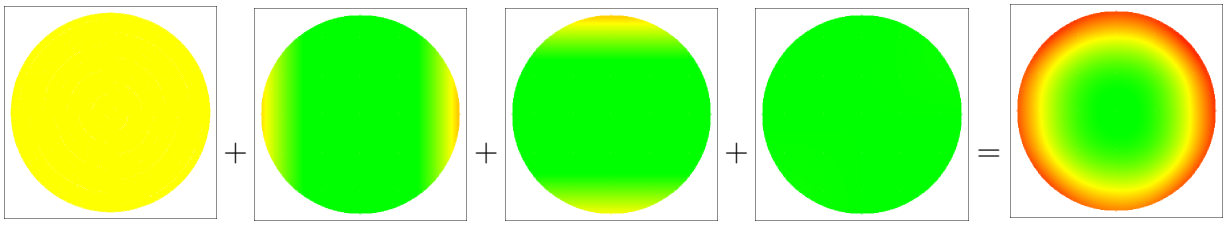
Dataset



DoE

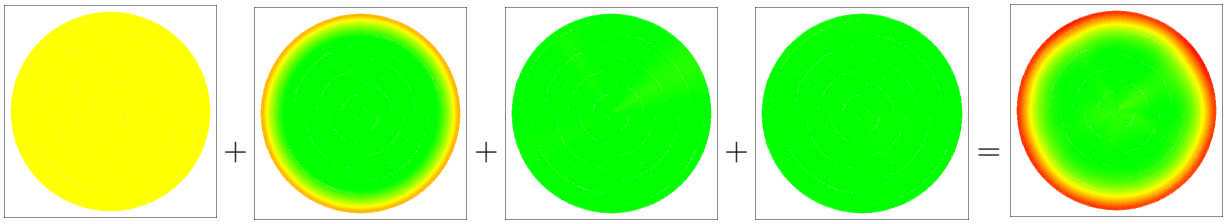
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	34	16.3
Cartesian	$k_{\text{fad:}[-1,1]^2}$	21.7	20.7
Polar	$k_{\text{prod+add}}$	0.3	78.2
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	12.1	21.3
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	0	131.8

Assessment of different GP models



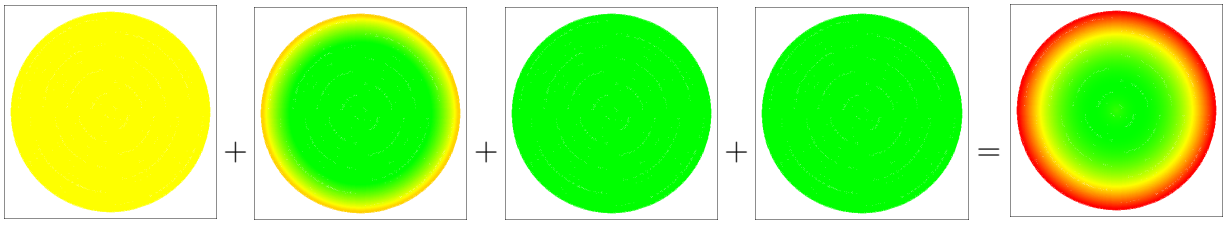
$\mu(x,y)$	$m_1(x)$	$m_2(y)$	$m_{12}(x,y)$	$m(x,y)$
–	$S = 51.9$	$S = 48.1$	$S = 0.1$	Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 99.4$	$S = 0.6$	$S = 0.1$	Kriging mean

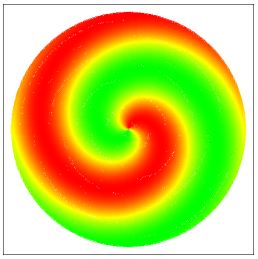
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



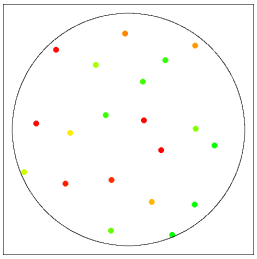
$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 100$	$S = 0$	$S = 0$	Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 3



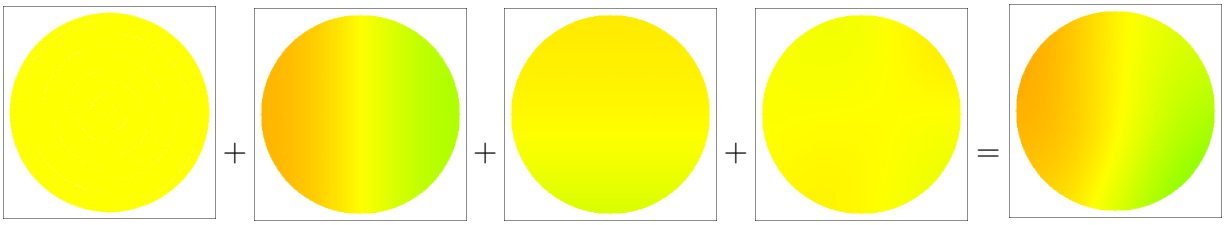
Dataset



DoE

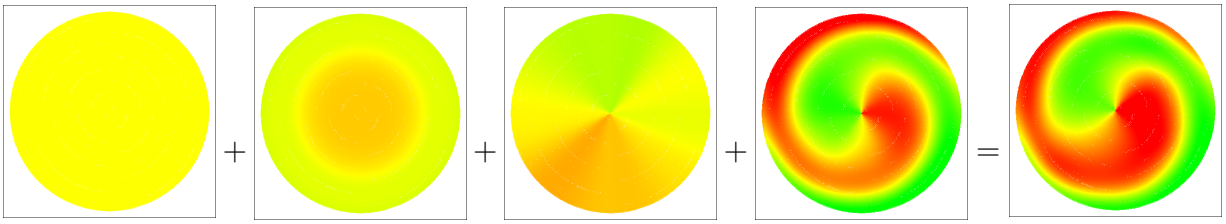
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	102.5	-22.4
Cartesian	$k_{\text{fad:}[-1,1]^2}$	90.2	-22.5
Polar	$k_{\text{prod+add}}$	39.6	-19.9
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	41	-19.2
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	18.5	-6.3

Assessment of different GP models



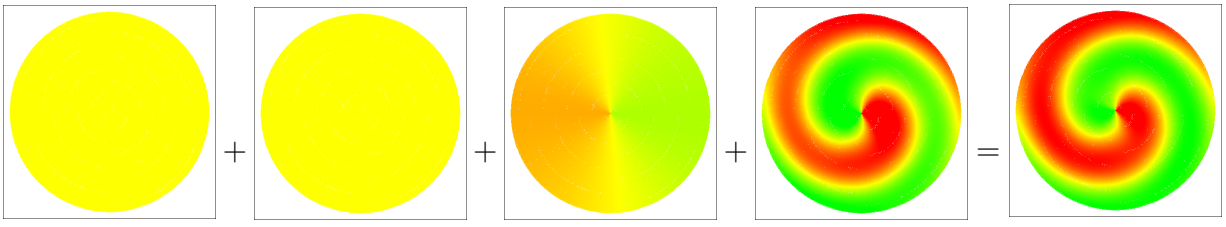
$\mu(x,y)$ — | $m_1(x)$ $S = 88.6$ | $m_2(y)$ $S = 9.4$ | $m_{12}(x,y)$ $S = 2$ | $m(x,y)$ Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$ — | $m_1(\rho)$ $S = 3.8$ | $m_2(\theta)$ $S = 9.7$ | $m_{12}(\rho,\theta)$ $S = 86.4$ | $m(\rho,\theta)$ Kriging mean

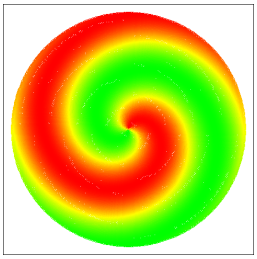
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



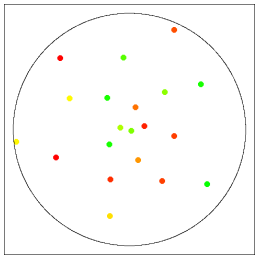
$\mu(x,y)$ — | $m_1(\rho)$ $S = 0$ | $m_2(\theta)$ $S = 10.4$ | $m_{12}(\rho,\theta)$ $S = 89.6$ | $m(\rho,\theta)$ Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 3



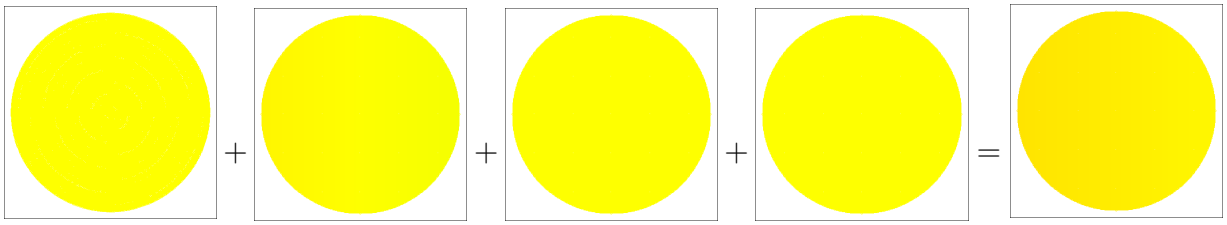
Dataset



DoE

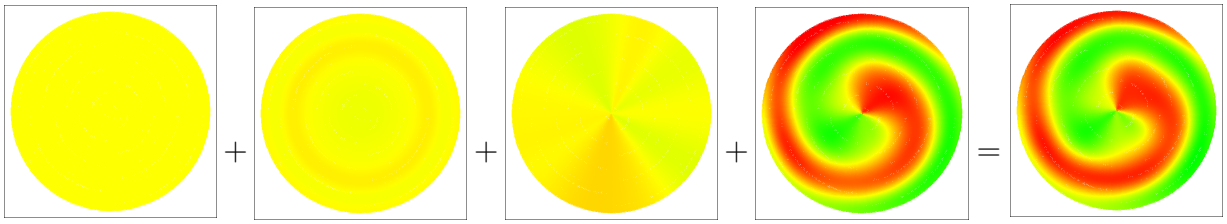
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	100.5	-21.3
Cartesian	$k_{\text{fad:}[-1,1]^2}$	99.3	-21.2
Polar	$k_{\text{prod+add}}$	100.5	-21.3
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	29.8	-20.4
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	9.9	-11.6

Assessment of different GP models



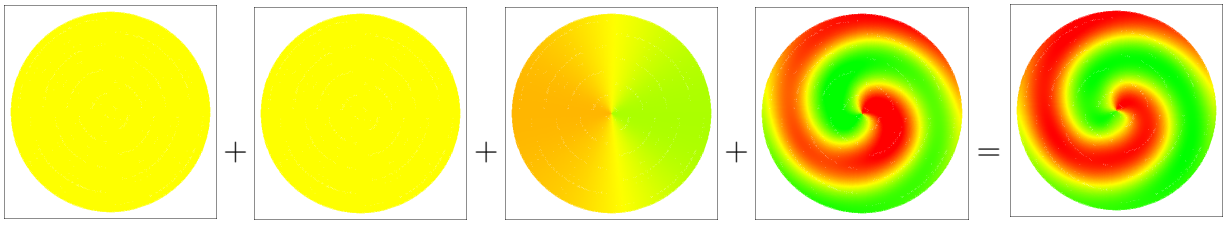
$\mu(x,y)$	$m_1(x)$	$m_2(y)$	$m_{12}(x,y)$	$m(x,y)$
$-$	$S = 100$	$S = 0$	$S = 0$	Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
$-$	$S = 0.3$	$S = 1.8$	$S = 97.8$	Kriging mean

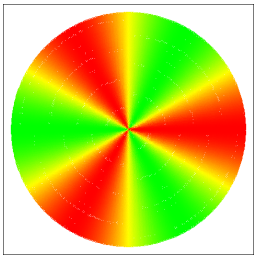
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



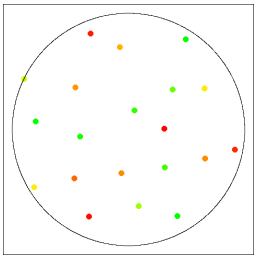
$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
$-$	$S = 0$	$S = 10.7$	$S = 89.3$	Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 4



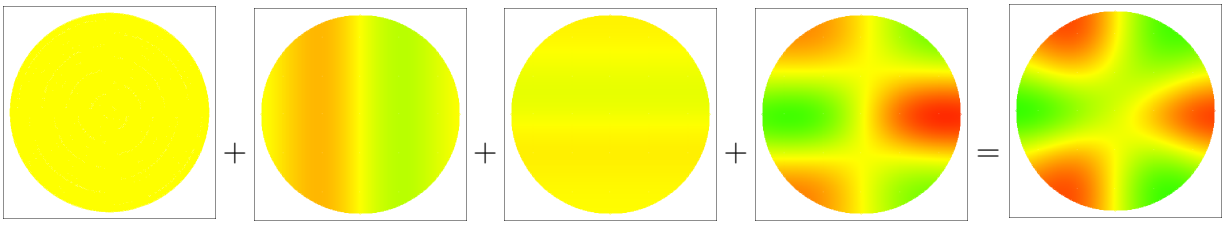
Dataset



DoE

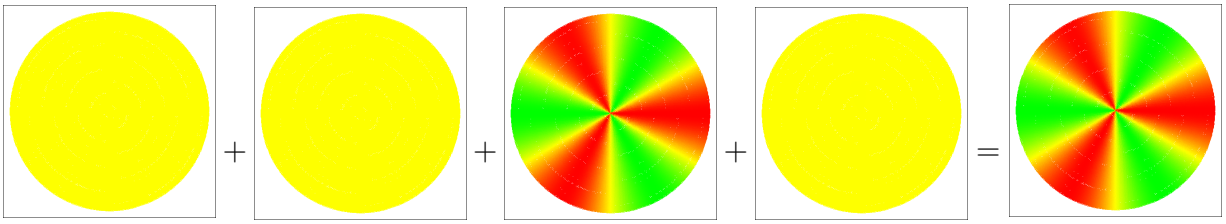
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	97.2	-21.4
Cartesian	$k_{\text{fad:}[-1,1]^2}$	60.6	-21.1
Polar	$k_{\text{prod+add}}$	1.3	-8.2
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	0.9	-3.8
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	0.9	-2.3

Assessment of different GP models



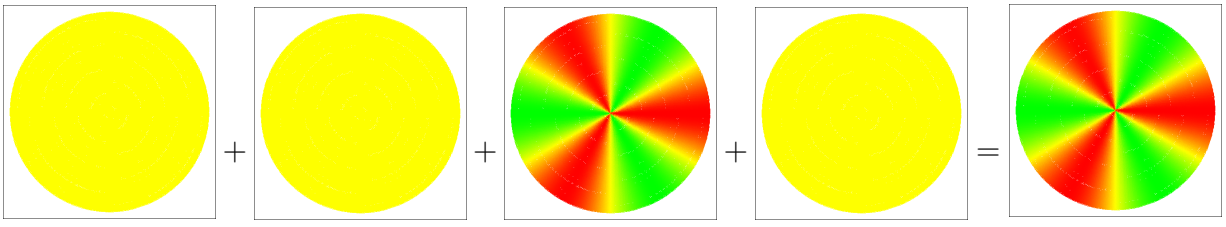
$\mu(x,y)$	$m_1(x)$	$m_2(y)$	$m_{12}(x,y)$	$m(x,y)$
–	$S = 20.3$	$S = 1.3$	$S = 78.5$	Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 0$	$S = 100$	$S = 0$	Kriging mean

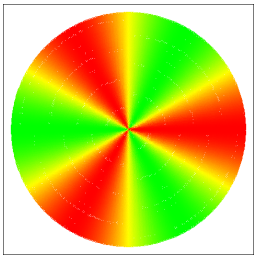
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



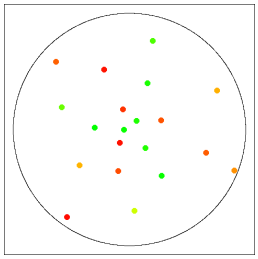
$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 0$	$S = 100$	$S = 0$	Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 4



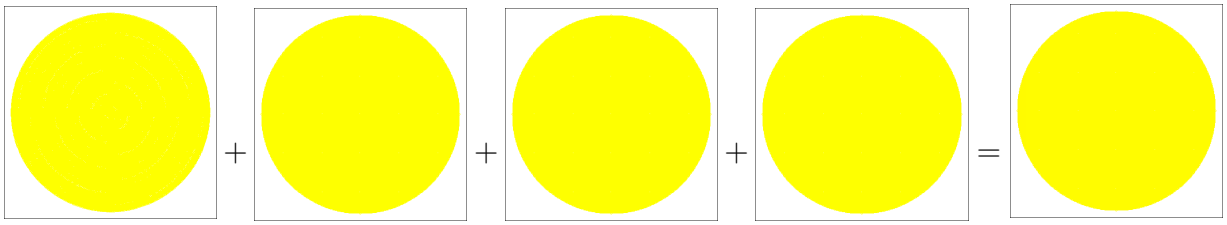
Dataset



DoE

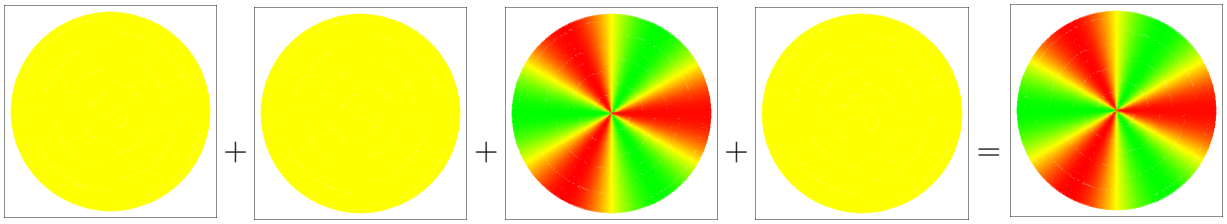
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	99.4	-22.8
Cartesian	$k_{\text{fad:}[-1,1]^2}$	100	-22.8
Polar	$k_{\text{prod+add}}$	2.8	-3.4
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	1.4	-0.8
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	1.4	2.1

Assessment of different GP models



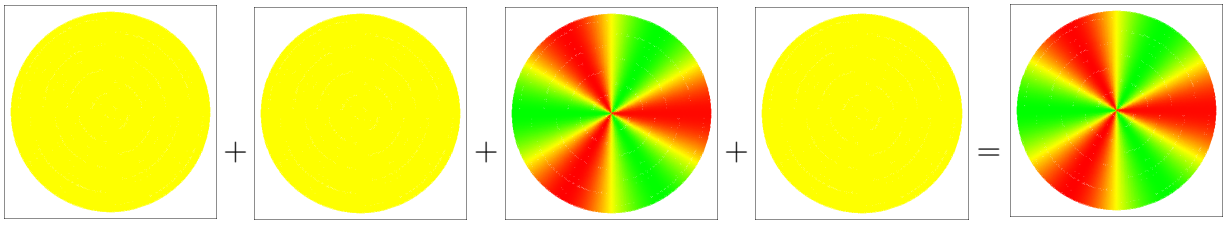
$\mu(x,y)$	$m_1(x)$	$m_2(y)$	$m_{12}(x,y)$	$m(x,y)$
–	$S = 100$	$S = 0$	$S = 0$	Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 0$	$S = 100$	$S = 0$	Kriging mean

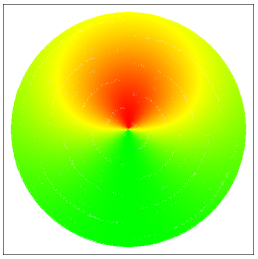
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



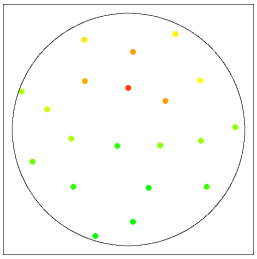
$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 0$	$S = 100$	$S = 0$	Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 5



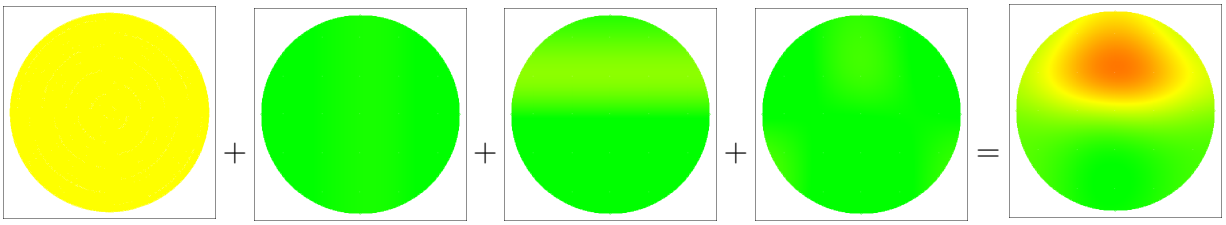
Dataset



DoE

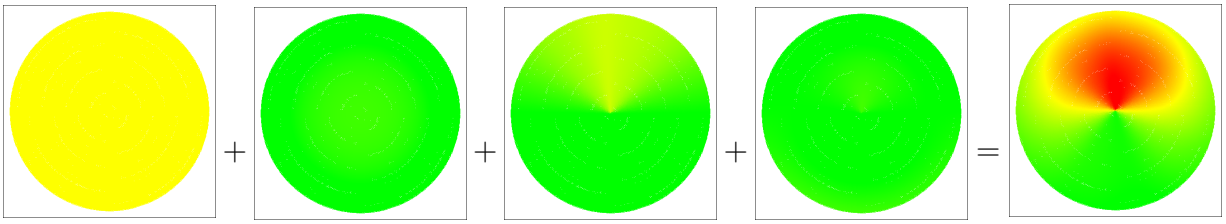
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	23.3	-2.7
Cartesian	$k_{\text{fad:}[-1,1]^2}$	27.2	-2.3
Polar	$k_{\text{prod+add}}$	5.7	5.9
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	3.8	4.4
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	2.7	16.4

Assessment of different GP models



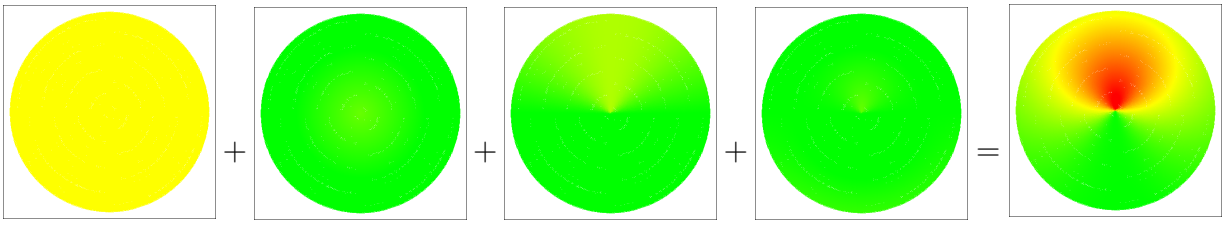
$\mu(x,y)$ — | $m_1(x)$ $S = 2.8$ | $m_2(y)$ $S = 85.1$ | $m_{12}(x,y)$ $S = 12.2$ | $m(x,y)$ Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$ — | $m_1(\rho)$ $S = 8.1$ | $m_2(\theta)$ $S = 88.6$ | $m_{12}(\rho,\theta)$ $S = 3.3$ | $m(\rho,\theta)$ Kriging mean

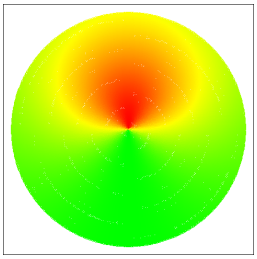
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



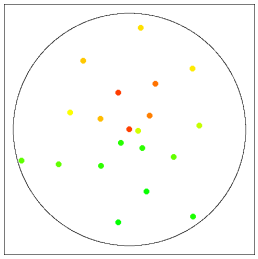
$\mu(x,y)$ — | $m_1(\rho)$ $S = 7.8$ | $m_2(\theta)$ $S = 88.5$ | $m_{12}(\rho,\theta)$ $S = 3.7$ | $m(\rho,\theta)$ Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 5



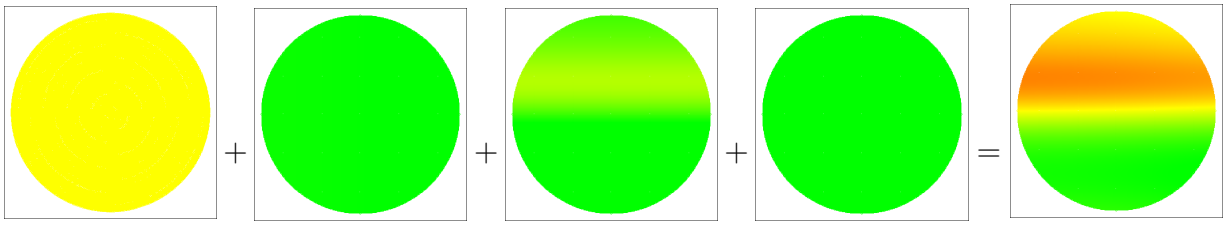
Dataset



DoE

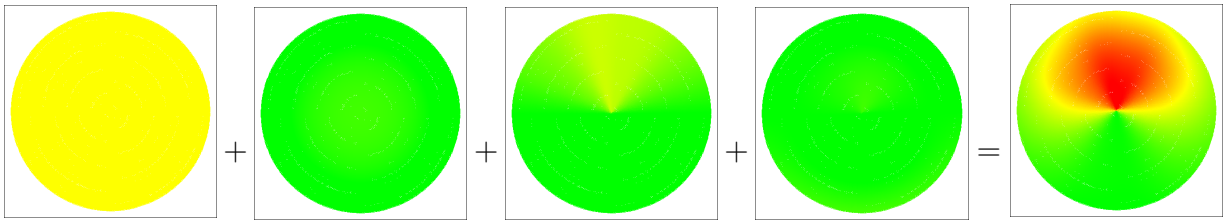
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	49.5	-11.2
Cartesian	$k_{\text{fad:}[-1,1]^2}$	50	-9.2
Polar	$k_{\text{prod+add}}$	2.9	4.7
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	3.9	8.2
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	4.2	13.1

Assessment of different GP models



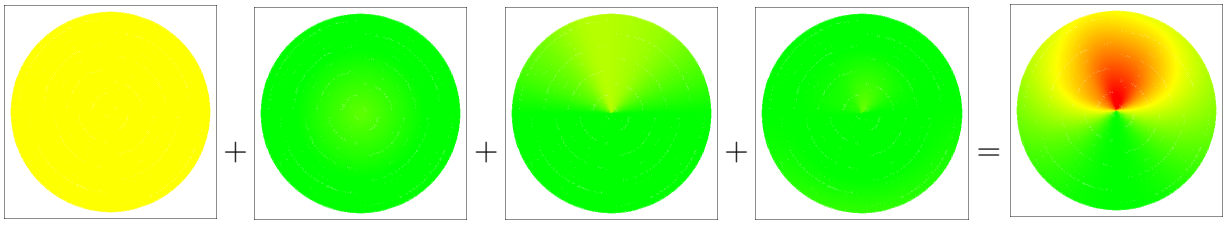
$\mu(x,y)$	$m_1(x)$	$m_2(y)$	$m_{12}(x,y)$	$m(x,y)$
–	$S = 0.3$	$S = 99.7$	$S = 0$	Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 7.5$	$S = 89.1$	$S = 3.3$	Kriging mean

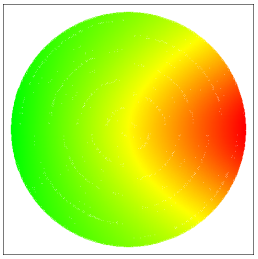
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



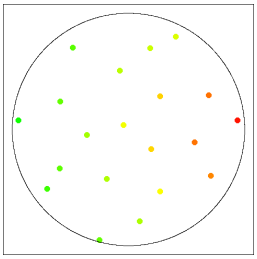
$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 6.8$	$S = 89.7$	$S = 3.5$	Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 6



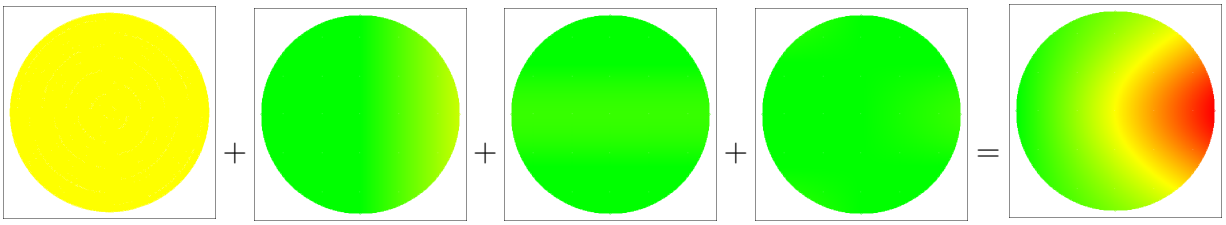
Dataset



DoE

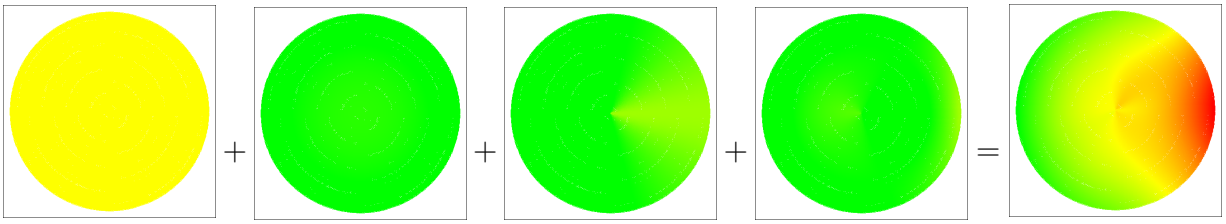
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	1.3	21.2
Cartesian	$k_{\text{fad:}[-1,1]^2}$	0.3	39.1
Polar	$k_{\text{prod+add}}$	12.9	2.1
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	4.5	11.1
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	5.1	11.6

Assessment of different GP models



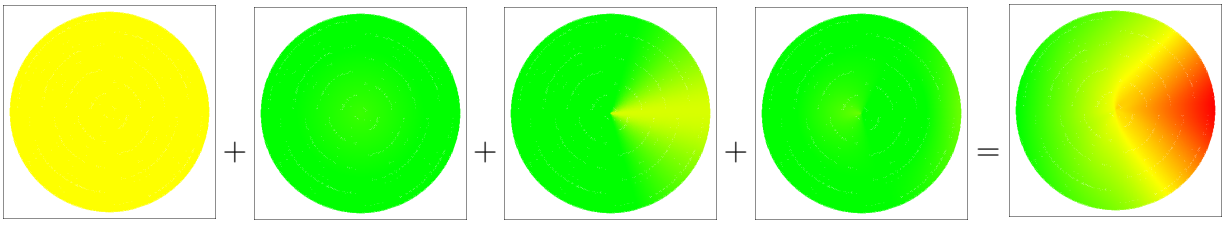
$\mu(x,y)$ — | $m_1(x)$ $S = 85.6$ | $m_2(y)$ $S = 10.8$ | $m_{12}(x,y)$ $S = 3.7$ | $m(x,y)$ Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$ — | $m_1(\rho)$ $S = 5.9$ | $m_2(\theta)$ $S = 74$ | $m_{12}(\rho,\theta)$ $S = 20.1$ | $m(\rho,\theta)$ Kriging mean

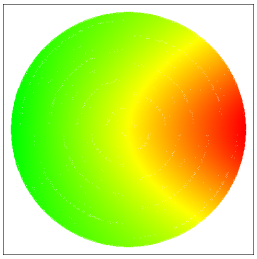
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



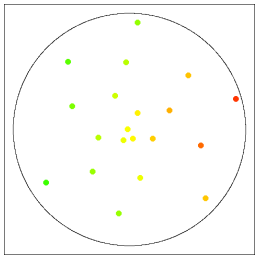
$\mu(x,y)$ — | $m_1(\rho)$ $S = 3$ | $m_2(\theta)$ $S = 88.1$ | $m_{12}(\rho,\theta)$ $S = 9$ | $m(\rho,\theta)$ Kriging mean

Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Test function 6



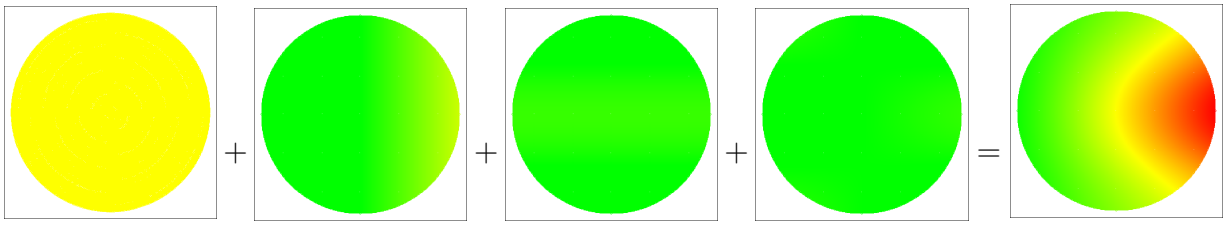
Dataset



DoE

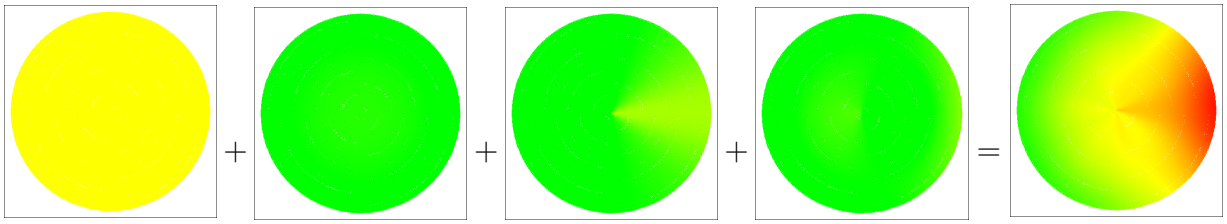
GP type	kernel	RMSE	Log-likelihood
Cartesian	$k_{\text{prod+add}}$	6	30.1
Cartesian	$k_{\text{fad:}[-1,1]^2}$	2.3	42.3
Polar	$k_{\text{prod+add}}$	9.2	4.9
Polar	$k^{\dagger}_{\text{fad:}[0,1]\times\mathbb{S}}$	8.5	8.5
Polar	$k^{*}_{\text{fad:}\mathcal{D}}$	8.3	10

Assessment of different GP models



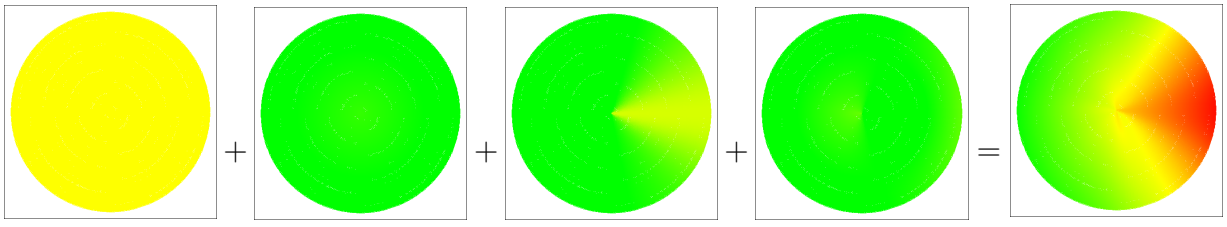
$\mu(x,y)$	$m_1(x)$	$m_2(y)$	$m_{12}(x,y)$	$m(x,y)$
–	$S = 86.7$	$S = 10.6$	$S = 2.7$	Kriging mean

Sobol decomposition of the Kriging mean (Cartesian GP, uniform measure over $[-1,1]^2$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 5.2$	$S = 75.7$	$S = 19.1$	Kriging mean

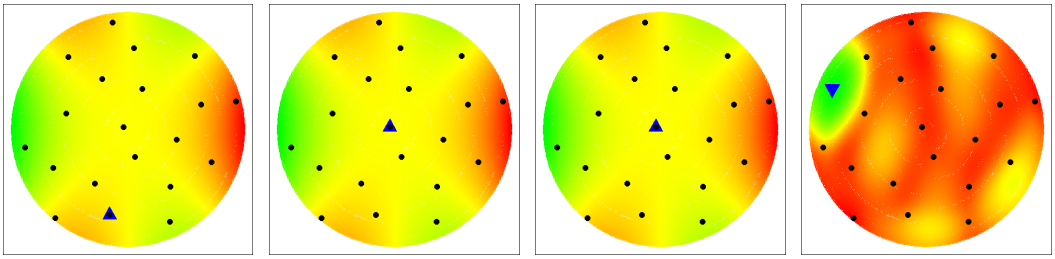
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $[0,1] \times \mathbb{S}$)



$\mu(x,y)$	$m_1(\rho)$	$m_2(\theta)$	$m_{12}(\rho,\theta)$	$m(\rho,\theta)$
–	$S = 2.6$	$S = 87.9$	$S = 9.5$	Kriging mean

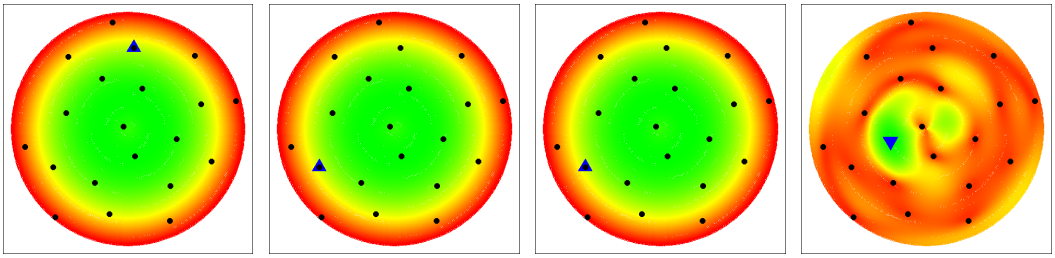
Sobol decomposition of the Kriging mean (Polar GP, uniform measure over $\mathcal{D}$)

Appendix: relocations



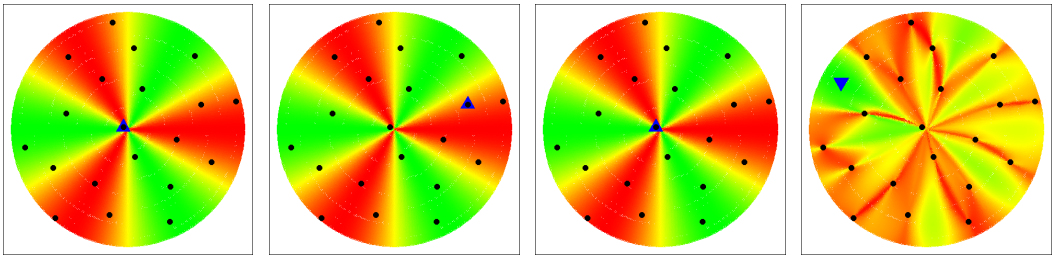
Leave-one-out 1 | Leave-one-out 2 | Leave-one-out 3 | $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $x^3 - xy^2$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



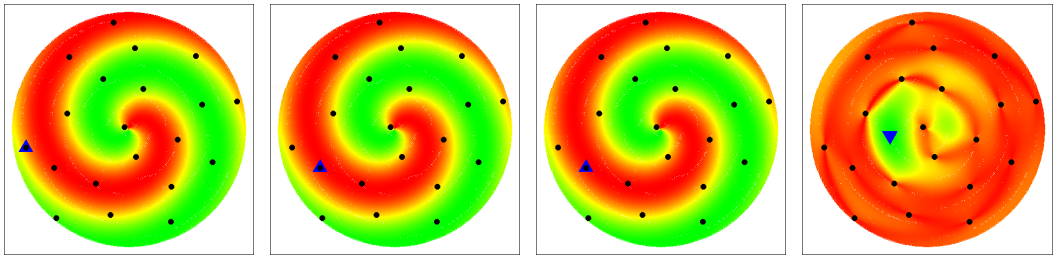
Leave-one-out 1 | Leave-one-out 2 | Leave-one-out 3 | $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $(\rho - \frac{1}{4})^2$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



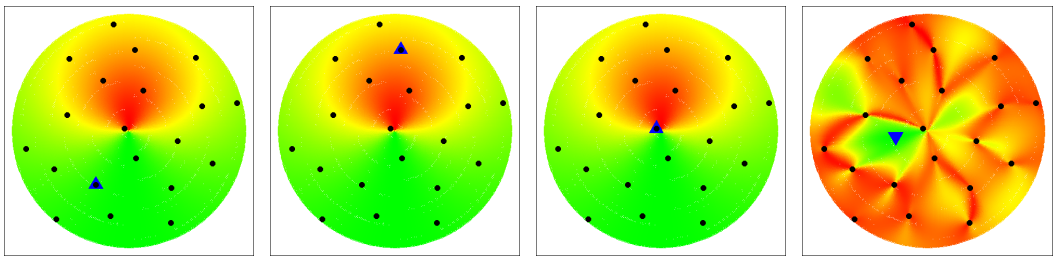
Leave-one-out 1 | Leave-one-out 2 | Leave-one-out 3 | $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $\cos(3\theta)$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



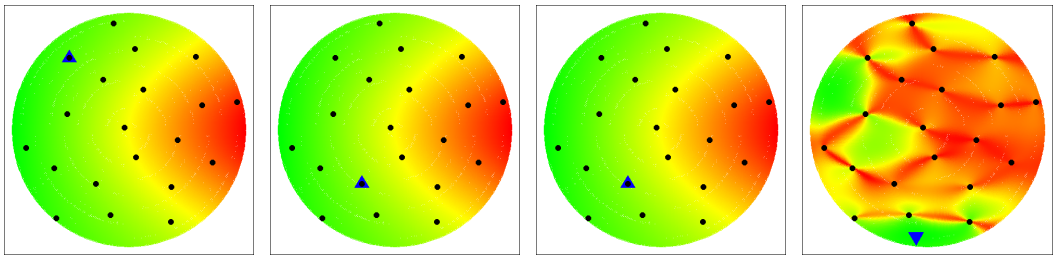
Leave-one-out 1 Leave-one-out 2 Leave-one-out 3 $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $\sin(2\pi\rho + \theta)$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



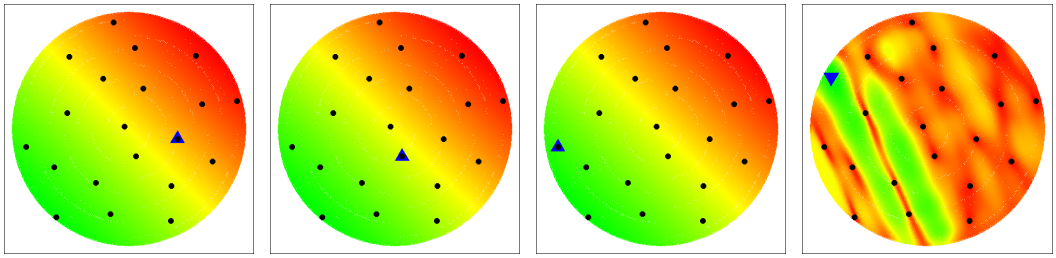
Leave-one-out 1 Leave-one-out 2 Leave-one-out 3 $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $\frac{1+\sin(\theta)}{1+\rho^2}$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



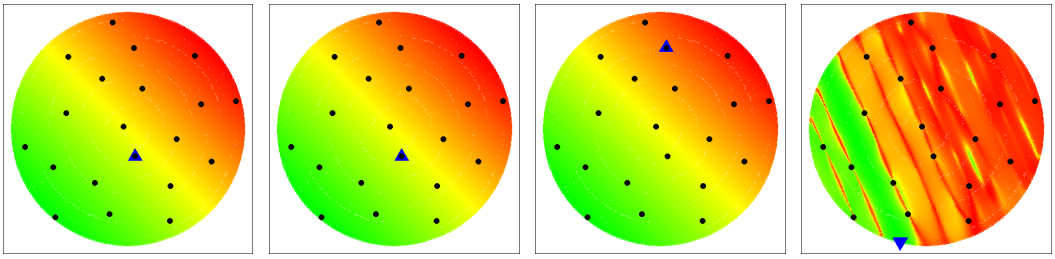
Leave-one-out 1 Leave-one-out 2 Leave-one-out 3 $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $\frac{1+x}{1+y^2}$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



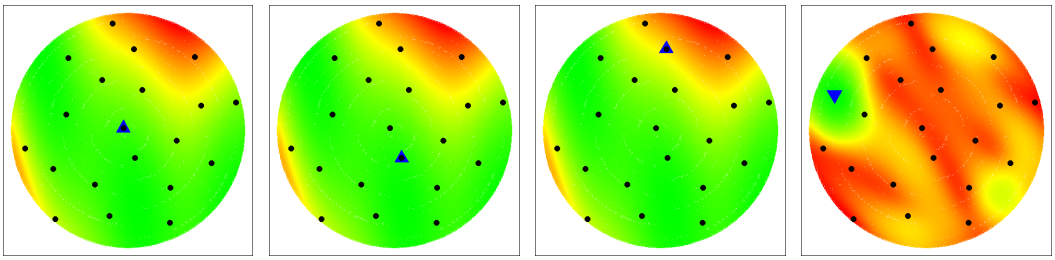
Leave-one-out 1 Leave-one-out 2 Leave-one-out 3 $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $\mathrm{S}(x + y)$, *S being the sigmoid*. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



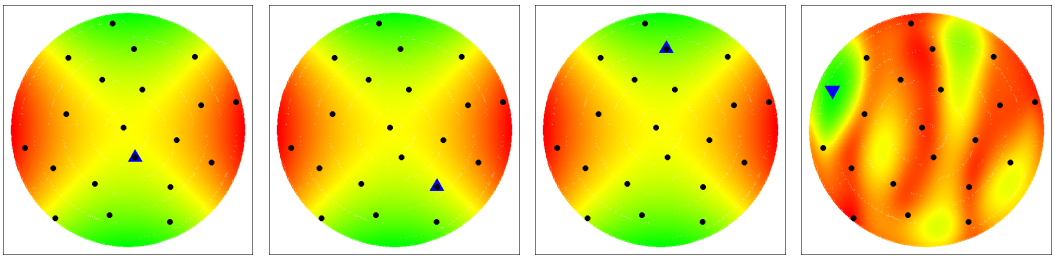
Leave-one-out 1 Leave-one-out 2 Leave-one-out 3 $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $\mathrm{sh}\Big(5(x_1+x_2)\Big)$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



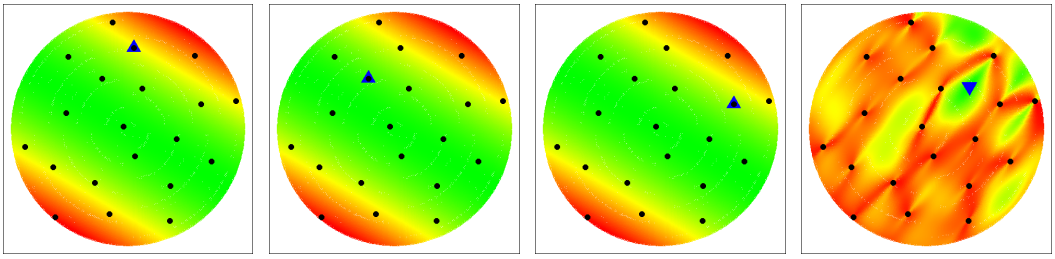
Leave-one-out 1 Leave-one-out 2 Leave-one-out 3 $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $\mathrm{B}\left(\frac{x+1}{2}, \frac{y+1}{2}\right)$, *B being the Branin function*. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



Leave-one-out 1 Leave-one-out 2 Leave-one-out 3 $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $Z_2^{+2}(x,y)$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.



Leave-one-out 1 Leave-one-out 2 Leave-one-out 3 $I_{n+1}(\mathbf{u})$

3 relocation strategies for the function $f_c\left(x,y\right) = \Big(x\cos\left(\frac{\pi}{3}\right) + y\sin\left(\frac{\pi}{3}\right)\Big)^{\beta}$. The point-up triangles are proposals for relocation, and the point-down triangle indicates the new location.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT: *2016LYSEM009*

Espéran PADONOU

STATISTICAL LEARNING ON CIRCULAR DOMAINS FOR ADVANCED PROCESS CONTROL IN MICROELECTRONICS

Speciality: Applied mathematics

Keywords: Disk, Gaussian process, Time series, Designs of experiments, Statistical Process Control, Sensitivity analysis

Abstract:

Driven by industrial needs in microelectronics, this thesis is focused on probabilistic models for spatial data and Statistical Process Control.

The spatial problem has the specificity of being defined on circular domains. It is addressed through a Kriging model where the deterministic part is made of orthogonal polynomials and the stochastic term represented by a Gaussian process. Defined with the Euclidean distance and the uniform measure over the disk, traditional Kriging models do not exploit knowledge on manufacturing processes.

To take rotations or diffusions from the center into account, we introduce polar Gaussian processes over the disk. They embed radial and angular correlations in Kriging predictions, leading to significant improvements in the considered situations. Polar Gaussian processes are then interpreted via Sobol decomposition and generalized in higher dimensions. Different designs of experiments are developed for the proposed models. Among them, Latin cylinders reproduce in the space of polar coordinates the properties of Latin hypercubes.

To model spatial and temporal data, Statistical Process Control is addressed by monitoring Kriging parameters, based on standard control charts. Furthermore, the monitored time – series contain outliers and structural changes, which cause bias in prediction and false alarms in risk management. These issues are simultaneously tackled with a robust and adaptive smoothing.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : *2016LYSEM009*

Espéran PADONOU

APPRENTISSAGE STATISTIQUE EN DOMAINE CIRCULAIRE POUR LA PLANIFICATION DE CONTROLES EN MICROELECTRONIQUE

Spécialité: Mathématiques appliquées

Mots clefs : Disque, Processus gaussiens, Séries temporelles, Plans d'expériences, Analyse de sensibilité, Maitrise Statistique de Procédés

Résumé :

Motivés par des besoins en industrie microélectronique, ces travaux apportent des contributions en modélisation probabiliste de données spatiales, et en maîtrise statistique de procédés.

Le problème spatial a pour spécificité d'être posé sur un domaine circulaire. Il se représente par un modèle de krigeage dont la partie déterministe est constituée de polynômes orthogonaux et la partie stochastique de processus gaussiens. Traditionnellement définis avec la norme euclidienne et la mesure uniforme sur le disque, ces choix n'exploitent pas les informations a priori sur les procédés d'usinage.

Pour tenir compte des mécanismes de rotation ou de diffusion à partir du centre, nous formalisons les processus gaussiens polaires sur le disque. Ces processus intègrent les corrélations radiales et angulaires dans le modèle de krigeage, et en améliorent les performances dans les situations considérées. Ils sont ensuite interprétés par décomposition de Sobol et généralisés en dimension supérieure. Des plans d'expériences sont proposés dans le cadre de leur utilisation. Au premier rang figurent les cylindres latins qui reproduisent en coordonnées polaires les caractéristiques des hypercubes latins.

Pour intégrer à la fois les aspects spatiaux et temporels du problème industriel, la maîtrise statistique de procédé est abordée en termes d'application de cartes de contrôle aux paramètres des modèles spatiaux. Les séries temporelles suivies ont aussi la particularité de comporter des données atypiques et des changements structurels, sources de biais en prévision, et de fausses alarmes en suivi de risque. Ce problème est traité par lissage robuste et adaptatif.