

UNIVERSITE CATHOLIQUE DE LOUVAIN
INSTITUT DE STATISTIQUE



STATISTICAL CONTRIBUTION TO THE
VIRTUAL MULTICRITERIA OPTIMISATION OF
COMBINATORIAL MOLECULES LIBRARIES AND TO THE
VALIDATION AND APPLICATION OF QSAR MODELS

Membres du jury:

Dr. Benoît Beck (*Axiosis*)
Prof. Bernadette Govaerts (*UCL, Promoteur*)
Prof. Philippe Lambert (*ULg*)
Prof. Rainer von Sachs (*UCL, Président du jury*)
Prof. Léopold Simar (*UCL, Co-promoteur*)
Prof. Claus Weihs (*Universität Dortmund*)

Thèse présentée en vue de
l'obtention du grade de
Docteur en Sciences
(orientation statistique) par :

Céline Le Bailly de Tillegem

Louvain-la-Neuve, December 2007

à Philippe et Clémence

Je voudrais remercier tous ceux qui ont contribué, de près ou de loin, à l'élaboration de ma thèse.

Je tiens tout d'abord à remercier ma promotrice, le Professeur B. Govaerts, de m'avoir transmis sa passion de la statistique appliquée. Elle a été, tout au long de ces années de thèse, d'une grande disponibilité et d'une grande écoute. Ses précieux conseils et son soutien permanent m'ont permis de mener ce travail à bien.

Je souhaite également exprimer toute ma gratitude aux membres du jury, qui ont accepté de lire ma thèse avec autant de soin. Je remercie en particulier le Professeur L. Simar, le Professeur Ph. Lambert et le Professeur R. von Sachs pour leurs conseils et leur accompagnement au sein de notre Institut. I also gratefully thank Professor C. Weihs for his scientific contribution and all the relevant comments and suggestions which have helped to improve this research.

Je voudrais également exprimer tout ma reconnaissance aux Dr. B. Beck et Dr. B. Boulanger pour m'avoir proposé de travailler sur ce projet, ainsi que pour leur suivi. Avec ma promotrice, ils m'ont également encouragé à participer à de multiples conférences (dont la première pour moi, mémorable, à Barcelone !). Cela m'a permis de partager ma recherche avec d'autres scientifiques, me motivant à apporter des réponses efficaces à des problèmes concrets. Je pense en particulier au Professeur D. Steinberg, au Dr. H. Trautmann et au Dr. F. Vandenhende.

Mes remerciements vont également à tous les membres de l'Institut, personnel administratif, académique et scientifique, qui ont contribué à une atmosphère de travail très agréable. Je tiens à remercier particulièrement le Pôle d'Attraction Interuniversitaire (PAI) pour son soutien financier, Monsieur Jean-Luc Marrion pour son aide informatique précieuse ainsi que Astrid, Bianca, Céline, Maria, Nancy, Oana et Réjane pour leurs encouragements et tous les bons moments partagés.

Je tiens finalement à remercier ma famille et mes amis pour leur soutien continu. Ils m'ont portée et supportée... Merci particulièrement à Philippe pour sa patience, sa compréhension, son soutien et toute son aide logistique à la maison ainsi qu'à Clémence qui, sans le savoir, m'a transmis une incroyable force et une immense motivation pour mener ce travail à terme.

Contents

INTRODUCTION:

context, goals and contents of the thesis	1
0.1 Context: drug discovery	1
0.2 Deciding on an area for pharmaceutical research	4
0.3 Drug discovery stages	5
0.3.1 Target identification and validation	5
0.3.2 Hit identification	7
0.3.3 Lead generation	7
0.3.4 Lead optimisation	7
0.4 Drug development phases	8
0.4.1 Phase I	9
0.4.2 Phase II	10
0.4.3 Phase III	10
0.4.4 Phase IV	11
0.5 Contents of the thesis	11
0.5.1 QSAR modelling	13
0.5.2 Multiresponse optimisation criterion	14
0.5.3 Optimisation algorithm	15
0.6 Summary	16
1 Quantitative Structure-Activity Relationship (QSAR) models: review and perspectives	23
1.1 History of QSAR models	23
1.2 The responses modelled in QSAR	31
1.3 The explanatory variables entered in QSAR models	32
1.4 Different classes of QSAR models	35
1.4.1 Multiple Linear Regression (MLR)	36
1.4.2 Principal Components Regression (PCR)	37
1.4.3 Partial Least Squares Regression (PLSR)	38

1.4.4	Ridge Regression (RR)	39
1.4.5	Regression Tree (RT)	40
1.4.6	k -Nearest Neighbours (k NN)	42
1.4.7	Neural Networks (NN)	43
1.4.8	Choice of complexity parameters	44
1.5	Data collection and model fit	45
1.6	Internal model performance	47
1.6.1	Validation of underlying assumptions	48
1.6.2	Goodness of fit criteria	48
1.6.3	Robustness and internal predictivity	49
1.7	External performance	51
1.7.1	Global external performance	51
1.7.2	Local external performance	52
1.8	Applicability domain	53
1.8.1	Definition of the applicability domain	53
1.8.2	Applicability domain based on the ranges	54
1.8.3	Applicability domain based on the convex hull	55
1.8.4	Applicability domain based on the Euclidean, Mahalanobis or L_1 distance	55
1.8.5	Applicability domain based on leverages	57
1.8.6	Applicability domain based on DModX	57
1.8.7	Applicability domain based on the Hotelling's T^2	60
1.8.8	Applicability domain based on the density	60
1.8.9	Comparison of different applicability domains on a QSAR model example	61
1.9	Conclusion	61
2	Assessing prediction error in Quantitative Structure-Activity Relationship (QSAR) models	75
2.1	Introduction	75
2.2	Expected Prediction Error (EPE)	76
2.3	Bias-variance tradeoff	78
2.4	Estimates of the global Expected Prediction Error	80
2.4.1	Model selection: AIC and BIC	81
2.4.2	Model assessment: cross-validation and bootstrap	82
2.5	Estimates of $V[\hat{E}(Y \mathbf{x}_0)]$, the prediction variance for a new data point	85
2.5.1	Bootstrap estimate of $V[\hat{E}[Y \mathbf{x}_0]]$	86
2.5.2	Estimate of $V[\hat{E}[Y \mathbf{x}_0]]$ for MLR	86
2.5.3	Estimate of $V[\hat{E}[Y \mathbf{x}_0]]$ for PLSR	87
2.5.4	Estimate of $V[\hat{E}[Y \mathbf{x}_0]]$ for RT	88

2.5.5	Bootstrap estimate <i>versus</i> analytical formula?	89
2.6	Estimates of $Cov[\hat{E}(Y \mathbf{x}_1), \hat{E}(Y \mathbf{x}_2)]$, the covariance of two data points estimations	89
2.6.1	Estimate of $Cov[\hat{E}[Y \mathbf{x}_1], \hat{E}[Y \mathbf{x}_2]]$ for MLR	89
2.6.2	Estimate of $Cov[\hat{E}[Y \mathbf{x}_1], \hat{E}[Y \mathbf{x}_2]]$ for PLSR	89
2.6.3	Estimate of $Cov[\hat{E}[Y \mathbf{x}_1], \hat{E}[Y \mathbf{x}_2]]$ for RT	90
2.7	Conclusion	90
3	Multicriteria optimisation using a desirability index	93
3.1	Introduction	93
3.2	Desirability function and desirability index	94
3.2.1	Classical desirability functions	95
3.2.2	New Normal <i>cdf</i> based desirability functions	95
3.2.3	Comparison of desirability functions	98
3.3	Stochastic approach to multicriteria optimisation	99
3.4	Distribution of the desirability index	100
3.4.1	<i>Pdf</i> of the desirability function $d(Y) \mathbf{x}$	100
3.4.2	<i>Pdf</i> of the weighted desirability function $(d(Y))^w \mathbf{x}$	103
3.4.3	<i>Pdf</i> of the weighted desirability function $w \cdot d(Y) \mathbf{x}$	107
3.4.4	Distribution of the weighted geometric mean of DFs	109
3.4.5	Distribution of the weighted arithmetic mean of DFs	110
3.4.6	Distribution of the minimum of DFs	111
3.5	Distribution of DI when models are estimated	112
3.6	Conclusion	113
4	Uncertainty propagation on desirability index	117
4.1	Introduction	117
4.2	Estimation of the expected desirability index $D^S(\mathbf{x}) = E[D(\mathbf{Y}) \mathbf{x}]$	120
4.3	Quantification of the uncertainty of the expected desirability index $\hat{D}^S(\mathbf{x})$	122
4.4	Set up of an Indistinguishable Optimal Zone (IOZ)	124
4.4.1	Objective of an Indistinguishable Optimal Zone (IOZ)	124
4.4.2	Hypothesis test of $H_0 : D^S(\mathbf{x}) \geq D^S(\hat{\mathbf{x}}_{opt}^S)$ for a fixed point $\mathbf{x}$	125
4.4.3	Multiple tests over χ	127
4.4.4	Discussion	129
4.5	Prediction interval for the desirability index $D(\mathbf{x})$	130
4.6	Extension to correlated models residuals	130
4.7	Illustrative example	132
4.7.1	Example description	132
4.7.2	Comparison of classical and stochastic use of DI	135
4.7.3	Confidence interval on $D^S(x)$	136

4.7.4	Indistinguishable Optimal Zone (IOZ)	138
4.8	Application to Derringer and Suich data and simulation study	140
4.8.1	Derringer and Suich classical optimisation	140
4.8.2	Stochastic optimisation of the simplified Derringer and Suich problem	141
4.8.3	Simulations on the simplified Derringer and Suich problem	147
4.8.4	Simulations study around the simplified Derringer and Suich problem	154
4.9	Conclusion	166
4.10	Appendices	168
5	A fast exchanges algorithm for lead optimisation	177
5.1	Introduction	177
5.2	Quantifying the quality of a set of m molecules	180
5.3	A Weighted Exchanges Algorithm for Library Design (<i>WEALD</i>)	183
5.3.1	Principles	183
5.3.2	Exchanges methodology	184
5.3.3	Application on the <i>antidepressant library</i>	185
5.3.4	What happens when applying the <i>WEALD</i> algorithm without using selection probabilities?	186
5.4	Comparison with the <i>Ultrafast</i> algorithm	187
5.5	Comparison with the <i>Piccolo</i> algorithm	191
5.6	Algorithms comparison on a second combinatorial library	194
5.7	Simulations with different combinatorial library dimensions	195
5.8	<i>WEALD</i> as a tool to search for the best compounds in a combinatorial library	197
5.9	Conclusion	199
6	An integrated methodology for virtual molecules optimisation with an application to the screening of the <i>diabetes</i> combinatorial library	205
6.1	Introduction	205
6.2	Summary of the integrated methodology for virtual molecules optimisation	206
6.2.1	Problem description	206
6.2.2	QSAR models development	207
6.2.3	Molecules optimisation	207
6.3	Application of the integrated methodology to optimise the <i>diabetes</i> combinatorial library	208
6.3.1	Problem description	209
6.3.2	QSAR models development	210
6.3.3	Molecules optimisation	220
6.3.4	Analysis of the whole combinatorial library	226
6.4	Conclusion	227
6.5	Appendices	229

INTRODUCTION:

context, goals and contents of the thesis

0.1 Context: drug discovery

The medical research of the last century has an enormous impact on our quality of life. As it can be seen on Figure 1, life expectancy has significantly increased in the United States (like in other countries) since the year 1900. The drugs developed by pharmaceutical companies are one of the main cause for this increase in longevity.

According to the United State Code, *drugs* are defined as:

- (A) articles recognised in the official United States Pharmacopoeia, official Homeopathic Pharmacopoeia of the United States, or official National Formulary, or any supplement to any of them; and
- (B) articles intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease in man or other animals; and
- (C) articles (other than food) intended to affect the structure or any function of the body of man or other animals; and
- (D) articles intended for use as a component of any article in clause (A), (B), or (C).

More traditionally, a *drug* in the sense of a *medicine* is defined as a substance that prevents, cures or treats a disease. That is what pharmaceutical companies are looking for.

Ideally, a drug must be safe and effective. It should also be orally well absorbed by the organism and bioavailable¹. In addition, a drug needs to be metabolically stable to main-

¹*Bioavailability* refers to the rate and extent to which an active drug enters the general circulation, permitting access to the site of action.

tain a reasonably long half-life² and should distribute selectively to the target tissue(s). Finally, an ideal drug should be non-toxic and cause minimal or no adverse effects. Finding compounds that possess all these desirable features is a challenging task!

Throughout history, people have discovered by trial and error which berries, roots and barks could be used for medicinal purposes to reduce symptoms of illness (Verpoorte 1998). Some natural products are still in use today as these explained below and represented on Figure 2.

Snowdrop contains *galanthamine* used probably since the Antiquity, as written by Homer in the Odyssey, to protect Ulysses from the poisoned beverages of the magician. According to medicinal literature, in the 1960s, galanthamine was first used in neurological and neuromuscular indications (Venturi *et al.* 1965). Now it has been proved that galanthamine significantly improves cognitive function and slow functional decline in patients with Alzheimer's disease (Wilcock *et al.* 2000).

Morphine was discovered in 1809 and has a variety of useful effects like analgesia, euphoria and cough suppression. Morphine is a naturally occurring substance in the opium poppy, *Papaver Somniferum* (Marculescu 2001).

The *vincristine* was found in the 1960s, isolated from *Catharanthus Roseus* (Madagascar Periwinkle). It is given as a treatment for some types of cancer including leukemia, lymphoma, breast and lung cancer (Haim *et al.* 1994).

Fortunately, it has come a long way since the day of extracting roots and chewing leaves were the only medicinal preparations. Now drug discovery is inevitably a team effort requiring constant interactions based on experimental findings, planning and synergies across many different scientific disciplines (biology, chemistry, informatics, statistics, . . .) and marketing.

The long process of creating/discovering and evaluating drugs for the safe and effective treatment of human disease is referred as *the drug discovery and development process*. This long process and the high percentage of failed projects mean that each new drug requires between 10 to 15 years to be developed. This long development costs on average \$400 million and sometimes more than \$1 billion. Between 50 000 and 5 000 000 compounds are often screened to find a few (less than 10) drug candidates and finally one new drug.

²The *half-life* is the time required for the potency of a drug to fall to half of its potency or to be eliminated from the body.

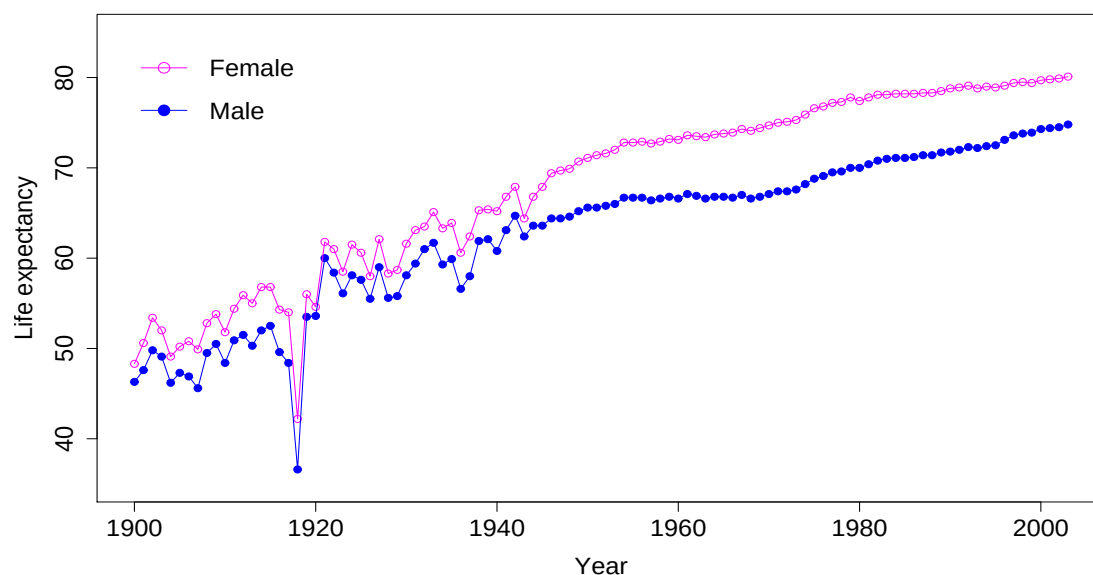
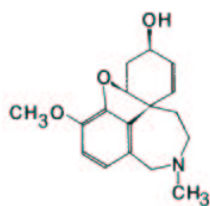


Figure 1: Increase of life expectancy in the United States of America according to the data of the National Center for Health Statistics (<http://www.cdc.gov/nchs/fastats/lifexpec.htm>)



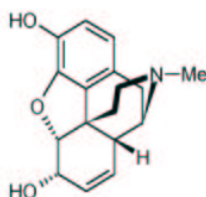
SNOWDROP

GALANTHAMINE



PAPAVER SOMNIFERUM

MORPHINE



CATHARANTHUS ROSEUS

VINCISTINE

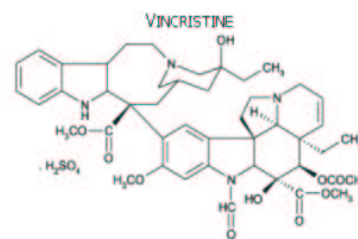


Figure 2: Example of natural products still in use today. On the left, galanthamine extracted from snowdrop. In the middle, morphine extracted from Papaver Somniferum. On the right, vincristine extracted from Catharanthus Roseus.

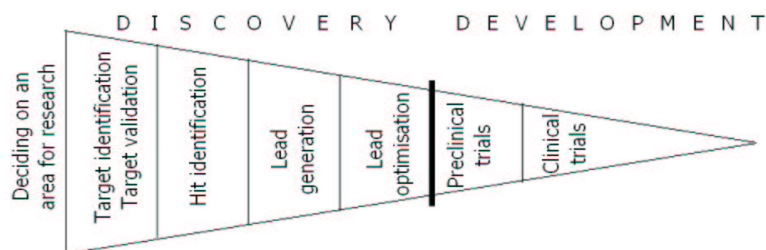


Figure 3: General scheme of drug discovery and development process.

Drug discovery and development process is often depicted as a multi-stage process summarised in Figure 3. It is divided in two main parts (not always distinct in practice):

- (i) the discovery of a drug candidate by selecting a target and finding an active compound toward this target by screening of large collections of compounds
- (ii) and the long development to confirm the effects of the new drug.

This thesis falls under the first part of this process. The different steps of this long evolution are summarised in the next sections. For a more detailed review on drug discovery and development, see Neamati (2002); The Global Biodiversity Institute / International Institute of Tropical Agriculture (2000); Ohlstein *et al.* (2000); Cunningham (2000); Debouck and Metcalf (2000); Drew (2000).

0.2 Deciding on an area for pharmaceutical research

There are important diseases for which no efficient treatment has yet been found. There are also illnesses for which efficient drugs exist but have undesirable side effects. According to Minino *et al.* (2006), there were 2 398 365 deaths registered in 2004 in the United States. The 10 leading causes were:

- Heart disease: 654 092
- Cancer: 550 270
- Stroke (cerebrovascular diseases): 150 147
- Chronic lower respiratory diseases: 123 884
- Accidents (unintentional injuries): 108 694
- Diabetes: 72 815
- Alzheimer's disease: 65 829
- Influenza/Pneumonia: 61 472
- Nephritis, nephrotic syndrome, and nephrosis: 42 762
- Septicemia: 33 464

These statistics clearly demonstrate that a lot of work has yet to be done in some areas.

Before starting any new project, pharmaceutical companies must wonder about the present medical needs and especially what the severity of the life conditions for the targeted disease . Indeed patients may not bother buying drug for conditions that are minor, but if it is life threatening then the wallet comes out more quickly.

Pharmaceutical companies must also take into account the existing therapies. What are the drawbacks of these current therapies that a new drug could improve (poor efficiency, side effects, large costs,...)?

According to competitors activity, pharmaceutical compagnies must decide whether to seek improvements within an existing drugs class or follow a novel mechanistic approach.

And last but not least, the new product must of course be commercially viable, taking into account the patients number, the duration of the proposed therapy, the price at which the new drug would be sold. This depends mainly on the price of the drugs in the same area and of the investment during the long drug discovery and development process to find the new drug, demonstrate its efficiency and safety and also its advantages against other therapies.

In practice, finding a new drug is so complex that most of the time each pharmaceutical company carries on working on the same area as the past few years, avoiding extra expenses to understand a new targeted disease.

0.3 Drug discovery stages

0.3.1 Target identification and validation

A *target* is a body molecule (usually a protein encoded by a gene) that is thought to play a fundamental role in the onset or progression of a particular disease.

Pharmaceutical companies have to find targets against which they can then develop drugs. Finding a good target is a crucial step in the drug discovery process. Still now, a significant number of projects fail in the clinical stages because a wrong target was chosen, leading to an enormous loss of time and money.

This important search involves two steps:

- *target identification* : identify molecules that clearly play a role in a disease process;

- *target validation* : determine whether intervening on this target will affect an appropriate biological response.

These two steps are quickly described bellow. For a more precise description, see Butcher (2003).

Target identification

There are several ways to identify targets. Information pharmaceutical companies gather about the disease can give an insight into the normal and pathological state of a cell or organism and help to formulate hypotheses about targets for intervention.

Until recently, pharmaceutical researchers were limited to study around 500 known targets. The number of available biological targets is being vastly expanded with the publication of the entire human genome sequence (Venter *et al.* 2001; International Human Genome Sequence Consortium 2001). The key issue is therefore to identify those genes that are involved in the susceptibility or progression of human diseases. This is the objective of *genomics*. Genomics greatly helps the process of target identification (Ohlstein *et al.* 2000; Cunningham 2000).

Statistics has an important role to play in that field (Walker 2001): gene profiles or protein profiles from diseased and healthy tissues can be fractionated and compared to see which ones are involved in a particular disease. An undeniably powerful method for studying differential gene expression is the use of *microarrays* (Speed 2003). Microarrays are simply small glass slides upon which thousands of genes are arrayed (usually between 500 and 20 000) and compared.

Target validation

Having formulated the hypothesis that intervention at a certain pathway may lead to a desired biological effect, biologists and chemists have to prove it by *in vitro* testing or *in vivo* testing (often in small rodents).

A well-known method is the one based on *targeted gene disruption*. Targeted gene disruption relates to production of knockout or transgenic animals to study the effect of removing a particular gene coding for the putative molecular target. See Nicole *et al.* (1998) for an example of targeted gene disruption application. For a recent review on target validation techniques, see Szymkowski (2003).

When a target has been validated, pharmaceutical companies can start searching compounds that are active against this target. This is the *hit identification* stage.

0.3.2 Hit identification

A *hit* is a compound that interacts with a target. Such a compound can be derived from a number of techniques: natural product screening, high throughput screening, structure based design, ...

There are good reasons to search for a hit in *natural products* (Irth and Van Elswijk 2002): many existing drugs are based upon natural products and those molecules are often structurally diverse. Iterative biological tests are used to predict the efficiency of tried substance. Simple and cheap tests are first used to eliminate non-active compounds and further tests are done on compounds that have passed the first tests. Often *in vitro* tests precede *in vivo* tests. In practice, screening all natural compounds is of course impossible and a hit can only be found by a good mixture of screening, popular knowledge and ... chance.

High throughput screening (Rubenstein and Coty 2001; Bajorath 2002) is the method by which very large numbers of compounds (100 000 – 500 000) are screened against a putative drug target in either cell-free (*in vitro*) or whole-cell (*in vivo*) assays. Those compounds come from past projects or from *combinatorial libraries*. A combinatorial library is a huge collection of compounds obtained by a systematic combination of reagents (complex product, protein, or ligand) (Dean and Lewis 1999).

Another popular technique to identify hit compounds is *structure based design* (Bailey and Brown 2001). The 3-dimensional structure of the tested target is determined thanks to modern computers and specialised softwares. Structural information is used for building completely novel molecules (*de novo design*) or to search in molecular databases for molecules complementary to the target.

0.3.3 Lead generation

Once a promising hit has been identified, diverse combinatorial libraries are designed to explore the structural activity space around that compound and refine its pharmacological profile to produce leads. A lead is an active compound with improved chemical characteristics (*e.g.* potency and selectivity). Lead generation involves the chemical modification of hits by repeated cycles of synthesis and testing of analogs to produce leads thereby increasing their suitability as potential drug. For more details about the techniques used in lead generation, see Bleicher *et al.* (2003).

0.3.4 Lead optimisation

Additional repeated modifications are made to the leads to produce drug development candidates by optimising some drug characteristics: ADME properties (Absorption, Distri-

bution, Metabolism, Elimination), bioavailability, pharmacokinetics, solubility, blood/brain barrier permeability, ...

The most widely used method is based on screening large combinatorial libraries to find compounds that optimise these drugability properties. There is increasing demand to design computer programs that can accurately predict these physicochemical parameters and use them to perform virtual screening (Leach and Hann 2000), *i.e.* select compounds in the library according to their predicted features. In this way, chemists in laboratories focus on testing only promising compounds.

In this stage of the drug discovery process, statisticians have also a key role to play to provide as good prediction models as possible. Indeed, approximately 40% of drugs in clinical trials are discarded because they do not show the correct ADME properties. Those kinds of models are known under the name *Quantitative Structure Activity Relationship* (QSAR). There is an increasing interest in statistical modeling tools (regression, neural networks, tree, ...) that can be used in QSAR and also some questioning about an efficient technique of data points collections (experimental design) on which models may be estimated.

The development and statistical validation of QSAR models and techniques to select compounds in the combinatorial library that optimise predicted characteristics are the two major concerns of this thesis. An integrated methodology is developed to take QSAR models uncertainty into account in a new algorithm aimed to select the most promising compound(s) out of a huge combinatorial library. The contents of the thesis are detailed in Section 5.

The drug discovery process can take three to five years and, if everything is positive and promising, the project is ready to pass on to the development phases where the effect of the drug candidate will be tested in humans.

0.4 Drug development phases

Before a drug can be tested on humans, the drug candidate is administered to animals (mice, rats, dogs, monkeys) daily for a shorter period in order to prove that the drug is not carcinogenic (causes cancer), mutagenic (causes mutations) or teratogenic (causes malformations) and also to understand how the drug is absorbed and excreted. This is the preclinical phase. For more details about this phase, see The European Agency for the Evaluation of Medicinal Products (1997).

Once the pharmaceutical company proves that the drug appears to be safe and possibly

effective in animals, a package containing information on the safety, activity and chemical properties of the drug candidate must be submitted to the FDA in the US (Food and Drug Administration) or to the EMEA in Europe (European Medicines Agency).

Doctors and researchers running a clinical trial develop a written plan (*protocol*) detailing exactly how the trials will be conducted. If the FDA agrees, the pharmaceutical company may proceed with clinical trials. These are normally organised in four distinct (not always) phases, each one having its own objective.

If the drug candidate behaved perfectly in phase I, II and III, the pharmaceutical company may receive the market authorisation. Clinical trials are performed under severe control of the FDA and any unexpected problem may stop them and create an enormous loss of time and money.

These different phases are explained with more details in the next sections and the main statistical contributions are emphasised. Indeed, statisticians are essential in planning the clinical trials and in analysing, interpreting and presenting the results of the trials. For more details about the use of statistics in the clinical phases see ICH E9 Expert Working Group (1999); Peace (1987); Bland (2000); Cleophas *et al.* (2002); Zwinderman *et al.* (2003); Curtis and Tonascia (1986).

0.4.1 Phase I

Aim: evaluation of safety and tolerability in healthy human volunteers

Number of involved subjects: less than one hundred

Duration: several months

Phase I consists of initial testing of the drug candidate in humans, usually in healthy volunteers and the study of its chemical action. Safety evaluations are the primary objective. Often the attempt is made to establish the approximate levels of patient tolerance for acute and usually multiple dosing. Sometimes also some preliminary efficiency data may be collected.

Statistical tools are used in pharmacokinetic (“what the body does to the drug”) studies and pharmacodynamic (“what the drug does to the body”) studies. Pharmacokinetics evaluates the evolution of the drug concentration in the body. Statistics is used to estimate from the data what is called the *maximum tolerable dose*. Pharmacodynamics relates the response to the circulation drug concentration. Statistical modelling is applied to the concentration-effect relationship.

About seventy percent of experimental drugs passes this initial phase of testing and enters phase II.

0.4.2 Phase II

Aim: evaluation of the drug efficiency in patients

Number of involved subjects: up to several hundreds

Duration: from several months to two years

Phase II studies are designed to evaluate and prove efficiency when the drug candidate is administrated to patients (persons with the treated disease). During this phase, the therapeutic dose range is established as well as the dose-response curve, the route and frequency of compound administration. Efficiency is determined using placebo control. Eventually some short-term side-effects can yet be observed.

Statistical design (for instance the one proposed by Gehan (1961)) is applied to determine which sample sizes to use in the study. Hypothesis tests are performed about the truly anti-disease activity or when the drug is considered in combination with a standard therapy.

About one third of experimental drugs successfully completes both phase I and phase II studies.

0.4.3 Phase III

Aim: comparison of the new drug *versus* placebo or other existing treatments

Number of involved subjects: several hundreds to several thousands

Duration: two to three years

Phase III is the last stage to obtain the market authorisation. It is often described in two parts: phase IIIa and phase IIIb.

In phase IIIa, a broad and representative patients population is enrolled to test for long-term safety and explore if any rare adverse event occurs after long-term use. This study usually consists in comparing the drug to existing treatment(s). The purpose is, apart from safety, to establish data for superiority and regards efficiency. In parallel, a number of clinical pharmacology studies will be executed to test the drug among interaction with other drugs, food, ... Study reports will be made to demonstrate safety and efficiency. It is submitted to health authorities for marketing authorisation approval.

Then phase IIIb is initiated. This involves the study of efficiency and safety in specific subpopulation (children, elderly, pregnant...).

Statistics is even more important in this third phase. The main statistical tools used in this phase are the design of the treated group and control group (cross-over design, sequential design, parallel design, stratified or unstratified design...) as well as the choice of their sample size (often chosen to be able to detect a clinically meaningful difference between treatments or to estimate treatment difference with enough precision). Some more hypotheses tests are performed.

Seventy to ninety percent of drugs that enter phase III studies successfully complete this phase of testing.

0.4.4 Phase IV

Aim: long-term surveillance and search for new indications

Specific safety concerns can be demonstrated in the more long term trials of phase IV. These are post-marketing studies to elucidate common side effects and to gather data on additional indications for the drug candidate.

0.5 Contents of the thesis

Combinatorial chemistry is widely used in the drug discovery process (Gallop *et al.* 1994; Gordon *et al.* 1994; Floyd *et al.* 1999; Lam 1997; Agrafiotis *et al.* 2002; Leach and Hann 2000; Venton and Woodbury 1999). This technique systematically combines a variety of molecular building blocks into a huge collection of compounds called *combinatorial library*. Combinatorial chemistry can be used in the lead optimisation stage to automatically modify the lead and search for the most promising modified compound.

Once a lead has been generated, chemists in pharmaceutical companies try to chemically modify the lead to obtain compounds having still better properties. They virtually divide the lead into N basic parts and set up a list of chemicals that could react on each part. Each chemical used in reaction with the lead is called a *reagent* (or *reactant*) and each set of reagents on a part of the lead is called a *R-group* (or a *pool*) (Agrafiotis 2002; Agrafiotis and Lobanov 2000; Zheng *et al.* 2000).

Let N be the number of R-groups and N_i the number of reagents in the i^{th} R-group ($i = 1, 2, \dots, N$), $R_i = \{R_{ik}; k = 1, 2, \dots, N_i\}$. The number of possible compounds obtained by combining a reagent from each R-group is $M = \prod_{i=1}^N N_i$. They form the combinatorial

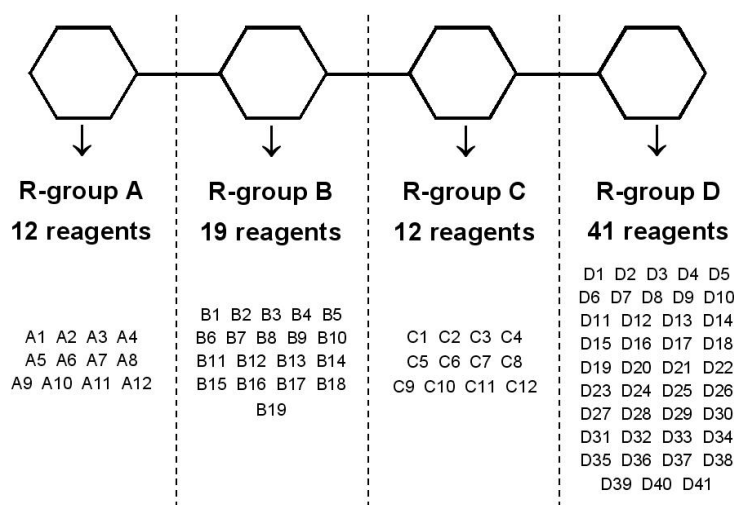


Figure 4: *antidepressant library*: real combinatorial library composed of 4 R-groups and $12 \times 19 \times 12 \times 41 = 112\,176$ compounds.

library. The main objective of the chemists is to find the molecule that optimises some drugability properties. Another more general objective may be to find a subset of $m > 1$ good molecules, either the m best molecules, or the best combinatorial sub-library of size $m = \prod_{i=1}^N n_i$ obtained by combining all the n_i selected reagents in the i^{th} R-group. The combinatorial nature of the compounds in the sub-library allows an easy synthesis as far as their number is not too large compared to used plates.

Two combinatorial libraries are used to illustrate this thesis. They have been provided by Eli Lilly and Company. The first one has been developed in the context of a new serotonergic antidepressant and is referred as the *antidepressant library*. The library is composed of $N = 4$ R-groups. The different R-groups are represented on Figure 4. There are $N_1 = 12$, $N_2 = 19$, $N_3 = 12$ and $N_4 = 41$ reagents in each R-group. The size of the whole combinatorial library is $M = \prod_{i=1}^N N_i = 12 \times 19 \times 12 \times 41 = 112\,176$.

The second combinatorial library has been built up in the context of the development of a new molecule acting on glucose regulation in type II diabetes. It will be further referred as the *diabetes library*. This combinatorial library is composed of three R-groups, with respectively 47, 50 and 47 reagents leading to $M = 47 \times 50 \times 47 = 110\,450$ possible compounds.

In most cases, the size M of the combinatorial library is so huge that it is impossible to synthesise and test all the compounds in the library to find the best one. **This thesis proposes a methodology to virtually screen the combinatorial library and retain**

the most promising compound(s).

The quality of a molecule is, most of the time, not only measured by only one response but by a set of $P > 1$ responses. It is first essential to define precisely all the drugability properties of interest and how they can be measured. The optimisation of molecules is then, by definition, a multicriteria optimisation problem. **The methodology proposed in the thesis can be applied to any other multicriteria optimisation problem, often occurring in industry.**

The different elements of this methodology are the modelling of each optimised property, the quantification of molecules quality by a desirability index and the search of the most promising molecule(s) with an efficient algorithm. They are explained next in a few words. The importance of taking model prediction error into account is pointed out.

0.5.1 QSAR modelling

The central emerging technique that allows to speed up the drug discovery process is the development of QSAR models. For each drugability property Y_i ($i = 1, 2, \dots, P$) that has to be optimised, a QSAR model of the form $Y_i = f_i(\mathbf{x}, \boldsymbol{\beta}_i) + \epsilon_i$ is assumed to be the true relationship between the activity Y_i and a set of K chemical descriptors $\mathbf{x} = (x_1, x_2, \dots, x_K)$ characterising a given molecule of interest. f_i is a link function, $\boldsymbol{\beta}_i$ is a vector of parameters and ϵ_i is a random error term. Multiple linear regression, partial least squares regression and regression tree are the three most often used classes of models in QSAR.

There exists a lot of features to chemically describe a molecule: the molecular weight, the number of hydrogen atoms, some measure of volume. . . Most of them can be computed using specific softwares. Up to 8000 chemical descriptors can be computed thanks to the *SMILES* coding (ASCII format) of a molecule.

Each theoretical model is fitted on experimental data collected in laboratories and then validated to insure a good predictive power. The fitted models $\hat{Y}_i = \hat{f}_i(\mathbf{x}, \hat{\boldsymbol{\beta}}_i)$ can then be applied to predict the optimised properties for any molecule in the combinatorial library, after checking that this new molecule is not too different from the ones used to train the models.

Chapter 1 details the development of QSAR models, the choice of the drugability properties and the descriptors. It also points out the increasing concern about the statistical validation of QSAR models and the definition of an applicability domain. Prediction error for new compounds is usually neglected in QSAR literature. Chapter 2 presents the different

sources of error when predicting the activity of new compounds with QSAR models. Some estimators are reviewed from the statistical literature to quantify those different sources with multiple linear regression, partial least squares regression or regression tree models.

0.5.2 Multiresponse optimisation criterion

It's improbable that one molecule in the combinatorial library is the best for all optimised drugability properties Y_i , $i = 1, 2, \dots, P$. In front of such multiresponse optimisation problem, the literature often proposes to aggregate the multiple responses into a unique value between 0 and 1 that is a compromise of all their values. This is the *desirability* principle introduced by Harrington (1965).

First, all drugability properties are rescaled using *desirability functions*: $d_i(Y_i) \in [0, 1]$ for $i = 1, 2, \dots, P$. The most often used desirability functions are the ones proposed by Derringer and Suich (1980), presented on Figure 5. Then, the desirability functions are aggregated by a *desirability index* in a single value between 0 and 1. Derringer and Suich proposed, for instance, to use the geometric mean: $D = (\prod_{i=1}^P d_i(Y_i))^{1/P}$.

The desirability index can be used to balance the P modeled drugability properties and rank the molecules of the combinatorial library. The optimal molecule is the one that maximises the desirability index.

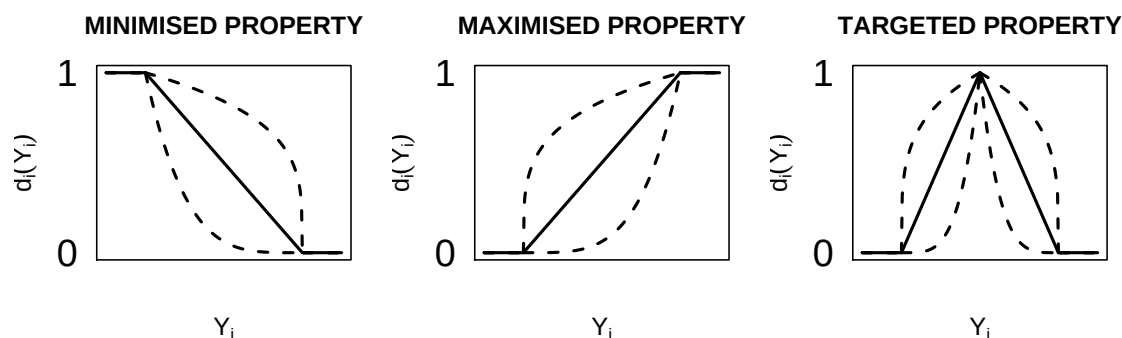


Figure 5: Different kinds of desirability functions d_i proposed by Derringer and Suich (1980). The first and the second graphs show decreasing and increasing functions that can be used when the i^{th} property has to be minimised or maximised respectively. The last graph shows desirability functions d_i that can be used when the i^{th} property must reach an optimal value.

Chapter 3 reviews desirability functions and desirability indexes developed in the context of multiresponse optimisation. New and easy-to-use differentiable desirability functions

based on the Normal distribution are also presented. In this chapter, a new stochastic optimisation considers the desirability index as a random variable because, for a given molecule with descriptors $\mathbf{x}$, each response $Y_i|\mathbf{x}$ ($i = 1, 2, \dots, P$) is a random variable. Its analytical distribution is derived in order to compute its expectation $E[D(\mathbf{Y})|\mathbf{x}]$.

We propose, in Chapter 4, a general approach based on the Delta method to estimate $E[D(\mathbf{Y})|\mathbf{x}]$ from the fitted QSAR models, encompassing different papers on the treatment of desirability index uncertainty (Steuer 2000; Cahya 2002; Weber and Weihs 2003; Trautmann and Weihs 2004; 2006). Moreover, an asymptotic confidence interval allows to quantify the propagation of the models prediction error on the estimation $\hat{E}[D(\mathbf{Y})|\mathbf{x}]$ and the definition of the *indistinguishable optimal zone* allows to quantify the uncertainty of the corresponding optimum by gathering not significantly worse solutions than this optimum found. Those two tools constitute really a great improvement as, in multiresponse optimisation literature, desirability indexes are traditionally compared assuming they are exact and the reliability of the optimum found by maximising this estimated index is not questioned. Under some general conditions, our new approach can be applied to molecules optimisation and also to other multicriteria optimisation.

0.5.3 Optimisation algorithm

If χ denotes the descriptors space corresponding to the molecules in the combinatorial library, the lead optimisation problem can be written

$$\arg \max_{\mathbf{x} \in \chi} E[D(\mathbf{Y})|\mathbf{x}] .$$

In practice, for any molecule, the expected desirability is unknown and can be estimated by $\hat{E}[D(\mathbf{Y})|\mathbf{x}]$ as proposed in Chapter 4.

If the number of descriptors in the QSAR models is quite huge, their computation may take a long time. With most combinatorial library, it is not possible to compute descriptors, predict activities and estimate the expected desirability automatically for each compound. There is a clear need for algorithms to explore partially but efficiently a combinatorial library.

In Chapter 5, a new algorithm called *WEALD* (Weighting Exchange Algorithm for Library Design) is proposed to screen a combinatorial library searching for the most promising molecule(s). This algorithm starts with a molecule chosen at random and iterates using exchanges between reagents until no more increase in desirability can be obtained. A subtle weighting of reagents allows a faster and more efficient exploration of the combinatorial library compared to other existing algorithms like the *Ultrafast algorithm* (Agrafiotis and

Lobanov 2000) and the *PICCOLO algorithm* (Zheng *et al.* 2000).

The new integrated methodology including QSAR modelling techniques, desirability index use and optimisation of molecules using the *WEALD* algorithm, is illustrated in Chapter 6 on the diabetes library provided by Eli Lilly and Company. The importance of taking models prediction error into account in this molecules optimisation is emphasised.

Bibliographical references are listed at the end of each chapter and not gathered at the end of the thesis as the contents are quite different from one chapter to another.

0.6 Summary

This thesis develops an integrated methodology based on the desirability index and QSAR models to virtually optimise molecules. Statistical and algorithmic tools are proposed to search in huge collections of compounds obtained by combinatorial chemistry the most promising ones.

First, once the drugability properties of interest have been precisely defined, QSAR models are developed to mimic the relationship between those optimised properties and chemical descriptors of molecules. The literature on QSAR models is reviewed and the statistical tools to validate the models, analyse their fit and their predictive power are detailed. Even if a QSAR model has been validated and sounds highly predictive, we emphasise the importance of measuring extrapolation by the definition of its applicability domain and quantifying the prediction error for a given molecule. Indeed, QSAR models are often massively applied to predict drugability properties for libraries of new compounds without taking care of the reliability of each individual prediction.

Then, a desirability index measures the compromise between the multiple estimated drugability properties and allows to rank the molecules in the combinatorial library in preference order. The propagation of the models prediction error on the desirability index is quantified by a confidence interval that can be constructed under general conditions for linear regression, PLS regression or regression tree models. This fulfills an important lack of the desirability index literature that considers it as exact.

Finally, a new efficient algorithm (*WEALD*) is proposed to virtually screen the combinatorial library and retain the molecule with the highest desirability indexes. For each explored molecule, it is checked if it belongs to the applicability domain of each QSAR models. In addition, the uncertainty of the desirability index of each explored molecule is taken into account by gathering molecules that can not be distinguished from the optimal

one due to the propagation of QSAR models prediction error. Those molecules do not have a significantly smaller desirability than the optimal molecule found by *WEALD*. This constitutes another important improvement in the use of desirability index as a tool to compare solutions in a multicriteria optimisation problem.

This integrated methodology has been developed in the context of lead optimisation and is illustrated on a real combinatorial library provided by Eli Lilly and Company. This is the main application of the thesis. Nevertheless, as the results on desirability index uncertainty are applicable under general conditions, they can be applied to any multicriteria optimisation problem, like it often occurs in industry.

Bibliography

- Agrafiotis, D.K (2002), "Multiobjective Optimization of Combinatorial Libraries", *Journal of Computer-Aided Molecular Design*, 16, 335-356.
- Agrafiotis, D.K. and Lobanov, V.S. (2000), "Ultrafast algorithm for designing focused combinatorial arrays", *Journal of Chemical Information and Computer Sciences*, 40, 1030-1038.
- Agrafiotis, D.K., Lobanov, V.S. and Salemme, F.R. (2002), "Combinatorial informatics in the post-genomics era", *Nature reviews. Drug discovery.*, 1, 337-346.
- Bailey, D. and Brown, D. (2001), "High-throughput chemistry and structure-based design: survival of the smartest", *Drug Discovery Today*, 6, 57-59.
- Bajorath, J. (2002), "Integration of virtual and high-throughput screening", *Nature Reviews Drug Discovery*, 1, 882-894.
- Bland, M. (2000), *An Introduction to Medical Statistics (3rd. ed.)*, Oxford University Press.
- Bleicher, K.H., Bohm, H.J., Muller, K. and Alanine, A.I. (2003), "Hit and lead generation: beyond high-throughput screening", *Nature Reviews Drug Discovery*, 2, 369-378.
- Butcher, S.P. (2003), "Target discovery and validation in the post-genomic era", *Neurochemical research*, 28(2), 367-371.
- Cahya, S. (2002), *Sampling properties of optimal operating conditions of single and multiple response surface systems*, Thesis of the Pennsylvania State University.
- Cleophas, T.J., Zwinderman, A.H. and Cleophas, T.F. (2002), *Statistics applied to clinical trials (2nd ed.)*, Kluwer Academic Publishers.
- Cunningham, M.J. (2000), "Genomics and proteomics: the new millenium of drug discovery and development", *Journal of Pharmacological and Toxicological Methods*, 44(1), 291-300.

- Curtis, L.M. and Tonascia, S. (1986), *Clinical trials: design, conduct, and analysis (monographs in epidemiology and biostatistics, Vol 8)* , Oxford University Press.
- Dean, P.M. and Lewis, R.A. (1999), *Molecular diversity in drug design*, P.M. Dean & R.A. Lewis.
- Debouck, C. and Metcalf, B. (2000), "The impact of genomics on drug discovery", *Annual Review of Pharmacology and Toxicology* , 40, 193-207.
- Derringer, G.C. and Suich, R. (1980), "Simultaneous optimization of several response variables", *Journal of Quality Technology*, 12(4), 214-219.
- Drew, J. (2000), "Drug discovery today and tomorrow", *Drug Discovery Today*, 5, 2-4.
- The European Agency for the Evaluation of Medicinal Products (Human medicines evaluation unit) (1997), *Notes for guidance on preclinical safety evaluation of biotechnology-derived pharmaceuticals (CPMP/ICH/302/95)*, Oxford University Press.
- Floyd, C.D., Leblanc, C. and Whittaker, M. (1999), "Combinatorial chemistry as a tool for drug discovery", *Progress in medicinal chemistry*, 36, 91-168.
- Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gordon, E.M. (1994), "Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries", *Journal of Medicinal Chemistry*, 37(9), 1233-1251.
- Gehan, E.A. (1961), "The determination of the number of patients in a preliminary and a follow-up : trial of a new chemotherapeutic agent", *Journal of Chronic Diseases*, 13, 346-353.
- The Global Biodiversity Institute / International Institute of Tropical Agriculture (1-24 March 2000), *Training course on biodiversity, biotechnology, and law*, Ibadan, Nigeria.
- Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gallop, M.A. (1994), "Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic-synthesis, library screening strategies, and future-directions", *Journal of Medicinal Chemistry*, 37(10), 1385-1401.
- Haim, N., Epelbaum, R., Ben-Shahar, M., Yarnitsky, D. (1994), "Full dose vincristine (without 2-mg dose limit) in the treatment of lymphomas", *Cancer*, 73, 2515-2519.
- Harrington, J. (1965), "The desirability function", *Industrial Quality Control*, 21(10), 494-498.
- ICH E9 Expert Workin Group (1999), "Statistical principles for clinical trials: ICH harmonised tripartite guideline", *Statistics in Medecine*, 18, 1905-1942.

- International Human Genome Sequence Consortium (2001), "Initial sequencing and analysis of the human genome", *Nature*, 409, 860-921.
- Irth, H. and Van Elswijk, D.A. (2002), "Back to the roots: nature as a source of novel drug structures", *European Biopharmaceutical Review*
- Lam, K.S. (1997), "Application of combinatorial library methods in cancer research and drug discovery", *Anti-cancer Drug Design*, 12(3), 145-167.
- Minino, A.M., Heron, M.P. and Smith, B.L. (2006), "Deaths: Preliminary Data for 2004", *National Vital Statistics Reports*, 54(19), June 28, 2006.
- Leach, A.R. and Hann, M.M. (2000), "The in silico world of virtual libraries", *Drug Discovery Today*, 5(8), 326-336.
- Marculescu, A. and Bobit, D. (2001), "Studies on the morphine content of L", *Roumanian Biotechnology Letters*, 6(5), 403-409.
- Neamati, N. and Barchi, J.J. (2002), "New paradigms in drug design and discovery", *Current Topics in Medicinal Chemistry*, 2, 211-227.
- Nicole, J., Rauchenberger, R., Hacker, U., Kast, T. and Maniak, M. (1998), "Targeted gene disruption reveals a role for vacuolin B in the late endocytic pathway and exocytosis", *Journal of Cell Science*, 111, 61-70.
- Ohlstein, E.H., Ruffolo, R.R. and Elliott, J.D. (2000), "Drug discovery in the next millennium", *Annu. Rev. Pharmacol. Toxicol.*, 40, 177-191.
- Peace, K.E. (1987), *Biopharmaceutical Statistics for Drug Development*, Marcel Dekker.
- Rubenstein, K. and Coty, C. (2001), *High throughput screening: redefining the mission (2nd edition)*, Drug & market development publications.
- Speed, T.P. (2003), *Statistical analysis of gene expression microarray data*, Chapman & Hall.
- Steuer, D. (2000), *An improved optimisation procedure for Desirability Indices*, Technical Report 27/00, SFB 475, Dortmund University.
- Szymkowski, D. (2003), "Hitting the target", *Nature*, 422, 341-347.
- Trautmann, H. and Weihs, C. (2004), *Uncertainty of the optimum influence factor levels in multicriteria optimization using the concept of desirability*, Technical Report 23/04, SFB 475, Dortmund University.

- Trautmann, H. and Weihs, C. (2004), *On the Distribution of the Desirability Index using Harrington's Desirability Function*, *Metrika*, 63, 207-213.
- United State Code, Title 21, Chapter 9, *Federal Food, Drug and Cosmetic Act Definitions Sections* (g)(1)
- Venter, J.C. *et al.* (2001), "The sequence of the human genome", *Science*, 291, 1304-1351.
- Venton, D.L. and Woodbury, C.P. (1999), "Screening combinatorial libraries", *Chemometrics and Intelligent Laboratory Systems*, 48(2), 131-150.
- Venturi, V.M., Piccinin, G.L. and Taddei, I. (1965), "Pharmacognostic of self-snow *Galanthus nivalis* (var. *gracilis*) in Italy", *Bollettino - Societa Italiana Biologia Sperimentale (Napoli)*, 41(11), 593-597.
- Verpoorte, R. (1998), "Exploration of nature's chemodiversity: the role of secondary metabolites as leads in drug development", *Drug Discovery Today*, 3, 232-238.
- Walker, M.G. (2001), "Pharmaceutical target identification by gene expression analysis", *Mini Reviews in Medicinal Chemistry*, 1, 197-205.
- Weber, H. and Weihs, C. (2003), *On the distribution of the desirability index using Harrington's desirability function*, *Forschungsbericht 2003/3*, Statistics Department, Dortmund University.
- Wilcock, G.K., Lilienfeld, S. and Gaens, E. (2000), "Efficacy and safety of galantamine in patients with mild to moderate Alzheimer's disease: multicentre randomized controlled trial", *British Medical Journal*, 321 (dec.), 1-7.
- Zheng, W., Hung, S.T., Saunders, J.T. and Seibel, G.L. (2000), "PICCOLO: a tool for combinatorial library design via multicriterion optimization", *Pacific Symposium on Biocomputing*, 588-599.
- Zwinderman, A.H., Cleophas, T.F., Cleophas, T.J.M. and Schwab, C.R. (2003), *Statistics Applied to Clinical Trials Self-Assessment Book*, Kluwer Academic Publishers.

Chapter 1

Quantitative Structure-Activity Relationship (QSAR) models: review and perspectives

This work constitutes the discussion paper DP0719 by Le Bailly de Tillegem and Govaerts (2007) of the Institute of Statistics, Université catholique de Louvain, Belgium.

1.1 History of QSAR models

Quantitative Structure-Activity Relationships (QSARs) are mathematical models approximating the often complex link between chemical properties and biological activities of compounds. They are used in many different fields such as agrochemistry, pharmaceutical chemistry or toxicology. This chapter reviews the existing literature about QSAR modelling, more particularly concerning the statistical tools used to evaluate their quality. Those are quite basic statistical tools, not necessarily appropriate and not necessarily well applied in practice but, the aim of this chapter is to present the state of the art about statistics in QSAR.

The development of a QSAR is based on the assumption expressed more than a century ago by Crum-Brown and Fraser (1868), that the physiological action of a substance is a function of its chemical composition. This leads to the idea that similar structures have similar biological properties and a small change in chemical structure is accompanied by a proportionally small shift in biological activity.

During the following hundred years, researchers tried to formalise some of those relationships. Richardson (1868) showed that the toxicities of ethers and alcohols were inversely

related to their water solubility. Richet (1893) demonstrated a relationship between the narcotic¹ effect of alcohols and their molecular weight and Overton (1897) and Meyer (1899) independently showed that the narcotic action of many compounds was dependent on their oil/water partition coefficients². Then, great strides were being made in the delineation of substituent effects on organic reactions, mainly by the work of Hammett (1937) and Taft (1952). Those contributions constitute together the mechanistic basis for the development of the QSAR paradigm in the 60's by Corwin Hansch, the first chemist that really succeeds in quantifying relationships between compound physicochemical properties and biological activities. His series of papers about the relationships of plant growth regulators and their dependency on Hammett constants³ and hydrophobicity laid the foundation of QSARs (Hansch *et al.* 1962; 1963; 1964; Fujita *et al.* 1964).

Since then, with the increasing knowledge in chemistry and the exploding power of computers, researchers are developing and using more and more QSAR models. QSARs development has now become an important branch of chemometrics, the science of the application of mathematical or statistical methods to chemical data.

There are two common objectives of QSARs (Eriksson *et al.* 2003). A first goal can be to understand the link between structure and activity and extract clues of which chemical properties are likely determinants for the biological activities of a compound. To achieve this objective, the models should be transparent and as simple as possible. A second aim can be to allow prediction of biological activity for untested and sometimes yet unavailable compounds. In this case, a QSAR model should have a good predictive power, whatever its complexity.

The modelling technique to apply in QSAR development depends on the kind of available data. A distinction is done by Livingstone (1995) between a SAR (Structure-Activity Relationship) and a QSAR (Quantitative Structure-Activity Relationship) model:

- A SAR is an association between a chemical substructure and a biological activity, a kind of structural alert. As an example, the presence of a carboxylic group or an amino group in a molecule is known to impart skin corrosion potential.
- A QSAR is a mathematical relationship between a quantitative measure of chemical structure, or a quantitative measure of a physicochemical property, and a biological activity. Two types of QSARs can be distinguished: classification models for which

¹**Narcotic** refers to a variety of substances that induced sleep.

²The **oil/water partition coefficient** is a measure of the hydrophobicity (repelled by water) and hydrophilicity (able to bond with water) of a substance.

³**Hammett constants** are measures of electron-withdrawing or electron-donating effects.

the response variable is on a categorical scale, and regression models, for which the response is continuous.

All statistical methods to model the link between a quantitative or a qualitative response and a set of quantitative and/or qualitative explanatory variables can then be applied to QSAR data. The most popular techniques for quantitative endpoints are simple or multiple linear regression, principal components regression, partial least squares regression, as well as regression trees and neural networks. If the activity property is qualitative, discriminant analysis, decision trees or distance-based similarity analysis are often applied to model its relationship with the physicochemical explanatory variables (Jaworska *et al.* 2003).

The types of statistical models used in QSAR as well as the techniques to assess their qualities depend on the nature of the response. In this chapter, only quantitative activity endpoints are considered. The different kinds of activity responses and chemical explanatory variables are detailed in Section 2 and 3 respectively. The different modelling techniques as well as the selection of an explanatory variables subset are reviewed in Section 4.

The main area of QSAR application is the pharmaceutical industry (Tong *et al.* 2004). QSAR is now an inexorably embedded tool in drug discovery, from lead discovery to lead optimisation (the background of the thesis) (Hopfinger and Tokarski 1997; Kubinyi *et al.* 1998). Indeed, a successful new drug requires many properties, such as a low toxicity, and an appropriate absorption, distribution, metabolism and elimination (*ADME*) profile (Hansch *et al.* 2004). The chemical entity should also be easily and, if possible, cheaply synthesised. If a drug candidate fails one of those drug-like properties, it can never come to the market. As, once developed, QSAR models are easily and rapidly applied to predict the activity endpoint for any new compound (not yet synthesised) if its chemical structure is known, there is increasing use of QSARs early in the drug discovery process as a tool to virtually screen large chemical databases. On the basis of the QSAR predictions, chemists can eliminate from further development those chemicals lacking drug-like properties and construct a priority list with the most promising compounds for further testing.

Authorities, industries and other institutions also use QSAR models for assessing the risks of chemicals released to the environment and allowing their regulation. For example, in 1996, the Environmental Protection Agency⁴ used QSAR modelling for developing and implementing a screening and testing program for chemicals that may disrupt the endocrine system. More than 87 000 chemicals were initially selected for evaluation. It was of course impossible to synthesise and test all those compounds. QSAR models were used to virtually screen this huge database and rank the compounds for priority testing (Tong *et al.* 2003).

⁴The **Environmental Protection Agency** is an agency charged by the United States federal government with protecting human health and with safeguarding the natural environment.

The great advantage of QSAR models is that predicting activity instead of measuring it allows to save time and money in chemical management and to speed up managerial decision. In addition, QSAR models can be used to reduce, refine, or replace the use of animals for an experimental purpose. To avoid animal testing is an important request of international associations (Cronin *et al.* 2003a;b) and QSAR models as well as *in vitro*⁵ tests constitute alternative methods.

As with any estimated statistical model, predictions provided by QSAR models are always uncertain. The current challenge is no longer in fitting a model that is statistically able to predict the activity within the chemical training set, but in developing a model with the capability to accurately predict the activity of untested compounds. The expected quality of a QSAR model depends on the application domain and the achieved goals. But there are growing international concerns to standardise criteria to measure the quality of a QSAR model with scientific basis.

The European Economic Community (1996) presented, in a technical guidance document, a general framework in which QSARs can be used within the risk assessment process. In this document, it is pointed out that QSAR models are only approximating methods and that it is important to perform further analysis to validate the QSAR models. In order to ensure transparency of any QSAR model, the European Commission suggested to provide minimal information as presented in Table 1.1.

A number of principles for assessing the validity of QSARs were proposed at an international workshop on the "Regulatory Acceptance of QSARs for Human Health and Environmental endpoints", organised by the International Council of Chemical Associations (ICCA)⁶ and the European Chemical Industry Council (CEFIC)⁷, held in Setubal, Portugal, on 4-6 March, 2002 (Worth 2002). But the workshop did not produce any guidance on how to interpret and apply these principles.

⁵An **in vitro** (Latin: "within glass") test is an experimental technique where the experiment is performed in a test tube, or generally outside a living organism or cell.

⁶The **International Council of Chemical Associations** is the world-wide voice of the chemical industry, representing chemical manufacturers and producers all over the world. ICCA has a central role in the exchange of information within the international industry, and in the development of position statements on matters of policy.

⁷The **European Chemical Industry Council** is aimed to maintain and develop a prosperous chemical industry in Europe. CEFIC promotes the best possible economic, social and environmental conditions to bring benefits to society with a commitment to the continuous improvement of all its activities including the safety, health and environmental performance.

General information	Reference: Process modelled: Domain of model:
Y-variable (dependent variable)	Species Type: Other information: Test method Experimental procedure: End-point modelled Type: Reliability: Data source: Units:
X-variable (independent variable)	Descriptors ↳ of initial descriptors: List of initial descriptors: descriptors: ↳ of final descriptors: List of final descriptors: descriptors: Data source: Other remarks:
Model	Samples: ↳ initial compounds: ↳ final compounds: Presentation of data: Design of training set: Outliers: Technique: Model Statistics r-squared: q-squared (x-val): External validation: ratio ↳ compounds / ↳ descriptors(initial): ratio ↳ compounds / ↳ descriptors(final): Validation: Range of validity:
Accuracy	
Remarks	

Table 1.1: Minimal information for an approach based on QSAR models. European Economic Community (1996)

The Organisation for Economic Co-operation and Development⁸ (OECD) started a program in 2003 in order to enhance the use of QSARs in the regulatory assessment of chemicals (report from the expert group on (Q)SAR, OECD 2004). The expert group worked on the basis of the Setubal principles and added a check list to provide guidance on the interpretation of those principles and to encourage consistency in their application to individual QSARs.

According to the OECD principles for (Q)SAR validation, “a (Q)SAR should be associated with the following information:

- (1.) a defined endpoint,
- (2.) an unambiguous algorithm,
- (3.) a defined domain of applicability,
- (4.) appropriate measures of goodness-of-fit, robustness and predictivity and
- (5.) a mechanistic interpretation, if possible.”

The intent of Principle 1 (defined endpoint) is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. It is therefore important to identify the experimental system that is being modelled by the (Q)SAR.

The intent of Principle 2 (unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. The issue of reproducibility of the predictions is covered by this principle. The different kinds of QSAR models as well as the choice of entering descriptors and other complexity parameters are reviewed in Section 4.

The need to define an applicability domain (Principle 3) expresses the fact that (Q)SARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions. The applicability domain is the set of molecules for which the QSAR model is valid. If a new molecule is outside the applicability domain, the prediction provided by the QSAR model may be unreliable as this molecule is not similar to the ones used to train the model. The different definitions of QSAR applicability domain are reviewed in Section 8.

The wording of the Principle 4 (appropriate measures of goodness-of-fit, robustness and predictivity) is intended to point out the distinction between the internal performance of

⁸The **Organisation for Economic Co-operation and Development** is an intergovernmental organisation in which representatives of 30 industrialised countries in North America, Europe and the Pacific, as well as the European Commission.

a model (goodness-of-fit and robustness) and the predictive power of a model (external validation). Such statistical tools are reviewed in Sections 6 and 7.

It is recognised that it is not always possible, from a scientific viewpoint, to provide a mechanistic interpretation of a given (Q)SAR (Principle 5), or that there can even exist multiple mechanistic interpretations of a given model. The absence of a mechanistic interpretation for a model does not mean that a model is not potentially useful in the regulatory context. The intent of Principle 5 is to ensure that some consideration is given to the possibility of a mechanistic association between the physicochemical properties used in a model and the endpoint being predicted, and to ensure that this association is documented.

The OECD provides also a check list to help the application of those principles. This check list is presented in Table 1.2. When developing a QSAR model, it is recommended to fill in this list and join it to the document describing the model to ensure transparency as done by Gramatica in the report from the expert group on (Q)SAR, OECD (2004).

This international effort to transparently validate QSAR models is a consequence of the increasing interest in developing and using those modelling techniques. As the Figure 1.1 shows, the number of QSAR models publications has exponentially increased in the past twenty years.

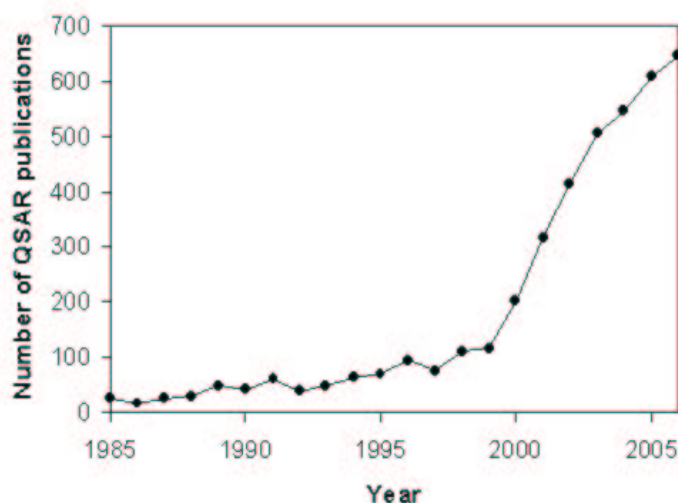


Figure 1.1: Number of references per year of publications which were retrieved using QSAR keyword with Pubmed (see <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>).

CONSIDERATIONS	YES/NO
1) Defined endpoint - Does the model have a clearly defined scientific purpose? (i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?) - Does the model have the potential to address (or partially address) a clearly defined regulatory need? (i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline)? - Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)? - Are the units of measurement of the endpoint given?	
2) Defined algorithm - In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents? - In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used?	
3) Domain of applicability - In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals? - In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment? - In the case of a (Q)SAR, are the descriptor and response variables with inclusion and/or exclusion rules that define the variable associated ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	
4) Internal performance - Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables? - If the data used to develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided? - Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)? - Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models) - Is the QSAR associated with any statistics based on cross-validation or resampling? - If yes, is the number or samples used indicated?	
5) Mechanistic basis - In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region) - In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)? - Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	

Table 1.2: Check list in applying the OECD principles. report from the expert group on (Q)SAR, OECD (2004)

QSARs have been subject to a number of excellent reviews: Cronin (2000); Dearden *et al.* (1997); Hulzebos *et al.* (1999); Walker *et al.* (2002). For a review of QSAR modelling in virtual screening and data mining, see Oprea *et al.* (2005). For applications of QSARs in modelling mutagenicity and carcinogenicity, see Benigni (2003) and in modelling toxicity and fate, see Cronin and Livingstone (2004).

1.2 The responses modelled in QSAR

In most QSARs definitions, the kind of modelled responses are said to be any biological activity of the compounds. In this section, some frequent biological endpoints are reviewed.

To be efficient, a drug candidate should bind to the targeted receptor (see Section 3.1 of the Introduction Chapter). A measure of the efficacy of the candidate that could be modeled by a QSAR is its chemical affinity with the targeted receptor. Once binded to the receptor, the drug should induce a positive response like the production or the inhibition of a given protein. A QSAR can try to model such functional response.

QSARs have been developed for a large number of toxic endpoints in both pharmaceutical and environment fields. There is a great interest in modelling the deleterious effect of a substance on organisms, individual organs, cells or plants. One of these most important acute toxic endpoints is the *LC50*, *i.e.* the concentration of a compound that causes 50% lethality of the animal in a test batch. Several countries have taken steps to ban the oral evaluation of *LC50* on animals as it is judged cruel and sometimes unnecessary. As the OECD abolished the requirement of this oral test (Test guideline 401 2003), alternative methods such as QSARs are required. For more details about toxic endpoints and measurements, see Cronin (2002) or Schultz *et al.* (2003)

To become a drug candidate, a compound should verify multiple other properties than being non-toxic. A compound should optimise the crucial *ADME* properties. *ADME* stands for *Absorption*, *Distribution*, *Metabolisation*, and *Elimination*. Before a compound can become biologically active, it has to be taken into the bloodstream (*absorption*), to spread in the body, sometimes breaking barriers such as the blood-brain barrier, to finally arrive in the target organs or cells (*distribution*). Once the compound has worked its effect, it has to be broken down by biochemical reaction in the body (*metabolisation*) and the metabolised compound has then to be eliminated (*elimination*) in order to be not accumulated in the tissues.

In the past, a lot of drug investigations were failing in the last stages of the drug discovery and development process because of poor *ADME* properties. QSARs offer the opportunity

to explore those properties even in the beginning of the process. That's why QSARs are increasingly used in pharmaceutical industries. For more details about QSARs predictions of ADME properties, see Selick *et al.* (2002).

The modelled endpoints can be of different nature: continuous, categorical ordinal or categorical nominal. The toxicity of a compound, for instance, can be continuously measured by its concentration that causes 50% lethality (LC_{50}) as explained upper. The toxicity can also be an ordinal property like a compound is non-toxic, acceptably slightly toxic or too much toxic to become a new medicine. Finally, a response can be nominal as the fact that a compound can either break the blood-brain barrier or not. According to the nature of the response, the statistical tools applied to model it as well as the QSAR validation process are different. This chapter only concentrates on continuous responses.

Whatever the type of response is, it should be clearly defined as recommended by the first OECD principle. The conditions in which the endpoint is measured should be detailed as well as the protocol(s) applied. This is an important requirement to ensure appropriate QSAR predictions by other users than the model builders.

1.3 The explanatory variables entered in QSAR models

There are two approaches in modelling the activity-structure relationships: fragment-based QSAR and descriptor-based QSAR. The first approach consists in dividing each molecule of the database in all possible fragments (atoms or group of atoms) and modelling the activity of molecules only with the fragments counts. The second approach makes use of the whole structure of the compounds by modelling the activity as a function of chemical descriptors. Those descriptors could be simply the number of such atom, as well as properties characterising the links between atoms or the three-dimensional shape of the molecule. This section focuses on chemical descriptors and the choice of descriptors that enter QSAR models.

Seven years ago, Todeschini and Consonni (2000) have already listed more than 3000 molecular descriptors in their Handbook of molecular descriptors. Nowadays, around 8000 chemical properties can be used to describe a molecule. Descriptors are often classified according to the molecule dimensions taken into account as represented in Figure 1.2:

- 1D: one-dimensional linear representation of the molecule
This class of descriptors contains simple indexes that can be deduced from the molecular formula, for instance, the number of such atoms or groups, the molecular weight⁹

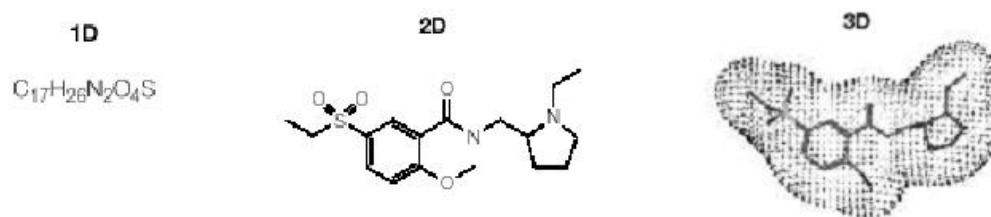


Figure 1.2: Number of dimensions taken into account in descriptors computation.

or the molar refractivity¹⁰.

- 2D: two-dimensional planar representation of the molecule
Two-dimensional descriptors regroup mainly topological and connectivity indexes. One of the most well-known 2D-descriptor is the octanol-water partition coefficient, K_{OW} , that is a measure of hydrophobicity¹¹ and hydrophilicity¹² of a substance.
- 3D: three-dimensional spatial representation of the molecule
Three-dimensional descriptors summarise the geometry, the surface and the volume, as well as other electrostatic properties of the molecule such as the HOMO¹³ or the LUMO¹⁴. The difference of the energies of the HOMO and LUMO can sometimes serve as a measure of the excitability of the molecule.

For more examples of descriptors, see (Livingstone 2000; Todeschini and Consonni 2000)

Some descriptors have to be measured in laboratories by performing experiences on the molecule, such as the acidity constant K_a . They are often referred to as *empirical descriptors*. These descriptors can be heavy to use in practice, especially if the goal of QSAR modelling is to speed up the drug discovery process. Hopefully, more and more descriptors can be computed exactly by applying new softwares such as ADAPT (Jurs 2002; Stuper and Jurs 1976), OASIS (Mekenyan and Bonchev 1986), CODESSA (Katritzky *et al.* 1994), and DRAGON (Consonni *et al.* 2005) .

⁹The **molecular weight** can be calculated as the sum of the atomic weights of all the atoms constituting the molecule.

¹⁰**Molar refractivity** is a measure of the volume occupied by an atom or group.

¹¹**Hydrophobicity** is the property of a molecule that is repelled by water.

¹²**Hydrophilicity** is the property of a molecule that can bond with water.

¹³**HOMO** stands for Highest Occupied Molecular Orbital. It is the highest-energy orbital of a molecule with one or two electrons.

¹⁴**LUMO** stands for Lowest Unoccupied Molecular Orbital. It is the lowest-energy orbital with no electrons.

Those software can theoretically compute a set of descriptors for any molecule coded using a universal text representation called SMILES. The name *SMILES* stands for “Simplified Molecular Input Line Entry Specification”. This universal text coding of a molecule was first introduced by Weininger (1988). It allows to specify the structure of chemical substances in an ASCII format that can be understood by most chemical softwares that compute descriptors. On Figure 1.3 are shown the SMILES coding for the acetic acid and the benzene molecules.

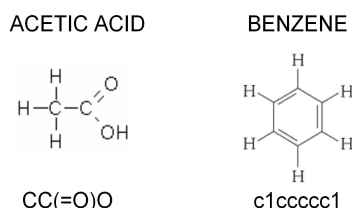


Figure 1.3: Two examples of SMILES coding: acetic acid and benzene molecules.

The nature of the descriptors may be either continuous (*e.g.* molecular weight), discrete (*e.g.* number of such atom), ordinal or nominal (*e.g.* solvent type). Continuous descriptors can be also transformed by centering or scaling, to ensure the same weight for each descriptor in the QSAR models. To avoid explanatory variables with skewed distribution, it is not rare to work with the logarithm of descriptors. Other transformations are possible such as any power of the descriptors.

The number of available molecular descriptors is great. Allowing transformations and cross-products for interactions make the number of possible models exploding. And it is often frequent that the number of explanatory variables exceeds the number of observations, leading the failure of multiple linear regression models. With any kind of modelling techniques, using too many explanatory variables may cause overfitting and results in poor predictive power for new compounds. Not all available explanatory variables can enter the model and a choice between them has to be performed.

In practice, for some defined endpoint, important descriptors that influence this endpoint are known. For instance, it is well-known that the octanol-water partition coefficient, K_{OW} , influences greatly the $LC50$, *i.e.* the concentration of a substance that causes 50% lethality (report from the expert group on (Q)SAR, OECD 2004). For those situations, one can hope to obtain good QSAR only with one or several of those few known descriptors. But, most of the time, the structure-activity relationship is much more complex, ambiguous and it is often necessary to use a greater number of chemical descriptors to be able to explain it. It is then necessary to apply descriptors selection methods to the huge

set of available descriptors. But trying all the subsets of explanatory variables is impossible.

A first rapid screening can be made by discarding constant values and paired correlated variables that could lead to collinearity and instability problems with multiple linear regression models. Then, there exists a number of various techniques to select a descriptors subset that will enter the regression models, such as stepwise regression, backward elimination, forward selection, simulated annealing, evolutionary and genetic algorithm. Subset variables selection constitutes the subject of a great part of QSAR papers (Kubinyi 1994; 1996; Waller and Bradley 1999; Whitley *et al.* 2000). A recent comparison has given a demonstration of the advantages and success of genetic algorithm (Xu and Zhang 2001).

Whatever is the subset variable selection method, the second OECD principle states that the descriptors choice should be transparent, saying which variables were first considered and finally which variables are kept in the model and why. Eriksson *et al.* (2003) noted that the choice of descriptors used is far more important than the specific modelling method employed. Nevertheless other methods than multiple linear regression are used in QSAR modelling, less sensitive to the correlation between the descriptors. The main used modelling methods are reviewed in the next section.

1.4 Different classes of QSAR models

The most common modelling techniques in QSAR to link the descriptors to the activity response are based on regression analysis. Among those techniques, Multiple Linear Regression (abbreviated *MLR*) is the most classical one. But, as explained upper, MLR is not adapted to the existing correlations between descriptors. Other alternatives exist. Multivariate projection methods to a subspace of orthogonal latent variables have become more and more popular in QSAR. Partial Least Squares Regression (*PLSR*) and Principal Components Regression (*PCR*) are two such projection techniques increasingly used. Another alternative, close to MLR, is the Ridge Regression (*RR*) which imposes a penalty to the size of the coefficient in the linear regression model. MLR, PLSR, PCR and RR are reviewed in details below. Other modelling techniques such as Regression Trees (*RT*), k-Nearest Neighbours (*kNN*) and Neural Networks (*NN*) are also briefly introduced.

The following notations are used :

- Y is the (random) activity response;
- $\mathbf{x}$ is a K -vector of (deterministic) descriptor variables (or any transformation of them, including power, cross-product, dummy variables encoding a categorical descriptor, constant term...);
- $\mathbf{x}^c$ and $\mathbf{z}$ are respectively the mean-centered and the standardised $\mathbf{x}$;

- $\mathbf{y} = (y_1, y_2, \dots, y_N)^T$ is the N -vector of observed responses;
- $\mathbf{y}^c$ is the mean-centered $\mathbf{y}$;
- $\mathbf{x}_i$ is the i^{th} observed K -vector of transformed descriptor variables ($i = 1, 2, \dots, N$);
- $\mathbf{x}_i^c$ and $\mathbf{z}_i$ are respectively the mean-centered and the standardised $\mathbf{x}_i$ ($i = 1, 2, \dots, N$);
- $\mathbf{X}$ is a $(N \times K)$ matrix where each row is an observed vector of transformed descriptor variables;
- $\mathbf{X}^c$ and $\mathbf{Z}$ are respectively the mean-centered and standardised matrix $\mathbf{X}$.

1.4.1 Multiple Linear Regression (MLR)

MLR assumes that the link between the activity response and the standardised descriptors is given by the equation

$$Y = \mathbf{z}'\boldsymbol{\beta} + \epsilon \quad (1.1)$$

where $\boldsymbol{\beta}$ is a K -vector of coefficients and ϵ is the random error term with zero mean, constant variance σ_ϵ^2 and independent from one observation to another.

This model is traditionally fitted to the data by least squares method, estimating the regression coefficients $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})$. If the inverse of the matrix $\mathbf{Z}'\mathbf{Z}$ exists, the solution of this minimisation problem is unique and depends explicitly on both observed responses and observed descriptors: $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$. Using this least squares estimate, the predicted responses $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ are the orthogonal projections of the observed responses $\mathbf{y}$ on the space spanned by the columns of standardised transformed descriptors $\mathbf{Z}$.

MLR is very popular in QSAR modelling because of its simplicity and its well-known theoretical background (Cronin and Schultz 2003; Schultz and Cronin 2003). Through the estimated equation

$$\hat{Y} = \mathbf{z}'\hat{\boldsymbol{\beta}}, \quad (1.2)$$

the effect of each descriptor used on the response is understood as well as its amplitude. The hypothesis testing on the parameters allows to check if those effects are significant. The confidence interval and the prediction interval for the expected response and for the individual response respectively can be used to quantify the uncertainty of their estimations.

Concurrently to those advantages, MLR has its own limitations. A main drawback of MLR in the field of QSAR modelling is that it can not be used when the number of explanatory variables exceeds the number of observations. In addition, MLR assumes the predictor variables to be not linearly dependent, *i.e.* the rank of the matrix $\mathbf{Z}$ is K . If the data exhibit collinearity among the variables $\mathbf{z}$, the estimated regression coefficients get unstable. As a consequence, MLR can be used but a careful selection within the set of

available descriptors has to be performed as proposed in the previous section.

Examples of the application of MLR coupled with a variables selection by a genetic algorithm can be found in the report from the expert group on (Q)SAR, OECD (2004).

1.4.2 Principal Components Regression (PCR)

Principal Components Regression is an alternative to MLR when the explanatory variables are correlated (Draper and Smith 1981; Myers 1986). A Principal Components Analysis (*PCA*) is first applied to construct orthogonal latent variables, called the Principal Components (*PCs*), that can then enter an MLR model without any more collinearity problem.

The PCs are computed as linear combinations of the mean-centered descriptors, $\mathbf{t}_K = \mathbf{W}_K \mathbf{x}^c$, where the weight matrix $\mathbf{W}$ of dimension $(K \times K)$ is composed of the K eigenvectors of the covariance matrix of the observed mean-centered descriptors $\mathbf{X}^c$. One may use all the PCs (ordinary least squares results) as they are orthogonal or one may select only a few ones. A selection can be done by discarding the less important PCs and keeping only the $M < K$ first PCs such that the retained explained variance is large enough. The more the explanatory variables are correlated, the less PCs are necessary to recover most of the $\mathbf{X}^c$ information.

The M first PCs are not necessarily the most related to the response Y . An alternative approach to PCR, called the inferential approach, selects only the set of principal components that are significant in the regression model. See, for example, Massey (1965) and Mason and Gunst (1985).

The M uncorrelated selected PCs, $\mathbf{t}_M$, are used to model the response:

$$Y = \mathbf{t}_M' \boldsymbol{\beta} + \epsilon = \mathbf{x}^c' \mathbf{W}_M' \boldsymbol{\beta} + \epsilon \quad (1.3)$$

where $\boldsymbol{\beta}$ is a M -vector of parameters that can be estimated using least squares. The corresponding predicted responses

$$\hat{\mathbf{y}} = \mathbf{T}_M (\mathbf{T}_M' \mathbf{T}_M)^{-1} \mathbf{T}_M' \mathbf{y} = \mathbf{X}^c \mathbf{W}_M' (\mathbf{W}_M \mathbf{X}^c' \mathbf{X}^c \mathbf{W}_M')^{-1} \mathbf{W}_M \mathbf{X}^c' \mathbf{y} \quad (1.4)$$

are the orthogonal projections of the observed responses $\mathbf{y}$ on the space spanned by the M PCs. The effect of the original descriptors on the activity is not as direct as with MLR models. One has to inspect the scores and the loadings to interpret the PCs as a function of the original descriptors and analyse the most impacting descriptors.

While fitting an MLR model to explain the response by the M first PCs or all the PCs, one can then discard the PCs that have non-significant effect on the activity response or if their eliminations do not decrease the predictive power. Indeed, the PCs are only summarising the observed explanatory variables ignoring their potential link with the response. There is no guarantee that the M first PCs are the M best ones in predicting the response. Partial Least Squares Regression attempts to take also the response into account in the construction of the new regressors.

1.4.3 Partial Least Squares Regression (PLSR)

Partial Least Squares were first developed in 1966 in the field of econometrics by a Swedish statistician, Herman Wold (Wold, H. 1966). PLS Regression was finalised in 1982 by Herman Wold and his son, Svante Wold (Wold S. *et al.* 1982). Svante Wold introduced this modelling technique in chemometrics (Wold and Dunn 1983). From that time, PLSR is increasingly used in QSAR modelling as it manages a great number of explanatory variables (even possibly greater than the number of available data) and the existing correlations between descriptors (Eriksson *et al.* 2001). PLSR also does not assume that the explanatory variables are exact and 100% relevant for modelling the response. For all those reasons, PLSR is now as popular as MLR.

The principle of PLSR is very similar to PCR as the link between the activity response and the descriptors is modeled through newly constructed latent variables. The specificity of PLSR is that those latent variables are recursively chosen to perform a simultaneous decomposition of the observed mean-centered descriptors $\mathbf{X}^c$ and the mean-centered observed responses $\mathbf{y}^c$ with the constraint that they explain as much as possible of the covariance between $\mathbf{X}^c$ and $\mathbf{y}^c$. More precisely, M latent variables are defined as linear combinations of the original descriptors, $\mathbf{t}_M = \mathbf{V}_M \mathbf{x}^c$, to model simultaneously $\mathbf{X}^c$ and $\mathbf{y}^c$:

$$\begin{cases} \mathbf{X}^c = \mathbf{T}_M \mathbf{P}_M + \mathbf{E} \\ \mathbf{y}^c = \mathbf{T}_M \mathbf{q}_M + \mathbf{f}. \end{cases} \quad (1.5)$$

$\mathbf{T}_M$ is the $(N \times M)$ matrix of scores and $\mathbf{P}_M$ and $\mathbf{q}_M$ are respectively the $(M \times K)$ matrix and M -vector of loadings.

An adequate recursive algorithm (Wold S. *et al.* 1982; Martens 1985; de Jong 1993) is used to define the scores $\mathbf{T}_M = \mathbf{X}^c \mathbf{V}_M' = \mathbf{X}^c (\mathbf{W}_M' \mathbf{P}_M)^{-1} \mathbf{W}_M'$ based on the singular value decomposition of $\mathbf{X}^c' \mathbf{y}^c$ where $\mathbf{W}_M$ is a weight matrix of dimension $M \times K$ (similar to PCR). The first latent variable is selected as the linear combination of the explanatory variables $\mathbf{X}^c$ that maximises its covariance with the response $\mathbf{Y}^c$. It is first used to model the response. Then this first dimension is subtracted from $\mathbf{X}^c$ and $\mathbf{Y}^c$ and the second latent variable is selected as the linear combination of the residuals of $\mathbf{X}^c$ that maximises its

covariance with the residuals of Y_c . This process can be repeated K times.

As those scores are linear combinations of the original descriptors, the model for the response in (1.5) can be rewritten similarly to an MLR model:

$$\mathbf{y}^c = \mathbf{X}^c \mathbf{b}_M + \mathbf{f} \quad (1.6)$$

with the M -vector $\mathbf{b}_M = \mathbf{V}_M' \mathbf{q}_M = (\mathbf{W}_M' \mathbf{P}_M)^{-1} \mathbf{W}_M' \mathbf{q}_M$.

Using the definition of the loadings $\mathbf{q}_M = (\mathbf{T}_M' \mathbf{T}_M)^{-1} \mathbf{T}_M' \mathbf{y}^c$ provided by the PLS algorithm, the N responses are predicted by

$$\hat{\mathbf{y}}^c = \mathbf{T}_M (\mathbf{T}_M' \mathbf{T}_M)^{-1} \mathbf{T}_M' \mathbf{y}^c = \mathbf{X}^c \mathbf{V}_M' (\mathbf{V}_M \mathbf{X}^c' \mathbf{X}^c \mathbf{V}_M')^{-1} \mathbf{V}_M \mathbf{X}^c' \mathbf{y}^c, \quad (1.7)$$

which is similar to MLR or PCR predictions. Those predictions can also be interpreted as the orthogonal projections of the observed responses $\mathbf{y}^c$ on the space spanned by the columns of the observed latent variables $\mathbf{T}_M = \mathbf{X}^c \mathbf{V}_M'$.

The problem of variable selection remains as in MLR modelling but PLSR solves the collinearity problem of the descriptors. A lot of different methods exist to choose the adequate number latent variables M . A quite often used technique in the field of QSAR modelling is the one of Baroni *et al.* (1993) called GOLPE (Generating optimal linear PLS estimations). It mixes a preliminary descriptors selection by means of D-optimal design in the loadings space and an iterative evaluation of the effects of the individual variables on the model predictivity.

The effect of original descriptors on the activity response is not as direct as with MLR models. But the scores and loadings can be used to interpret the latent variables as a function of the original descriptors and analyse the most impacting descriptors. For examples of PLSR applications to QSAR modelling, see for instance the work of Eriksson *et al.* (1997) or Eriksson *et al.* (2003)

1.4.4 Ridge Regression (RR)

Instead of selecting a variables subset and dropping descriptors from an MLR model, one can use a modification of the classical least squares MLR method, called Ridge Regression (RR). The assumed model of RR is exactly the same as the MLR model, $Y = \mathbf{z}'\boldsymbol{\beta} + \epsilon$. The difference stands in the least squares criteria that is penalised by a multiple of the sum of squared regression coefficients. For a fixed constant λ , the K least squares RR coefficients estimates are given by

$$\hat{\boldsymbol{\beta}}^{RR} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{ (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}' \boldsymbol{\beta} \} = (\mathbf{Z}' \mathbf{Z} + \lambda \mathbf{I})^{-1} \mathbf{Z}' \mathbf{y}. \quad (1.8)$$

The activity response of a new molecule with transformed descriptors $\mathbf{x}$ is predicted by

$$\hat{y} = \mathbf{z}'\hat{\boldsymbol{\beta}}^{RR} = \mathbf{z}'(\mathbf{Z}'\mathbf{Z} + \lambda I)^{-1}\mathbf{Z}'\mathbf{y}. \quad (1.9)$$

This RR sum of squares is equivalent to the minimisation of the ordinary sum of squares under the constraint that the sum of the squared coefficients does not exceed a given size. This constraint protects the gradients of the response surface in the direction of the smallest PCs in correlated $\mathbf{Z}$ -space against potentially high variance. These directions are exactly the causes of trouble in MLR. That is how the additional penalty term has the effect to shrink the vector of coefficients and limits the two cited risks.

It is well-known that RR produces biased estimates of $\boldsymbol{\beta}$. RR modifies $\hat{\boldsymbol{\beta}}^{MLR}$ by introducing a bias, which decreases the variance of the estimates by more than the squared bias, so that the RR estimates of $\boldsymbol{\beta}$ have lower mean squared error (MSE) than $\hat{\boldsymbol{\beta}}^{MLR}$. More precisely, the ridge existence theorem (Vinod 1981) shows that, for $0 < \lambda < \frac{2\sigma^2}{\boldsymbol{\beta}'\boldsymbol{\beta}}$, $MSE(\hat{\boldsymbol{\beta}}^{RR}) < MSE(\hat{\boldsymbol{\beta}}^{MLR})$ where σ^2 is the residual variance. That's why RR performs pretty well in practice for prediction. Hastie *et al.* (2001) summarise: "For minimising prediction error, RR is generally preferable to MLR with variables subset selection, PCR and PLS. However the improvement over the latter two methods is only slight."

For a comparison of RR, PLS and PCR in QSAR applications, see Hawkins *et al.* (2001).

1.4.5 Regression Tree (RT)

The principle of Regression Trees is really different from the previous regression methods (MLR, PLSR, PCR or RR) as the nature of the relationship between the activity response and the descriptors is not pre-specified. RTs are nonparametric models.

The most simple technique for fitting an RT to QSAR data consists in recursively partitioning the data into successively smaller groups (called *nodes*) with binary splits based on a single descriptor (example: $x_j \leq c$ and $x_j > c$ for the j^{th} transformed descriptor and a constant c). At each step, splits for all of the descriptors are examined by an exhaustive search procedure and the best split is chosen. The idea of Breiman *et al.* (1984) is to define the best split at stage M as the one that minimises the total residuals sum of squares over all the current M nodes, $\sum_{m=1}^M \sum_{i=1}^{N_m} (y_i - \hat{y}_i)^2$. At each step, this is achieved by finding the split that maximises the decrease in the residuals sum of squares between the parent splitted node (RSS_P) and the left (RSS_L) and right (RSS_R) child nodes: $\Delta RSS = RSS_P - (RSS_L + RSS_R)$.

In the m^{th} node, the predicted activity $\hat{y}_i$ of the i^{th} molecule can be simply defined as the average value of the N_m observed activities of the compounds in that node. Indeed, this is the constant value that minimises the sum of squares. The activity response of a new molecule can be predicted by

$$\hat{y} = \sum_{m=1}^M \bar{y}_m I_m(\mathbf{x}) \quad (1.10)$$

where $\mathbf{x}$ is the K -vector of transformed descriptors, $I_m(\mathbf{x})$ is the indicator of the node (1 if the observation $\mathbf{x}$ belongs to the node m ; 0 otherwise) and $\bar{y}_m$ is the average observed value at node m . One can also imagine to build local predictive model in each node of the tree, such as an MLR model, to obtain at the end a smoother regression surface.

The tree can recursively be splitted till all terminal nodes (called *leaves*) contain only one molecule. Techniques such as cross-validation can then be used to prune the overfitted tree to an optimal size (see Therneau and Atkinson (1997)). Breiman *et al.* (1984) proposed to associate to each number of split a cross-validated estimate of the prediction error and select the optimal one as the number of split that minimises the cross-validated prediction error, eventually adding 1 standard deviation of this estimate (defined as the *1 SE rule*).

The output is a tree diagram with the branches determined by the splitting rules and a series of terminal nodes that contain the mean response or a local model. An example of a regression tree with average response in the leaves is shown in Figure 1.4.

RT is a very flexible modelling technique. Indeed, there are various methods for growing the tree, allowing not only binary splitting, various criteria to optimise while splitting and also various tools for pruning (Hastie *et al.* 2001). For all those variation of the presented basic RT, there is no *a priori* assumption about the kind of activity-structure relationship. RT is appreciated in QSAR applications as it can handle large data sets, it allows interactions and nonlinear relation between descriptors and response, and also because it produces sequences of prediction rules that are readily interpretable. Both continuous and categorical descriptors can be handled when splitting.

Beside those advantages, RTs have a main drawback: those are unstable predictors. Breiman (1996) proposed to use bootstrap to reduce the sensitivity of the RT to the training set and obtain a smoother predictor. On a series of bootstrapped versions of the training set, an RT is fitted by the same procedure as the original one. A new predictor is constructed by aggregating the replicated RTs. The new predictions are defined by the predictions averaged over all replicates.

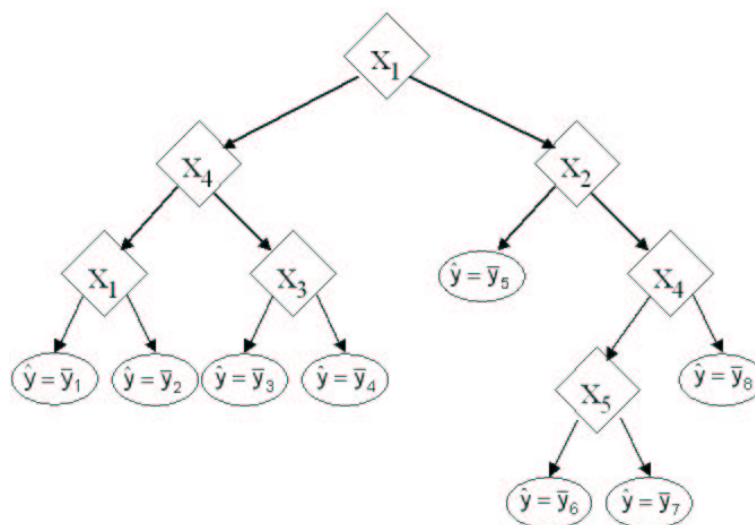


Figure 1.4: Schematic example of regression tree using the 5 explanatory variables x_1, x_2, x_3, x_4 and x_5 to recursively split the data set in 8 final leaves where the response is predicted by the mean response of observations belonging to that leaf.

For applications of RT to QSAR modelling, see Dzeroski (2001), Izrailev and Agrafiotis (2001) or Blockeel *et al.* (2004).

1.4.6 k -Nearest Neighbours (k NN)

The k -nearest neighbours method consists in predicting the activity of a molecule as the average (or weighted average) of the observed activity values of the k nearest molecules. k NN is greatly appreciated by QSAR practitioners as it is really intuitive. Indeed, its principle reflects the ancestral idea that similar compounds reveal similar activity property. Like RT, k NN does not make any *a priori* assumption about the nature of the activity-structure relationship. The descriptors are not directly used to model the activity but to define the neighbourhood. More precisely, for any molecule with observed descriptors $\mathbf{x}$, the predicted value is

$$\hat{y} = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i, \quad (1.11)$$

i.e. the average response of the k closest observations $\mathbf{x}_i$ in the neighbourhood $N_k(\mathbf{x})$ of $\mathbf{x}$.

Even if the k NN principle seems simple, its application in practice requires some crucial choices about the following parameters (Hastie *et al.* 2001):

- Choice of k , the number of points belonging to the neighbourhood.

The traditional rule of thumb suggests to adapt k to the number of observations: $k = \sqrt{N}$. One can also use cross-validation to choose k to obtain the best bias-variance tradeoff.

- Choice of the distance between two observations.

Some possible choices are the Euclidean distance, the Mahalanobis distance or the L_1 -norm for quantitative descriptors and an indicator (Tutz 1990) for qualitative descriptors (0 if the two observations share the same category, 1 otherwise). Those distances are precisely defined in Section 8.

- Choice of the weighting method.

Computing the average of the observed activity in the neighbourhood provides the same weight to the k observations, the closest as the furthest observations of the neighbourhood. One can also weight the response values by a function of their distances or by using kernels.

Due to the curse of dimensionality, the k NN method (like most of non-parametric methods) performs very poorly in a high-dimensional descriptors space like it is often the case in QSAR modelling. Before applying the k NN method, it is necessary to perform a subset variables selection to eliminate irrelevant descriptors. Zheng and Tropsha (2000) proposed a variables selection based on simulated annealing to apply k NN to QSAR data. For another application of k NN to QSAR modelling, see Hoffman *et al.* (1999) or Baurin *et al.* (2002).

1.4.7 Neural Networks (NN)

The most common (Artificial) Neural Network model equation is of the form

$$f(\mathbf{x}) = G\left(\sum_j (w_j g_j(\mathbf{x}))\right) \quad (1.12)$$

where $\mathbf{x}$ is the K -vector of transformed descriptors. Like the neurons in the brain, the descriptors $\mathbf{x}$ (continuous or binary), the functions g_j and the function G are multiple layers that are interconnected to produce the prediction of the activity response by $f(\mathbf{x})$. When the number and the shape of functions g_j (often logistic functions) and the function G (often *tanh*) have been fixed, the weights w_i are estimated to minimise the squared prediction error. The gradient-based backpropagation method can achieve this task (Werbos 1974).

NN models are nonlinear models that can be represented as a network structure as depicted in Figure 1.5. There exist different architectures of NN according to the number of hidden layers, the number of nodes for each layer and the connections existing between all the different layers. The two main forms of NN are the feedforward NN and NN using

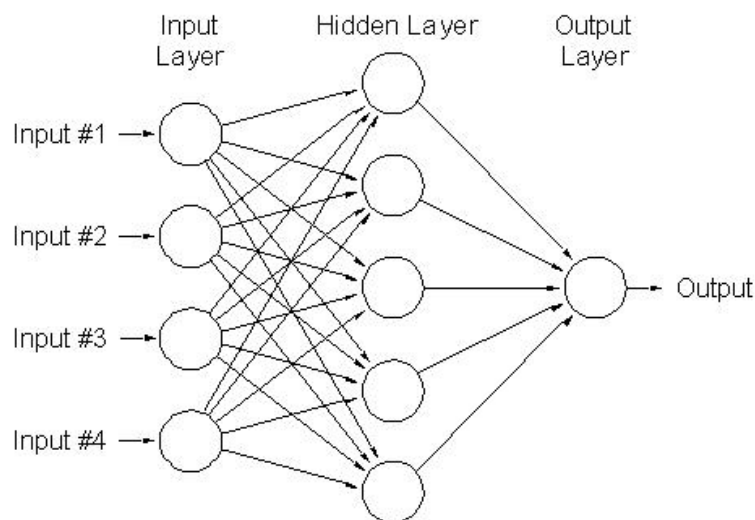


Figure 1.5: Schematic example of feedforward Neural Network (from the website http://smig.usgs.gov/SMIG/features0902/tualatin_nn.fig3.gif).

radial basis functions.

Neural Networks are described with more details in Bishop (1995). For examples of QSAR Neural Network applications, see (Andrea and Kalayeh 1991), (Devillers 1996), (Duprat *et al.* 1998), (Kovesdi *et al.* 1999), (Manallack and Livingstone 1999) or (Huskonen *et al.* 2000)

1.4.8 Choice of complexity parameters

All the presented models depend on some complexity or tuning parameters: the number of explanatory variables (descriptors) in MLR, the number of latent variables kept in PLSR or the number of PCs in PCR, the penalty parameter λ in RR, the level of pruning in RT, the number k of neighbours in k NN or the number of layers or nodes in NN. The idea is to try different values for those parameters and select the value that yields to the best model performance.

If the criterion measuring the model performance is only based on the fit to the training set (such as the squared error loss summarising the difference between the observed responses and the fitted values), then there is a clear risk of overfitting. Jaworska *et al.* (2005a;b) suggest to use other analytical criteria (Mallow Cp statistic, Akaike Information Criterion, Bayesian or Schwartz Information Criterion) or sample re-use techniques (Y-scrambling, cross-validation or bootstrap). Those different techniques to measure the

performance of a QSAR are detailed in Chapter 2.

With the recent computer advances, it's no more expensive to fit the data with a selected modelling technique using different complexity parameters. It is even also recommended to try different modelling techniques and to choose the most performing one (Eriksson *et al.* 2003).

1.5 Data collection and model fit

The biological data to which modelers have access are very limited in terms of chemical and biological space. Chemists often have to collect their own data for QSAR modelling. There is a lack of international dash to gather biological database, with adequate explanation of the way they have been obtained. This is one of the goal of the OECD principles.

The selection of the training set is crucial whatever is the field of application. The intuitive idea is to collect data similar to the data for which predictions will be made with the fitted QSAR model. The objective is to select molecules that well span the chemical domain of interest.

In theory, design of experiments should be used to generate the training set to ensure maximum variance in the descriptors space and produce a representative training set. But the powerful tools of experimental design can not directly be applied to construct a training set. Indeed, when a molecule is even not yet synthesised, the corresponding descriptors can be computed thanks to its SMILES coding, but the opposite is not possible as no chemist is able to construct a molecule with such given values of descriptors.

In many QSAR papers, nothing is said about the choice of the molecules in the training set. Most of the time the only remark is that the QSAR is fitted on molecules having the same mechanism of action (for instance, all esters, all antioxidant, all disinfectant,...). This ensures the adequacy of a unique QSAR to model the activity response of those molecules because a "universal" QSAR able to model chemicals with significantly different structures and modes of action may not seem plausible. A set of such predefined molecules is often called a *library* in chemical jargon.

A chemical library may be composed either of compounds having the same mechanism of action or compounds obtained by systematic reactions of a small number of starting compounds with a larger number of reagents (*combinatorial library*). The chemist may choose as a starting data set such a library. Often, as the compounds are not yet synthesised, the collection of compounds is referred to as a *virtual library*.

In practice, if this virtual library contains too many compounds to synthesise, the statistician may help the chemist to select a sub-library of reasonable size with diverse molecules. The statistical techniques to select a subset of compounds from a greater starting set is called multivariate design (Wold S. *et al.* 1986) or statistical molecular design (Eriksson *et al.* 1996; Linusson 2000). Eriksson *et al.* (2000) proposed to apply fractional factorial designs and De Aguiar *et al.* (1995) proposed to apply D-optimal designs to obtain a representative data set from a virtual library. Those techniques of experimental design (Box *et al.* 1978) are based on the descriptors computed for the compounds in the virtual library or on summaries of those descriptors such as the latent variables obtained with PLS or the principal components obtained with PCA.

In the first OECD principle, it is recommended to use the same assay protocol to obtain high quality data and to explain this protocol with the QSAR equation. Some preliminary analysis can be performed on the collected data to ensure their quality before applying the modelling techniques. One can first check if the standard deviation of experimental error is constant over the domain of interest. This is possible if replicates have been tested but this is not always the case.

When the quality of the collected data is insured, a QSAR model can be fitted. The data used to fit the QSAR model is called the *training set*. One can use the PLS or PCA techniques to project the training set in a space of smaller dimensions to analyse their distribution, looking for outliers or clusters. Compounds are often gathered in clusters and it has to be decided if a QSAR model for each cluster may be better than a global model on all data. The necessary homogeneity of the data and the absence of outliers can also be graphically analysed for any kind of models by plotting the observed responses (y_i , $i = 1, \dots, N$) against the predicted ones ($\hat{y}_i$, $i = 1, \dots, N$) or the residuals ($e_i = y_i - \hat{y}_i$, $i = 1, \dots, N$) against the data numbers (i).

Different other tools can be used to detect outliers in the training set according to the QSAR model class. The standard instrument for MLR models is the *leverage* (Atkinson 1985) that measures the distance from one molecule to the whole training set in the descriptors space. It is traditionally computed for the i^{th} compound of the training set as the i^{th} diagonal element of the Hat matrix

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'. \quad (1.13)$$

An observation with a leverage higher than $2K/N$ for great samples (Belsley *et al.* 1980) or $3K/N$ for small samples (Vellman and Welsch 1981) may be considered potentially as an influential point and the effect of removing such an observation from the training set has to

be analysed. The leverage can be computed for any kind of model but is more meaningful for MLR and RR.

The *standardised residual in prediction* is another tool to detect outliers with MLR models. It is defined for the i^{th} observation in the training set as

$$e_i^{SRP} = \frac{e_i^{CV}}{\hat{\sigma} \sqrt{(1 - h_{ii})}} \quad (1.14)$$

where e_i^{CV} is the difference between the observed i^{th} response and its prediction using the model fitted on the rest of the data, $\hat{\sigma}$ is an estimate of the standard deviation of the residuals and h_{ii} is the leverage of the i^{th} observation. Molecules with a standardised residual in prediction greater than 2 or 3 are considered as outliers.

Leverages and standardised residuals in prediction can be simultaneously represented on a graph called *William plot* to visualize outliers and define the applicability domain of the MLR model, as detailed in the sub-section 8.5.

For other classes of models, other distances (Euclidean, Mahalanobis, L_1 , $DModX$, Hotelling's T^2) can be defined to measure the distance of a molecule to the training set and detect outliers. They are reviewed in section 8 on the applicability domain.

When outliers and influential points have been removed from the training set, further analyses have to be performed to assess the validity of the model. There are three main points in the validation of a QSAR model: the analysis of the model performance on the training set and on an external data set (OECD Principle 4), and the definition of its applicability domain (OECD principle 3), *i.e.* the group of compounds for which the QSAR model predictions are valid. A review of the QSAR literature about those three validation steps constitute the three next sections.

1.6 Internal model performance

There is an increasingly international interest to validate QSAR models (Cronin and Livingstone 2004; report from the expert group on (Q)SAR, OECD 2004). OECD principle 4 intends to make the distinction between the internal performance of a QSAR model characterised by goodness of fit and robustness measurements and its external performance characterised by its predictive power for new chemicals. The aim of this section is to review internal performance criteria for the practical application of OECD principle 4.

1.6.1 Validation of underlying assumptions

The assumptions underlying the QSAR model must be validated. In regression analysis, the residuals are often assumed to be independently normally distributed with zero mean and constant variance. To validate the normality assumption, an histogram or a normal probability plot of the observed residuals e_i 's can be made. A graph of the observed residuals (e_i) against the predicted values ($\hat{y}_i$) validates the other underlying assumptions of the regression model (MLR, RR, PLSR and PCR).

1.6.2 Goodness of fit criteria

An intuitive idea is that a good model should provide predicted values similar to the observed ones. This can be visualised for any kind of models by plotting the observed responses ($y_i, i = 1, \dots, N$) against the predicted ones ($\hat{y}_i, i = 1, \dots, N$). The most well-known index for measuring the alignment of those points ($y_i, \hat{y}_i$) is the coefficient of determination

$$R_Y^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (1.15)$$

This coefficient, bounded by 0 and 1, represents the percentage of the variation of the observed activity endpoint explained by the model. The higher it is, the better the fit is. This coefficient was introduced in the field of MLR but can be computed for any other kind of model with the same interpretation.

The main drawback of R_Y^2 is that entering new explanatory variables in the model always increases its value. According to the parsimony principle, not too many variables should be entered in the model, especially if the number of observations is small. With MLR models, the R_Y^2 is adjusted to take the number of parameters, K , into account as well as the number of observations, N :

$$R_{Y_{ADJ}}^2 = R_Y^2 - \frac{K-1}{N-K}(1 - R_Y^2). \quad (1.16)$$

$R_{Y_{ADJ}}^2$ is upper-bounded by 1 but can decrease if the contribution of an additional variable is less than the impact on the degrees of freedom. The higher it is, the better the model is.

Other indexes are used with MLR models in the QSAR literature to compare the observed and the predicted data such as the standard error of estimate, s , or the Fisher statistics, F . The standard error of estimate, s , is the classical estimator for the residual standard deviation in MLR:

$$s = \sqrt{\frac{1}{N-K} \sum_{i=1}^N (y_i - \hat{y}_i)^2}. \quad (1.17)$$

If the model fits well the data, s should be small. If replicates have been performed, the experimental variance can be estimated and the estimated residual variance s^2 should be of the same order otherwise this indicates that the model suffers from a lack of fit. The Fisher statistics, F is used in MLR analysis to test the global utility of the model by comparing the part of the response variation that is explained by the model to the part that left unexplained:

$$F = \frac{\frac{\sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N (y_i - \hat{y}_i)^2}{K-1}}{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N-K}} = \frac{\frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{K-1}}{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N-K}} = \frac{\frac{R_Y^2}{K-1}}{\frac{1-R_Y^2}{N-K}}. \quad (1.18)$$

The higher is F , the better is the fit of the model. The F statistics can be compared to the 95% quantile of the Fisher distribution, $F_{K-1, N-K}^{0.95}$, and if $F > F_{K-1, N-K}^{0.95}$, the MLR model is significantly useful.

With other QSAR models classes, such criteria can be generalised by an adequate measure of model complexity (K for MLR). With PLSR, the complexity can be measured by the number of latent variables and with RT, by the number of nodes.

To summarise, the higher are the indexes R_Y^2 , $R_{Y_{ADJ}}^2$ and F , or the smaller is the residual standard deviation s , the better is the fit of the data by the model. But the model with the best fit is not surely the best model for prediction. If the model is changed to improve too much the fitting, the model will explain also the noise contained in the observed activity responses y_i 's. A such overfitted model, even if it seems to be very good, may be useless to predict the activity response for new molecules not included in the training set. As the objective of the QSAR models developer is to apply them for the activity prediction of new compounds, even not yet synthesised, the developer should find a model with a maximised predictive power and not only with good fitting criteria. The predictive power of a model can be measured either on the training set (next subsection) or on an external set (next section).

1.6.3 Robustness and internal predictivity

The most important method in QSAR modelling to quantify the predictive power on the basis only of the training set is the *cross-validation* technique. The principle of cross-validation is to simulate predictions for new molecules not used in the fitting of the model. The training set is divided in distinct subgroups. A series of models are fitted on reduced datasets constructed by omitting each subgroup in its turn from the training set and each fitted model is applied on the left subgroup for predicting the Y response. In this way, a prediction is associated to every molecule of the original training set using a model trained on other data.

Let's denote $\hat{y}_i^{CV}$ the prediction of the activity response of the i^{th} molecule obtained by cross-validation. The classical index to summarise the internal predictive power is the cross-validated R_Y^2 defined as

$$Q_Y^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i^{CV})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (1.19)$$

This is a cross-validation estimate of the percentage of the variance of the activity response explained by the model. The higher Q_Y^2 is, the more predictive the model is. The value of Q_Y^2 must be compared to the value of R_Y^2 (see equation (1.15)). If they are not of the same order and the difference $R_Y^2 - Q_Y^2$ is larger than 0.2 or 0.3, this indicates that the model overfits the training data or that there are some outliers (Eriksson *et al.* 2003).

There is two main practices in defining the subgroups. The *leave-one-out* method considers each compound as a subgroup. Each compound is omitted in its turn. The model is fitted on the $N - 1$ other observations of the training set and the activity response of the remaining compound is predicted using the obtained model. To simulate predictions for new molecules not used in the fitting of the model, the leave-one-out principle is the most intuitive idea as the most information as possible is taken into account in the training of the model by removing only one observation. But leave-one-out is too optimistic and tends to overestimate the predictive power for independent new compounds.

The alternative is to delete more than one compound at each turn by defining subgroups with up to 50% of the data set. This method is called *leave-many-out*. The reasonable number of compounds to be included in each subgroup depends on the sample size. Let's denote G the number of groups, generally selected between 2 and 10. $\frac{N}{G}$ is the size of groups. Every group in its turn is left out of the training set and the model is fitted on the $N - \frac{N}{G}$ other compounds. This fitted model is then applied to the remaining group to predict the responses of its $\frac{N}{G}$ compounds.

Another technique to quantify the predictive power is the *bootstrap*. N compounds are drawn at random with replacement from the original dataset. Some of the original molecules might appear more than once in the resample while other molecules might not be included. The model is fitted on this resample and applied on the remaining compounds to predict the endpoint. Let's denote $\hat{y}_i^{(b)}$ the prediction of the activity response of the i^{th} molecule obtained with this bootstrap model. An estimate of the percentage of the variance of the activity response explained by the model is

$$Q_Y^{2(b)} = 1 - \frac{\sum_i (y_i - \hat{y}_i^{(b)})^2}{\sum_i (y_i - \bar{y}^{(b)})^2}. \quad (1.20)$$

This resampling is repeated a great number of times. Then the predictive power can be summarised by averaging the $Q_Y^{2(b)}$'s of each bootstrap model.

A final tool to insure that the model has not been obtained by chance correlation is the response permutation testing or *Y*-scrambling (Eriksson *et al.* 2003). The *N*-vector of observed responses is permuted at random and the model is fitted on these permuted responses keeping intact the matrix of observed descriptors **X**. The two indexes R_Y^2 and the cross-validated Q_Y^2 are computed. This permutation procedure is repeated a great number of times providing reference distributions of both R_Y^2 and Q_Y^2 . If the R_Y^2 and Q_Y^2 of the initial model are higher than the values of the simulated distributions, this constitutes a strong indication that the initial model correlation was not obtained by chance.

The Q_Y^2 , either computed by cross-validation (1.19) or by bootstrapping (1.20), is the most used index to quantify the internal predictive power of a QSAR model. Using leave-one-out cross-validation, Eriksson *et al.* (2001) suggested that $Q_Y^2 > 0.5$ is good and $Q_Y^2 > 0.9$ is excellent. But Golbraikh and Tropsha (2002) recommended to beware of Q_Y^2 , especially if it is computed by leave-one-out cross-validation. Indeed, the only thing that is sure is that a small value of Q_Y^2 is the sign of a poor predictive power. But a high value of Q_Y^2 does not imply automatically a high predictive power for new compounds not used in the training set. The right way to estimate the predictive power is to test the model on a sufficiently large external data set.

1.7 External performance

1.7.1 Global external performance

According to Doweyko (2004), only about half of the QSAR models published in the last decade made reasonable predictions about test compounds not used to create the model. Tropsha *et al.* (2003) pointed out that the quality of QSAR models is often typically measured on the training set alone, but this approach does not necessarily generate good predictive QSAR models. The most demanding way to quantify the predictive power of a QSAR model should be based on an external data set, by making predictions for an independent set of data, not used in the model calibration. This provides a more rigorous evaluation of the model predictive capability for untested chemicals than cross-validation or bootstrap on the training set.

An external data set may be found in the QSAR literature, but this is quite rare. Indeed, such data should concern the same endpoint, measured with the same protocol and the whole data set with all the computed descriptors should be known. This kind of data is often not publically available. If it is, there is always a risk of interlaboratory and assay

variability among the different data sources.

If no similar data can be found in the QSAR literature, an external data set can be obtained by splitting the original data into a training set and a test set. The training set is the set of data which is used to construct the QSAR model and the test set is the set of data which is used to validate the QSAR model. The best splitting of the original data set can be found by an experimental design procedure (Marengo and Todeschini 1992; Eriksson *et al.* 2000).

Whatever the origin of the external set is, it has first to be checked that this external set is composed of compounds similar to the training data (*cfr* Section 8 on applicability domain) and contains no outlier. Then the estimated model can be applied to predict the activity response of each of the N' molecules in the external set. Those predictions ($\hat{y}'_i$, $i = 1, \dots, N'$) are finally compared to the observed responses (y'_i , $i = 1, \dots, N'$) graphically and using

$$R_Y^{2(ext)} = 1 - \frac{\sum_{i=1}^{N'} (y'_i - \hat{y}'_i)^2}{\sum_{i=1}^{N'} (y'_i - \bar{y}')^2}. \quad (1.21)$$

Setting aside an external data set is not reasonable when the number of available data is small. If no similar data can be found in the QSAR literature, only internal predictivity can be assessed using cross-validation or bootstrap technique. But this overestimates the predictive power for new molecules.

Like any other kind of models, a QSAR model is never better than the series of measurements it was obtained from. As Aristotle was saying: "It is the mark of an instructed mind to rest easy with the degree of precision which the nature of the subject permits and not to seek an exactness where only an approximation of the truth is possible." This means that, whatever is the degree of validation of the model, the QSAR user has to keep in mind that it is never possible to predict the activity response of an unknown chemical with absolute certainty. This is then crucial to measure the uncertainty of the predictions provided by QSAR models and verify that this uncertainty is adequate for the purpose of the QSAR application.

1.7.2 Local external performance

Once the predictive power of the QSAR model has been assessed on an external test set, it can be used to predict the activity of a new compound. Only a small part of QSAR papers mention the uncertainty of such pointwise prediction. This subsection reviews this rare QSAR literature on prediction uncertainty for a given new compound. The next chapter reviews the statistical literature and presents tools that can be used to quantify such

local uncertainty with MLR, PLSR or RT predictions.

For MLR, a 95% confidence interval for the expected activity value of the new compound $E[Y|\mathbf{x}]$ is centered around its estimated value $\mathbf{z}'\hat{\boldsymbol{\beta}}$ and is constructed as $\mathbf{z}'\hat{\boldsymbol{\beta}} \pm t_{N-K}^{0.975} \sqrt{s^2 \mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}}$. A 95% prediction interval for the activity of the new compound $Y|\mathbf{x}$ is centered around its predicted value $\mathbf{z}'\hat{\boldsymbol{\beta}}$ and is constructed as $\mathbf{z}'\hat{\boldsymbol{\beta}} \pm t_{N-K}^{0.975} \sqrt{s^2(1 + \mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z})}$.

In the report from the expert group on (Q)SAR, OECD (2004), the standard deviation of the residuals of the external set $s^{(ext)} = \sqrt{\frac{1}{N'-K} \sum_{i=1}^{N'} (y_i - \hat{y}_i)^2}$ is computed and used to construct a 95% prediction interval for the activity response y_0 of any new molecule as $\hat{y}_0 \pm z_{0.975} \cdot s^{(ext)}$. This seems not correct as the standard error of prediction is considered constant over the chemical domain although it is certainly greater at the edges than in the center of the experimental domain. But no other QSAR paper evokes the subject of prediction uncertainty of a new particular molecule.

1.8 Applicability domain

1.8.1 Definition of the applicability domain

Wold and Dunn (1983) have shown that QSAR models normally have local validity only. They can embrace only compounds with similar chemical and biological properties as the compounds in the training set. Whereas excellent fit to the training data may be attainable, often QSAR models fail to predict accurately chemicals that differ substantially from the training set molecules. In order to state clearly this intrinsic limitation of any QSAR model, the OECD recommended, in the report from the expert group on (Q)SAR, OECD (2004), to define the applicability domain of the QSAR model. This way, any chemist that wants to use a QSAR model for the prediction of a new compound with descriptors $\mathbf{x}_0$ is able to first check if it is included in its applicability domain and the corresponding prediction is reliable.

The applicability domain is the group of substances for which the QSAR model is valid. Any new compound that lies in the chemical space beyond this boundary may possess a different structure-activity relationship than the molecules in the training set and the prediction provided by the QSAR model may be unreliable.

The applicability domain can be defined regarding the descriptors space and/or the response as well as the chemical class, the mechanism of action or the species considered. In practice, an applicability domain is defined for the descriptors space and another applicability domain is defined for the response. The QSAR model can then be applied for a new molecule if it is included in the two applicability domains.

Simple tools such as the descriptors ranges can be used or more complex tools such as the convex hull, the leverages and the William plot, some other distance measurements (Euclidean, Mahalanobis distance or DModX), the Hotelling T^2 or density measurements. Their exact definitions can often become complex due to the highly multivariate nature of the data.

The different definitions are given below and applied on a simple QSAR data set provided by Veith and Mekenyan (1993). An MLR was developed to model the 96h LC_{50} of organic chemicals to a fish species, the fathead minnow (*Pimephales promelas*). The 96h LC_{50} is the concentration (in moles per litre) causing 50% lethality in *Pimephales promelas* after an exposure of 96 hours. Two descriptors are used to model the 96h LC_{50} : K_{OW} , the octanol-water partition coefficient and E_{LUMO} , the energy of the lowest unoccupied molecular orbital. The training set is composed of 114 observations and an MLR model was fitted on those data, transforming the response and the first descriptor with the logarithm function: $\hat{\log}(LC_{50}) = -2.414 + 0.579\log(K_{OW}) - 0.473E_{LUMO}$.

1.8.2 Applicability domain based on the ranges

The most simple definition for the applicability domain can be achieved using the ranges of the descriptors (see Figure 1.6). The range of a variable is, by definition, the difference between the observed maximum and minimum. Based on this idea, a new compound is considered to belong to the applicability domain if its computed descriptors $\mathbf{x}_0$ are within the minimum and the maximum of each descriptor. If the QSAR model uses only one descriptor, the applicability domain reduces to a simple interval. With two descriptors, it is a rectangle and with $K > 2$ descriptors, the applicability domain is a K -dimensional hyper-rectangle.

This definition of the applicability domain takes the descriptors into account only separately. The minimum and the maximum of each descriptor are listed and sometimes also an histogram of observed values is constructed for each descriptor. But the applicability domain should take into account the overall distribution of the data in the structure space. In addition, this definition relies on the assumption that the data are uniformly distributed. This is not always the case with QSAR data as observations are often clustered and the K -dimensional hyper-rectangle defined with the ranges contains often empty spaces, especially in the edges.

An alternative consists in applying first PCA and then defining the applicability domain on the ranges of the principal components scores. The resulting hyper-rectangle may also include empty spaces, but the volume enclosed will be most of the time less empty than

in the original descriptors ranges case, especially if the descriptors are correlated. This is not the case with the data of Veith and Mekenyan as $\log(K_{OW})$ and E_{LUMO} are not at all correlated (see Figure 1.6).

1.8.3 Applicability domain based on the convex hull

To insure no extrapolation, the convex hull of the descriptors in the training set can be computed (see Figure 1.7). The convex hull is defined as the smallest convex polyhedra that contains the data. If a new molecule with descriptors $\mathbf{x}_0$ is outside this boundary, the prediction may be unreliable.

Efficient algorithms for convex hull calculation are available for dimension 2 or 3 (Graham 1972; Jarvis 1973). But the complexity increases rapidly in higher dimension. Moreover, this approach does not consider the global distribution of the data as the convex hull definition is only based on the boundary of the data set. This does not prevent from empty spaces in the convex hull.

1.8.4 Applicability domain based on the Euclidean, Mahalanobis or L_1 distance

The intuitive idea of distance-based applicability domain is that a compound too far from the compounds in the training set may be badly predicted. The distance of a new molecule to the training set may be computed as its distance to the mean of the elements of the training set, the averaged distance between this new molecule and all the molecules of the training set or the maximum distance between the new molecule and all the molecules of the training set. A definition of the distance between two substances is, for instance, already used with the k NN modelling technique. Each compound is represented by a K -vector of observed transformed descriptors $\mathbf{x}$. The distance between two substances with descriptors $\mathbf{x}_1$ and $\mathbf{x}_2$ can be defined using the Euclidean distance (d^E), the Mahalanobis distance (d^M) or the L_1 -distance (d^{L_1}) defined as follows (see respectively Figure 1.8, 1.9 and 1.10):

$$d^E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{k=1}^K (z_{1k} - z_{2k})^2}$$
$$d^M(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2) S^{-1} (\mathbf{x}_1 - \mathbf{x}_2)'$$
$$d^{L_1}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{k=1}^K |z_{1k} - z_{2k}|$$

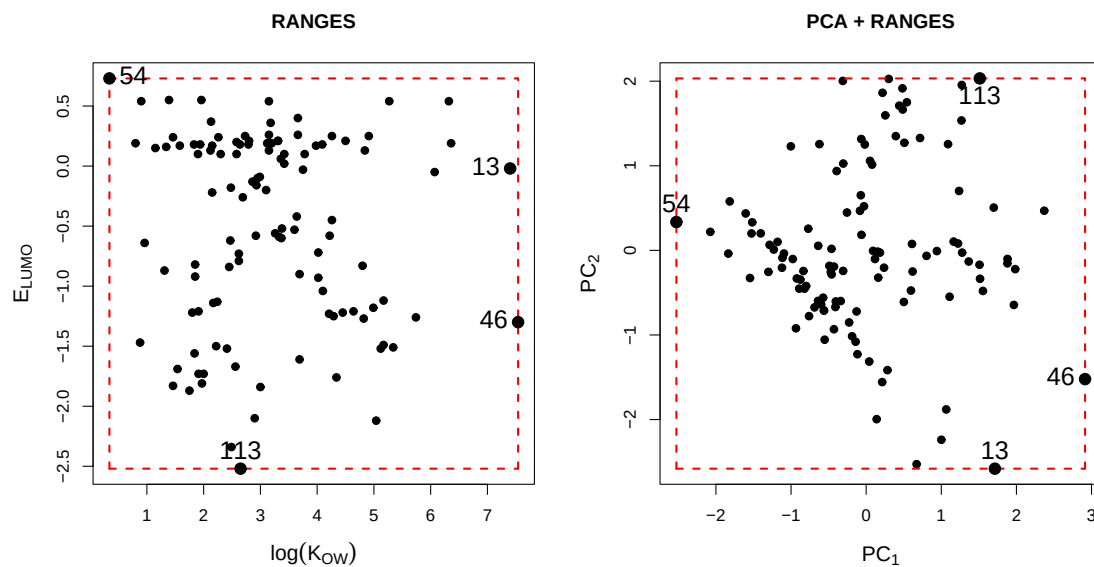


Figure 1.6: Applicability domain based on the ranges of the two observed descriptors (left) or on the ranges of the two principal components scores (right). Data of Veith and Mekenyan (1993).

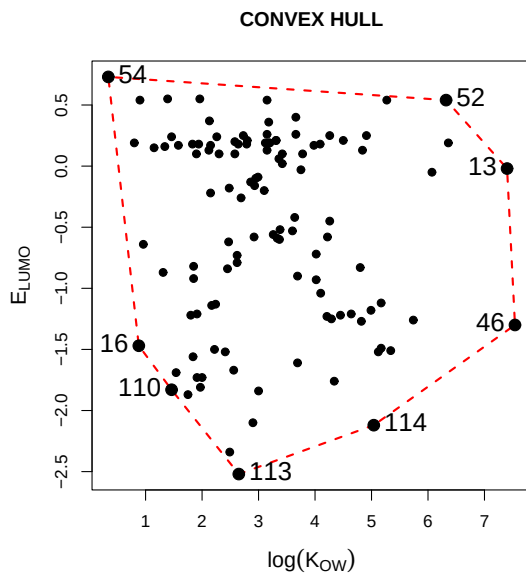


Figure 1.7: Applicability domain based on the convex hull of the 114 observed pairs of descriptors. Data of Veith and Mekenyan (1993).

As S denotes the covariance matrix of the observed transformed descriptors, the Mahalanobis distance takes into account the correlation between the transformed descriptors which is not the case with the two other definitions. For Euclidean and L_1 distance, it is necessary to standardise the descriptors ($\mathbf{z}$) otherwise the descriptors with wider scales would have higher weights in the distance calculation. For those two distances, it is also recommended to apply PCA and measure the distances in the principal components space where new variables are uncorrelated.

A new molecule with descriptors $\mathbf{x}_0$ is considered outside of the applicability domain of the QSAR model if its distance to the molecules of the training set exceeds a certain threshold fixed by the QSAR developer. This threshold can be defined as, for instance, the largest distance of any molecule in the training set to the rest of the molecules.

The Euclidean and Mahalanobis distances are specially adapted for observations that are normally distributed Seber (1984); Fukunaga (1990); Jaworska *et al.* (2005a). The L_1 distance assumes that the observations are uniformly distributed. Any of those three distances take into account possible empty regions in the structure space.

1.8.5 Applicability domain based on leverages

The leverage is another measure of the distance between a new molecule and the molecules in the training set. As explained in Section 5, the leverage (1.13) is originally defined in the MLR literature to detect outliers in the training set. This definition can be extended to any new molecule with the standardised descriptors contained in the K -vector $\mathbf{z}_0$ by $h_0 = \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0$.

The leverage of each molecule in the data set can be represented on a graph referred to as the William plot (report from the expert group on (Q)SAR, OECD 2004). The William plot is a scatterplot representing the leverages and the standardised residuals in prediction as defined in Section 5 (see Figure 1.11). This graph allows to detect outliers and defines the boundary of the applicability domain with the limits 2 or 3 for the standardised residual in prediction and $2K/N$ or $3K/N$ for the leverage.

1.8.6 Applicability domain based on DModX

For PLSR or PCR, another tool is used to detect outliers and define the applicability domain: DModX, the distance from one observation to the model in the space of original explanatory variables space (SIMCA 1998). More formally, for the i^{th} compound of the training set, DModX is computed as $DModX_i = \sqrt{\frac{1}{K-M} \sum_{k=1}^K e_{ik}^2}$ where e_{ik} is the element ik of the residual matrix $\mathbf{E} = \mathbf{X}^c - \mathbf{T}_M \mathbf{W}_M^{-1}$, $\mathbf{T}_M$ contains the M latent variables or PCs

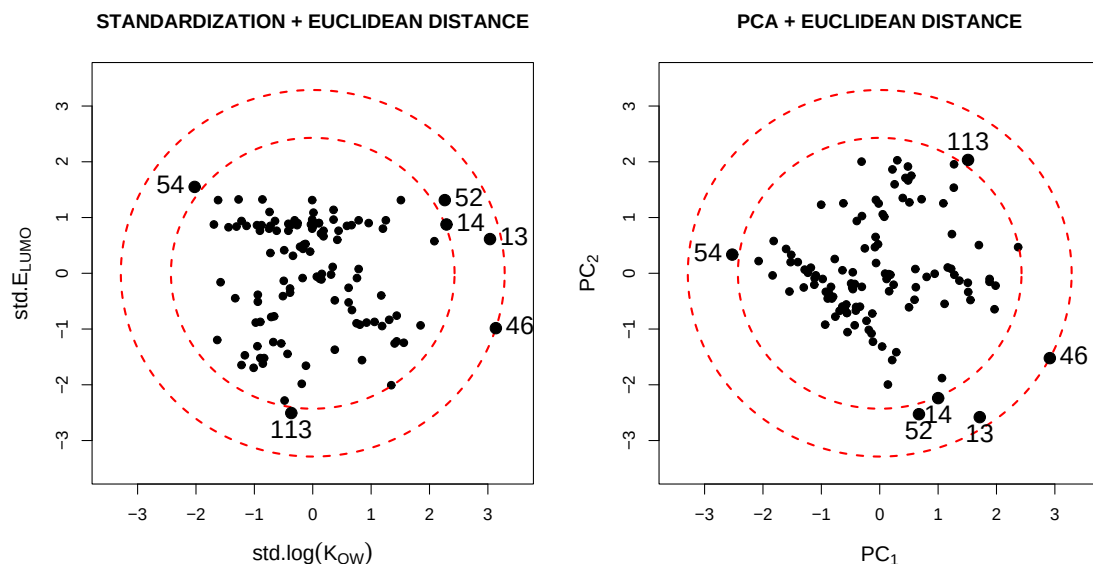


Figure 1.8: Applicability domain based on the Euclidean distance to the center of the dataset after a standardisation of the two descriptors (left) or after the transformation to the two PCs scores (right). The inner (resp. outer) dashed circle is the applicability domain considering a maximal Euclidean distance such that it contains 95% (resp. 100%) of the data.

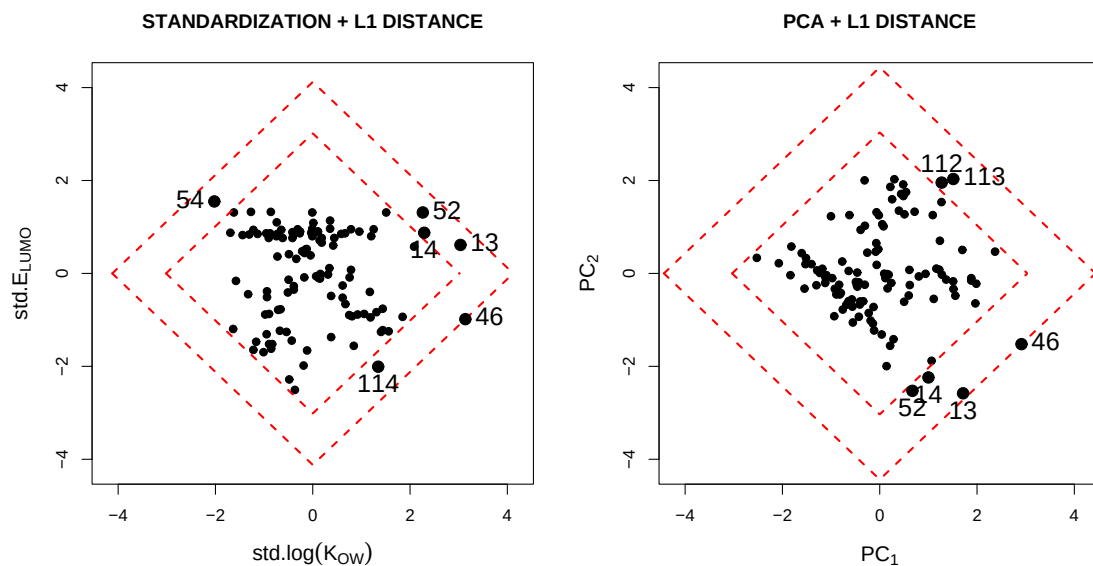


Figure 1.9: Applicability domain based on the L_1 -distance to the center of the data after a standardisation of the two descriptors (left) or after the transformation to the two PCs scores (right). The inner (rep. outer) dashed circle is the applicability domain considering a maximal L_1 -distance such that it contains 95% (resp. 100%) of the data. Data of Veith and Mekenyan (1993).

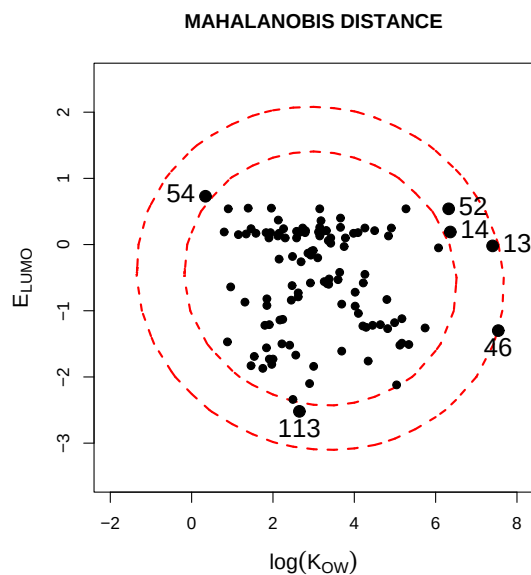


Figure 1.10: Applicability domain based on the Mahalanobis distance to the center of the data. The inner dashed circle is the applicability domain considering a maximal Mahalanobis distance such that it contains 95% of the data and the outer circle corresponds to 100% (the largest Mahalanobis distance). Data of Veith and Mekenyan (1993).

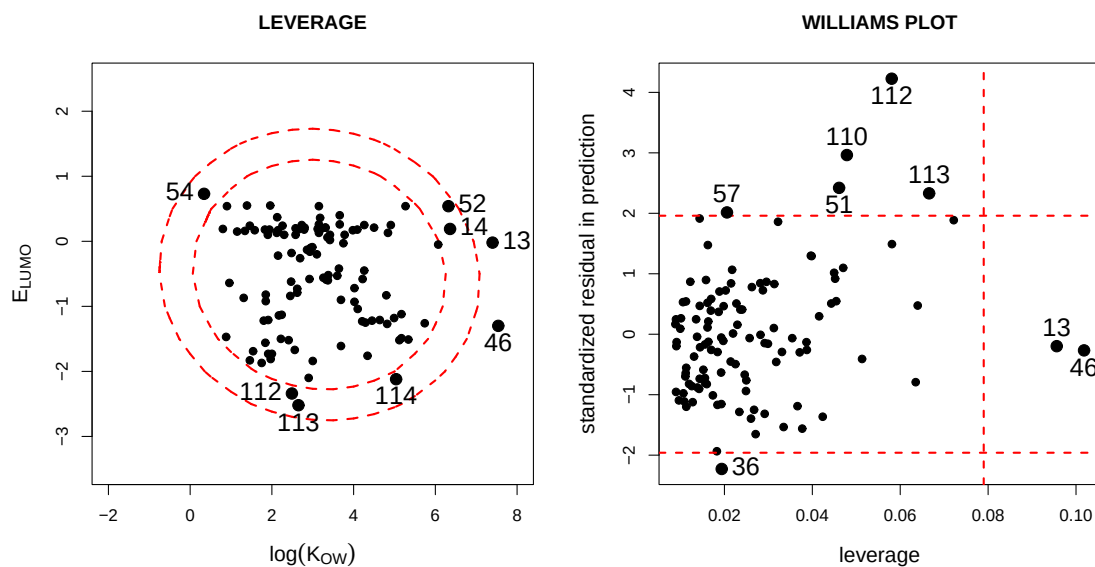


Figure 1.11: Left: Applicability domain based on the leverages. The inner dashed circle corresponds to the limit $2K/N$ and the outer circle to $3K/N$. Right: Williams plot. The vertical dashed line represents the limit $3K/N$ and the two horizontal dashed lines are the limits in ± 2 . Data of Veith and Mekenyan (1993).

scores vectors of the model and $\mathbf{W}_M$ is the weight matrix used to construct the latent variables as linear combination of the original variables $\mathbf{X}^c$. This definition can be generalised to any new compound for which the descriptors can be computed as well as the scores in the new latent subspace.

A high value of $DModX$ indicates that the compound is quite different from the compounds in the training set. An applicability domain can be defined by normalising the distances and considering as extrapolating points observations having standardised distances greater than 3.

This definition of the applicability domain is not illustrated on the data of Veith and Mekenyan (1993) because the two latent variables or the two principal components are significant in both PLSR and PCR models and the dimension of the data is not reduced.

1.8.7 Applicability domain based on the Hotelling's T^2

Another distance tool to detect strong outliers and define the applicability domain is given by the Hotelling's T^2 (Jackson 1991). This statistic is a multivariate generalisation of Student's t-test, and provides a check for observations adhering to multivariate normality. For the i^{th} compound of the training set, the Hotelling's T^2 is given by $T_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})'$ where $\mathbf{x}_i$ is the observed K -vector of descriptors, $\bar{\mathbf{x}}$ is the K -vector of mean descriptors and $\mathbf{S}$ is the variance-covariance matrix of the observed descriptors in the training set. This definition can be generalised to any other new molecule using the corresponding observed descriptors $\mathbf{x}_0$. If the descriptors are centered, the Hotelling's T^2 is the same distance as the Mahalanobis one and the leverage.

A high value of T_i^2 denotes a compound that is far from the central molecule (see Figure 1.12). The 95% significance limits of the Hotelling distribution (proportional to the Fisher) correspond to 95% of the data and can detect extreme compounds. Assuming normality of the data, the Hotelling's T^2 can be computed with any modelling technique but is mainly used with PLSR and PCR.

1.8.8 Applicability domain based on the density

Most of the tools used for the definition of applicability domain assume implicitly an underlying distribution of the data, either uniform or normal. Those assumptions are not realistic with QSAR data as such data are often clustered, with empty regions in the descriptors space. Nonparametric density estimation is an approach capable of identifying such empty regions (Jaworska *et al.* 2005a;b).

The most well-known method for nonparametric density estimation is the kernel-based method. To simplify the multivariate density estimation, a PCA is first applied to the QSAR data to obtain orthogonal variables. Jaworska *et al.* (2005a;b) proposed to estimate each PC density using kernel estimators and take their product.

The applicability domain is then defined by high density regions of level $(1 - \alpha)$, with $0 < \alpha < 1$ (see Figure 1.13). A $(1 - \alpha)$ high density region is the smallest multidimensional region comprising $(1 - \alpha) * 100$ percents of the probability mass. In practice, delimiting exactly this region may be quite complicated with high dimensional data.

1.8.9 Comparison of different applicability domains on a QSAR model example

Jaworska *et al.* (2005a;b) compared the different cited approaches and concluded that the range-based definition with a preliminary PCA rotation is the most adapted practical approach. In addition, only the PCA transformation and the ranges need to be stocked to be able to check if any new molecule is in the applicability domain of the QSAR model.

The report from the expert group on (Q)SAR, OECD (2004), recommend that further work is carried out to define more precisely the tools that can be used to define QSAR applicability domain. There is a need to reconcile the method used for the model development and the applicability domain definition. Whatever is the statistical tool used, the exact definition of the applicability domain should be stated. If a new compound reveals to be outside the applicability domain, this information should be considered as a warning. Indeed, there is always the possibility that the QSAR model extrapolates correctly.

Nearly all the presented methods are only applicable for quantitative descriptors. Jaworska *et al.* (2005a;b) suggest to divide the data set according to the levels of categorical descriptors and construct an applicability domain for each level. Only the $L1$ -distance or the leverages can manage dummy variables for encoding the categorical descriptors.

1.9 Conclusion

QSAR modeling is an emerging tool in pharmaceutical industry. MLR, PLSR and RT are the three most often used classes of models. NN is becoming also more and more popular. In practice, it is recommended to fit different model families, to enter different subsets of descriptors, and select the “best” model. There is an increasing concern in statistically validating QSAR models and measuring their performance. The coefficient of determination (R^2) and its cross-validated version (Q^2) are the two most important statistical tools

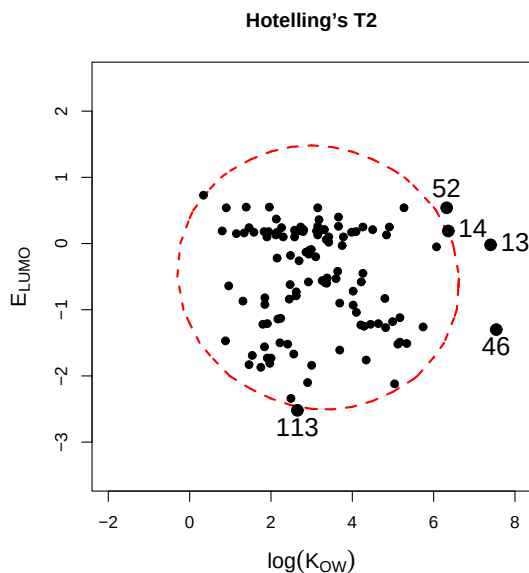


Figure 1.12: Applicability domain based on the Hotelling's T^2 . Data of Veith and Mekenyan (1993).

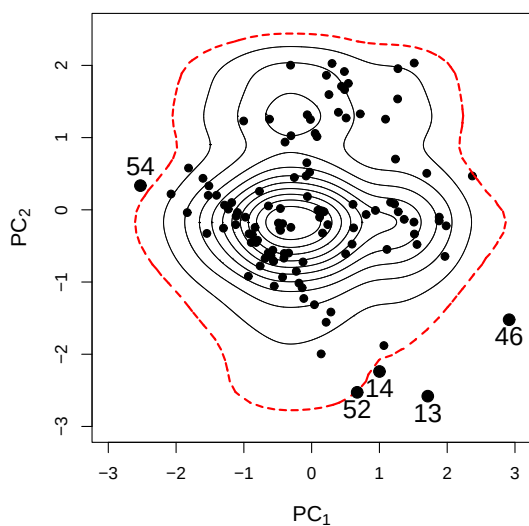


Figure 1.13: Applicability domain based on the joint density estimated using kernels after the transformation to the two principal components scores. The outer dashed line represents the high density region of level 0.95. The other inner lines represent different levels of density. Data of Veith and Mekenyan (1993).

to measure respectively the goodness-of-fit and the predictive power of the model. Nevertheless, the importance of measuring the predictive power on an independent test set is not yet ingrained in mind of all practitioners.

The definition of the applicability domain of a QSAR model is encouraged to be able to check if the QSAR model is valid for new molecules and provides reliable predictions. However, in practice, a QSAR model is still often used for the prediction of new compounds without comparing them to the training set and without measuring the reliability of those local predictions. Chapter 2 is aimed to fill in this gap in quantifying prediction error.

Acknowledgment

We express our gratitude to Eli Lilly and Company for its support to attend the conference on validation of QSAR models and methods for applicability domain estimation organised by the UK QSAR and Chemoinformatics Group.

Bibliography

- De Aguiar, P.F., Bourguignon, B., Khots, M.S., Massart, D.L. and Phan-Than-Luu, R. (1995), "D-Optimal Designs", *Chemometrics and Intelligent Laboratory Systems*, 30, 199-210.
- Andrea, T.A. and Kalayeh, H. (1991), "Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors", *Journal of Medicinal Chemistry*, 34, 2824-2836.
- Atkinson, A.C. (1985), *Plots, transformations and regression*, Oxford:Clarendon Press.
- Baroni, M., Costantino, G., Cruciani, G., Riganelli, D. , Valigi, R. and Clementi S. (1993), "Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems", *Quantitative Structure-Activity Relationships*, 12(1), 9-20.
- Baurin, N., Mozziconacci, J.C., Arnoult, E., Chavatte, P., Marot, C. and Morin-Allory, L., (2002), "Insights into 2D-QSAR consensus prediction Virtual Screening of a 1992430 compounds database", available on the web at [http : //www.univ – orleans.fr/SCIENCES/ICOA/communications/com2002/baurin.pdf](http://www.univ-orleans.fr/SCIENCES/ICOA/communications/com2002/baurin.pdf)
- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), *Regression diagnostics: Identifying influential data and sources of collinearity*, New-York, John Wiley and Sons.
- Benigni, R. (2003), *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, Ed., CRC Press, Boca Raton, FL, USA.
- Bishop, C.M. (1986), *Neural Networks for Pattern Recognition*, Oxford University Press.
- Blockeel, H., Dzeroski, S., Kompare, B., Kramer, S., Pfahringer, B. and Van Laer W. (2004), "Experiments In Predicting Biodegradability", *Applied Artificial Intelligence*, 18(2), 157-181.
- Box, G.E.P., Hunter, W.G., and Hunter, J.S., (1978), *Statistics for Experimenters*, John Wiley & Sons, Inc., New York.

- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and regression trees*, Belmont, CA: Wadsworth International Group.
- Breiman, L. (1996), *Bagging predictors*, Machine Learning, 24:123-140.
- Consonni, V. Mauri, A. and Pavan M. (2005), *DRAGON, version 5 (2005)*, Milano, Italy. Program for the calculation of molecular descriptors from HyperChem, Tripos, MDL file, SYBYLmolfile formats from ChemOffice and Tripos molecular design software. Download available at: http://www.taletе.mi.it/main_exp.htm.
- Crum-Brown, A. and Fraser, T. R. (1989), "On the connection between chemical constitution and physiological action. Part I. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia", *Transactions of the Royal Society of Edinburgh*, 25, 151.
- Cronin, M.T.D. (2000), "Computational methods for the prediction of drug toxicity", *Current Opinion in Drug Discovery & Development*, 3, 292-297.
- Cronin, M.T.D. (2002), "The current status and future applicability of quantitative structure-activity relationships (QSARs) in predicting toxicity", *Alternatives to Laboratory Animals*, 30(2), 81-84.
- Cronin, M.T.D. and Schultz, T.W. (2003), "Pitfalls in QSAR", *Journal of Molecular Structure*, 622, 39-51.
- Cronin, M.T.D., Jaworska, J., Walker, J.D., Comber, M., Watts, C.D., and Worth, A.P. (2003), "Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances", *Environmental Health Perspectives*, 111, 1391-1401.
- Cronin, M.T.D., Walker, J.D., Jaworska, J., Comber, M., Watts, C.D. and Worth, A.P. (2003), "Use of QSAR relationships in international decision-making frameworks to predict health effects of chemical substances", *Environmental Health Perspectives*, 111, 1376-1390.
- Cronin, M.T.D. and Livingstone, D.J. (2004), *Predicting Chemical Toxicity and Fate*, Ed., CRC Press, Boca Raton, FL, USA.
- Dearden, J.C., Barratt, M.D., Benigni, R., Bristol, D.W., Combes, R.D., Cronin, M.T.D. et al. (1997), "The development and validation of expert systems for predicting toxicity—the report and recommendations of an ECVAM/ECB workshop (ECVAM workshop 24)", *Alternatives to laboratory animals*, 25, 223-252.
- Devillers, J. (1996), *Neural networks in QSAR and drug design*, London: Academic Press.

- Doweyko, A.M. (2004), "3D-QSAR illusions", *Journal of computer-aided molecular design*, 18(7-9), 587-596.
- Draper, N.R. and Smith, H. (1981), *Applied regression analysis*, 2nd edition. New York: John Wiley.
- Duprat, A.F., Huynh, T. and Dreyfus, G. (1998), "Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of logP", *Journal of Chemical Information and Computer Sciences*, 38, 586-594.
- Dzeroski, S. (2001), "Applications of symbolic machine learning to ecological modelling", *Ecological Modelling*, 146, 263-273.
- Efron, B. (1982), *The jackknife, the bootstrap and other resampling planes*, Philadelphia : Society for Industrial and Applied Mathematics.
- Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, London:Chapman & Hall.
- Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (1996), "Multivariate Design and Modelling in QSAR", *Chemometrics and Intelligent Laboratory Systems*, 34, 1-19.
- Eriksson, L., Johansson, E. and Wold, S. (1997), QSAR model validation. In: *Quantitative Structure-Activity Relationships in Environmental Sciences-VII*, (Chen, F., Schüürmann, G., eds), Pensacola, FL:SETAC Press, 381-397. Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C., and Wold, S., (2000),
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C. and Wold, S. (2000), *Design of Experiments - principles and applications*, Umetrics AB.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. (2001), *Multi- and megavariable data analysis - principles and applications*, Umetrics AB.
- Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (2003), "Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs", *Environmental Health Perspectives*, 111(10), 1361-1375.
- European Commission (1996), "Technical Guidance Document in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation (EC) No 1488/94 on Risk Assessment for Existing Substances", Luxembourg: European Commission, Office for Official Publications of the European Communities, <http://ecb.jrc.it>.

- Fujita, T., Iwasa, J. and Hansch, C. (1964), "A new substituent constant, p , derived from partition coefficients", *Journal of the American Chemical Society*, 86, 5175-5180.
- Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, Computer Science and Scientific Computing Series. New York, NY, USA: Academic Press.
- Golbraikh, A. and Tropsha, A. (2002), "Beware of q^2 !", *Journal of Molecular Graphics and Modelling*, 20(4), 269-276.
- Graham, R. (1972), "An efficient algorithm for determining the convex hull of a finite point set", *Information Processing Letters*, 1, 132-133.
- Gramatica, P., Navas, N. and Todeschini, R. (1998), "3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs)", *Chemometrics and Intelligent Laboratory Systems*, 40, 53-63.
- Gramatica, P., Consonni, V. and Todeschini, R. (1999), "QSAR study of the tropospheric degradation of organic compounds", *Chemosphere*, 38, 1371-1378.
- Gramatica P, Corradi M, Consonni V. (2000), "Modeling and prediction of soil sorption coefficients of non-ionic organic pesticides by different sets of molecular descriptors", *Chemosphere*, 41, 763-777.
- Gramatica, P. and Papa, E. (2003), "QSAR modeling of bioconcentration factor by theoretical molecular descriptors", *Quantitative Structure-Activity Relationships*, 22, 374-385.
- Hammett, L.P. (1937), "The effect of structure upon the reactions of organic compounds. Benzene derivatives", *Journal of the American Chemical Society*, 59, 96.
- Hansch, C., Maloney, P.P., Fujita, T., and Muir, R.M. (1962), "Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients", *Nature*, 194, 178-180.
- Hansch, C., Muir, R.M., Fujita, T., Maloney, P.P., Geiger, C.F. and Streich, M. (1963), "The Correlation of Biological Activity of Plant Growth-Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients", *Journal of the American Chemical Society*, 85, 2817-2824.
- Hansch, C., Leo A., Suresh Babu Mekapati and Kurup A. (2004), "QSAR and ADME", *Bioorganic & Medicinal Chemistry*, 12(12), 3391-3400.
- Hansch, C., and Fujita, T. (1964), "R-s-p analysis - A method for the correlation of biological activity and chemical structure", *Journal of the American Chemical Society*, 86, 1616-1626.

- Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The elements of statistical learning. Data Mining, inference and predictions*, Springer.
- Hawkins, D.M., Basak, S.C. and Shi, X. (2001), "QSAR with Few Compounds and Many Features", *Journal of Chemical Information and Computer Sciences*, 41, 663-670.
- Hoffman, B. , Cho, S.J. and Zheng, W. (1999), "Quantitative structure-activity relationship modeling of dopamine D1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and k nearest neighbor methods", *Journal of Medicinal Chemistry*, 42, 3217-3226.
- Hopfinger, A.J. and Tokarski, J.S. (1997), *Practical applications of computer-aided drug design. In: Practical Applications of Computer-Aided Design*, (Charifson PS, ed) New York:Marcel Dekker, 105-164.
- Hulzebos, E.M., Schielen, P.C.J.I. and Wijkhuizen-Maslankiewicz, L. (1999), "(Q)SARs for human toxicological endpoints: a literature search", *RIVM Report 601516.00. A report by RIVM (Research for Man and Environment)*. Bilthoven:RIVM.
- Huuskonen, J.J., Livingstone, D.J. and Tetko, I.V. (2000), "Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices", *Journal of Chemical Information and Computer Sciences*, 40, 947-955.
- Izrailev, S. and Agrafiotis, D. (2001), "A novel method of building regression tree models for QSAR based on artificial ants", *Journal of Chemical Information and Computer Sciences*, 41(1), 176-180.
- Jackson, J.E. (1991), *A User's Guide to Principal Components*, Wiley: New York.
- Jarvis, R.A. (1973), "On the Identification of the Convex Hull of of a Finite Set of Points in the Plane", *Information Processing Letters*, 2, 18-21.
- Jaworska, J.S., Comber, M., Auer, C. and Van Leeuwen, C.J. (2003), "Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints", *Environmental Health Perspectives*, 111(10), 1358-1360.
- Jaworska, J.S., Aldenberg, T. and Nikolova, N. (2005), "Review of methods for assessing the applicability domains of SARs and QSARs", available on the web site of the European Chemicals Bureau at <http://ecb.jrc.it/home.php?CONTENU=/DOCUMENTS/QSAR/>
- Jaworska, J.S., Nikolova, N. and Aldenberg, T. (2005), "QSAR applicability domain estimation by projection of the training set in descriptor space: a review", *Alternatives to Laboratory Animals*, 33, 445-459.

- de Jong, S. (1993), "SIMPLS: An alternative approach to partial least squares regression", *Chemometrics and Intelligent Laboratory Systems*, 18, 251-263.
- Jurs, P.C. (2003), *ADAPT—Automated Data Analysis and Pattern Recognition Toolkit*, University Park, PA: Pennsylvania State University. Available: <http://research.chem.psu.edu/pcjgroup/ADAPT.html> [accessed 23 April 2002].
- Katritzky, A.R., Lobanov, V.S. and Karelson, M. (1994), *CODESSA, Reference Manual*, Gainesville, FL: University of Florida. Available: <http://www.semichem.com/codessarefs.html> [accessed 19 April 2002].
- Kovesdi, I., Dominguez-Rodriguez, M.F. and Orfi, L. (1999), "Application of neural networks in structure-activity relationships", *Medicinal Research Reviews*, 19, 249-269.
- Kubinyi, H. (1994), "Variable selection in QSAR studies. I. An evolutionary algorithm", *Quantitative Structure-Activity Relationships*, 13, 285-294.
- Kubinyi, H. (1996), "Evolutionary variable selection in regression and PLS analyses", *Journal of Chemometrics*, 10, 119-133.
- Kubinyi, H., Folkers, G. and Martin, Y.C. (1998), "3D QSAR in drug design—recent advances", *Perspectives in Drug Discovery and Design*, 12, R5-R7.
- Le Bailly de Tilleghe, C. and Govaerts, B. (2007), *A review of quantitative structure-activity relationship (QSAR) models*, Discussion Paper 0719, Institut de Statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium.
- Linusson, A. (2000), *Efficient Library Selection in Combinatorial Chemistry*, Ph.D. Thesis, Umea University, Umea, Sweden.
- Livingstone, D. (1995), *Data Analysis for Chemists. Applications to QSAR and Chemical Product Design*, Oxford Science Publications, Oxford.
- Livingstone, D.J. (2000), "The characterization of chemical structures using molecular properties. A survey.", *Journal of Chemical Information and Computer Sciences*, 40, 195-209.
- Loehlin, J.C. (1998), *Latent variable models: an introduction to factor, path, and structural analysis*, Hillsdale, NJ: Lawrence Erlbaum Associates..
- Manallack, D.T. and Livingstone D.J. (1999), "Neural networks in drug discovery: Have they lived up to their promise?", *European Journal of Medicinal Chemistry*, 34, 195-208.
- Marengo, E. and Todeschini, R. (1992), "A new algorithm for optimal, distance-based experimental design", *Chemometrics and Intelligent Laboratory Systems*, 16, 37-44.

- Martens, H. (1985), *Multivariate Calibration*, Dr. techn. Thesis. Technical 25 University of Norway, Trondheim
- Massey, W. F. (1965), "Principal components regression in exploratory statistical research", *Journal of the American Statistical Association*, 309, 234-256.
- Mason, R. L. and Gunst, R. F. (1985), "Selecting principal components in regression", *Statistics and Probability Letters*, 3, 299-301.
- Mekenyan, O. and Bonchev, D. (1986), "OASIS method for predicting biological activity of chemical compounds", *Acta Pharmaceutica Jugoslavica*, 36, 225-237.
- Meyer, H. (1899), "Zur Theorie der Alkoholnarkose. 1. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung?", *Archiv für experimentelle Pathologie und Pharmakologie*, 42, 109.
- Myers, R.H. (1986), *Classical and modern regression with applications*, Boston: Duxbury Press.
- Oprea, T.I., Mannhold, R., Kubinyi, H. and Folkers, G (2005), *Cheminformatics in Drug Discovery*, eds. Hoboken, NJ: John Wiley & Sons, Inc.
- Overton, E. (1897), "Ueber die Osmotischen Eigenschaften der Zelle in ihrer Bedeutung für die Toxicologie und Pharmakologie", *Zeitschrift für physik. Chemie*, 22, 189.
- OECD series on testing and assessment No.49 (November 2004), "The report from the expert group on (Quantitative) Structure-Activity Relationship [(Q)SARs] on the principles for the validation of (Q)SARs", *ENV/JM/MOMNO(2004)24. Paris, France, OECD*
- Richardson, J. (1868), *Medical Times and Gazette*, 2, 703.
- Richet, C. (1893), *Compt. Rend. Seances Soc. Biol.*, 9, 775.
- Schultz, T.W. and Cronin, M.T.D. (2003), "Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships", *Environmental Toxicological and Chemistry*, 22, 599-607.
- Schultz, T.W., Cronin, M.T.D., Walker, J.D. and Aptula, A.O. (2003), "Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective", *Journal of Molecular Structure*, 622, 1-22.
- Seber, G.A.F. (1984), *Multivariate Observations*, New York, NY, USA: John Wiley & Sons.
- Selick, H.E., Beresford A.P. and Tarbit, M.H. (2002), "The emerging importance of predictive ADME simulation in drug discovery", *Drug Discovery Today*, 7(2), 7-109.

- SIMCA 7.0. A new standard in multivariate data analysis*, Manual, Umetrics AB, Umea, Sweden.
- Sokal, R.R. and Rohlf, F.J. (1995), *Biometry: The Principles and Practice of Statistics in Biological Research*, Freeman, New York.
- Stuper, A.J. and Jurs, P.C. (1976), "ADAPT: A computer system for automated data analysis using pattern recognition techniques", *Journal of Chemical Information and Computer Sciences*, 16, 99-105.
- Taft, R.W.Jr. (1952), "Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters", *Journal of the American Chemical Society*, 74, 3120-3128.
- Test guideline 401, 2001, *Trends in Pharmacological Sciences*, 22(2), 2375-2388.
- Therneau, T.M. and Atkinson, E.J. (1997), "An introduction to recursive partitioning using the RPART routines", Technical Report #61. Mayo Clinic. Rochester, Minnesota.
- Todeschini, R. and Consonni, V. (2000), *Handbook of molecular descriptors*, Wiley-VCH: Weinheim (Germany).
- Todeschini, R. and Gramatica, P. (1997), "3D-modelling and prediction by WHIM descriptors. Part 6. Applications of WHIM descriptors in QSAR studies", *Quantitative Structure-Activity Relationships*, 16, 120-125.
- Tong, W., Fang, H., Hong, H., Xie, Q., Perkins, R., Anson, J. and Sheehan, D. (2003), "Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: A perspective", *Pure and Applied Chemistry*, 75(11-12), 2375-2388.
- Tong, W., Xie, Q., Hong, H., Shi, L., Fang, H. and Perkins, R. (2004), "Assessment of prediction confidence and domain extrapolation of two structure activity relationship models for predicting estrogen receptor binding activity", *Environmental Health Perspectives*, 112(12), 1249-1254.
- Tropsha, A., Gramatica, P. and Gombar V.K. (2003), "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models", *QSAR & Combinatorial Science*, 22(1), 69-77.
- Tutz, G. (1990), "Smoothed categorical regression based on direct kernel estimates", *Journal of Statistical Computation and Simulation*, 36, 139-156.
- Veith, G.D. and Mekenyan, O.G. (1993), "A QSAR approach for estimating the aquatic toxicity of soft electrophiles", *Quantitative Structure-Activity Relationships*, 12, 349-356.

- Vellman, P.F. and Welsch, R. (1981), "Efficient computing of regression diagnostics", *The American Statistician*, 35, 234-242.
- Vinod, H.D. and Ullah, A. (1981), *Regression Methods*, New York: Marcel Dekker.
- Walker, J.D., Rodford, R. and Patlewicz, G. (2002), "Quantitative structure-activity relationships for predicting percutaneous absorption rates", *Environmental toxicology and chemistry*, 22(8), 1870-1884.
- Waller, C.L. and Bradley, M.P. (1999), "Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies", *Journal of Chemical Information and Computer Sciences*, 39, 345-355.
- Weininger, D. (1988), "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules", *Journal of Chemical Information and Computer Sciences*, 28(1), 31-36.
- Werbos, P. (1974), *Beyond regression: New tools for prediction and analysis in the behavioral sciences*, Ph.D. dissertation, Dept. of Applied Mathematics, Harvard University.
- Whitley, D.C., Ford, M.G. and Livingstone, D.J. (2000), "Unsupervised forward selection: A method for eliminating redundant variables.", *Journal of Chemical Information and Computer Sciences*, 40, 1160-1168.
- Worth, A. (2002), "ECVAM's Activities on Computer Modelling and Integrated Testing", *Alternatives to Laboratory Animals*, 30(2), 133-137.
- Wold, H. (1966), *Estimation of principal components and related models by iterative least squares*, in *Multivariate analysis* Academic Press, New York.
- Wold, S., Martens, H. and Wold H. (1982), *The multivariate calibration problem in chemistry solved by the PLS method*, Proceedings of the Conference on Matrix Pencils, March 1982, (A. Ruhe and B. Kagstrom, eds), Lecture Notes in Mathematics, 286-293. Heidelberg: Springer Verlag.
- Wold, S. and Dunn, W.J. (1983), "Multivariate Quantitative Structure-Activity Relationships (QSAR): Conditions for Their Applicability", *Journal of Chemical Information and Computer Sciences*, 23, 6-23.
- Wold, S., Sjöström, M., Carlson, R., Lundstedt, T., Hellberg, S., Skagerberg, B., Wikström, C., and Öhman, J. (1986), "Multivariate Design", *Analytica Chimica Acta*, 191, 17-32.

- Worth, A.P. and Cronin, M.T.D (2004), "Report of the workshop on the validation of QSARs and other computational prediction models", *Alternatives to Laboratory Animals*, 32(1), 703-706.
- Xu, L., and Zhang, W.J. (2001), "Comparison of different methods for variable selection", *Analytica Chimica Acta*, 446, 477-483.
- Wold, S. and Dunn, W.J. (1983), "Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle", *Journal of Chemical Information and Computer Sciences*, 40, 185-194.

Chapter 2

Assessing prediction error in Quantitative Structure-Activity Relationship (QSAR) models

We have the project to submit part of this chapter in a chemometrics journal with an illustration on a published QSAR model.

2.1 Introduction

Chapter 1 reviews the chemometrical literature about QSAR models development and validation. Even if there is a growing interest for a statistical validation of QSAR models, in practice, most of the times, only goodness-of-fit measures (the (adjusted) coefficient of determination, the standard error of estimate, the Fisher statistic) are used. The predictive power of the QSAR model is sometimes quantified (cross-validated or bootstrapped Q^2), but the reliability of the prediction provided by a QSAR model for a new given molecule is never assessed.

This chapter reviews the statistical literature about “error” when applying fitted models for prediction. Let $Y = g(\mathbf{x}) + \epsilon$ be the true relationship between the activity response Y and the chemical descriptors $\mathbf{x}$ with the error term ϵ supposed to have zero mean and constant variance σ_ϵ^2 . QSAR models, like any statistical model, attempt to find a useful function $f(\mathbf{x})$ for predicting the activity Y given values of the descriptors $\mathbf{x}$. This assumed link function $f(\mathbf{x})$ is fitted on a training set $\tau = \{(\mathbf{x}_i, y_i), i = 1, 2, \dots, N\}$, providing the predictor $\hat{y} = \hat{f}(\mathbf{x})$.

This chapter focuses on Multiple Linear Regression (MLR), Partial Least Squares Regression (PLS) and Regression Tree (RT) models, the three most often used classes of QSAR models. The aim is to understand the different sources of error when predicting the activity response or estimating its expected value for a new molecule with descriptors $\mathbf{x}_0$ by $\hat{f}(\mathbf{x}_0)$ and to provide tools to measure each source.

2.2 Expected Prediction Error (EPE)

Using a squared error loss, Hastie *et al.* (2001) define the Expected Prediction Error (abbreviated *EPE*) of the assumed link function $f(\mathbf{X})$ as

$$EPE(f) = E_{\mathbf{X},Y}[(Y - f(\mathbf{X}))^2] = E_{\mathbf{X}}E_{Y|\mathbf{X}}[(Y - f(\mathbf{X}))^2|\mathbf{X}]. \quad (2.1)$$

The first expectation is taken over the random vector $(\mathbf{X}, Y)$ and is then splitted using conditional density into two parts, $Y|\mathbf{X}$ and $\mathbf{X}$. Indeed, the descriptors $\mathbf{X}$ are real valued random inputs and conditionally on $\mathbf{X}$, the response Y is a random variable due to the error term ϵ . Note that, in QSAR modeling, the descriptors of a given molecule are deterministic but the descriptors are considered as random variables because they vary over the chemical space.

The best function $f(\mathbf{X})$ for predicting Y given values of $\mathbf{X}$ is defined as the one that minimises the EPE (2.1). This reduces to a pointwise minimisation problem and it can be proven that the best constant function $f(\mathbf{x})$ to approximate the response Y for given values of the inputs $\mathbf{x}$ is

$$f(\mathbf{x}) = \operatorname{argmin}_c E_{Y|\mathbf{x}}[(Y - c)^2|\mathbf{x}] = E[Y|\mathbf{x}], \quad (2.2)$$

i.e. the expected response given $\mathbf{x}$.

According to the model class, different assumptions are made about the unknown expectation $E[Y|\mathbf{x}]$. With MLR models, $E[Y|\mathbf{x}]$ is assumed to be linear, having the form $\mathbf{z}'\beta$. The parameters β are chosen in order to minimise the EPE (2.1). As the expectation over $\mathbf{X}$ is not known, the exact minimum cannot be found and is estimated by minimising the training error, *i.e.* the average residual sum of squares observed over the training set τ

$$\hat{\beta} = \operatorname{argmin}_{\beta} Err_{tr}(\beta) = \operatorname{argmin}_{\beta} \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{z}'_i \beta)^2. \quad (2.3)$$

The resulting estimated linear predictor is then

$$\hat{f}(\mathbf{x}) = \mathbf{z}'\hat{\beta} = \mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}. \quad (2.4)$$

With RT models, $E[Y|\mathbf{x}]$ is assumed to be constant in a region defined by the successive splits of the inputs space around $\mathbf{x}$. As the expectation over $\mathbf{X}$ is also unknown, the exact minimum of the EPE (2.1) can be estimated by minimising the training error. $E[Y|\mathbf{x}]$ is then estimated by the average observed responses in the same node

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^M \bar{y}_m I_m(\mathbf{x}) \quad (2.5)$$

where $I_m(\mathbf{x})$ is the indicator of the node and $\bar{y}_m$ is the average observed value at node m .

As the assumed link function $f(\mathbf{x})$ is fitted on a training set τ , there is an additional source of randomness induced by τ and the global EPE of the fitted predictor $\hat{f}(\mathbf{x})$ is

$$EPE(\hat{f}) = E_{\mathbf{X}} E_{Y|\mathbf{X}} E_{\tau} [(Y - \hat{f}(\mathbf{X}))^2 | \mathbf{X}] \quad (2.6)$$

The expectation over τ averages over all possible training sets of size N . Indeed, different training sets would provide different predictors $\hat{f}$.

For most classes of assumed models, an estimate of EPE (2.6) may be necessary to select the best value for the complexity parameters and the resulting function $\hat{f}(\mathbf{x})$. For MLR, the number of descriptors selected among all available descriptors (and transformed descriptors) to be entered into the model has to be chosen. For PLSR, the number of entering latent variables has to be fixed. For RT, the tree depth has to be chosen.

It is quite intuitive that increasing model complexity (*e.g.* more descriptors, more latent variables, deeper partitioning) may reduce the bias of the predictor but may increase its variance. The training error is not a good estimate of the EPE as it decreases if the model complexity increases. On the opposite, it is expected that the error on an independent test set may not always decrease with model complexity but, for some level of complexity, the prediction error may start to increase again. This bias-variance tradeoff represented on Figure 2.1 is reviewed in Section 3. Estimates of the global EPE of a predictor (2.6) are proposed in Section 4. Such estimates may be also useful to assess model quality. But different estimates must be used for the two different tasks, model selection and model assessment, otherwise there is a clear risk to overestimate the predictive power of the model. Often, in practice, analytical criteria such as AIC or BIC are used for model selection and resampling techniques are applied for model assessment.

One may also be interested in measuring the error when using the estimated predictor $\hat{f}(\mathbf{x}_0)$ for a new molecule having descriptors $\mathbf{x}_0$. The Expected Prediction Error for a new data point $\mathbf{x}_0$ is defined as

$$EPE(\mathbf{x}_0) = E_{Y|\mathbf{x}_0} E_{\tau} [(Y - \hat{f}(\mathbf{x}_0))^2 | \mathbf{x}_0] \quad (2.7)$$

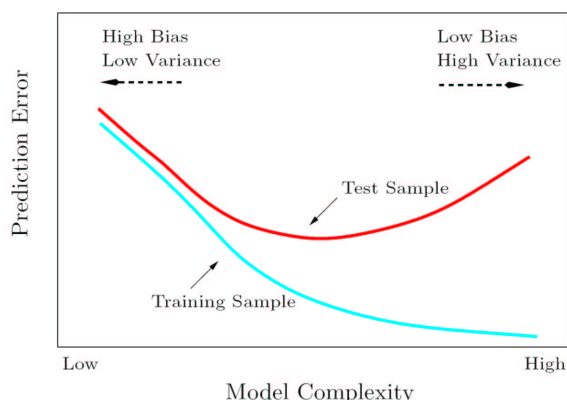


Figure 2.1: Behavior of test sample and training sample error as the model complexity is varied.

The expectation is taken over two sources of randomness, $Y|\mathbf{x}_0$ and the training set τ . Estimates for this local quantity are reviewed in Section 5.

2.3 Bias-variance tradeoff

Let $\hat{f}(\mathbf{x})$ be a regression fit for the assumed link $f(\mathbf{x})$. Hastie *et al.* (2001) proved that the EPE (2.7) for a new molecule having descriptors $\mathbf{x}_0$ can be decomposed in three sources

$$\begin{aligned}
 EPE(\mathbf{x}_0) &= E_{Y|\mathbf{x}_0} E_{\tau} [(Y - \hat{f}(\mathbf{x}_0))^2 | \mathbf{x}_0] \\
 &= E_{Y|\mathbf{x}_0} [(Y - g(\mathbf{x}_0))^2 | \mathbf{x}_0] + (E_{\tau}[\hat{f}(\mathbf{x}_0)] - g(\mathbf{x}_0))^2 + E_{\tau}[(\hat{f}(\mathbf{x}_0) - E_{\tau}[\hat{f}(\mathbf{x}_0)])^2] \\
 &= \sigma_{\epsilon}^2 + Bias_{\tau}^2[\hat{f}(\mathbf{x}_0)] + V_{\tau}[\hat{f}(\mathbf{x}_0)].
 \end{aligned} \tag{2.8}$$

The first term is the variance of the activity of the new molecule around its true expectation (left graph of Figure 2.2). This noise does not depend on the training set τ and the quality of the fit. It is the consequence of the inner randomness of the activity response (*e.g.* experimental variance) and can not be avoided. The second term is the squared bias of the predictor, *i.e.* the difference between the average of the predictor and the true average of the activity (central graph of Figure 2.2). The last term is the variance over possible training sets of the fitted predictor around its mean (right graph of Figure 2.2).

This bias itself may arise from two sources: the model bias $f(\mathbf{x}_0) - g(\mathbf{x}_0)$, *i.e.* the difference between the assumed average and the true average of the activity, and the estimation bias $E_{\tau}[\hat{f}(\mathbf{x}_0)] - f(\mathbf{x}_0)$, *i.e.* the difference between the average of the predictor and the assumed average. Those different sources of error are represented on Figure 2.3, extracted from the book of Hastie *et al.* (2001).

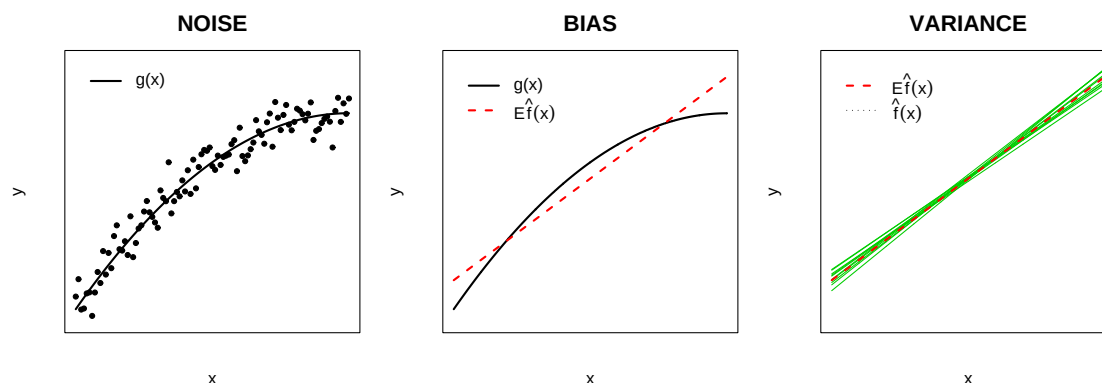


Figure 2.2: Three sources to the Expected Prediction Error: the irreducible noise, the bias and the variance of the predictor.

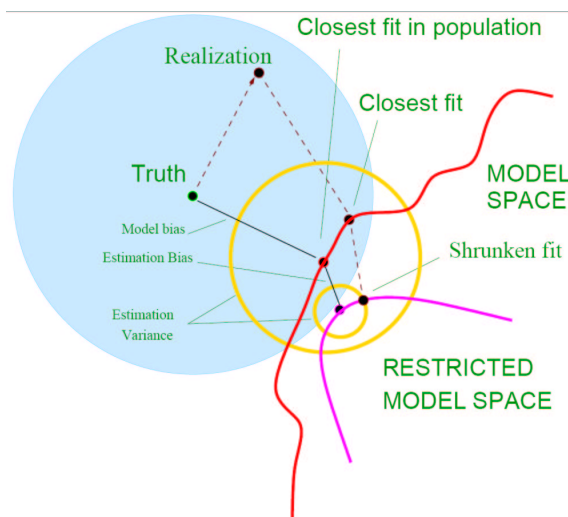


Figure 2.3: Schematic of the behavior of bias and variance. The model space is the set of all possible predictions from the model, with the “closest fit” labeled with a black dot. The model bias from the truth is shown, along with variance, indicated by the large yellow circle centered at the black dot labeled “closest fit in population”. A shrunken or regularized fit is also shown, having additional estimation bias, but smaller prediction error due to its decreased variance.

For MLR models, as a consequence of the analytical expression of the predictor, the three sources of error for a new molecule can be identified as

$$EPE(\mathbf{x}_0) = \sigma_\epsilon^2 + (E_\tau[\mathbf{z}_0' \hat{\beta}] - g(\mathbf{x}_0))^2 + V_\tau[\mathbf{z}_0' \hat{\beta}] \quad (2.9)$$

with the least squares estimate $\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ depending on the training set τ through the matrix of observed transformed descriptors $\mathbf{Z}$ and the vector of observed responses $\mathbf{y}$,

and $V_\tau[\mathbf{z}_0'\hat{\beta}] = \sigma_\epsilon^2 \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0$. By averaging (2.9) over the observations of the training set τ , the average squared bias is $\frac{1}{N} \sum_{i=1}^N (E_\tau[\mathbf{z}_0'\hat{\beta}] - g(\mathbf{x}_0))^2$ and the average variance is $\frac{K}{N}\sigma_\epsilon^2$. One can conclude that entering more descriptors in the MLR model (increasing K) increases the variance of the resulting predictor. On the other hand, this allows the predictor to reach nearly interpolation and obviously decreases the bias.

One may think of selecting complexity parameters such that the tradeoff between bias and variance is optimum. This tradeoff is measured by the sum of the two terms in (2.8) depending on the fitted model, called *Mean Squared Error of Prediction*

$$MSEP(\hat{f}(\mathbf{x})) = Bias_\tau^2[\hat{f}(\mathbf{x})] + V_\tau[\hat{f}(\mathbf{x})]. \quad (2.10)$$

MSEP has to be minimised. Different estimates of the MSEP are discussed in the next section.

2.4 Estimates of the global Expected Prediction Error

The global EPE of a predictor (2.6) is a measure of generalisation error, quantifying what the average prediction error would be on an independent test set. It includes the three sources of error, the irreducible noise, the squared bias and the variance. If the interest is to measure error when estimating the expected activity response given values of descriptors, $E[Y|\mathbf{x}]$, the noise term vanishes and $EPE(\hat{f})$ (2.6) is reduced to $MSEP(\hat{f}(\mathbf{x}))$ (2.10).

An ideal situation would be to have enough data to split them into three separate sets: a training set of size N_{tr} to fit the models, a validation set of size N_v to select the best model and a test set of size N_{te} to assess its predictive power, with $N_{tr} + N_v + N_{te} = N$. Different estimates of the prediction error can be defined on each subset by computing the observed average squared error

$$Err_{tr} = \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} (y_i - \hat{f}(\mathbf{x}_i))^2 \quad (2.11)$$

$$Err_v = \frac{1}{N_v} \sum_{i=1}^{N_v} (y_i - \hat{f}(\mathbf{x}_i))^2 \quad (2.12)$$

$$Err_{te} = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} (y_i - \hat{f}(\mathbf{x}_i))^2. \quad (2.13)$$

The training error (2.11) is used to fit the assumed predictor f . With MLR model, minimising (2.11) allows to estimate the parameters β . With PLSR model, (2.11) is used to

define the first latent variable. The best successive splits in an RT are selected to minimise (2.11).

The validation error (2.12) is used to select the best level of complexity parameters. The number of descriptors entering the MLR model, the number of latent variables entering the PLSR model or the depth of the RT can be chosen to minimise (2.12).

The test error (2.13) estimates $EPE(\hat{f})$ and assesses the predictive power of the fitted predictor $\hat{f}$ for new data neither used to train nor select the model.

But in practice, the dimensions of QSAR datasets do not usually allow this splitting in training, validation and test sets. The number of observations is quite small compared to the huge number of descriptors and all those observations are necessary to train the assumed model. But the training error can not be used to select complexity parameters (overfitting risk) or to assess predictive power (overly optimistic to do that on the same data as those used for training). Analytical criteria such as *AIC* and *BIC* are reviewed in the next subsection as tools to select the best model and trade bias and variance off. Then resampling techniques such as cross-validation and bootstrapping are presented to assess the predictive power of the selected model.

The training error, the test error, the *AIC* and *BIC* are estimates of the $EPE(\hat{f})$, including the three sources of error, noise, bias and variance. They are mainly used to select complexity parameters and choose the best model. Cross-validation and bootstrap can be used to estimate $MSEP(\hat{f}(\mathbf{x}))$ and assess the predictive power of the selected model.

2.4.1 Model selection: *AIC* and *BIC*

The principle behind the definition of *AIC* and *BIC* is the same. The training error is obviously too optimistic in estimating the average prediction error for new data $EPE(\hat{f})$ (2.6). Indeed, additional uncertainty in the generalisation error arises from the fact that the predictor may be used for a new molecule with descriptors different from those in the training set τ . The two proposed criteria are constructed by adding an estimate of this optimism to the training error.

More precisely, let's define the in-sample error as the average prediction error for N new observed responses Y_i^{new} at each of the training points $\mathbf{x}_i$, $i = 1, 2, \dots, N$, by

$$Err_{in} = \frac{1}{N} \sum_{i=1}^N E_{Y^{new}} [(Y_i^{new} - \hat{f}(\mathbf{x}_i))^2].$$

The optimism of the training error is defined as the difference between this in-sample error and the expected training error given the observed responses in the training set $op = Err_{in} - E_{\mathbf{y}}(Err_{tr})$. For linear models with additive errors (like MLR and PLSR), it can be proved (Hastie *et al.* 2001) that the optimism depends on the link between the estimated predictor and the observed responses quantified through the equation

$$op = \frac{2}{N} \sum_{i=1}^N Cov(y_i, \hat{f}(\mathbf{x}_i)). \quad (2.14)$$

For an MLR model with K estimated parameters, $\sum_{i=1}^N Cov(y_i, \hat{f}(\mathbf{x}_i)) = K\sigma_{\epsilon}^2$ leads to an optimism of $op = 2\frac{K}{N}\sigma_{\epsilon}^2$. The definition of the Akaike Information Criterion

$$AIC = Err_{tr} + 2\frac{d}{N}\hat{\sigma}_{\epsilon}^2 \quad (2.15)$$

comes from a generalisation of this MLR results to the effective number of parameters, denoted d . For any linear fit of the form $\hat{y} = Sy$ where the $(N \times N)$ matrix S depends only on the observed descriptors $\mathbf{X}$ and not on the observed activity response, the effective number of parameters is defined as $d = trace(S)$.

For an MLR model, the effective number of parameters is the number of estimated parameters $d = K$ (the AIC corresponds to the so-called Cp -statistic) and for a PLSR model, it is the number of latent variables $d = M$. One may think of using the number of nodes for the effective number of parameters of an RT.

The Bayesian Information Criterion is similar to AIC but has its origin in a Bayesian framework:

$$BIC = \frac{N}{\hat{\sigma}_{\epsilon}^2} [Err_{tr} + \frac{d \log(N)}{N} \hat{\sigma}_{\epsilon}^2]. \quad (2.16)$$

BIC is proportional to AIC but it results in more simple models than the AIC . It has better asymptotical properties like the consistency (BIC selects the correct model among a family of models with probability approaching 1 as N tends to ∞).

Both AIC and BIC are estimates of the in-sample error that can be used for model selection. To assess the model quality, other tools are necessary to estimate $EPE(\hat{f})$ (2.6), the average prediction error for new data.

2.4.2 Model assessment: cross-validation and bootstrap

The idea of cross-validation (abbreviated CV) or bootstrap (abbreviated BS) is to simulate a test set on the basis of the training set τ and estimate $MSEP(\hat{f})$ like the test error

(2.13). CV and BS can be applied for any class of predictor (Efron and Tibshirani 1993; Hjorth 1994; Shao and Tu 1995), like RT models (Breiman *et al.* 1984) or PLSR (Mevik and Cederkvist 2004) for which no exact analytical formula is available.

The principle of CV is simple, which makes this technique very popular in QSAR modeling. The training set τ is divided into V distinct subsets τ_v , $v = 1, 2, \dots, V$, of approximately equal size ($\#\tau_v \approx N/V$). The assumed model $f(\mathbf{x})$ is fitted on a series of reduced data constructed by omitting each subgroup in its turn from the training set, $\tau \setminus \tau_v$. Each resulting fitted model, $\hat{f}^{-v}(\mathbf{x})$, is applied on the left subset τ_v for predicting the activity of its molecules. This way, a prediction is associated to every molecule of the original training set using a model trained on other data and the $MSEP$ can be estimated by

$$\begin{aligned} MSEP_{CV} &= \frac{1}{V} \sum_{v=1}^V \frac{1}{\#\tau_v} \sum_{i \in \tau_v} (y_i - \hat{f}^{-v}(\mathbf{x}_i))^2 \\ &\approx \frac{1}{N} \sum_{v=1}^V \sum_{i \in \tau_v} (y_i - \hat{f}^{-v}(\mathbf{x}_i))^2. \end{aligned} \quad (2.17)$$

If V equals N , each molecule constitutes a subset and the assumed model $f(\mathbf{x})$ is fitted on the whole training set minus each observation at its turn. This is called *leave-one-out* CV, abbreviated LOO-CV. $MSEP_{LOO-CV}$ is a nearly unbiased estimate of $MSEP$ but its variance can be important. Decreasing V may reduce the variance of the estimate at the cost of a higher bias. Mevik and Cederkvist (2004) suggests that $V = 10$, called *10-fold* CV is a good compromise in practice.

As the assumed model $f(\mathbf{x})$ is fitted on a series of datasets $\tau \setminus \tau_v$ of size $N - N/V$ smaller than the initial size of the training set N , it is expected that $MSEP_{CV}$ overestimates $MSEP(\hat{f})$. The *adjusted* CV estimate is defined to correct this positive bias:

$$\begin{aligned} MSEP_{adj\ CV} &= MSEP_{CV} + MSEP_{adj} \quad \text{where} \\ MSEP_{adj} &\approx Err_{tr} - \frac{1}{N} \sum_{v=1}^V \frac{\#\tau_v}{N} \sum_{i \notin \tau_v} (y_i - \hat{f}^{-v}(\mathbf{x}_i))^2. \end{aligned} \quad (2.18)$$

$MSEP_{adj}$ measures the difference between the training error when the model is fitted over the whole training set τ and the weighted average of training error when the model is fitted over each training subset $\tau \setminus \tau_v$. It can be proved that $MSEP_{adj\ CV}$ has both a smaller bias and a smaller variance than $MSEP_{CV}$.

According to Efron (1983), bootstrapping performs better, in many cases, than cross-validation. There exist different BS techniques. In its simplest form (bootstrapping on

pairs), BS consists in selecting a random sample τ^* of size N with replacement from the training set τ . The assumed model $f(\mathbf{x})$ is fitted on this resample and applied to predict the activity response over the training set $\hat{f}^*(\mathbf{x}_i)$, $i = 1, 2, \dots, N$. This resampling process is repeated B times and the MSEP is naively estimated by averaging the training error over all resamples (τ^{*b} , $b = 1, 2, \dots, B$):

$$MSEP_{naive} = \frac{1}{B} \sum_{b=1}^B \frac{1}{N} \sum_{i=1}^N (y_i - \hat{f}^{*b}(\mathbf{x}_i))^2. \quad (2.19)$$

This estimate is expected to underestimate the true $MSEP$ because each BS predictor $\hat{f}^{*b}(\mathbf{x})$ is fitted and then partially tested on the same subset of observations. An improvement over this naive BS estimate can be obtained by avoiding this overlap similarly to what leave-one-out CV does. Indeed, some of the original molecules appear more than once in each resample τ^{*b} while other molecules might not be included. For each molecule in the original data set τ , let's denote by B_{-i} the number of resamples that do not include the i^{th} molecule ($i = 1, 2, \dots, N$). The corresponding B_{-i} BS predictors are used to predict the activity of the i^{th} molecule and those predictions are averaged. The *leave-one-out* bootstrap estimate finally averages over all compounds in the training set:

$$MSEP_{LOO-BS} = \frac{1}{N} \sum_{i=1}^N \frac{1}{B_{-i}} \sum_{b: i \notin \tau^{*b}} (y_i - \hat{f}^{*b}(\mathbf{x}_i))^2. \quad (2.20)$$

This LOO-BS estimate is expected to be less variable than the LOO-CV because $MSEP_{LOO-CV}$ is averaged over all B resamples.

But, like V -fold CV, $MSEP_{LOO-BS}$ may overestimate the true error because each BS predictor is fitted on a smaller subset of distinct observations than the original predictor that is fitted on the whole training set τ . As each BS sample contains on average 63.2% of distinct observations, Efron (1983) suggests to use this percentage to balance the LOO-BS estimate that overestimates $MSEP$ and the training error Err_{tr} that underestimates $MSEP$:

$$MSEP_{.632} = 0.632 MSEP_{LOO-BS} + (1 - 0.632) Err_{tr}. \quad (2.21)$$

The weight 0.632 is the limit of $1 - (1 - 1/n)^n$ when n tends to infinity, where $1 - (1 - 1/n)^n$ is the probability for a given point to be actually included in a given bootstrapped dataset. Mevik and Cederkvist (2004) concluded on a simulations study that either $MSEP_{LOO-CV}$ or $MSEP_{.632}$ should be used for estimating $MSEP$ to assess the predictive power of PLSR models.

Lendasse *et al.* (2005) has written an excellent recent review on CV and BS estimates of the generalisation error for regression model selection. Their main conclusion is that most

recent bootstrap methods (like .632 BS) are almost unbiased and also have a low variance.

Those different estimates of the *MSEP* are not used in the QSAR literature. The AIC (2.15) and the BIC (2.16) are sometimes used for model selection but the standard tool to assess the global predictive power of a QSAR model is the cross-validated Q^2 (1.19). Nevertheless, the definitions of the $MSEP_{CV}$ (2.17) and the Q^2 are quite similar as they are both based on the sum of squared differences between the observed responses and the cross-validated predictions. $MSEP_{CV}$ averages those differences, dividing this sum by N while Q^2 compares this average to the variance of the observed responses to obtain a dimensionless number. No statistical study of the properties of the cross-validated Q^2 has been performed yet, but Golbraikh and Tropsha (2002) suggested to beware of Q^2 because some QSAR models with high values of Q^2 revealed in practice to badly predict the activity response of new compounds.

2.5 Estimates of $V[\hat{E}(Y|\mathbf{x}_0)]$, the prediction variance for a new data point

In Section 3, the EPE of a regression fit for a new molecule $\hat{f}(\mathbf{x}_0)$ was decomposed in three sources, noise, squared bias and variance (see equation (2.8)). However, in QSAR literature, those three error components are neglected. Once the global predictive power of a QSAR model has been assessed, it is used for the prediction of new compounds after checking their inclusion in the applicability domain and nothing is said about the uncertainty of those predicted values. This section focuses on a general procedure, applicable for any class of QSAR model, to quantify the uncertainty of the estimation $\hat{E}(Y|\mathbf{x}_0) = \hat{f}(\mathbf{x}_0)$ for a new molecule with descriptors $\mathbf{x}_0$.

For any class of models, a procedure similar to LOO-BS, can be used to estimate the bias and the variance of a fitted predictor for each observed molecule of the training set $\tau = \{(\mathbf{x}_i, y_i), i = 1, 2, \dots, N\}$. Let's first select B bootstrapped samples from the training set, τ^{*b} , $b = 1, 2, \dots, B$. Then, the assumed model is fitted on each resampled training set τ^{*b} and used to predict the activity response of molecules in the rest of the original training set $\tau \setminus \tau^{*b}$. Finally, for each molecule in the original training set, $i = 1, 2, \dots, N$, the observed activity y_i can be compared to the collections of predicted values and the bias and the variance of the predictor can be respectively estimated by

$$\hat{Bias}(\hat{f}(\mathbf{x}_i)) = \overline{\hat{f}^*(\mathbf{x}_i)} - y_i \quad (2.22)$$

$$\hat{V}(\hat{f}(\mathbf{x}_i)) = \frac{1}{B-i-1} \sum_{b:i \notin \tau^{*b}} (\hat{f}^{*b}(\mathbf{x}_i) - \overline{\hat{f}^*(\mathbf{x}_i)})^2 \quad (2.23)$$

where $\overline{\hat{f}^*}(\mathbf{x}_i) = \frac{1}{B_{-i}} \sum_{b:i \notin \tau^{*b}} \hat{f}^{*b}(\mathbf{x}_i)$ is the average of the B_{-i} predictions for the i^{th} molecule corresponding to the B_{-i} resamples that do not include it.

Unfortunately, this presented BS-based procedure can not be used to estimate the bias and the variance of a predictor for a new molecule not observed in the training set as it requires to have an observed activity value. In the next point, a similar BS technique is applied to estimate the variance of the estimated expected activity of a new molecule, assuming no noise and no bias in the predictor $\hat{f}(\mathbf{x}_0) = \hat{E}[Y|\mathbf{x}_0]$.

2.5.1 Bootstrap estimate of $V[\hat{E}[Y|\mathbf{x}_0]]$

As proposed by Efron (1979), bootstrapping on pairs can be used to simulate the distribution of any estimator. First, B random samples of size N are drawn with replacement from the training set, τ^{*b} , and used to refit the model, $\hat{f}^{*b}(\mathbf{x})$, with $b = 1, 2, \dots, B$. Then, the bootstrapped predictors are applied to estimate the activity response of the new molecule, $\hat{f}^{*b}(\mathbf{x}_0)$. This simulates the distribution of the estimator $\hat{E}[Y|\mathbf{x}_0]$. Finally, the variance of the predictor for this new molecule can be estimated by the sample variance of the bootstrapped estimates

$$\hat{V}[\hat{E}[Y|\mathbf{x}_0]] = \frac{1}{B-1} \sum_{b=1}^B (\hat{f}^{*b}(\mathbf{x}_0) - \overline{\hat{f}^{*b}}(\mathbf{x}_0))^2 \quad (2.24)$$

where $\overline{\hat{f}^{*b}}(\mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(\mathbf{x}_0)$ is the average of bootstrapped estimate for the new molecule. As Hastie *et al.* (2001) noticed, (2.24) can be thought of as a Monte-Carlo estimate of $V[\hat{E}[Y|\mathbf{x}_0]]$ under sampling from the empirical distribution function for the data $(\mathbf{X}, Y)$. Efron and Tibshirani (1993) showed that B in the range 50 to 200 makes a good standard error estimator.

This BS estimate can be applied for any class of QSAR models. It is especially useful when the variance of prediction $V[\hat{E}[Y|\mathbf{x}_0]]$ can not directly be derived from the analytical expression of the estimator $\hat{E}[Y|\mathbf{x}]$. This question is analysed in the next subsections for MLR, PLSR or RT models.

2.5.2 Estimate of $V[\hat{E}[Y|\mathbf{x}_0]]$ for MLR

The Multiple Linear Regression model has been reviewed in the previous chapter (Chapter 1, Section 4.1). The expected activity response for a new molecule is estimated on the basis of the least squares fit by

$$\hat{E}[Y|\mathbf{x}_0] = \mathbf{z}_0' \hat{\beta} = \mathbf{z}_0' (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{y} \quad (2.25)$$

where $\mathbf{z}_0$ are the transformed descriptors of the new molecule, $\mathbf{Z}$ is the model matrix and $\mathbf{y}$ is the corresponding vector of observed responses. If there is no model bias, (2.25) is an unbiased estimator for the expected response $E[Y|\mathbf{x}_0]$.

As $\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is a constant that does not depend on $\mathbf{y}$, the MLR estimate (2.25) is a linear transformation of the observed responses $\mathbf{y}$ and the variance of prediction for a new molecule can be estimated by

$$\hat{V}[\hat{E}[Y|\mathbf{x}_0]] = \hat{\sigma}_\epsilon^2 \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 \quad (2.26)$$

where $\hat{\sigma}_\epsilon^2$ is an unbiased estimator of the residual variance comparing the observed and estimated responses by

$$\hat{\sigma}_\epsilon^2 = \hat{V}[Y|\mathbf{x}_0] = \frac{1}{N-K} \sum_{i=1}^N (y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}})^2. \quad (2.27)$$

Those three estimators (2.25-2.26-2.27) are the usual ones with an MLR model. Their derivations are demonstrated in many statistics books like the one of Hastie *et al.* (2001).

2.5.3 Estimate of $V[\hat{E}[Y|\mathbf{x}_0]]$ for PLSR

The Partial Least Squares Regression estimate (Chapter 1, Section 4.3) for the expected response of a new molecule has an analytical form close to the MLR one (2.25),

$$\hat{E}[Y^c|\mathbf{x}_0] = \mathbf{t}_{\mathbf{M}_0}'(\mathbf{T}_{\mathbf{M}}'\mathbf{T}_{\mathbf{M}})^{-1}\mathbf{T}_{\mathbf{M}}'\mathbf{y}^c. \quad (2.28)$$

where $\mathbf{t}_{\mathbf{M}_0} = \mathbf{V}_{\mathbf{M}}\mathbf{x}_0^c$ are the scores of the new molecule and $\mathbf{T}_{\mathbf{M}} = \mathbf{X}^c\mathbf{V}_{\mathbf{M}}'$ and $\mathbf{y}^c$ are respectively the scores of the observed molecules and the corresponding observed centered activity responses for the training set.

The term $\mathbf{t}_{\mathbf{M}_0}'(\mathbf{T}_{\mathbf{M}}'\mathbf{T}_{\mathbf{M}})^{-1}\mathbf{T}_{\mathbf{M}}'$ in (2.28) depends nonlinearly on $\mathbf{y}^c$ through the recursive algorithm for the construction of the latent variables. Nevertheless, it is a common practice (Höskuldsson 1988) to consider this term constant and estimate $V[\hat{E}[Y^c|\mathbf{x}_0]]$ similarly to (2.26) for MLR

$$\hat{V}[\hat{E}[Y^c|\mathbf{x}_0]] = \hat{\sigma}_\epsilon^2 \mathbf{t}_{\mathbf{M}_0}'(\mathbf{T}_{\mathbf{M}}'\mathbf{T}_{\mathbf{M}})^{-1}\mathbf{t}_{\mathbf{M}_0}. \quad (2.29)$$

Höskuldsson (1988) proposed a similar estimator as in classical MLR (2.27) for the noise variance σ_ϵ^2 , adjusting the degrees of freedom according to the number M of latent variables entered into the model and the centering of the responses:

$$\hat{\sigma}_\epsilon^2 = \hat{V}[Y^c|\mathbf{x}_0] = \frac{1}{N-M-1} \sum_{i=1}^N (y_i^c - \mathbf{t}_{\mathbf{M}_i}'(\mathbf{T}_{\mathbf{M}}'\mathbf{T}_{\mathbf{M}})^{-1}\mathbf{T}_{\mathbf{M}}'\mathbf{y}^c)^2. \quad (2.30)$$

The estimator (2.29) is sometimes presented as exact in the literature although it is based on a zeroth-order linearisation of the PLSR estimator (2.28). Phatak *et al.* (1993), Faber and Kowalski (1997) and Serneels *et al.* (2004) proposed to refine this approximation using the first-order linearisation of the PLSR estimator. This requires the expression of the Jacobian matrix of $\mathbf{t}_{\mathbf{M}_0}'(\mathbf{T}_{\mathbf{M}}'\mathbf{T}_{\mathbf{M}})^{-1}\mathbf{T}_{\mathbf{M}}'$ which is possible thanks to the approach to matrix differential calculus developed by Magnus and Neudecker (1988).

As there is no exact analytical expression for the variance of the PLSR estimator, bootstrap is also commonly used (Wehrens and van der Linden 1997). In addition, Denham (1997) suggested to derive the confidence interval for the expected activity response of the new molecule on the basis of the bootstrapped distribution of the estimator.

2.5.4 Estimate of $V[\hat{E}[Y|\mathbf{x}_0]]$ for RT

Consider that a Regression Tree with M terminal nodes has been fitted on a QSAR training set. The activity response for a new molecule is then estimated by the average response of the observations belonging to the same node

$$\hat{E}(Y|\mathbf{x}_0) = \sum_{m=1}^M \bar{y}_m I_m(\mathbf{x}_0) \quad (2.31)$$

where $I_m(\mathbf{x}_0)$ is the indicator of the m^{th} node (1 if the compound belongs to the node m ; 0 otherwise) and $\bar{y}_m$ is the average response value at node m .

This RT estimator is a nonlinear function of the observations $\mathbf{y}$ as the successive splits have been defined to minimise the residual sum of squares. Each indicator $I_m(\mathbf{x}_0)$ is a random variable depending on $\mathbf{y}$. As a consequence, no analytical formula exists for the variance of the RT estimator.

The only direct estimator proposed in the literature (Breiman *et al.* 1984) is the variance of the observed responses composing the node to which the new observation belongs to. But as the successive splits are constructed to minimise the within-node variance, this certainly underestimate the true variance.

Bootstrap is the only procedure to reliably quantify the uncertainty of the RT estimate (Breiman 1996). On each bootstrapped sample τ^{*b} ($b = 1, 2, \dots, B$), an RT is constructed according to the same principles of splitting and pruning as the original RT. Each bootstrapped RT is then used to estimate the activity of the new molecule of interest, $\hat{E}^{*b}(Y|\mathbf{x}_0)$. The variance of all the simulated estimations can be computed to estimate $V[\hat{E}(Y|\mathbf{x}_0)]$ like (2.24).

2.5.5 Bootstrap estimate *versus* analytical formula?

This question arises for MLR or PLSR models because a well-established estimator of $V[\hat{E}[Y|\mathbf{x}_0]]$ exists. The MLR estimator (2.26) and the PLSR estimator (2.29) have the great advantage in the QSAR context to provide quickly an estimate of the local prediction variance as only one formula has to be computed. On the opposite, the bootstrap estimate requires to refit the QSAR model B times (between 50 and 200) and can be computer intensive due to the huge dimensions of QSAR data.

Nevertheless, if the assumed MLR or PLSR model is biased, the estimator (2.26) and (2.29) are not reliable and it is better to use the bootstrap estimate, especially as bootstrapping on pairs is not sensitive to model assumptions. To limit the computations, the B bootstrapped versions of the QSAR model, $\hat{f}^{*b}(\cdot)$ with $b = 1, 2, \dots, B$, can be stored before starting to use the fitted model $\hat{f}(\mathbf{x})$ for the prediction of hundreds to thousands new compounds, in order to avoid repeating the resampling each time an estimate of $V[\hat{E}[Y|\mathbf{x}_0]]$ would be needed.

2.6 Estimates of $Cov[\hat{E}(Y|\mathbf{x}_1), \hat{E}(Y|\mathbf{x}_2)]$, the covariance of two data points estimations

In this last section, an estimate of the covariance between the estimated responses for two molecules is studied for the three most often used classes of QSAR models, *i.e.* MLR, PLSR and RT. Such an estimate is required to apply the methodology developed in Chapter 4 about the uncertainty of desirability index, more precisely for the construction of the indistinguishable optimal zone.

2.6.1 Estimate of $Cov[\hat{E}[Y|\mathbf{x}_1], \hat{E}[Y|\mathbf{x}_2]]$ for MLR

On the basis of the MLR estimated response (2.25), an estimate of the covariance between the estimated responses for two molecules is provided by

$$\hat{Cov}[\hat{E}[Y|\mathbf{x}_1], \hat{E}[Y|\hat{\mathbf{x}}_2]] = \hat{\sigma}_\epsilon^2 \mathbf{z}_1' (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{z}_2, \quad (2.32)$$

where the residual variance estimate $\hat{\sigma}_\epsilon^2$ is defined by (2.27).

2.6.2 Estimate of $Cov[\hat{E}[Y|\mathbf{x}_1], \hat{E}[Y|\mathbf{x}_2]]$ for PLSR

An estimator for the covariance is not discussed in the literature on PLSR but we may think to apply the same zeroth-order approximation as Höskuldsson (1988) used for $V[\hat{E}(Y|\mathbf{x})]$ (2.29). This leads to the following estimator:

$$\hat{Cov}[\hat{E}[Y|\mathbf{x}_1], \hat{E}[Y|\hat{\mathbf{x}}_2]] = \hat{\sigma}_\epsilon^2 \mathbf{t}_{\mathbf{M}_1}' (\mathbf{T}_{\mathbf{M}}' \mathbf{T}_{\mathbf{M}})^{-1} \mathbf{t}_{\mathbf{M}_2}, \quad (2.33)$$

where the residual variance estimate $\hat{\sigma}_\epsilon^2$ is defined by (2.30).

A bootstrap approach can also be applied but a rapid analytical approach is preferred for the virtual molecules optimisation of forthcoming Chapter 6.

2.6.3 Estimate of $Cov[\hat{E}[Y|\mathbf{x}_1], \hat{E}[Y|\mathbf{x}_2]]$ for RT

The only tool to estimate the covariance between the expected responses of two molecules is to apply bootstrap (Breiman 1996). Each bootstrapped RT is used to estimate the responses $E[Y|\mathbf{x}_1]$ and $E[Y|\mathbf{x}_2]$ for the two molecules of interest. Finally the covariance of this series of estimations pairs can be computed.

2.7 Conclusion

Eriksson *et al.* (2003) suggest to test several classes of QSAR models, and for each class of model to fit various complexity parameters (the number of entered descriptors with MLR, the number of entered latent variables with PLSR and the number of terminal nodes with RT). Whatever the fitted model is, the corresponding predictions will always be subject to three sources of error: the irreducible noise of the activity response and the bias and the variance of the fitted predictor.

The best predictive model can be defined as the one that minimises the global expected prediction error (2.6). The AIC (2.15) or the BIC (2.16) criterion are estimates of this global error and the complexity parameters can be selected in order to minimise it. Then, cross-validation or bootstrap can be applied to estimate the Mean Square Error of Prediction (2.10) in order to assess the predictive power of the QSAR model for new molecules. If several classes of models are fitted, the best model can be defined as the one that minimises the estimated Mean Square Error of Prediction. Those estimates are not popular in QSAR literature but are similar to the cross-validated or the bootstrapped Q^2 (1.19).

Finally, the selected QSAR model can be applied to estimate the activity of new compounds after checking that they are in its applicability domain. We point out the importance of quantifying the uncertainty of those estimations, which is generally neglected in QSAR applications. Bootstrapping or analytical estimators of the prediction variance for a new molecule are available with MLR, PLSR or RT models.

Bibliography

- Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984), *Classification and regression trees*, Wadsworth, Belmont, California.
- Breiman, L. and Spector, P. (1992), "Submodel selection and evaluation in regression: the X-random case", *International Statistical Review*, 60, 291-319.
- Breiman, L. (1996), "Prediction intervals in Partial Least Squares", *Machine Learning*, 26, 123-140.
- Denham, M.C. (1997), "Prediction intervals in Partial Least Squares", *Journal of Chemometrics*, 11, 39-52.
- Efron, B. (1979), "Bootstrap methods: Another look at the jackknife", *The annals of Statistics*, 7, 1-26.
- Efron, B. (1983), "Estimating the error rate of a prediction rule: improvement on cross-validation", *Journal of the American Statistical Association*, 78, 316-331.
- Efron, B. and Tibshirani, R.J. (1993), *An introduction to the bootstrap*, London: Chapman & Hall.
- Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (2003), "Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs", *Environmental Health Perspectives*, 111(10), 1361-1375.
- Faber, N.M. and Kowalski, B.R. (1997), "Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares", *Journal of Chemometrics*, 11, 181-238.
- Golbraikh, A. and Tropsha, A. (2002), "Beware of q²!", *Journal of Molecular Graphics and Modelling*, 20(4), 269-276.

- Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The elements of statistical learning. Data Mining, inference and predictions*, Springer.
- Hjorth, J.S.U. (1994), *Computer intensive statistical methods validation, model selection, and bootstrap*, London: Chapman & Hall.
- Höskuldsson, A. (1988), "PLS regression methods", *Journal of Chemometrics*, 2, 211-228.
- Lendasse, A., Simon, G., Wertz, V. and Verleysen, M. (2005), "Fast bootstrap methodology for regression model selection", *Neurocomputing*, 64, 161-181.
- Magnus, J.R. and Neudecker, H. (1988), *Matrix differential calculus with applications in statistics and econometrics*, John Wiley and sons.
- Mevik, B.-H. and Cederkvist, H.R. (2004), "Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR)", *Journal of Chemometrics*, 18(9), 422-429.
- Phatak, A., Reilly, P.M. and Penlidis, A. (1993), "An approach to interval estimation in partial least squares regression", *Analytica Chimica Acta*, 277, 495-501.
- Serneels, S., Lemberge, P. and Van Espen, P.J. (2004), "Calculation of PLS prediction intervals using efficient recursive relations for the Jacobian matrix", *Journal of Chemometrics*, 18(2), 76-80.
- Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag.
- Wehrens, R., and van der Linden, W.E. (1997), "Bootstrapping principal component regression models", *Journal of Chemometrics*, 18(2), 76-80.

Chapter 3

Multicriteria optimisation using a desirability index

Part of this chapter constitutes the discussion paper DP0531 by Govaerts and Le Bailly de Tillegheem (2005) of the Institute of Statistics, Université catholique de Louvain, Belgium.

3.1 Introduction

Improving the quality of a product is a challenge that industries face daily. A food industry may want to improve the quality of its cakes, a rubber industry the quality of a tire tread compound, a pharmaceutical industry the quality of drug candidates, *etc.*

The definition of the quality of a product is most of the time based on several criteria $\mathbf{Y} = (Y_1, Y_2, \dots, Y_p)$. If a set of factors, $\mathbf{x} = (x_1, x_2, \dots, x_K)$, can be chosen during manufacturing leading to different quality responses, modifying the x_j 's makes the Y_i 's varying. Producers want to find the factors settings that simultaneously optimise the p quality criteria. This is multicriteria optimisation (abbreviated *MCO*).

The motivating application field for this chapter is the search of the most promising molecule in a library of possible drug candidates, in the early stages of the drug discovery process, as explained in the Introduction and in Chapter 1. The factors are physico-chemical properties that describe the compounds and the responses are biological endpoints that measure their activity. Finding the best molecule is an MCO problem. This chapter is written for any MCO context.

The classical approach to MCO assumes for each response that the statistical model

$$Y_i = f_i(\mathbf{x}, \beta_i) + \epsilon_i \quad (3.1)$$

captures the relationship between the i^{th} response and the factors over the domain of interest, denoted χ . Most of the time, the function f_i is assumed to be the true link function and the error term ϵ_i is assumed to follow a $N(0, \sigma_{\epsilon_i}^2)$ distribution. The unknown model parameters β_i are estimated on data collected following an experimental design.

The most well-known class of models with MCO is the multiple linear regression (Box and Draper 1986; Khuri and Cornell 1987; Myers *et al.* 1989; Myers and Montgomery 2002). The link between the i^{th} response and the transformed factors (possibly different for each response) is assumed to follow an equation of the form

$$Y_i = \mathbf{z}'\beta_i + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, \sigma_{\epsilon_i}^2). \quad (3.2)$$

As the optimal x_j 's for the Y_i 's are usually different, the Y_i 's responses may be competitive and it is often impossible to improve one response without deteriorating another. The best compromise has to be found taking into account the relative importance of each response. A possible approach to transform the multicriteria optimisation into a single objective optimisation is to use *distance* or *loss* function measuring the departure from the individual responses optima. This methodology is not treated in this chapter. For a review of this subject, see Khuri and Conlon (1981) or Ko *et al.* (2005).

In the context of industrial statistics and experimental design, the common approach to MCO, implemented in most experimental design softwares (JMP, MODDE, Design-Expert, . . .), consists in summarising the different responses by a *desirability index*, a concept introduced by Harrington (1965) and presented in the next section. Desirability functions can be seen as a kind of truncated inverse distance or loss function as demonstrated by Jessenerger (1999), particularly in the case of a targeted property. Section 3 compares two approaches in using desirability indexes. In the classical approach of Harrington and Derringer and Suich (1980), the desirability of the expected responses is computed. On the opposite, the stochastic approach proposed by Steuer (2000) considers the desirability index as a random variable and the expected desirability is computed. The distribution of the desirability index is studied in Section 4 using the density transformation theorem (Mood *et al.* 1974).

3.2 Desirability function and desirability index

Harrington suggests to associate a value belonging to $[0, 1]$ to each combination of factors levels, $\mathbf{x}$, representing the desirability of the resulting product quality. This transforms the MCO into a unique objective optimisation, as the desirability index has then simply to be maximised.

Harrington proceeds in two steps. First, each response Y_i is transformed to the same scale using a desirability function (abbreviated *DF*), denoted by d_i , such that $d_i(Y_i) \in [0, 1]$. If $d_i(Y_i) = 0$, the product is not at all acceptable according to the specifications of the i^{th} property and if $d_i(Y_i) = 1$, the product fulfils them perfectly. Secondly, the properties transformed by desirability functions are aggregated in a single value still in the $[0, 1]$ interval, the desirability index (abbreviated *DI*), representing the overall desirability of the product. The weighted geometric mean or the weighted arithmetic mean of the DFs (Derringer 1994), as well as their minimum (Kim and Lin 2000) are the three most often used indexes, denoted by D :

$$D(\mathbf{Y}) = \prod_{i=1}^p (d_i(Y_i))^{w_i} \quad \text{or} \quad D(\mathbf{Y}) = \sum_{i=1}^p w_i \cdot d_i(Y_i) \quad \text{with} \quad w_i > 0 \quad \text{and} \quad \sum_{i=1}^p w_i = 1$$

(3.3)

$$\text{or} \quad D(\mathbf{Y}) = \min d_i(Y_i).$$

DI allows to balance the optimised properties one against the other, taking into account their target value, their relative importance and their scale. It has to be noticed that a sum of weights equalled to 1 in the arithmetic or geometric mean is a convention and is not essential. Other indexes exist to take into account the possible correlation of responses or desirabilities as those proposed by Dehhaoui (1996) and Mao (2007), but this possible dependence is usually neglected by practitioners.

3.2.1 Classical desirability functions

The most well-known DFs are the Harrington's ones (1965) based on the exponential function of a linear transformation of the response and the Derringer and Suich's ones (1980) based on a power of a linear transformation of the response. Gibb *et al.* (2001) proposed also smoother and differentiable DFs using the logit function.

These three types of DFs are presented in Table 3.1 for the cases where the response Y (the sub-index i has been removed to simplify notations) must be maximised, minimised or reach a target value. DFs depend on parameters (a , b , s or T) that have to be fixed by a specialist of the product to define which responses values are desirable. As it can be seen on Figure 3.1, the three DFs may provide similar desirability curves.

3.2.2 New Normal *cdf* based desirability functions

Govaerts and Le Bailly de Tillegem (2005) proposed a new class of DFs to transform a quality property Y in the $]0, 1[$ interval. It is based on the cumulative distribution function, abbreviated *cdf*, of the standard Normal:

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y \exp\left(-\frac{t^2}{2}\right) dt$$

	Maximum	Minimum	Target Value
Harrington (1965)	$\exp(-\exp(-a-bY))$	$1-\exp(-\exp(-a-bY))$	$\exp(- \frac{Y-T}{b} ^n)$
Derringer and Suich (1980)	$\begin{cases} 0 & \text{if } Y < a \\ (\frac{Y-a}{b-a})^s & \text{if } a \leq Y \leq b \\ 1 & \text{if } Y > b \end{cases}$	$\begin{cases} 1 & \text{if } Y < b \\ (\frac{a-Y}{a-b})^s & \text{if } b \leq Y \leq a \\ 0 & \text{if } Y > a \end{cases}$	$\begin{cases} 0 & \text{if } Y < a_1 \\ (\frac{Y-a_1}{T-a_1})^{s_1} & \text{if } a_1 \leq Y \leq T \\ (\frac{a_2-Y}{a_2-T})^{s_2} & \text{if } T \leq Y \leq a_2 \\ 0 & \text{if } Y > a_2 \end{cases}$
Gibb <i>et al</i> (2001)	$(1+\exp(-\frac{Y-a}{b}))^{-1}$	$(1-\exp(-\frac{Y-a}{b}))^{-1}$	$\exp(-\frac{1}{2}(\frac{Y-T}{b})^2)$

Table 3.1: Examples of desirability functions. Y is a response; the target value T and the parameters a , b and s have to be adjusted according to the specifications.

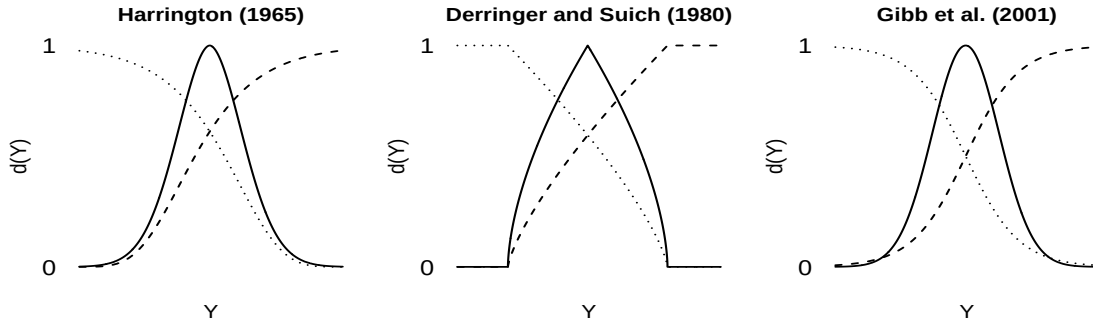


Figure 3.1: Different desirability functions. The continuous lines represent cases of a targeted property and the dotted and dashed lines represent cases of a minimised and a maximised property respectively.

Other *cdf* could be used, but the Normal one is the most wellknown. This function is continuous and differentiable, which will be useful to apply the density transformation theorem further. According to the desirable values of the property Y , the following transformations can be used as depicted on Figure 3.2:

$$d(Y) = \Phi\left(\frac{Y-a}{b}\right) \text{ if } Y \text{ has to be maximised,} \quad (3.4)$$

$$d(Y) = 1 - \Phi\left(\frac{Y-a}{b}\right) \text{ if } Y \text{ has to be minimised,} \quad (3.5)$$

$$d(Y) = \sqrt{\Phi\left(\frac{Y-a_1}{b_1}\right) \cdot \left[1 - \Phi\left(\frac{Y-a_2}{b_2}\right)\right]} \text{ if } Y \text{ has to reach some optimal value.} \quad (3.6)$$

The parameters a and b in equations (3.4-3.5-3.6) have to be chosen according to the knowledge of both the statistician and the specialist of the product. They can be fixed using one of the two simple following rules :

- If the specialist notifies that under the value LL (Lower Limit) and over the value UL

(Upper Limit) changes for the property Y have no additional interest, the parameters a and b can be chosen as a function of those limits according to the properties of the standard Normal distribution:

- If Y has to be maximised, take $a = \frac{LL+UL}{2}$ and $b = \frac{UL - \frac{LL+UL}{2}}{2}$ such that $d(LL) \approx 0.023$ and $d(UL) \approx 0.977$
 - If Y has to be minimised, take $a = \frac{LL+UL}{2}$ and $b = \frac{UL - \frac{LL+UL}{2}}{2}$ such that $d(LL) \approx 0.977$ and $d(UL) \approx 0.023$
 - If Y has to reach some optimal value T , take $a_1 = \frac{LL+T}{2}$, $b_1 = \frac{T - \frac{LL+T}{2}}{2}$, $a_2 = \frac{T+UL}{2}$ and $b_2 = \frac{UL - \frac{T+UL}{2}}{2}$ such that $d(LL) = d(UL) \approx 0.151$ and $d(T) = 0.977$
- If the specialist does not notify any limit for the property Y and just desires to maximise or minimise it, or to reach a target value T , the two parameters a and b can be chosen according to the values of the response Y observed on a small sample of the experimental domain:
- If Y has to be maximised or minimised, take a and b respectively as the sample average and the sample standard deviation.
 - If Y has to reach some optimal value T , divide the sample into a first subset with smaller values than T and a second subset with higher value than T . Take a_1 and b_1 , as the arithmetic mean and the standard deviation of the first subset and a_2 and b_2 , as the arithmetic mean and the standard deviation of the second subset respectively.

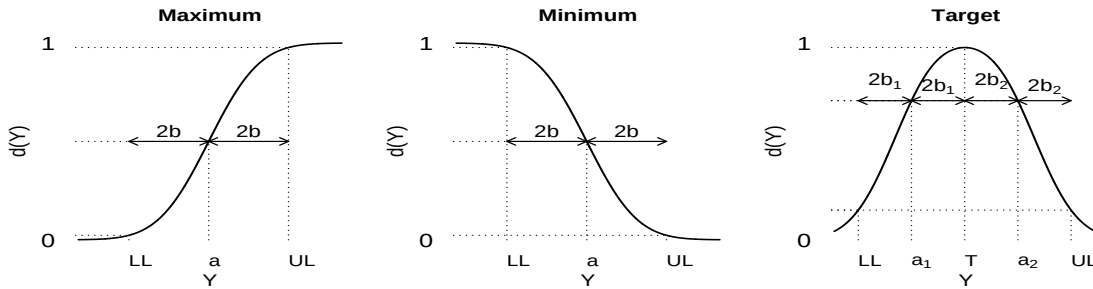


Figure 3.2: Desirability functions based on the standard Normal *cdf*. The three graphs correspond to the case of maximised property (left), minimised property (center) and targeted property (right). The dashed lines correspond to the limits (LL and USL). Decreasing or increasing parameter a shifts the desirability curves respectively to the left or to the right and decreasing or increasing parameter b makes the curves respectively more or less stiff.

3.2.3 Comparison of desirability functions

All proposed DFs are aimed at transforming each property to optimise on a utility scale: higher is the desirability, better is the product. Harrington is the first author to define a DF. He proposed to use exponential functions, arguing their interest in an optimisation context. Then, Derringer and Suich generalised the Harrington DFs to obtain more flexible curves but at the cost of a heavier parametrization. Finally, Gibb *et al.* presented new DFs as differentiable versions of the Derringer and Suich DFs. The Normal *cdf* based DFs proposed in the thesis are close to the Gibb *et al.* ones, but all DFs can have very similar shapes.

Those 4 types of DFs are compared in Table 3.2 analysing the resulting desirability scale, the number of parameters to be fixed by the user, the symmetry of the curve and the differentiability. One can point out some practical drawbacks when using those DFs in an optimisation algorithm:

- *ex aequos* are possible with the Derringer and Suich DFs,
- the Govaerts and Le Bailly DFs in the maximisation/minimisation case are necessary symmetric around the a parameter and the Harrington and the Gibb *et al.* DFs are necessary symmetric around the target value.
- an expert has to be consulted to fix the parameters of the Harrington, the Derringer and Suich and the Gibb *et al.* DFs.
- the Derringer and Suich DFs and the Harrington DF in the target case are not two times continuously differentiable.

		Harrington	Derringer and Suich	Gibb <i>et al.</i>	Govaerts and Le Bailly
Scale	(max-min)	]0,1[	[0,1]	]0,1[	]0,1[
	(target)	]0,1]	[0,1]	]0,1]	]0,1]
Parameters	(max-min)	2	3	2	2
	(target)	3	5	2	4
Symmetric	(max-min)	No	Can be	No	Yes
	(target)	Yes	Can be	Yes	Can be
2 times differentiable	(max-min)	Yes	No	Yes	Yes
	(target)	No	No	Yes	Yes

Table 3.2: Comparison of the desirability functions proposed by Harrington (1965); Derringer and Suich (1980); Gibb *et al.* (2001); Govaerts and Le Bailly de Tillegem (2005a).

3.3 Stochastic approach to multicriteria optimisation

If the statistical model of equation (3.1) is assumed as the true link between each optimised response and the factors, there are two possibilities to associate a desirability value to any factors settings: applying DFs on expected responses or taking the expectation of the DI. The first case is the classical use of desirability, $D^C(x)$, as Harrington (1965) and Derringer and Suich (1980) proposed in their papers and the second case is the stochastic approach introduced by Steuer (2000), $D^S(x)$. Maximising the expected desirability of the responses is intuitively and mathematically more correct than looking for the factors settings for which the desirability of expected responses is maximum.

If, for instance, the weighted geometric mean (3.3) is used as desirability index, the optimisation problem can be formalised as below:

$$\begin{aligned} \max_{\mathbf{x} \in \mathcal{X}} D^C(\mathbf{x}) \quad & \text{with} \\ D^C(\mathbf{x}) = D(E[\mathbf{Y}|\mathbf{x}]) &= \prod_{i=1}^p [d_i(E[Y_i|\mathbf{x}])]^{w_i} = \prod_{i=1}^p [d_i(f_i(\mathbf{x}, \boldsymbol{\beta}_i))]^{w_i} \quad \text{or} \end{aligned} \quad (3.7)$$

$$\begin{aligned} \max_{\mathbf{x} \in \mathcal{X}} D^S(\mathbf{x}) \quad & \text{with} \\ D^S(\mathbf{x}) = E[D(\mathbf{Y})|\mathbf{x}] &= E \left[\prod_{i=1}^p [d_i(Y_i)]^{w_i} | \mathbf{x} \right] = E \left[\prod_{i=1}^p [d_i(f_i(\mathbf{x}, \boldsymbol{\beta}_i) + \epsilon_i)]^{w_i} \right] \end{aligned} \quad (3.8)$$

As the DI is not a linear transformation of the responses, most of the time, $D(E[\mathbf{Y}|\mathbf{x}]) \neq E[D(\mathbf{Y})|\mathbf{x}]$ and the corresponding optima, $\mathbf{x}_{opt}^C$ and $\mathbf{x}_{opt}^S$, are different.

The idea of maximising the expected DI instead of the DI of expected responses is the same as in the utility theory field (von Neumann and Morgenstern 1944). The expected utility function, u , is maximised instead of the utility function of expected results, R , to take risk factors into account: $\max E[u(R)]$ instead of $\max u(E[R])$. Similarly, as it takes into account the propagation of the uncertainty of the response ϵ_i on the DI, Steuer proposed to associate to each factors setting $\mathbf{x}$ the average desirability of the resulting product quality, $E[D(\mathbf{Y})|\mathbf{x}]$.

The computation of $E[D(\mathbf{Y})|\mathbf{x}]$ requires the knowledge of the distribution of the DI considered as a random variable for given values of the factors settings, $D(\mathbf{Y})|\mathbf{x}$. This is the subject of the next section.

3.4 Distribution of the desirability index

For given factors settings $\mathbf{x}$, if the error term in equation (3.1) is assumed to be normally distributed $\epsilon_i \sim N(0, \sigma_{\epsilon_i}^2)$, and if the link function f_i and the model parameters β_i are known as well as the error term variance $\sigma_{\epsilon_i}^2$, the distribution of $D(\mathbf{Y})|\mathbf{x}$ can be known, exactly or approximately.

Steuer (2000) proposed to approximate the distribution of $D(\mathbf{Y})|\mathbf{x}$ by Monte-Carlo simulations on the basis of the normal distribution for the models errors. This is a heavy procedure to explore the experimental domain χ , especially if there are more than two factors.

To avoid intensive use of Monte-Carlo simulations, Weber and Weihs (2003) and Trautmann and Weihs (2006), suggested to derive analytically the distribution of $D(\mathbf{Y})|\mathbf{x}$ on the basis of the model error distribution using the density transformation theorem. If the probability density function (abbreviated *pdf*) of $D(\mathbf{Y})|\mathbf{x}$ is known, its expectation $E[D(\mathbf{Y})|\mathbf{x}]$ can be computed by analytical or numerical integration. They derived the analytical expressions of the Harrington DFs distributions and deduced, for special cases, the DI distribution.

In this section, the probability density function (abbreviated *pdf*) of DFs based on the Normal *cdf* are derived, as well as the DI based on the geometric mean, the arithmetic mean or the minimum of such DFs. For simplicity, the i index standing for the i^{th} optimised criteria Y_i is removed.

3.4.1 Pdf of the desirability function $d(Y)|\mathbf{x}$

Most of the following results are based on the univariate density transformation theorem.

Theorem 1 (Density transformation). *Let U have a pdf $f_U(u)$. If $h(u)$ is either increasing or decreasing for all u such that $f_U(u) > 0$, then $V = h(U)$ has a pdf given by*

$$f_V(v) = f_U(h^{-1}(v)) \cdot \left| \frac{d(h^{-1}(v))}{dv} \right|. \quad (3.9)$$

Given the assumption that the error term ϵ of the model (3.1) follows a $N(0, \sigma_\epsilon^2)$ distribution, we know that $Y|\mathbf{x} \sim N(E[Y|\mathbf{x}], \sigma_\epsilon^2)$ and $\frac{Y-a}{b}|\mathbf{x} \sim N\left(\frac{E[Y|\mathbf{x}]-a}{b}, \frac{\sigma_\epsilon^2}{b^2}\right)$. As the DFs based on the Normal *cdf* for the maximisation and the minimisation cases (see equations (3.4-3.5)) are respectively increasing and decreasing functions of $\frac{Y-a}{b}|\mathbf{x}$, we can derive their *pdf*'s using the density transformation theorem.

In the following $E[Y|\mathbf{x}]$ is abbreviated μ_Y and $\phi_{M;S^2}$ and $\Phi_{M;S^2}$ denote respectively the *pdf* and the *cdf* of the Normal with expectation M and variance S^2 and ϕ and Φ without any index denote respectively the *pdf* and the *cdf* of the standard Normal.

Proposition 1 (DF - maximisation). *The pdf of $U \equiv d(Y)|\mathbf{x} = \Phi\left(\frac{Y-a}{b}\right)|\mathbf{x}$ is given by*

$$f_U(u) = \begin{cases} \frac{\phi_{\frac{\mu_Y-a}{b}; \frac{\sigma_\epsilon^2}{b^2}}(\Phi^{-1}(u))}{\phi(\Phi^{-1}(u))} & \text{if } 0 < u < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (3.10)$$

Proposition 2 (DF - minimisation). *The pdf of $U \equiv d(Y)|\mathbf{x} = 1 - \Phi\left(\frac{Y-a}{b}\right)|\mathbf{x}$ is given by*

$$f_U(u) = \begin{cases} \frac{\phi_{\frac{\mu_Y-a}{b}; \frac{\sigma_\epsilon^2}{b^2}}(\Phi^{-1}(1-u))}{\phi(\Phi^{-1}(1-u))} & \text{if } 0 < u < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (3.11)$$

For both maximisation and minimisation cases, the DFs *pdf*'s are the ratio of two Normal *pdf*'s and are then easy to compute in practice with any statistical software. The shapes of those two densities are similar and vary in the same way with parameters μ_Y , a , b and σ_ϵ^2 as presented in Figure 3.3.

The *pdf* of the DF in the case of a targeted property can not be derived directly using the density transformation theorem (Theorem 1) as $U \equiv d(Y)|\mathbf{x} = \sqrt{\Phi\left(\frac{Y-a_1}{b_1}\right) \cdot [1 - \Phi\left(\frac{Y-a_2}{b_2}\right)]}|\mathbf{x}$ does not satisfy the monotonicity assumption. Nevertheless, $d(Y)$ is strictly increasing in Y below the value target T and strictly decreasing above the value target T . Thanks to this particular shape, as represented in Figure 3.4, realisations of U smaller than a given u correspond either to realisations of $Y|\mathbf{x}$ smaller than a certain $y_1 \in (-\infty, T]$ such that $d(y_1) = u$, or realisations of $Y|\mathbf{x}$ higher than a certain $y_2 \in [T, \infty)$ such that $d(y_2) = u$. This allows to compute first the *cdf* of $d(Y)|\mathbf{x}$ and then derive this analytical expression to obtain its *pdf*.

Proposition 3 (DF - target). *The cdf and the pdf of $U \equiv d(Y)|\mathbf{x} = \sqrt{\Phi\left(\frac{Y-a_1}{b_1}\right) \cdot [1 - \Phi\left(\frac{Y-a_2}{b_2}\right)]}|\mathbf{x}$ are respectively given by*

$$F_U(u) = \begin{cases} 0 & \text{if } u \leq 0, \\ \Phi_{\mu_Y; \sigma_\epsilon^2}(y_1) + 1 - \Phi_{\mu_Y; \sigma_\epsilon^2}(y_2) & \text{if } 0 < u < d(T) = 0.9772, \\ 1 & \text{if } u \geq d(T) = 0.9772, \end{cases} \quad (3.12)$$

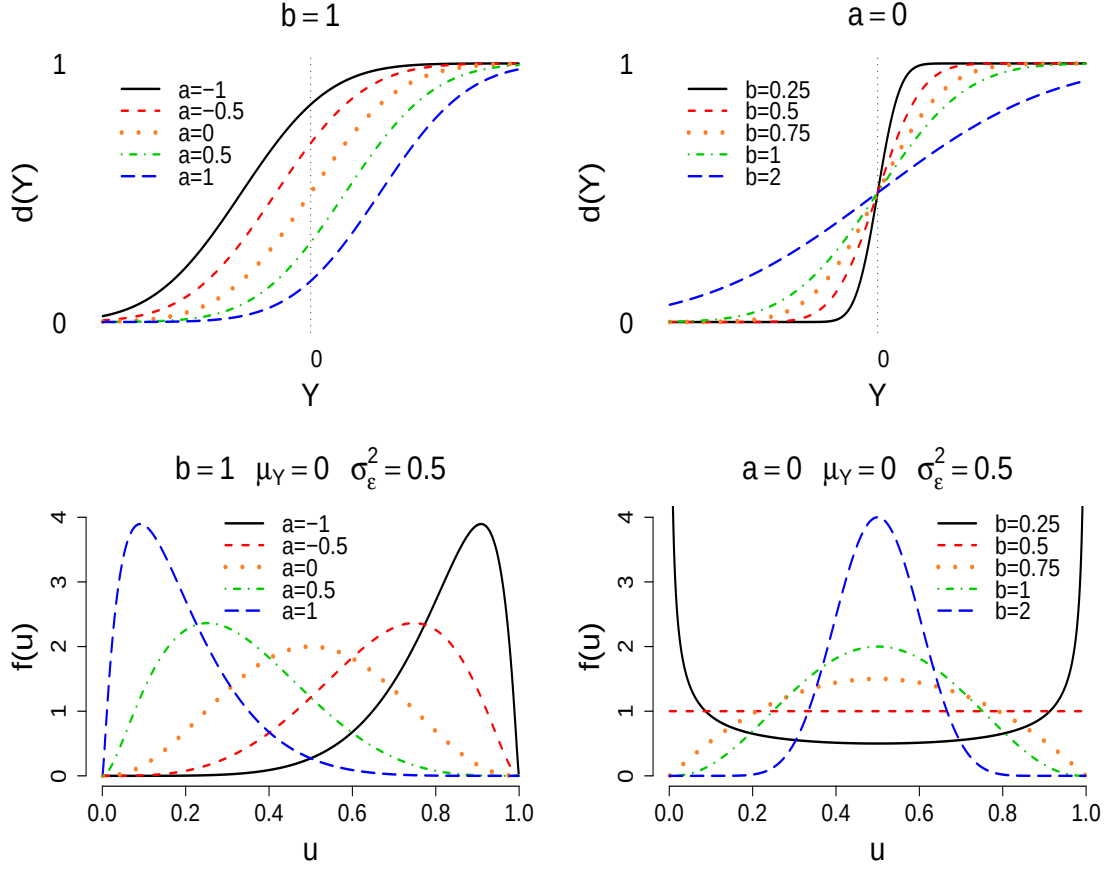
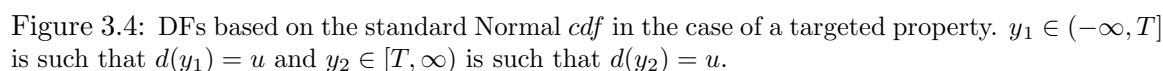


Figure 3.3: First line: different shapes of the Normal *cdf* based DF in the case of maximised property (equation (3.4)). The left graph shows how the shape varies with a ($b = 1$) and the right graph shows how the shape varies with b ($a = 0$). Second line: corresponding *pdf* with $\mu_Y = 0$ and $\sigma_\epsilon^2 = 0.5$ (equation (3.10)). The left graph shows how the shape varies with a (inverse behaviour with μ_Y) and the right graph shows how the shape varies with b (inverse behaviour with σ_ϵ^2).

and

$$f_U(u) = \begin{cases} \frac{1}{2}\phi_{\mu_Y; \sigma_\epsilon^2}(y_1) \left[\frac{\phi(\frac{y_1-a_1}{b_1})\frac{1}{b_1}}{\sqrt{\Phi(\frac{y_1-a_1}{b_1})}} \sqrt{1 - \Phi(\frac{y_1-a_2}{b_2})} - \frac{\phi(\frac{y_1-a_2}{b_2})\frac{1}{b_2}}{\sqrt{1 - \Phi(\frac{y_1-a_2}{b_2})}} \sqrt{\Phi(\frac{y_1-a_1}{b_1})} \right]^{-1} \\ - \frac{1}{2}\phi_{\mu_Y; \sigma_\epsilon^2}(y_2) \left[\frac{\phi(\frac{y_2-a_1}{b_1})\frac{1}{b_1}}{\sqrt{\Phi(\frac{y_2-a_1}{b_1})}} \sqrt{1 - \Phi(\frac{y_2-a_2}{b_2})} - \frac{\phi(\frac{y_2-a_2}{b_2})\frac{1}{b_2}}{\sqrt{1 - \Phi(\frac{y_2-a_2}{b_2})}} \sqrt{\Phi(\frac{y_2-a_1}{b_1})} \right]^{-1} \\ 0 \end{cases} \quad \text{if } 0 < u < d(T) = 0.9772, \\ \text{otherwise,} \end{cases} \quad (3.13)$$


$$\sqrt{\Phi\left(\frac{y_1 - a_1}{b_1}\right) \cdot [1 - \Phi\left(\frac{y_1 - a_2}{b_2}\right)]} = u \quad \text{and} \quad \sqrt{\Phi\left(\frac{y_2 - a_1}{b_1}\right) \cdot [1 - \Phi\left(\frac{y_2 - a_2}{b_2}\right)]} = u.$$

The shapes of this density vary with parameters μ_Y , a_1 and a_2 , b_1 and b_2 , and σ_ϵ as presented in Figure 3.5.

Propositions 1, 2 and 3 provide the *pdf* of DFs for the three optimisation cases, $d(Y)|\mathbf{x} = \Phi(\frac{Y-a}{b})|\mathbf{x}$, $d(Y)|\mathbf{x} = 1 - \Phi(\frac{Y-a}{b})|\mathbf{x}$ and $d(Y)|\mathbf{x} = \sqrt{\Phi(\frac{Y-a_1}{b_1}) \cdot [1 - \Phi(\frac{Y-a_2}{b_2})]}|\mathbf{x}$. The DF based on the geometric mean (equation (3.3)) makes use of such functions weighted by a certain exponent $w \in (0, 1)$. The following question arises then: what is the *pdf* of a weighted DF $(d(Y))^w|\mathbf{x}$? The density transformation theorem (Theorem 1) provides straightforward results as $(d(Y))^w|\mathbf{x}$ is an increasing function of $d(Y)|\mathbf{x}$:

Proposition 4 (weighted DF - maximisation). *The pdf of $V \equiv (d(Y))^w | \mathbf{x} = [\Phi(\frac{Y-a}{b})]^w | \mathbf{x}$*

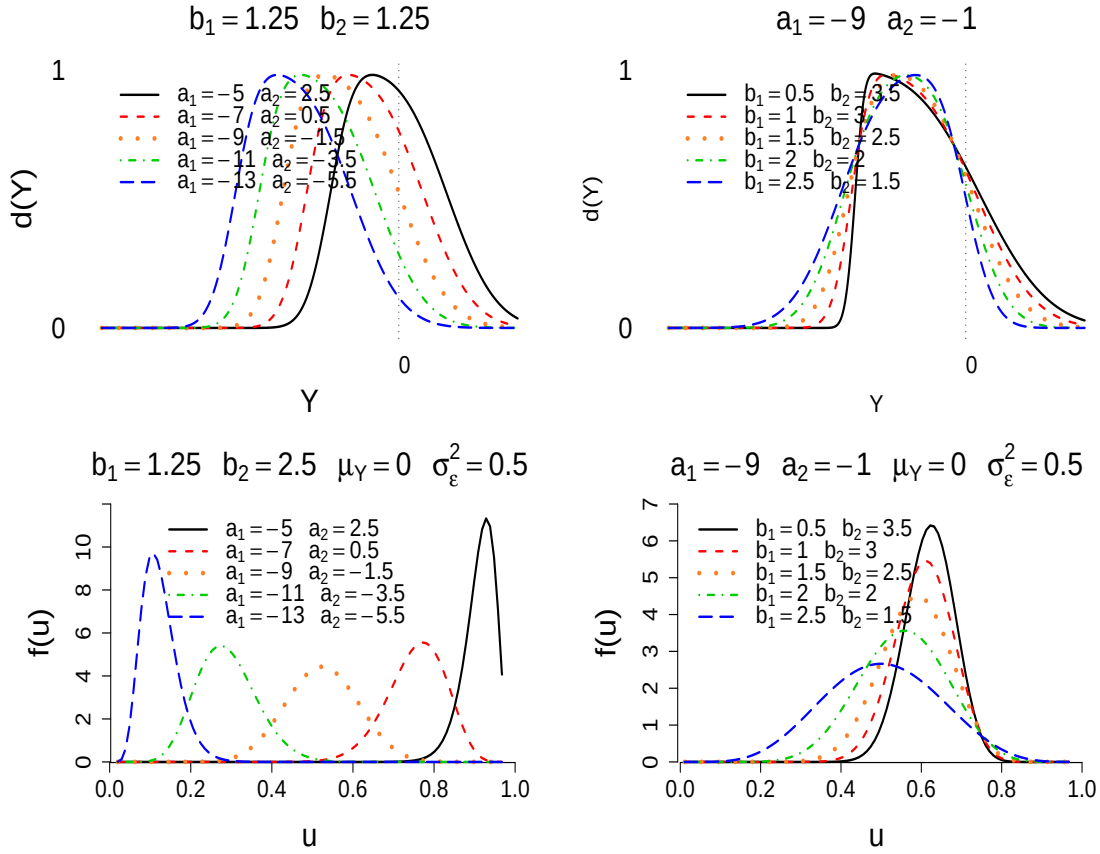


Figure 3.5: First line: different shapes of the Normal *cdf* based DF in the case of targeted property (equation (3.6)). The left graph shows how the shape varies with a_1 and a_2 ($b_1 = 1.25$ and $b_2 = 2.5$) and the right graph shows how the shape varies with b_1 and b_2 ($a_1 = -9$ and $a_2 = -1$). Second line: corresponding *pdf* with $\mu_Y = 0$ and $\sigma_\epsilon^2 = 0.5$ (equation (3.13)). The left graph shows how the shape varies with a_1 and a_2 (inverse behaviour with μ_Y) and the right graph shows how the shape varies with b_1 and b_2 (inverse behaviour with σ_ϵ^2).

is given by

$$f_V(v) = \begin{cases} \frac{\phi_{\frac{\mu_Y - a}{b}; \frac{\sigma_\epsilon^2}{b^2}}(\Phi^{-1}(v^{1/w}))}{\phi(\Phi^{-1}(v^{1/w}))} \cdot \frac{1}{w} v^{\frac{1}{w}-1} & \text{if } 0 < v < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (3.14)$$

Proposition 5 (weighted DF - minimisation). *The pdf of $V \equiv (d(Y))^w | \mathbf{x} = [1 - \Phi(\frac{Y-a}{b})]^w | \mathbf{x}$ is given by*

$$f_V(v) = \begin{cases} \frac{\phi_{\frac{\mu_Y - a}{b}; \frac{\sigma_\epsilon^2}{b^2}}(\Phi^{-1}(1-v^{1/w}))}{\phi(\Phi^{-1}(1-v^{1/w}))} \cdot \frac{1}{w} v^{\frac{1}{w}-1} & \text{if } 0 < v < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (3.15)$$

For both maximisation and minimisation cases, the *pdf*'s are the ratio of two Normal *pdf*'s times a simple factor depending on the realisation v and the weight w . Those two *pdf*'s are still easy to compute with any statistical software. The shapes of the densities for both minimisation and maximisation are similar. The influence of the weight w is presented in Figure 3.6 for the maximisation case.

The *cdf* and the *pdf* of the weighted DF in the case of a targeted property can be obtained in the same way as proposition 3, using the particular increasing-decreasing shape of $(d(Y))^w | \mathbf{x}$ like $d(Y) | \mathbf{x}$ (Figure 3.4).

Proposition 6 (weighted DF - target). *The cdf and the pdf of*

$$V \equiv (d(Y))^w | \mathbf{x} = \left(\sqrt{[\Phi(\frac{Y-a_1}{b_1})] \cdot [1 - \Phi(\frac{Y-a_2}{b_2})]} \right)^w | \mathbf{x} \text{ are respectively given by}$$

$$F_V(v) = \begin{cases} 0 & \text{if } v \leq 0, \\ \Phi_{\mu_Y; \sigma_\epsilon^2}(y_1) + 1 - \Phi_{\mu_Y; \sigma_\epsilon^2}(y_2) & \text{if } 0 < v < d(T)^w = 0.9772^w, \\ 1 & \text{if } v \geq d(T) = 0.9772^w, \end{cases} \quad (3.16)$$

and

$$\begin{aligned} f_V(v) &= \phi_{\mu_Y; \sigma_\epsilon^2}(y_1) \left[\frac{w}{2} \frac{\phi(\frac{y_1 - a_1}{b_1}) \frac{1}{b_1}}{[\Phi(\frac{y_1 - a_1}{b_1})]^{1-\frac{w}{2}}} [1 - \Phi(\frac{y_1 - a_2}{b_2})]^{\frac{w}{2}} - \frac{w}{2} \frac{\phi(\frac{y_1 - a_2}{b_2}) \frac{1}{b_2}}{[1 - \Phi(\frac{y_1 - a_2}{b_2})]^{1-\frac{w}{2}}} [\Phi(\frac{y_1 - a_1}{b_1})]^{\frac{w}{2}} \right]^{-1} \\ &\quad - \phi_{\mu_Y; \sigma_\epsilon^2}(y_2) \left[\frac{w}{2} \frac{\phi(\frac{y_2 - a_1}{b_1}) \frac{1}{b_1}}{[\Phi(\frac{y_2 - a_1}{b_1})]^{1-\frac{w}{2}}} [1 - \Phi(\frac{y_2 - a_2}{b_2})]^{\frac{w}{2}} - \frac{w}{2} \frac{\phi(\frac{y_2 - a_2}{b_2}) \frac{1}{b_2}}{[1 - \Phi(\frac{y_2 - a_2}{b_2})]^{1-\frac{w}{2}}} [\Phi(\frac{y_2 - a_1}{b_1})]^{\frac{w}{2}} \right]^{-1} \\ &\quad \text{if } 0 < v < d(T)^w = 0.9772^w, \quad 0 \quad \text{otherwise,} \end{aligned} \quad (3.17)$$

where $y_1 \in (-\infty, T]$ and $y_2 \in [T, \infty)$ depend on v by the equations

$$\left(\sqrt{[\Phi(\frac{y_1 - a_1}{b_1})] \cdot [1 - \Phi(\frac{y_1 - a_2}{b_2})]} \right)^w = v \quad \text{and} \quad \left(\sqrt{[\Phi(\frac{y_2 - a_1}{b_1})] \cdot [1 - \Phi(\frac{y_2 - a_2}{b_2})]} \right)^w = v.$$

The shape of this density varies with parameters μ_Y , a_1 and a_2 , b_1 and b_2 , and σ_ϵ as presented in Figure 3.7.

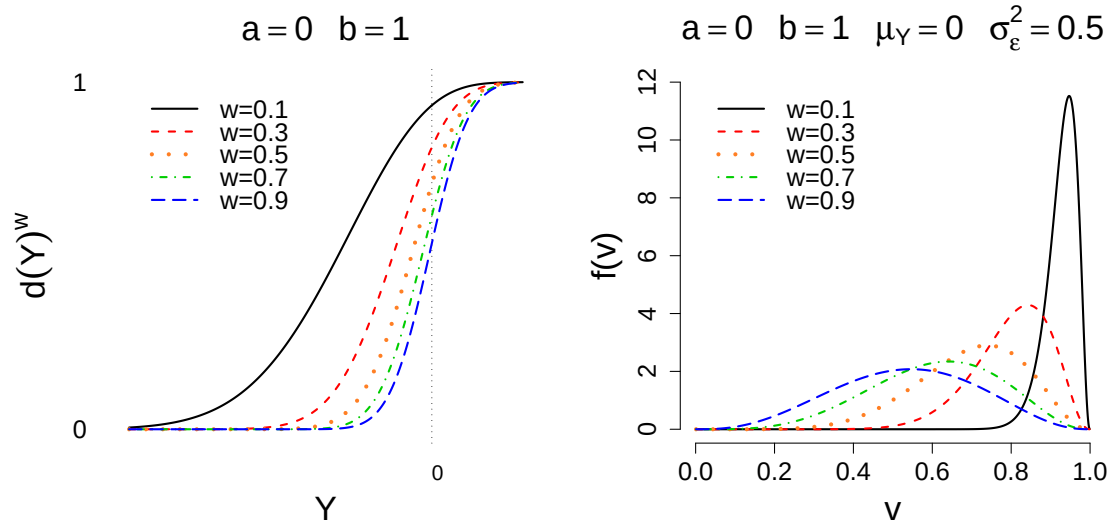


Figure 3.6: The left graph shows how the shape of the weighted Normal *cdf* based DF in the case of maximised property (equation (3.4)) varies with w ($a = 0$ and $b = 1$). The right graph shows how the shape of the corresponding densities varies with w .

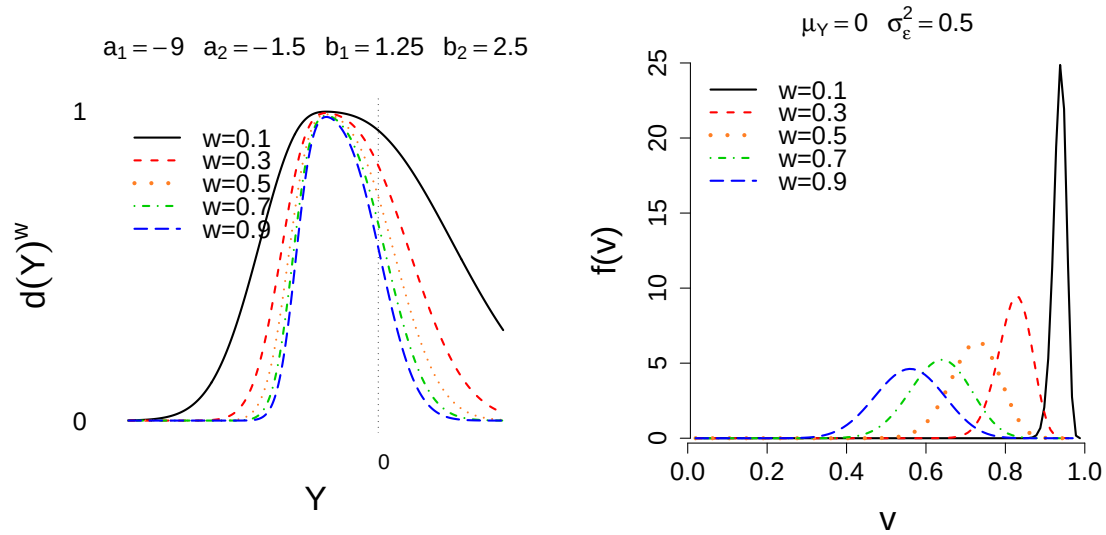


Figure 3.7: The left graph shows how the shape of the weighted Normal *cdf* based DF in the case of targeted property (equation (3.4)) varies with w ($a_1 = -9$ and $a_2 = -1.5$, $b_1 = 1.25$ and $b_2 = 2.5$). The right graph shows how the shape of the corresponding densities varies with w .

3.4.3 Pdf of the weighted desirability function $w \cdot d(Y)|\mathbf{x}$

With the DI based on the arithmetic mean (equation (3.3)), DFs can be multiplied by a constant $w \in (0, 1)$ to take into account the relative importance of each response. The pdf of $w \cdot d(Y)|\mathbf{x}$ can be easily derived by applying once again the density transformation theorem for the maximisation and the minimisation cases where $d(Y)|\mathbf{x} = \Phi\left(\frac{Y-a}{b}\right)|\mathbf{x}$ and $d(Y)|\mathbf{x} = 1 - \Phi\left(\frac{Y-a}{b}\right)|\mathbf{x}$. The variation of the shape of the weighted DF with w can be viewed on Figure 3.8 for the maximisation case.

For the targeted case, the double monotonic shape of the Normal cdf based DF allows to derive the pdf of $w \cdot d(Y)|\mathbf{x}$. The variation of the shape of the weighted DF with w can be viewed on Figure 3.9 for this targeted case.

Proposition 7 (weighted DF - maximisation). *The pdf of $V \equiv w \cdot d(Y)|\mathbf{x} = w \cdot \Phi\left(\frac{Y-a}{b}\right)|\mathbf{x}$ is given by*

$$f_V(v) = \begin{cases} \frac{\phi_{\frac{\mu_Y-a}{b}; \frac{\sigma_\epsilon^2}{b^2}}(\Phi^{-1}(\frac{v}{w}))}{\phi(\Phi^{-1}(\frac{v}{w}))} \cdot \frac{1}{w} & \text{if } 0 < v < w, \\ 0 & \text{otherwise.} \end{cases} \quad (3.18)$$

Proposition 8 (weighted DF - minimisation). *The pdf of $V \equiv w \cdot d(Y)|\mathbf{x} = w \cdot [1 - \Phi\left(\frac{Y-a}{b}\right)]|\mathbf{x}$ is given by*

$$f_V(v) = \begin{cases} \frac{\phi_{\frac{\mu_Y-a}{b}; \frac{\sigma_\epsilon^2}{b^2}}(\Phi^{-1}(1-\frac{v}{w}))}{\phi(\Phi^{-1}(1-\frac{v}{w}))} \cdot \frac{1}{w} & \text{if } 0 < v < w, \\ 0 & \text{otherwise.} \end{cases} \quad (3.19)$$

Proposition 9 (weighted DF - target). *The cdf and the pdf of*

$V \equiv w \cdot d(Y)|\mathbf{x} = w \cdot [\Phi\left(\frac{Y-a_1}{b_1}\right)]^{0.5} [1 - \Phi\left(\frac{Y-a_2}{b_2}\right)]^{0.5}|\mathbf{x}$ *are respectively given by*

$$F_V(v) = \begin{cases} 0 & \text{if } v \leq 0 \\ \Phi_{\mu_Y; \sigma_\epsilon^2}(y_1) + 1 - \Phi_{\mu_Y; \sigma_\epsilon^2}(y_2) & \text{if } 0 < v < w \cdot d(T) = w \cdot 0.9772, \\ 1 & \text{if } v \geq w \cdot d(T) = w \cdot 0.9772, \end{cases} \quad (3.20)$$

and

$$f_V(v) = \phi_{\mu_Y; \sigma_\epsilon^2}(y_1) \left[2w \left(\frac{\phi(\frac{y_1-a_1}{b_1}) \frac{1}{b_1}}{\sqrt{\Phi(\frac{y_1-a_1}{b_1})}} \sqrt{1-\Phi\left(\frac{y_1-a_2}{b_2}\right)} - \frac{\phi(\frac{y_1-a_2}{b_2}) \frac{1}{b_2}}{\sqrt{1-\Phi(\frac{y_1-a_2}{b_2})}} \sqrt{\Phi\left(\frac{y_1-a_1}{b_1}\right)} \right) \right]^{-1} \\ - \phi_{\mu_Y; \sigma_\epsilon^2}(y_2) \left[2w \left(\frac{\phi(\frac{y_2-a_1}{b_1}) \frac{1}{b_1}}{\sqrt{\Phi(\frac{y_2-a_1}{b_1})}} \sqrt{1-\Phi\left(\frac{y_2-a_2}{b_2}\right)} - \frac{\phi(\frac{y_2-a_2}{b_2}) \frac{1}{b_2}}{\sqrt{1-\Phi(\frac{y_2-a_2}{b_2})}} \sqrt{\Phi\left(\frac{y_2-a_1}{b_1}\right)} \right) \right]^{-1} \\ \text{if } 0 < v < w \cdot d(T) = 0.9772w, \quad 0 \quad \text{otherwise,} \quad (3.21)$$

where $y_1 \in (-\infty, T]$ and $y_2 \in [T, \infty)$ depend on v by the equations

$$w \cdot \sqrt{\Phi\left(\frac{y_1 - a_1}{b_1}\right) \cdot \left[1 - \Phi\left(\frac{y_1 - a_2}{b_2}\right)\right]} = v \quad \text{and} \quad w \cdot \sqrt{\Phi\left(\frac{y_2 - a_1}{b_1}\right) \cdot \left[1 - \Phi\left(\frac{y_2 - a_2}{b_2}\right)\right]} = v.$$

Those expressions are similar to the *pdf* of the unweighted cases (Propositions 1-2-3) but evaluated in $\frac{v}{w}$ and multiplied by $\frac{1}{w}$.

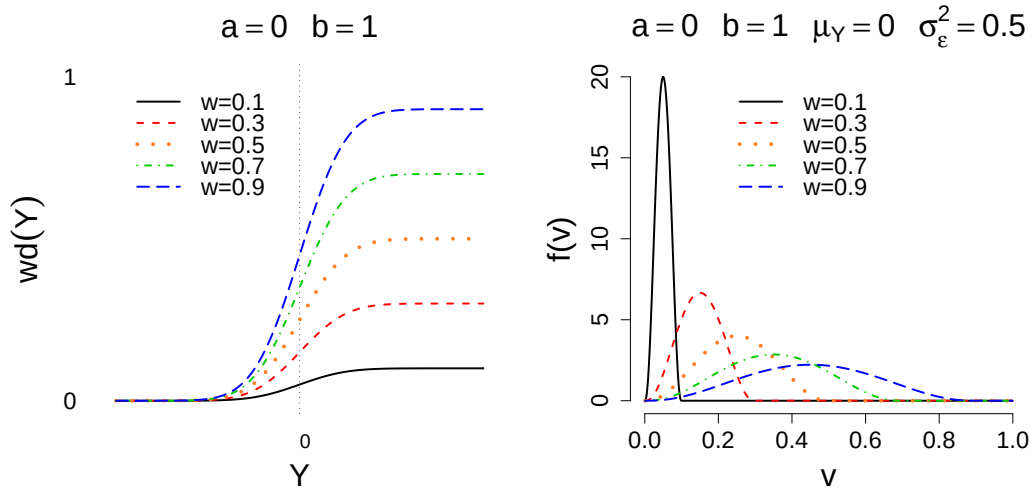


Figure 3.8: The left graph shows how the shape of the weighted Normal *cdf* based DF in the case of maximised property (equation (3.4)) varies with w ($a = 0$ and $b = 1$). The right graph shows how the shape of the corresponding densities varies with w .

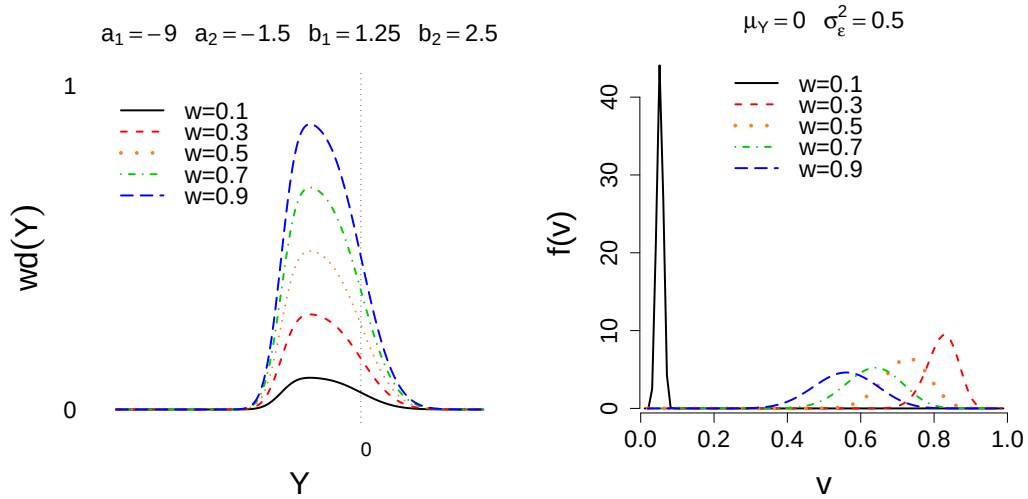


Figure 3.9: The left graph shows how the shape of the weighted Normal *cdf* based DF in the case of targeted property (equation (3.4)) varies with w ($a_1 = -9$ and $a_2 = -1.5$, $b_1 = 1.25$ and $b_2 = 2.5$). The right graph shows how the shape of the corresponding densities varies with w .

3.4.4 Distribution of the weighted geometric mean of DFs

The weighted geometric mean is defined as the product of the weighted individual DFs: $D(\mathbf{Y})|\mathbf{x} = \prod_{i=1}^p (d_i(Y_i))^{w_i}|\mathbf{x}$ with $w_i \in (0, 1)$ and $\sum_{i=1}^p w_i = 1$. This DI is a random variable, abbreviated D , constructed by the product of p continuous random variables $V_i \equiv d_i(Y_i)^{w_i}|\mathbf{x}$: $D = \prod_{i=1}^p V_i$. The *pdf* of a product of such p random variables can be generalised from the following theorem that provides the *pdf* of the product of two random variables. A demonstration can be found in (Mood *et al.* 1974).

Theorem 2 (Density of the product of two random variables). *Let (V_1, V_2) be a vector of two continuous random variables with known joint distribution $f_{V_1, V_2}(\cdot, \cdot)$ then $W \equiv V_1 \cdot V_2$ is a continuous random variable with *pdf* given by.*

$$f_W(w) = \int_{-\infty}^{\infty} f_{V_1, V_2}(t, \frac{w}{t}) \frac{1}{|t|} dt = \int_{-\infty}^{\infty} f_{V_1, V_2}(\frac{w}{t}, t) \frac{1}{|t|} dt \quad (3.22)$$

The density of the product $V_1 \cdot V_2$ depends on the joint density of the vector (V_1, V_2) . If the realistic assumption is made that, conditionally to $\mathbf{x}$, the two responses are independent, *i.e.* $(Y_1|\mathbf{x}) \perp (Y_2|\mathbf{x})$, then $V_1 \equiv d_1(Y_1)^{w_1}|\mathbf{x} \perp V_2 \equiv d_2(Y_2)^{w_2}|\mathbf{x}$ and the joint density of (V_1, V_2) is simply the product of the two known marginal densities: $f_{V_1, V_2}(v_1, v_2) = f_{V_1}(v_1) \cdot f_{V_2}(v_2)$. Generalising that, by recurrence, to the product of p independent random variables leads to the new proposition:

Proposition 10 (*Pdf of the weighted geometric mean*). *Under the assumption that $Y_1|\mathbf{x}, Y_2|\mathbf{x}, \dots, Y_p|\mathbf{x}$ are mutually independent, the *pdf* of $D \equiv D(\mathbf{Y})|\mathbf{x} = \prod_{i=1}^p (d_i(Y_i))^{w_i}|\mathbf{x}$ is given by*

$$f_D(d) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{V_1}(t_1) f_{V_2}(\frac{t_2}{t_1}) \dots f_{V_{p-1}}(\frac{t_{p-1}}{t_{p-2}}) f_{V_p}(\frac{d}{t_{p-1}}) \frac{1}{|t_1 \cdot t_2 \dots t_{p-1}|} dt_1 dt_2 \dots dt_{p-1} \quad (3.23)$$

where $f_{V_i}(\cdot)$ is the *pdf* of the random variable $V_i \equiv d_i(Y_i)^{w_i}|\mathbf{x}$ for $i = 1, 2, \dots, p$.

This results can not be extended to the case of non-independent random variables. It has also to be noted that the range of each integral can be simplified as the marginal densities $f_{V_i}(v_i)$ are positive for $0 < v_i < 1$ if the i^{th} response must be maximised or minimised and for $0 < v_i < d(T)^w = 0.9772^w$ if the i^{th} response must reach some optimal target value T (see propositions 4-5-6). According to those domains to have $f_{V_i}(v_i) > 0$, the density $f_D(d)$ is non zero for $d \in (0, 1)$ without any targeted property but could be more narrow otherwise.

The *pdf* of $D(\mathbf{Y})|\mathbf{x}$ (3.23) is difficult to handle analytically as the multiple integral to solve makes use of complex marginal *pdf*'s $f_{V_i}(\cdot)$. The quantity of interest $E[D(\mathbf{Y})|\mathbf{x}]$ can, on the other hand, be computed without using the *pdf* of $D(\mathbf{Y})|\mathbf{x}$ as the expectation of a product of independent random variables is the product of the marginal expectations.

Proposition 11 (Expectation of the weighted geometric mean). *If $Y_1|\mathbf{x}, Y_2|\mathbf{x}, \dots, Y_p|\mathbf{x}$ are mutually independent, then*

$$E[D(\mathbf{Y})|\mathbf{x}] = E\left[\prod_{i=1}^p (d_i(Y_i))^{w_i}|\mathbf{x}\right] = \prod_{i=1}^p E[(d_i(Y_i))^{w_i}|\mathbf{x}] \quad (3.24)$$

where each individual expectation $E[(d_i(Y_i))^{w_i}|\mathbf{x}]$, $i = 1, 2, \dots, p$, can be computed using the pdf of $(d_i(Y_i))^{w_i}|\mathbf{x}$ presented in proposition 4, 5 or 6.

Proposition 11 can be used to associate to any combination of factors levels $\mathbf{x}$ the expected desirability of the resulting product $E[D(\mathbf{Y})|\mathbf{x}]$ and find the corresponding optimum. This approach is different from the classical use of DI, $\prod_{i=1}^p (d_i(E[Y_i|\mathbf{x}]))^{w_i}$, as $(d_i(\cdot))^{w_i}$ is a non linear function.

3.4.5 Distribution of the weighted arithmetic mean of DFs

The weighted arithmetic mean is constructed by summing the individual DFs multiplied by a constant weight: $D(\mathbf{Y})|\mathbf{x} = \sum_{i=1}^p w_i \cdot d_i(Y_i)|\mathbf{x}$ with $w_i \in (0, 1)$ and $\sum_{i=1}^p w_i = 1$. This DI is a random variable, abbreviated D , constructed by the sum of p continuous random variables $V_i \equiv w_i \cdot d_i(Y_i)|\mathbf{x}$: $D = \sum_{i=1}^p V_i$. There is a well-known result (Mood *et al.* 1974) for the case of the sum of two random variables (Mood *et al.* 1974).

Theorem 3 (Density of the sum of two random variables). *Let (V_1, V_2) be a vector of two continuous random variables with known joint distribution $f_{V_1, V_2}(\cdot, \cdot)$ then $W \equiv V_1 + V_2$ is a continuous random variable with pdf given by*

$$f_W(w) = \int_{t=-\infty}^{\infty} f_{V_1, V_2}(t, w-t)dt = \int_{t=-\infty}^{\infty} f_{V_1, V_2}(w-t, t)dt \quad (3.25)$$

This theorem is easily proven by deriving the cdf of $W \equiv V_1 + V_2$. Unfortunately theorem 3 can not be generalised to obtain the pdf of the sum of p continuous random variables. The only result about the pdf of the sum of p continuous random variables necessitates the independence assumption. This leads to the next corollary:

Corollary 1 (Density of the sum of two independent random variables). *Let (V_1, V_2) be a vector of two continuous independent random variables with known marginal distributions $f_{V_1}(\cdot)$ and $f_{V_2}(\cdot)$ then $W \equiv V_1 + V_2$ is a continuous random variable with pdf given by the convolution of $f_{V_1}(\cdot)$ and $f_{V_2}(\cdot)$:*

$$\begin{aligned} f_W(w) &= f_{V_1} * f_{V_2}(x) = \int_{t=-\infty}^{\infty} f_{V_1}(t) \cdot f_{V_2}(w-t)dt \\ &= f_{V_2} * f_{V_1}(x) = \int_{t=-\infty}^{\infty} f_{V_1}(w-t) \cdot f_{V_2}(t)dt \end{aligned} \quad (3.26)$$

This corollary can be extended, by recurrence, to any number of independent random variables. This way, the *pdf* of the weighted arithmetic mean can be obtained only under the independence assumption.

Proposition 12 (weighted arithmetic mean). *Under the assumption that $Y_1|\mathbf{x}, Y_2|\mathbf{x}, \dots, Y_p|\mathbf{x}$ are mutually independent, the *pdf* of $D \equiv D(\mathbf{Y})|\mathbf{x} = \sum_{i=1}^p w_i \cdot d_i(Y_i)|\mathbf{x}$ is given by*

$$f_D(d) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{V_1}(v_1) \cdots f_{V_{p-1}}(v_{p-1}) f_{V_p}(d - v_{p-1} - \cdots - v_1) dv_1 dv_2 \cdots dv_{p-1} \quad (3.27)$$

where $f_{V_i}(\cdot)$ is the *pdf* of the random variable $V_i \equiv w_i \cdot d_i(Y_i)|\mathbf{x}$ for $i = 1, 2, \dots, p$.

The range of each integral can be simplified according to the values of v_i such that $f_{V_i}(v_i) > 0$ (see propositions 7-8-9). According to those domains to have $f_{V_i}(v_i) > 0$, the density $f_D(d)$ is non zero for $d \in (0, 1)$ without any targeted property but could be more narrow otherwise.

The *pdf* of $D(\mathbf{Y})|\mathbf{x}$ is difficult to handle analytically as the multiple integral to solve makes use of complex marginal *pdf*'s $f_{V_i}(\cdot)$. The quantity of interest $E[D(\mathbf{Y})|\mathbf{x}]$ can, on the other hand, be computed without using the *pdf* of $D(\mathbf{Y})|\mathbf{x}$ as the expectation of a sum is the sum of the expectations for independent or non-independent random variables.

Proposition 13 (Expectation of the weighted arithmetic mean).

$$E[D(\mathbf{Y})|\mathbf{x}] = E\left[\sum_{i=1}^p w_i d_i(Y_i)|\mathbf{x}\right] = \sum_{i=1}^p E[w_i d_i(Y_i)|\mathbf{x}] = \sum_{i=1}^p w_i E[d_i(Y_i)|\mathbf{x}] \quad (3.28)$$

where each individual expectation $E[d_i(Y_i)|\mathbf{x}]$, $i = 1, 2, \dots, p$, can be computed using the *pdf* of $d_i(Y_i)|\mathbf{x}$ presented in propositions 1, 2 and 3.

This proposition can be used to associate to any combination of factors levels $\mathbf{x}$ the expected desirability of the resulting responses $E[D(\mathbf{Y})|\mathbf{x}]$ and to find the corresponding optimum. Like the geometric mean DI, $E[\sum_{i=1}^p w_i d_i(Y_i)|\mathbf{x}]$ is different from the classical use of DI $\sum_{i=1}^p w_i d_i(E[Y_i])|\mathbf{x}$ as $d_i(\cdot)$ is a non linear function.

3.4.6 Distribution of the minimum of DFs

A third often used way to summarise the individual DFs is to take their minimum: $D(\mathbf{Y})|\mathbf{x} = \min_{i=1,2,\dots,p} d_i(Y_i)|\mathbf{x}$. This desirability index is a random variable, abbreviated D , constructed as the min of p continuous random variables $V_i \equiv d_i(Y_i)|\mathbf{x}$: $D = \min_{i=1,2,\dots,p} V_i$. In most cases, nothing can be said about the density of the minimum of random variables without assuming their independence (Mood *et al.* 1974).

Theorem 4 (Density of the minimum of p random variables). *If $V_1, V_2, \dots, V_p$ are independent random variables with pdf $f_{V_1}(\cdot), f_{V_2}(\cdot), \dots, f_{V_p}(\cdot)$ and cdf $F_{V_1}(\cdot), F_{V_2}(\cdot), \dots, F_{V_p}(\cdot)$, then the pdf of $D = \min_{i=1,2,\dots,p} V_i$ is given by*

$$f_D(d) = \sum_{i=1}^p \left[f_{V_i}(d) \cdot \prod_{\substack{j=1 \\ j \neq i}}^p (1 - F_{V_j}(d)) \right] \quad (3.29)$$

Applying this result to the minimum of DFs leads to the following proposition:

Proposition 14 (Pdf of the minimum). *Under the assumption that $Y_1|\mathbf{x}, Y_2|\mathbf{x}, \dots, Y_p|\mathbf{x}$ are mutually independent, the pdf of $D \equiv D(\mathbf{Y})|\mathbf{x} = \min_{i=1,2,\dots,p} d_i(\mathbf{Y}_i)|\mathbf{x}$ is given by*

$$f_D(d) = \sum_{i=1}^p \left[f_{V_i}(d) \cdot \prod_{\substack{j=1 \\ j \neq i}}^p (1 - F_{V_j}(d)) \right] \quad (3.30)$$

where $f_{V_i}(\cdot)$ and $F_{V_i}(\cdot)$ are respectively the pdf's and the cdf's of the random variables $V_i \equiv d_i(Y_i)|\mathbf{x}$ for $i = 1, 2, \dots, p$.

The values of d for which $f_D(d)$ is non zero depends on the types of individual DFs $d_i(\mathbf{Y}_i)|\mathbf{x}$. Without any targeted response, $d \in (0, 1)$, but this could be more narrow otherwise.

The expected minimum desirability can then be computed by intergration of its density for any combination of factors levels $\mathbf{x}$ and the optimum can be found using an adequate optimisation algorithm.

3.5 Distribution of DI when models are estimated

All the densities presented in the preceding sections depend on the quantities $E[Y_i|\mathbf{x}]$ and $\sigma_{\epsilon_i}^2$ ($i = 1, 2, \dots, p$), that are unknown in practice. Indeed, the models given by equation (3.1) are not known and have to be estimated. Using powerful tools of experimental design, a set of n experiments within the domain of interest χ can be adequately chosen and performed to collect data on which the assumed models can be fitted. For each property, the estimated model is denoted by

$$\hat{Y}_i = \hat{f}_i(\mathbf{x}, \hat{\beta}_i). \quad (3.31)$$

Thanks to equation (3.31), the responses can be predicted for any point $\mathbf{x}$ in the experimental domain χ , and the unknown quantities, $E[Y_i|\mathbf{x}]$ and $\sigma_{\epsilon_i}^2$ ($i = 1, 2, \dots, p$), can be

estimated by

$$\hat{E}[Y_i|\mathbf{x}] = \hat{Y}_i = \hat{f}_i(\mathbf{x}, \hat{\beta}_i) \quad \text{and} \quad \hat{\sigma}_{\epsilon_i}^2 = \frac{1}{n - q_i} \sum_{j=1}^n (y_{ij} - \hat{y}_{ij})^2 \quad (3.32)$$

where n is the number of observations, q_i is the number of model parameters, y_{ij} ($j = 1, 2, \dots, n$) are the observed values for the i^{th} property and $\hat{y}_{ij}$ are the corresponding predicted values.

Before using the estimated models for prediction, it's crucial to analyze their adequacy for the training data (using R^2 , adjusted R^2 , goodness-of-fit tests, ...) and their performance for predicting responses of new points $\mathbf{x}$ (using a test set, cross-validation, bootstrap, ...). Nevertheless, even if the conclusions of these analyses are good, $\hat{E}[Y_i|\mathbf{x}]$ and $\hat{\sigma}_{\epsilon_i}^2$ are only uncertain estimation of the true parameters $E[Y_i|\mathbf{x}]$ and $\sigma_{\epsilon_i}^2$, leading to uncertain estimation of the densities of DFs and DI, as well as uncertain estimation of the expectation of interest: $E[D(\mathbf{Y})|\mathbf{x}]$.

There is no analytical method available to quantify the propagation of the uncertainty of the estimated parameters $E[Y_i|\mathbf{x}]$ and $\sigma_{\epsilon_i}^2$ on the estimated expected desirability index $E[D(\mathbf{Y})|\mathbf{x}]$. This is the subject of the next chapter. If the uncertainty is small, *i.e.* $E[Y_i|\mathbf{x}]$ and $\sigma_{\epsilon_i}^2$ are precisely estimated by $\hat{E}[Y_i|\mathbf{x}]$ and $\hat{\sigma}_{\epsilon_i}^2$, using the estimated densities provides quite accurate estimate of $E[D(\mathbf{Y})|\mathbf{x}]$.

3.6 Conclusion

New DFs based on the standard Normal distribution are presented for cases of maximised, minimised or targeted responses. The parameters of those functions are easy to choose in practice with or without the knowledge of a production specialist.

Due to the error terms of the assumed models linking the responses to the factors, DFs and DIs are random variables and the optimisation consists of maximising the expected desirability index to take the random errors into account, as suggested by Steuer (2000).

The densities of the Normal *cdf* based DFs are derived analytically using the univariate density transformation theorem under the assumption that the error terms are Normally distributed.

Under the assumption that conditionally to the factors levels, the responses are independent, the densities of the geometric mean and the arithmetic mean DIs are provided by applying two theorems of the statistical literature about the density of the product of

random variables and the sum of random variables (*convolution*). Those densities can be used to compute some kind of prediction interval for the DI of a new experiment. But they are difficult to handle when integrating to obtain the expected DI. Under the same independence assumption, the expected DI can be computed on the basis of the expected weighted DFs for which densities are also known and more easy to handle. The density of the minimum of DFs is also derived and can be integrated to obtain the expected minimum.

The developed densities assume that the models parameters as well as the errors variances are known. Nevertheless, in practice, only estimates of those quantities are available. If those estimates are quite accurate, the computation of the expected DI on the basis of estimated densities is of reasonable precision. Otherwise the estimated expected DI is suiled with error and the propagation of the models prediction error has to be taken into account. A way to quantify this uncertainty propagation through the use of the Delta method theorem is discussed in the next chapter.

Bibliography

- Box, G. and Draper, N. (1986), *Empirical model-building and response surface*, Wiley: New York.
- Dehhaoui, M. (1996), *Analyse multivarié des surfaces de réponse en relation avec les plans expérimentaux*, Thèse dfendue en 1996 - Faculté d'agronomie, Université catholique de Louvain, Belgium.
- Derringer, G.C. and Suich, R. (1980), "Simultaneous Optimization of Several Response Variables", *J. Qual. Tech.*, 12(4), 214-219.
- Derringer, G.C. (1994), "A balancing act: optimizing a product's properties", *Quality Progress*, June, 51-58.
- Khuri, A.I. and Cornell, J.A. (1987), *Response surfaces: designs and analyses*, Marcel Dekker, Inc., New York.
- Gibb, R.D., Carter, W.H., and Myers, R.H. (2001), *Incorporating experimental variability in the determination of desirable factor levels*, Unpublished manuscript.
- Govaerts, B., and Le Bailly de Tilleghe, C. (2005), *Distribution of Desirability Index in Multicriteria Optimization using Desirability Functions based on the Cumulative Distribution Function of the Standard Normal*, Discussion Paper 0531, Institut de Statistique, Univeristé catholique de Louvain, Louvain-la-Neuve, Belgium.
- Govaerts, B. and Le Bailly de Tilleghe, C. (2005), *Uncertainty Propagation in Multire-sponse Optimization using a Desirability Index*, Discussion Paper 0532, Institut de Statistique, Univeristé catholique de Louvain, Louvain-la-Neuve, Belgium.
- Harrington, E.C. (1965), "The desirability function", *Industrial Quality Control*, 21, 494-498.
- Jessenerger, J. (1999), *Prozessfaehigkeitsindizes in der Qualitätssicherung*, Libri Books on Demand; Dissertation at the Faculty of Statistics of the University of Dortmund.

- Kim, K. and Lin, D. (2000), "Simultaneous optimization of mechanical properties of steel by maximizing exponential desirability functions", *Applied Statistics*, 49, 311-325.
- Khuri, A. I. and Conlon, M. (1981), "Simultaneous Optimization of Multiple Responses Represented by Polynomial Regression Functions", *Technometrics*, 23, 363-376.
- Ko, Y.H., Kim, K.J. and Jun, C.H. (2005), "A new loss function-based method for multiresponse optimization", *Journal of Quality Technology*, 37(1), 50-59.
- Mao, M. (2007), *Optimisation multi-facteurs/multi-objectifs, application l'optimisation de profil organoleptique de produits secs par le contrôle des paramètres de matières premières et de procédés*, Thèse défendue en 2007 - Institut des sciences et industries du vivant et de l'environnement (AgroParisTech), France.
- Mood, A.M., Graybill F.A. and Boes, D.C (1974), *Introduction to the theory of statistics*, 3rd Edition, McGrawHill, Singapore.
- Myers, R.H., Khuri A.I. and Carter W.I. Jr (1989), "Response surface methodology: 1966-1988", *Technometrics*, 31(2), 137-157.
- Myers, R.H. and Montgomery, D.C. (2002), *Response surface methodology: process and product optimization using designed experiments*, 2nd Edition, Wiley: New York.
- von Neumann, J. and Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton NJ.
- Steuer, D. (2000), *An improved optimisation procedure for Desirability Indices*, Technical Report 27/00, SFB 475, Dortmund University.
- Trautmann, H. and Weihs, C. (2004), *Uncertainty of the optimum influence factor levels in multicriteria optimization using the concept of desirability*, Technical Report 23/04, SFB 475, Dortmund University.
- Trautmann, H. and Weihs, C. (2006), "On the Distribution of the Desirability Index using Harrington's Desirability Function", *Metrika*, 63, 207-213.
- Weber, H. and Weihs, C. (2003), *On the distribution of the desirability index using Harrington's desirability function*, Forschungsbericht 2003/3, Statistics Department, Dortmund University.

Chapter 4

Uncertainty propagation on desirability index

Part of this chapter constitutes the discussion paper DP0532 (Govaerts and Le Bailly de Tilleghe 2005b) of the Institute of Statistics, Université catholique de Louvain, Belgium. It has been submitted to the Journal of Quality Technology and Quantitative Management.

4.1 Introduction

Multicriteria optimisation (see Chapter 3, Section 1) is a common problem in industries. Each response defining the quality of the product, Y_i ($i = 1, 2, \dots, p$), is assumed to be related to the same set of varying factors, x_j 's ($j = 1, 2, \dots, K$), through an equation of the form

$$Y_i = f_i(\mathbf{x}, \boldsymbol{\beta}_i) + \epsilon_i, \quad (4.1)$$

where f_i is supposed to be the true link function depending on the unknown model parameters $\boldsymbol{\beta}_i$ and the error term ϵ_i is often assumed to follow a $N(0, \sigma_{\epsilon_i}^2)$ distribution. Moreover, in this chapter, the error terms ϵ_i and ϵ_j corresponding to two different responses are assumed to be independent. The objective is to find the factors settings $\mathbf{x} = (x_1, x_2, \dots, x_K)$ that simultaneously optimise the p responses $\mathbf{Y} = (Y_1, Y_2, \dots, Y_p)$.

The quality of a compromise between the responses can be measured by the *desirability* concept (see Chapter 3, Section 2) introduced by Harrington (1965). Using *desirability functions* (DFs), d_i , the adequacy of each of the p responses, Y_i , are first quantified by a value between 0 and 1 (the higher, the better). Those p values are then aggregated in a *desirability index* (DI), D , providing, for any combination of factors levels, a value between 0 and 1 measuring the desirability of the resulting product quality.

This chapter treats general desirability functions of the form $d_i : \mathfrak{R} \rightarrow]0, 1]$, that are two times continuously differentiable and supposes that they are aggregated with the weighted geometric mean desirability index (the most common DI). This differentiability property is not valid for the Derringer and Suich desirability functions and for the Harrington one when a target value is sought. Their DFs can easily be approximated by Logit and Normal *cdf* based DFs which have better statistical and optimisation properties. Similar results may also be obtained with the weighted arithmetic mean (3.3) using the same tools to develop analogue formulae.

As the i^{th} response for given factor levels, $Y_i|\mathbf{x}$, is a random variable (4.1), there are two possible treatments of this randomness when computing DI (see Chapter 3, Section 3), the classical approach (Harrington 1965; Derringer and Suich 1980) and the stochastic approach (Steuer 2000):

$$\max_{\mathbf{x} \in \chi} D^C(\mathbf{x}) = \max_{\mathbf{x} \in \chi} D(E[\mathbf{Y}|\mathbf{x}]) = \max_{\mathbf{x} \in \chi} \prod_{i=1}^p [d_i(E[Y_i|\mathbf{x}])]^{w_i} \quad \text{or} \quad (4.2)$$

$$\max_{\mathbf{x} \in \chi} D^S(\mathbf{x}) = \max_{\mathbf{x} \in \chi} E[D(\mathbf{Y})|\mathbf{x}] = \max_{\mathbf{x} \in \chi} E \left[\prod_{i=1}^p [d_i(Y_i)]^{w_i} | \mathbf{x} \right]. \quad (4.3)$$

Maximising the desirability of the expected responses, $D^C(\mathbf{x})$ (4.2), is the classical methodology to transform the multicriteria optimisation into a unique objective criteria.

In practice, the models assumed by equation (4.1) are unknown. Data are collected following an experimental design, possibly different for each response, and each theoretical model is fitted. Let us suppose that, for each property, the corresponding estimated model is denoted by

$$\hat{Y}_i = \hat{f}_i(\mathbf{x}, \hat{\beta}_i). \quad (4.4)$$

The factors included in each model can also be a subset $\mathbf{x}_i$ of the full vector $\mathbf{x}$. Each fitted model can be used to estimate the expected response for any point $\mathbf{x}$ in the experimental domain by $\hat{E}[Y_i|\mathbf{x}] = \hat{f}_i(\mathbf{x}, \hat{\beta}_i)$ and estimate the corresponding desirability by $\hat{D}^C(\mathbf{x}) = D(\hat{E}[\mathbf{Y}|\mathbf{x}])$. The optimal factors settings are traditionally found by applying an adequate screening algorithm over the experimental domain in order to maximise $\hat{D}^C(\mathbf{x})$.

In spite of this success, this classical methodology has a main drawback: it neglects the models prediction errors. For nearly all kinds of models, it is possible to quantify the uncertainty when using fitted models to estimate the expected responses for a new experimental domain point (see Chapter 2, Section 2.5). The uncertainty of estimated expected responses propagates on the desirability index, but this information is not taken into account in the classical optimisation approach. Desirability indexes of two solutions are compared as if

they were exact although they are random variables.

Steuer (2000) was the first author to discuss the problem of model uncertainty in the context of desirability optimisation. He suggests to take it into account by maximising the expected desirability index (4.3) over the factors domain instead of the desirability index of the estimated expected response (4.2) like in the classical approach.

Other problems follow from the stochastic character of the desirability index. First, due to response model uncertainty, the expected desirability index can only be estimated. A formula for this estimator must then be derived and one may also be interested in calculating a confidence interval to quantify the uncertainty of this estimator. Second, the uncertainty of the estimated expected desirability index leads to an uncertainty on the estimated optimal factors settings, and an *indistinguishable optimal zone* may also be derived to delimit the subset of the factors domain which contains not significantly worse solutions than the optimum found. Third, real experiments are always affected by experimental error. An experiment performed at the optimal factors settings will never give the expected response and desirability index even if the models fit “perfectly” the average responses. For a solution $\mathbf{x}$ in the factors domain, the calculation of a prediction interval for the (observed) desirability index is then also a question of interest.

These problems are partially discussed in the literature. When a single response is concerned, Box and Hunter (1954) proposed an exact method to derive a confidence region for the stationary point of an estimated second order polynomial model. Stablein *et al.* (1983) generalised it to the constraint experimental region using Lagrange multipliers. Böckenholt (1989) summarised the results of the former authors and proposed a method to test if the optima of different response surfaces are statistically different. Still for the single response problem, Peterson *et al.* (2002) proposed an improvement of the Stablein *et al.* method which can be used for general linear regression models and is more easy to apply provided that an adapted computer program is available.

For the multiple response case, Steuer (2000) pointed out the interest to work on the expected desirability index. Weber and Weihs (2003) and Trautmann and Weihs (2006) derived analytically the statistical distribution of the desirability index for Harrington DFs and Govaerts and Le Bailly de Tillegem (2005a) for Normal *cdf* based DFs (3.4-3.5-3.6). Cahya (2002) proposed a way to derive a confidence region for the optimal factors settings based on the Delta method theorem and Trautmann and Weihs (2004) proposed a way to derive an indistinguishable optimal zone based on a prediction interval for the desirability index.

This chapter proposes a general framework to take into account model and desirability

index uncertainty in multicriteria optimisation. This integrated methodology is applicable in the classical context of experimental design with regression models but also to different model classes and other area like molecules optimisation. Under general assumptions over the form of response models and DFs and for the geometric mean DI, this chapter proposes formulas to estimate the expected desirability index, $D^S(\mathbf{x}) = E[D(\mathbf{Y})|\mathbf{x}]$, calculate related confidence and prediction intervals and derive an indistinguishable optimal zone around the optimum found. All the related concepts discussed in the existing literature are encompassed in one framework. Most results are based on variations of the Delta method theorem that shows, through simulations, to perform very well.

In the next section, an estimator of the expected DI is proposed and Section 3 shows how to derive a corresponding confidence interval. Section 4 proposes a method to set up an indistinguishable optimal zone for the factors settings that do not have significantly smaller expected DI than the optimum found and Section 5 explains then how to construct a prediction interval for the DI when applying desirability index directly on observed responses. Section 6 discusses an extension of the methodology to the case of correlated models residuals.

Finally, in Section 7, the different concepts are illustrated on a simple simulated optimisation example. The method is then also applied in Section 8 to the Derringer and Suich (1980) classical dataset to demonstrate its interest on a well known example response surface methodology.

4.2 Estimation of the expected desirability index $D^S(\mathbf{x}) = E[D(\mathbf{Y})|\mathbf{x}]$

The Delta method is often used to approximate the variance of a function of random variables. This method can also be used to approximate expectation. Indeed, on the basis of the Taylor expansion of a differentiable function h of random variables $Y_1, Y_2, \dots, Y_p$, it can be shown that the following second order approximations hold (Hahn and Shapiro 1969):

$$\begin{aligned} E[h(Y_1, Y_2, \dots, Y_p)] &\approx h(E[Y_1], E[Y_2], \dots, E[Y_p]) \\ &+ \frac{1}{2} \sum_{i=1}^p \left[\frac{\partial^2 h}{\partial Y_i^2} V[Y_i] \right] + \sum_{i < j}^p \left[\frac{\partial^2 h}{\partial Y_i \partial Y_j} Cov(Y_i, Y_j) \right] \end{aligned} \quad (4.5)$$

$$V[h(Y_1, Y_2, \dots, Y_p)] \approx \sum_{i=1}^p \left[\left(\frac{\partial h}{\partial Y_i} \right)^2 V[Y_i] \right] + 2 \sum_{i < j}^p \left[\left(\frac{\partial h}{\partial Y_i} \right) \left(\frac{\partial h}{\partial Y_j} \right) Cov(Y_i, Y_j) \right] \quad (4.6)$$

If we assume that, conditionally to $\mathbf{x}$, the responses are independent, *i.e.* $(Y_i|\mathbf{x}) \perp (Y_j|\mathbf{x}) \forall i \neq j$, the approximation (4.5) can be rewritten as

$$E[h(Y_1, \dots, Y_p)|\mathbf{x}] \approx h(E[Y_1|\mathbf{x}], \dots, E[Y_p|\mathbf{x}]) + \frac{1}{2} \sum_{i=1}^p \left[\frac{\partial^2 h}{(\partial Y_i)^2} V[Y_i|\mathbf{x}] \right], \quad (4.7)$$

where the second derivatives are evaluated in the conditional expectations $E[Y_i|\mathbf{x}]$. In practice, in the context of experimental design, even if the responses are often correlated, the conditional independence between $Y_i|\mathbf{x}$ and $Y_j|\mathbf{x}$ is most of the time encountered. This assumption, its validation and an extension of the methodology to the case of correlated models residuals are discussed in Section 6.

As the DI for given factor settings $\mathbf{x}$ is a function of the random responses $Y_1|\mathbf{x}, Y_2|\mathbf{x}, \dots, Y_p|\mathbf{x}$, if the DFs are two times continuously differentiable and non-zero, formula (4.7) can be applied to approximate the expected DI conditionally to $\mathbf{x}$ to the second order as

$$D^S(\mathbf{x}) = E[D(\mathbf{Y})|\mathbf{x}] = E \left[\prod_{i=1}^p (d_i(Y_i))^{w_i} | \mathbf{x} \right] \approx \left[\prod_{i=1}^p (d_i(E[Y_i|\mathbf{x}]))^{w_i} \right] \cdot (1 + B(\mathbf{x})) \quad (4.8)$$

where $B(\mathbf{x})$ is given by

$$B(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^p c_i(\mathbf{x}) \cdot V[Y_i|\mathbf{x}]. \quad (4.9)$$

The first part of the approximation (4.8) is the classical desirability $D^C(\mathbf{x})$, and the half sum $B(\mathbf{x})$ in the second term is a kind of bias correction that takes into account the non-linearity of desirability index. The constants $c_i(\mathbf{x})$ weighting the residual variances in $B(\mathbf{x})$ (4.9) depend on the weights w_i , as well as the first and second derivatives of each DF:

$$c_i(\mathbf{x}) = w_i(w_i - 1) \left(\frac{\frac{\partial d_i(Y_i)}{\partial Y_i}}{d_i(E[Y_i|\mathbf{x}])} \right)^2 + w_i \frac{\frac{\partial^2 d_i(Y_i)}{\partial Y_i^2}}{d_i(E[Y_i|\mathbf{x}])}. \quad (4.10)$$

$D^S(\mathbf{x})$ can be estimated by replacing in the approximation (4.8) the unknown quantities $E[Y_i|\mathbf{x}]$ and $V[Y_i|\mathbf{x}]$ by their usual estimators, the estimators given by the fitted models $\hat{E}[Y_i|\mathbf{x}]$ and the estimators of the residual variances $\hat{V}[Y_i|\mathbf{x}] = \hat{\sigma}_{\epsilon_i}^2$:

$$\hat{D}^S(\mathbf{x}) = \hat{E}[D(\mathbf{Y})|\mathbf{x}] = \hat{D}^C(\mathbf{x}) \cdot (1 + \hat{B}(\mathbf{x})). \quad (4.11)$$

The approximation of $E[D(\mathbf{Y})|\mathbf{x}]$ (4.8) is proved in Appendix 1. The expressions of the derivatives of the Normal *cdf* based DFs are given in Appendix 3.

In the framework of multicriteria optimisation and design of experiments, the most well-known class of models is the multiple linear regression model (Section 4.1 of Chapter 1). On the basis of n_i independent experiments, possibly different for each response, each model (1.1) can be fitted to data by least squares and each expected response is estimated by

$$\hat{E}[Y_i|\mathbf{x}] = \mathbf{z}_i' \hat{\boldsymbol{\beta}}_i = \mathbf{z}_i' (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' \mathbf{y}_i \quad i = 1, 2, \dots, p, \quad (4.12)$$

where $\mathbf{Z}_i$ is the $(n_i \times q_i)$ model matrix for response i and $\mathbf{y}_i$ the corresponding $(n_i \times 1)$ vector of observed responses y_{ij} , $(j = 1, 2, \dots, n_i)$. The usual residual variance estimator in the regression context is $\hat{V}[Y_i|\mathbf{x}] = \hat{\sigma}_{\epsilon_i}^2 = \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2 / (n_i - q_i)$. For PLS Regression or Regression Trees, this formula can be adapted using the effective number of parameters (number of latent variables or number of nodes, respectively).

For any point $\mathbf{x} \in \chi$, formulas (4.8) and (4.11) provide then a general method to approximate and estimate the expected desirability $D^S(\mathbf{x})$ defined in (4.3). This allows to find the estimated optimum $\hat{\mathbf{x}}_{opt}^S$ over χ as the factors settings that maximise $\hat{D}^S(\mathbf{x})$ using an adequate screening algorithm. This optimal solution may be different from the classical one $\hat{\mathbf{x}}_{opt}^C$ that maximises $\hat{D}^C(\mathbf{x}) = \hat{D}(E[\mathbf{Y}|\mathbf{x}])$, due to the bias term $\hat{B}(\mathbf{x})$ present in the estimation of $D^S(\mathbf{x})$.

4.3 Quantification of the uncertainty of the expected desirability index $\hat{D}^S(\mathbf{x})$

Assuming that each estimator $\hat{E}[Y_i|\mathbf{x}]$ is asymptotically Normal and unbiased, the Delta method theorem (von Mises 1947) can be used to derive the asymptotic properties of our estimator $\hat{D}^S(\mathbf{x})$ (4.11) and build a confidence interval for $D^S(\mathbf{x})$.

The Delta method theorem: Suppose that $\boldsymbol{\theta}$ is a vector of parameters estimated on n observations by $\hat{\boldsymbol{\theta}}$ with $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and g is a continuously differentiable function then $\sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \frac{\partial g}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma} \frac{\partial g'}{\partial \boldsymbol{\theta}})$.

In order to apply this theorem to $\hat{D}^S(\mathbf{x}) = \hat{D}^C(\mathbf{x}) \cdot (1 + \hat{B}(\mathbf{x}))$, let us first note that the uncertainty on $\hat{B}(\mathbf{x})$ is negligible compared to the uncertainty of the term $\hat{D}^C(\mathbf{x})$. Moreover, as $\hat{D}^C(\mathbf{x})$ is a continuously differentiable function of $\hat{E}[\mathbf{Y}|\mathbf{x}] = (\hat{E}[Y_1|\mathbf{x}], \dots, \hat{E}[Y_p|\mathbf{x}])$, $\hat{E}[\mathbf{Y}|\mathbf{x}]$ is the vector of parameters $\hat{\boldsymbol{\theta}}$ of the Delta method theorem with expected value $E[\mathbf{Y}|\mathbf{x}]$ and the diagonal variance-covariance matrix $\boldsymbol{\Sigma} = \text{diag}(V[\hat{E}[Y_1|\mathbf{x}]], \dots, V[\hat{E}[Y_p|\mathbf{x}]])$ since, conditionally to $\mathbf{x}$, the Y_i 's are considered as independent.

4.3 Quantification of the uncertainty of the expected desirability index $\hat{D}^S(\mathbf{x})$ 23

In addition, note that the normality hypothesis on $\hat{D}^S(\mathbf{x})$ may be questionable for values of $D^S(\mathbf{x})$ near 0 or 1. In order to better fulfil this assumption, Cahya (2002) and Ding *et al.* (2003) suggest to apply a logit transformation to $\hat{D}^C(\mathbf{x})$ before applying the theorem to get an unbounded support. Moreover, this insures that resulting confidence intervals are included in the interval $[0, 1]$. Simulations of Section 8 confirm that $\text{logit}(\hat{D}^S(\mathbf{x}))$ is better approximated by a Normal distribution than $\hat{D}^S(\mathbf{x})$, especially for values near 0 or 1.

As $\text{logit}(\hat{D}^S(\mathbf{x}))$ is also a continuously differentiable function of $\hat{E}[\mathbf{Y}|\mathbf{x}]$, the Delta method theorem leads then to the following result:

$$\text{logit}(\hat{D}^S(\mathbf{x})) \xrightarrow{d} N(\text{logit}(\hat{D}^S(\mathbf{x})), V[\text{logit}(\hat{D}^S(\mathbf{x}))]) \quad (4.13)$$

The variance $V[\text{logit}(\hat{D}^S(\mathbf{x}))]$ can be estimated as a function of $\hat{E}[Y_i|\mathbf{x}]$ and $\hat{V}[\hat{E}[Y_i|\mathbf{x}]]$:

$$\hat{V}[\text{logit}(\hat{D}^S(\mathbf{x}))] \approx \frac{\hat{V}[\hat{D}^S(\mathbf{x})]}{(\hat{D}^S(\mathbf{x}) \cdot (1 - \hat{D}^S(\mathbf{x})))^2}, \quad (4.14)$$

where, neglecting $V[\hat{B}(\mathbf{x})]$,

$$\hat{V}[\hat{D}^S(\mathbf{x})] \approx (1 + \hat{B}(\mathbf{x}))^2 \cdot \hat{V}[\hat{D}^C(\mathbf{x})] \quad (4.15)$$

and

$$\hat{V}[\hat{D}^C(\mathbf{x})] \approx \sum_{i=1}^p \hat{a}_i^2(\mathbf{x}) \cdot \hat{V}[\hat{E}[Y_i|\mathbf{x}]]. \quad (4.16)$$

The squared constants weighting the variances of expected responses in (4.16) depend on the weights w_i , the first derivative of each DF and the estimated classical DI:

$$\hat{a}_i(\mathbf{x}) = w_i \cdot \frac{\frac{\partial d_i(Y_i)}{\partial Y_i}}{d_i(\hat{E}[Y_i|\mathbf{x}])} \cdot \hat{D}^C(\mathbf{x}). \quad (4.17)$$

An asymptotic $(1 - \alpha) \cdot 100\%$ confidence interval can then be derived for $\text{logit}(\hat{D}^S(\mathbf{x}))$ as

$$\text{logit}(\hat{D}^S(\mathbf{x})) \pm z_{1-\alpha/2} \cdot \sqrt{\hat{V}[\text{logit}(\hat{D}^S(\mathbf{x}))]}, \quad (4.18)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -percentile of the standard normal.

Then, a $(1 - \alpha) \cdot 100\%$ confidence interval for $D^S(\mathbf{x})$ is obtained by transforming (4.18) using the inverse of the logit function:

$$\frac{\exp \left\{ \text{logit}(\hat{D}^S(\mathbf{x})) \pm z_{1-\alpha/2} \cdot \sqrt{\hat{V}[\text{logit}(\hat{D}^S(\mathbf{x}))]} \right\}}{1 + \exp \left\{ \text{logit}(\hat{D}^S(\mathbf{x})) \pm z_{1-\alpha/2} \cdot \sqrt{\hat{V}[\text{logit}(\hat{D}^S(\mathbf{x}))]} \right\}}. \quad (4.19)$$

Such confidence interval can be constructed for any $\mathbf{x} \in \chi$ to quantify the uncertainty of the estimation $\hat{D}^S(\mathbf{x})$. This interval is included in the $[0, 1]$ interval and is not necessarily symmetric.

It requires an additional estimator of $V[\hat{E}[Y_i|\mathbf{x}]]$. With an MLR model, the variance of the estimated expected response $\hat{E}[Y_i|\mathbf{x}]$ (4.12) can be estimated by

$$\hat{V}[\hat{E}[Y_i|\mathbf{x}]] = \hat{\sigma}_{\epsilon_i}^2 \mathbf{z}_i' (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{z}_i \quad (4.20)$$

See Chapter 2, Section 5, for analogous formulas corresponding to PLSR and RT models.

4.4 Set up of an Indistinguishable Optimal Zone (IOZ)

4.4.1 Objective of an Indistinguishable Optimal Zone (IOZ)

The optimal factors settings $\hat{\mathbf{x}}_{opt}^S$ are found as the point in the experimental domain that maximises the estimated expected desirability $\hat{D}^S(\mathbf{x})$. Because of the uncertainty of the estimations provided by the fitted models, this estimated optimum is certainly not the true one and there exists a set of points in the experimental domain that are equivalent or even more desirable than the optimum found. Let us define the *True Optimal Zone* as the subset of such points:

$$TOZ(\hat{\mathbf{x}}_{opt}^S) \equiv \{\mathbf{x} \in \chi \text{ s.t. } D^S(\mathbf{x}) \geq D^S(\hat{\mathbf{x}}_{opt}^S)\}. \quad (4.21)$$

The true optimum, $\mathbf{x}_{opt}^S$, that maximises $D^S(\mathbf{x})$ over χ belongs to $TOZ(\hat{\mathbf{x}}_{opt}^S)$.

As the true desirability is unknown, the TOZ can only be approximated. Let us define the *Indistinguishable Optimal Zone* as the subset of points in the experimental domain for which the hypothesis that they are equivalent or better than $\hat{\mathbf{x}}_{opt}^S$ can not be rejected at some significant level α :

$$IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S) \equiv \{\mathbf{x} \in \chi \text{ s.t. } H_0 : D^S(\mathbf{x}) \geq D^S(\hat{\mathbf{x}}_{opt}^S) \text{ is not rejected at level } \alpha\}. \quad (4.22)$$

The two zones are illustrated for a one factor example on Figure 4.1.

If a point is included in $IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$, the hypothesis that this point belongs to $TOZ(\hat{\mathbf{x}}_{opt}^S)$ can not be rejected. This point can not be significantly distinguished from the optimum found because of the uncertainty of the models that propagates on the desirability index. The IOZ might sometimes be pretty large in the case of high experimental variance and/or in the case of a small dataset. Nevertheless, for the points outside the IOZ, it is quite sure that

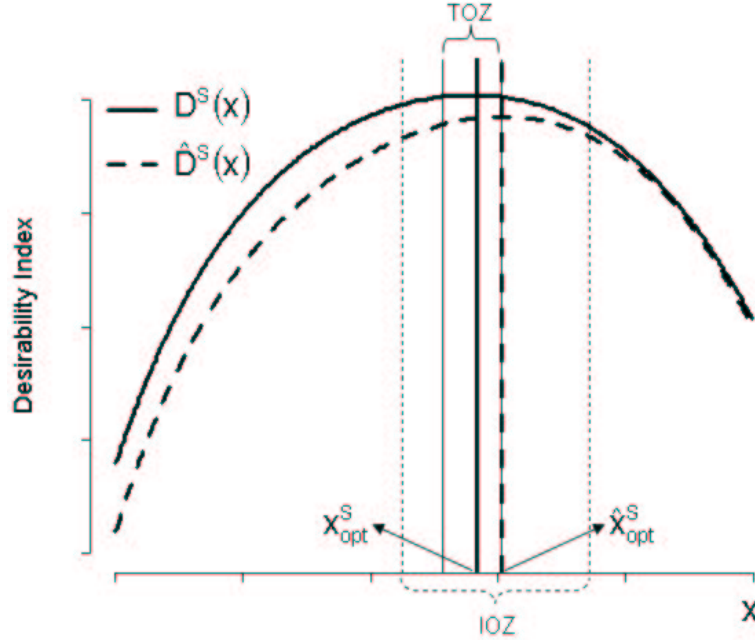


Figure 4.1: Example of True Optimal Zone and Indistinguishable Optimal Zone for a one-factor optimisation. The thick continuous and dashed curves are respectively the true expected desirability $D^S(x)$ and an estimated expected desirability $\hat{D}^S(x)$. The thick continuous and dashed vertical lines represent respectively the corresponding true optimum x_{opt}^S and the corresponding estimated optimum $\hat{x}_{opt}^S$. The thin continuous and dashed vertical lines delimit respectively the True Optimal Zone, $TOZ(\hat{x}_{opt}^S)$ (4.21), and the Indistinguishable Optimal Zone, $IOZ_{\alpha=0.05}(\hat{x}_{opt}^S)$ (4.22).

they are worse than the optimum found and, consequently, not interesting for the producers.

For any point $\mathbf{x} \in \chi$, a hypothesis test is required to decide if $\mathbf{x}$ is included in $IOZ_{\alpha}(\hat{\mathbf{x}}_{opt}^S)$. Such a test is developed in the next subsection.

4.4.2 Hypothesis test of $H_0 : D^S(\mathbf{x}) \geq D^S(\hat{\mathbf{x}}_{opt}^S)$ for a fixed point $\mathbf{x}$

For a given point $\mathbf{x} \in \chi$, a hypothesis test is built up to study the likelihood of

$$H_0 : D^S(\mathbf{x}) \geq D^S(\hat{\mathbf{x}}_{opt}^S) \quad \text{meaning that } \mathbf{x} \in TOZ(\hat{\mathbf{x}}_{opt}^S) \quad \text{versus} \quad (4.23)$$

$$H_1 : D^S(\mathbf{x}) < D^S(\hat{\mathbf{x}}_{opt}^S) \quad \text{meaning that } \mathbf{x} \notin TOZ(\hat{\mathbf{x}}_{opt}^S). \quad (4.24)$$

A test statistic could be constructed on the basis of the estimated difference $\hat{D}^S(\mathbf{x}) - \hat{D}^S(\hat{\mathbf{x}}_{opt}^S)$, but, as discussed in Section 3, the Normal distribution of the desirability estimator may be questionable for desirability near 0 or 1, and a monotonic transformation should

be applied. Section 8 confirm that $\text{logit}(\hat{D}^S(\mathbf{x})) - \text{logit}(\hat{D}^S(\hat{\mathbf{x}}_{opt}^S))$ is better approximated by a Normal distribution than $\hat{D}^S(\mathbf{x}) - \hat{D}^S(\hat{\mathbf{x}}_{opt}^S)$, especially for values of desirability near 0 or 1.

Let us then define the true and the estimated differences

$$\Delta_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S) = \text{logit}(D^S(\mathbf{x})) - \text{logit}(D^S(\hat{\mathbf{x}}_{opt}^S)) \quad \text{and} \quad (4.25)$$

$$\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S) = \text{logit}(\hat{D}^S(\mathbf{x})) - \text{logit}(\hat{D}^S(\hat{\mathbf{x}}_{opt}^S)). \quad (4.26)$$

As, by the Delta method theorem, both $\text{logit}(\hat{D}^S(\mathbf{x}))$ and $\text{logit}(\hat{D}^S(\hat{\mathbf{x}}_{opt}^S))$ are asymptotically Normal, their difference is also asymptotically Normal:

$$\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S) \xrightarrow{d} N\left(\Delta_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S), V[\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S)]\right), \quad (4.27)$$

with

$$\begin{aligned} V[\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S)] &= V[\text{logit}(\hat{D}^S(\mathbf{x}))] + V[\text{logit}(\hat{D}^S(\hat{\mathbf{x}}_{opt}^S))] \\ &\quad - 2 \cdot \text{Cov}(\text{logit}(\hat{D}^S(\mathbf{x})), \text{logit}(\hat{D}^S(\hat{\mathbf{x}}_{opt}^S))). \end{aligned} \quad (4.28)$$

Formula (4.14) can be applied to estimate $V[\text{logit}(\hat{D}^S(\mathbf{x}))]$ and $V[\text{logit}(\hat{D}^S(\hat{\mathbf{x}}_{opt}^S))]$. The covariance between the two *logit*'s cannot be neglected and can also be estimated by the Delta method theorem:

$$\begin{aligned} &\hat{\text{Cov}}(\text{logit}(\hat{D}^S(\mathbf{x})), \text{logit}(\hat{D}^S(\hat{\mathbf{x}}_{opt}^S))) \\ &\approx \frac{1 + \hat{B}(\mathbf{x})}{\hat{D}^S(\mathbf{x}) \cdot (1 - \hat{D}^S(\mathbf{x}))} \cdot \hat{\text{Cov}}(\hat{D}^C(\mathbf{x}), \hat{D}^C(\hat{\mathbf{x}}_{opt}^S)) \cdot \frac{1 + \hat{B}(\hat{\mathbf{x}}_{opt}^S)}{\hat{D}^S(\hat{\mathbf{x}}_{opt}^S) \cdot (1 - \hat{D}^S(\hat{\mathbf{x}}_{opt}^S))} \end{aligned} \quad (4.29)$$

where

$$\hat{\text{Cov}}(\hat{D}^C(\mathbf{x}), \hat{D}^C(\hat{\mathbf{x}}_{opt}^S)) \approx \sum_{i=1}^p \hat{a}_i(\mathbf{x}) \cdot \hat{\text{Cov}}(\hat{E}[Y_i|\mathbf{x}], \hat{E}[Y_i|\mathbf{x}_{opt}^S]) \cdot \hat{a}_i(\mathbf{x}_{opt}^S). \quad (4.30)$$

The constants $\hat{a}_i(\mathbf{x})$ and $\hat{a}_i(\mathbf{x}_{opt}^S)$ are defined in (4.17). In the case of linear regression models, the covariance between the estimations of $E[Y_i]$ in $\mathbf{x}$ and $\mathbf{x}_{opt}^S$ is given by

$$\hat{\text{Cov}}(\hat{E}[Y_i|\mathbf{x}], \hat{E}[Y_i|\mathbf{x}_{opt}^S]) = \hat{\sigma}_i^2 \mathbf{z}'(\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \hat{\mathbf{z}}_{opt}^S. \quad (4.31)$$

For PLSR and RT, see Chapter 2, Section 6.

Let us define the test statistic

$$T = \frac{\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S)}{\sqrt{V[\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S)]}} \xrightarrow{d} N(0, 1) \quad \text{under } H_0. \quad (4.32)$$

As the power function of the test decreases with $\Delta_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S)$, the composite hypothesis $\Delta_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S) \geq 0$ (4.23) is equivalent to the simple hypothesis $\Delta_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S) = 0$ and the asymptotic distribution (4.27) provides the asymptotic normality of the proposed test statistic.

In order to control the type I error with a probability of α , one may reject the null hypothesis (4.23) if the observed test statistic $T_{obs} = \frac{\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S)}{\sqrt{\hat{V}[\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S)]}}$ is smaller than z_α or, equivalently, if the p-value $P[Z < T_{obs}]$ is smaller than α . In such a case, $\mathbf{x} \notin IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$ and one can conclude that $\mathbf{x}$ is significantly worse than $\hat{\mathbf{x}}_{opt}^S$. Otherwise $\mathbf{x} \in IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$ and one can not conclude that $\mathbf{x}$ is significantly less desirable than $\hat{\mathbf{x}}_{opt}^S$.

This hypothesis test can be conducted for any point $\mathbf{x} \in \chi$ to check if $\mathbf{x}$ is significantly worse than the optimum found $\hat{\mathbf{x}}_{opt}^S$.

4.4.3 Multiple tests over χ

Theoretically, an infinity of hypothesis tests have to be conducted over the experimental domain χ to construct the IOZ (4.22). In practice, however, only a limited number of tests are performed for a grid defined over χ . As the number of grid points, denoted by m , can be huge, especially with a large number of factors, the global type I error rate for those multiple tests should be controlled by adapting the α level used for each individual test. In this context, the type I error consists in excluding a point $\mathbf{x}$ from $IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$ (concluding that $\mathbf{x}$ is significantly worse than $\hat{\mathbf{x}}_{opt}^S$) although $\mathbf{x}$ is in $TOZ(\hat{\mathbf{x}}_{opt}^S)$ ($\mathbf{x}$ is better than $\hat{\mathbf{x}}_{opt}^S$).

Different approaches are presented in the literature to solve this type I error rate explosion in multiple testing. The Bonferroni procedure is maybe the most well-known and the most simple solution (Bonferroni 1936). As the type I error rate for m independent tests is $1 - (1 - \alpha^*)^m$ if α^* is the significance level of each individual test, the level for each single test should be $\alpha^* = 1 - (1 - \alpha)^{1/m} \approx \alpha/m$ to control the type I error at a global significance level of α . Bonferroni proposed to take an adjusted significance level for each comparison equal to α/m .

The Bonferroni adjustment procedure controls the probability of erroneously rejecting even one of the true null hypotheses, which is called the familywise error-rate. Insuring

that all null hypotheses are jointly true is extremely conservative. There are two additional drawbacks of the Bonferroni method in this context. First, the adjusted significance level depends on the size of the grid, m . Refining the grid should not influence the regions of χ included in the IOZ. Second, this correction procedure relies on the assumption of independent multiple tests, which is certainly not the case for the grid of points prone to a spatial dependence.

Benjamini and Hochberg (1995) proposed a less conservative and more powerful approach by controlling the expected proportion of incorrectly rejected null hypotheses among all rejected null hypotheses, called *False Discovery Rate* and abbreviated *FDR*. In the context of the construction of the IOZ, the FDR is the expected proportion of points excluded from $IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$ (concluding that $\mathbf{x}$ is significantly worse than $\hat{\mathbf{x}}_{opt}^S$) that are however in $TOZ(\hat{\mathbf{x}}_{opt}^S)$ ($\mathbf{x}$ is better than $\hat{\mathbf{x}}_{opt}^S$) among all points excluded from $IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$. Their adjustment procedure works as follows:

The Benjamini-Hochberg procedure: Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the ordered observed p -values corresponding to the m tested null hypotheses $H_0^{(1)}, H_0^{(2)}, \dots, H_0^{(m)}$. Define $k = \max\{i : p_{(i)} \leq \frac{i}{m}q\}$ and reject $H_0^{(1)}, \dots, H_0^{(k)}$. If no such i exists, reject no hypothesis.

If the test statistics are independent, this procedure controls the FDR at level $\frac{m_0}{m}q \leq q$ where m_0 is the number of true null hypotheses and q is the user-defined control level. As already noticed previously, the test statistics (4.32) are certainly not independent. Benjamini and Yekutieli (2001) proposed an adaptation of the Benjamini-Hochberg procedure to any dependence structure by replacing q by $\frac{q}{\sum_{i=1}^m 1/i}$ in the definition of k . This modified procedure controls the FDR at level less than or equal to $\frac{m_0}{m}q$. The ratio $\frac{m_0}{m}$ and the definition of the rejection level $\frac{i}{m} \frac{q}{\sum_{i=1}^m 1/i}$ are quite stable with the size of the grid used to construct the IOZ.

Both approaches, familywise error rate (Bonferroni (1936)) or false discovery rate (Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001)), are aimed to decrease the number of erroneous rejections by adjusting the significance level of each individual test. Instead of comparing each p -value to a fixed global threshold of α or q , smaller thresholds are used (either α/m for Bonferroni either $\frac{i}{m}\alpha$ or $\frac{i}{m} \frac{\alpha}{\sum_{i=1}^m 1/i}$ for Benjamini) leading to less rejections and an increase of the IOZ size. Those multiplicity corrections insure a control of the type I error that is certainly too conservative for Bonferroni and possibly conservative for Benjamini and Yekutieli as the ratio $\frac{m_0}{m}$ might be small.

Those different approaches are compared by simulations in Sections 7 and 8 to check

that a multiplicity correction is necessary to insure a good coverage of the TOZ and which one is the best.

4.4.4 Discussion

The definition of $IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$ (4.22) is a way to take into account the propagation of the models uncertainty on the optimum found $\hat{\mathbf{x}}_{opt}^S$. The whole experimental domain or only a portion of interest is screened to gather factors settings that are associated to expected desirabilities not significantly smaller than the one associated to $\hat{\mathbf{x}}_{opt}^S$. Such factors settings can not be distinguished from $\hat{\mathbf{x}}_{opt}^S$.

A point $\mathbf{x}$ ideally belongs to $IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$ because its true desirability $D^S(\mathbf{x})$ is indeed nearly equal or higher to the one of the optimum found $D^S(\hat{\mathbf{x}}_{opt}^S)$, but this may also be due to the high variability of a response, the sparseness of data or to a bad choice of the experimental design providing uncertain predictions. Such situations lead to high variance in the estimated desirability index and do not allow to reject the hypothesis $H_0 : D^S(\mathbf{x}) \geq D^S(\hat{\mathbf{x}}_{opt}^S)$. On the other hand, points outside $IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$ are significantly worse than the optimum found.

Another methodology to measure the uncertainty of the optimum found has been developed by Cahya (2002), but he works with the classical index $D^C(\mathbf{x})$, the linear regression model and applies the Delta theorem to transformations of the regression parameters instead of transformations of the expected predicted responses. Moreover, his goal is to build a confidence region for the true optimum $\mathbf{x}_{opt}^C$, as done in all former literature, *i.e.* a zone that contains the true optimum with a probability $1 - \alpha$. The IOZ is usually larger than the confidence region developed by Cahya as it should include the true optimum $\mathbf{x}_{opt}^C$ as well as all other points of the TOZ that have a higher expected desirability than the optimum found $\hat{\mathbf{x}}_{opt}^S$.

The Cahya methodology necessitates optimising an objective function on χ for each $\mathbf{x}$ to decide if the hypothesis $H_0 : D^C(\mathbf{x}) = D^C(\mathbf{x}_{opt}^C)$ is rejected or not and, consequently, if it belongs or not to the confidence region for the true optimum. This is a quite heavy procedure when the number of factors exceeds 2 or 3. The advantage of the approach proposed in this chapter is that the IOZ is much more straightforward to calculate. In addition, the IOZ is easily derived over a slice of the experimental domain requiring no computation for the rest of the experimental domain. This is specially appealing to implement the procedure in a design of experiment software.

4.5 Prediction interval for the desirability index $D(\mathbf{x})$

As experimental responses are random variables, the result of a single experiment at a point $\mathbf{x}$ of χ can not exactly be predicted and individual values prediction intervals are usually calculated to quantify this uncertainty. In the MLR model, the prediction variance for a single observation Y_i is given by

$$\hat{V}[\hat{Y}_i|\mathbf{x}] = \hat{\sigma}_{\epsilon_i}^2 \cdot (1 + \mathbf{z}_i'(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{z}_i) \quad (4.33)$$

where $\hat{Y}_i|\mathbf{x}$ is given by (4.12).

This implies that the DI for a single observation is also an unpredictable quantity and one may be interested to derive a prediction interval for the desirability in $\mathbf{x}$, $D(\mathbf{Y})|\mathbf{x}$.

In this context, Trautmann and Weihs (2004) proposed a methodology to construct such prediction interval from the analytical distribution function of $D(\mathbf{Y})|\mathbf{x}$ when DFs are of Harrington type. From this interval, they derive a sort of optimal zone as a subset of χ around the optimum $\mathbf{x}_{opt}^S$ such that the probability of having $D(\mathbf{Y})|\mathbf{x}$ lying in a prediction interval around $D^S(\mathbf{x}_{opt}^S)$ is high. In their computations, they only take into account the randomness of the model error term, ϵ , but neglect the model estimation error.

Similarly, Govaerts and Le Bailly de Tillegheem (2005a) have derived the analytical distribution of $D(\mathbf{Y})|\mathbf{x}$ for Normal *cdf* based DFs (3.4-3.5-3.6) and can derive similar results.

When no analytical form for the distribution function of the DI exists, the best approach is to derive the prediction interval through simulations. In the classical regression model, one may simulate S possible responses in $\mathbf{x}$, $Y_i|\mathbf{x}$, using the statistical distribution of $\hat{Y}_i$:

$$\frac{Y_i - \hat{Y}_i}{\sqrt{\hat{\sigma}_{\epsilon_i}^2 \cdot (1 + \mathbf{z}_i'(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{z}_i)}} \sim t_{n_i - q_i}. \quad (4.34)$$

Each simulated vector $\mathbf{Y}^s = (Y_1^s, Y_2^s, \dots, Y_p^s)$, $s = 1, 2, \dots, S$, will allow to calculate one possible DI $D(\mathbf{Y}^s)|\mathbf{x}$ and the prediction interval will be derived from the quantiles of the empirical distribution of $(D(\mathbf{Y}^1)|\mathbf{x}, \dots, D(\mathbf{Y}^S)|\mathbf{x})$.

4.6 Extension to correlated models residuals

The results presented in this chapter are based on the assumption that, conditionally on $\mathbf{x}$, the responses $(Y_1, Y_2, \dots, Y_p)$ are independent. This hypothesis is often realistic and always verified if models are established on the basis of independent sets of experiments.

Moreover, if such conditional dependence exists between two responses, the responses themselves will often be very dependent and one may question the interest to include both responses in the optimisation exercise.

Nevertheless, in all cases, the independence assumption must be validated because, if it is not encountered, the estimator of the expected desirability (4.11) is biased and the confidence intervals (4.19) and the test statistic (4.32) for the construction of the indistinguishable optimal zone (4.22) are unreliable. The dependence between the vectors of residuals of the different models can be studied by a scatterplot matrix or the correlation coefficients. If those tools demonstrate a dependence between the models residuals, then the next extension to correlated models residuals can be used.

Let us assume a global modeling of the p responses by

$$\mathbf{Y} = \mathbf{f}(\mathbf{x}, \boldsymbol{\beta}) + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathbf{N}_p(\mathbf{0}, \boldsymbol{\Sigma}), \quad (4.35)$$

where $\mathbf{Y} = (Y_1, Y_2, \dots, Y_p)'$, $\mathbf{f}(\mathbf{x}, \boldsymbol{\beta}) = (f_1(\mathbf{x}, \boldsymbol{\beta}_1), f_2(\mathbf{x}, \boldsymbol{\beta}_2), \dots, f_p(\mathbf{x}, \boldsymbol{\beta}_p))'$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_p)$ and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the error terms $\boldsymbol{\epsilon}$. If the conditional independence is not assumed any more, the variance-covariance matrix $\boldsymbol{\Sigma}$ of dimension $(p \times p)$ is not diagonal any more. It is also supposed that the same experiments are performed to set up all individual models in order to estimate $\boldsymbol{\Sigma}$. Under these notations, the main results of this chapter can be generalised as follows.

Applying the approximation of the expectation of a differentiable function of random variables (4.5) to the expected desirability index with the model (4.35) provides the following approximation (proof in Appendix 2)

$$D^S(\mathbf{x}) \approx D^C(\mathbf{x}) \cdot (1 + B(\mathbf{x})) \quad (4.36)$$

where the bias term $B(\mathbf{x})$ is defined in matrix notations differently than in the independence case (4.9) by $B(\mathbf{x}) = \frac{1}{2} \mathbf{1}_p(\mathbf{C}(\mathbf{x}) \bullet \boldsymbol{\Sigma}) \mathbf{1}_p'$, with $\bullet$ denoting the Hadamard product of two matrices (element by element). $\mathbf{1}_p$ is the p -vector of 1 and the $(p \times p)$ matrix $\mathbf{C}(\mathbf{x})$, derived from the Hessian matrix of $D(\mathbf{Y})|\mathbf{x}$, is defined as

$$\mathbf{C}(\mathbf{x}) = \begin{pmatrix} c_{11}(\mathbf{x}) & c_{12}(\mathbf{x}) & \cdots & c_{1p}(\mathbf{x}) \\ c_{21}(\mathbf{x}) & c_{22}(\mathbf{x}) & \cdots & c_{2p}(\mathbf{x}) \\ \vdots & \vdots & \cdots & \vdots \\ c_{p1}(\mathbf{x}) & c_{p2}(\mathbf{x}) & \cdots & c_{pp}(\mathbf{x}) \end{pmatrix} \quad (4.37)$$

with the constants $c_i(\mathbf{x})$ computed as (4.10) and $c_{ij}(\mathbf{x})$ as

$$c_{ij}(\mathbf{x}) = w_i w_j \frac{\frac{\partial d_i(Y_i)}{\partial Y_i}}{d_i(E[Y_i|\mathbf{x}])} \frac{\frac{\partial d_j(Y_j)}{\partial Y_j}}{d_j(E[Y_j|\mathbf{x}])}. \quad (4.38)$$

An estimator for $D^S(\mathbf{x})$ in the case of correlated models residuals can be obtained by replacing the unknown quantities in (4.36) by their usual estimators:

$$\hat{D}^S(\mathbf{x}) = \hat{D}^C(\mathbf{x}) \cdot (1 + \hat{B}(\mathbf{x})) \quad (4.39)$$

Thanks to the similar form of this estimator (4.39), the construction of a confidence interval and the indistinguishable optimal zone can be conducted like in the independence case. The Delta method theorem can also be applied in the case of correlated models residuals and a confidence interval for $\text{logit}(D^S(\mathbf{x}))$ can still be based on the asymptotic Normal distribution of $\text{logit}(\hat{D}^S(\mathbf{x}))$ (4.13). The estimation of $V[\hat{D}^C(\mathbf{x})]$ (4.16) necessary to estimate $V[\text{logit}(\hat{D}^S(\mathbf{x}))]$ is similar to the independence case:

$$\hat{V}[\hat{D}^C(\mathbf{x})] = \hat{\mathbf{a}}(\mathbf{x})' \hat{V}(\hat{E}[Y|\mathbf{x}]) \hat{\mathbf{a}}(\mathbf{x}) \quad \text{where} \quad (4.40)$$

$\hat{\mathbf{a}}(\mathbf{x}) = (\hat{a}_1(\mathbf{x}), \hat{a}_2(\mathbf{x}), \dots, \hat{a}_p(\mathbf{x}))'$ is a p -vector of constants defined in (4.17) resulting from the gradient of $D(\mathbf{Y})|\mathbf{x}$. The estimated variance-covariance matrix $\hat{V}(\hat{E}[\mathbf{Y}|\mathbf{x}])$ of dimension $(p \times p)$ equals $\mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}\hat{\Sigma}$ in the case of linear regression models if it is assumed that the same set of factors and transformed factors are entered into all individual models.

Finally the same test statistics (4.32) is used for the construction of the IOZ. It is still Normally distributed. The components of the variance $V[\hat{\Delta}_D(\mathbf{x}, \hat{\mathbf{x}}_{opt}^S)]$ defined in (4.28) are estimated like in equation (4.29) with $Cov(\text{logit}(\hat{D}^S(\mathbf{x})), \text{logit}(\hat{D}^S(\hat{\mathbf{x}}_{opt}^S)))$ depending on

$$\hat{Cov}(\hat{D}^C(\mathbf{x}), \hat{D}^C(\hat{\mathbf{x}}_{opt}^S)) \approx \hat{\mathbf{a}}(\mathbf{x})' \hat{Cov}(\hat{E}[Y|\mathbf{x}], \hat{E}[Y|\mathbf{x}_{opt}^S]) \hat{\mathbf{a}}(\mathbf{x}_{opt}^S). \quad (4.41)$$

The estimated variance-covariance matrix $\hat{Cov}(\hat{E}[Y|\mathbf{x}], \hat{E}[Y|\mathbf{x}_{opt}^S])$ of dimension $(p \times p)$ equals $\mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_{opt}^S\hat{\Sigma}$ in the case of linear regression models if the same set of factors and transformed factors are included in all individual models.

4.7 Illustrative example

Simulations were performed on a simple example in order to illustrate the concepts introduced in this chapter and show that the Delta method based confidence interval formula and the indistinguishable optimal zone give accurate results. A more complex and more realistic application is presented on the Derringer and Suich data in the Section 8.

4.7.1 Example description

Two responses, Y_1 and Y_2 , are simultaneously optimised on a one factor domain, $x \in [0, 1]$. The first response, Y_1 , has to be maximised and the second, Y_2 , minimised. Two third order models are used to simulate data. As shown in Figure 4.2, these models are not

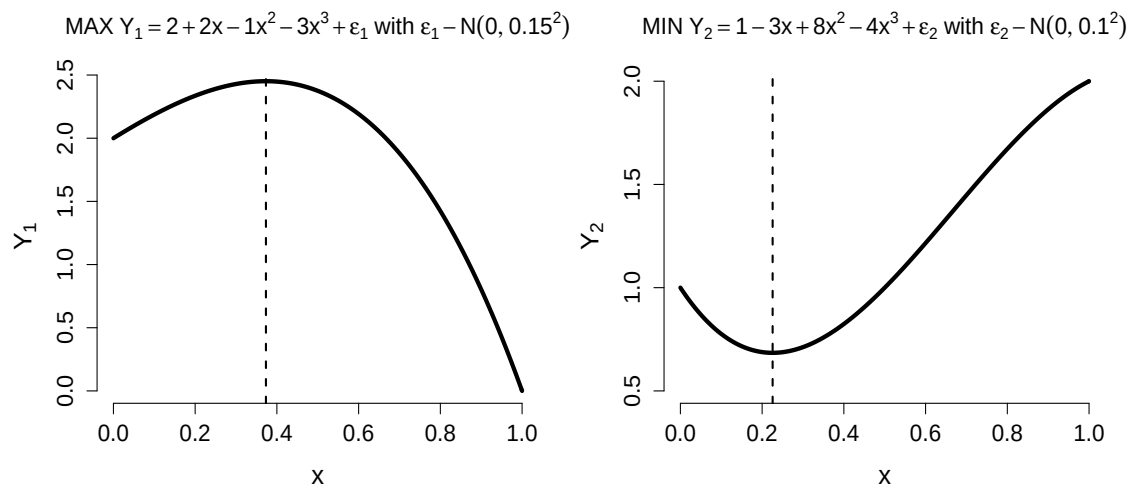


Figure 4.2: Two responses, Y_1 and Y_2 , to be respectively maximised and minimised on $[0, 1]$. The respective optima are marked by vertical dashed lines and are 0.373 for Y_1 and 0.226 for Y_2

too simple and their respective optima are not too competitive (0.373 and 0.226).

The Normal *cdf* is used to construct DFs for Y_1 and Y_2 (3.4-3.5-3.6). It depends on parameters a and b fixed according to the acceptable responses values. The first response has to be maximised and we suppose that all values smaller than 0 are not at all desired and all values greater than 3 are equivalently hugely desired. Those lower and upper limits, $LL_1 = 0$ and $UL_1 = 3$, are used to compute the two parameters of the DF as $a_1 = (LL_1 + UL_1)/2 = 1.5$ and $b_1 = (UL_1 - a_1)/2 = 0.75$. In this way, accordingly to the properties of the standard Normal *c.d.f.*, $d_1(0) = 0.023$ and $d_1(3) = 0.977$ and between 0 and 3, the desirability increases. This DF is represented in Figure 4.3 as a function of Y_1 .

The second response has to be minimised. Let us suppose that all values greater than $UL_2 = 2$ are not at all desired and all values smaller than $LL_2 = 0.5$ are equivalently hugely desired. Those limits are used to compute the two parameters of the DF, $a_2 = (LL_2 + UL_2)/2 = 1.25$ and $b_2 = (UL_2 - a_2)/2 = 0.375$. In this way, $d_2(2) = 0.023$ and $d_2(0.5) = 0.977$ and between 0.5 and 2, the desirability decreases. This DF is represented in Figure 4.3 as a function of Y_2 . Both DFs are also represented in Figure 4.4 as a function of x and, over the $[0, 1]$ interval, the respective optima are 0.373 for Y_1 and 0.226 for Y_2 .

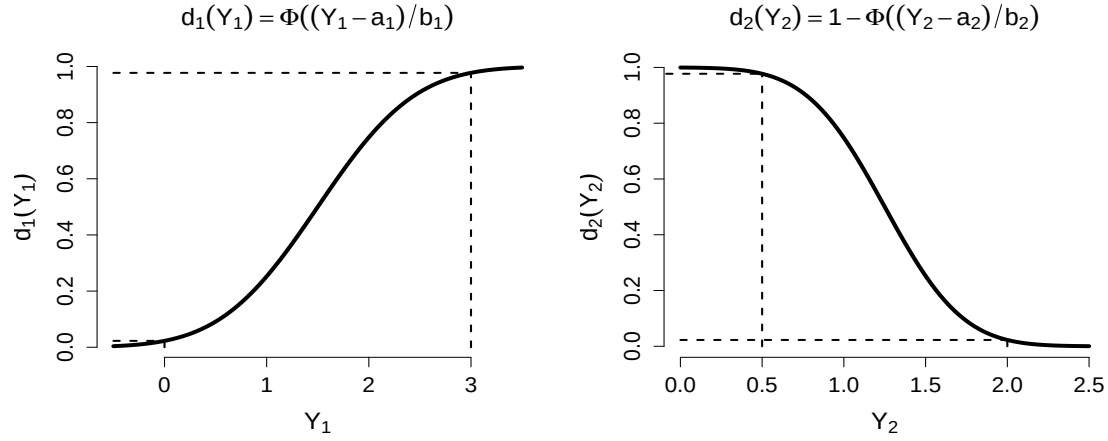


Figure 4.3: Desirability functions for Y_1 and Y_2 based on the Normal *cdf*, $d_1(Y_1)$ and $d_2(Y_2)$, with $a_1 = 1.5$, $b_1 = 0.75$, $a_2 = 1.25$ and $b_2 = 0.375$.

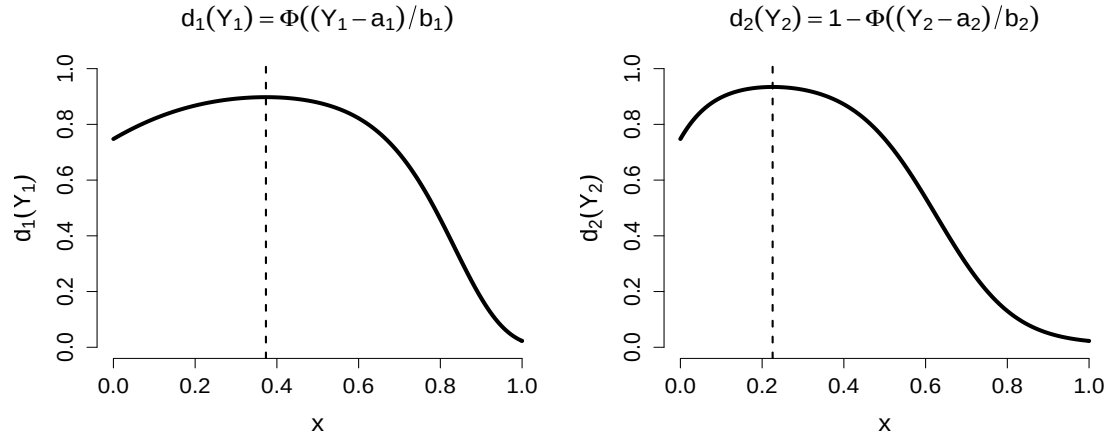


Figure 4.4: Desirability functions for Y_1 and Y_2 based on the Normal *cdf* as functions of x . The respective optima are marked by vertical dashed lines and are 0.373 for Y_1 and 0.226 for Y_2 .

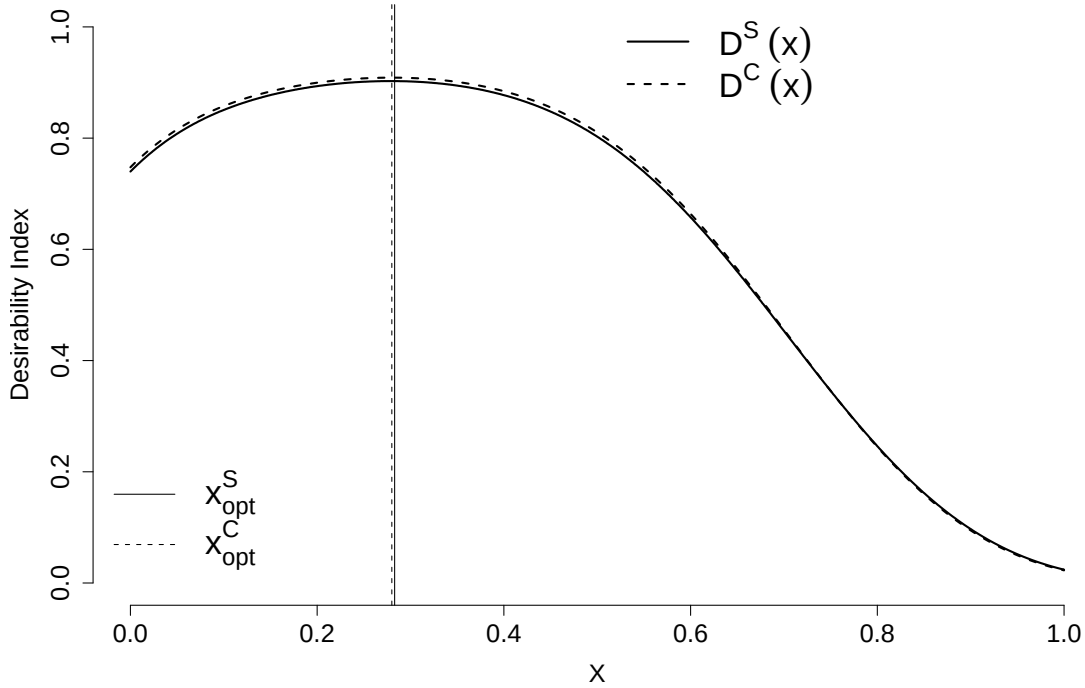


Figure 4.5: Difference between the classical use and the new use of desirability index: $D^S(x)$ and $D^C(x)$ are represented by a continuous and a dashed thick curves respectively. The respective optima are marked by vertical thin continuous and dashed lines, $x_{opt}^S = 0.283$ and $x_{opt}^C = 0.280$.

4.7.2 Comparison of classical and stochastic use of DI

The geometric mean DI is used to summarise the two DFs, giving the same weight to both properties. In this example, the difference between the classical index $D^C(x)$ and the modified one $D^S(x)$ is small as shown in Figure 4.5. The resulting theoretical optima are then very closed to each other, $x_{opt}^S = 0.283$ and $x_{opt}^C = 0.280$.

Note that, in all performed simulations, the difference between $D^C(\mathbf{x})$ and $D^S(\mathbf{x})$ is small compared to the uncertainty of the corresponding quantities $\hat{D}^C(\mathbf{x})$ and $\hat{D}^S(\mathbf{x})$. More precisely, the difference between $D^C(x)$ and $D^S(x)$ on Figure 4.5 is indeed negligible compared to the size of confidence intervals of Section 7.3 (Figure 4.8). Similarly, the difference between the two optima x_{opt}^C and x_{opt}^S is small compared to the size of the IOZ developed in Section 7.4 (Figure 4.10).

4.7.3 Confidence interval on $D^S(x)$

In order to first check the adequacy of the confidence interval for $D^S(x)$ (4.19), the next simulation was repeated 10 000 times:

Step 1: For each response, Y_1 and Y_2 , generation of a sample of size $n = 10$ (three replicates in $x = 0$ and $x = 1$ and two replicates in $x = 1/3$ and $x = 2/3$) following the models given in Figure 4.2 and fit of a cubic regression model to the simulated data (Figure 4.6).

Step 2: Estimation of $D^S(x)$ over the factor domain $\chi = [0, 1]$ with a grid of stride 0.001 using the fitted models and our estimator (4.11) (Figure 4.7).

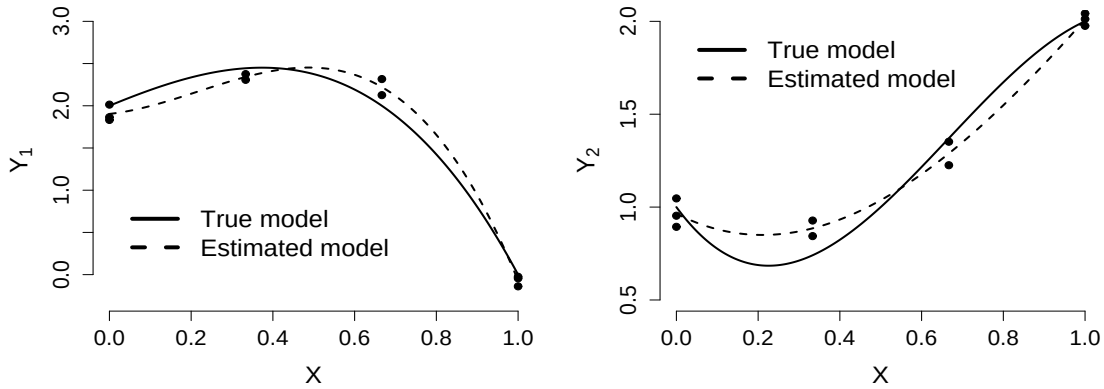


Figure 4.6: Example of the first step for one simulation run: the dots represent the generated datasample on which the models are fitted. The continuous and the dashed lines represent respectively the true and the estimated models.

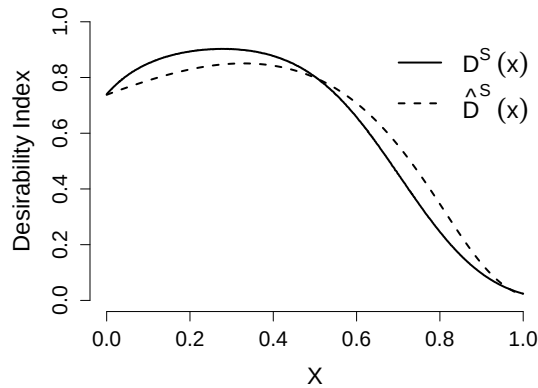


Figure 4.7: Example of the second step for one simulation run: the continuous and the dashed lines represent respectively the true and the estimated expected desirabilities.

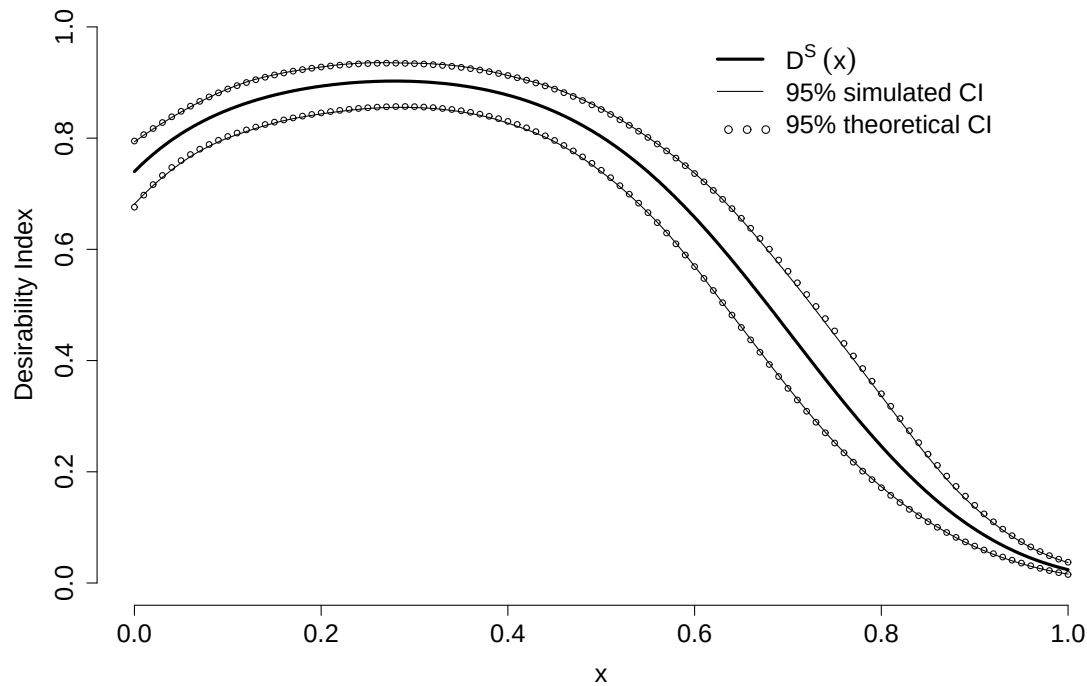


Figure 4.8: Validation of the confidence interval: comparison of the 95% theoretical confidence interval for $D^S(x)$ given by formula (4.19) using the true $D^S(x)$ - empty dots - and 95% simulated confidence interval constructed on the basis of the quantiles 2.5% and 97.5% of the simulated distribution for the estimator $\hat{D}^S(x)$ - continuous thin lines -.

From run to run, the estimated expected desirability can have very different shapes. For each x of the grid over $[0, 1]$, we have 10 000 realisations of the distribution of our estimator $\hat{D}^S(x)$. Figure 4.8 compares the 2.5% and 97.5% quantiles of this simulated distribution and the theoretical 95% confidence interval for $D^S(x)$ given by formula (4.19), using the true $D^S(x)$ (4.8). The theoretical formula provides confidence intervals really similar to the simulations.

We can conclude that the approximated asymptotic distribution given by the Delta method theorem provides, in this example, quite accurate confidence intervals even if the number of data used to fit the regression models is small.

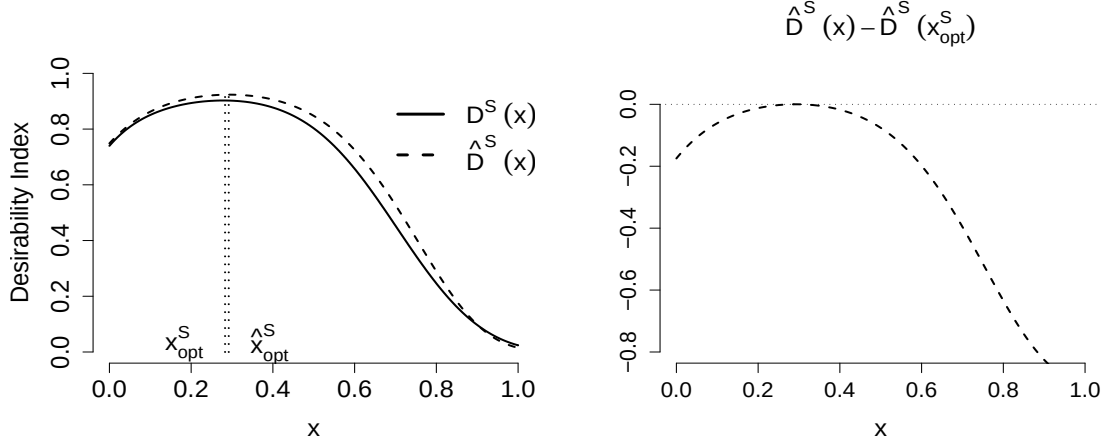


Figure 4.9: Example of the second and third steps for one simulation run: the first graph compares the true expected desirability index $D^S(x)$ (continuous curve) and the estimated one $\hat{D}^S(x)$ (dashed curve) and the second graph presents the resulting estimated difference of expected desirability between any point x in $[0, 1]$ and the optimum x_{opt}^S , $\hat{D}^S(x) - \hat{D}^S(x_{opt}^S)$.

4.7.4 Indistinguishable Optimal Zone (IOZ)

In order to check the adequacy of the Delta method based Indistinguishable Optimal Zone (4.22), the next simulation was repeated 10 000 times:

- Step 1:** For each response, Y_1 and Y_2 , generation of a sample of size $n = 10$ (three replicates in $x = 0$ and $x = 1$ and two replicates in $x = 1/3$ and $x = 2/3$) following the models given in Figure 4.2 and fit of a cubic regression model to the simulated data (Figure 4.6).
- Step 2:** Estimation for a grid of stride 0.001 over $\chi = [0, 1]$ of the desirability $\hat{D}^S(x)$ and search for the optimum $\hat{x}_{opt}^S$ (Figure 4.9).
- Step 3:** Computation of $\hat{D}^S(x) - \hat{D}^S(x_{opt}^S)$ for a grid of stride 0.001 over $\chi = [0, 1]$ (Figure 4.9).
- Step 4:** Computation of $IOZ_\alpha(\hat{\mathbf{x}}_{opt}^S)$ (with $\alpha = 0.01$ or 0.05) and $TOZ(\hat{\mathbf{x}}_{opt}^S)$ (Figure 4.1).

First, thanks to Step 3, the Normal distribution of the test statistics used to conduct the hypothesis test defining the IOZ is assessed. For each x of the grid over $[0, 1]$, we have 10 000 realisations of the distribution of the estimator $\hat{D}^S(x) - \hat{D}^S(x_{opt}^S)$. The simulated IOZ around the optimum x_{opt}^S is constructed by retaining points x for which the one-sided intervals determined by the 95% quantiles of this simulated distribution contain 0, meaning that, for those points x , $D^S(x)$ is not significantly smaller than $D^S(x_{opt}^S)$. This simulated

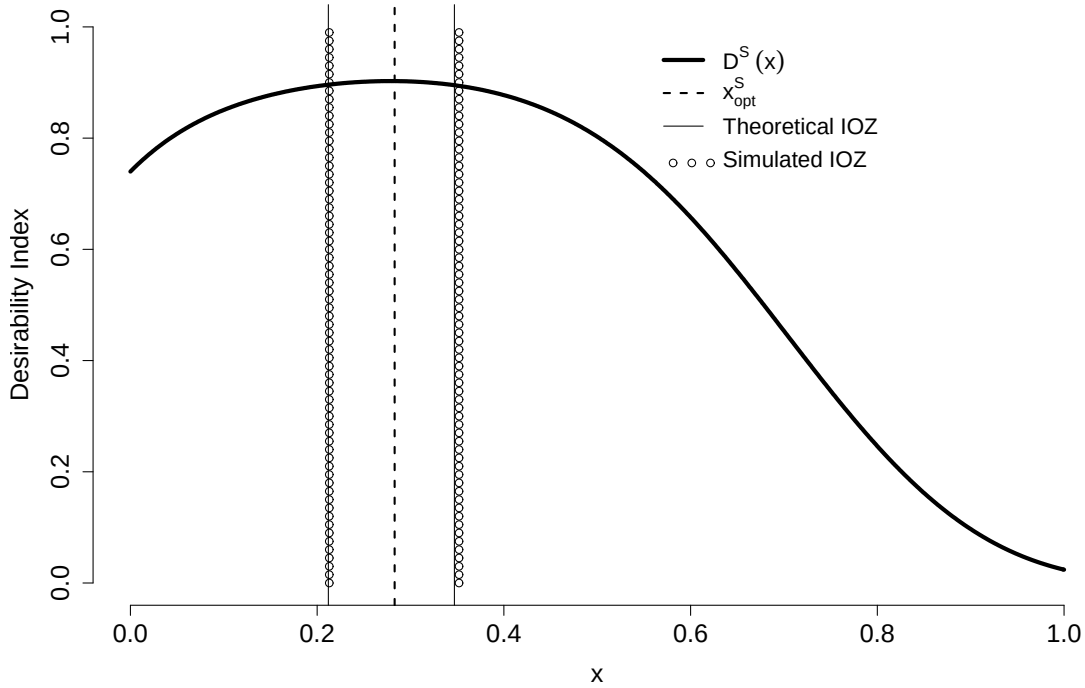


Figure 4.10: Comparison of the IOZ obtained by simulations of the distribution of estimated differences $\hat{D}^S(x) - \hat{D}^S(x_{opt}^S)$ ($[0.213, 0.352]$, continuous thin vertical line) and the IOZ obtained with the theoretical formula (4.22) ($[0.214, 0.350]$, dotted vertical line). The thick continuous curve is $D^S(x)$ and the thick dashed vertical line is the true optimum $x_{opt}^S = 0.283$.

IOZ is compared in Figure 4.10 with the theoretical IOZ around x_{opt}^S , *i.e.* as defined by (4.22) with the Normally distributed test statistics (4.32) computed with the true quantities instead of estimators. On this example, the theoretical IOZ $[0.214, 0.350]$ is nearly the same as the one obtained by simulations $[0.213, 0.352]$: all the points x in this interval are indistinguishable from the optimum $x_{opt}^S = 0.283$. Those simulations confirms that the test statistics and its asymptotic Normal distribution are appropriate to construct the IOZ.

The Step 4 of the simulations allows to check if the indistinguishable optimal zone (computed without multiplicity correction or with one of the corrections proposed in Section 4.3) adequately covers the true optimal zone and test the adequacy of the different multiplicity corrections. The type I error rate (*i.e.* $1 - \text{coverage}$) as well as the False Discovery Rate are presented in Table 4.1 for $\alpha = q = 0.01$ or 0.05 .

α	Type I error		FDR	
	0.01	0.05	0.01	0.05
No correction	0.0006	0.0112	0.0004	0.0105
Bonferroni	0	0	0	0
FDR indep	0.0003	0.0077	0.0001	0.0073
FDR dep	0	0	0	0.0003

Table 4.1: Coverage study of the IOZ.

On this example, no multiplicity correction sounds necessary. Most of the simulated IOZs cover entirely the TOZs. But if the desirability curve becomes flatter or stiffer the type I error rate and the FDR both increase and a multiplicity correction is required.

4.8 Application to Derringer and Suich data and simulation study

The objective of this section is to illustrate the relevance of the confidence interval for the desirability and the definition of the IOZ on a realistic application with several factors and optimised responses. The methodology presented in this chapter is applied on a well-known multicriteria optimisation problem derived from the paper of Derringer and Suich (1980) and a simulation study is performed to analyse the behaviour of the methodology on variations of this example.

4.8.1 Derringer and Suich classical optimisation

In their introductory article on multicriteria optimisation, Derringer and Suich were interested in optimising the quality of a tire tread compound. It is characterised by four properties submitted to given quality constraints:

Y_1 = PICO abrasion index	one-sided (max)	$120 < Y_1$
Y_2 = 200% modulus	one-sided (max)	$1000 < Y_2$
Y_3 = elongation at break	two-sided (target)	$400 < Y_3 < 600$
Y_4 = hardness	two-sided (target)	$60 < Y_4 < 75$.

Three ingredients can be varied independently: x_1 , the hydrated silica level; x_2 , the silane coupling agent level; and x_3 , the sulfur level. The three factors are standardised according to their ranges of interest: $z_1 = (x_1 - 1.2)/0.5$, $z_2 = (x_2 - 50)/10$ and $z_3 = (x_3 - 2.3)/0.5$. A three-variable, rotatable, central composite design with six center points was carried out (see Figure 4.11). For each of those 20 experiments, the four responses are measured and reported in Table 4.2.

For each response, a second degree polynomials regression model was fitted to the collected data using least squares. The estimated models $\hat{Y}_i = b_0^i + \sum_{k=1}^3 b_k^i z_k + \sum_{k=1}^3 \sum_{l=k}^3 b_{kl}^i z_k z_l$ are summarised in Table 4.3. Some parameters are not significant at a level 0.05, especially in the second model.

The objective of Derringer and Suich was to optimise simultaneously the four responses over the experimental domain $[-1, 1]^3$. They defined their own DFs (see Figure 4.12) and aggregated them according to the classical desirability concept introduced by Harrington (1965). The optimal factors settings were found by a grid search as the point $\mathbf{x}$ that maximises $\hat{D}^C(\mathbf{x})$: $\hat{\mathbf{x}}_{opt}^C = (-0.05, 0.145, -0.868)$. In this classical optimisation example, nothing is said about the reliability of the fitted models and prediction error.

In the following, the second property Y_2 has been dropped from the optimisation because of a poor quality model (adjusted $R^2 = 0.51$, compared to 0.95, 0.96 and 0.92 for the three other models and a p -value for the F -test of 0.0422 compared to order 10^{-6} , 10^{-7} and 10^{-5} for the three other models). The white noise of the second response, with an estimated standard deviation of 329, propagates on the corresponding DF (see Figure 4.12) in such a way that the DF could range from 0 to 1. Other quality statistics are presented in Table 4.4.

To simplify the graphical presentation of the optimisation results and allow faster simulations, the third factor is fixed at its optimal value found by Derringer and Suich, $x_3 = -0.868$.

The methodology proposed in this chapter is applied to optimise the three responses Y_1 , Y_3 and Y_4 as a function of the two factors x_1 and x_2 and take into account the propagation of the uncertainty of the regression models developed by Derringer and Suich on the DI (Section 8.2). Simulations, similar to the previous sections, are also performed to analyse the behaviour of the proposed methodology on this simplified Derringer and Suich problem (Section 8.3) and when varying the residual variances, the sample size, the DF shapes as well as the weights used in the DI (Section 8.4).

4.8.2 Stochastic optimisation of the simplified Derringer and Suich problem

In order to apply the stochastic optimisation procedure, some applicability conditions have to be validated. First, we validate the underlying assumptions about the error term in all regression models: $\epsilon \sim N(0, \sigma_\epsilon)$. The graphs of the standardised residuals *versus* the predicted values and the qqplots of the residuals are presented on Figure 4.13 for each model. No graph shows a strange pattern, except the qqplot for the fourth response where some extreme negative and positive observed residuals are highlighted. Nevertheless, there

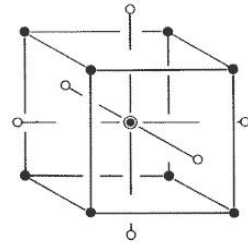


Figure 4.11: Central composite design with six center points used in the paper of Derringer and Suich (1980) to collect data.

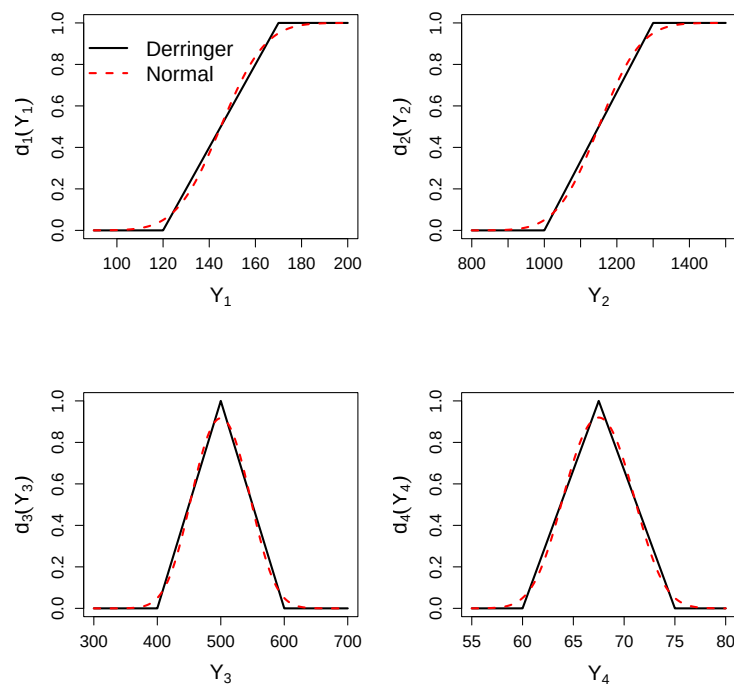


Figure 4.12: Desirability functions of Derringer and Suich (1980) or Normal *cdf* based for the three optimised responses (see Table 3.1).

n	z_1	z_2	z_3	Y_1	Y_2	Y_3	Y_4
1	-1	-1	+1	102	900	470	67.5
2	1	-1	-1	120	860	410	65
3	-1	+1	-1	117	800	570	77.5
4	1	+1	+1	198	2294	240	74.5
5	-1	-1	-1	103	490	640	62.5
6	1	-1	+1	132	1289	270	67
7	-1	+1	+1	132	1270	410	78
8	1	+1	-1	139	1090	380	70
9	-1.633	0	0	102	770	590	76
10	1.633	0	0	154	1690	260	70
11	0	-1.633	0	96	700	520	63
12	0	1.633	0	163	1540	380	75
13	0	0	-1.633	116	2184	520	65
14	0	0	1.633	153	1784	290	71
15	0	0	0	133	1300	380	70
16	0	0	0	133	1300	380	68.5
17	0	0	0	140	1145	430	68
18	0	0	0	142	1090	430	68
19	0	0	0	145	1260	390	69
20	0	0	0	142	1344	390	70

Table 4.2: 20 experiments dictated by a central composite design with six center points from the paper of Derringer and Suich (1980)

	b_0	b_1	b_2	b_3	b_{11}	b_{22}	b_{33}	b_{12}	b_{13}	b_{23}
Y_1	139.2***	16.5***	17.9***	10.9***	-4.0*	-3.4*	-1.6	5.1*	7.1**	7.9**
Y_2	1261.1***	268.1*	246.5*	139.5	-83.5	-124.7	199.1	69.4	94.1	104.4
Y_3	400.4***	-99.7***	-31.4***	-73.9***	7.9	17.3*	0.4	8.7	6.2	1.2
Y_4	68.9***	-1.4**1	4.3***	1.6***	1.6**	0.1	-0.3	-1.6**	0.1	-0.2

Table 4.3: Equations of the fitted models. Significant levels: *** < 0.001, ** < 0.01 and * < 0.05

	σ_ϵ	R^2	adjusted R^2	F -statistic	p-value
Y_1	5.61	0.97	0.95	38.65	$1.34 \cdot 10^{-6}$
Y_2	328.7	0.74	0.51	3.20	0.0422
Y_3	20.55	0.98	0.96	58.94	$1.75 \cdot 10^{-7}$
Y_4	1.27	0.96	0.92	25.39	$9.77 \cdot 10^{-6}$

Table 4.4: Analysis of the quality of the four fitted regression models from the paper of Derringer and Suich (1980)

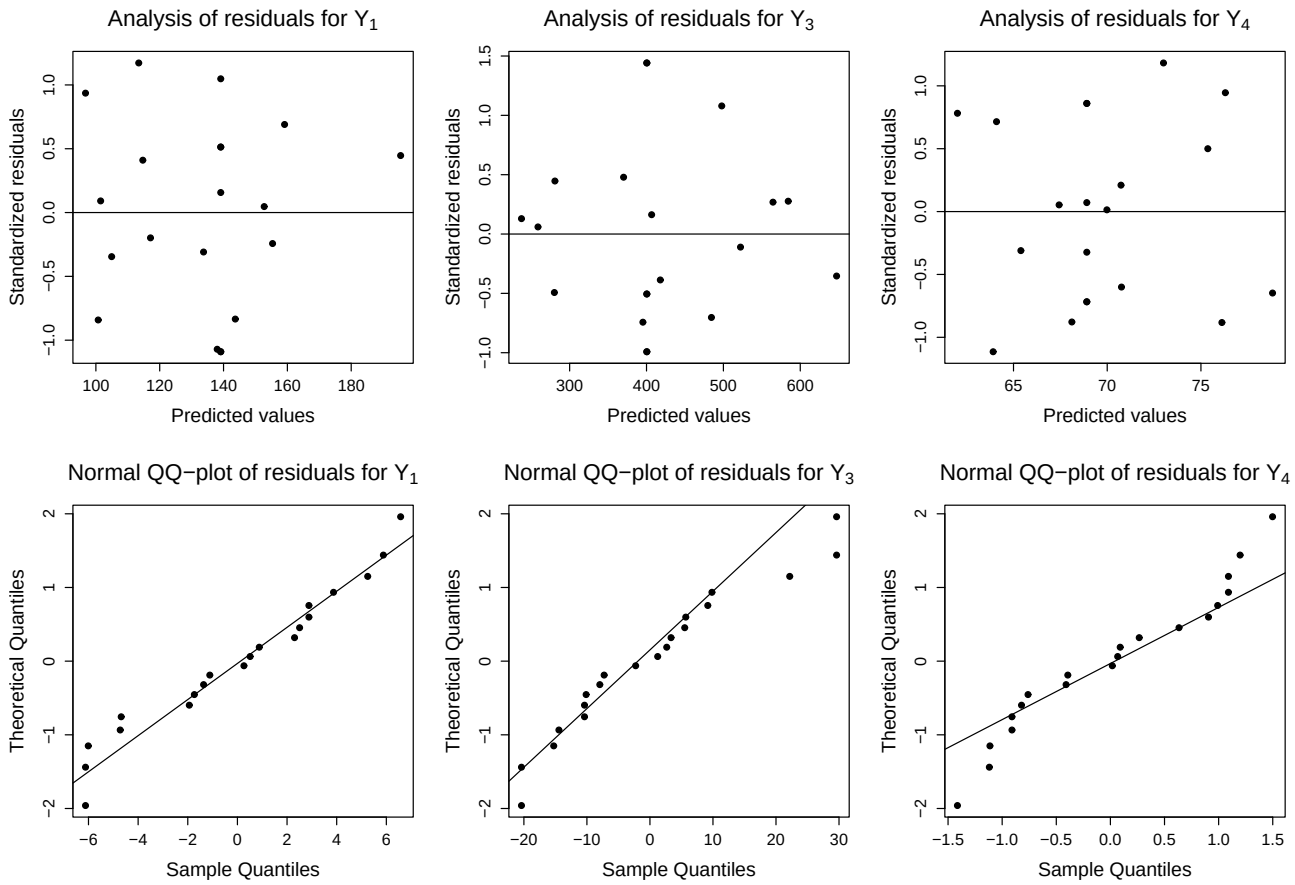


Figure 4.13: Residuals analysis for the polynomial regression models of Derringer and Suich (1980). The first line shows the standardised residuals *versus* predicted values for each of the three models. The second line shows the Normal qq-plot of residuals for each of the three models.

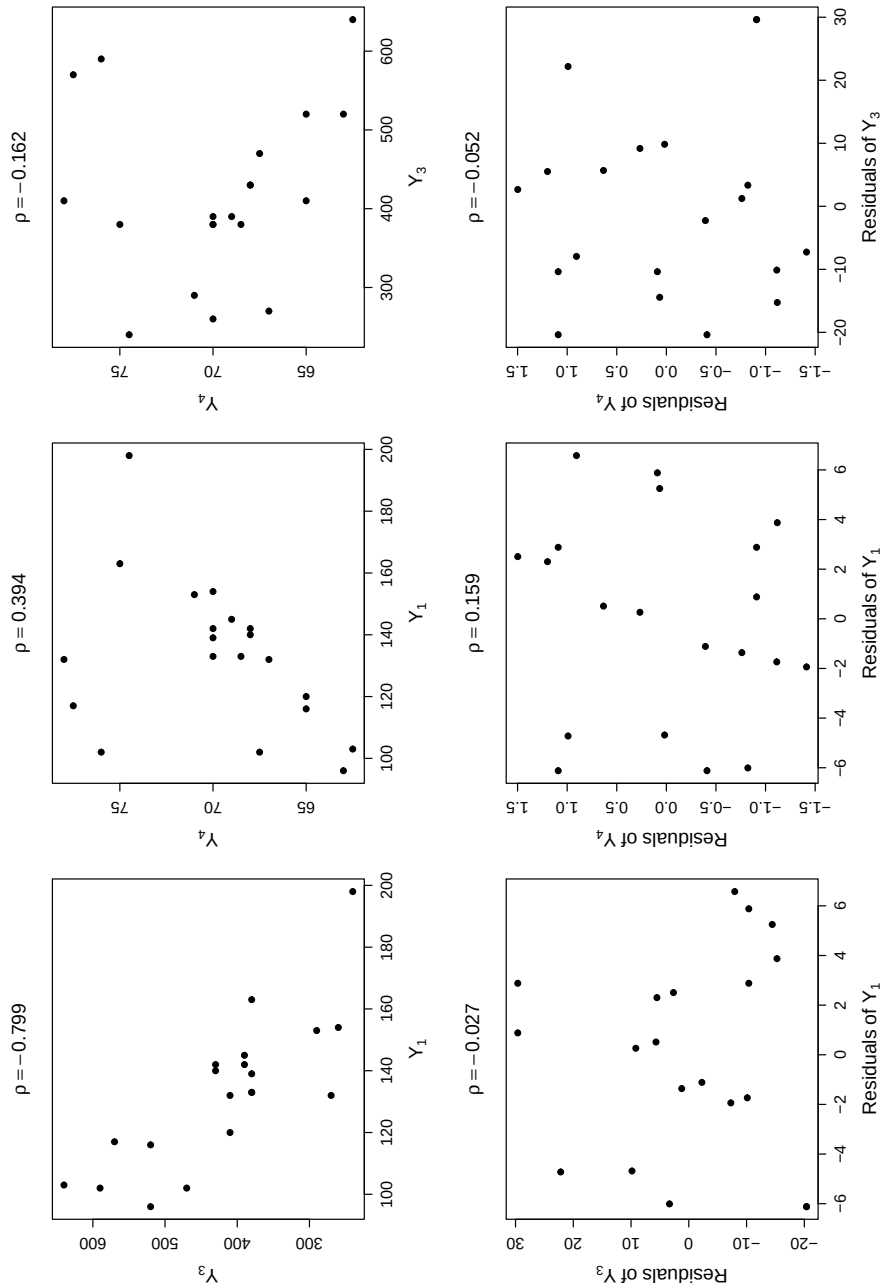


Figure 4.14: Independence analysis for the polynomial regression models of Derringer and Suich (1980). The first line shows the observed values for Y_1 vs Y_3 , for Y_1 vs Y_4 and for Y_3 vs Y_4 to analyse the independence of the responses. The second line shows the observed residuals for Y_1 vs Y_3 , for Y_1 vs Y_4 and for Y_3 vs Y_4 to analyse the conditional independence of the responses.

is only 20 observations and those points do not correspond to outliers.

To estimate $E[D(\mathbf{Y})|\mathbf{x}]$ by (4.11), the assumption of independence between the residuals of responses pairs is validated. Even if the first and the third observed responses are significantly strongly correlated ($\rho = -0.799$), the conditional independence is encountered with no correlation between residuals higher than 0.159, which is not significant. Scatterplots of responses pairs and residuals pairs are presented on Figure 4.14. The observations and the fitted regression models provide also estimates for $E[Y_i|\mathbf{x}]$ and $V[Y_i|\mathbf{x}]$ that are necessary for estimating $E[D(\mathbf{Y})|\mathbf{x}]$.

Finally, the stochastic optimisation requires that the DFs are two times continuously differentiable. The DFs proposed by Derringer and Suich are approximated by differentiable DFs based on the Normal distribution function (3.4-3.5-3.6). These approximated DFs are represented on Figure 4.12.

Using a grid search algorithm to maximise $\hat{D}^S(\mathbf{x})$ over $[-1, 1]^2$, the optimum found is $\hat{\mathbf{x}}_{opt}^S = (-0.13, 0.40)$ corresponding to an estimated desirability of $\hat{D}^S(\hat{\mathbf{x}}_{opt}^S) = 0.4822$. With the classical optimisation methodology, the optimum found would have been really similar: $\hat{\mathbf{x}}_{opt}^C = (-0.11, 0.45)$ corresponding to an estimated desirability of $\hat{D}^C(\hat{\mathbf{x}}_{opt}^C) = 0.5100$. This result is different from the optimum found in the paper of Derringer and Suich, $(-0.05, 0.145)$, as Y_2 is not optimised and the DFs are slightly different. The contour plot of $\hat{D}^C(\mathbf{x})$ and $\hat{D}^S(\mathbf{x})$ as a function of the two factors are presented on Figure 4.15.

Formula (4.19) was used to construct a confidence interval for $D^S(\hat{\mathbf{x}}_{opt}^S)$ based on the regression estimator (4.20) of $V[\hat{E}[Y_i|\hat{\mathbf{x}}_{opt}^S]]$: $IC_{90\%}(D^S(\hat{\mathbf{x}}_{opt}^S)) = [0.4040, 0.5547]$. The validity of this confidence interval is analysed by simulations in the next subsection (see Figure 4.19).

Using the available theory on regression model, (4.20) and (4.31) estimate the variance of estimated responses $V[\hat{E}[Y_i|\mathbf{x}]]$ and the covariance between the estimated responses of two design points $Cov(\hat{E}[Y_i|\mathbf{x}_1], \hat{E}[Y_i|\mathbf{x}_2])$ for the three fitted models. Those estimations were used to construct the IOZ around the optimum found (4.22). For a grid of stride 0.1 over $[-1, 1]^2$, the hypothesis test (4.23-4.24) was conducted without applying a multiplicity correction or with a multiplicity correction (Bonferroni (1936), Benjamini and Hochberg (1995) or Benjamini and Yekutieli (2001)) with a significance level of $\alpha = 0.1$. The corresponding IOZ's are represented on Figure 4.19 in the next subsection. Among the 441 tested points, 27 are included in the IOZ without multiplicity correction and 29, 76 and 142 are in the IOZ with the multiplicity correction of Benjamini and Hochberg (1995), of Benjamini and Yekutieli (2001) and of Bonferroni (1936), respectively. The adequacy of the IOZ's is studied by simulations in the next subsection.

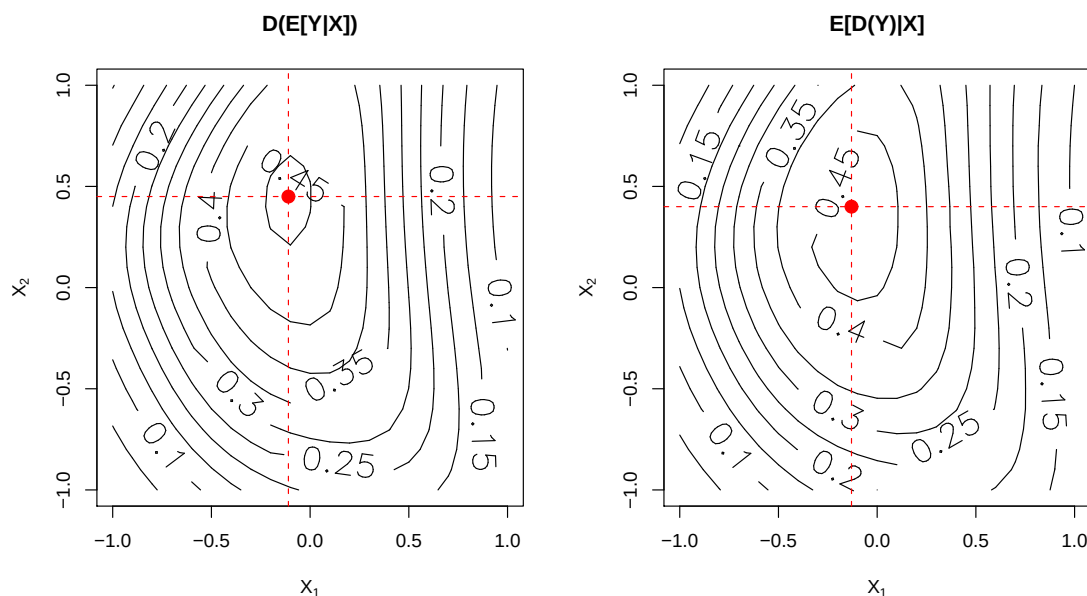


Figure 4.15: Left: Contour plot of $\hat{D}^C(\mathbf{x})$ as a function of the two factors x_1 and x_2 with $\hat{\mathbf{x}}_{opt}^C = (-0.11, 0.45)$ corresponding to $\hat{D}^C(\hat{\mathbf{x}}_{opt}^C) = 0.5100$. Right: Contour plot of $\hat{D}^S(\mathbf{x})$ as a function of the two factors x_1 and x_2 with $\hat{\mathbf{x}}_{opt}^S = (-0.13, 0.40)$ corresponding to $\hat{D}^S(\hat{\mathbf{x}}_{opt}^S) = 0.4822$

4.8.3 Simulations on the simplified Derringer and Suich problem

Objectives and principles of the simulations

Some simulations were performed to analyse the quality of the proposed estimators $\hat{D}^S(\mathbf{x})$ and $\hat{V}(\hat{D}^S(\mathbf{x}))$, the (Normal?) distribution of $\hat{D}^S(\mathbf{x})$, $\log(\hat{D}^S(\mathbf{x}))$ or $\text{logit}(\hat{D}^S(\mathbf{x}))$, as well as the difference of two estimated desirabilities, the coverage of the proposed confidence interval for $D^S(\mathbf{x})$ and the coverage of the TOZ by the IOZ. The estimated models of Table 4.3 were assumed to be the real ones and the next simulation was repeated 10 000 times:

Step 1: For each response, Y_1 , Y_3 and Y_4 , generation of a sample of size $n = 20$ corresponding to a two-variable central composite design (2 replicates at each point and 4 replicates in the center).

Step 2: Estimation for a grid of stride 0.1 over $\chi = [-1, 1]^2$ of the desirability $D^S(\mathbf{x})$ and its variance $V(\hat{D}^S(\mathbf{x}))$, as well as the corresponding 90% confidence interval for $D^S(\mathbf{x})$.

Step 3: Selection of the optimum $\hat{\mathbf{x}}_{opt}^S$ as the grid-point $\mathbf{x}$ that maximises $\hat{D}^S(\mathbf{x})$.

Step 4: Computation of $TOZ(\hat{\mathbf{x}}_{opt}^S)$ and construction of $IOZ_{\alpha=0.1}(\hat{\mathbf{x}}_{opt}^S)$ without multiplicity correction and with the multiplicity correction of Benjamini and Yekutieli (2001).

The main results of those simulations are presented below.

Bias of $\hat{D}^S(\mathbf{x})$ and $\hat{V}[\hat{D}^S(\mathbf{x})]$

For each of the 441 grid-points $\mathbf{x}$, the bias of the proposed estimator (4.11) for the expected desirability $D^S(\mathbf{x})$ can be estimated by

$$Bias(\hat{D}^S(\mathbf{x})) = \overline{\hat{D}^S(\mathbf{x})} - D^S(\mathbf{x}) \quad (4.42)$$

where $\overline{\hat{D}^S(\mathbf{x})}$ is the average of the 10 000 estimations of $D^S(\mathbf{x})$ (Step 2). The left graph of Figure 4.16 representing $\overline{\hat{D}^S(\mathbf{x})}$ as a function of $D^S(\mathbf{x})$ illustrates the adequacy of the proposed estimator as, for most points $\mathbf{x}$ in $[-1, 1]^2$, $\overline{\hat{D}^S(\mathbf{x})}$ nearly equals $D^S(\mathbf{x})$.

In order to summarise the bias of the proposed estimator by one value, those biases (4.42) can then be averaged over the grid-points $\mathbf{x}$: $Bias(\hat{D}^S(\mathbf{x})) = -0.0015$. This averaged bias is small relatively to the value of the desirabilities ($D^S(\mathbf{x})$ is, on average, equal to 0.27 over $[-1, 1]^2$) and compared to the variability of the estimator quantified by its averaged standard error over $[-1, 1]^2$, $\sqrt{V[\hat{D}^S(\mathbf{x})]} = 0.0428$. One can conclude that, on the Derringer and Suich data, the bias of the estimator $\hat{D}^S(\mathbf{x})$ induced by the second order approximation of $D^S(\mathbf{x})$ is negligible.

A similar analysis can be performed to measure the bias of the proposed estimator (4.15) of the standard error $\sqrt{V[\hat{D}^S(\mathbf{x})]}$. For each grid-point $\mathbf{x}$, let us define

$$Bias(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]}) = \sqrt{\overline{\hat{V}[\hat{D}^S(\mathbf{x})]}} - \sqrt{V[\hat{D}^S(\mathbf{x})]} \quad (4.43)$$

where the first term $\overline{\hat{V}[\hat{D}^S(\mathbf{x})]}$ is the average of the 10 000 estimations of $V[\hat{D}^S(\mathbf{x})]$ and the second term $V[\hat{D}^S(\mathbf{x})]$ is estimated by the variance of the 10 000 estimations of $D^S(\mathbf{x})$ (Step 2). The right graph of Figure 4.16 representing $\sqrt{\overline{\hat{V}[\hat{D}^S(\mathbf{x})]}}$ as a function of $\sqrt{V[\hat{D}^S(\mathbf{x})]}$ shows a small positive bias of the estimator $\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]}$.

Those biases are averaged over the grid-points $\mathbf{x}$: $Bias(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]}) = 0.0027$. This bias is not negligible but reasonable relatively to the size of the standard error $\sqrt{V[\hat{D}^S(\mathbf{x})]}$ ($\sqrt{V[\hat{D}^S(\mathbf{x})]}$ is, on average, equal to 0.0428 over $[-1, 1]^2$). One can conclude that, on the

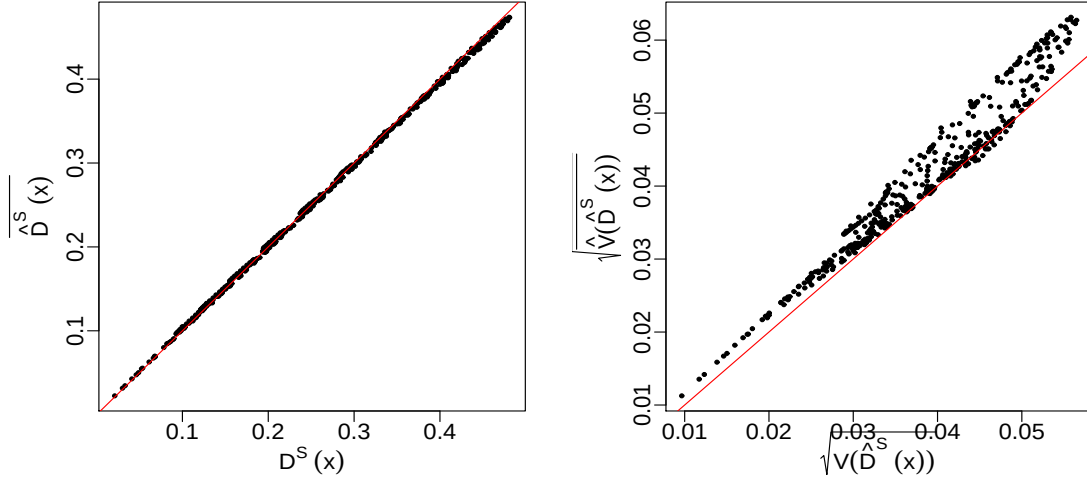


Figure 4.16: Left: $\hat{D}^S(\mathbf{x})$ as a function of $D^S(\mathbf{x})$ for each of the 441 grid-points $\mathbf{x}$ over $[-1, 1]^2$. Right: $\hat{V}[\hat{D}^S(\mathbf{x})]$ as a function of $V[\hat{D}^S(\mathbf{x})]$ for each of the 441 grid-points $\mathbf{x}$ over $[-1, 1]^2$.

Derringer and Suich data, the estimator $\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]}$ is slightly upward biased due to the second order approximation of $V[\hat{D}^S(\mathbf{x})]$.

Distribution of $\hat{D}^S(\mathbf{x})$

First, the distribution of the estimator $\hat{D}^S(\mathbf{x})$ is analysed. The relevance of a Normal distribution for $\hat{D}^S(\mathbf{x})$ is brought into question when applying the Delta method theorem for the confidence interval construction (Section 3). Kolmogorov-Smirnov tests are performed on the simulated distribution of $\hat{D}^S(\mathbf{x})$ for each grid-point $\mathbf{x}$. The grid-points $\mathbf{x}$ for which the Normal distribution of $\hat{D}^S(\mathbf{x})$ is rejected are plotted on the contour-plot of $D^S(\mathbf{x})$ (see Figure 4.17). As expected, for desirabilities near 0, the Normality assumption is rejected due to a boundary effect (distribution skewed to the right). If there were desirabilities near 1, the Normality assumption of $\hat{D}^S(\mathbf{x})$ would probably also be rejected (distribution skewed to the left).

The reliability of a Normal distribution for $\log(\hat{D}^S(\mathbf{x}))$ or $\text{logit}(\hat{D}^S(\mathbf{x}))$ is then also tested for each grid-point $\mathbf{x}$. The grid-points $\mathbf{x}$ for which the Normal distribution of $\log(\hat{D}^S(\mathbf{x}))$ and $\text{logit}(\hat{D}^S(\mathbf{x}))$ is rejected by a Kolmogorov-Smirnov test are plotted on the contour-plot of $D^S(\mathbf{x})$ (see Figure 4.17). One conclude that the Normal distribution is more plausible for $\text{logit}(\hat{D}^S(\mathbf{x}))$ than for $\hat{D}^S(\mathbf{x})$ itself or $\log(\hat{D}^S(\mathbf{x}))$. Both $\log$ and logit functions are

similar for desirabilities near 0 and both decrease the boundary effect (see Figure 4.18). For estimators $\hat{D}^S(\mathbf{x})$ possibly Normal, a logarithmic transformation is disastrous although Normality remains plausible with a logit transformation. One can expect an even worse behaviour of $\log(\hat{D}^S(\mathbf{x}))$ for desirabilities near 1 that are skewed to the left and an adequate decrease of the boundary effect for $\text{logit}(\hat{D}^S(\mathbf{x}))$.

Finally, the distributions of the differences $\hat{D}^S(\mathbf{x}) - \hat{D}^S(\mathbf{x}_{\text{opt}}^S)$, $\log(\hat{D}^S(\mathbf{x})) - \log(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))$ and $\text{logit}(\hat{D}^S(\mathbf{x})) - \text{logit}(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))$ are analysed to confirm the adequacy of the hypothesis test developed for the IOZ construction (Section 4). One conclude on the basis of Figure 4.17 that, on the Derringer and Suich data, the difference $\text{logit}(\hat{D}^S(\mathbf{x})) - \text{logit}(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))$ used in the IOZ can be assumed to follow a Normal distribution, but no transformation is really necessary as $\hat{D}^S(\mathbf{x}) - \hat{D}^S(\mathbf{x}_{\text{opt}}^S)$ is, most of the time, normally distributed.

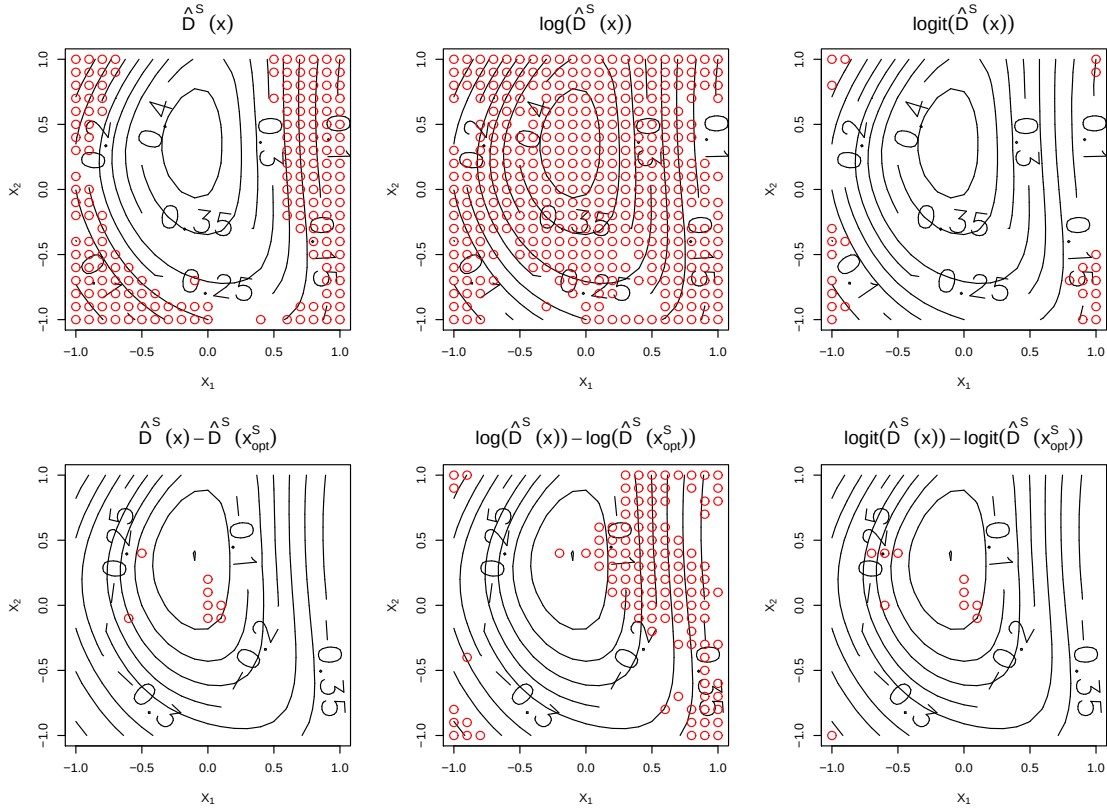
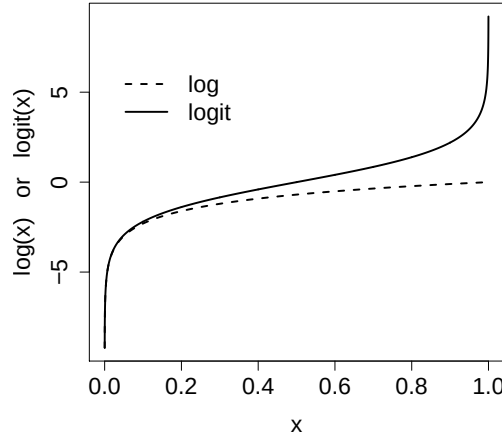


Figure 4.17: First line: contour-plot of $D^S(\mathbf{x})$ with red grid-points corresponding to Normality rejection by Kolmogorov-Smirnov tests of the simulated distribution of $\hat{D}^S(\mathbf{x})$ (left), $\log(\hat{D}^S(\mathbf{x}))$ (center) or $\text{logit}(\hat{D}^S(\mathbf{x}))$ (right). Second line: contour-plot of $D^S(\mathbf{x}) - D^S(\mathbf{x}_{\text{opt}}^S)$ with red grid-points corresponding to Normality rejection by Kolmogorov-Smirnov tests of the simulated distribution of $\hat{D}^S(\mathbf{x}) - \hat{D}^S(\mathbf{x}_{\text{opt}}^S)$ (left), $\log(\hat{D}^S(\mathbf{x})) - \log(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))$ (center) or $\text{logit}(\hat{D}^S(\mathbf{x})) - \text{logit}(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))$ (right)

Figure 4.18: $\log$ and logit transformation of x between 0 and 1.

Coverage of the confidence interval for $D^S(\mathbf{x})$

First, the adequacy of the proposed asymptotic confidence interval for $D^S(\mathbf{x})$ (4.19) is checked at the optimum $\mathbf{x}_{\text{opt}}^S = (-0.13, 0.40)$ by comparing this asymptotic confidence interval to the interval based on the quantiles of the simulated distribution of $\hat{D}^S(\mathbf{x}_{\text{opt}}^S)$. Figure 4.19 shows that the 90% asymptotic confidence interval ("theoretical CI" = [0.4040, 0.5547]) is similar to the one obtained by computing the 5% and 95% quantiles of the simulated distribution of $\hat{D}^S(\hat{\mathbf{x}}_{\text{opt}}^S)$ ("simulated CI" = [0.3973, 0.5507]). The approximation of the Delta method theorem works once more well.

The coverage rate of the confidence interval for $D^S(\mathbf{x})$ is analysed over the grid using the results of step 2 simulations. For each grid-point $\mathbf{x}$, the coverage rate of the confidence interval is estimated as the percentage among the 10 000 confidence intervals that contains the true value $D^S(\mathbf{x})$ (denoted $CR_{CI}(\mathbf{x})$). The average of those coverage rates over the grid is $\overline{CR_{CI}(\mathbf{x})} = 90.30\%$. One conclude that the coverage rate of the proposed asymptotic confidence interval for $D^S(\mathbf{x})$ is slightly higher than the expected level. The confidence interval is also slightly longer than expected: the ratio of the length of the "theoretical" confidence intervals (with formula (4.19)) averaged over $[-1, 1]^2$ and the length of the "simulated" confidence intervals (with the 5% and 95% quantiles of the simulated distribution of $\hat{D}^S(\mathbf{x})$) averaged over $[-1, 1]^2$ is 1.0501. Those results are a direct consequence of the upward bias of $\hat{V}(\hat{D}^S(\mathbf{x}))$, but the coverage rate and the length of the confidence interval for $D^S(\mathbf{x})$ are reasonably acceptable.

Coverage of the Indistinguishable Optimal Zone

The coverage rate and the false discovery rate of the IOZ are analysed using the results of Step 4 simulations, with or without multiplicity correction. For one simulation run, the coverage is computed as the percentage of grid-points in TOZ that are included in IOZ. Those coverages are averaged over all simulation runs, providing the global coverage rate displayed in Table 4.5. One can conclude that, without multiplicity correction, the TOZ is not covered at a correct level as the coverage rate does not reach 90%.

The FDR is computed for each simulation run as the percentage of grid-points not included in IOZ that are however in TOZ. Those FDRs are averaged over all simulations and the global FDR is displayed in Table 4.5. The FDR is small, especially with the multiplicity correction for dependent tests. This can be explained by the expected control level boundary $\frac{m_0}{m}q$ with $\frac{m_0}{m} = 0.02$ and $q = 0.1$ ($m = 441$ is the number of hypothesis tests and $m_0 \simeq 9$ is the average number of true null hypothesis) when applying the Benjamini and Yekutieli (2001) multiplicity correction.

The size of the IOZ is also studied. For each simulation run, the size of the IOZ is defined as the percentage of grid-points included in the IOZ and can be compared to the size of the TOZ defined as the percentage of grid-points in the TOZ. The IOZ sizes averaged over all simulations are displayed in Table 4.5. As TOZ contains on average 2.12% of the grid-points (*i.e.* 9 grid-points), the IOZ without multiplicity correction is 3 times larger than TOZ and the IOZ with the multiplicity correction for dependent tests is nearly 8 times larger than TOZ. Nevertheless, this is the price to pay to insure a good coverage of the TOZ. The IOZ with the multiplicity correction is totally coherent with the zone determined by the optima found in each Step 3 of the simulations as shown on Figure 4.19 for the "real world" case. One can conclude that the FDR correction for dependent tests statistics is necessary to correct the multiple testing issue and construct the IOZ.

	Coverage (in %)	FDR (in %)	Size (in %)
No correction	87.30	0.60	6.60
Benjamini and Yekutieli (2001)	98.44	0.12	16.67

Table 4.5: Coverage rate (percentage of grid-points in TOZ that are included in IOZ), false discovery rate (percentage of grid-points no included in IOZ that are however in TOZ) and size of the IOZ (percentage of grid-points in IOZ).

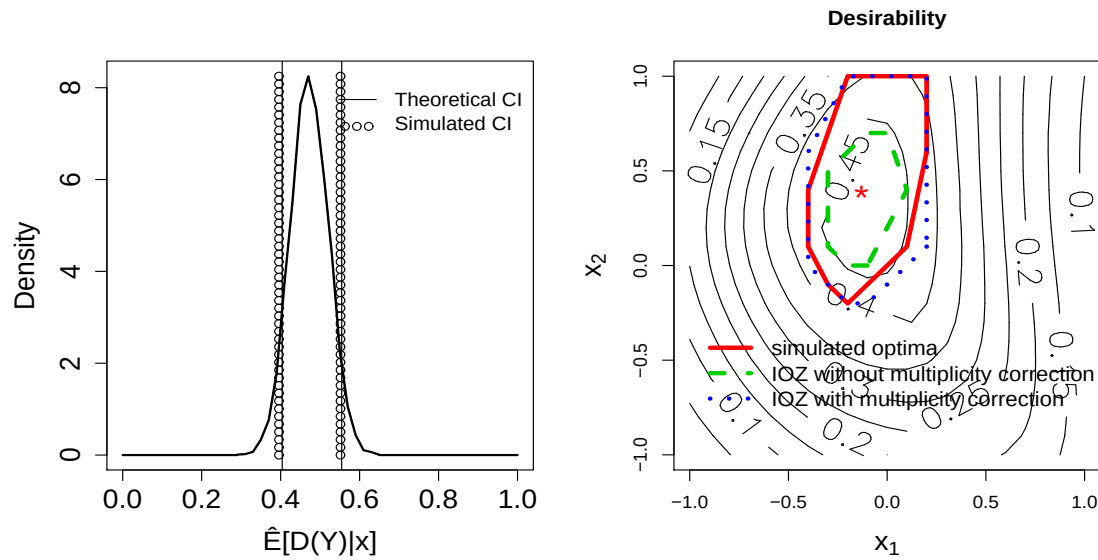


Figure 4.19: Left: simulated density of the estimator $\hat{D}^S(\mathbf{x}_{opt}^S)$. The continuous vertical lines represent the 90% confidence interval for $D^S(\mathbf{x}_{opt}^S)$ according to formula (4.19), [0.4040, 0.5547] and the dashed vertical lines represent the interval computing the 5% and 95% quantiles of the simulated distribution of $\hat{D}^S(\mathbf{x}_{opt}^S)$, [0.3973, 0.5507]. Right: IOZ without multiplicity correction or with the multiplicity correction of Benjamini and Yekutieli (2001) can be compared to the optima found over the 10 000 simulated optimisation runs (Step 3).

Conclusion of the simulations

Those simulations demonstrate a good behaviour of the Delta method on the Derringer and Suich data and the size of the IOZ show that, even with well fitted models, the propagation of the prediction error on the estimated expected desirability index is not negligible. Due to this uncertainty, a great part of the experimental domain is not significantly worse than the optimum found. If the second response Y_2 was kept, nearly all the experimental domain would have been included in the IOZ. Those simulations reveal the importance of taking prediction error into account instead of giving only a pointwise optimum as originally done by Derringer and Suich.

4.8.4 Simulations study around the simplified Derringer and Suich problem

Objectives and principles of the simulations

In order to confirm the good behaviour of the Delta method on other situations or to analyse its limits, further simulations are run around the simplified Derringer and Suich problem presented in Section 8.3 (referred to as basic situation). The proposed tools are tested when

- decreasing/increasing the model standard errors of residuals σ_{ϵ_1} , σ_{ϵ_3} and σ_{ϵ_4} (two or three times smaller than in the basic situation or two or three times higher),
- decreasing/increasing the number of available data to fit the model ($n = 10, 30$ or 100),
- shifting the DFs (subtracting/adding b or $2b$ to the parameter a of Normal *cdf* based DFs),
- varying the DFs stiffness (dividing/multiplying parameter b of Normal *cdf* based DFs by 2 or 4),
- varying the DFs weights in the DI ($w_1 = 1/2$, $w_3 = 1/4$ and $w_4 = 1/4$ or $w_1 = 3/5$, $w_3 = 1/5$ and $w_4 = 1/5$)

Those parameters are varied in a univariate way as represented on Figure 4.20. For each situation, 10 000 simulations, similar to simulations on the basic situation, were performed. The main results concerning the bias of $\hat{D}^S(\mathbf{x})$ and $\hat{V}[\hat{D}^S(\mathbf{x})]$, the distribution of $\hat{D}^S(\mathbf{x})$, the confidence interval coverage and length and the IOZ coverage and size are presented in this section.

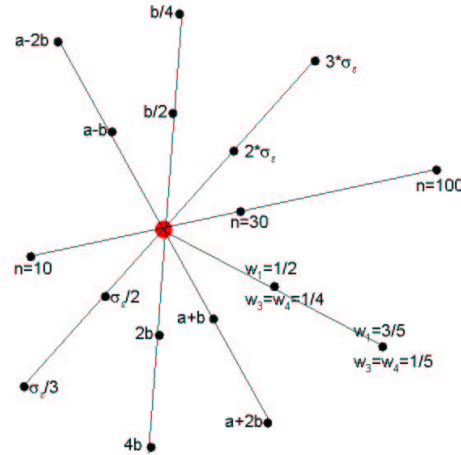


Figure 4.20: Simulation plan. The red center point represents the simplified Derringer and Suich problem (Section 8.3). Around this situation, some parameters are varied (in a univariate way): the model standard errors of residuals (σ_ϵ), the number of available data to fit the models (n), the parameters a and b of the Normal *cdf* based DFs, as well as the DFs weights in the DI (w).

Varying the model standard errors of residuals

In the basic situation, the standard errors of residuals are $\sigma_{\epsilon_1} = 5.61$, $\sigma_{\epsilon_3} = 20.55$ and $\sigma_{\epsilon_4} = 1.27$ respectively for the three regression models. Four other situations were simulated by decreasing those standard errors of residuals ($\forall i, \sigma_{\epsilon_i}/2$ and $\forall i, \sigma_{\epsilon_i}/3$) and increasing them ($\forall i, 2\sigma_{\epsilon_i}$ and $\forall i, 3\sigma_{\epsilon_i}$).

First the impact of varying the standard errors of residuals is studied on the DFs and the DI. By increasing σ_ϵ , $D^S(\mathbf{x})$ is becoming flatter over $[-1, 1]^2$ (see Figure 4.22). This can be explained by looking at the effect of increasing σ_ϵ on the DFs (see Figure 4.21): for a given grid-point $\mathbf{x}$, the corresponding responses $\mathbf{Y}|\mathbf{x}$ and the corresponding DF $d(\mathbf{Y})|\mathbf{x}$ are more variable and $E[d(\mathbf{Y})|\mathbf{x}]$ increases if $d(\mathbf{Y})|\mathbf{x}$ was on average near 0 or decreases if $d(\mathbf{Y})|\mathbf{x}$ was on average near 1.

For the $3\sigma_\epsilon$ situation, the estimator $\hat{D}^S(\mathbf{x})$ fails, resulting in a negative desirability for some simulations runs (0.44% of cases). The approximation of the Delta method is not enough accurate. But, it has to be noticed that optimising a desirability index with such noisy responses does not make sense as the standard errors of residuals ($\sigma_{\epsilon_1} = 16.83$, $\sigma_{\epsilon_3} = 61.64$ and $\sigma_{\epsilon_4} = 3.80$) exceeds the standard deviation of the observed responses over $[-1, 1]^2$ (9.51, 67.04 and 3.00, respectively).

The behaviour of the proposed methodology was studied on the $\sigma_\epsilon/3$, $\sigma_\epsilon/2$ and $2\sigma_\epsilon$ situations and summarised by some indices in Table 4.6. The approximations of $D^S(\mathbf{x})$ and $V[\hat{D}^S(\mathbf{x})]$ work very well for smaller standard errors of residuals, but become less accurate for the $2\sigma_\epsilon$ situation, especially when estimating $V[\hat{D}^S(\mathbf{x})]$. Nevertheless, this is not a so bad result for standard errors of residuals nearly equal to the standard deviation of the observed responses over $[-1, 1]^2$.

By computing the p-values of the Kolmogorov-Smirnov averaged over the grid, the adequacy of the Normal distribution of $\text{logit}(\hat{D}^S(\mathbf{x}))$ and $\text{logit}(\hat{D}^S(\mathbf{x})) - \text{logit}(\hat{D}^S(\mathbf{x}_{\text{opt}}^{\mathbf{S}}))$ seems to decrease with σ_ϵ but not enough to influence significantly the coverage of the confidence interval for $D^S(\mathbf{x})$ or the IOZ.

For $\sigma_\epsilon/3$, $\sigma_\epsilon/2$ and σ_ϵ , the coverage rate of the 90% confidence interval for $D^S(\mathbf{x})$ nearly equals 90% but is higher for the $2\sigma_\epsilon$ situation (longer CI). This is a direct consequence of the upward bias of $\hat{V}[\hat{D}^S(\mathbf{x})]$. As expected, the CI length and the IOZ size increases with σ_ϵ . The coverage of IOZ exceeds 90%, but the multiplicity correction of Benjamini and Yekutieli (2001) is necessary (without it, the coverage of IOZ is smaller than 90%) and the corrected IOZ is coherent with the optima found in the different simulations.

One can conclude that, with reasonable standard error of residuals, the proposed methodology based on the Delta method performs well.

	$\sigma_\epsilon/3$	$\sigma_\epsilon/2$	σ_ϵ	$2\sigma_\epsilon$
$\frac{Bias(\hat{D}^S(\mathbf{x}))}{\hat{D}^S(\mathbf{x})}$	-0.000020	-0.000501	-0.001459	-0.007062
	0.277	0.276	0.270	0.248
$\frac{Bias(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]})}{\sqrt{V[\hat{D}^S(\mathbf{x})]}}$	0.000091	0.000327	0.002730	0.018802
	0.015	0.022	0.043	0.084
$pvalue\ KS\ logit(\hat{D}^S(\mathbf{x}))$	0.71	0.63	0.45	0.36
$pvalue\ KS\ logit(\hat{D}^S(\mathbf{x})) - logit(\hat{D}^S(\mathbf{x}_{opt}^S))$	0.70	0.66	0.54	0.20
$CR_{theo\ CI_{90\%}}(\mathbf{x})$ in %	88.70	88.95	90.30	92.80
$\frac{length(theo\ CI_{90\%})}{length(simu\ CI_{90\%})}$	0.045	0.068	0.134	0.254
	0.994	1.004	1.055	1.275
Coverage IOZ in %	99.15	98.65	98.44	94.82
Points in TOZ in %	0.41	0.66	2.12	11.14
Points in IOZ in %	1.94	4.33	16.67	62.56
Optima found in %	2.49	4.76	13.83	48.98

Table 4.6: Effect of varying the standard errors of residuals on the proposed methodology.

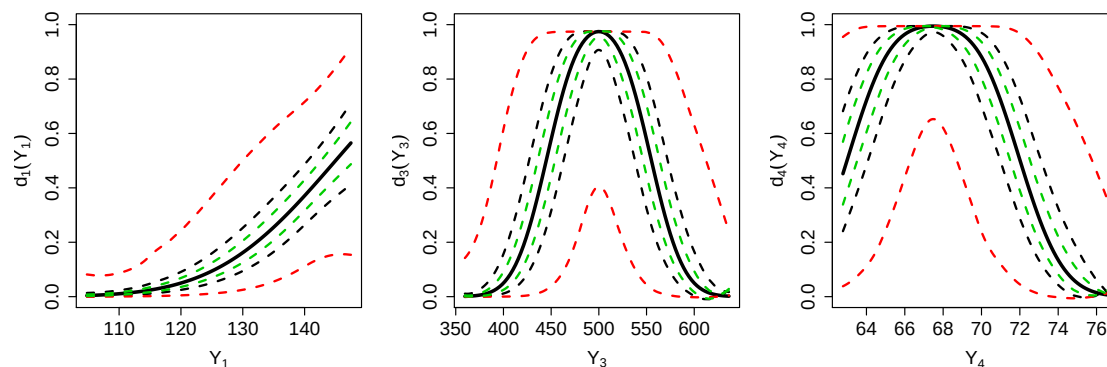


Figure 4.21: Continuous black lines: Normal *cdf* based DFs as a function of the expected values of the three responses to optimise over $[-1, 1]^2$ for the basic situation (σ_ϵ). Dashed curves: 5% and 95% quantiles of the simulated distribution of the Normal *cdf* based DFs as a function of the estimated responses to optimise over $[-1, 1]^2$ (black: basic σ_ϵ , red: $3\sigma_\epsilon$ and green: $\sigma_\epsilon/3$).

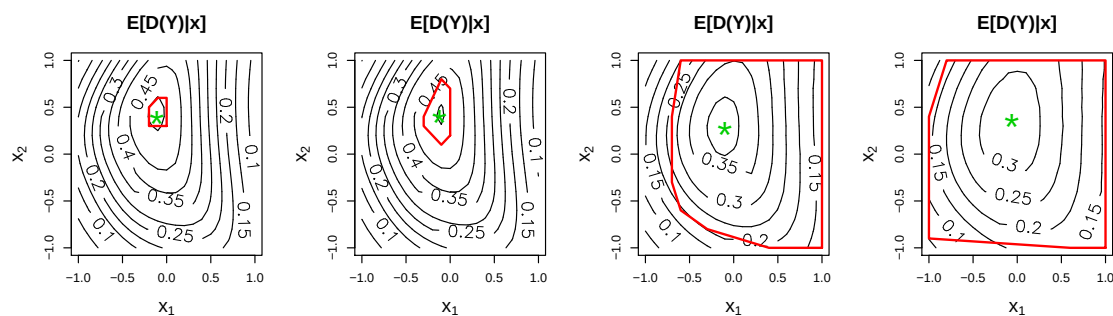


Figure 4.22: Contour-plot of $D^S(\mathbf{x})$ for different standard errors of residuals: $\sigma_\epsilon/3$, $\sigma_\epsilon/2$, $2\sigma_\epsilon$ and $3\sigma_\epsilon$ (from left to right). The star corresponds to the optimum $\mathbf{x}_{\text{opt}}^S$ and the red line delimits the zone of simulated optima (Step 3 of the simulations).

Varying the sample size

In the basic situation, $n = 20$ measurements of each response were simulated according to a two-variable central composite design (4 replicates in the center and 2 replicates at each other point). Three other situations were considered by generating datasets of different sizes: $n = 10$ (two-variable central composite design with 2 experiments in the center and 1 at each other point), $n = 30$ (two-variable central composite design with 6 experiments in the center and 3 at each other point) and $n = 100$ (two-variable central composite design with 20 experiments in the center and 10 at each other point).

The DFs and the DI remain the same (see Figures 4.23 and 4.7). According to the regression theory, the variance $V[\hat{E}[Y|\mathbf{x}]]$ decreases with the sample size and it is expected that the Delta method will better work as the sample size increases.

Table 4.7 summarises the results of the proposed methodology for the four sample sizes ($n = 10, 20, 30$ or 100). As expected, the bias of $\hat{D}^S(\mathbf{x})$ and $\hat{V}[\hat{D}^S(\mathbf{x})]$ decreases with n , and the Normal distribution of $\text{logit}(\hat{D}^S(\mathbf{x}))$ and $\text{logit}(\hat{D}^S(\mathbf{x})) - \text{logit}(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))$ is more and more plausible. Confidence intervals for $D^S(\mathbf{x})$ and IOZs are smaller. The coverage of the CI is close to 90%, except for $n = 10$ where it is 86.54%. Nevertheless, this is reasonable with so few data to estimate a second order polynomial regression. The coverage of the IOZ is higher than 90% but coincides with the size of the optima found zone (except for $n = 10$).

	$n = 10$	$n = 20$	$n = 30$	$n = 100$
$\text{Bias}(\hat{D}^S(\mathbf{x}))$	-0.0031	-0.0015	-0.0009	-0.0006
$\frac{\text{Bias}(\hat{D}^S(\mathbf{x}))}{\hat{D}^S(\mathbf{x})}$	0.270	0.270	0.270	0.270
$\text{Bias}(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]})$	0.0042	0.0027	0.0021	0.0012
$\frac{\text{Bias}(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]})}{\sqrt{V[\hat{D}^S(\mathbf{x})]}}$	0.061	0.043	0.035	0.019
$p\text{value } KS \text{ logit}(\hat{D}^S(\mathbf{x}))$	0.40	0.45	0.51	0.76
$p\text{value } KS \text{ logit}(\hat{D}^S(\mathbf{x})) - \text{logit}(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))$	0.43	0.54	0.55	0.76
$CR_{theo CI_{90\%}}(\mathbf{x})$ in %	86.54	90.30	90.97	91.62
$\text{length}(\text{theo } CI_{90\%})$	0.184	0.134	0.111	0.061
$\text{length}(\text{theo } CI_{90\%})/\text{length}(\text{simu } CI_{90\%})$	1.025	1.055	1.058	1.067
Coverage IOZ in %	96.48	98.44	98.68	99.17
Points in TOZ in %	4.36	2.12	1.46	0.67
Points in IOZ in %	31.19	16.67	12.04	3.86
Optima found in %	23.36	13.83	11.56	3.63

Table 4.7: Effect of varying the sample size on the proposed methodology.

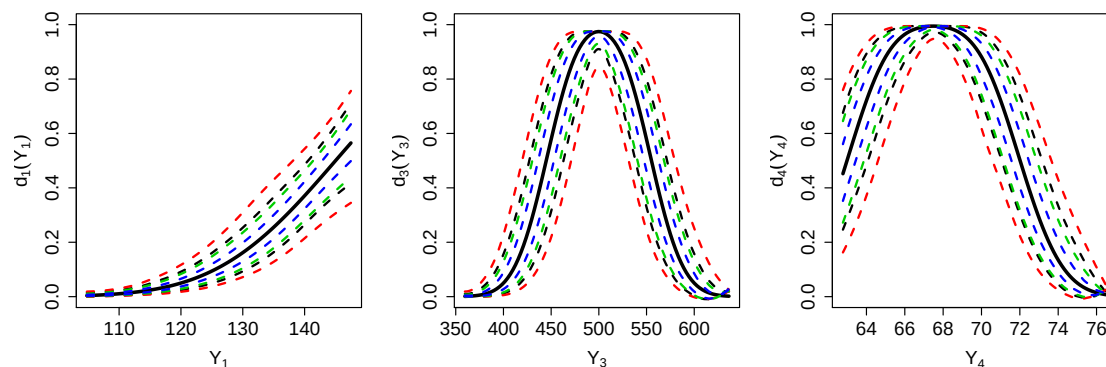


Figure 4.23: Continuous black lines: Normal *cdf* based DFs as a function of the expected values of the three responses to optimise over $[-1, 1]^2$ for the basic situation ($n = 20$). Dashed curves: 5% and 95% quantiles of the simulated distribution of the Normal *cdf* based DFs as a function of the estimated responses to optimise over $[-1, 1]^2$ (red: $n = 10$, black: basic $n = 20$, green: $n = 30$ and blue: $n = 100$).

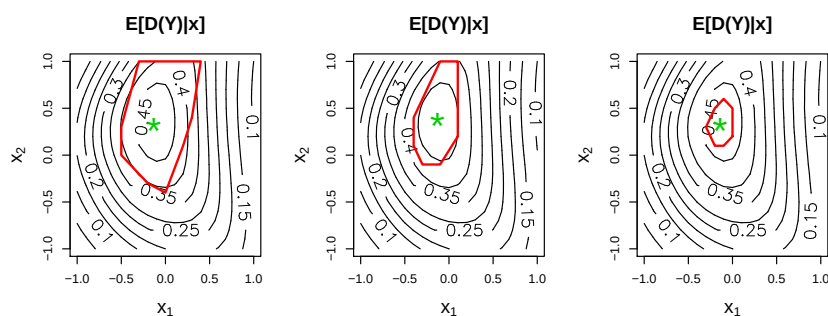


Figure 4.24: Contour-plot of $D^S(\mathbf{x})$ for different sample size: $n = 10$, $n = 30$ and $n = 100$ (from left to right). The star corresponds to the optimum $\mathbf{x}_{\text{opt}}^S$ and the red line delimits the zone of simulated optima (Step 3 of the simulations).

Shifting the DFs

Denoting by $a = (a_1, a_{3_1}, a_{3_2}, a_{4_1}, a_{4_2})$ and $b = (b_1, b_{3_1}, b_{3_2}, b_{4_1}, b_{4_2})$ the parameters of the Normal *cdf* based DFs in the basic situation, four other cases are considered by shifting the three DFs to the left ($a - b$ or $a - 2b$) or to the right ($a + b$ or $a + 2b$) as shown on Figure 4.25. The corresponding effect on the DI can be visualised on Figure 4.26. $D^S(\mathbf{x})$ decreases on average when shifting the DFs to the left, mainly because $d_1(Y_1)$ is decreasing. The corresponding optimum is, of course, also shifted. The $a + 2b$ situation is maybe too much extreme as the desirability is very small over $[-1, 1]^2$ with a maximum of $D^S(\mathbf{x}_{\text{opt}}^S) = 0.0852$.

Table 4.8 summarises the results of the methodology for the different DF shapes. The bias of $\hat{D}^S(\mathbf{x})$ is always negligible compared to the desirability value itself $\hat{D}^S(\mathbf{x})$. The bias of $\hat{V}^S[\hat{D}^S(\mathbf{x})]$ is small (20 times smaller than the corresponding true quantity $V^S[\hat{D}^S(\mathbf{x})]$) but not negligible, leading to confidence intervals for $D^S(\mathbf{x})$ a little bit too long and IOZ's a little bit too large. Nevertheless, the order of magnitude of those two tools corresponds reasonably to the simulations and the results of the Delta method are acceptable.

	$a - 2b$	$a - b$	a	$a + b$	$a + 2b$
$\frac{\text{Bias}(\hat{D}^S(\mathbf{x}))}{\hat{D}^S(\mathbf{x})}$	-0.0046	-0.0045	-0.0015	0.0002	0.0005
$\frac{\text{Bias}(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]})}{\sqrt{V[\hat{D}^S(\mathbf{x})]}}$	0.0028	0.0033	0.0027	0.0019	0.0007
$\frac{pvalue\ KS\ logit(\hat{D}^S(\mathbf{x}))}{pvalue\ KS\ logit(\hat{D}^S(\mathbf{x})) - logit(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))}$	0.62	0.54	0.45	0.38	0.21
$\frac{CR_{theo\ CI_{90\%}}(\mathbf{x})\text{ in \%}}{length(theo\ CI_{90\%})}$	88.87	90.10	90.30	90.57	90.05
$\frac{length(theo\ CI_{90\%})}{length(simu\ CI_{90\%})}$	0.172	0.167	0.134	0.078	0.031
Coverage IOZ in %	92.59	96.30	98.44	98.80	98.87
Points in TOZ in %	1.71	1.76	2.12	2.90	2.26
Points in IOZ in %	7.37	11.36	16.67	19.38	20.06
Optima found in %	10.66	10.43	13.83	13.38	11.34

Table 4.8: Effect of varying the parameter a of the Normal *cdf* based DFs on the proposed methodology.

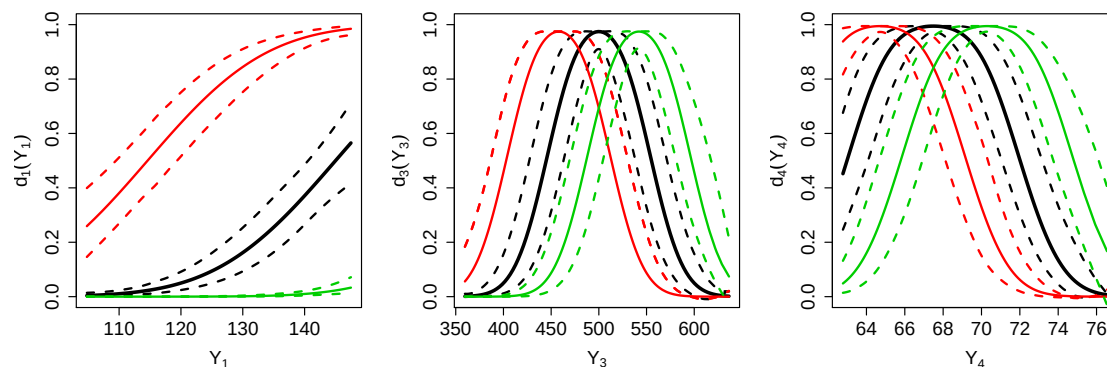


Figure 4.25: Continuous black lines: Normal *cdf* based DFs as a function of the expected values of the three responses to optimise over $[-1, 1]^2$ for the basic situation (a and b). Dashed curves: 5% and 95% quantiles of the simulated distribution of the Normal *cdf* based DFs as a function of the estimated responses to optimise over $[-1, 1]^2$ (red: $a - 2b$, black: basic a , green: $a + 2b$).

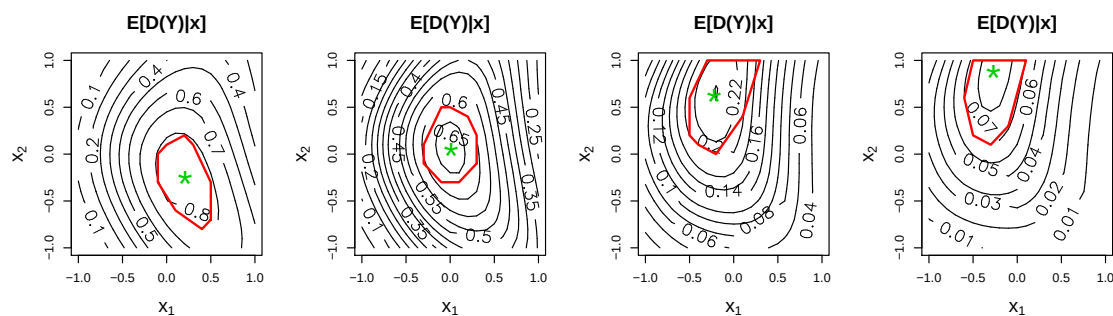


Figure 4.26: Contour-plot of $D^S(\mathbf{x})$ for different parameter a of the Normal *cdf* based DFs: $a - 2b$, $a - b$, $a + b$ and $a + 2b$ (from left to right). The star corresponds to the optimum $\mathbf{x}_{\text{opt}}^S$ and the red line delimits the zone of simulated optima (Step 3 of the simulations).

Varying the DFs stiffness?

Decreasing (resp. increasing) the parameters $b = (b_1, b_{3_1}, b_{3_2}, b_{4_1}, b_{4_2})$ of the Normal *cdf* based DFs of the basic situation leads to stiffer (resp. flatter) DF curves (see Figure 4.27) and DI surface (see Figure 4.28). Four new sets of simulations were performed for $b/4$, $b/2$, $2 * b$ and $4 * b$.

The situation with $b/4$ corresponds nearly to a binary case with either totally acceptable product for Y_3 or Y_4 or totally unacceptable product for Y_3 or Y_4 . Considering the DFs as random variables, $D^S(\mathbf{x})$ is, in that extreme situation, always close to 0 over $[-1, 1]^2$ except around the optimum where it reaches $D^S(\mathbf{x}_{\text{opt}}^S) = 0.1086$. The proposed estimator $\hat{D}^S(\mathbf{x})$ did not work for most points $\mathbf{x}$ in $[-1, 1]^2$ because of numerical problems: one DF value $d_i(\hat{E}[Y_i|\mathbf{x}])$ was equaled to zero due to a low precision of the software, resulting in an undetermined $\hat{D}^S(\mathbf{x})$. The same numerical problems occurred in some simulations with $b/2$ but rarely. Those cases were suppressed to obtain the presented results of Table 4.9.

Both $\hat{D}^S(\mathbf{x})$ and $\hat{V}[\hat{D}^S(\mathbf{x})]$ are highly biased in the $b/2$ situation (also close to binary DFs), leading to too long confidence interval for $D^S(\mathbf{x})$ (and not well centered) and too huge IOZ. One conclude that flatter the DFs are (larger parameter b), better the approximations work.

	$b/2$	b	$2b$	$4b$
$Bias(\hat{D}^S(\mathbf{x}))$	0.0054	-0.0015	-0.0025	-0.0008
$D^S(\mathbf{x})$	0.088	0.270	0.449	0.524
$Bias(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]})$	0.01046	0.00273	0.00038	-0.00003
$\sqrt{V[\hat{D}^S(\mathbf{x})]}$	0.051	0.043	0.026	0.011
$pvalue KS logit(\hat{D}^S(\mathbf{x}))$	0.56	0.45	0.71	0.40
$pvalue KS logit(\hat{D}^S(\mathbf{x})) - logit(\hat{D}^S(\mathbf{x}_{\text{opt}}^S))$	0.69	0.54	0.77	0.57
$CR_{theo CI_{90\%}}(\mathbf{x})$ in %	93.05	90.30	89.22	88.87
$length(theo CI_{90\%})$	0.136	0.134	0.084	0.037
$length(theo CI_{90\%})/length(simu CI_{90\%})$	1.524	1.055	1.006	0.990
Coverage IOZ in %	98.38	98.44	98.40	98.27
Points in TOZ in %	3.49	2.12	1.79	2.29
Points in IOZ in %	27.90	16.67	15.28	17.34
Optima found in %	12.47	13.83	12.93	14.97

Table 4.9: Effect of varying the parameter b of the Normal *cdf* based DFs on the proposed methodology.

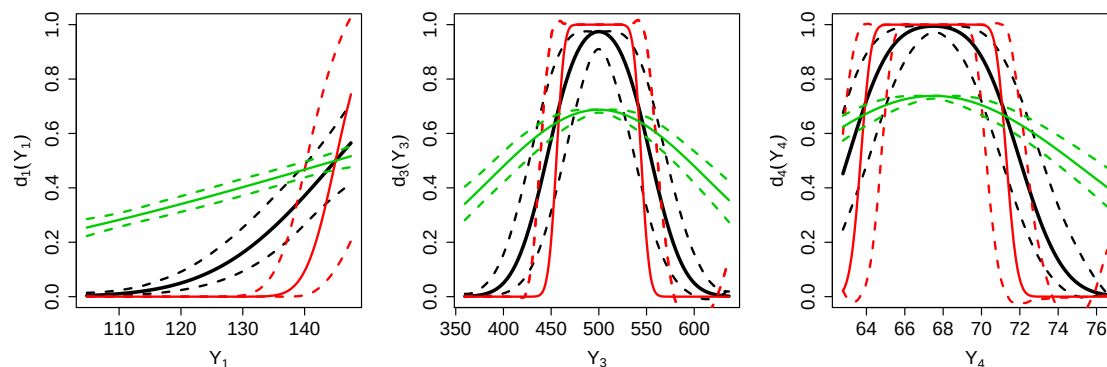


Figure 4.27: Continuous black lines: Normal *cdf* based DFs as a function of the expected values of the three responses to optimise over $[-1, 1]^2$ for the basic situation (a and b). Dashed curves: 5% and 95% quantiles of the simulated distribution of the Normal *cdf* based DFs as a function of the estimated responses to optimise over $[-1, 1]^2$ (red: $b/4$, black: basic b , green: $4b$).

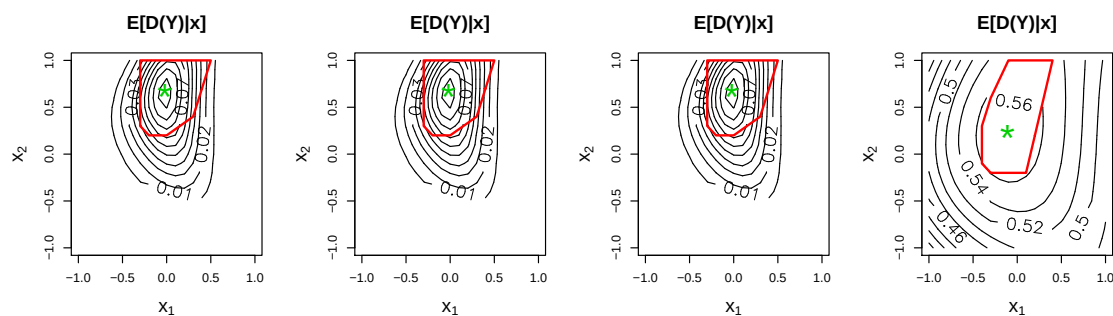


Figure 4.28: Contour-plot of $D^S(\mathbf{x})$ for different parameter b of the Normal *cdf* based DFs: $b/4$, $b/2$, $2b$ and $4 * b$ (from left to right). The star corresponds to the optimum $\mathbf{x}_{\text{opt}}^S$ and the red line delimits the zone of simulated optima (Step 3 of the simulations).

Varying the weighting of the DFs in the DI

In the basic situation, the same weight is used to aggregate the three DFs in a DI by their geometric mean: $w_1 = w_3 = w_4 = 1/3$. Two other situations were simulated according more importance to the first response: $w_1 = 1/2$, $w_3 = 1/4$ and $w_4 = 1/4$ or $w_1 = 3/5$, $w_3 = 1/5$ and $w_4 = 1/5$. As DFs are taking values between 0 and 1, increasing (resp. decreasing) the exponent w decreases (resp. increases) the weighted DF $d(Y)^w$ and leads to stiffer curves (resp. flatter) (Figure 4.29). As a consequence, the desirability surface $D^S(\mathbf{x})$ changes, as well as the corresponding optimum (Figure 4.30). In this example, $D^S(\mathbf{x})$ decreases if w_1 increases (and w_3 and w_4 decrease) because $d_1(Y_1)$ is taking particularly small values compared to $d_3(Y_3)$ and $d_4(Y_4)$.

Table 4.10 summarises the results of the methodology for those three different weighting situations. The bias of $\hat{D}^S(\mathbf{x})$ is negligible, but the bias of $\hat{V}[\hat{D}^S(\mathbf{x})]$ is becoming more important as w_1 increases, relatively to the true quantity $V[\hat{D}^S(\mathbf{x})]$. One conclude that increasing the weight of an already restrictive property leads to worse behaviour of the Delta method based approach.

	$w_1 = 1/3$ $w_3 = 1/3$ $w_4 = 1/3$	$w_1 = 1/2$ $w_3 = 1/4$ $w_4 = 1/4$	$w_1 = 3/5$ $w_3 = 1/5$ $w_4 = 1/5$
$Bias(\hat{D}^S(\mathbf{x}))$	-0.00146	-0.00006	0.00039
$\frac{Bias(\hat{D}^S(\mathbf{x}))}{D^S(\mathbf{x})}$	0.270	0.214	0.191
$Bias(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]})$	0.0027	0.0021	0.0029
$\frac{Bias(\sqrt{\hat{V}[\hat{D}^S(\mathbf{x})]})}{\sqrt{V[\hat{D}^S(\mathbf{x})]}}$	0.0428	0.0385	0.0371
$pvalue\ KS\ logit(\hat{D}^S(\mathbf{x}))$	0.45	0.38	0.59
$pvalue\ KS\ logit(\hat{D}^S(\mathbf{x})) - logit(\hat{D}^S(\mathbf{x}_{opt}^S))$	0.54	0.61	0.77
$CR_{theo\ CI_{90\%}}(\mathbf{x})$ in %	90.30	90.00	90.59
$length(theo\ CI_{90\%})$	0.134	0.119	0.113
$\frac{length(theo\ CI_{90\%})}{length(simu\ CI_{90\%})}$	1.055	1.047	1.075
Coverage IOZ in %	98.44	98.62	98.87
Points in TOZ in %	2.12	2.93	2.26
Points in IOZ in %	16.67	18.84	18.76
Optima found in %	13.83	13.15	11.11

Table 4.10: Effect of varying the w weighting of the DF in the DI on the proposed methodology.

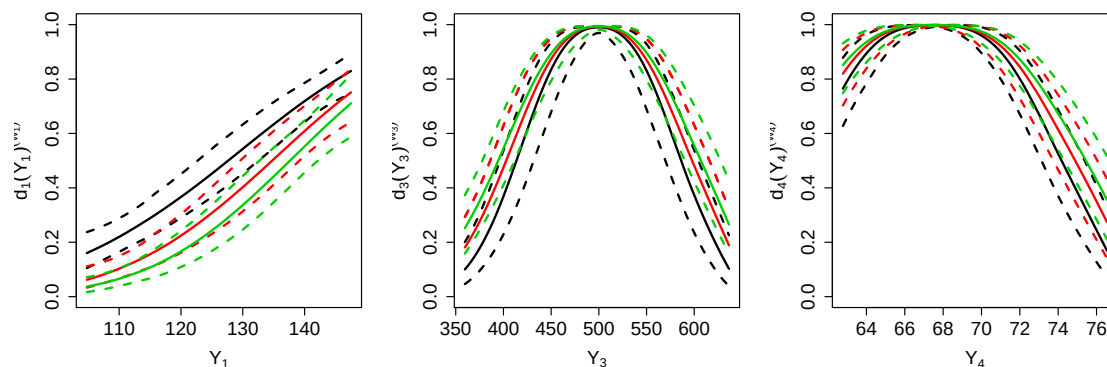


Figure 4.29: Continuous black lines: Normal *cdf* based DFs as a function of the expected values of the three responses to optimise over $[-1, 1]^2$ for the basic situation weighted by $w_1 = 1/3$, $w_3 = 1/3$ and $w_4 = 1/3$. Dashed curves: 5% and 95% quantiles of the simulated distribution of the weighted Normal *cdf* based DFs as a function of the estimated responses to optimise over $[-1, 1]^2$ (black: basic $w_1 = w_3 = w_4 = 1/3$; red: $w_1 = 1/2$, $w_3 = 1/4$ and $w_4 = 1/4$; green: $w_1 = 3/5$, $w_3 = 1/5$ and $w_4 = 1/5$).

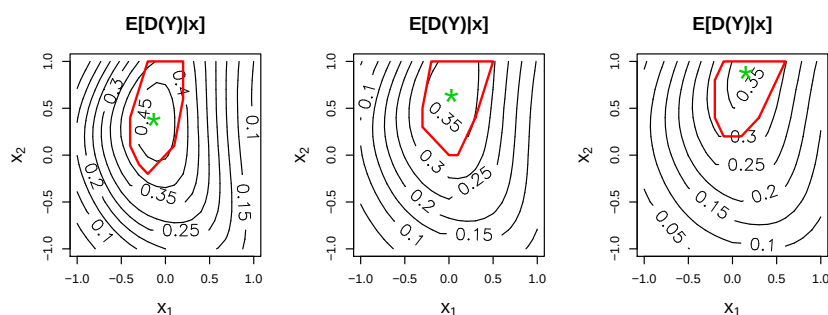


Figure 4.30: Contour-plot of $D^S(\mathbf{x})$ for different weighting of the DFs in the DI: basic case $w_1 = w_3 = w_4 = 1/3$; $w_1 = 1/2$, $w_3 = 1/4$ and $w_4 = 1/4$ or $w_1 = 3/5$, $w_3 = 1/5$ and $w_4 = 1/5$ (from left to right). The star corresponds to the optimum $\mathbf{x}_{\text{opt}}^S$ and the red line delimits the zone of simulated optima (Step 3 of the simulations).

4.9 Conclusion

This chapter proposes a stochastic use of desirability indexes in multicriteria optimisation and a new methodology to quantify uncertainty of estimated desirability index and estimated optimum. It generalises and integrates existing literature on the subject. Moreover, it is applicable to very general regression models (multiple linear regression, logistic regression, partial least squares regression,...) and to general desirability functions provided that they are two times continuously differentiable.

It first emphasises that, due to model uncertainty, a desirability index is a random variable and suggests to take the expected desirability $D^S(\mathbf{x}) = E[D(\mathbf{Y})|\mathbf{x}]$ as objective function to be optimised. As this expectation is in most cases very difficult, even impossible, to derive analytically, a Delta method based approach is proposed to approximate it. This approximation depends only on expected responses $E[Y_i|\mathbf{x}]$ and on response variances $V[Y_i|\mathbf{x}]$. By replacing these usually unknown quantities by their classical estimators, an estimator for the expected DI is obtained ($\hat{D}^S(\mathbf{x})$).

The Delta method theorem is then applied to derive the asymptotic distribution of $\hat{D}^S(\mathbf{x})$ and build a confidence interval for the true expected DI. This theorem may also be used to derive the asymptotic distribution of the difference of the estimated expected DI of two solutions and construct an indistinguishable zone around the estimated optimum, gathering factors levels having non-significantly smaller expected DI as the optimum found.

In most cases, the difference between the classical use of DI, $D^C(\mathbf{x})$, and the new one, $D^S(\mathbf{x})$, is small as well as the resulting optima, $\mathbf{x}_{opt}^C$ and $\mathbf{x}_{opt}^S$. The most important improvement is the new methodology based on the Delta method theorem to measure the prediction error propagation on the estimated expected DI. This allows to compare the estimated expected DI of two solutions taking into account their uncertainty and constructing an indistinguishable zone around the optimum found. Simulations on the Derringer and Suich (1980) data, demonstrate that the Delta method provides, in most situations, a good second order approximation of the expected DI, as well as adequate confidence intervals and IOZ's. Nevertheless, the method may fail if the DFs or the weighted DFs are too stiff or the standard error of residuals is large, especially compared to the number of available data to fit the models. Those simulations reveal that this uncertainty propagation is much more important in practice than the bias correction between $D^C(\mathbf{x})$ and $D^S(\mathbf{x})$.

Acknowledgment

We want to express our gratitude to Dr H. Trautmann and Prof. C. Weihs from the university of Dortmund for stimulating discussion. We also gratefully thank the co-guest editors of the response surface methodology special issue for Journal of Quality Technology and Quantitative Management for their comments which have helped to improve this research.

4.10 Appendices

Appendix 1: proof of $E[D(\mathbf{Y})|\mathbf{x}]$ approximation (4.8)

Approximation (4.8) of $E[D(\mathbf{Y})|\mathbf{x}]$ is a nearly direct application of approximation (4.7)

$$E[h(Y_1, \dots, Y_p)|\mathbf{x}] \approx h(E[Y_1|\mathbf{x}], \dots, E[Y_p|\mathbf{x}]) + \frac{1}{2} \sum_{i=1}^p \left[\frac{\partial^2 h}{(\partial Y_i)^2} V[Y_i|\mathbf{x}] \right]$$

to the multivariate function $h(Y_1, \dots, Y_p) = \prod_{i=1}^p (d_i(Y_i))^{w_i}$.

Let us first compute the corresponding terms appearing in (4.7): $h(E[Y_1|\mathbf{x}], \dots, E[Y_p|\mathbf{x}])$, and its second derivatives $\frac{\partial^2 h}{(\partial Y_i)^2}$ via its first derivatives $\frac{\partial h}{\partial Y_i}$.

- $h(E[Y_1|\mathbf{x}], \dots, E[Y_p|\mathbf{x}]) = \prod_{i=1}^p (d_i(E[Y_i|\mathbf{x}]))^{w_i}$
- $$\begin{aligned} \frac{\partial h}{\partial Y_i} &= \frac{\partial}{\partial Y_i} \left(\prod_{j=1}^p (d_j(Y_j))^{w_j} \right) \\ &= \left(\prod_{j=1; j \neq i}^p (d_j(Y_j))^{w_j} \right) \cdot \left(\frac{\partial}{\partial Y_i} (d_i(Y_i))^{w_i} \right) \\ &= \left(\prod_{j=1; j \neq i}^p (d_j(Y_j))^{w_j} \right) \cdot \left(w_i \cdot (d_i(Y_i))^{w_i-1} \cdot \frac{\partial d_i(Y_i)}{\partial Y_i} \right) \end{aligned}$$
- $$\begin{aligned} \frac{\partial^2 h}{(\partial Y_i)^2} &= \frac{\partial}{\partial Y_i} \left(\frac{\partial h}{\partial Y_i} \right) = \left(\prod_{j=1; j \neq i}^p (d_j(Y_j))^{w_j} \right) \cdot \frac{\partial}{\partial Y_i} \left(w_i \cdot (d_i(Y_i))^{w_i-1} \cdot \frac{\partial d_i(Y_i)}{\partial Y_i} \right) \\ &= \left(\prod_{j=1; j \neq i}^p (d_j(Y_j))^{w_j} \right) \cdot \left(w_i \cdot (w_i - 1) \cdot (d_i(Y_i))^{w_i-2} \cdot \left(\frac{\partial d_i(Y_i)}{\partial Y_i} \right)^2 + w_i \cdot (d_i(Y_i))^{w_i-1} \cdot \frac{\partial^2 d_i(Y_i)}{(\partial Y_i)^2} \right) \\ &= \left(\prod_{j=1}^p (d_j(Y_j))^{w_j} \right) \cdot \left(w_i \cdot (w_i - 1) \cdot \frac{\left(\frac{\partial d_i(Y_i)}{\partial Y_i} \right)^2}{(d_i(Y_i))^2} + w_i \cdot \frac{\frac{\partial^2 d_i(Y_i)}{(\partial Y_i)^2}}{(d_i(Y_i))} \right) \end{aligned}$$

By plugging those expressions in (4.7), evaluating the second derivatives in the conditional expected values $E[Y_i|\mathbf{x}]$ and emphasising $\prod_{i=1}^p (d_i(E[Y_i|\mathbf{x}]))^{w_i} = D^C(\mathbf{x})$, approximation (4.8) is obtained:

$$D^S(\mathbf{x}) = E[D(\mathbf{Y})|\mathbf{x}] = E \left[\prod_{i=1}^p (d_i(Y_i))^{w_i} | \mathbf{x} \right] \approx \left[\prod_{i=1}^p (d_i(E[Y_i|\mathbf{x}]))^{w_i} \right] \cdot (1 + B(\mathbf{x}))$$

with

$$B(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^p c_i(\mathbf{x}) \cdot V[Y_i|\mathbf{x}] \quad \text{as defined by (4.9) and}$$

$$c_i(\mathbf{x}) = w_i(w_i - 1) \left(\frac{\frac{\partial d_i(Y_i)}{\partial Y_i}}{d_i(E[Y_i|\mathbf{x}])} \right)^2 + w_i \frac{\frac{\partial^2 d_i(Y_i)}{\partial Y_i^2}}{d_i(E[Y_i|\mathbf{x}])} \quad \text{as defined by (4.10)}$$

Appendix 2: proof of $E[D(\mathbf{Y})|\mathbf{x}]$ approximation (4.36) in the case of correlated model residuals

Approximation (4.36) of $E[D(\mathbf{Y})|\mathbf{x}]$ in the case of correlated model residuals is a nearly direct application of approximation (4.5)

$$\begin{aligned} E[h(Y_1, Y_2, \dots, Y_p)|\mathbf{x}] &\approx h(E[Y_1|\mathbf{x}], E[Y_2|\mathbf{x}], \dots, E[Y_p|\mathbf{x}]) \\ &+ \frac{1}{2} \sum_{i=1}^p \left[\frac{\partial^2 h}{\partial Y_i^2} V[Y_i|\mathbf{x}] \right] + \sum_{i < j}^p \left[\frac{\partial^2 h}{\partial Y_i \partial Y_j} Cov(Y_i, Y_j|\mathbf{x}) \right] \end{aligned}$$

to the multivariate function $h(Y_1, \dots, Y_p) = \prod_{i=1}^p (d_i(Y_i))^{w_i}$.

In the preceeding proof, the expressions of $h(E[Y_1|\mathbf{x}], \dots, E[Y_p|\mathbf{x}])$ and its first and second derivatives $\frac{\partial^2 h}{(\partial Y_i)^2}$ are given considering $h(Y_1, \dots, Y_p) = \prod_{i=1}^p (d_i(Y_i))^{w_i}$. To prove approximation (4.36), expression $\frac{\partial^2 h}{\partial Y_i \partial Y_j}$ appearing in the last term of approximation (4.5) is analysed.

- $h(E[Y_1|\mathbf{x}], \dots, E[Y_p|\mathbf{x}]) = \prod_{i=1}^p (d_i(E[Y_i|\mathbf{x}]))^{w_i}$
- $\frac{\partial h}{\partial Y_i} = \left(\prod_{j=1; j \neq i}^p (d_j(Y_j))^{w_j} \right) \cdot \left(w_i \cdot (d_i(Y_i))^{w_i-1} \cdot \frac{\partial d_i(Y_i)}{\partial Y_i} \right)$
- $\frac{\partial^2 h}{(\partial Y_i)^2} = \left(\prod_{i=1}^p (d_i(Y_i))^{w_i} \right) \cdot \left(w_i \cdot (w_i - 1) \cdot \frac{\left(\frac{\partial d_i(Y_i)}{\partial Y_i} \right)^2}{(d_i(Y_i))^2} + w_i \cdot \frac{\frac{\partial^2 d_i(Y_i)}{(\partial Y_i)^2}}{(d_i(Y_i))} \right)$
- $\begin{aligned} \frac{\partial^2 h}{\partial Y_i \partial Y_j} &= \frac{\partial}{\partial Y_i} \left(\frac{\partial h}{\partial Y_j} \right) \\ &= \left(\prod_{k=1; k \neq i; k \neq j}^p (d_k(Y_k))^{w_k} \right) \cdot \left(w_i \cdot (d_i(Y_i))^{w_i-1} \cdot \frac{\partial d_i(Y_i)}{\partial Y_i} \right) \cdot \left(w_j \cdot (d_j(Y_j))^{w_j-1} \cdot \frac{\partial d_j(Y_j)}{\partial Y_j} \right) \\ &= \left(\prod_{k=1}^p (d_k(Y_k))^{w_k} \right) \cdot \left(w_i \cdot (d_i(Y_i))^{-1} \cdot \frac{\partial d_i(Y_i)}{\partial Y_i} \right) \cdot \left(w_j \cdot (d_j(Y_j))^{-1} \cdot \frac{\partial d_j(Y_j)}{\partial Y_j} \right) \\ &= \left(\prod_{k=1}^p (d_k(Y_k))^{w_k} \right) \cdot \left(w_i \cdot \frac{\partial d_i(Y_i)}{\partial Y_i} \right) \cdot \left(w_j \cdot \frac{\partial d_j(Y_j)}{\partial Y_j} \right) \end{aligned}$

By evaluating the derivatives of $h(Y_1, \dots, Y_p) = \prod_{i=1}^p (d_i(Y_i))^{w_i}$ in $E[Y_i|\mathbf{x}]$, approximation (4.5)

can be written

$$\begin{aligned}
E[D(\mathbf{Y})|\mathbf{x}] &\approx \prod_{k=1}^p (d_k(E[Y_k|\mathbf{x}]))^{w_k} \\
&+ \frac{1}{2} \sum_{i=1}^p \left[\left(\prod_{k=1}^p (d_k(E[Y_k|\mathbf{x}]))^{w_k} \right) \cdot \left(w_i \cdot (w_i - 1) \cdot \frac{\left(\frac{\partial d_i(Y_i)}{\partial Y_i} \right)^2}{(d_i(E[Y_i|\mathbf{x}]))^2} + w_i \cdot \frac{\frac{\partial^2 d_i(Y_i)}{(\partial Y_i)^2}}{(d_i(E[Y_i|\mathbf{x}]))} \right) \cdot V[Y_i|\mathbf{x}] \right] \\
&+ \sum_{i < j}^p \left[\left(\prod_{k=1}^p (d_k(E[Y_k|\mathbf{x}]))^{w_k} \right) \cdot \left(w_i \cdot \frac{\frac{\partial d_i(Y_i)}{\partial Y_i}}{d_i(E[Y_i|\mathbf{x}])} \right) \cdot \left(w_j \cdot \frac{\frac{\partial d_j(Y_j)}{\partial Y_j}}{d_j(E[Y_j|\mathbf{x}])} \right) \text{Cov}(Y_i, Y_j|\mathbf{x}) \right].
\end{aligned}$$

By emphasising $\prod_{k=1}^p (d_k(E[Y_k|\mathbf{x}]))^{w_k} = D^C(\mathbf{x})$ and defining the constants $c_i(\mathbf{x})$ and $c_{ij}(\mathbf{x})$ by (4.10) and (4.38) we obtain

$$E[D(\mathbf{Y})|\mathbf{x}] \approx \prod_{k=1}^p (d_k(E[Y_k|\mathbf{x}]))^{w_k} \cdot \left[1 + \frac{1}{2} \sum_{i=1}^p (c_i(\mathbf{x}) \cdot V[Y_i|\mathbf{x}]) + \sum_{i < j}^p (c_{ij}(\mathbf{x}) \text{Cov}(Y_i, Y_j|\mathbf{x})) \right].$$

Finally, by defining $B(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^p (c_i(\mathbf{x}) \cdot V[Y_i|\mathbf{x}]) + \sum_{i < j}^p (c_{ij}(\mathbf{x}) \text{Cov}(Y_i, Y_j|\mathbf{x}))$, we obtain approximation (4.36):

$$E[D(\mathbf{Y})|\mathbf{x}] \approx D^C(\mathbf{x}) \cdot (1 + B(\mathbf{x})).$$

The bias term can be written in matrix notation by $B(\mathbf{x}) = \frac{1}{2} \mathbf{1}_p(\mathbf{C}(\mathbf{x}) \bullet \Sigma) \mathbf{1}_p'$, with $\bullet$ denoting the Hadamard product of two matrices (element by element), $\mathbf{1}_p$ being the p-vector of 1 and the $(p \times p)$ matrix $\mathbf{C}(\mathbf{x})$ defined as

$$\mathbf{C}(\mathbf{x}) = \begin{pmatrix} c_1(\mathbf{x}) & c_{12}(\mathbf{x}) & \cdots & c_{1p}(\mathbf{x}) \\ c_{21}(\mathbf{x}) & c_2(\mathbf{x}) & \cdots & c_{2p}(\mathbf{x}) \\ \vdots & \vdots & \cdots & \vdots \\ c_{p1}(\mathbf{x}) & c_{p2}(\mathbf{x}) & \cdots & c_p(\mathbf{x}) \end{pmatrix}.$$

Appendix 3: first and second derivatives of the Normal *cdf* based DFs

The first and second derivatives of the Normal *cdf* based DFs are given bellow (index i is removed to simplify notations).

- if Y has to be maximised and $d(Y) = \Phi\left(\frac{Y-a}{b}\right)$ then

$$\begin{aligned}\frac{\partial d(Y)}{\partial Y} &= \phi\left(\frac{Y-a}{b}\right) \cdot \frac{1}{b} \quad \text{and} \\ \frac{\partial^2 d(Y)}{\partial Y^2} &= -\phi\left(\frac{Y-a}{b}\right) \cdot \frac{Y-a}{b^3}\end{aligned}$$

- if Y has to be minimised and $d(Y) = 1 - \Phi\left(\frac{Y-a}{b}\right)$ then

$$\begin{aligned}\frac{\partial d(Y)}{\partial Y} &= -\phi\left(\frac{Y-a}{b}\right) \cdot \frac{1}{b} \quad \text{and} \\ \frac{\partial^2 d(Y)}{\partial Y^2} &= \phi\left(\frac{Y-a}{b}\right) \cdot \frac{Y-a}{b^3}\end{aligned}$$

- if Y has to reach some optimal value and $d(Y) = \sqrt{\Phi\left(\frac{Y-a_1}{b_1}\right) \cdot [1 - \Phi\left(\frac{Y-a_2}{b_2}\right)]}$ then

$$\begin{aligned}\frac{\partial d(Y)}{\partial Y} &= \frac{1}{2} \cdot \left(1 - \Phi\left(\frac{Y-a_2}{b_2}\right)\right)^{-\frac{1}{2}} \cdot \left(-\phi\left(\frac{Y-a_2}{b_2}\right) \cdot \frac{1}{b_2}\right) \cdot \sqrt{\Phi\left(\frac{Y-a_1}{b_1}\right)} \\ &\quad + \frac{1}{2} \cdot \left(\Phi\left(\frac{Y-a_1}{b_1}\right)\right)^{-\frac{1}{2}} \cdot \left(\phi\left(\frac{Y-a_1}{b_1}\right) \cdot \frac{1}{b_1}\right) \cdot \sqrt{1 - \Phi\left(\frac{Y-a_2}{b_2}\right)} \quad \text{and} \\ \frac{\partial^2 d(Y)}{\partial Y^2} &= \frac{1}{2} \cdot \frac{-1}{2} \cdot \left(1 - \Phi\left(\frac{Y-a_2}{b_2}\right)\right)^{-\frac{3}{2}} \cdot \left(-\phi\left(\frac{Y-a_2}{b_2}\right) \cdot \frac{1}{b_2}\right)^2 \cdot \sqrt{\Phi\left(\frac{Y-a_1}{b_1}\right)} \\ &\quad + \frac{1}{2} \cdot \left(1 - \Phi\left(\frac{Y-a_2}{b_2}\right)\right)^{-\frac{1}{2}} \cdot \left(\phi\left(\frac{Y-a_2}{b_2}\right) \cdot \frac{Y-a_2}{b_2^3}\right) \cdot \sqrt{\Phi\left(\frac{Y-a_1}{b_1}\right)} \\ &\quad + \frac{1}{2} \cdot \left(1 - \Phi\left(\frac{Y-a_2}{b_2}\right)\right)^{-\frac{1}{2}} \cdot \left(-\phi\left(\frac{Y-a_2}{b_2}\right) \cdot \frac{1}{b_2}\right) \cdot \left(\frac{1}{2} \cdot \left(\Phi\left(\frac{Y-a_1}{b_1}\right)\right)^{-\frac{1}{2}} \cdot \phi\left(\frac{Y-a_1}{b_1}\right) \cdot \frac{1}{b_1}\right) \\ &\quad + \frac{1}{2} \cdot \frac{-1}{2} \cdot \left(\Phi\left(\frac{Y-a_1}{b_1}\right)\right)^{-\frac{3}{2}} \cdot \left(\phi\left(\frac{Y-a_1}{b_1}\right) \cdot \frac{1}{b_1}\right)^2 \cdot \sqrt{1 - \Phi\left(\frac{Y-a_2}{b_2}\right)} \\ &\quad + \frac{1}{2} \cdot \left(\Phi\left(\frac{Y-a_1}{b_1}\right)\right)^{-\frac{1}{2}} \cdot \left(-\phi\left(\frac{Y-a_1}{b_1}\right) \cdot \frac{Y-a_1}{b_1^3}\right) \cdot \sqrt{1 - \Phi\left(\frac{Y-a_2}{b_2}\right)} \\ &\quad + \frac{1}{2} \cdot \left(\Phi\left(\frac{Y-a_1}{b_1}\right)\right)^{-\frac{1}{2}} \cdot \left(\phi\left(\frac{Y-a_1}{b_1}\right) \cdot \frac{1}{b_1}\right) \cdot \left(\frac{1}{2} \cdot \left(1 - \Phi\left(\frac{Y-a_2}{b_2}\right)\right)^{-\frac{1}{2}} \cdot \left(-\phi\left(\frac{Y-a_2}{b_2}\right) \cdot \frac{1}{b_2}\right)\right)\end{aligned}$$

Bibliography

- Benjamini, Y., and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1), 289-300.
- Benjamini, Y., and Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, 29(4), 1165-1188.
- Böckenholt, U. (1989), "Analyzing Optima in the Exploration of Multiple Response Surfaces," *Biometrics*, 45, 1001-1008.
- Bonferroni, C.E. (1936), "Teoria statistica delle classi e calcolo delle probabilità," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3-62.
- Box, G. E., and Hunter, J. S. (1954), "A Confidence Region for the Solution of a Set of Simultaneous Equations with an Application to Experimental Design," *Biometrika*, 41, 190-199.
- Box, G., and Draper, N. (1986), *Empirical Model-Building and Response Surface*, Wiley: New York.
- Cahya, S. (2002), *Sampling Properties of Optimal Operating Conditions of Single and Multiple Response Surface Systems*, Thesis of the Pennsylvania State University.
- Del Castillo, E., and Cahya, S. (2001), "A Tool for Computing Confidence Regions on the Stationary Point of a Response Surface," *The American Statistician*, 55(4), 358-365.
- Del Castillo, E., Montgomery, D. C., and McCarville, D. R. (1996), "Modified desirability functions for multiple response optimization," *Journal of Quality Technology*, 28(3), 337-345.
- Derringer, G. C. (1994), "A Balancing Act: Optimizing a Product's Properties," *Quality Progress*, June 1994, 51-58.

- Derringer, G. C., and Suich, R. (1980), "Simultaneous Optimization of Several Response Variables," *Journal of Quality Technology*, 12(4), 214-219.
- Ding, R., Lin, D. K. J., and Peterson, J. J. (2003), *A Large-Sample Confidence Band for Multi-Response Ridge Path of a Desirability Functions*, Technical Report of the Pennsylvania State University.
- Khuri, A. I., and Cornell, J. A. (1987), *Response surfaces: designs and analyses*, Marcel Dekker, Inc.: New York.
- Gibb, R. D., Carter, W. H., and Myers, R. H. (2001), *Incorporating Experimental Variability in the Determination of Desirable Factor Levels*, Unpublished Manuscript.
- Govaerts, B., and Le Bailly de Tillegheem, C. (2005), *Distribution of Desirability Index in Multicriteria Optimization using Desirability Functions based on the Cumulative Distribution Function of the Standard Normal*, Discussion Paper 0531, Institut de Statistique, Univeristé catholique de Louvain, Louvain-la-Neuve, Belgium.
- Govaerts, B., and Le Bailly de Tillegheem, C. (2005), *Uncertainty propagation in multiresponse optimization using a desirability index*, Discussion Paper 0532, Institut de Statistique, Univeristé catholique de Louvain, Louvain-la-Neuve, Belgium.
- Le Bailly de Tillegheem, C., Trautmann, H., Weihs, C., and Govaerts, B. (2006), "Comparison of Multiresponse Optimization Methodologies to Take Into Account the Propagation of Models Uncertainty on the Desirability Index," *to appear in 2006*.
- Hahn, G. J., and Shapiro, S. S. (1969), *Statistical models in engineering*, John Wiley, Inc.: New York, 228-236.
- Harrington, E. C. (1965), "The Desirability Function," *Industrial Quality Control*, 21, 494-498.
- Kim, K., and Lin, D. (2000), "Simultaneous Optimization of Mechanical Properties of Steel by Maximizing Exponential Desirability Functions," *Applied Statistics*, 49, 311-325.
- Khuri, A. I., and Conlon, M. (1981), "Simultaneous Optimization of Multiple Responses Represented by Polynomial Regression Functions," *Technometrics*, 23, 363-376.
- Ko, Y. H., Kim, K. J., and Jun, C. H. (2005), "A New Loss Function-Based Method for Multiresponse Optimization", *Journal of Quality Technology*, 37(1), 50-59.
- Le Bailly de Tillegheem, C., Beck, B., Boulanger, B. and Govaerts, B. (2005) "A Fast Exchange Algorithm for Designing Focused Libraries in Lead Optimization," *Journal of Chemical Information and Computer Sciences*, 45(3), 758-767.

- von Mises, R. (1947), "On the Asymptotic Distribution of Differentiable Statistical Functions," *Annals of Mathematical Statistics*, 18, 309-348.
- Myers, R. H., Khuri, A. I. and Carter, W. I. Jr (1989), "Response Surface Methodology: 1966-1988," *Technometrics*, 31(2), 137-157.
- Myers, R. H., and Montgomery, D. C. (2002), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd Edition, Wiley: New York.
- von Neumann, J., and Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton NJ.
- Peterson, J. J., Cahya, S., and Del Castillo, E. (2002), *A General Approach to Confidence Regions for Optimal Factor Levels of Response Surfaces*, *Biometrics*, 58(2), 422-431.
- Stablein, D. L., Carter W. H. Jr., and Wampler, G. L (1983), "Confidence Regions for Constrained Optima in Response Surface Experiments," *Biometrics*, 39, 759-763.
- Steuer, D. (2000), *An Improved Optimisation Procedure for Desirability Indices*, Technical Report 27/00, SFB 475, Dortmund University.
- Trautmann, H., and Weihs, C. (2004), *Uncertainty of the Optimum Influence Factor Levels in Multicriteria Optimization Using the Concept of Desirability*, Technical Report 23/04, SFB 475, Dortmund University.
- Trautmann, H., and Weihs, C. (2006), *On the Distribution of the Desirability Index using Harrington's Desirability Function*, *Metrika*, 63, 207-213.
- Weber, H., and Weihs, C. (2003), *On the Distribution of the Desirability Index using Harrington's Desirability Function*, Forschungsbericht 2003/3, Statistics Department, Dortmund University.

Chapter 5

A fast exchanges algorithm for lead optimisation

This work has been published in the Journal of Chemical Information and Modeling, 45(3) pp 758-767, by C. Le Bailly de Tillegem, B. Beck, B. Boulanger, and B. Govaerts (2005)

5.1 Introduction

The process of drug discovery generally implies the following successive stages. Once a biological target has been defined, the screening of large collections of compounds is performed to identify *hits* *i.e.* compounds that interact with the biological target. Then *leads* are generated by chemically modifying the hits to improve chemical properties. Further chemical alterations of leads are performed to optimise criteria and convert leads into drug development candidates. Finally, preclinical and clinical trials confirm the characteristics of the new potential drug. Those different stages are summarised in Figure 5.1 and have been detailed in the Introduction of the thesis.

Combinatorial chemistry is widely used in this process (Gallop *et al.* 1994; Gordon *et al.* 1994; Floyd *et al.* 1999; Lam 1997; Agrafiotis *et al.* 2002; Leach. and Hann 2000; Venton and Woodbury 1999). This technique systematically combines a variety of molecular building blocks into a huge collection of compounds called *combinatorial library*. There is an increasing need of methods to design combinatorial libraries (Stanton *et al.* 2000; Terrett *et al.* 1995; Pickett *et al.* 2000).

A first possible objective is to span a wide range of chemical properties. This is known under the name *diverse library* or *exploratory library* (Agrafiotis 2002; Agrafiotis and Lobanov 2000; Zheng *et al.* 2000). It can be used in the hit identification stage, when

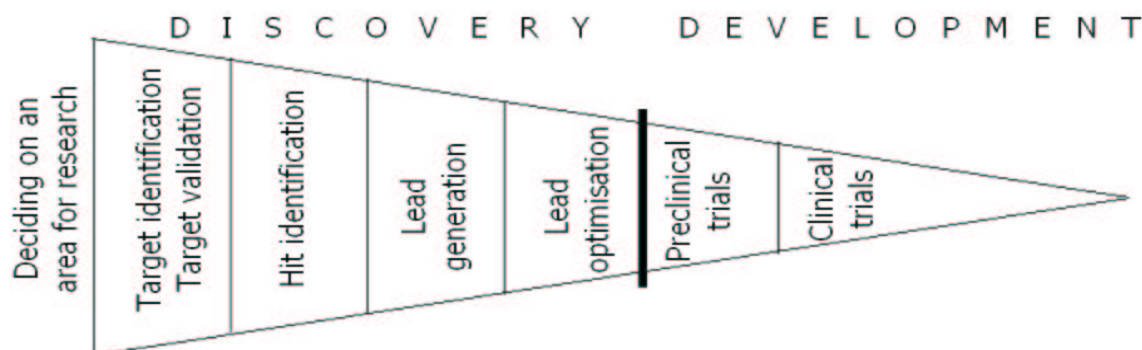


Figure 5.1: General scheme of drug discovery and development process.

chemists want to select active compounds as diverse as possible. The design of libraries using some diversity measurements is often discussed in literature (Gillet *et al.* 1999; 1997; Hassan *et al.* 1996; Agrafiotis 1997a; Cummins *et al.* 1996; Agrafiotis *et al.* 2000; Agrafiotis 1997b; Lewis *et al.* 1997).

A second possible objective is to obtain a library biased toward a specific target or optimising some chemical properties. This is known under the name *focused library* or *targeted library* (Agrafiotis 2002; Agrafiotis and Lobanov 2000; Zheng *et al.* 2000). It can be used in the lead optimisation stage. The design of focused library in the context of lead optimisation is the concern of this chapter.

Once a lead has been generated, chemists in pharmaceutical industries try to chemically modify the lead to obtain compounds having still better properties. They virtually divide the lead into N basic parts and set up a list of chemicals that could react on each part. Each chemical used in reaction with the lead is called a *reagent* (or *reactant*) and each set of reagents on a part of the lead is called an *R-group* (or *pool*) (Agrafiotis 2002; Agrafiotis and Lobanov 2000; Zheng *et al.* 2000).

Let N be the number of R-groups and N_i the number of reagents in the i^{th} R-group ($i = 1, 2, \dots, N$), $R_i = \{R_{ik}; k = 1, 2, \dots, N_i\}$. The number of possible compounds obtained by combining a reagent from each R-group is $M = \prod_{i=1}^N N_i$. They form the combinatorial library. The main objective of the chemists is to find the molecule that optimises some drugability properties.

Another more general objective may be to find a subset of $m > 1$ good molecules, either the m best molecules, or the best combinatorial sub-library of m molecules. Those best molecules should optimise some drugability properties (efficiency, toxicity, ...). A

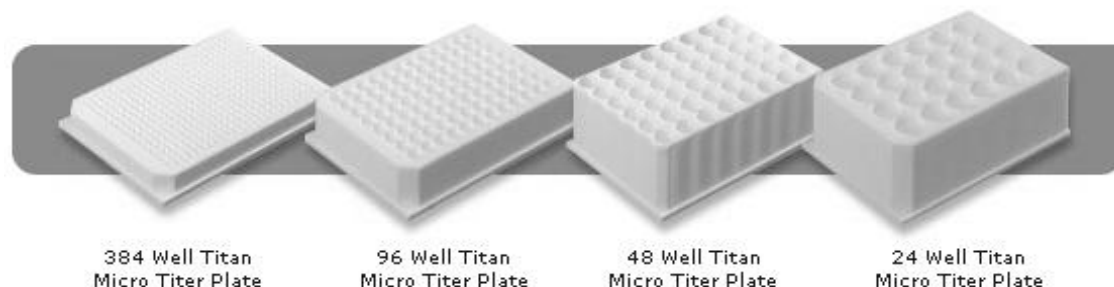


Figure 5.2: Standard 24, 48, 96 and 384 well PTFE micro titer plates, downloaded from [http : //www.radleys.com/pages/products/titan_plates.shtml](http://www.radleys.com/pages/products/titan_plates.shtml).

combinatorial sub-library is obtained by selecting in each R-group n_i reagents ($n_i \leq N_i$, $i = 1, 2, \dots, N$) and combining them in a full fashion to obtain a sub-library of reasonable size. The chemists fix the desired number of selected reagents in each R-group according to their knowledge (not necessarily the same number in each R-group) and according to the number of compounds in the resulting sub-library, $m = \prod_{i=1}^N n_i$. Indeed, the m selected compounds would then have to be tested in laboratories. The combinatorial nature of the compounds in the sub-library allows an easy synthesis as far as their number is not too large compared to used plates, as represented on Figure 5.2. The case $n_i = 1 \forall i = 1, 2, \dots, N$ corresponds to the search for the best molecule.

As there are $\binom{N_i}{n_i} = \frac{N_i!}{(N_i - n_i)!n_i!}$ different ways to select n_i reagents in the i^{th} R-group, the number of all possible sub-libraries is $\prod_{i=1}^N \binom{N_i}{n_i}$, which is often huge in practice. There is a clear need of methods to select the best sub-library among those possibilities. Part of existing techniques are borrowed from the field of operational research, based on simulated annealing algorithms or genetic algorithms (Zheng *et al.* 2000; Sheridan and Kearsley 1995; Venkatasubramanian *et al.* 1995; Singh *et al.* 1996; Weber *et al.* 1995; Gillet *et al.* 2002; Clark and Westhead 1996; Judson 1997).

This chapter presents a new algorithm based on exchanges (called *Weighted Exchanges Algorithm for Library Design*, abbreviated *WEALD*) to design focused library in lead optimisation. This iterative algorithm finds rapidly in the combinatorial library a sub-library of reasonable size that is composed of promising molecules that could be easily synthesised in laboratories thanks to their combinatorial nature.

Section 2 presents the library design as a multicriteria optimisation problem. QSAR models, desirability index and loss allow to quantify the quality of a sub-library and transform this multicriteria optimisation problem into a single objective optimisation. The prin-

ciples of *WEALD* are explained in depth in Section 3 and illustrated on the *antidepressant library* provided by Eli Lilly. *WEALD* is also compared on this library to two algorithms of the combinatorial chemistry literature, the *ultrafast* algorithm (Agrafiotis and Lobanov 2000) in Section 4 and the *Piccolo* algorithm (Zheng *et al.* 2000) in Section 5. The three algorithms are also compared on the *diabetes library* in Section 6. The good performance of *WEALD* is demonstrated in Section 7 on simulated combinatorial libraries of different dimensions. A modified version of *WEALD* is also implemented in Section 8 to select the $m \geq 1$ best compound(s) not in a combinatorial fashion.

5.2 Quantifying the quality of a set of m molecules

The *antidepressant library* (see Figure 4 in the Introduction) is composed of $N = 4$ R-groups and $M = \prod_{i=1}^N N_i = 12 \times 19 \times 12 \times 41 = 112\,176$ compounds. Chemists at Eli Lilly want to select $n_i = 3$ ($i = 1, 2, 3, 4$) reagents in each R-group to obtain a sub-library of $m = \prod_{i=1}^4 n_i = 3^4 = 81$ compounds that optimise eight properties of interest. Library design is a multicriteria optimisation problem (see Chapter 3).

A unique criterion has to be defined to quantify the quality of a sub-library. In the *WEALD* algorithm, exchanges between reagents are repeatedly performed to optimise this unique criterion. Figure 5.3 presents the different steps that we propose to transform the multicriteria optimisation problem into a single objective problem:

- Step 1:** Specific softwares compute some chemical descriptors for each molecule in the sub-library on the basis of the SMILES coding (Chapter 1, Section 3);
- Step 2:** The properties to be optimised are estimated on the basis of the computed descriptors by QSAR models built on historical data (Chapter 1, Section 4);
- Step 3:** The estimated properties are aggregated by a desirability index in $[0, 1]$ that quantifies the quality of each molecule in the sub-library (the higher, the better) (Chapter 3, Section 2);
- Step 4:** The desirability indexes of the m compounds in the sub-library, $D_1, D_2, \dots, D_m$, are aggregated by a loss that summarises the differences between the desirabilities of the m compounds and their ideal value 1: $Loss = \frac{1}{m} \sum_{j=1}^m [1 - D_j]^2$ (the smaller, the better).

The 112176 molecules of the *antidepressant library* are encoded using SMILES and chemical descriptors are computed using internal softwares at Eli Lilly. Bagging of regression trees is used to model the drugability properties of interest, $Y_1, Y_2, \dots, Y_8$, as functions of the chemical descriptors.

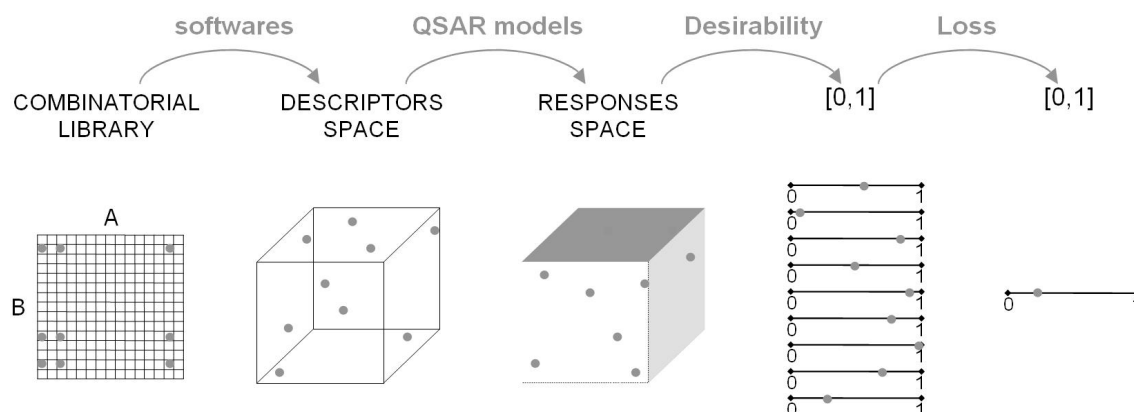


Figure 5.3: The different steps to quantify the quality of a sub-library by a loss value in $[0, 1]$.

The chemists are interested in two targeted receptors. Properties Y_1 and Y_2 measure respectively the binding to the first targeted receptor and the functional assay on the second targeted receptor. Both properties have to be maximised. On the other side, Y_3 , Y_4 , Y_5 and Y_6 measure the binding to four other undesirable receptors and those four criteria have to be minimised as the drug candidate has to be selective. Finally properties Y_7 and Y_8 are respectively the probability of nonmutagenicity (a mutagen is an agent that changes the genetic information of an organism) and the opposite of the metabolism rate (the speed to which a chemical compound is biochemically modified in living organisms or cells), both to be maximised. Those eight drugability properties were predicted for all the molecules in the *antidepressant library* (usually, those predictions are not available for the whole library as it is too huge). The graphs of Figure 5.4 displays those predictions for all the compounds in the combinatorial library.

Desirability functions map each predicted drugability property for a molecule in the interval $[0, 1]$ (see Section 2 of Chapter 3). Then the desirability functions are aggregated in one value between 0 and 1 using their geometric mean. This way a desirability value can be associated to each compound in the combinatorial library, as shown for some molecules of the antidepressant library on Figure 5.5. Any other desirability functions and desirability index can be equivalently used.

The desirability of expected properties (classical methodology) has been used in the next sections, but the expected desirability proposed in Chapter 4 can also be used. In this chapter, the quality of the QSAR models is not analysed and the reliability of the resulting predictions is not quantified. The integration of the stochastic optimisation proposed in the previous chapter in *WEALD* is detailed in the next chapter.

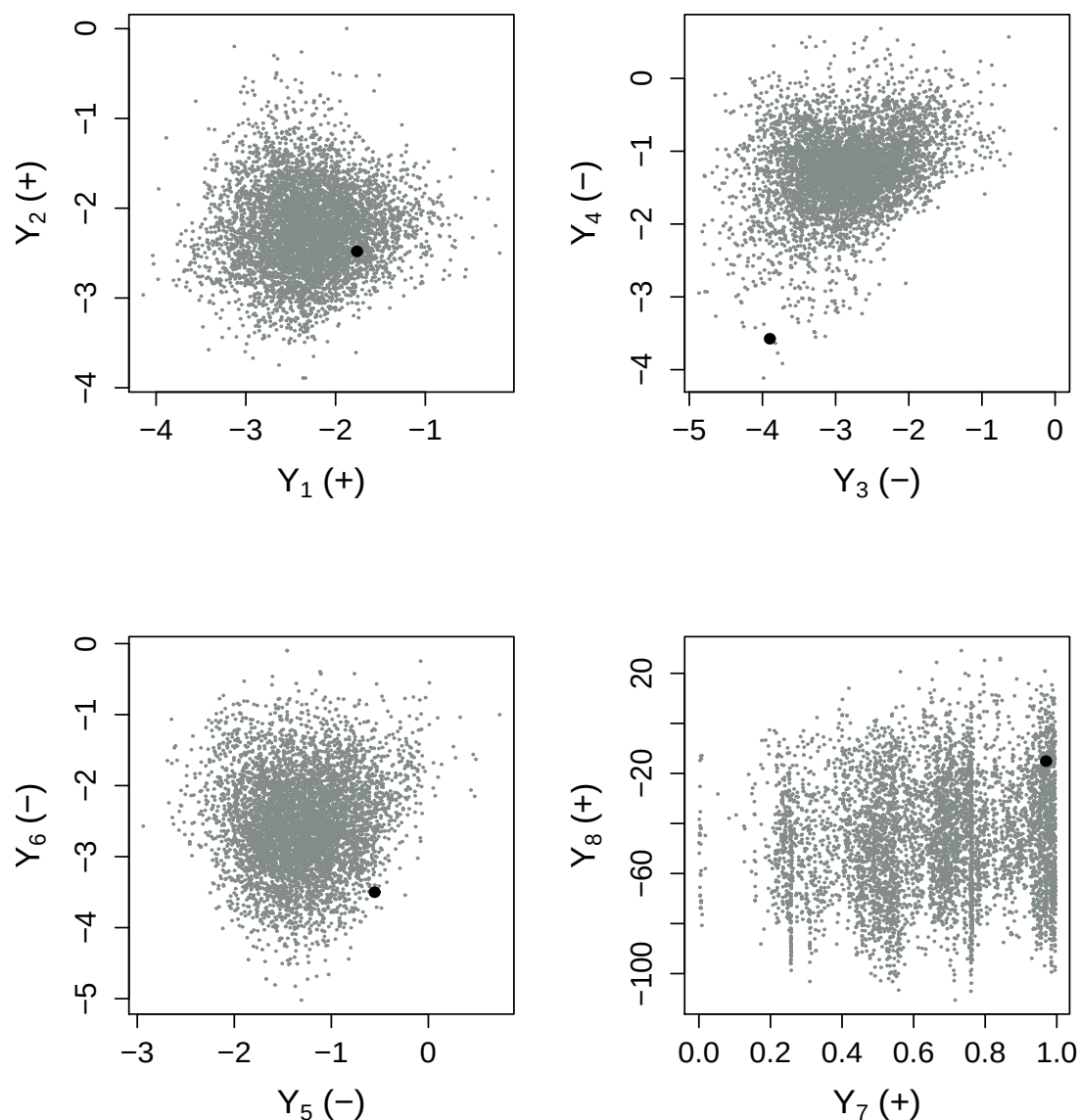


Figure 5.4: Values of the 8 properties $Y_1, Y_2, \dots, Y_8$ computed for the 112 176 compounds of the *antidepressant library*. Y_1 = binding to a targeted receptor; Y_2 = functional assay on another targeted receptor; Y_3, Y_4, Y_5 and Y_6 = binding to other receptors; Y_7 = probability of nonmutogenicity and Y_8 = - metabolism rate. Besides properties names, (+) stands for maximised property and (-) stands for minimised property. The dark circle represents the best molecule (desirability=0.6908).

N	R_A	R_B	R_C	R_D	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$	$\hat{Y}_6$	$\hat{Y}_7$	$\hat{Y}_8$	Desirability
1	a7	b16	c3	d3	-1.76	-2.48	-3.90	-3.57	-0.55	-3.50	0.97	-15.09	0.6908
2	a1	b3	c2	d8	-1.80	-2.65	-2.39	-1.45	-0.53	-3.05	0.93	-31.89	0.6023
3	a11	b1	c8	d3	-1.29	-1.59	-1.89	-0.69	-0.92	-2.06	1.00	11.63	0.5956
4	a8	b15	c8	d18	-1.20	-2.54	-2.88	-1.14	-1.13	-2.27	0.77	-10.90	0.5910
5	a3	b2	c1	d13	-2.48	-2.34	-2.73	-1.31	-0.29	-2.52	0.65	-33.49	0.5406
6	a2	b14	c4	d3	-2.13	-2.46	-2.56	-1.76	-1.47	-2.78	0.74	-39.79	0.5129
7	a8	b6	c6	d30	-2.08	-2.86	-2.94	-1.31	-1.22	-2.29	0.76	-66.83	0.4891
8	a10	b1	c7	d40	-1.60	-2.25	-2.17	-0.84	-1.62	-2.66	0.86	-72.06	0.4732
9	a6	b19	c5	d17	-2.13	-2.56	-3.37	-0.92	-2.15	-1.64	0.60	-22.29	0.4498
10	a6	b17	c7	d37	-1.77	-2.26	-3.21	-1.59	-2.24	-2.72	0.57	-66.78	0.4310
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 5.5: Ten compounds of the combinatorial *antidepressant library* with the corresponding estimated drugability properties and the corresponding desirability ($D(\hat{E}[\mathbf{Y}|\mathbf{x}])$).

The *antidepressant library* is used to analyze the performance of *WEALD* and to compare it to other algorithms of the combinatorial literature.

5.3 A Weighted Exchanges Algorithm for Library Design (*WEALD*)

5.3.1 Principles

This section proposes a new method to select reagents in R-groups to obtain compounds that optimise some properties. The algorithm is iterative as most of the algorithms in this domain.

Firstly, n_i reagents are selected at random in the i^{th} R-group (for $i = 1, 2, \dots, N$) and combined to form an initial sub-library with $m = \prod_{i=1}^N n_i$ compounds. Then exchanges between reagents are performed on the current sub-library to improve the quality of its compounds, *i.e.* decrease the loss. The principle of exchanges was introduced in the context of D-optimal experimental designs by Fedorov (1972). This principle is applied to select a reagent in the current sub-library and replace it by a new one. But exchanges are not done in a naive way at random. A weighting method summarises all the information accumulated till each iteration to guide the reagents exchanges. This is explained with more details in the next section.

5.3.2 Exchanges methodology

In an exchange, one reagent is ejected from the current sub-library and replaced by another one of the same R-group to decrease the loss of the current sub-library.

First the reagent to be ejected out of the current sub-library is selected. For each current selected reagent, the mean of the desirability associated to the compounds in the sub-library containing this reagent is computed. The worst one is ejected, *i.e.* the one with the smallest average desirability.

In a second step, a new reagent is selected to replace the ejected one in the set of reagents of the same R-group that are not yet in the current sub-library. This choice is done randomly according to probabilities of selection associated to each reagent. At the beginning of the algorithm, in each R-group, all reagents have the same probability to enter the sub-library. Then, at each iteration, probabilities of selection are adapted using all the desirabilities already computed according to the following rules:

Rule 1: If a reagent has never been tested for an exchange, it is associated to a selection probability of 1.

Rule 2: If a reagent has already been tested for an exchange, it is associated to a probability of selection which is the ratio between the average desirability of all the compounds already explored that contains this reagent and the maximum desirability computed till that step.

Rule 3: If the average desirability of the compounds obtained using a reagent is lower than the 10th percentile of all the desirabilities computed till that step and if more than 1% of the compounds in the whole combinatorial library have been explored, the probability of selection for this reagent will be set at 0.

All these selection probabilities are of course standardised to sum to 1 in each R-group. Rule 1 insures that a reagent not explored yet will have the highest selection probability for next exchanges. According to rule 2, exchanges with reagent that provides good compounds will be preferentially tested. And with rule 3, a bad reagent will definitely be excluded from the sub-library and never more tried for an exchange.

Finally, the exchange between the entering reagent drawn randomly according to selection probabilities and the ejected worst one is applied if the loss of the new sub-library is lower than the loss of the current sub-library. If it is not the case, another reagent of the same R-group is tried to replace the ejected one and if no exchange with the ejected reagent decreases the loss it is kept till the convergence of the algorithm.

The algorithm ends when there is no more possible exchange with reagent having strictly positive selection probability.

The different steps of the *WEALD* algorithm are summarised next:

Step 0: - Initialise a sub-library at random $S^0 = S_1^0 \times S_2^0 \times \dots \times S_N^0$ where $S_i^0 = \{R_{ik}; k = 1, 2, \dots, n_i\}$ is a set of selected reagents of the i^{th} R-group, R_i , and record its loss L^0 ;
- Set the best sub-library $S^b = S^0$ and its loss $L^b = L^0$;
- Initialise the selection probabilities for each reagent, R_{ij} , by $w_{ij} = \frac{1}{N_i}$ ($j = 1, \dots, N_i$ and $i = 1, \dots, N$) ;
- Go to step 1.

Step 1: If $w_{ij} = 0 \forall i$ and $\forall j$ stop and output S^b , else execute step 2.

Step 2: - Select the worst reagent of S^b , denoted R_{IK} ;
- Select the set of possible reagents for an exchange, denoted $R^* = R_I \setminus S_I^b = \{R_{It}, t = 1, \dots, N_i - n_i\}$;
- Go to step 3.

Step 3: - If $R^* = \emptyset$ or if $w_{It} = 0 \forall t$, keep R_{IK} in S^b till convergence and never try it again for an exchange; update the selection probabilities; return to step 1.
- else generate a trial sub-library S^t by replacing R_{IK} in S^b by a reagent of R^* , denoted R_{IT} , chosen according to the selection probabilities; compute its loss L^t ; go to step 4;

Step 4: - If $L^t < L^b$ set $S^b = S^t$ and $L^b = L^t$; update the selection probabilities; return to step 1;
- If $L^t \geq L^b$ set $R^* = R^* \setminus R_{IT}$ and return to step 3;

5.3.3 Application on the *antidepressant library*.

The *WEALD* algorithm (implemented in the R-software) is applied on the *antidepressant library* ($12 \times 19 \times 12 \times 41$) to select $n_i = 3$ reagents in each R-group. By combining them, $m = 81$ compounds are obtained and they have to optimise 8 criteria of interest. As described in Section 2, a desirability index is used to quantify the quality of compounds and a loss value to measure the quality of a sub-library. As *WEALD* is not deterministic and as it may depend on the starting sub-library, it is applied 100 times, each time with a different random initial sub-library. This section analyses the rate of convergence of the *WEALD* algorithm and the quality of the sub-libraries obtained.

The first important result is that the *WEALD* algorithm converges rapidly. On average, 2613 desirabilities are computed before the algorithm stops. This represents only 2.33% of the whole combinatorial library. In addition, this fast rate of convergence is quite stable,

whatever is the initial sub-library. Indeed, each time *WEALD* is applied, the convergence is reached by computing between 2220 and 3114 desirabilities (between 2% and 2.8% of the whole library). This can be seen on the histogram (1) of Figure 5.6 representing the distribution of the numbers of desirability computations for the 100 runs of the algorithm.

The second important result is that the quality of the sub-libraries obtained with the *WEALD* algorithm is quite high. Indeed, losses of the 100 sub-libraries are quite small, on average 0.1382. They are also stable if we analyse the 100 losses represented on histogram (2) of Figure 5.6.

The desirabilities of the compounds in the 100 sub-libraries are high, on average 0.6291, which is the 99,71th percentile of the desirabilities of the whole combinatorial library. One can conclude by analysing those numbers or comparing histograms (3) and (4) of Figure 5.6 that the *WEALD* algorithm provides a sub-library containing most of the best compounds (highest desirabilities) of the initial combinatorial library.

Those 100 runs of the *WEALD* algorithm allow to conclude that it converges rapidly, provides a sub-library with the most promising compounds whatever is the initial sub-library. In real cases applications, the *WEALD* algorithm can be applied only one time as it is stable. Nevertheless, to prevent a local minimum it can be recommended to repeat the application a few times (2 or 3) with other initial sub-libraries.

5.3.4 What happens when applying the *WEALD* algorithm without using selection probabilities?

The principles of the *WEALD* algorithm are applied on the *antidepressant library* but without using selection probabilities (algorithm called *Exchanges Algorithm for Library Design - EALD*) to confirm that those proposed weights help finding more rapidly a sub-library ³⁴ containing compounds that optimise the eight properties of interest.

Both algorithms, *WEALD* and *EALD*, are applied on 100 different initial sub-libraries, and the rates of convergence and the quality of the obtained sub-libraries are compared. The results of those simulations are summarised on Figure 5.7. We conclude that using probabilities of selection accelerates the convergence because *WEALD* only computes on average 2651 desirabilities which is only 2,36% of the whole combinatorial although *EALD* computes on average 3553 desirabilities which corresponds to 3,17% of the whole combinatorial library. Moreover, *WEALD* provides slightly better sub-libraries than *EALD* according to the average losses that are smaller with *WEALD* (0.1372) than with *EALD* (0.1392) or according to the desirabilities of the selected compounds that are a little higher with *WEALD* (0.6304) than with *EALD* (0.6279).

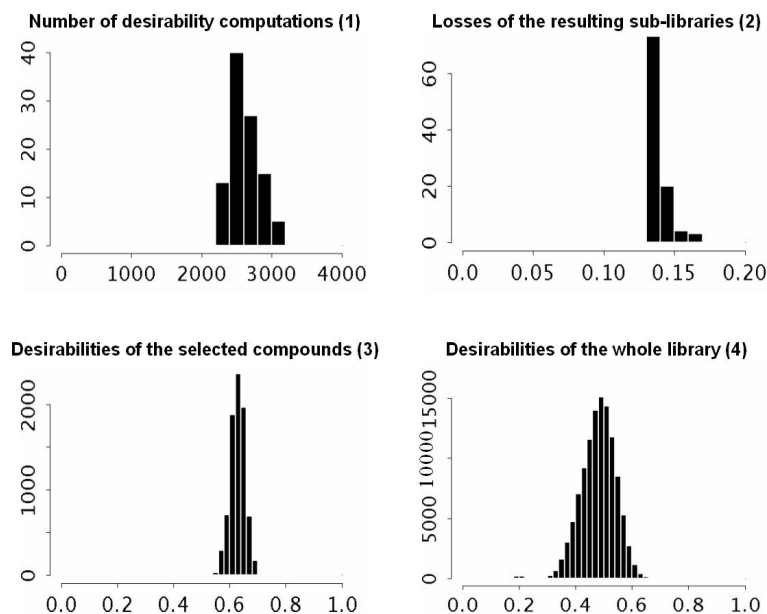


Figure 5.6: Results of the application of the *WEALD* algorithm on 100 different initial sub-libraries. (1) Number of desirability computations, (2) Losses of the final sub-libraries, (3) Desirabilities of the compounds in the final sub-libraries and (4) Desirabilities of the 112176 compounds in the whole library.

In conclusion, using the probabilities of selection to guide exchanges allows a faster convergence through better sub-libraries.

5.4 Comparison with the *Ultrafast* algorithm

Agrafiotis and Lobanov (2000) proposed an iterative algorithm to design focused library. As most of the algorithms in this domain, it starts with a randomly chosen sub-library and its quality is evaluated by a single value that summarises the properties to optimise. Then it optimises each R-group in sequence until no further improvement is possible.

More precisely the different steps are summarised as follows:

Step 0: - Initialise a sub-library at random: $S^0 = S_1^0 \times S_2^0 \times \dots \times S_N^0$ where $S_i^0 = \{R_{ik}; k = 1, 2, \dots, n_i\}$ is a set of selected reagents of R-group R_i and record its loss L^0 ;
- Set the best sub-library $S^b = S^0$ and its loss $L^b = L^0$;
- Go to step 1.

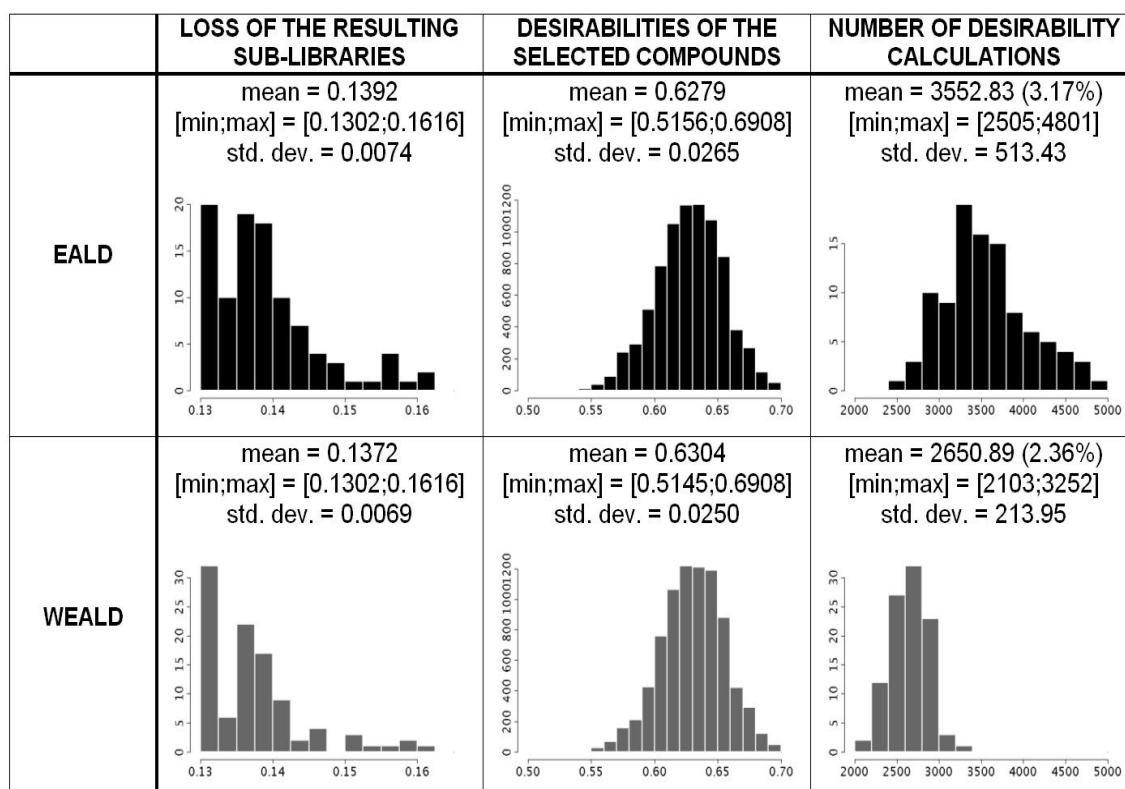


Figure 5.7: Results of the application of the *EALD* algorithm (without selection probabilities) and the *WEALD* (with selection probabilities) algorithm on 100 initial sub-library.

Step 1: Perform steps 2 to 4 for $i = 1, \dots, N$ (*i.e.* for each R-group R_i) then go to step 5;

Step 2: Perform step 3 for $j = 1, \dots, N_i$ (*i.e.* for each reagents R_{ij} of the i^{th} R-group R_i)

Step 3: Evaluate the sub-library $S = S_1^b \dots \times S_{i-1}^b \times R_{ij} \times S_{i+1}^b \dots \times S_N^b$ and record its loss L_{ij} .

Step 4: Sort the reagents in the i^{th} R-group R_i in ascending order of loss L_{ij} and select the n_i reagents with the smallest losses: $S_i^t = \{R_{ik}; k = 1, 2, \dots, n_i\}$

Step 5: Evaluate the trial sub-library $S^t = S_1^t \times S_2^t \times \dots \times S_N^t$ and record its loss L^t . If $L^t < L^b$ set $S^b = S^t$ and $L^b = L^t$ then return to step 1, else stop and output S^b .

This is a systematic search through the solutions space: for each R-group R_i , all possible reagents are combined with the current selected ones of the other R-groups and the n_i best reagents are selected.

The *WEALD* algorithm and the *Ultrafast* algorithm are compared on the *antidepressant library* to select three reagents in each R-group. Both algorithms are applied on 50 different initial sub-libraries chosen at random. By analysing the sub-libraries obtained with each algorithm, one can conclude that the *WEALD* algorithm converges 3 times faster than the *Ultrafast* algorithm, but the quality of the sub-libraries is nevertheless slightly higher (losses of the obtained sub-libraries smaller and desirabilities of the selected compounds in these sub-libraries higher). In addition, the *WEALD* algorithm is less sensitive to the choice of the initial sub-library (standard deviations smaller). Those results are summarised on Figure 5.8.

Figure 5.9 shows the evolution of the loss according to the number of desirability computations for the two algorithms applied on the same initial sub-library. The losses at the end of the two methods are quite the same, but the rate of convergence of the *WEALD* algorithm is clearly higher than the *Ultrafast* algorithm. With the *Ultrafast* method, only 6 iterations are performed but for each one, $(12-3) \times (3^3) + (19-3) \times (3^3) + (12-3) \times (3^3) + (41-3) \times (3^3) = 1944$ desirabilities are first computed to select the three best reagents in each R-group and then $3^4 = 81$ desirabilities are computed to evaluate the quality of the sub-library obtained by combining the three best reagents in each R-group. At each iteration around 2000 desirabilities have to be computed with the *Ultrafast* algorithm although with the *WEALD* algorithm, this number could be only $1 \times 3^3 = 27$ if only one exchange is tried or $N_e \times 3^3$ if N_e exchanges are tried before having an exchange that decreases the loss.

A drawback of the *Ultrafast* algorithm is that the desirabilities computed at each iteration are entirely neglected when performing next iterations. This is not the case with

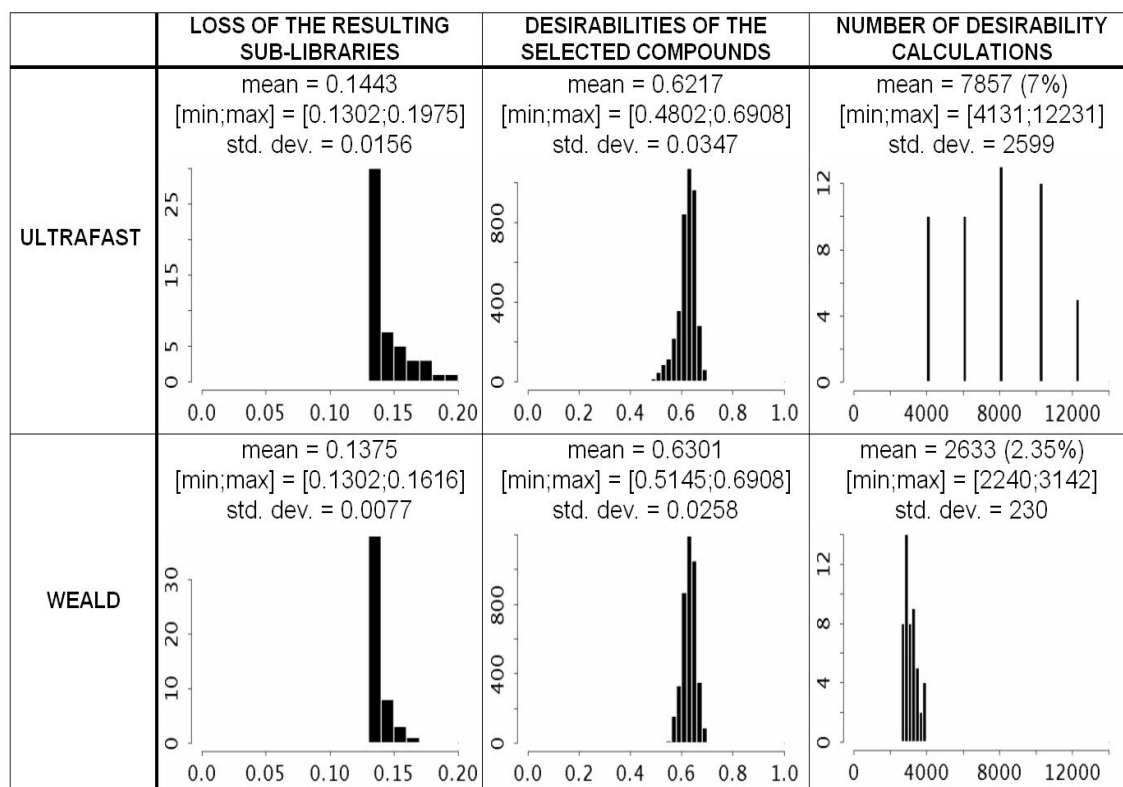


Figure 5.8: Results of the application of the *Ultrafast* algorithm and the *WEALD* algorithm on 50 different initial sub-library chosen at random.

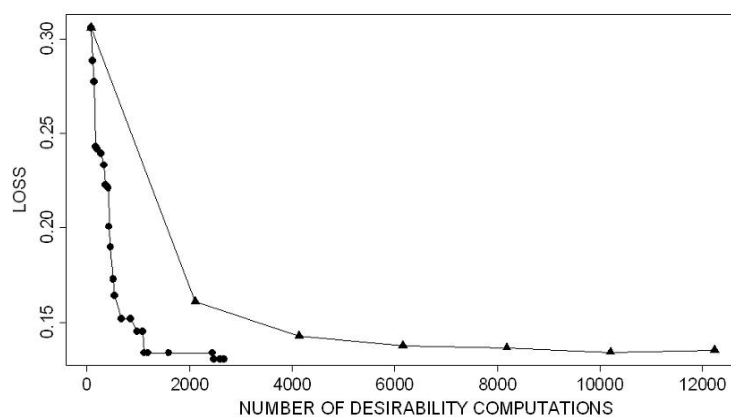


Figure 5.9: Loss as a function of the number of desirability computations when applying the *Ultrafast* algorithm (triangles) and the *WEALD* algorithm (circles) on the same initial sub-library.

WEALD that uses computed desirabilities to guide exchanges by using the selection probabilities as explained in Section 3.2. In addition, with those selection probabilities, when *WEALD* has converged, one do not need to perform all possible exchanges to see that, indeed, the algorithm has converged because part of the reagent are associated to a selection probability of 0.

5.5 Comparison with the *Piccolo* algorithm

Simulated annealing is a well known iterative method to solve discrete and combinatorial optimisation problems. Zheng *et al.* (2000) suggest to apply its principles to solve the problem of focused library design by also using a single criterion to summarise the properties of interest.

Let t_0 , r_0 and μ be respectively the initial temperature, the number of iterations in an annealing series and the temperature reducing factor, the general framework of the simulated annealing algorithm is the following:

Step 0: - Generate an initial sub-library S^0 randomly and record its loss L^0 ;

- Set $t = t_0$, $r = r_0$ and flag = False;
- Set the current sub-library $S^c = S^0$ and its loss $L^c = L^0$;
- Set the best sub-library $S^b = S^0$ and its loss $L^b = L^0$;
- Go to step 1.

Step 1: Generate a trial sub-library S^t by perturbing the current one S^c (*cfr* below). Calculate $\Delta L = L^t - L^c$. If $\Delta L < 0$ then execute step 3; otherwise execute step 2.

Step 2: Compute $P = e^{-\Delta L/t}$ and compare it with a random value y from a uniform distribution in $[0, 1]$. If $y < P$ then execute step 4; otherwise execute step 3.

Step 3: Set $S^c = S^t$. If $L^b \geq L^t$ then set $S^b = S^t$. If $\Delta L < 0$ then set flag = True.

Step 4: Set $r = r - 1$. If $r > 0$ then return to step 1.

Step 5: If flag = True, then set flag = False, $r = r_0$, $t = t\mu$ and repeat step 1; otherwise stop and return S^b .

The perturbation scheme used when applying those principles to design focused library is also based on exchanges. Zheng *et al.* propose to first choose the R-group with a probability proportional to the average of $N_i/\sum_{i=1}^N N_i$ (the number of reagents in the i^{th} R-group *vs* the total number of reagents) and $n_i/\sum_{i=1}^N n_i$ (the number of selected reagents in the i^{th} R-group *vs* the total number of selected reagents). In this chosen R-group, they use a

uniform random sampling approach to select the reagent that enters the sub-library and the one that is ejected.

Three parameters must be fixed to start the *Piccolo* algorithm: t_0 , r_0 and μ . The initial temperature t_0 is generally determined to ensure that the initial probability for accepting bad moves is approximately equal to a prescribed value P_0 . We choose t_0 such that $P_0 = 0.5$. For the number of iterations in an annealing series, r_0 , and the temperature reducing factor, μ , 6 different combinations of those parameters are tested: $\mu = 0.6$ or 0.9 and $r_0 = 12$, 20 or 41 (the minimum, the mean or the maximum number of reagents per R-group).

The *WEALD* algorithm and the *Piccolo* algorithm with the six different combinations of parameters are applied on 30 different initial sub-libraries. The extreme results are displayed on Figure 5.10.

On the quality point of view, the smallest losses and the highest desirabilities are obtained with $r_0 = 41$ and $\mu = 0.9$. But the sub-libraries obtained with *WEALD* algorithm are slightly better with less desirability computations. This may be due to the fact that, thanks to the selection probabilities used in the *WEALD* algorithm, sub-libraries are faster guided through sub-libraries with compounds having good properties than the *Piccolo* algorithm that uses random sampling approach to select reagents that enter the sub-library.

On the convergence rate point of view, the smallest number of desirability computations was reached with $r_0 = 12$ and $\mu = 0.6$: less than 1% of the desirabilities of the whole library had been calculated. It is faster than the *WEALD* algorithm but too fast to give good results: the average loss of the 30 sub-libraries is 0.1723 against 0.1375 with *WEALD* and the average desirability of the compounds in the 30 sub-libraries is 0.5868 against 0.6301 with *WEALD*. In addition, *Piccolo* algorithm is unstable, suffering from the random perturbation scheme. There exist more intelligent versions of simulated annealing algorithm, like the one of Nelder and Mead (1965), that could reduce this sensitivity.

We conclude that, on the *antidepressant library*, *WEALD* has the best quality/price (loss/number of desirability computations) ratio, as it can also be seen on Figure 5.11 representing the evolution of the loss according to the number of desirability computations for the two algorithms applied on the same initial sub-library.

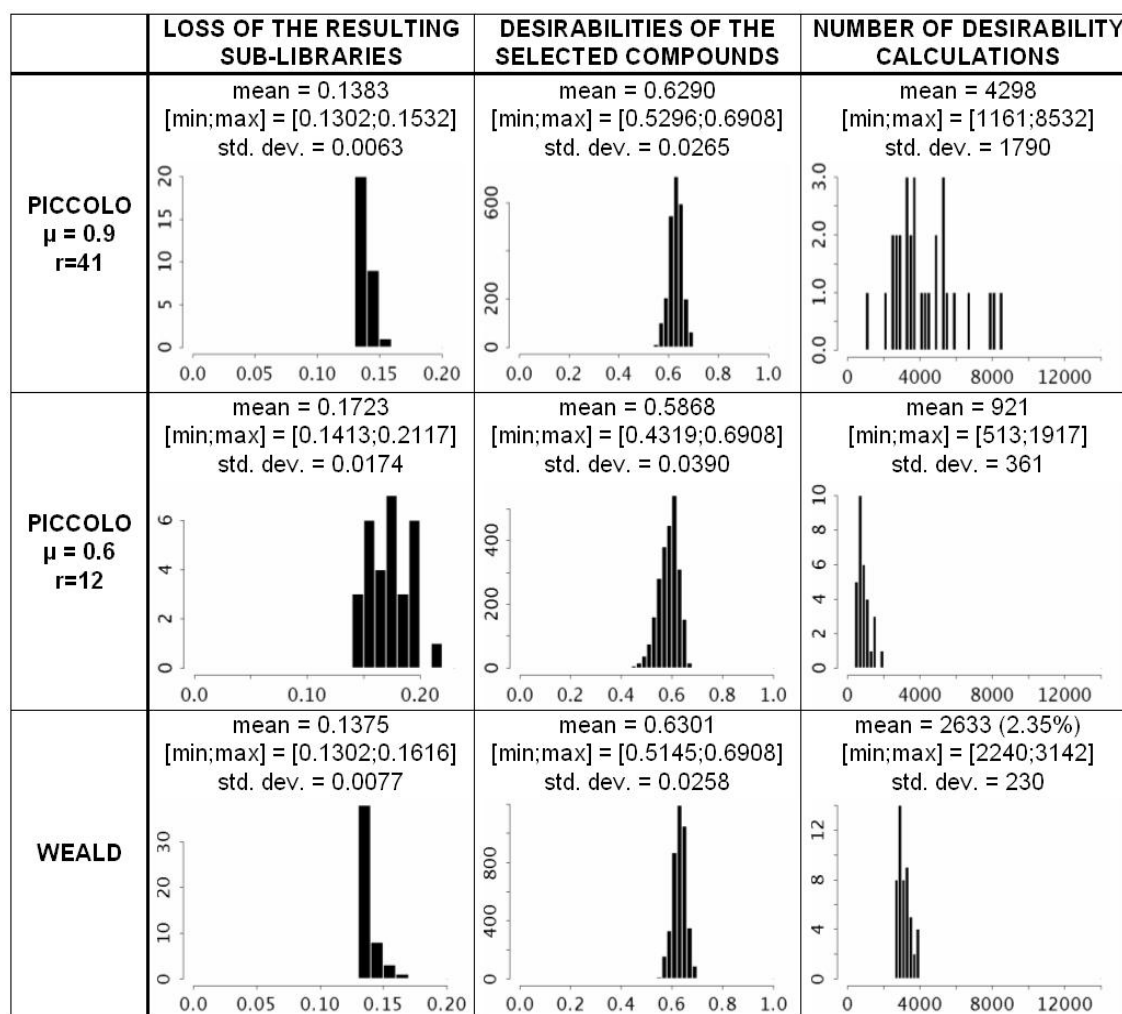


Figure 5.10: Results of the application of the *Piccolo* algorithms with the two extremes parameters r_0 and μ combinations ($r_0 = 41$ and $\mu = 0.9$; $r_0 = 12$ and $\mu = 0.6$) and the *WEALD* algorithm on 30 initial sub-libraries.

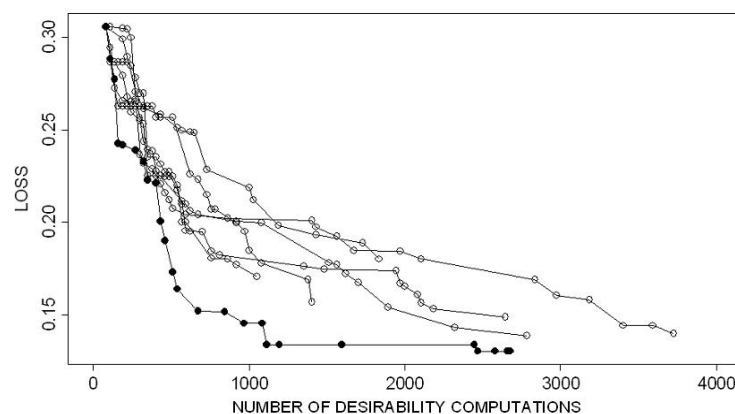


Figure 5.11: Loss as a function of desirability computations when applying the *Piccolo* algorithm (empty circles) with the 6 parameters combinations ($\mu = 0.6$ or 0.9 and $r_0 = 12, 20$ or 41) and the *WEALD* algorithm (dark circles) on the same initial sub-library.

5.6 Algorithms comparison on a second combinatorial library

The same conclusions are verified if the three algorithms, *Ultrafast*, *Piccolo* and *WEALD*, are applied on another real combinatorial library. This second library has been built up at Eli Lilly in the context of the development of a new molecule acting on regulation of glucose in type II diabetes (the *diabetes library*). This combinatorial library is composed of three R-groups, with respectively 47, 50 and 47 reagents leading to $M = 47 \times 50 \times 47 = 110\,450$ possible compounds. The three algorithms are used to select a combinatorial sub-library with $m = 5 \times 5 \times 5 = 125$ compounds optimising ten criteria of interest, $Y_1, Y_2, \dots, Y_{10}$.

In the same context, two molecules are well known and chemists want to obtain a new molecule that has a similar structure as the first one but a completely different structure as the second one. Properties Y_1 and Y_2 are appropriate distance measures and have to be respectively minimised and maximised. Properties Y_3 and Y_4 measure binding and functional assay on a targeted receptor and have both to be maximised. To quantify selectivity, properties Y_5, Y_6, Y_7, Y_8 and Y_9 measure the binding to five other undesirable receptors. Finally property Y_{10} is the opposite of the metabolism rate and has to be maximised. The distances Y_1 and Y_2 are based on chemical descriptors and are computed although properties $Y_3, Y_4, \dots, Y_{10}$ are predicted using QSAR models developed on historical data. The multicriteria optimisation is converted to a single objective optimisation using a similar desirability index as the one used for the *antidepressant library*.

WEALD converges nearly 4 times faster than *Ultrafast* and 3 times faster than *Piccolo* (with parameters $r_0 = 300$ and $\mu = 0.95$), with a number of desirability computations of 3549, 13 830 and 9890 respectively. *WEALD* provides also better sub-libraries than *Ultrafast* with average desirability for the selected compounds of 0.5815 and 0.5309 respectively. With a higher number of desirability computations, *Piccolo* provides slightly better sub-libraries with an average desirability of 0.5990 (the 99,64th percentile of desirabilities distribution) against 0.5815 with *WEALD* (the 99,53th percentile of desirabilities distribution).

Figure 5.11 shows the evolution of the loss according to the number of desirability computations for the three algorithms applied on the same initial sub-library that has a loss of 0.689. The losses at the convergence of *Piccolo* and *WEALD* are quite the same, but the rate of convergence of *WEALD* is clearly higher than *Piccolo*. One can also see that *WEALD* provides faster a better sub-library than *Ultrafast*.

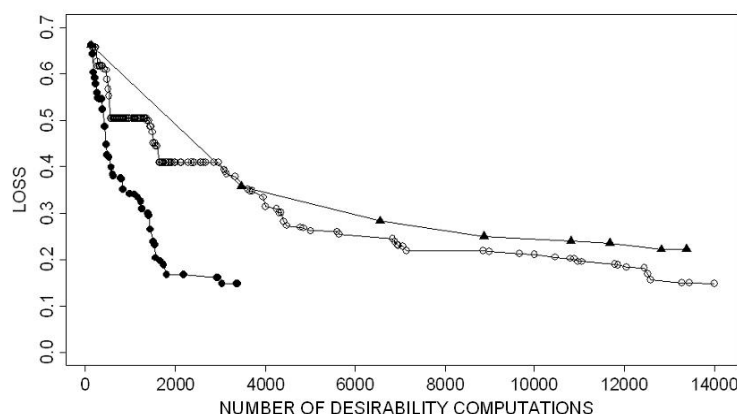


Figure 5.12: Loss as a function of the number of desirability computations when applying the *WEALD* algorithm (dark circles), the *Ultrafast* algorithm (triangles), and the *Piccolo* algorithm (empty circles) on the same initial sub-library.

5.7 Simulations with different combinatorial library dimensions

The good performance of the *WEALD* algorithm is only demonstrated through the previous sections by the mean of the *antidepressant library* and the *diabetes library*. This section compares the three algorithms on other combinatorial libraries to check if the *WEALD* algorithm has still the best quality/price ratio. More precisely, this section addresses the

following question: what is the impact of the combinatorial library dimensions (the number of R-groups, the number of reagents, the number of selected reagents...) on the performance of the three algorithms? A set of simulations is performed to check that *WEALD* insures a good balance between the loss of the resulting sub-library and the corresponding number of desirability computations to attain it whatever the dimensions of the whole combinatorial library are.

Simulations are performed on a set of variations of the *antidepressant library* used in sections 2 to 6. The four next factors vary to obtain different combinatorial library dimensions:

Nreag: the average number of reagents per R-group. It takes three values, 10, 30 or 100, by resampling reagents of the *antidepressant library*.

MaxMin: the ratio of the maximum and minimum number of reagents per R-group. It takes two values: 1 or 3.

N: the number of R-groups. It takes three values: 4, 3 (by choosing at random one of the 4 R-groups of the *antidepressant library* and fixing at random a reagent in this R-group) or 2 (by choosing at random two of the 4 R-groups of the *antidepressant library* and fixing at random a reagent in each of this two R-groups).

Nselec: the number of selected reagents per R-group. It takes values 2, 3, 4 or 5.

A complete factorial design was used and for each of the $3 \times 2 \times 3 \times 4 = 72$ combinations of factors 100 initial sub-libraries are initialised at random and the three algorithms applied.

By analysing the average loss and the average number of desirability computations for each of the 72 cases, one can conclude that *WEALD* converges always faster than the *Ultrafast* and the *Piccolo* algorithms. The *WEALD* algorithm and the *Piccolo* algorithm reach very similar losses (same quality of resulting sub-libraries), smaller than losses attained with the *Ultrafast* algorithm.

Those simulations are also used to build a prediction model of the number of computed desirabilities, *Ndes*, as a function of the four factors, *Nreag*, *MaxMin*, *N* and *Nselec*. A full second-order regression model is fitted on the data using the logarithm of *Ndes* as the response and non-significant effects are removed. The following model is finally obtained: $\ln(Ndes) = 2.044 + 0.026 \times Nreag - 0.393 \times Nselec + 0.530 \times N + 0.329 \times Nselec \times N$. We conclude that, as $N \geq 2$, all three factors, *Nreag*, *N* and *Nselec* have a positive effect on the number of desirability computations. We also conclude that *MaxMin*, the ratio of the maximum and minimum number of reagents per R-group, seems to have no significant

effect on the number of computed desirabilities.

As R^2 is 0.950 (Adjusted R^2 is 0.947), one may hope to use this model for prediction when applying the *WEALD* algorithm to solve a new combinatorial library design optimisation.

5.8 *WEALD* as a tool to search for the best compounds in a combinatorial library

In the previous sections, the three presented algorithms focused on the selection of combinatorial sub-libraries that optimise some drugability properties. The combinatorial nature of the selected compounds makes the synthesis in laboratories directly in plates easier, but this is not always a priority. One may also want to select the $m \geq 1$ best compound(s) in a sub-library whatever reagents are used. In such context, the *WEALD* algorithm may also be used after some slight modifications.

Three points are adapted to this more general optimisation problem:

Initialisation: m compounds are selected at random in the set of all possible compounds.

Choice of the ejected reagent: the worst compound in the current sub-library is first selected (the one with the smallest desirability) and one of its reagents is selected at random to be ejected.

Selection probabilities: the selection probabilities are computed as before except that probabilities are not any more set at zero when a reagent is considered as “bad” (see rule 3 of Section 3.2).

This adaptation of the *WEALD* algorithm is applied on the *antidepressant library* to select a sub-library with $m = 100$ best compounds. Hundred simulations were performed, each time with a different initial sub-library. We conclude that, by computing on average 6937 desirabilities (6.18% of the whole library), the sub-library obtained at the end of the algorithm contains on average 80 of the 100 best compounds. Those numbers are averaged over the 100 runs.

For one sub-library provided by this adapted *WEALD* algorithm, Figure 5.13 shows the rank of the selected compounds in the full library. This shows that the 100 selected compounds are in the set of the 120 best compounds and around 80 of the 100 best compounds were found.

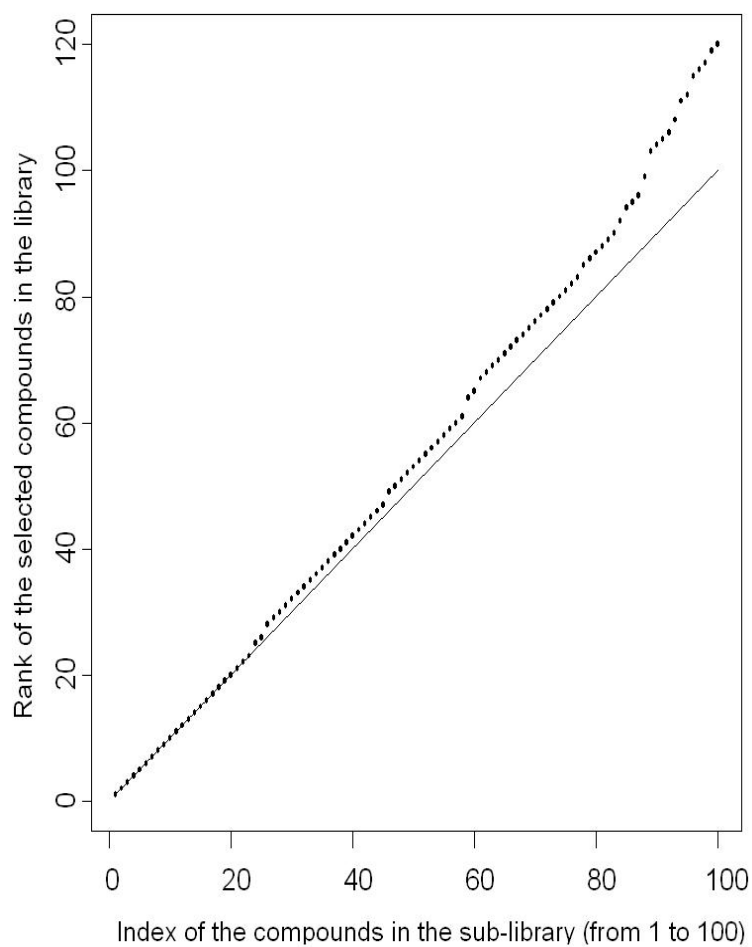


Figure 5.13: WEALD applied to select the 100 best compounds: the points are the ranks of the selected compounds (indexed by numbers from 1 to 100) in the whole combinatorial library and the straight line is the ideal ranks (1 to 100).

5.9 Conclusion

Combinatorial chemistry has revolutionised the drug discovery process. It can be used in screening for a hit or a lead, as well as in the lead optimisation stage. It allows to generate a huge compounds library by systematically combining building blocks composed of reagents.

The *WEALD* algorithm is a new iterative method based on exchanges to explore a combinatorial library and select a sub-library of reasonable size that optimise druglike properties. The *WEALD* algorithm can be used to obtain either a combinatorial sub-library containing compounds generated by combinations of a selected reagents subsets or either a sub-library composed of the m best compounds whatever reagents are used. With both objectives, the same methodology is applied to summarise the criteria to optimise: the quality of a molecule is quantified by a desirability index and the quality of a sub-library is quantified by a corresponding loss value.

The *WEALD* algorithm is compared to two other algorithms developed for library design, the *Ultrafast* algorithm of Agrafiotis and Lobanov (2000) and the *Piccolo* algorithm of Zheng *et al.* (2000). The three algorithms are applied on two real combinatorial libraries, the *antidepressant library* and the *diabetes example*, to select a combinatorial sub-library. The *WEALD* algorithm reveals to have the best quality/price ratio: it converges rapidly, exploring only a small part of the whole combinatorial library (2.33% with the *antidepressant library* and 3.21% with the *diabetes library*), and provides sub-libraries with most of the best compounds, *i.e.* compounds with highest desirability indexes. Comparisons on other simulated combinatorial libraries allow to conclude that the *WEALD* algorithm converges the fastest and also provides sub-libraries of quality similar to the ones obtained with the *Piccolo* algorithm whatever the dimensions of combinatorial libraries are.

The advantage of the *WEALD* algorithm is the weighting methodology developed to guide the exchanges between reagents. According to some fixed rules, a probability of selection is associated to each reagent and they are updated at each iteration taking into account the quality of the explored compounds. Those selection probabilities are used to choose the new reagent that enters the sub-library and replaces the worst actual one. Those weights allow a faster convergence through better sub-libraries.

Acknowledgement

We thank Dr. B. Beck and Dr. B. Boulanger from Eli Lilly and Company for motivating this research and contributing to it, as well as for providing the two real combinatorial libraries, the *antidepressant library* and the *diabetes example*. A special thank to Philippe

Remusat who spent a lot of his time to implement the *WEALD* algorithm in *SAS* and made its application possible at Eli Lilly, linking the *WEALD* algorithm to all other softwares used in the lead optimisation stage.

Bibliography

- Agrafiotis, D.K. and Lobanov, V.S. (2000), "Ultrafast algorithm for designing focused combinatorial arrays", *Journal of Chemical Information and Computer Sciences*, 40, 1030-1038.
- Agrafiotis, D.K., Lobanov, V.S. and Salemme, F.R. (2002), "Combinatorial informatics in the post-genomics era", *Nature reviews. Drug discovery.*, 1, 337-346.
- Agrafiotis, D.K (2002), "Multiobjective Optimization of Combinatorial Libraries", *Journal of Computer-Aided Molecular Design*, 16, 335-356.
- Agrafiotis, D.K (1997), "Stochastic Algorithms for maximizing molecular diversity", *Journal of Chemical Information and Computer Sciences*, 37, 841-851.
- Agrafiotis, D.K., Lobanov, V.S., Rassokhin, D.N. and Izrailev, S. (2000), *The measurement of molecular diversity in virtual screening of bioactive molecules*, Bhm, H.-J., and Schneider, G., Eds., Wiley-VCH, Weinheim.
- Agrafiotis, D.K (1997), "On the use of information theory for assessing molecular diversity", *Journal of Chemical Information and Computer Sciences*, 37(3), 576-580.
- Clark, D.E. and Westhead, D.R. (1996), "Evolutionary algorithms in computer-aided molecular design", *Journal of Computer-Aided Molecular Design*, 10(4), 337-358.
- Cummins, D.J., Andrews, C.W., Bentley, J.A. and Cory, M. (1996), "Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds.", *Journal of Chemical Information and Computer Sciences*, 36, 750-763.
- Derringer, G.C. and Suich, R. (1980), "Simultaneous optimization of several response variables", *Journal of Quality Technology*, 12(4), 214-219.
- Fedorov, V.V (1972), *Theory of optimal experiments*, Academic Press, NY.

- Floyd, C.D., Leblanc, C. and Whittaker, M. (1999), "Combinatorial chemistry as a tool for drug discovery", *Progress in medicinal chemistry*, 36, 91-168.
- Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gordon, E.M. (1994), "Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries", *Journal of Medicinal Chemistry*, 37(9), 1233-1251.
- Gillet, V.J., Willett, P., Bradshaw, J. and Green, D.V.S. (1999), "Selecting combinatorial libraries to optimize diversity and physical properties", *Journal of Chemical Information and Computer Sciences*, 39, 169-177.
- Gillet, V.J., Willett, P. and Bradshaw, J. (1997), "The effectiveness of reactant pools for generating structurally diverse combinatorial libraries", *Journal of Chemical Information and Computer Sciences*, 37, 731-740.
- Gillet, V.J., Willett, P., Fleming, P.J. and Green, D.V. (2002), "Designing focused libraries using MoSELECT", *Journal of Molecular Graphics & Modelling*, 20(6), 491-498.
- Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gallop, M.A. (1994), "Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic-synthesis, library screening strategies, and future-directions", *Journal of Medicinal Chemistry*, 37(10), 1385-1401.
- Harrington, J. (1965), "The desirability function", *Industrial Quality Control*, 21(10), 494-498.
- Hassan, M., Bielawski, J.P., Hempel, J.C. and Waldman, M. (1996), "Optimization and visualization of molecular diversity of combinatorial libraries", *Molecular Diversity*, 2, 64-74.
- Judson, R. (1997), "Genetic algorithms and their use in chemistry", *Reviews in Computational Chemistry*, 10, 1-73.
- Lam, K.S. (1997), "Application of combinatorial library methods in cancer research and drug discovery", *Anti-cancer Drug Design*, 12(3), 145-167.
- Leach, A.R. and Hann, M.M. (2000), "The in silico world of virtual libraries", *Drug Discovery Today*, 5(8), 326-336.
- Lewis, R.A., Mason, J.S. and McLay, I.M. (1997), "Similarity measures for rational set selection and analysis of combinatorial libraries: The diverse property-derived (DPD) approach", *Journal of Chemical Information and Computer Sciences*, 37(3), 599-614.
- Nelder, J.A. and Mead, R. (1965), "A Simplex Method for Function Minimization", *Computer Journal*, 7(4), 308-313.

- Pickett, S.D., McLay, I.M. and Clark, D.E. (2000), "Enhancing the hit-to-lead properties of lead optimization libraries", *Journal of Chemical Information and Computer Sciences*, 40(2), 263-272.
- Sheridan, R.P. and Kearsley, S.K. (1995), "Using a genetic algorithm to suggest combinatorial libraries", *Journal of Chemical Information and Computer Sciences*, 35(2), 310-320.
- Singh, J., Ator, M.A., Jaeger, E.P., Allen, M.P., Whipple, D.A., Solowej, J.E., Chowdhary, S. and Treasurywala, A.M. (1996), "Application of genetic algorithms to combinatorial synthesis: a computational approach for lead identification and lead optimization", *Journal of the American Chemical Society*, 118, 1669-1676.
- Stanton, R.V., Mount, J. and Miller, J.L. (2000), "Combinatorial library design: maximizing model fitting compounds with matrix synthesis constraints", *Journal of Chemical Information and Computer Sciences*, 40, 701-705.
- Terrett, N.K., Gardner, M. and Gordon, D.W. (1995), "Combinatorial synthesis - The design of compound libraries and their application to drug discovery", *Tetrahedron*, 51, 8135-8173.
- Venkatasubramanian, V., Chan, K. and Caruthers, J.M. (1995), "Evolutionary design of molecules with desired properties using the genetic algorithm", *Journal of Chemical Information and Computer Sciences*, 35(2), 188-195.
- Venton, D.L. and Woodbury, C.P. (1999), "Screening combinatorial libraries", *Chemometrics and Intelligent Laboratory Systems*, 48(2), 131-150.
- Weber, L., Wallbaum, S., Broger, C. and Gubernator, K. (1995), "Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm", *Angewandte Chemie (International ed. in English)*, 34, 2280-2282.
- Weber, L., Wallbaum, S., Broger, C. and Gubernator, K. (1995), "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules", *Journal of Chemical Information and Computer Sciences*, 28(1), 31-36.
- Zheng, W., Hung, S.T., Saunders, J.T. and Seibel, G.L. (2000), "PICCOLO: a tool for combinatorial library design via multicriterion optimization", *Pacific Symposium on Bio-computing*, 588-599.

Chapter 6

An integrated methodology for virtual molecules optimisation with an application to the screening of the *diabetes* combinatorial library

We have the project to write a paper presenting the integrated methodology for a chemometrics journal.

6.1 Introduction

The development and statistical validation of QSAR models and the techniques to select compounds in a combinatorial library that optimise predicted characteristics are the two major concerns of this thesis. QSAR models (Chapter 1 and 2) and desirability indexes (Chapter 3 and 4) can be integrated in the *WEALD* algorithm (Chapter 5) to virtually screen a combinatorial library and select the most promising molecules. An integrated methodology is presented in this chapter and illustrated on the *diabetes* combinatorial library (the R-code is available under request).

The different steps of this methodology to virtually optimise molecules are summarised on Figure 6.1. First, it is important to construct a meaningful library and clearly define the objective of a sub-library design. Then, QSAR models are developed in order to be able to predict the drugability properties of interest for any compound in the combinatorial library. Finally, the *WEALD* algorithm is applied to screen the combinatorial library, retaining the compounds with the highest expected desirability. The great improvement

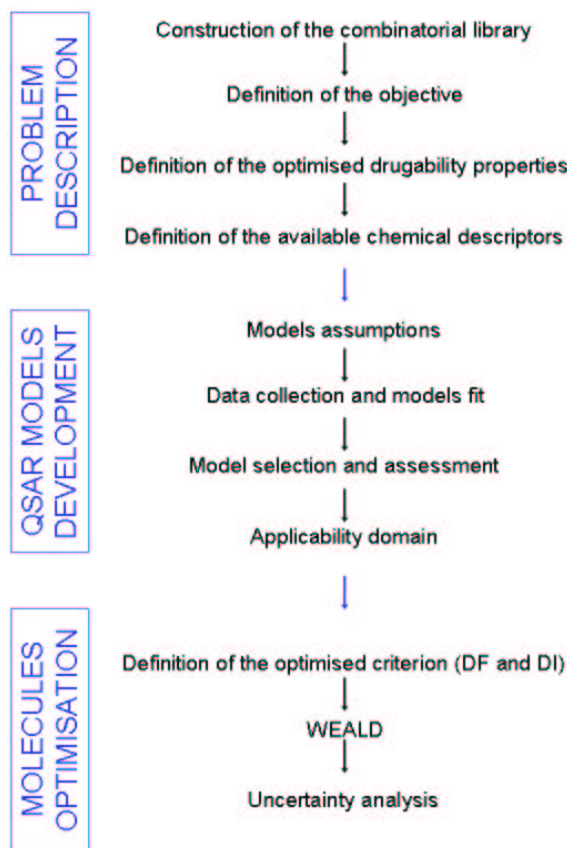


Figure 6.1: Schematic summary of the integrated methodology.

over existing methods is that the QSAR models uncertainty is not neglected in the ranking of the molecules provided by *WEALD*.

6.2 Summary of the integrated methodology for virtual molecules optimisation

6.2.1 Problem description

The first part of the optimisation process is mainly the task of the chemists. They have the crucial responsibility to construct the combinatorial library around the lead by defining the possible R-groups and listing the relevant reagents. It is among this collection of compounds that a better drug candidate will be searched out.

6.2 Summary of the integrated methodology for virtual molecules optimisation

The chemists have to define the desired design for the sub-library of molecules that will then be tested in laboratories. If a combinatorial sub-library is preferred, the number of reagents selected in each R-group should be stated. If a non-combinatorial sub-library is preferred, its number of compounds should be fixed.

Finally, the chemists have to determine the drugability properties that the drug candidate should encounter. The statisticians should understand those properties and know their desirable values. The chemists can use their *a priori* knowledge to help the statisticians to select, among the set of all chemical descriptors provided by the specific softwares, the descriptors or the groups of descriptors that are likely related to the drugability properties.

6.2.2 QSAR models development

QSAR models like Multiple Linear Regression (MLR), Partial Least Squares Regression (PLSR) or Regression Tree (RT) allow to mimic the link between the biological activity of a molecule and a set of chemical descriptors. With MLR, the linear equation linking the drugability property and the chemical descriptors should be stated. The statisticians have also to define a method to select relevant descriptors that enter the MLR model, as well as a criterion (*e.g.* least squares) to estimate the parameters. With PLSR, an algorithm should be chosen to construct the latent variables and select the ones that enter the model. With RT, there is no *a priori* assumption about the nature of the relationship between the drugability property and the descriptors. The splitting rules and the pruning rules should be stated.

If no data are available to build the models, a subset of molecules are synthesised in laboratories and the responses of interest measured. Specific softwares compute the corresponding chemical descriptors. After a potential pretreatment or cleaning, those data are finally used to select determinant chemical descriptors and fit the assumed QSAR models.

It is recommended to try different classes of models and different QSAR models, measure their goodness-of-fit and their predictivity, and select the most predictive one as, for example, the model with the highest Q^2 , computed by cross-validation or bootstrap, or preferably on an external independent set. Once the predictive power of the selected QSAR model has been assessed, it can be used to predict the drugability properties for any compound in the combinatorial library. It is also important to define the applicability domain of the QSAR model in order to measure extrapolation.

6.2.3 Molecules optimisation

The desirability functions and the desirability index formulate the compromise between the different drugability properties and allows a ranking of the molecules in the combina-

torial library. Any DF two times continuously differentiable and non zero can be used (see Section 2 of Chapter 3). The weighted geometric mean is advised as it allows to penalise a molecule with at least one bad drugability property.

Once the desirability index has been defined, the estimator (4.11) of the expected desirability index can be applied to quantify the quality of any molecule in the combinatorial library, based on the QSAR models, allowing their ranking. Finally, the confidence interval around the expected desirability index allows to quantify the uncertainty of its estimation. The size of those confidence intervals gives also an idea of the reliability of the optimisation process. Large intervals mean uncertain estimations of desirability index and optimisation of noise.

The *WEALD* algorithm can be applied to screen the combinatorial library and optimise a sub-library, in a combinatorial fashion or not. An initial sub-library is selected at random and exchanges between reagents are performed to increase the expected desirabilities of its compounds summarised by a loss value. Those exchanges are guided by some judicious selection probabilities that are associated to each reagent depending on the quality of molecules constructed with the reagent. *WEALD* converges fast and provides a sub-library with the most promising compounds.

The concept of indistinguishable optimal zone (4.22) can be applied to select explored molecules that are not significantly worse than the best molecule found, *i.e.* the molecule with the highest estimated desirability. This can help the chemists to choose the molecules that will finally be tested in laboratories, among the sub-library provided by *WEALD* or among the molecules not distinguishable from the optimal molecule, taking also into account if the molecules are in the applicability domains of the QSAR models.

6.3 Application of the integrated methodology to optimise the *diabetes* combinatorial library

This section illustrates the proposed methodology on the *diabetes* combinatorial library. The initial library optimisation problem of Eli Lilly has been simplified (originally, there were more responses to optimise and also even more compounds in the library constructed with more reagents than what is presented there). This section does not present the real application made at Eli Lilly. It illustrates on real data the interest of the proposed methodology and also some difficulties often encountered by QSAR practitioners such as working with high dimensional data (more descriptors than observed responses, requiring a descriptors subset selection method), building models on different training sets (different sets of molecules, different sets of computed descriptors), and optimising several drugability prop-

erties simultaneously with models fitted on possibly totally different molecules than the ones in the library.

6.3.1 Problem description

Construction of the combinatorial library

The *diabetes* combinatorial library has been built up at Eli Lilly in the context of the development of a new molecule acting on glucose regulation in type II diabetes. It has been constructed on the basis of three R-groups (R-group A, R-group E and R-group W), with respectively 47, 50 and 47 reagents. It contains $M = 47 \times 50 \times 47 = 110\,450$ compounds that can be obtained by combining one reagent in each R-group. The *diabetes* combinatorial library is schematically represented on Figure 6.2.

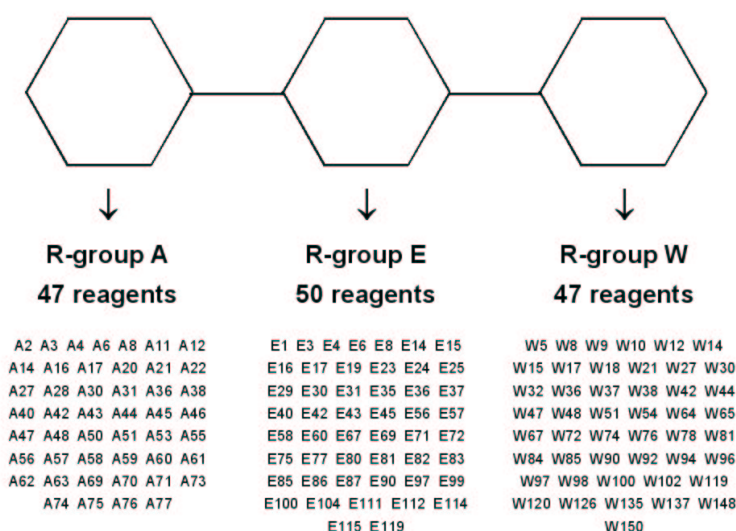


Figure 6.2: *diabetes library*: real combinatorial library composed of three R-groups and $M = 47 \times 50 \times 47 = 110\,450$ compounds.

Definition of the objective

Let us suppose that the objective was to find a subset of $m = 100$ most promising molecules that could then be synthesised and tested in laboratories. Those 100 compounds should not necessarily be obtained in a combinatorial fashion.

Definition of the optimised drugability properties

The 100 compounds selected in the sub-library should optimise ten drugability properties, but only four are retained in this chapter to simplify the illustration of the methodology.

A drug candidate should be selective. In order to act on glucose regulation, chemists were looking for compounds that bind to two targeted receptors, R_1 and R_2 , but not to two other undesired receptors, R_3 and R_4 . The four properties (Y_1, Y_2, Y_3, Y_4) to be optimised were measuring, in log-scale, the quantity of substance to inject around a receptor to have a binding. The two first properties, Y_1 and Y_2 , had to be minimised as the molecules should bind receptors R_1 and R_2 and the two last properties, Y_3 and Y_4 , had to be maximised as the molecules should not bind receptors R_3 and R_4 .

Definition of the available chemical descriptors

Chemical descriptors can be computed using a specific software on the basis of the SMILES coding technique of molecules. They are computed by groups. For instance, a first class of descriptors interesting to model the selectivity is a group of simple chemical properties describing the molecule as a whole. It contains variables such as the number of atoms, the number of rings, the number of different elements in the molecule, the average molecular weight, the number of singly connected atoms, the fraction of atoms that are singly connected, the number of doubly connected atoms, the fraction of atoms that are doubly connected,... It is clear that inside this group, some molecular descriptors are redundant.

Another set of available descriptors quantifies the overall charge distribution for the molecule: the total absolute charge, the total positive charge, the total negative charge... For this group again, it is obvious that some molecular descriptors are redundant.

Other groups of descriptors measure electrotopological properties, molecular surface properties, connectivity properties,... In general, the available descriptors are discrete or continuous variables. More than 9000 molecular descriptors can be computed at Eli Lilly, but a selection for entering the QSAR models should be done.

6.3.2 QSAR models development

Models assumptions

For each optimised response, an MLR model was assumed to explain the response as a linear function of the standardised descriptors. A forward regression based on the BIC criterion was applied in order to select the subset of descriptors to put in the model. Results

were obtained with the function *stepAIC* of the *MASS* R-package.

A PLSR model was also fitted to each training set as this modeling technique is popular in chemometrics. The first 100 latent variables were constructed using the *SIMPLS* algorithm of de Jong (1993) and the number of them kept in the regression model was chosen in order to minimise the bias-corrected 10-fold cross-validation estimate of the Mean Squared Error of Prediction (2.10) as proposed by Mevik and Cederkvist (2004). The results were obtained using the function *pls* of the *pls* R-package.

Finally, binary RTs were fitted to each training set in order to obtain predictive models. First, recursive partitioning was applied till all terminal nodes contain at least two observations. At each step, the best binary split was selected as the one that produces the greatest change in the residuals sum of squares, *i.e.* $RSS_P - (RSS_L + RSS_R)$ where RSS_P is the residuals sum of squares at parent node, RSS_L and RSS_R are the residuals sums of squares at left and right child nodes respectively. Pruning and bagging were also tested. The results were obtained using the function *rpart* of the *rpart* R-package.

It has to be noticed that it would be more relevant to use the same criterion to build all the different models. Nevertheless, it is not the case in this application as it was not possible with the default implementations of the different R-packages.

Data collection and models fit

Each response was measured in laboratories for different sets of compounds: 558, 519, 985 and 1037 molecules for Y_1 , Y_2 , Y_3 and Y_4 , respectively. The choice of the tested molecules is not explained by Eli Lilly. Those measurements were gathered in four text files, one for each response. They are composed of two columns: a first column with the Lilly molecules number code and a second column with the measurements.

Corresponding to each observed response file, a text file stocks the computed descriptors in columns for each tested molecule in rows. The same descriptors, 3394, are available for the two first responses Y_1 and Y_2 . The number of available descriptors varies for the two last properties Y_3 and Y_4 with 2030 and 1914 respectively.

For the 110 450 compounds in the combinatorial library, 9716 molecular descriptors were computed and stocked in a huge text file of more than 2.5GB. Note that this file cannot be entirely opened neither in TextPad nor in R. It can only be read row by row, *i.e.* molecule by molecule. From a practical point of view, it is not necessary to compute the descriptors for the whole library. Descriptors can only be punctually computed while molecules are explored by the *WEALD* algorithm.

Before starting to study the link between observed responses and descriptors, the different data files were scrupulously analysed and cleaned. First, the files format was analysed. All are space-delimited, but there were double spaces occurring in some places, shifting the descriptors values. Those double spaces were replaced by a single space. The molecules order was not the same in both response file and descriptors file. The response file rows were reordered such that the rows of both files correspond to the same molecules.

Secondly, the descriptors files were cleaned in order to work with smaller files in the modeling step. Some descriptors were only computed for the laboratory-tested molecules but were not available for the whole combinatorial library. Those were removed of the descriptors files because if one of them was entering a QSAR model, this would then have been impossible to predict the response for molecules in the combinatorial library.

Other descriptors were removed from the descriptors files because they were useless for the QSAR model building. This was the case of constant descriptors, *i.e.* descriptors for which an identical value was observed for all molecules tested in laboratory. This was also the case for pairwise redundant descriptors, *i.e.* two descriptors that take exactly same values or, more generally, two descriptors that have an observed absolute correlation of 1. For those pairwise redundant descriptors, one of the two was removed.

Finally, for each optimised property, a unique text file was created merging the observed responses and the descriptors kept. Each row corresponds to a laboratory-tested molecule, the first column to the observed responses and the following columns to computed descriptors. Those four files were considered as the available training sets for QSAR model building. Table 6.1 summarises their dimensions.

The three classes of models, MLR, PLSR and RT, were fitted on those four training sets, selecting entered explanatory variables as explained in Section 3.2. The most predictive models were chosen to carry on the optimisation process.

	Observed molecules	Available descriptors	Descriptors left after cleaning
Y_1	558	3394	1563
Y_2	519	3394	1769
Y_3	985	2030	1964
Y_4	1037	1914	1844

Table 6.1: Training sets for each optimised response: number of observed molecules and number of descriptors left after cleaning

Model selection and assessment

Firstly, the MLR models were fitted on the Eli Lilly data. Even with a reasonable number of variables (40, 30, 48 and 60 for modeling Y_1 , Y_2 , Y_3 and Y_4 respectively), the adjusted R^2 's are high, 0.700, 0.589, 0.892 and 0.825 respectively, reflecting a quite good fit. This can be visualised on Figure 6.3 presenting the fitted values *versus* the observed ones for each drugability property. Other summary statistics are presented in Table 6.2. The graphs of the evolution of the BIC criterion as a function of the number of entered descriptors are also presented in Appendix 1.

A residuals analysis was also performed to validate the underlying assumptions of the MLR model. For each model, a plot of the studentised residuals *versus* the fitted values is reported in Appendix 1 (Figure 6.19). A quantile-quantile plot is also drawn for each model in Appendix 1 to check the Normality of the residuals.

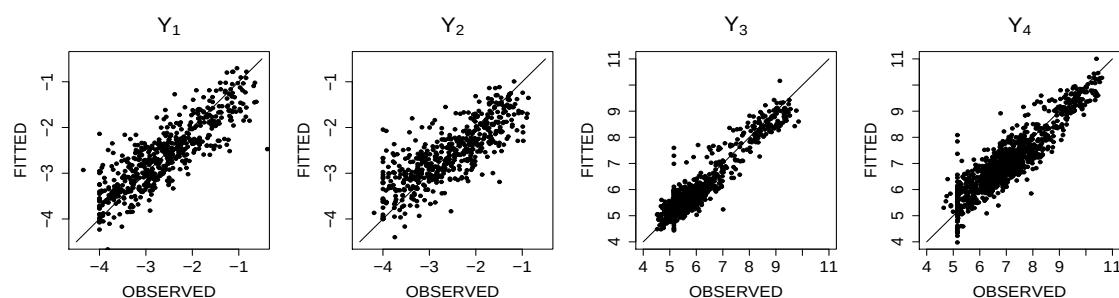


Figure 6.3: Graphs of the fitted values versus the observed values for each MLR model.

	N	$K - 1$	R^2	R^2_{adj}	S	F p-value
Y_1	558	40 (out of 1563)	0.722	0.700	0.470	$< 2.2 * 10^{-16}$
Y_2	519	30 (out of 1769)	0.613	0.589	0.527	$< 2.2 * 10^{-16}$
Y_3	985	48 (out of 1964)	0.897	0.892	0.392	$< 2.2 * 10^{-16}$
Y_4	1037	60 (out of 1844)	0.836	0.825	0.548	$< 2.2 * 10^{-16}$

Table 6.2: Summary statistics for the four fitted MLR models: N is the number of observed molecules, $K - 1$ is the number of selected descriptors, R^2 is the multiple determination coefficient, R^2_{adj} is the adjusted multiple determination coefficient, S is the estimated standard error of residuals and F p-value is the p-value of the F-test for model significance.

Then, PLSR models were also fitted to the same data. Some summary statistics are displayed for each fitted PLSR model in Table 6.3. The fit is good (too good?), especially for Y_3 and Y_4 as it can also be viewed from Figure 6.4 comparing the observed and the fitted values. Other graphs are presented in Appendix 2, such as the estimated MSEP *versus* the number of latent variables entered (Figure 6.20), and the graphs of a residuals analysis (Figure 6.21).

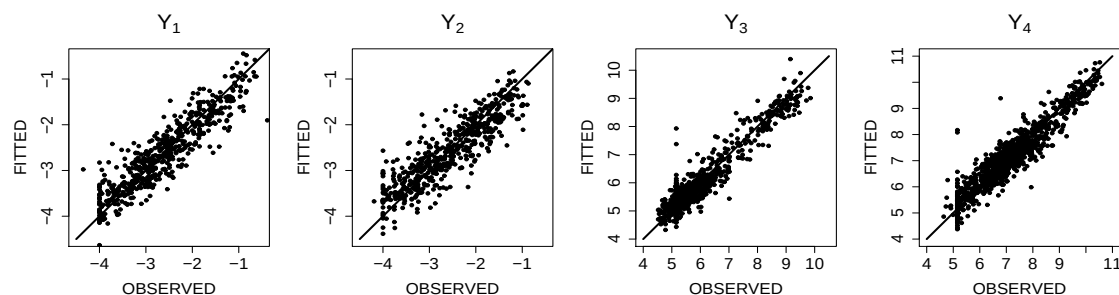


Figure 6.4: Graphs of the fitted values versus the observed values for each PLSR model.

	N	K	R_Y^2	R_X^2	S
Y_1	558	54	83.28	100	0.369
Y_2	519	51	77.04	100	0.414
Y_3	985	37	91.55	100	0.353
Y_4	1037	44	88.80	100	0.449

Table 6.3: Summary statistics for the four fitted PLSR models: N is the number of observed molecules, K is the number of entered latent variables, R_Y^2 is the percentage of Y-variance explained by the latent variables, R_X^2 is the percentage of X-variance explained by the latent variables and S is the estimated standard error of residuals.

Finally, the fit of RTs on the QSAR data was analysed. As expected, the trees "without pruning" (partitioning to decrease the residuals sum of squares till at least 6 observations in a node to allow further splitting and at least 2 observations per node) are overfitted as it can be seen on the graphs of Figure 6.5 that compare the fitted values to the observed ones.

The dendrograms of Figure 6.6 represent the successive splits with respectively 110, 95, 116 and 134 descriptors used. The size of the last branches representing a small decrease in the residuals sum of squares is also the sign of the overfit.

The graphs of the residuals analysis (Figure 6.25) are displayed in Appendix 3

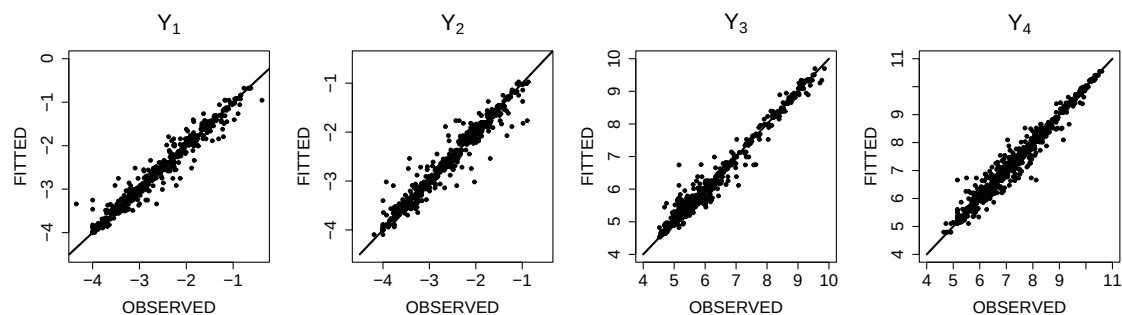


Figure 6.5: Graphs of the fitted values versus the observed values for each RT without pruning.

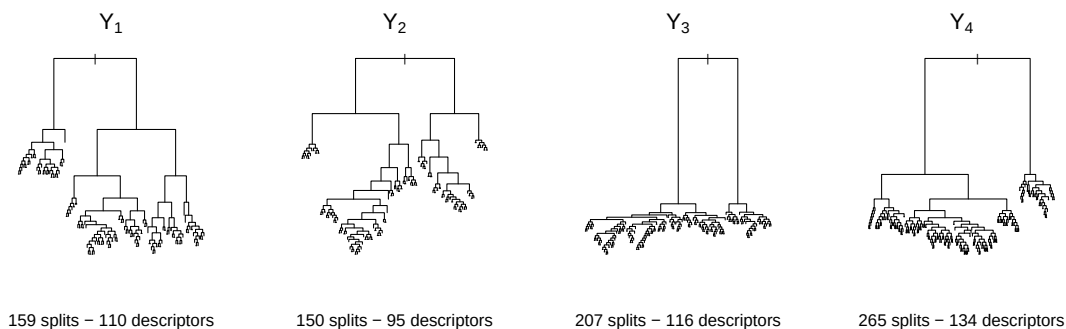


Figure 6.6: Dendrograms of each RT without pruning.

In a second step, those too deep RTs were pruned by selecting the number of splits that minimises a 10-fold cross-validated estimate of the prediction error (2.17). The evolution of the CV prediction error as a function of the number of splits is presented in Appendix 3 (Figure 6.22).

By looking at the corresponding dendrograms of Figure 6.8, one can conclude that the starting trees of Figure 6.6 have been intensively pruned, especially for the RT of the third response that contains only four terminal nodes after pruning. This results in poor fit, as it can clearly be seen on the graphs of Figure 6.7 comparing the fitted values to the observed ones.

The graphs of the residuals analysis (Figure 6.26) are displayed in Appendix 3

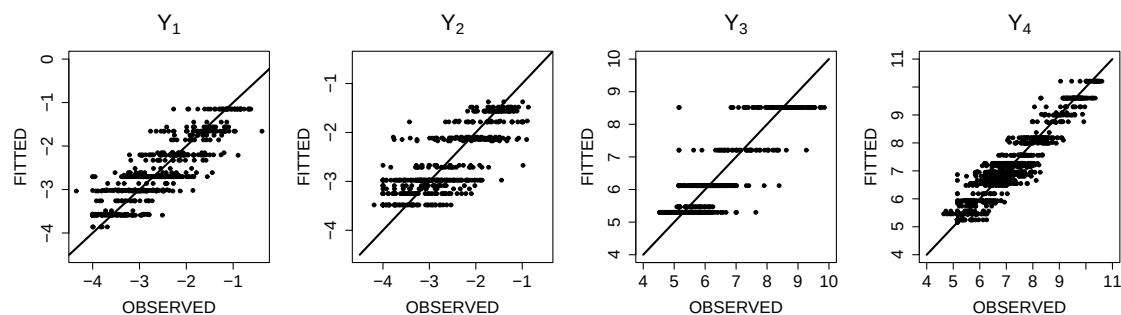


Figure 6.7: Graphs of the fitted values versus the observed values for each RT with pruning.

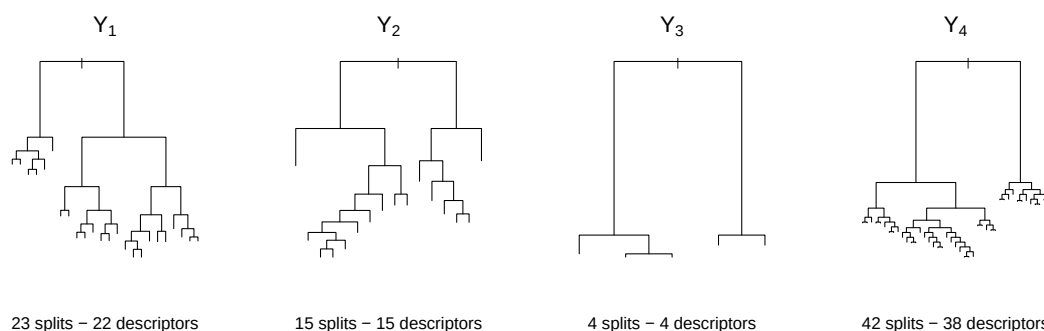


Figure 6.8: Dendrograms of each RT after pruning according to a cost complexity measured based on a cross-validated Residual Sum of Squares.

In a third step, the power of *bagging* (bootstrap aggregation) was used to reduce the sensitivity of the trees to the training sets and obtain smoother predictors. In his paper about bagging predictors, Breiman (1996) used 25 replicates of a regression tree and he judged the results reasonable. As the fit of RT on such high dimensional datasets is intensive in R-software, only ten replicates of each training set were drawn at random with replacement (as demonstrated later, the predictive power has already significantly increased compared to the original predictor and more replicates would probably not provide significant additional improvement). Bootstrapped versions of the training set were used, one at its turn, to fit an RT like in the first step of the analysis. This bagging procedure was also performed for pruned trees, according to the same number of splits as in the second step of the analysis. Dendograms of the ten bagged RTs for the first optimised property are presented in Appendix 3, as well as the ten bagged pruned RTs (Figures 6.23 and 6.24).

The resulting predictor is not a tree any more, but a collection of ten RTs. The fitted values are defined as the average prediction over the ten replicated RTs. Without pruning, fitted values are not closer to the observed ones with bagging than with a simple RT (see Figure 6.9). But with pruning, fitted values are closer to the observed ones with bagging than with a simple pruned RT (see Figure 6.10). Bagging improves the fit of pruned RTs but deteriorates the fit of non-pruned RTs. Nevertheless, it can be expected that bagged trees have a higher predictive power for new data. The fitting performances of the four RTs types are summarised in Table 6.4. A residuals analysis is performed in Appendix 3 (Figures 6.25, 6.26, 6.27 and 6.28).

The goodness-of-fit varies from one model to another. By averaging the R^2 over the 4 responses, one conclude that RT without pruning provides the best fit and MLR the worst fit. Figure 6.3 for MLR models shows a high variability of the fitted values compared to Figure 6.5 for RT. Some graphs of fitted *vs* observed values or graphs of residuals *vs* fitted values disclose some blocks of data (mainly for Y_3), some strata (because of constant observed values due to some threshold or because of constant fitted values as it is the case with RT) and some outlying points that are not well fitted but in small number compared to the number of data. Those points do not seem to influence the model fit. Most of the Normal qqplots of residuals (especially for RT) show an 'S' pattern reflecting that the majority of points are well fitted, but some extremer points are missed. For some molecules with high (or small) observed responses, the predicted value is closer to the mean than it should. This phenomenon especially occurs with RT as the RT predictions are defined as the average of observed responses. Nevertheless, there is no worrying pattern in the residuals analyses that could questioned the validity of the models.

	N	No pruning			Pruning			Bagging No pruning		Bagging Pruning	
		K	R^2	S	K	R^2	S	R^2	S	R^2	S
Y_1	558	110	0.962	0.197	22	0.749	0.438	0.901	0.332	0.798	0.412
Y_2	519	95	0.950	0.218	15	0.621	0.513	0.896	0.332	0.763	0.445
Y_3	985	116	0.979	0.193	4	0.818	0.509	0.962	0.258	0.847	0.468
Y_4	1037	134	0.972	0.253	38	0.889	0.446	0.947	0.351	0.897	0.435

Table 6.4: Summary statistics for the four fitted RTs models: N is the number of observed molecules, K is the number of used descriptors, R_Y^2 is the percentage of Y-variance explained by the RT and S is the estimated standard error of residuals.

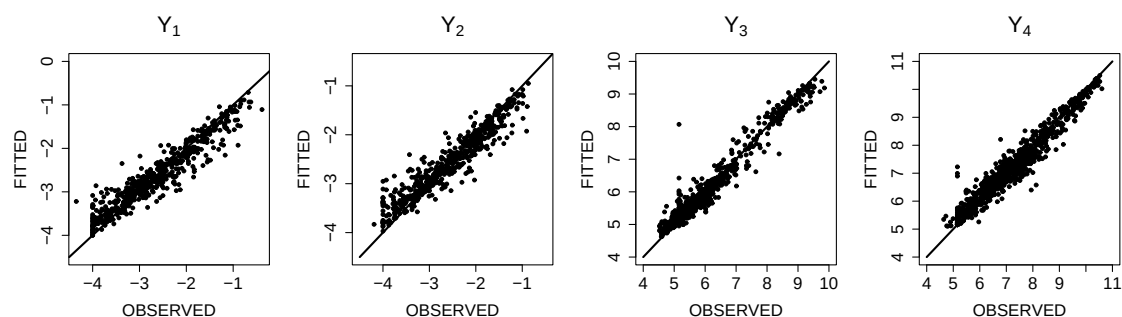


Figure 6.9: Graphs of the fitted values versus the observed values for bagged RTs without pruning.

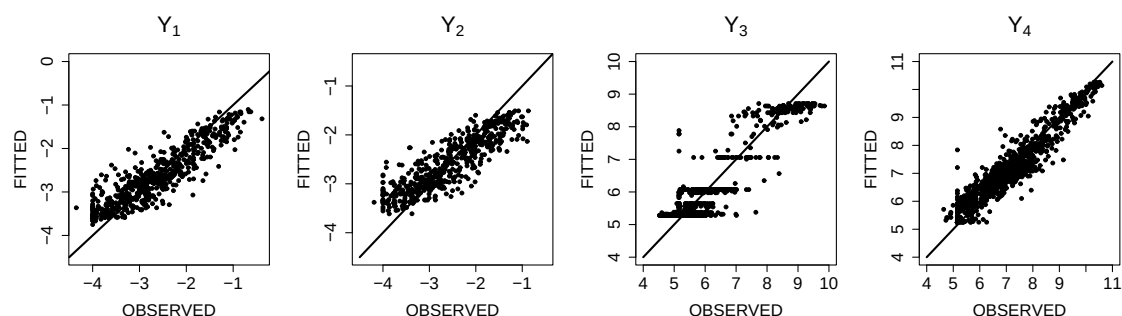


Figure 6.10: Graphs of the fitted values versus the observed values for bagged RTs with pruning.

As discussed in the literature review on QSAR modeling of Chapter 1, QSAR developers are not looking for a model with the best fit as possible, but a model that has a high predictive power. When no additional test set is available, a common measure of internal predictivity is the cross-validated R^2 denoted as Q^2 (1.19). Q^2 allows to manage the bias-variance tradeoff of fitted models as a function of their complexity and is similar to the cross-validated MSE (2.17) generally used in statistical applications.

For each optimised response and for each type of fitted model, 10-fold cross-validation was applied 100 times to obtain 100 values of Q^2 that were then averaged. The average Q^2 's are presented in Table 6.5. For all modeled responses, the highest predictive power (*i.e.* the highest Q^2) is obtained with bagged RTs without pruning. Nevertheless, we prefer to carry on this application with MLR models, which is the third best model, as formulas are available to measure prediction error. Bagged trees would require a second loop of bootstrap procedure to estimate the variance of prediction (see Section 5.4 of Chapter 2), which can be heavy in practice to apply the error propagation methodology in this context.

It has to be noted that the results for PLSR models are not as good as expected. An explanation for this bad predictive performance may be the lack of robustness of such methods (like PCR) as demonstrated by Vanden Branden and Hubert (2004). Not surprisingly, from this prediction power point of view, the worst model is the RT without pruning, which is overfitting.

The four fitted MLR models were used to predict the responses to be optimised for any compound in the combinatorial library as a function of the computed values of the 40, 30, 48 or 60 entered descriptors respectively.

	Y_1	Y_2	Y_3	Y_4
MLR	0.722 (40)	0.613 (30)	0.897 (48)	0.835 (60)
PLSR	0.542 (54)	0.359 (51)	0.804 (37)	0.727 (44)
RT no pruning	0.307 (110)	0.232 (95)	0.762 (116)	0.682 (134)
RT pruning	0.403 (22)	0.357 (15)	0.795 (4)	0.714 (38)
RT bagging - no pruning	0.896	0.871	0.961	0.939
RT bagging - pruning	0.783	0.717	0.844	0.895

Table 6.5: For each fitted models, average Q^2 computed by 10-fold cross-validation (number of variables entered in the models).

Applicability Domain

In agreement with the choice of QSAR models class, the applicability domain of each QSAR model was defined on the basis of the leverage. A molecule is in the applicability domain if its leverage is smaller than $5K/N$ where K is the number of parameters in the MLR model and N is the number of molecules used to train it.

A graph of the leverages for each training set is displayed on Figure 6.11. A new molecule upper the cyan horizontal dashed line at $5K/N$ was considered outside the applicability domain of the QSAR model. The QSAR model was used anyway to predict the response in the optimisation, but a warning was emitted, meaning that the MLR model was extrapolating and the prediction might be unreliable.

6.3.3 Molecules optimisation

Definition of the optimised criterion (DF and DI)

Normal *cdf* based desirability functions (3.4-3.5-3.6) were first used to map each optimised response in the same interval $]0,1[$. The two first responses, Y_1 and Y_2 , had to be minimised and a desirability function of the form $d(Y) = 1 - \Phi(\frac{Y-a}{b})$ was used (Φ is the cumulative distribution function of the standard Normal). The two other responses, Y_3 and Y_4 , to be maximised were transformed with a desirability function of the form $d(Y) = \Phi(\frac{Y-a}{b})$.

The parameters a and b of those functions were fixed respectively as the mean and the standard deviation of the predicted responses for 100 molecules chosen at random in the combinatorial library. These constituted an initial sub-library for the *WEALD* algorithm (see Appendix 4). The resulting desirability functions are displayed on Figure 6.12.

In a second step, the desirability functions were aggregated using a desirability index defined as their geometric mean, $D(\mathbf{Y}) = \prod_{i=1}^4 (d_i(Y_i))^{1/4}$, assigning the same importance to each of the four optimised responses. This way a molecule that is miserable for at least one response would have at least one desirability function value near zero, resulting in an index also near zero.

Figure 6.13 compares the stochastic criterion $\hat{D}^S(\mathbf{x}) = \hat{E}[D(\mathbf{Y})|\mathbf{x}]$ to the classical criterion $\hat{D}^C(\mathbf{x}) = D(\hat{E}[\mathbf{Y}|\mathbf{x}])$ on the 100 molecules randomly selected in the combinatorial library. It shows that there is nearly no difference between them. The stochastic criterion was optimised in the *WEALD* algorithm. It was estimated by (4.11) for any molecule in the combinatorial library on the basis of the MLR estimations $\hat{E}[Y_i|\mathbf{x}]$ and $\hat{V}[Y_i|\mathbf{x}]$ ($i = 1, 2, 3, 4$) provided by the four QSAR models.

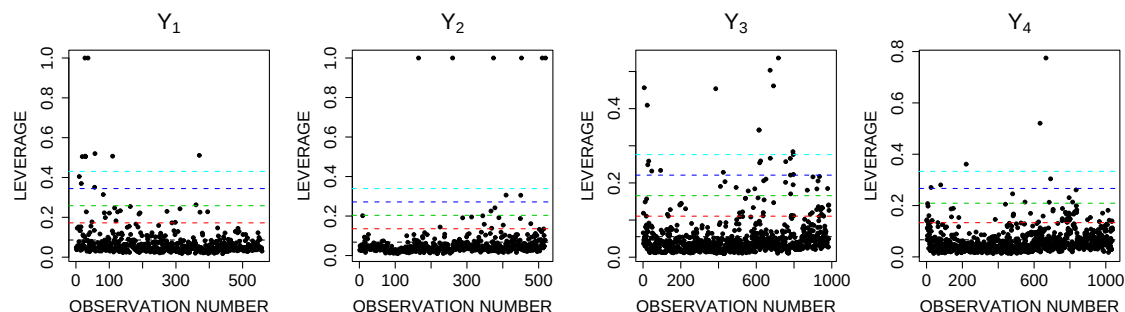


Figure 6.11: Graphs of the leverages of observations in each training set. Five different thresholds are represented as dashed horizontal lines, $1K/N$ (black), $2K/N$ (red), $3K/N$ (green), $4K/N$ (blue) and $5K/N$ (cyan).

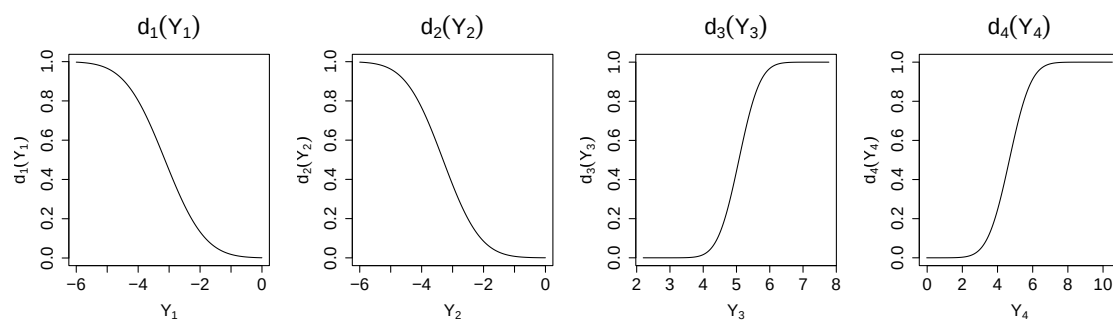


Figure 6.12: Graphs of the four desirability functions. $d(Y_1) = 1 - \Phi(\frac{Y_1 + 3.1351}{1.0231})$, $d(Y_2) = 1 - \Phi(\frac{Y_2 + 3.3086}{0.9535})$, $d(Y_3) = \Phi(\frac{Y_3 - 5.0730}{0.5022})$ and $d(Y_4) = \Phi(\frac{Y_4 - 4.6749}{0.9758})$.

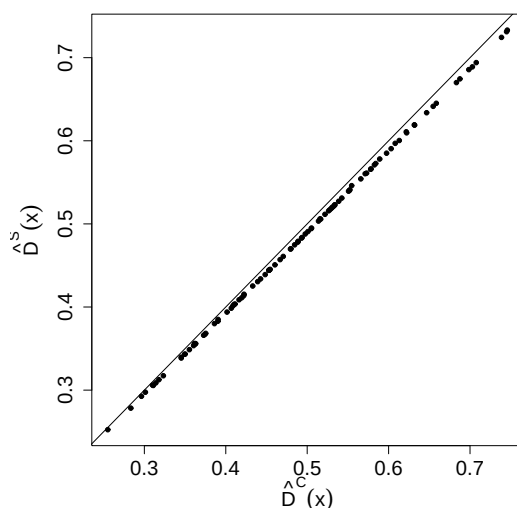


Figure 6.13: Graphs of $\hat{D}^S(\mathbf{x}) = \hat{E}[D(\mathbf{Y})|\mathbf{x}]$ versus $\hat{D}^C(\mathbf{x}) = \hat{D}(E[\mathbf{Y}|\mathbf{x}])$ for the 100 compounds chosen at random in the combinatorial library.

WEALD

The *WEALD* algorithm started with an initial sub-library composed of 100 molecules randomly selected in the combinatorial library (the ones selected to define the desirability functions parameters, in the previous section). For each molecule, the corresponding expected desirability $\hat{E}[D(\mathbf{Y})|\mathbf{x}]$ were computed using the estimator (4.11). The loss of the initial sub-library was computed to summarise the global quality of its compounds as $\frac{1}{100} \sum_{j=1}^{100} (1 - \hat{E}[D(\mathbf{Y})|\mathbf{x}_j])^2 = 0.2833$.

Then, the exchanges principles of the *WEALD* algorithm to optimise a sub-library that is not combinatorial (Section 8 of Chapter 5) were applied in order to decrease the loss until no more exchange could decrease it. The evolution of the loss is displayed on Figure 6.14. 4729 molecules (only 4.28% of the whole combinatorial library) have been explored until algorithm convergence. The resulting final sub-library has a loss of 0.0417. The estimated desirabilities of its compounds ranges from 0.7783 to 0.8448. The most promising molecule with $\hat{E}[D(\mathbf{Y})|\mathbf{x}_{\text{opt}}] = 0.8448$ is obtained by combining reagents A69, E45 and W97. The top 100 compounds provided by *WEALD* are listed in Appendix 4 with the corresponding estimated responses and desirabilities. The estimated responses of this top 100 molecules are displayed on Figure 6.15 among all explored molecules.

However, the uncertainty of the QSAR models predictions propagates on the estimated expected desirabilities and the impact of this uncertainty on the top 100 molecules and on all explored molecules has to be analysed.

Uncertainty analysis

Uncertainty results of the 100 selected molecules are summarised on Figure 6.16. First, for each molecule, the leverages for each QSAR model were computed on the basis of the standardised descriptors $\mathbf{z}$ to check if the molecule was in its applicability domain. If it is not the case for at least one applicability domain, a warning was emitted (red point). 56 compounds in the final sub-library are out of the applicability domain for at least one QSAR model. This is the case for the most promising molecule obtained by combining reagents A69, E45 and W97. It is outside the applicability domain of the first QSAR model with a leverage of 0.95, more than two times further than the limit fixed at $5K/N = 0.43$.

The leverages were also computed for the 4729 explored molecules, leading to the alarming results that 1948 molecules (more than 41%) are outside at least one applicability domain. The whole combinatorial library was analysed in order to confirm this high rate of compounds outside applicability domains. Results are presented in the next section.

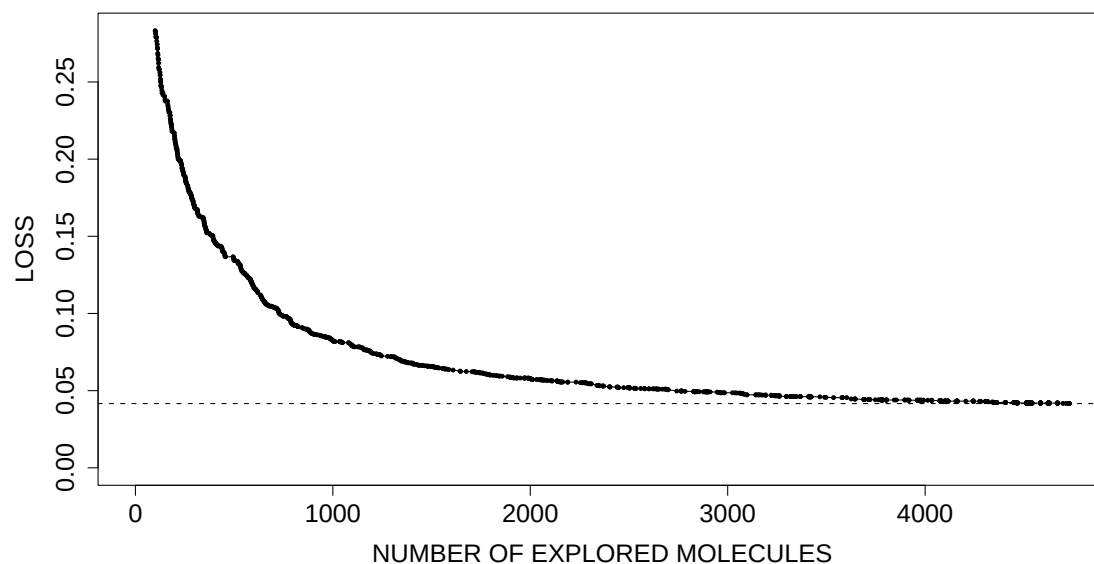


Figure 6.14: Graph of the loss as a function of the number of explored molecules. Each point corresponds to an exchange that decreases the loss. The horizontal dashed line corresponds to the loss of the 100 best molecules (the smallest possible loss), 0.0416.

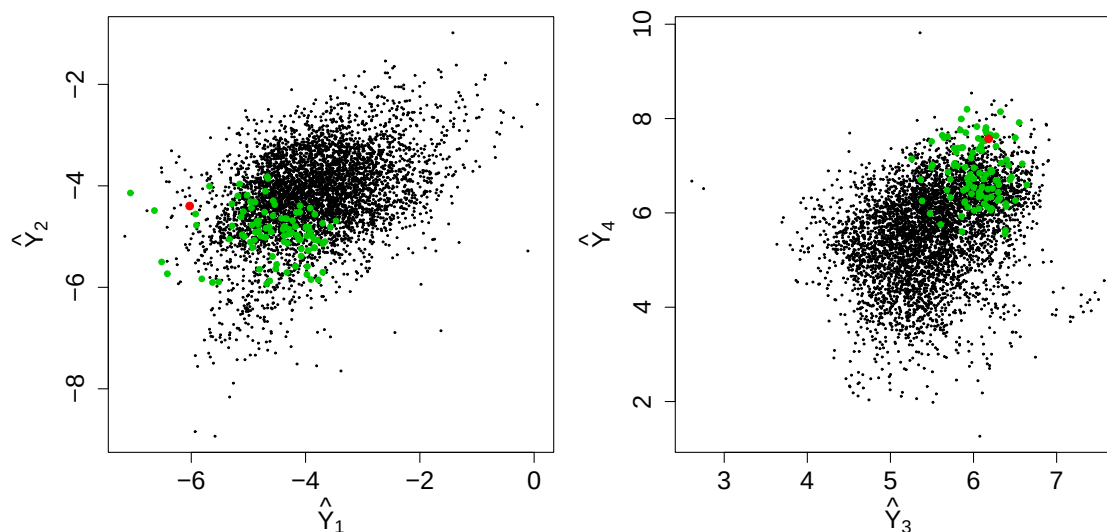


Figure 6.15: Graphs of the estimated responses ($\hat{Y}_1$ vs $\hat{Y}_2$ to be minimised on the left and $\hat{Y}_3$ vs $\hat{Y}_4$ to be maximised on the right) for the 4729 explored molecules. The green points correspond to the top 100 molecules and the red point to the best molecule.

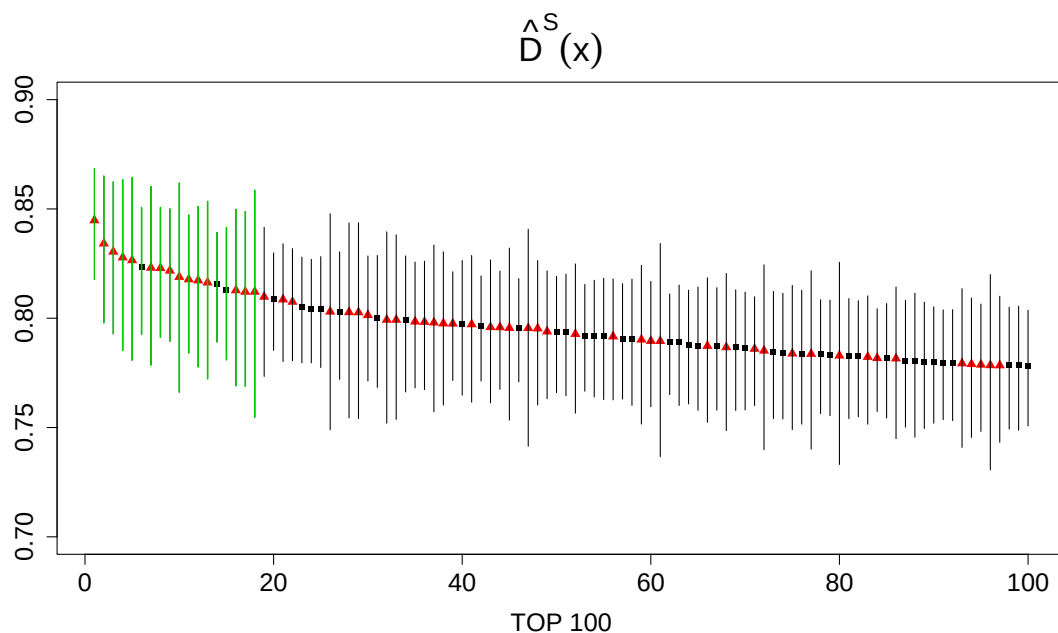


Figure 6.16: Graphs of the estimated expected desirability for the selected 100 best compounds. Red triangles correspond to molecule out of at least one applicability domain and black squares to molecule included in all applicability domains. Green confidence intervals for $E[D(\mathbf{Y})|\mathbf{x}]$ correspond to molecules equivalent to the optimal one and black confidence intervals for $E[D(\mathbf{Y})|\mathbf{x}]$ to compounds significantly worse than the optimum.

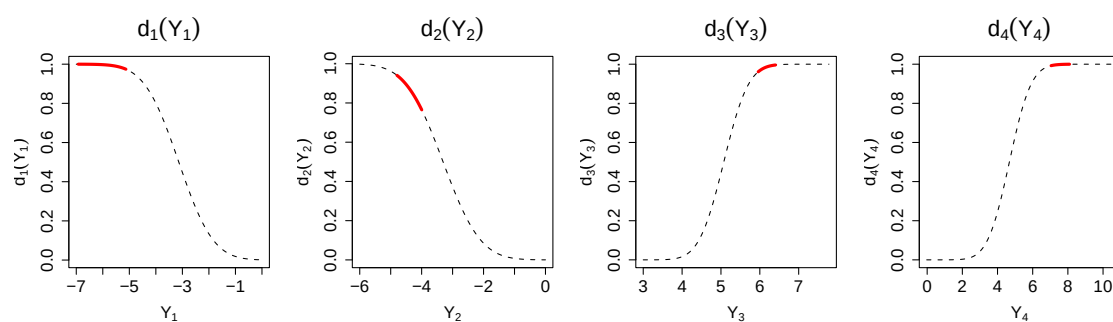


Figure 6.17: Graphs of the four desirability functions where the thick continuous lines correspond to the confidence interval around the estimated response for the optimal molecule.

Then, a confidence interval for $E[D(\mathbf{Y})|\mathbf{x}]$ was associated to each selected molecule using the construction formula (4.19) based on the MLR estimators $\hat{E}[Y_i|\mathbf{x}]$, $\hat{V}[Y_i|\mathbf{x}]$ and $\hat{V}(\hat{E}[Y_i|\mathbf{x}])$ ($i = 1, 2, 3, 4$). The intervals for the top 100 found are listed in Appendix 4 and displayed on Figure 6.16 around the estimated expected desirability $\hat{E}[D(\mathbf{Y})|\mathbf{x}]$. Their lengths vary from 0.04 to 0.11, reflecting a reasonable impact of the propagation of QSAR models predictions errors on the estimated expected desirabilities.

This may be quite surprising as the QSAR models are not that predictive (Q^2 from 0.6 to 0.9). It can be explained by the fact that a great uncertainty for an estimated response may result in small uncertainty of the desirability function because this corresponds to a flat region of the desirability function curve, near 0 or 1. This is the case of the optimum found, for instance, that is quite uncertain for the first response (out of applicability domain). The confidence interval $[-6.93, -5.13]$ around the estimated value -6.03 for the first response corresponds to similar desirability values near 1 (between $d_1(-6.93) = 0.9999$ and $d_1(-5.13) = 0.9744$), as it can be seen on the first graph of Figure 6.17. As this situation is similar for the three other responses, the confidence interval for $E[D(\mathbf{Y})|\mathbf{x}_{opt}]$ is not that large: $[0.8177, 0.8685]$. In the formula for the construction of the confidence interval for $E[D(\mathbf{Y})|\mathbf{x}]$ (4.19), the different variances $V[\hat{E}[Y_i|\mathbf{x}]]$ ($i = 1, 2, \dots, P$) are weighted by the first derivative of the corresponding desirability functions, which is nearly 0 in such flat regions.

But for some compounds, the uncertainty around the estimated expected desirability is not at all negligible. The confidence intervals for $E[D(\mathbf{Y})|\mathbf{x}]$ were computed for all the explored molecules. One can conclude that their lengths are on average 0.12 but may vary from 0.04 to nearly 1! The confidence intervals for the whole combinatorial library were also analysed in order to confirm that QSAR models predictions errors may have an important impact and the *WEALD* algorithm is sometimes optimising noise. Full results are detailed in the next section.

Finally, the concept of the Indistinguishable Optimal Zone (4.22) has been applied to retain explored molecules not significantly worse than the optimal one (A69, E45 and W97) using the multiplicity corrections of Benjamini and Yekutieli for dependent tests (with $\alpha = 0.1$). The test statistic requires the MLR estimate of the covariance between the estimated responses of two molecules $\hat{Cov}(\hat{E}[Y_i|\mathbf{x}_1], \hat{E}[Y_i|\mathbf{x}_2])$, in addition to $\hat{E}[Y_i|\mathbf{x}]$, $\hat{V}[Y_i|\mathbf{x}]$ and $\hat{V}(\hat{E}[Y_i|\mathbf{x}])$ ($i = 1, 2, 3, 4$). One can conclude that 17 compounds in the sub-library have an estimated expected desirability not significantly smaller than the optimal molecule. The confidence intervals for $E[D(\mathbf{Y})|\mathbf{x}]$ are displayed on Figure 6.16 in green for those molecules having a desirability not significantly smaller than the optimal one (otherwise in black).

Among the 4729 explored molecules, only 116 cannot be distinguished from the opti-

mum found due to the QSAR models error. This may sound a few if one looks at the often huge size of the confidence intervals for $E[D(\mathbf{Y})|\mathbf{x}]$. But the covariance between the two estimated desirabilities $E[D(\mathbf{Y})|\mathbf{x}]$ and $E[D(\mathbf{Y})|\mathbf{x}_{\text{opt}}]$ is often as important as their own variance for the explored compounds.

This uncertainty analysis provides complementary information to the sub-library provided by *WEALD*. One could also take those measurements into account directly in *WEALD*, for instance by performing an exchange between reagents only if the resulting new molecule belongs to all applicability domains or/and only if this new molecule is not significantly worse than the current one. But, as a lot of molecules are outside at least one applicability domain and the desirability indexes may be highly uncertain for some molecules, it sounds more logical to let *WEALD* evolving without constraint in the chemical space and, once converged, provide uncertainty information to the chemists and let them choose the molecules to test in laboratories.

6.3.4 Analysis of the whole combinatorial library

In practice, the combinatorial library is so huge that it is too much computer-intensive to generate the descriptors for all compounds and predict the corresponding optimised properties. Nevertheless, the descriptors for the whole *diabetes* library have been computed at Eli Lilly. This allowed to estimate the expected desirability $E[D(\mathbf{Y})|\mathbf{x}]$ for all the 110 450 compounds. The 100 selected compounds in the final sub-library obtained with *WEALD* were compared to the whole library and one can conclude that the 100 selected compounds are within the 105 best compounds of the library. Only five compounds in the true top 100 have been missed by the algorithm. This result confirms again that the *WEALD* algorithm converges fast (only 4729 explored molecules) through excellent sub-libraries.

The confidence interval around $\hat{E}[D(\mathbf{Y})|\mathbf{x}]$ were computed for the whole library. One can conclude that 50% of their lengths are between 0.10 and 0.29, even if this ranges from 0.04 to 1. This confirms that the propagation of QSAR model prediction error on the estimated expected desirability cannot be neglected.

The whole combinatorial library was screened to look for molecules not significantly worse than optimum found. One concludes that 11 416 compounds (more than 10% of the library) cannot be distinguished from the optimal molecule due to the QSAR models error (at a significance level of $\alpha = 0.1$). By defining slices in the chemical space around the optimum found, one can visualise the factor ranges that are covered by the IOZ or compute the proportion of grid-points in that slice that are in the IOZ (like in the simplified Derringer and Suich example), even this is not necessary representative of the global IOZ.

Finally, the result that 56 compounds of the final sub-library are outside at least one applicability domain seems to generalise to the whole library. By computing the leverages for all its compounds, one can conclude that 57 543 compounds (52% of the library) are outside of at least one applicability domain. This is mainly due to the fourth model for which more than a quarter of molecules are outside the applicability domain (see Table 6.6). Chemists are aware of this huge percentage of extrapolation, but they have no other solution than applying the QSAR models to build their priority list for testing compounds.

Y_1	Y_2	Y_3	Y_4
22 266 (20%)	16 228 (15%)	12 431 (11%)	29 407 (27%)

Table 6.6: Number of molecules in the library out of the applicability domain for each QSAR model.

6.4 Conclusion

QSAR models, desirability indexes and *WEALD* algorithm can be combined in a global methodology to construct a priority list for testing compounds in laboratory. The application of the integrated methodology on the *diabetes* library points out some important practical issues.

First, the construction of the whole combinatorial library is a crucial initial step. This is among those molecules that a better drug candidate will be searched out. The choice of the molecules used to train the QSAR models is also essential to insure good models and reliable predictions for the molecules in the combinatorial library.

QSAR data are not easy to handle, mainly because of the great number of descriptors for a small number of measured responses and, also, because the descriptors levels for the observed molecules are, most of the time, not chosen according to an optimal experimental design. Even after a first cleaning of the training sets, removing the numerous constant or redundant descriptors, the number of available descriptors may exceed the number of observed molecules. A method to select the descriptors that enter the QSAR model is necessary and should be based on a prediction criterion instead of a goodness-of-fit measurement to avoid overfitting. It takes a long time in R to fit a forward MLR model, a PLSR model with the selection of entering latent variables or an RT with pruning. Nevertheless it is recommended to fit different QSAR models classes and select the most predictive one.

The virtual screening of the combinatorial library with *WEALD* is very efficient. But in its application, it is important to check if the desirabilities of the explored molecules based on the fitted QSAR models are reliable. The size of the confidence intervals for the desirability of the molecules in the initial sub-library provides a first idea of the impact

of the QSAR models error on the desirability. It is also interesting to compute the confidence intervals for the DI of the explored molecules to check if *WEALD* is not optimising noise. If the confidence intervals for the DI of the molecules in the final sub-library are huge, this raises the question of the interest of testing these molecules in laboratories. The concept of the indistinguishable optimal zone allows to gather molecules that, due to QSAR models error, cannot be distinguished from the optimal one. On the *diabetes* library, some molecules were totally uncertain with a huge confidence interval for the DI and, due to this uncertainty, were not significantly worse than the optimum found.

Finally, the definition of the applicability model of a QSAR model is crucial to check if the prediction is extrapolated. This may be often the case (half of the *diabetes* library), but chemists are forced to use such uncertain tools as they have no other safer techniques available to select the most promising compounds among the huge combinatorial library.

6.5 Appendices

Appendix 1: MLR models

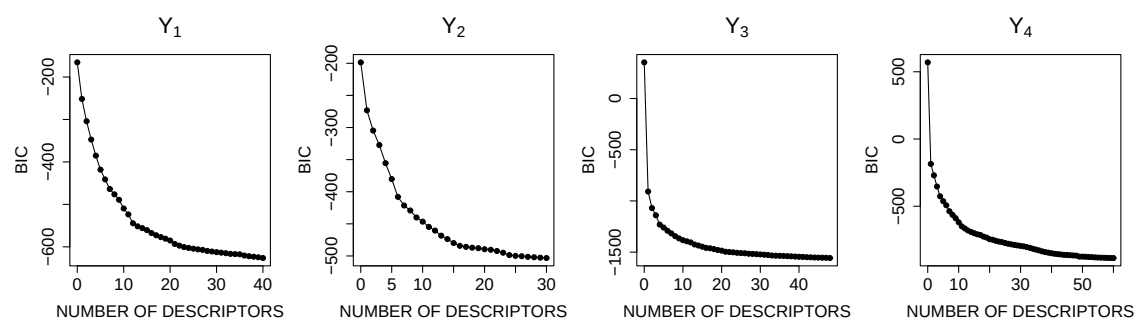


Figure 6.18: Evolution of the BIC criterion as a function of the number of entered descriptors in each MLR model fitted using forward method.

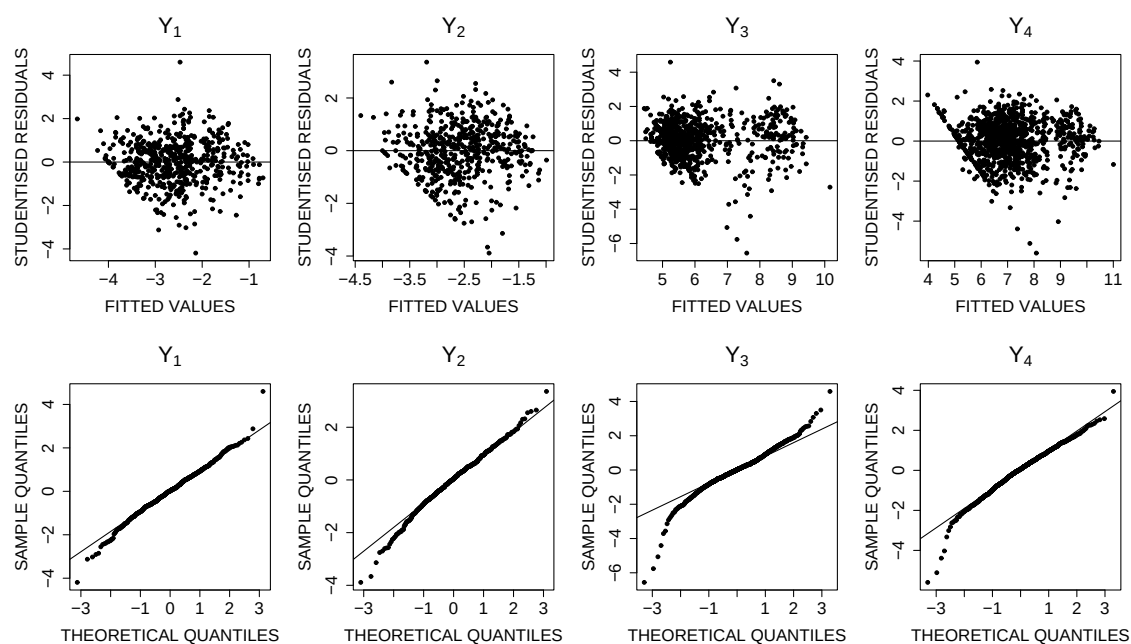


Figure 6.19: Residuals analysis for each MLR models

Appendix 2: PLSR models

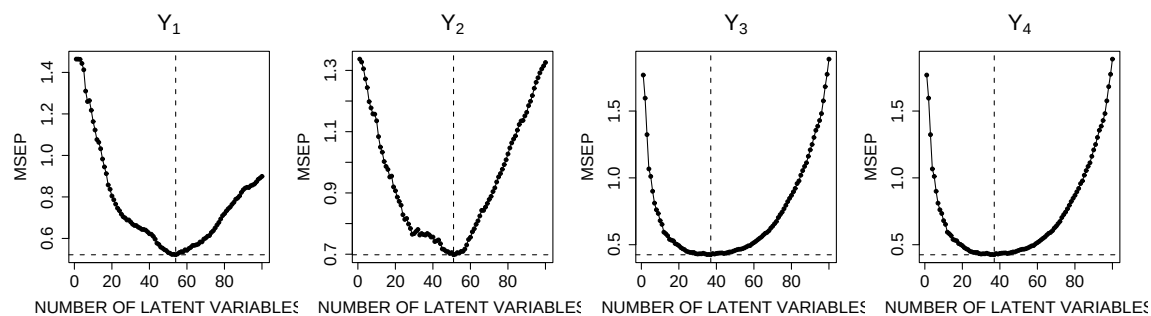


Figure 6.20: Evolution of the MSEP criterion (bias-corrected 10-foldCV) as a function of the number of entered latent variables in each PLSR model.

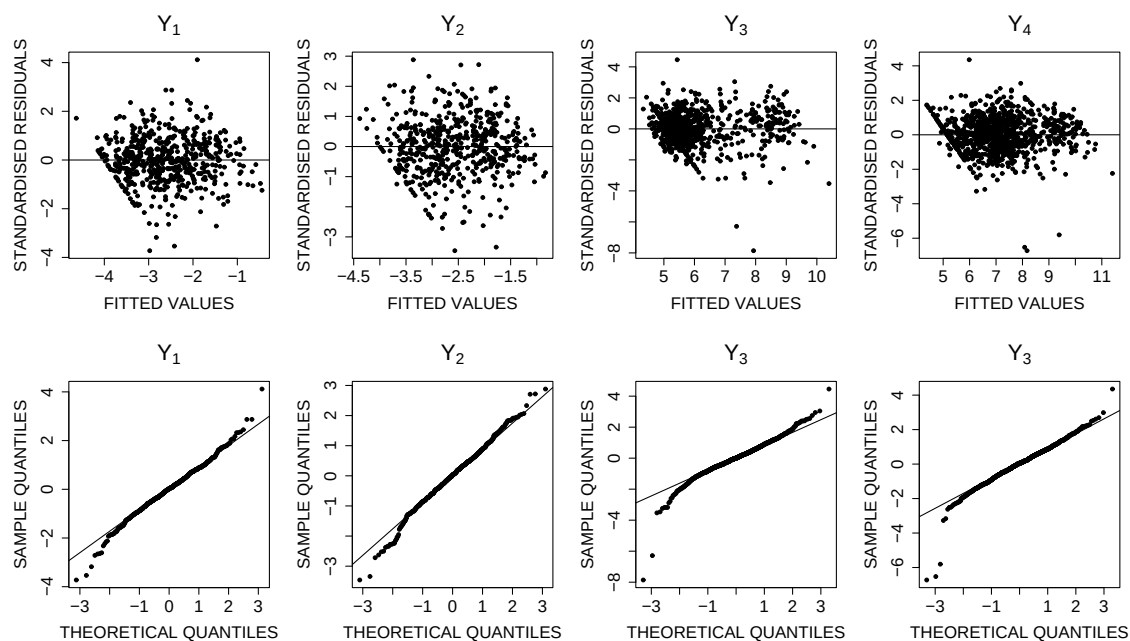


Figure 6.21: Residuals analysis for each PLSR models

Appendix 3: RT models

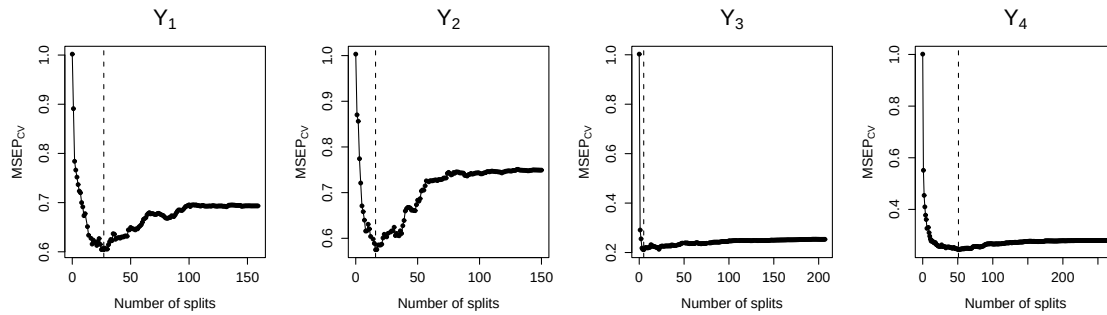


Figure 6.22: Evolution of the Mean Square Error of Prediction estimated by 10-fold Cross-Validation (2.17) as a function of the number of splits.

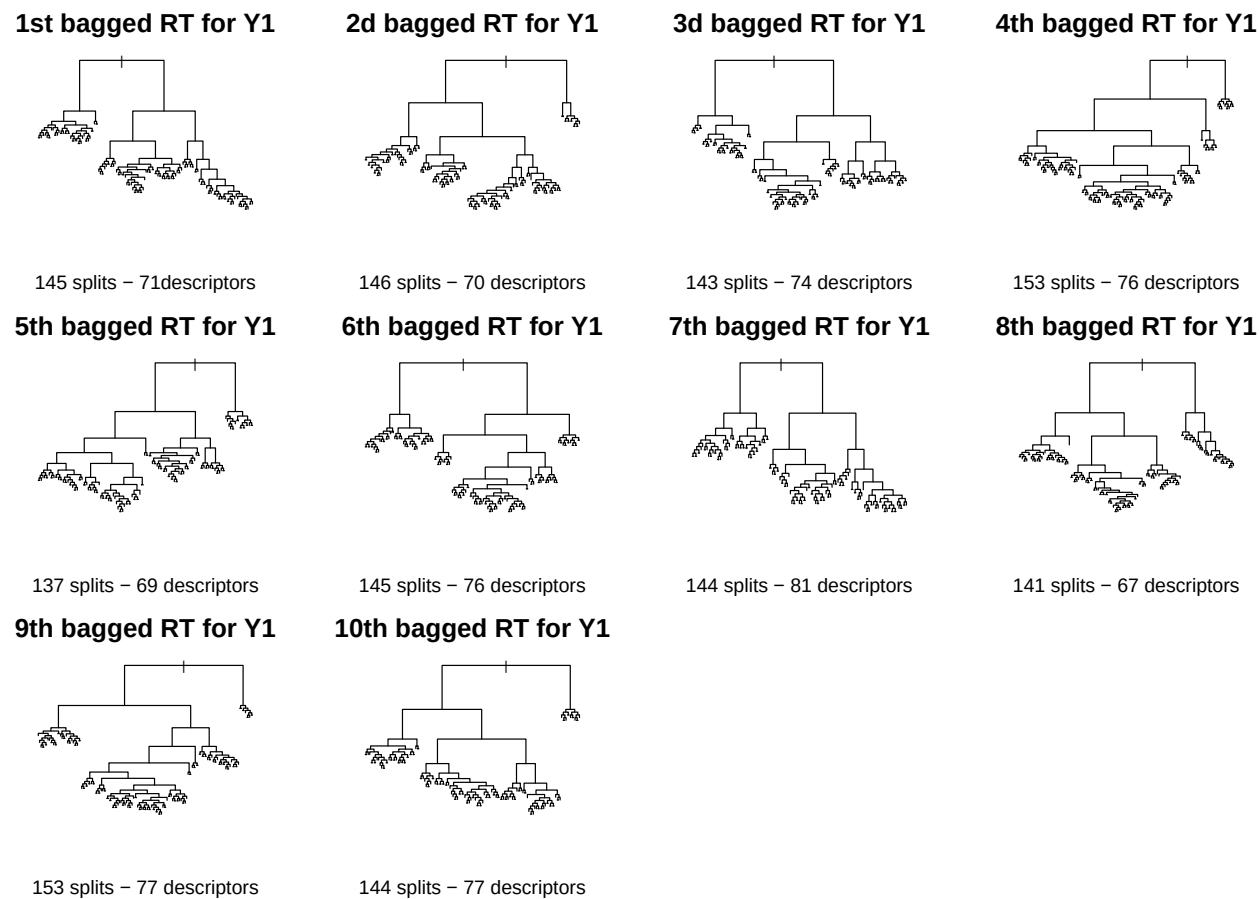
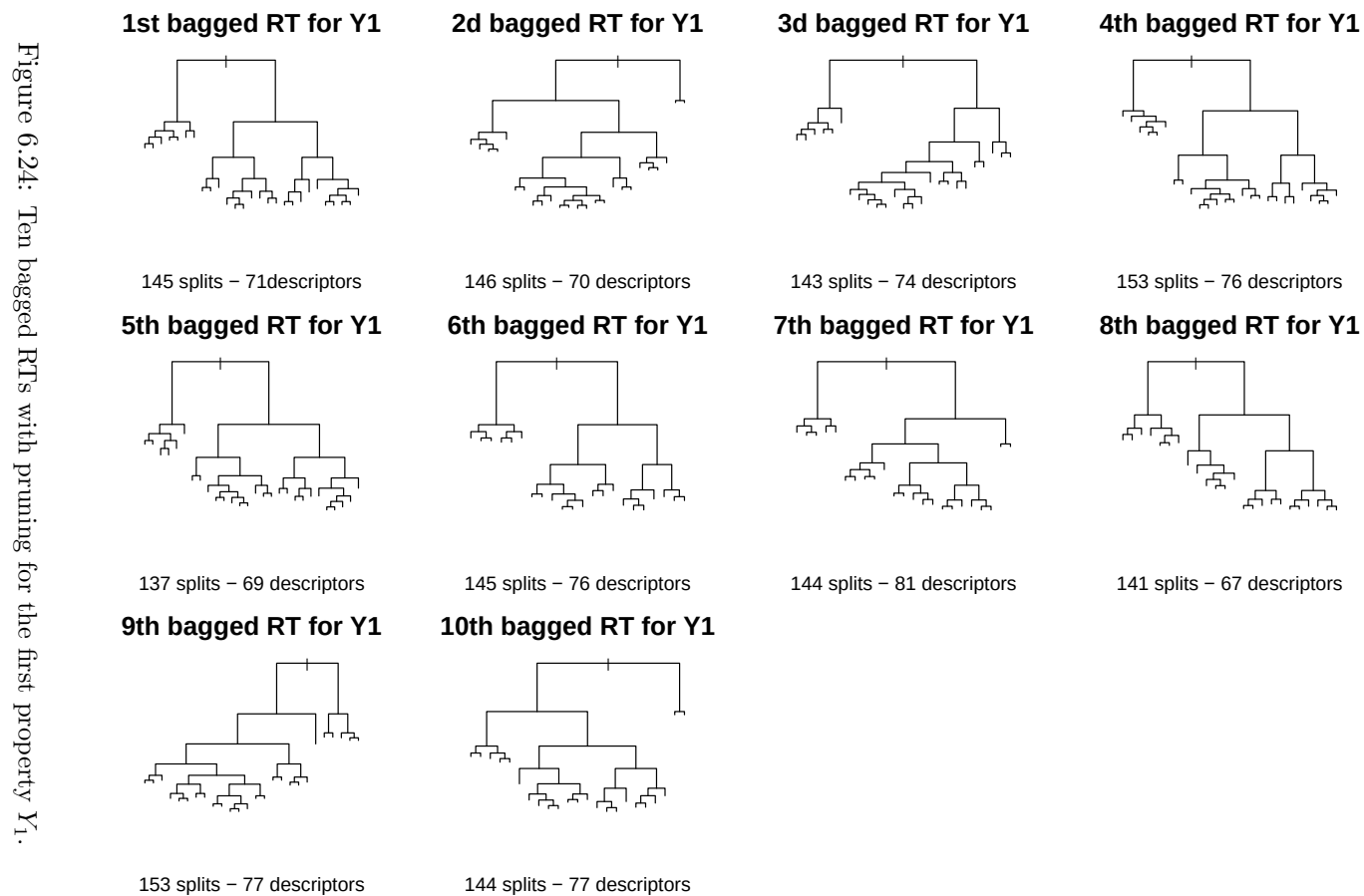


Figure 6.23: Ten bagged RTs without pruning for the first property Y1.

Figure 6.24: Ten bagged RTs with pruning for the first property Y_1 .

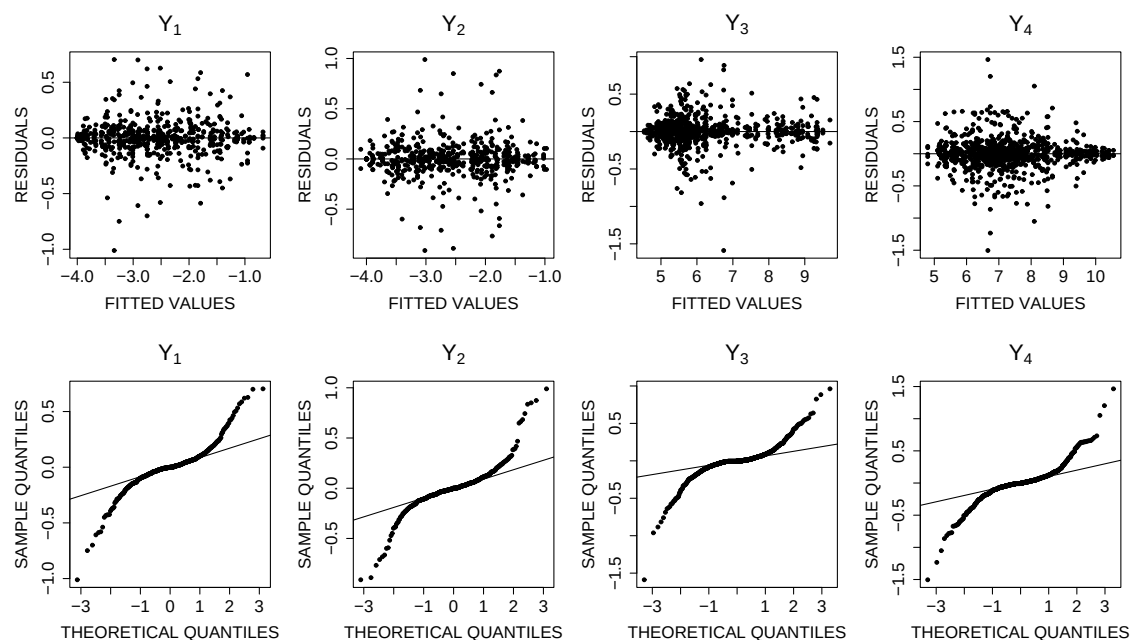


Figure 6.25: Residuals analysis for each RT models without pruning

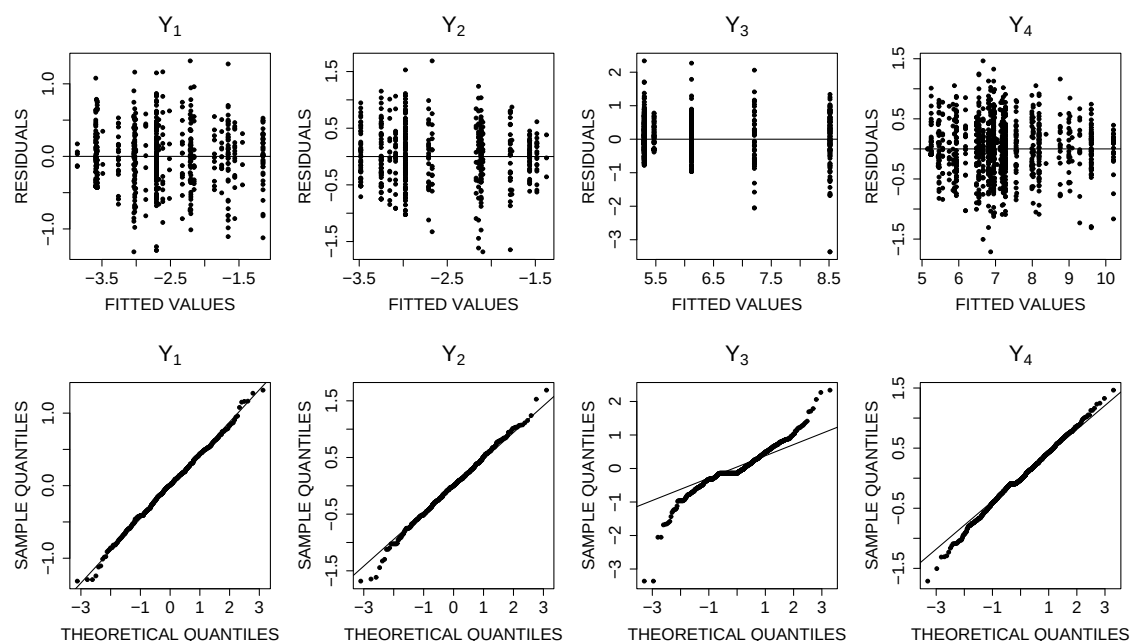


Figure 6.26: Residuals analysis for each RT models with pruning

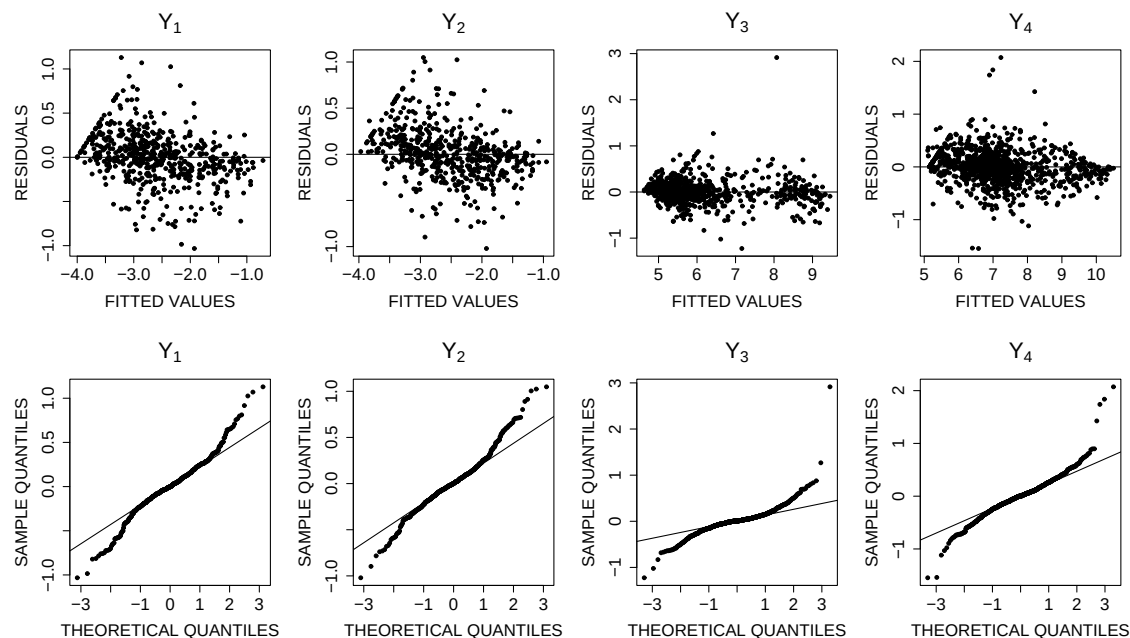


Figure 6.27: Residuals analysis for bagged RT models without pruning

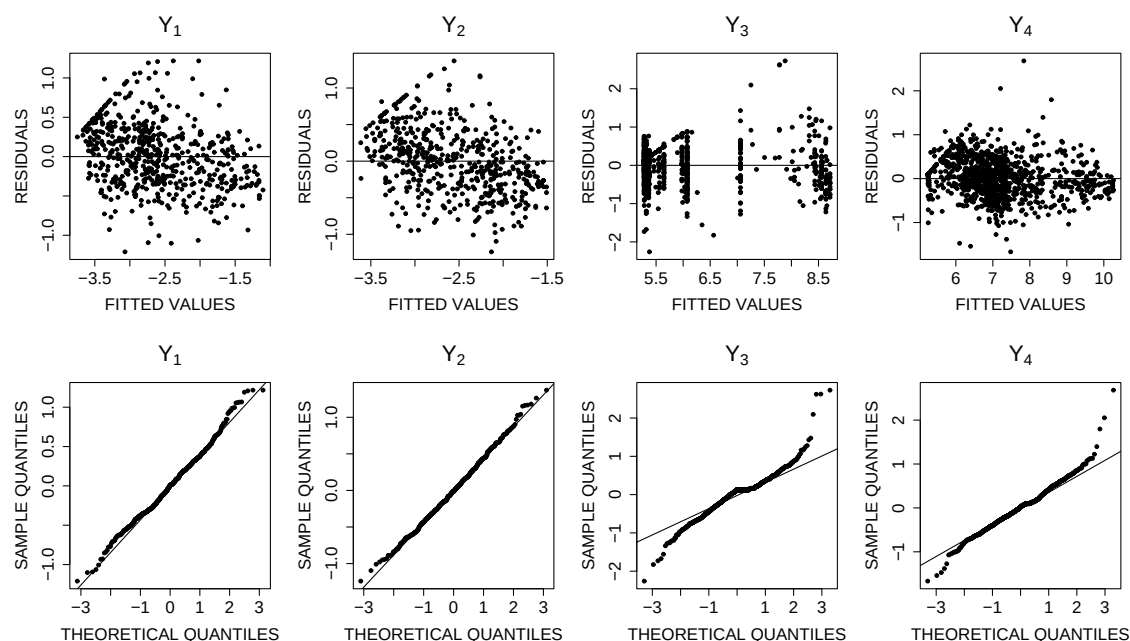


Figure 6.28: Residuals analysis for bagged RT models with pruning

Appendix 4: WEALD application

REAGENTS			ESTIMATED PROPERTIES				DESIRABILITY	
A	E	W	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{D}^C(\mathbf{x})$	$\hat{D}^S(\mathbf{x})$
A73	E35	W27	-1.31	-2.91	3.63	5.33	0.255	0.252
A11	E100	W74	-0.36	-2.66	5	3.85	0.283	0.278
A59	E112	W37	-2.29	-2.49	3.82	5.22	0.296	0.293
A55	E111	W47	-3.33	-1.83	3.92	5.2	0.301	0.298
A30	E17	W100	-1.67	-1.71	4.84	4.62	0.311	0.306
A59	E37	W21	-2.62	-2.26	3.88	5.59	0.31	0.306
A21	E40	W38	-3.49	-1.92	4.99	2.39	0.314	0.309
A42	E42	W32	-1.62	-1.86	4.6	5.43	0.318	0.313
A40	E17	W51	-0.69	-3.23	5.22	3.36	0.323	0.317
A30	E30	W36	-2.22	-2.6	4.92	3.14	0.345	0.339
A61	E42	W30	-2.97	-2.24	4.24	4.74	0.345	0.34
A28	E15	W36	-1.93	-1.97	5.6	3.52	0.349	0.343
A4	E42	W84	-2	-2.24	4.96	3.95	0.35	0.343
A3	E87	W150	-2.73	-2.04	4.77	4.02	0.355	0.349
A46	E19	W21	-1.72	-3.77	4.84	2.99	0.36	0.354
A21	E114	W64	-3.55	-2.52	5.51	1.98	0.362	0.355
A27	E77	W32	-2.57	-2.32	4.94	3.55	0.363	0.356
A28	E43	W30	-2.63	-2.38	4.79	3.69	0.363	0.356
A2	E17	W96	-1.02	-2	6.27	4.6	0.362	0.356
A28	E81	W42	-1.69	-2.45	5.83	3.36	0.373	0.366
A59	E81	W84	-2.25	-2.89	4.64	3.99	0.374	0.367
A74	E112	W15	-2.23	-2.33	4.67	4.99	0.375	0.368
A60	E24	W126	-3.53	-3.4	3.99	4.5	0.386	0.38
A38	E17	W38	-2.28	-2.01	5.04	5.09	0.39	0.383
A30	E35	W54	-1.63	-6.86	4.83	2.84	0.391	0.385
A56	E16	W38	-2.29	-2.6	4.6	5.45	0.402	0.394
A38	E81	W21	-2.49	-2.47	5	4.33	0.407	0.399
A14	E71	W120	-3.24	-2.83	4.32	4.95	0.409	0.401
A22	E25	W36	-3.23	-2.61	6.05	2.44	0.41	0.403
A43	E36	W135	-3.62	-3.25	4.47	3.64	0.411	0.404
A36	E14	W37	-2.57	-1.77	5.42	5.29	0.416	0.409
A76	E14	W74	-1.99	-2.56	5.15	4.9	0.42	0.411
A30	E6	W15	-1.53	-2.91	5.5	4.44	0.422	0.414
A12	E14	W42	-2.6	-1.54	5.74	5.66	0.422	0.415
A46	E80	W67	-3.42	-4.7	5.13	2.11	0.433	0.425
A12	E40	W119	-3.28	-2.98	4.59	4.56	0.439	0.431
A43	E90	W38	-3.92	-3.24	4.97	3.05	0.443	0.434
A77	E67	W84	-2.11	-3.41	5.05	4.32	0.448	0.439
A14	E4	W32	-3.29	-3.09	4.49	5.1	0.453	0.444
A45	E30	W38	-3.78	-3.29	4.72	3.73	0.453	0.444
A14	E119	W17	-1.02	-3.51	5.51	5.5	0.453	0.444
A43	E71	W38	-3.87	-2.9	4.9	3.78	0.455	0.445
A4	E112	W64	-1.98	-3.27	5.24	4.6	0.46	0.451
A56	E87	W14	-2.37	-3.68	4.9	4.44	0.467	0.457
A75	E112	W81	-2.28	-3	5.01	5.46	0.471	0.461
A12	E1	W119	-3.52	-3.14	4.66	4.87	0.479	0.47
A58	E16	W8	-2.43	-2.94	5.51	4.44	0.48	0.47
A4	E75	W30	-3.63	-2.36	5.3	4.69	0.485	0.475
A14	E81	W102	-2.58	-3.5	5.12	4.39	0.488	0.478
A3	E119	W8	-2.79	-2.51	5.64	4.77	0.489	0.479

REAGENTS			ESTIMATED PROPERTIES				DESIRABILITY	
A	E	W	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{D}^C(\mathbf{x})$	$\hat{D}^S(\mathbf{x})$
A53	E87	W78	-3.12	-4.08	4.89	3.87	0.493	0.483
A56	E97	W47	-2.16	-3.88	4.87	5.4	0.494	0.484
A59	E90	W51	-3.85	-3.63	4.54	4.78	0.498	0.488
A11	E40	W100	-3.04	-3.46	4.68	5.47	0.501	0.491
A60	E29	W9	-3.46	-2.92	5.09	4.7	0.505	0.494
A11	E82	W67	-3.5	-3.84	4.83	4.15	0.505	0.495
A2	E6	W10	-2.94	-2.77	5.65	4.67	0.514	0.503
A36	E25	W84	-3.89	-3.4	5.2	3.78	0.517	0.506
A76	E4	W65	-3.5	-3.92	4.45	5.64	0.516	0.506
A45	E1	W51	-3.63	-4.47	5.23	3.2	0.522	0.512
A58	E57	W84	-4.13	-3.18	4.75	5.24	0.526	0.516
A4	E82	W37	-3.44	-2.59	5.49	5.13	0.528	0.517
A16	E30	W15	-2.62	-4.28	5.13	4.44	0.53	0.519
A51	E75	W21	-3.81	-2.94	4.89	5.65	0.531	0.52
A56	E77	W21	-2.68	-3.35	5.43	4.84	0.532	0.521
A31	E3	W38	-3.83	-3.22	5.01	4.78	0.534	0.522
A48	E6	W90	-3.62	-4.15	4.85	4.29	0.534	0.523
A31	E81	W32	-3.02	-3.23	5.26	5.12	0.539	0.527
A57	E60	W17	-3.28	-3.42	4.96	5.41	0.542	0.531
A2	E37	W10	-4.2	-2.9	5.28	4.8	0.551	0.539
A16	E56	W100	-3.38	-3.29	5	5.66	0.552	0.54
A63	E82	W67	-4.47	-3.16	5.28	4.26	0.552	0.541
A11	E29	W92	-5.34	-5.84	4.78	3.38	0.555	0.546
A74	E104	W90	-3.08	-3.42	5.73	4.53	0.566	0.554
A62	E31	W65	-4.47	-4.81	4.94	3.88	0.571	0.56
A75	E57	W90	-3.95	-3.09	5.07	5.81	0.573	0.561
A6	E69	W14	-3.48	-3.19	5.27	5.65	0.578	0.566
A42	E72	W78	-3.41	-3.85	5.12	5.05	0.578	0.566
A3	E85	W100	-4.66	-3.23	4.84	6.05	0.583	0.571
A31	E71	W47	-4.12	-3.7	4.73	6.17	0.584	0.573
A28	E57	W97	-1.91	-3.71	5.95	6.57	0.589	0.578
A2	E75	W100	-3.84	-4.15	5.27	4.43	0.597	0.585
A61	E75	W64	-4.31	-3.57	5.15	5.12	0.603	0.591
A2	E69	W148	-3.88	-5.37	5.39	3.77	0.608	0.597
A3	E114	W100	-4.31	-3.67	5.12	5.32	0.613	0.6
A69	E83	W119	-4.24	-3.9	4.94	5.91	0.622	0.61
A62	E67	W94	-4.62	-5.42	4.78	4.86	0.622	0.611
A56	E71	W9	-3.71	-4.22	5.3	5.12	0.632	0.619
A61	E25	W100	-4.76	-4.31	4.94	5.22	0.632	0.619
A74	E69	W102	-4.61	-3.17	5.64	5.53	0.647	0.634
A36	E8	W64	-4.15	-4.12	5.57	4.79	0.655	0.641
A17	E83	W102	-4.25	-3.6	5.6	5.39	0.659	0.645
A71	E3	W12	-3.59	-4.42	5.82	5.16	0.683	0.67
A11	E114	W97	-3.05	-4.28	5.9	6.08	0.688	0.674
A71	E90	W81	-4.31	-5.22	5.51	4.83	0.699	0.686
A70	E83	W64	-4.55	-4.2	5.35	6.01	0.703	0.689
A57	E57	W12	-3.93	-3.59	5.86	6.7	0.708	0.694
A20	E45	W97	-4.57	-4.27	5.61	6.13	0.739	0.724
A53	E25	W48	-5.2	-4.85	5.35	5.92	0.745	0.731
A48	E25	W98	-3.47	-4.64	6.29	5.96	0.746	0.733

Table 6.7: Initial sub-library of 100 compounds chosen at random in the *diabetes* combinatorial library

REAGENTS			ESTIMATED PROPERTIES				DESIRABILITY		95% CI for	
A	E	W	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{D}^C(\mathbf{x})$	$\hat{D}^S(\mathbf{x})$	$D^S(\mathbf{x})$	
A48	E43	W97	-3.47	-4.69	6.59	7.04	0.790	0.778	0.751	0.804
A55	E115	W48	-5.09	-4.51	5.72	6.71	0.793	0.778	0.749	0.806
A70	E43	W48	-5.29	-4.79	5.49	6.97	0.792	0.778	0.749	0.805
A70	E45	W98	-4.67	-3.82	6.15	7.81	0.792	0.778	0.743	0.810
A71	E45	W5	-3.7	-4.74	6.06	7.58	0.792	0.779	0.731	0.820
A53	E23	W98	-3.86	-5.22	6.07	6.45	0.792	0.779	0.748	0.807
A48	E45	W8	-5.91	-4.78	5.86	5.6	0.793	0.779	0.746	0.809
A47	E45	W98	-4.66	-3.85	6.16	7.7	0.793	0.779	0.741	0.814
A69	E24	W97	-3.84	-5.06	6.29	6.26	0.792	0.780	0.753	0.804
A48	E24	W96	-3.97	-5.59	6.38	5.63	0.791	0.780	0.754	0.804
A57	E29	W97	-4.83	-4.69	5.9	6.24	0.794	0.780	0.752	0.805
A31	E104	W48	-4.14	-4.99	6.21	6.05	0.793	0.780	0.750	0.807
A69	E23	W12	-4.09	-5.06	6.13	6.19	0.793	0.780	0.746	0.811
A48	E90	W12	-3.7	-5.14	6.3	6.44	0.793	0.781	0.750	0.808
A48	E45	W98	-4.1	-4.39	6.09	7.42	0.795	0.782	0.745	0.814
A55	E29	W97	-4.95	-4.44	5.91	6.51	0.796	0.782	0.754	0.807
A47	E23	W48	-5.63	-5.9	5.38	6.25	0.794	0.782	0.757	0.804
A48	E115	W12	-3.57	-4.81	6.42	7.01	0.795	0.782	0.752	0.810
A51	E29	W97	-4.3	-4.95	5.98	6.4	0.796	0.783	0.755	0.808
A69	E4	W12	-4.35	-4.64	5.96	6.81	0.797	0.783	0.754	0.809
A44	E45	W12	-4.85	-4.71	6.38	5.56	0.796	0.783	0.733	0.826
A53	E29	W97	-4.29	-4.85	5.95	6.63	0.797	0.783	0.756	0.808
A47	E24	W97	-4.25	-4.93	6.07	6.33	0.797	0.784	0.756	0.809
A71	E45	W96	-5.13	-4.61	5.89	6.22	0.798	0.784	0.740	0.822
A69	E24	W12	-4.15	-4.97	6.08	6.39	0.797	0.784	0.752	0.813
A53	E45	W96	-4.97	-4.31	5.77	7.37	0.798	0.784	0.749	0.815
A48	E67	W12	-3.89	-5	6.12	6.72	0.797	0.784	0.754	0.811
A48	E1	W12	-4.38	-4.8	6.22	6.09	0.798	0.785	0.754	0.812
A57	E45	W102	-4.9	-5.07	5.68	6.48	0.799	0.785	0.740	0.824
A51	E23	W48	-5.52	-5.89	5.61	5.75	0.799	0.786	0.760	0.810
A47	E43	W97	-3.92	-4.44	6.4	7.12	0.799	0.786	0.758	0.812
A47	E23	W96	-4.7	-5.01	5.93	6.12	0.800	0.787	0.758	0.813
A53	E45	W97	-4.57	-4.29	5.85	7.99	0.800	0.787	0.749	0.821
A69	E1	W12	-4.16	-4.52	6.22	6.87	0.801	0.787	0.758	0.814
A71	E45	W48	-6.52	-5.5	5.48	5.99	0.800	0.787	0.752	0.819
A70	E23	W98	-3.94	-4.81	6.27	6.69	0.800	0.787	0.758	0.814
A70	E115	W48	-4.64	-4.81	5.62	7.64	0.802	0.788	0.761	0.813
A47	E23	W98	-3.92	-4.9	6.3	6.6	0.802	0.789	0.760	0.815
A56	E29	W97	-4.18	-4.8	6.31	6.31	0.802	0.789	0.765	0.811
A58	E45	W12	-3.7	-5.7	6.17	6.5	0.802	0.790	0.737	0.834
A71	E23	W98	-3.78	-5.85	6.16	6.31	0.802	0.790	0.760	0.817
A51	E45	W97	-4.54	-4.37	5.91	7.7	0.804	0.790	0.752	0.824
A48	E43	W12	-3.99	-4.81	6.29	6.69	0.803	0.791	0.760	0.818
A70	E24	W97	-4.35	-4.82	6	6.72	0.804	0.791	0.763	0.816
A71	E115	W48	-4.63	-5.87	5.51	7.02	0.804	0.792	0.763	0.818
A31	E56	W48	-5.33	-5.05	5.85	5.93	0.805	0.792	0.763	0.818
A47	E43	W48	-4.85	-4.71	5.77	7.01	0.806	0.792	0.764	0.817
A48	E24	W97	-3.95	-5.27	6.39	6.14	0.804	0.792	0.767	0.816
A48	E75	W12	-4.72	-4.52	6.21	6.34	0.806	0.793	0.757	0.825
A48	E69	W12	-3.84	-5.22	6.5	6.26	0.806	0.794	0.765	0.820

REAGENTS			ESTIMATED PROPERTIES				DESIRABILITY		95% CI for	
A	E	W	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{D}^C(\mathbf{x})$	$\hat{D}^S(\mathbf{x})$	$D^S(\mathbf{x})$	
A47	E115	W48	-4.69	-4.84	5.77	7.05	0.808	0.794	0.766	0.819
A38	E45	W97	-5.68	-4.01	5.84	7.76	0.808	0.794	0.763	0.822
A57	E45	W48	-5.82	-5.83	5.36	6.7	0.808	0.795	0.760	0.826
A55	E45	W5	-3.99	-4.82	5.92	8.2	0.809	0.796	0.741	0.841
A69	E29	W97	-4.39	-4.86	6.1	6.55	0.809	0.796	0.771	0.818
A31	E23	W5	-3.66	-5.1	6.64	6.59	0.807	0.796	0.753	0.832
A70	E45	W48	-6.64	-4.49	5.49	7.52	0.809	0.796	0.768	0.822
A70	E45	W97	-5.15	-3.96	6.04	7.83	0.809	0.796	0.761	0.827
A47	E56	W48	-4.78	-4.99	5.88	6.48	0.810	0.796	0.771	0.819
A55	E45	W48	-6.42	-5.73	5.26	7.14	0.809	0.797	0.762	0.829
A69	E29	W12	-4.38	-5.1	6.09	6.37	0.811	0.797	0.765	0.826
A71	E56	W48	-4.68	-5.93	5.72	6.33	0.810	0.798	0.772	0.821
A69	E45	W12	-4.7	-4.11	6.14	7.75	0.811	0.798	0.760	0.830
A71	E45	W12	-4.42	-4.64	6.15	6.81	0.811	0.798	0.757	0.834
A48	E57	W12	-4.23	-4.69	6.06	7.35	0.812	0.798	0.767	0.826
A16	E45	W97	-5.29	-4.36	5.81	7.37	0.812	0.799	0.768	0.826
A48	E23	W12	-4.33	-5.71	6.06	6.09	0.812	0.799	0.766	0.829
A55	E45	W102	-5.11	-4.97	5.69	6.87	0.813	0.799	0.754	0.838
A57	E45	W98	-4.08	-5.24	6.01	6.96	0.812	0.799	0.752	0.840
A48	E24	W12	-4.52	-5.64	6.01	6.05	0.813	0.800	0.768	0.829
A69	E45	W102	-7.06	-4.14	5.69	7.62	0.815	0.801	0.771	0.829
A56	E45	W12	-4.18	-5.59	5.79	7.28	0.816	0.803	0.754	0.844
A55	E45	W98	-4.37	-4.98	5.87	7.39	0.816	0.803	0.754	0.844
A48	E83	W12	-4.08	-5.38	6.39	6.2	0.815	0.803	0.772	0.830
A55	E45	W12	-4.33	-4.65	6.1	7.32	0.816	0.803	0.749	0.848
A48	E56	W12	-3.98	-4.99	6.44	6.7	0.816	0.804	0.777	0.828
A70	E56	W48	-4.88	-4.91	5.81	7	0.818	0.804	0.780	0.827
A47	E29	W97	-4.82	-4.89	6	6.54	0.819	0.805	0.780	0.828
A69	E115	W12	-3.84	-4.56	6.55	7.92	0.819	0.808	0.781	0.832
A47	E45	W48	-5.92	-4.55	5.64	7.6	0.822	0.809	0.780	0.834
A48	E29	W97	-4.32	-5.13	6.23	6.52	0.821	0.809	0.785	0.830
A48	E45	W96	-5.02	-4.18	6.33	7.06	0.823	0.810	0.773	0.842 *
A57	E45	W12	-4.1	-4.52	6.5	7.59	0.824	0.812	0.755	0.859 *
A56	E45	W97	-3.98	-5.74	5.99	7.27	0.824	0.812	0.769	0.849 *
A57	E45	W96	-4.58	-5.39	5.93	6.73	0.826	0.813	0.769	0.850 *
A31	E115	W48	-4.83	-4.76	6.07	6.84	0.827	0.813	0.781	0.842 *
A69	E83	W12	-4.12	-5.06	6.46	6.76	0.828	0.816	0.789	0.839 *
A58	E45	W97	-3.91	-5.85	6.21	6.91	0.828	0.816	0.772	0.854 *
A71	E45	W97	-5.07	-5	6.07	6.42	0.831	0.817	0.778	0.851 *
A47	E45	W97	-4.88	-4.32	6.27	7.41	0.831	0.818	0.784	0.847 *
A57	E45	W5	-3.79	-5.07	6.32	8.15	0.831	0.819	0.766	0.862 *
A48	E45	W97	-4.57	-4.59	6.21	7.66	0.835	0.822	0.790	0.850 *
A50	E45	W97	-5.19	-4.2	6.27	7.64	0.836	0.823	0.791	0.851 *
A55	E45	W96	-4.93	-5.22	5.9	6.97	0.836	0.823	0.779	0.860 *
A48	E29	W12	-4.81	-5.64	6.17	6.13	0.836	0.824	0.793	0.851 *
A55	E45	W97	-4.97	-5.11	5.86	7.38	0.840	0.827	0.781	0.864 *
A57	E45	W97	-4.51	-5.55	5.97	7.17	0.841	0.828	0.785	0.863 *
A71	E45	W98	-4.6	-4.91	6.14	7.53	0.843	0.830	0.793	0.862 *
A48	E45	W12	-4.94	-4.77	6.11	7.52	0.847	0.834	0.798	0.865 *
A69	E45	W97	-6.03	-4.4	6.18	7.57	0.858	0.845	0.818	0.869 optimum

Table 6.8: Final sub-library of 100 compounds selected by *WEALD* in the *diabetes* combinatorial library (* = molecule not significantly worse than the optimal one).

Bibliography

- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and regression trees*, Belmont, CA: Wadsworth International Group.
- Breiman, L. (1996), "Bagging predictors", *Machine Learning*, 24(2), 123-140.
- de Jong, S. (1993), "SIMPLS: An alternative approach to partial least squares regression", *Chemometrics and Intelligent Laboratory Systems*, 18, 251-263.
- Mevik, B.-H. and Cederkvist, H.R. (1993), "Mean Squared Error of Prediction (MSEP) estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR)", *Journal of Chemometrics*, 18(9), 422-429.
- Vanden Branden, K. and Hubert, M. (2004), "Robustness properties of a robust PLS regression method", *Analytica Chimica Acta*, 515, 229-241.

CONCLUSION

This thesis proposes an integrated methodology to optimise virtually combinatorial libraries of molecules in the early stages of the drug discovery process. This work was motivated by a real need of the pharmaceutical industry and a concrete library optimisation problem proposed by Eli Lilly and Company. It gathers in a coherent framework existing and new tools of statistics and chemometrics, mainly:

- (-) the development and validation of QSAR models to predict drugability properties,
- (-) the definition of a desirability index to summarise those properties and assessment of the propagation of QSAR models predictions, and
- (-) an efficient algorithm to screen the combinatorial library and select the most promising compounds.

The whole methodology has been applied for the screening of a real combinatorial library (the *diabetes* library) provided by Eli Lilly and Company.

The chemometrics literature about QSAR has been reviewed. The advantages and the failures of this emerging techniques have been emphasised. QSAR models constitute an extraordinary tool to speed up the drug discovery process by allowing the prediction of important drugability properties such as the ADME properties, without doing real chemical synthesis. But, even if there is a growing interest in the statistical validation of QSAR models, they are still often applied intensively for the screening of molecules without checking if those new molecules are similar to the molecules used to train the models and without assessing the reliability of those predictions.

This thesis underlines the importance of defining the applicability domain of a QSAR model to measure possible extrapolation and reviews the statistical literature to provide tools to estimate the (local) prediction variance for a new compound with linear regression, PLS regression or regression tree. The application on the *diabetes* library reveals that a lot of molecules may be far outside the applicability domain of a QSAR model and some predictions, even for a new molecule inside the applicability domain, may be highly uncertain.

The screening of a combinatorial library of molecules is a multicriteria optimisation problem as the quality of a compound is measured by several drugability properties. We propose to summarise the estimated drugability properties of a molecule by a desirability index, in $[0, 1]$, through the use of differentiable desirability functions, like the new Normal cdf based DFs. For a given molecule in the combinatorial library, we suggest to quantify its quality by the expected desirability of the corresponding responses, instead of the classical desirability of the corresponding expected responses. In all performed simulations and in the application on the *diabetes* library, this difference in the optimised criteria is small. An estimator for the expected desirability index based on the estimated expected responses and the estimated models error variances is presented, resulting from the application of the multivariate Taylor expansion.

The desirability index is usually applied in the design of experiments literature without taking into account the uncertainty of the estimated responses to optimise. This thesis proposes the construction of an asymptotic confidence interval based on the Delta method theorem to measure the propagation of the uncertainty of the estimated responses on the estimated desirability index. Although this is an approximative method, it reveals through simulations to perform well in practice if the responses are not too noisy and the desirability functions are not too stiff. We also define the concept of indistinguishable optimal zone to quantify the uncertainty of the optimal factors settings found. This methodology for the uncertainty of the desirability index and the corresponding optimum can be applied to the optimisation of molecules, as well as to any other multicriteria optimisation like the improvement of the quality of a tire tread compound conducted by Derringer and Suich.

On the basis of QSAR models predictions, one can associate to any compound in a combinatorial library a desirability index summarising its drugability potential. Nevertheless, the computations of chemical descriptors and the corresponding QSAR predictions are computer intensive and often too long for the whole combinatorial library. A new algorithm (*WEALD*) has been developed to select a subset of promising compounds in a combinatorial library. On the basis of QSAR models predictions and their aggregation in a desirability index, *WEALD* explores part of the available compounds to retain the compounds with the highest desirability indexes. The principle of this algorithm is to perform exchanges between reagents. A subtle weighting of the reagents allows to converge rapidly through a sublibrary with the most promising compounds.

WEALD has been applied on two different combinatorial libraries and reveals to find better final compounds than two other screening algorithms published in the chemometrics literature (ultrafast algorithm and PICCOLO algorithm). *WEALD* requires also less computations, converging more rapidly.

The methodology to measure uncertainty propagation on the desirability index can be integrated in the *WEALD* algorithm. If the confidence intervals for the desirability of explored compounds are large, it is the sign that the QSAR models are maybe not enough predictive to provide a meaningful ranking of the molecules. The effective superiority of the optimum compound against the explored molecules can be assessed by applying the concept of the indistinguishable optimal zone.

Those different tools constitute our statistical contribution to the virtual optimisation of combinatorial libraries and the validation and application of QSAR models. Some additional work can be done to improve the proposed methodology and boost its practical application in various fields.

First, applying applicability domain definition, statistical validation tools and (local) prediction error assessment on a published QSAR model (other than MLR?) would demonstrate the relevance and practical purpose of such tools when using QSAR models for the prediction of new compounds. One could also study the extension of the methodology to other classes of (QSAR) models, like the more and more popular neural networks or models for binary or categorial activity response like logistic regression.

The relevance of the conditional independence could be studied on other published multicriteria optimisation problems than the Derringer and Suich one or directly with practitioners in the industry. The desirability index users should be questioned about their interest in the proposed methodology to measure desirability index uncertainty and take it into account in comparing solutions.

Finally, the principles of *WEALD* can be tested on other chemistry problems, like the building of a chiral catalyst library studied by N. Vriamont in his PhD thesis (Chemistry department, Université catholique de Louvain, July 2007).

