UNIVERSITÀ DEGLI STUDI FIRENZE

UNIVERSITÀ DEGLI STUDI DI PERUGIA

Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

DOTTORATO DI RICERCA
IN MATEMATICA, INFORMATICA, STATISTICA

CURRICULUM IN STATISTICA
CICLO XXXI

Sede amministrativa Università degli Studi di Firenze
Coordinatore Prof. Graziano Gentili

Three projects for the design of experiments: choice experiments, Kriging and split-plot designs

Settore Scientifico Disciplinare SECS-S/03

Dottoranda:
Nedka Dechkova Nikiforova

Tutore
Prof.ssa Rossella Berni

Coordinatore
Prof. Graziano Gentili

Anni 2015/2018

Abstract

This PhD thesis is composed of three projects related to the design and analysis of experiments. The first project pertains to the construction of optimal designs for choice experiments. It has been developed also in collaboration with professor Jesús Fernando López-Fidalgo during my first PhD visiting period at the University of Navarra, Pamplona, Spain. Our thanks to Prof. Patrizia Pinelli, Department of Statistics, Computer Science and Applications - PHYTOLAB Laboratory, Scientific and Technological Pole, University of Florence, for providing us the HPLC analysis for the real case-study faced in this first project. The second project is related to computer experiments and Kriging modelling, applied to solve a complex engineering problem in the railway field. Regarding the second project about Kriging, we want to acknowledge Prof. Ing. Luciano Cantone, Department of Engineering for Enterprise "Mario Lucertini", University of Rome "Tor Vergata", for providing us the data and the engineering details for the problem under study. The third project deals with randomization issues in a split-plot design, due to production process requirements of a product in the field of mechanical engineering. This last project has been developed also in collaboration with professor G. Geoffrey Vining during my second PhD visiting period at the Virginia Tech University, Blacksburg, Virginia, USA. Moreover, we want also to acknowledge Dott. Ing. Francesco Bertocci and Esaote SpA, Florence, Italy for the collaboration and contribution to the engineering and technological features.

In what follows, a brief summary for each project is reported.

1. **OPTIMAL APPROXIMATE CHOICE DESIGNS WITH CORRELATED PREFERENCES THROUGH A COMPOUND CRITERION**

   In this project, we propose an innovative approach for the construction of heterogeneous choice designs with correlated choice preferences. Differently from existing research in the choice design literature that make use of an exact design framework to build optimal choice designs, we propose the construction of optimal heterogeneous choice designs based on an approximate design theory, and under the Panel Mixed Logit model structure that explicitly takes account of the correlation between the responses of a respondent facing a sequence of choice-sets. Our proposed approach allows us

to obtain optimal heterogeneous choice designs composed of groups of choice-sets to be administered to a proportion of respondents according to the optimal weights. We show the efficiency of our proposal through an application to a real case study that concerns the analysis of the consumers' preferences for coffee, integrating a choice experiment with the consumer sensory tests. To this end, we develop our proposal under a compound design criterion. Moreover, we present the estimation results of the Panel Mixed Logit model related to the proposed optimal heterogeneous choice design we applied to our real case study, which are very satisfactory, by further confirming the validity of our innovative proposal.

2. **LATIN HYPERCUBE DESIGNS BASED ON STRONG ORTHOGONAL ARRAYS AND KRIGING MODELLING TO IMPROVE THE PAYLOAD DISTRIBUTION OF TRAINS**

   This projects deals with computer experiments and Kriging modelling to improve the braking performance for freight trains. We focus on the payload distribution along the train, so as to reduce the effects of in-train forces, e.g. compression and tensile forces, among vehicles during a train emergency braking. The topic is particularly relevant for Railway Undertakings, especially in Europe, where a series of codes regulates international freight traffic. To this end, we propose a novel approach to improve the payload distribution of trains through a suitable design for the computer experiment and Kriging modelling. More precisely, we build a Latin Hypercube design based on strong orthogonal arrays for the computer experiment that achieves very good space-filling properties with a relatively low number of experimental runs. Kriging models with anisotropic covariance functions are subsequently applied to find the optimal payload distribution able to reduce the in-train forces. Moreover, differently from other researches in this field, where the entire train was characterized by a unique payload distribution, in the present application we consider that the train is divided in several sections, each one composed of different wagons. Therefore, each train section is characterized by its own payload distribution: having different train sections gives the possibility to optimize trains that deliver their payload along their route. Furthermore, it also allows for better understanding of the best payload distribution along the entire train, so as by further improving the freight train efficiency in terms of braking performance.

3. **THE IMPACT OF NOT RANDOMIZING A SPLIT-PLOT EXPERIMENT AND HOW TO DETECT ITS EFFECT**

This project deals with lacking of randomization in an industrial split-plot experiment. More precisely, a split-plot design is planned to improve the production process of an ultrasound transducer for medical imaging. Due to constraints on how the company could conduct the experiment, some of the factors in the design are not randomized. To this end, we focus on the possible consequent impact of lacking of randomization in the split-plot design. More precisely, we carry out a simulation study based on the real one, in order to examine the implications of lacking of randomization on the factor estimates and on the corresponding residual values.

Table of contents

List of figures

List of tables

Chapter 1

Literature review

1.1 Introduction

Design of experiments is a wide and fundamental methodology of the statistics theory. It is based on a theoretical background which relates to the main aspects of the literature in statistics and probability since 1920s. At the beginning, the classical experimental design was mainly directed to study specific problems sourcing in agriculture. Nevertheless, since 1950s, many theoretical developments have been formulated in order to extend the basic and simple classical designs and linear modelling; furthermore, these new theoretical issues have been differently developed by considering the specific fields of application. Optimal design criteria and Response Surface Methodology (RSM) date back to this period: the General Equivalence Theorem of Kiefer (1959) is a milestone and basic theorem for the optimal design theory, while the fundamental paper of Box and Wilson (1951) is the first introduction to the RSM. A great impulse to the evolution of the experimental design theory was further originated by the development of computer experiments (Sacks et al., 1989). Moreover, the design of experiments theory began to be succefully applied and further developed in new fields, such as marketing. For example, preference measurements and choice experiments are considered the main method to study and analyze the consumers' behaviour (Train, 2003).

Most of the fundamental elements and principles of the design of experiments have been introduced during the 1920's and 1930's by R.A. Fisher and F. Yates precisely in the field of the agricultural experimentation. Their contributions have been subsequently applied in all areas of scientific investigations, as well as in many areas of industrial production process researches and developments. Box and Wilson (1951) were the first to recognize the importance of the design and analysis of experiments in the technological field, in which they adapted classical agricultural experimentation methods. They founded the RSM, that in the years ahead completely changed the way that engineers, scientists and statisticians approached

to the industrial experimentation. Box (1999) referred to this process as a *sequential learning* in which a collection of statistical and mathematical techniques, combined with engineering knowledge, are applied to develop, improve and optimize industrial processes, in order to learn what factors are important to our investigation and which settings are required to improve one or more response variables. From 1980 and 1990 onwards, this research area has been largely studied and developed to efficiently plan, execute, analyze and optimize industrial experimentations. In this context of continuous development, take place the renewed attention toward the split-plot experiment, widely studied and used in the RSM setting.

More recently, some complex engineering and technological processes cannot be longer properly represented by physical experimentation due to costs, or even, to the impossibility to perform physical measurements. To this end, computer experiments are increasingly used in this field, where a computer code or simulator is run in order to mathematically represent the physical system under study. In the seminal contribution to computer experiments (Sacks et al., 1989), the authors introduced the concept of simulated designs, which were substantially different from physical and classical experimental designs (Cox and Reid, 2000). More precisely, there is a completely different approach to build the design for computer experiments with respect to classical design construction techniques. In fact, space-filling designs are the most used and preferred ones in this field, given that they spread the design points as uniformly as possible in order to observe the response in the entire design space. Despite the various methods to build space-filling designs (e.g. low-discrepancy sequencies (Niederreiter, 1992; Niederreiter, 2008), uniform designs (Fang et al., 2000)), the class of Latin Hypercube (LH) designs is one of the most commonly used and preferred one. Moreover, when considering computer experiments, the observations are predicted according to a simulated model of the process under study, in order to deeply analyse the relation between input and output variables. To this end, specific metamodels are used for the analysis of computer experiments that represent a valid approximation of the computer code, and act as statistical interpolators of the simulated input-output data. One of the most appropriate and widely used one is Kriging (Krige, 1951; Sacks et al., 1989), largely applied, especially recently, to study complex tecnological and engineering issues.

Optimal designs are a class of experimental designs in which the way to build the design is approached in a completely different way with respect to classical design of experiments. The seminal contributions of Kiefer (1959) and Kiefer and Wolfowitz (1960) laid the basis of the optimal design theory. Differently from classical design of experiments, optimal designs are model-dependent, that is one or several statistical models have to be selected *a priori* in order to build the design. The optimality of the design is achieved with respect to one or

several design optimality criteria, strictly related to the assumed statistical model(s). This strong dependence between statistical model(s) and underlying experimental design is one of the main criticism of the optimal design theory. This is due to the fact that a design could be good for a given model, and worse for a different one, which may be finally more appropriate for the data. Nevertheless, optimal designs are widely developed and applied in numerous research fields. Through a wide range of design optimality criteria, they allow to optimize specific design objective functions that comply with the final researcher's aim. Moreover, in situations in which the researcher has a substantial knowledge *a priori* of the model he wants to fit, the model-dependent nature of the optimal designs is surely an advantage rather than a disadvantage.

In this chapter a literature review related to each of the three topics of the thesis is presented. The projects are inserted in the thesis according to the relevance of each contribution, and the review follows this order. More precisely, in Section 1.2 the literature related to optimal designs for choice experiments is reviewed. More specifically, the various approaches developed to build optimal choice designs are described in details by also considering the main developments in the Random Utility class of models, applied for the analysis of the consumers' preferences. Section 1.3 contains the literature review related to the design and analysis for computer experiments. The review mainly focuses on the class of LH designs developed in literature. In the same Section, the Kriging methodology is also briefly reviewed by considering only recent issues strictly related to its application in the technological field. Section 1.4 includes a literature review on design and analysis of split-plot experiments. The review concentrates on the main developments of split-plot design and analysis for industrial experimentations.

1.2 Optimal designs for choice experiments

In the literature, a large number of researchers and practitioners are dealing with preference measurements which are considered as one of the main general methods in order to study and improve the consumer's behaviour intended as the consumer's decision about improving his/her utility in changing a service or a product. Various preference measurements' methods are defined, first of all according to the nature and definition of prefences, namely revealed or stated preferences. Differently from the revealed preferences, that are obtained by observing individual's behaviour in a real-world situations, stated preferences are used to elicit individuals' preferences in experimental or survey context, where respondents are presented with hypothetical scenarios (Train, 2003, p. 156). Despite the wide range of stated preference

methods developed in literature (Conjoint Analysis, Contingent Valuation), undoubtedly the method of choice experiments is the most applied and preferred one.

Since the seminal paper of Louviere and Woodworth (1983), choice experiments have been widely used to study and analyse the consumers' preferences for a new product or service, with applications in various fields like marketing, transportation, environmental, health and political sciences. In a choice experiment, respondents receive different choice-sets for the evaluation and are asked to express their preferences within the choice-sets provided, each of which is composed of a set of alternatives where every alternative is defined as a combination of the different levels of the attributes of the product (or service). All the alternatives submitted and included in the choice-sets form the experimental design. Respondents are supposed to be utility-maximisers, e.g. within each choice-set they choose the alternative that maximises their utility, and their choices are analysed through Random Utility (RU) models (McFadden, 1974). Therefore, two fundamental issues should be considered when dealing with the methods of choice experiments: i) the underlying experimental design consisting of the choice-sets that have to be administered to each respondent, and ii) the RU models applied for the analysis of the respondents' choices. To this end, the related theory on choice experiments has been largely developed by considering both the design and the related class of RU models, for the latter by also considering its developments (Revelt and Train, 1998; McFadden and Train, 2000; Wen and Koppelman, 2001; Boxall and Adamowicz, 2002).

Moreover, when considering choice experiments it must be noted that these two elements, e.g. the design and RU models, are closely connected: on one hand, the properties of the design affect the corresponding model; on the other hand, an appropriate experimental design should be built according to the chosen RU model. To this purpose, since 1990s, a considerable number of researchers built choice-sets based on optimal design theory (Kiefer and Wolfowitz, 1960; Fedorov, 1972). The construction of optimal choice designs has been largely evolved by mainly considering the following issues on which the review focuses: i) the applied design optimality criteria and the approaches to deal with the unknown parameter values in the Fisher information matrix, by also considering the developments in the RU class of models (Revelt and Train, 1998; McFadden and Train, 2000; Wen and Koppelman, 2001); ii) the exact and approximate design framework applied to build optimal choice designs; iii) the construction and the use of homogeneous versus heterogeneous choice designs.

The most applied design optimality criterion for choice experiments is the D-optimality, aiming to an efficient estimation of the choice design parameters. Strictly related to it, the A-optimality criterion that minimizes the sum of the variances of the parameter estimates (Atkinson et al., 2007), has also been largely studied. An issue to consider for both design criteria is the fact that the applied RU models are non-linear in the parameters, implying that

the choice design efficiency depends on the unknown parameter values. This issue contributed to a significant number of improvements in the methods to build optimal choice designs. These methodological improvements were primarly developed under the most simple RU model, e.g. the Multinomial Logit (MNL) model (McFadden, 1974), and subsequently extended to more complex RU models.

A first approach to deal with the dependence of the design efficiency on the unknown parameter values, consists of assuming zero values for these parameters (Lazari and Anderson, 1994; Kuhfeld et al., 1994; Street et al., 2001; Burgess and Street, 2003, 2005). Consequently, under this unrealistic assumption, it is as if optimal designs reduce to estimate only linear models rather than RU models. This type of choice design is also called "utility-neutral" to stress its underlying assumption that all the alternatives and attributes levels are equally preferred by the respondents. Widely representative contributions related to this approach are in Lazari and Anderson (1994), that built statistically efficient cross-effect designs, and in Kuhfeld et al. (1994), that built choice designs based on A- and D-optimal fractional factorial designs. Moreover, Street et al. (2001) developed several theoretical methods to find A- and D-optimal choice designs for two-level attributes and choice-sets of size two; Burgess and Street (2003, 2005) further extended this method to D-optimal choice designs for any choice-set size and for attributes with any number of levels.

To overcome this unrealistic assumption of complete respondents' indifference, Huber and Zwerina (1996) proposed the use of non-zero parameter values to generate A- and D-optimal choice designs through the Relabelling-Swapping algorithm. They demonstrated that these choice designs have the so-called *utility balance* property that improves the efficiency of the underlying design by balancing the utilities of the alternatives in each choice-set. Carlsson and Martinsson (2003) confirmed these findings by comparing several techniques to build D-optimal choice designs with an application in the field of health economics.

The work of Huber and Zwerina had an important impact on the experimental design for choice experiments: in fact, they linked the choice design construction to the underlying RU model, and hence to the use of optimal design techniques for non-linear models. Since then, the construction of optimal choice designs has been further improved by also considering the Bayesian design framework (Chaloner and Verdinelli, 1995). Sándor and Wedel (2001) was the first to propose Bayesian design techniques to build D-optimal choice designs assuming a prior parameter distribution elicited from managers. To obtain the optimal choice designs, the authors build *ad hoc* algorithm, called Relabelling-Swapping-Cycling algorithm based on the Relabelling-Swapping algorithm of Huber and Zwerina (1996). The Bayesian approach employed in Sándor and Wedel (2001) was further extended by Kessels et al. (2006) to the construction of optimal choice designs by also considering the A-, G- and V-optimality

criteria, and through the modified Wynn-Fedorov algorithm (Wynn, 1970; Fedorov, 1972). Moreover, in Kessels et al. (2008) some practical guidelines are suggested on how to properly specify the prior distribution for the unknown parameters; Kessels et al. (2009) deal with the computational burden of Bayesian optimal choice designs, and to this end, they apply a modified version of the coordinate-exchange algorithm (Meyer and Nachtsheim, 1995). Furthermore, in the same effort Toubia and Hauser (2007) suggested a new optimum criterion for choice designs, namely the M-optimality (Toubia and Hauser, 2007), that generalizes the standard A- and D-criteria by considering the covariance matrix of managerial interests rather than that of the attributes.

The approaches previously described have been subsequently extended to optimal choice designs for more complex RU models. The main reason is that, although very simple, the MNL model has a number of limiting assumptions. First of all, for this RU model, the limiting property of the Independence of Irrelevant Alternatives is assumed. This means that the choice probability in one choice-set is independent from the presence of any other attribute values or alternatives. Furthermore, the MNL model does not take account of differences in the consumers' behaviour, i.e. each respondent, with different baseline characteristics is treated in the same way based only on their judgment. To this end, in literature the Mixed Logit model has been introduced to improve these issues (McFadden and Train, 2000). This RU model allows for relaxing the limiting assumption of the MNL model by considering the attributes as random variables, and not as fixed ones. In particular, depending on how the Mixed Logit model is specified, it makes it possible to account for: i) the preference heterogeneity across consumers: the so-called Cross-Sectional Mixed Logit (C-MIXL) model (McFadden and Train, 2000), ii) the correlation that arises between the responses given by the same respondent: the so-called Panel Mixed Logit (P-MIXL) model, by also allowing to evaluate the respondents' preference heterogeneity (Revelt and Train, 1998).

Moreover, it must be also noted that the complexity in the estimation of the Mixed Logit model with respect to the MNL one (Train, 2003) requires a high quality data, in which the amount of the true choice behaviour should be captured (Hensher and Green, 2003). To this end, the construction of optimal designs for this class of RU models is a fundamental issue. Undoubtedly, the complexity of the Mixed Logit model leads to more complex methods to build optimal choice designs, by considering the high computational time required to find optimal choice designs and the no-closed form expression for its corresponding Fisher Information matrix. Sándor and Wedel (2002) were the first to build optimal choice designs for the C-MIXL model, assuming nominal values for the unknown parameters based on managers' prior belief, and through the Relabelling-Swapping-Cycling algorithm (Sándor and Wedel, 2001). Subsequently, Yu et al. (2009) built optimal choice designs for the

C-MIXL model in the Bayesian design framework. To this end, the authors specified a prior distribution for the mean vector and nominal values for the heterogeneity vector, applying the coordinate-exchange algorithm (Meyer and Nachtsheim, 1995). Moreover, Yu et al. (2009) also addressed the issue on the computationally intensive numerical integration over the prior parameter distributions. To this end, they applied Monte Carlo based approximation with Halton sequencies by obtaining a faster convergence and smaller simulation errors (Train, 2000; Bhat, 2001).

Optimal choice designs under the P-MIXL models have received less attention in the research literature. This is mainly due to the higher complexity to approximate the P-MIXL information matrix with respect to that of the C-MIXL model. This higher complexity consists in dealing with products of logit probabilities (P-MIXL model) with respect to summations (C-MIXL model). To this end, and in order to obtain A- and D-optimal choice designs, Bliemer and Rose (2010) carried out two types of simulations to approximate the P-MIXL information matrix. Yu et al. (2011) proposed an individually-adapted sequential Bayesian approach to build optimal choice designs for the P-MIXL model simulating the respondents' choices. Sandor (2013) and Zhang et al. (2013) have recently derived a more detailed expression for the P-MIXL Information matrix and addressed the issue on how it is efficiently approximated.

Simultaneously with the developments for building optimal choice designs, a greater emphasis was also placed on the complexity of choice experiments and choice-sets with respect to consumers, e.g. the number of choice-sets assigned to a respondent plays a fundamental role in preventing the respondents' fatigue, and consequently for obtaining a realistic evaluation of their preferences. This issue is particularly important when the number of choice-sets in the design is large. This is often the case of optimal choice designs for the Mixed logit model that involves a greater number of parameters to be estimated with respect to the MNL model. To this end, various researchers faced this issue directly in the design step through the construction of heterogeneous choice designs. Differently from homogeneous choice designs, in which all respondents get the same design, in a heterogeneous one, different groups of respondents receive different subdesigns extracted from a larger one. Surely, the computer search to obtain optimal heterogeneous choice designs is computationally much more intensive with respect to those for homogeneous choice designs. Sándor and Wedel (2005) were the first to propose the construction of heterogeneous choice designs. By considering Bayesian and local design framework, they built optimal heterogeneous choice designs for the MNL and C-MIXL models, by demonstrating a substantial improvement in the efficiency of the coefficients estimated through a heterogeneous choice design and

modelling with respect to the homogeneous one, event though with a small number of subdesigns.

Heterogeneous choice designs are very attractive by allowing to obtain a large amount of information about the respondents' preferences with respect to homogeneous ones. Nevertheless, the main difficulty is the extremely high computational time required to obtain them, by especially considering the fact that all the approaches previously described make use of an exact design framework for obtaining optimal choice designs. To this end, Liu and Tang (2015) recently proposed a new approach for the construction of heterogeneous choice designs for the C-MIXL model based on an approximate design theory (Kiefer and Wolfowitz, 1960). This approach allows for avoiding the high computational time in the computer search for heterogeneous designs, given the well-known mathematical tools for checking and guaranteeing the design optimality. More precisely, when considering the exact design framework, there are no mathematical tools for ensuring that the final design is globally optimal, unless a global check is made for each possible design belonging to the entire design space. Although the problem relates to both homogeneous and heterogeneous choice designs, it is particularly relevant when considering heterogeneous choice design with a prohibitevely huge design space. Conversely, in an approximate design framework, the General Equivalence Theorem provides the necessary and sufficient condition for checking optimality (Kiefer and Wolfowitz, 1960). Moreover, Liu and Tang (2015) proposed a modified version of the optimal weight exchange algorithm of Yang et al. (2015) to obtain optimal choice designs. This algorithm, making use of the Newton-Raphson optimisation method, substantially outperforms existing algorithms in terms of computation time, making it particularly suitable in the choice experiment context (Tian and Yang, 2017).

Finally, last two remarks relate to optimal choice designs developed when certain conditions and/or assumptions hold for the choice experiment. A first situation is when there is a large number of attributes: in this case, respondents express their choices only on a subset of these attributes. That is, it is implicitly assumed that when expressing their choices, the respondents concentrate only on a subset of the attributes, by ignoring the remaining ones. This type of experiments are called choice experiments with partial profiles. In this regard, the literature related to optimal choice designs with partial profiles has been largely developed (Grossman et al., 2006, 2014; Kessels et al., 2011, 2014; Cuervo et al., 2016). A second case regards a recent development in optimal choice designs related to the use of multi-objective design criterion (e.g. a compound design criterion) instead of a single-objective one. More precisely, a compound design criterion has been applied by Henderson and Liu (2016). The authors assume a selective choice process consisting of active and inactive attributes, and apply a compound design criterion in order to incorporate prior information for the joint

purpose of an efficient estimation of the coefficients for the active attributes, and detection of the effects for the inactive ones.

Moving on the considerations above, in Chapter 2 we propose an innovative approach for constructing heterogeneous choice designs with correlated choice preferences based on an approximate design theory and a compound design criterion. Moreover, we apply our proposal to a real case study, carried out by ourselves, in which we integrate a choice experiment with extra preference information.

1.3 Latin Hypercube designs and Kriging modelling for computer experiments

Nowadays, a physical experimentation for some complex scientific and technological processes is often time-consuming, costly or even impossible to be performed. Thus, computer experiments are performed instead of physical ones in order to investigate the deterministic relation between input and output variables. A fundamental issue for computer experiments is the programming of the underlying experimental design that substantially differs from the classical design of experiments. The deterministic nature of the computer experiments makes the three basic design principles (e.g. randomization, replication and local control of error) irrilevant in this context. For instance, we do not need replicates given that performing repeated runs of the computer code with the same input variables produces exactly the same output. Moreover, the lack of knowledge of the true relation between input and output variables requires to build an experimental design through which all the portions of the experimental region could be efficiently represented and explored. To this end, space-filling designs are increasingly used in this context. One of the most important and widely used class of space-filling designs for computer experiments is the class of LH designs.

LH designs have been introduced in the field of computer experiments by McKay et al. (1979) which compared random sampling, stratified sampling and LH sampling for selecting the input variables in terms of their associated mean and variance estimators. The authors found that, although all three methods provides empirically unbiased estimates of the response mean, the results obtained from the LH sampling have the smallest variance. Subsequently, Iman and Conover (1982) proposed a distribution-free method for LH sampling by "inducing" correlation for the input variables and for small sample size. In the same effort, and by further investigating the properties of LH designs, Stein (1987) confirmed the results of McKay et al.(1979) and developed a method to obtain a LH design that properly accounts for the dependence among the input variables when the sample size is large. Since then, LH

designs have been largely studied for the design of computer experiments. The main reason that makes them suitable for computer experiments is their one-dimensional space-filling property: namely, a LH design achieves the maximum uniformity when projected in any one dimension.

Nevertheless, in most of the real applications, the design involves a large number of input variables, and therefore it is desirable to achieve the uniformity in more than one dimension. Given that there is no guarantee that a random LH design attains the uniformity when projected in more than one dimension, various approaches have been developed to build LH designs with lower- or multi-dimensional space-filling properties. Although the distinction among the several types of LH designs is not always completely clear, we can generally distinguish between: i) LH designs based on some measure of distance (maximin LH designs), or, on one or more design optimality criteria (optimal LH designs), and ii) orthogonal LH designs and LH designs based on orthogonal and strong orthogonal arrays.

For optimal LH designs, we give only a non-exhaustive list of the most relevant contributions, given that they go beyond the scope of this thesis. For an excellent review on them, we refer to Pronzato and Müller (2012). Historically, two design optimality criteria have been primarily used to build optimal LH designs, namely the entropy criterion and the integrated mean squared error (IMSE) criterion. Shewry and Wynn (1987) built a LH design maximizing the entropy optimality criterion, so as maximizing the amount of the information of the experiment. Subsequently, Sacks et al. (1989) proposed the IMSE optimality criterion in order to minimize the variance of the prediction. Currin et al. (1991) constructed optimal LH designs in a Bayesian design framework through the entropy criterion of Shewry and Wynn (1987), and by also assuming equal correlation parameters. Park (1994) built optimal LH designs minimizing the IMSE criterion through a two-stage algorithm, exchange and Newton type. Pistone and Vicario (2010) developed the algebraic construction of LH designs minimizing the integrated Kriging variance for a specific correlation structure. Recently, Jourdan and Franco (2015) built optimal LH designs based on the Kullback-Leibler information criterion. Moreover, Dette and Pepelyshev (2010) proposed a new type of generalized LH designs by considering a transformation of the design points through the quantile function of the Beta distribution. It is also relevant to note that more recently, Jones et al. (2015) introduced the so-called bridge designs, that could be considered as a new type of LH designs. This is because bridge designs inglobates the property of LH designs, that guarantees a minimum distance among the design points, and of optimal designs, through the use of the D-optimality design criterion.

When considering maximin LH designs, it must be noted that they could be viewed as a special case of optimal LH designs, given that they use a maximin distance criterion through

an algorithmic optimization step. Morris and Mitchell (1995) were the first to propose the maximin LH design based on the maximin distance criterion introduced by Johnson et al. (1990). Through a search via the simulated annealing algorithm over various competing designs, the authors obtained the best LH design maximizing the minimum distance between the design points. They also demonstrated that their maximin LH design outperforms a randomly chosen LH design in terms of the mean squared error (MSE) and the maximum prediction error of the corresponding Gaussian process models. Subsequently, the approach of Morris and Mitchell (1995) has been further developed and improved by considering both the algorithms and the applied criteria. To this end, Ye et al. (2000) built an optimal symmetric LH design through the columnwise-pairwise algorithm, so as obtaining a design with better geometric properties; Van Dam et al. (2007) built a two-dimensional maximin LH design by considering the Euclidean distance in the maximin criterion; Moon et al. (2011) defined a two-dimensional maximin distance criterion and developed the smart swap algorithm for finding maximin LH design, while Chen et al. (2013) applied the particle swarm algorithm to this purpose. More recently, Yang et al. (2015) extended the approach to sliced maximin LH designs. One of the major advantage of the maximin LH designs is that they achieve very good one- and full-dimensional space-filling properties; however, when considering low-dimensional projections, the uniformity attained by the designs could be poor (Joseph et al., 2015). To deal with this issue, more recently Joseph et al. (2015) proposed the maximum projection LH design that ensurs very good space-filling properties in lower-dimensional projections.

Differently from optimal and maximin LH designs, orthogonal LH designs and LH designs based on orthogonal and strong orthogonal arrays are entirely model-independent. This issue is particularly important if the underlying statistical model is *a priori* unknown. Moreover, the design construction techniques for these types of LH designs are relatively simple, without the need to apply any complex optimization algorithm. Furthermore, orthogonal LH designs and LH designs based on orthogonal and strong orthogonal arrays achieve very good space-filling properties when projected in more that one dimension.

Orthogonal LH (OLH) and nearly OLH designs are based on the property of orthogonality, a milestone in the classical design of experiments, ensuring that the columns of a design matrix are uncorrelated. The main rationale for the use of OLH and nearly OLH designs is that space-filling designs should be orthogonal or nearly orthogonal, and therefore, it is reasonable to search for space-filling designs within this type of LH designs (Bingham et al., 2009; Lin and Tang, 2015). The construction of OLH was firstly proposed by Ye (1998), who developed this method for small run size through the use of permutation matrices. Subsequently, Cioppa and Lucas (2007) extended the Ye's approach by adding a new orthogonal column in order to

slightly increase the run size of the design. The method has been further improved through the use of rotation on factorial designs (Steinberg and Lin, 2006; Pang et al., 2009) and generalized orthogonal designs (Georgiou, 2009), but still applied to a limited sample size. Bingham et al. (2009) and Lin et al. (2009) extended the construction of OLH and nearly OLH design to more flexible runs size. More recently, the method has been extended to nested OLH and nearly OLH designs (Li and Qian, 2013; Yang et al., 2014; Yang et al., 2016), that are applied for conducting computer experiments that involve multiple level of cost and/or accuracy, and sliced OLH designs (Yang, 2013), useful for computer expriments with qualitative and quantitative factors, multiple computer experiments and data-pooling.

Although the construction of OLH designs is relatively easy, for a large number of input variables, the run size of the design becomes extremely large. To this end, LH designs based on orthogonal arrays (OA-based-LH) seems to be preferable by allowing more flexibility when considering the design run size. The use of orthogonal arrays to build LH designs with good low-dimensional space-filling properties has been proposed in the 1990's, contemporaneously but independently, by both Owen (1992) and Tang (1993). With the main aim to obtain low-dimensional projection properties, Owen (1992) developed randomized orthogonal arrays, while Tang (1993) developed OA-based-LH designs, also called U designs. The OA-based-LH design of Tang (1993) has been extended to nested OA-based-LH design by He and Qian (2011), useful to perform model adjustment between two sources of nested relation of high and low accuracy computer code. Moreover, Yin et al. (2014) developed sliced OA-based-LH designs by means of sliced random permutations that can be efficiently applied for computer experiments involving qualitative and quantitative variables, cross-validation and uncertainty quantification of computer models.

Recently, He and Tang (2013) developed a new class of orthogonal arrays called strong orthogonal arrays that results particularly useful for computer experiments. The authors built LH designs based on this new class of orthogonal arrays by demonstrating their excellent space-filling properties, also compared to OA-based-LH designs. The relatively easy way to generate this type of LH designs, without the use of any type of optimization algorithm, together with their excellent space-filling properties, contributed to further develop its related methodology. More precisely, He and Tang (2014) further developed the theory on strong orthogonal arrays of *strength* three describing a complete characterization of such a type of arrays. Moreover, sliced space-filling designs based on sliced strong orthogonal arrays have been developed by Liu and Liu (2015). More recently, in He et al. (2018) strong orthogonal arrays of strength two plus are described by also illustrating their connection with second-order saturated designs.

When considering the analysis for computer experiments to solve complex technological issues, undoubtedly the most applied metamodel is Kriging. The starting point for the Kriging methodology dates back to the geosciences in 1950's by the South African engineer Danie G. Krige (Krige, 1953) for the analysis of geostatistical data. In the 1960's the French mathematician and geologist Georges Matheron exploited the basic theory for geo-statistics (Matheron, 1962). G. Matheron was the first to use the term "Krigeage" in the honour of the work developed by Krige, subsequently translated in the English language as "Kriging" (Matheron, 1967), and which is now used worldwide. Since the seminal paper of Sacks et al. (1989), Kriging models have been widely used for the analysis of computer experiments.

When considering the Kriging methodology, it must be noted that in the last one decade, it has been further developed and applied in order to solve complex technological and engineering issues. In Roustant et al. (2010), the Kriging methodology is applied in order to analyze the output parameters in the field of the nuclear safety. More precisely, the authors proposed the Kriging as an alternative to the RSM approach to model the peak cladding temperature of a fuel. They demonstrated that Kriging provides more accurate prediction with respect to a polynomial response surface models, given its higher flexibility to handle the degree of smoothness. Through an application of the Kriging to study the production process of silicon wafers, Pistone and Vicario (2013) also discuss how to model the covariance structure when dealing with spatial data showing a strong correlation. A Kriging methodology with functional response is developed by Hung et al. (2015) in order to optimize the residual stresses in the matching of metals. Borrotti et al. (2016) compare polynomial regressions, Kriging and artificial neural network models for the optimization of the operational parameters of a Corona electrostatic separation process; they found that the most suitable metamodel for the optimization of the process under study is the Kriging, allowing to better describe the properties of the process. In Vicario et al. (2016), Kriging modelling is compared with artificial newral network metamodels in order to determine the most accurate predictive model in fluid dynamics experiments for low pressure turbines, while Vicario et al. (2018) apply the Kriging for the manufacturing process of a metal sheet bending, by confirming its high prediction accuracy. Recently, in Arcidiacono et al. (2016), Kriging models are successfully applied to also study the payload distribution of freight trains.

Fundamental issues to consider when dealing with the Kriging methodology are undoubtedly the choice of the covariance function, the estimation method and the inclusion of the nugget parameter. To this end, novel issues related to the the definition of the covariance structure for the stochastic part of the model are suggested in Del Castillo et al. (2015). Moreover, Sang et al. (2012) developed a new approach of full-scale approximations of

the covariance structure suitable for large data sets, while Durrande et al. (2012) deal with covariance functions for high dimensional additive Kriging models.

When considering the recent developments related to Kriging estimation, it is relevant to mention the contribution of Ginsbourger et al. (2009) in which recommendations on the choice and the estimation for Kriging models when the run size of the design is small are suggested. Moreover, Li and Sudjianto (2005) proposed a penalized likelihood approach for the estimation of the Kriging model to overcome the problem of a flat likelihood function near the optimum. In order to improve the Kriging predictions, a Bayesian approach is suggested in Deng et al. (2012) through the best definition of a model, and by applying a two-level Bayesian hierarchical prior distribution; in Hung (2011), a variable selection is suggested to find the best deterministic model. In Qian et al. (2008) and Zhou et al. (2011), the authors deal with Kriging modelling by also involving qualitative variables, and in Han et al. (2009) a Bayesian methodology for the prediction of computer experiments is developed for this purpose. Lastly, nugget issues are studied by Gramacy and Lee (2012) where the authors demonstrated that the estimation of the nugget allows for obtaining better statistical properties of the Kriging model, and by Peng and Wu (2014), where the choice of the nugget is analyzed to improve Kriging predictions and prevent numerical instabilities.

The second project of the thesis in Chapter 3 deals with computer experiments and related Kriging issues. More precisely, we aim to improve the payload distribution of freight trains through a suitable design for the computer experiment and Kriging modelling.

1.4 Split-plot design and analysis

Split-plot designs originate from the field of agricultural experimentation (Fisher, 1925), according to which large areas of land, known as Whole-Plots (WP) are subdivided into smaller areas, known as Sub-Plots (SP). Starting from this initial framework, nowadays in a split-plot design there are two types of factors, namely WP factors and SP factors, that generate a design structure involving restrictions on randomization. More precisely, we may speak of bi-randomization which is characterized by the distinction between the two set of factors, which, in turn, generates two types of experimental units with two separate error terms. Firstly, the WP factors are randomly assigned to WP experimental units, generating the WP error term. Then, the SP factors are randomly assigned within each WP, generating the SP error term. Therefore, the SP factors are nested and randomized within the WP units, and as a consequence, the experimental units for the SP factors become observational units for the WP factors. This means that we have more experimental units, and therefore more "information" for the SP factors with respect to the WP factors. Cox (1958) referred to the

WP factors as "classification factors", which are assumed not to be of major interest; the author highlighted that they are primarely included in the design to examine their effect in interaction with other factors considered more important, e.g. the SP factors. Therefore, through the use of a split-plot structure, less accurate WP estimates are accepted in order that more accurate SP estimates are obtained.

Often, many industrial experiments involve situations in which a complete randomization of all the factors in the design is difficult or also impossible to be performed. More precisely, some factors are considered as hard-to-change, because their levels are costly or time-consuming to be changed. Typical example of such a type of factors is the humidity. Instead, others are considered as easy-to-change factors because their levels could be easily changed during the experimentation. An excellent way to accomodate such a type of industrial experiment is through the use of the split-plot design, in which the hard-to-change factors are settled as WP factors, while the easy-to-change ones as SP factors. By this way, the use of the split-plot structure allows to reduce the manipulation of the hard-to-change factors. However, as a consequence, we have more degrees of freedom for the SP error term with respect to the WP one, and therefore higher power to detect significant the SP effects with respect to the WP ones.

It is also relevant to note that there are also many variations of the split-plot designs such as split-split plot, strip-plot, strip-block and split-block designs (Miller, 1997; Federer and King, 2007; Wu and Hamada, 2009; Montgomery, 2013), in which more complex experimental situations could be excellently accomodated. Moreover, the huge potential of the split-plot design allows to address further randomization issues, strictly related to certain technical requirements.

Nowadays, the split-plot design is considered as a basic design for industrial experimentation and for a robust design approach. By especially considering the latter one, it is relevant to note that it is since the seminal contribution of Box and Jones (1992) that the particular structure of the split-plot design has been exploited in the Robust Product Design (RPD) framework, in a context of fundamental developments related to the product quality improvement through the robust design approach (Vining and Myers, 1990; Myers et al., 1992). To this end, in their contribution Box and Jones (1992) analyzed three different arrangements of the design and environmental factors in order to assess the most efficient split-plot structure. They found that a design arrangement according to which the SP factors are assigned in strips across the WP factors (e.g. a strip-block design) allows to obtain more accurate estimates of the interaction between the design and environmental factors that is fundamental to study robustness (Box and Jones, 2001). Moreover, Box and Jones (1992)

also demonstrated the relative efficiency of the split-plot design in obtaining more accurate factor estimates with respect to completely randomized and random block designs.

Subsequently to this revisitation of the split-plot design by Box and Jones (1992), a fundamental issue for its further development was the definition of the split-plot design as a bi-randomized design (Letsinger et al., 1996), that inserted it effectively within the class of Response Surface designs. The introduction of the split-plot design in the RSM context contributed significantly to also address some concerns related to its proper inference and analysis. Certainly, this issue is not new. In fact, several authors already warned out that often in the past split-plot designs was erroneously analyzed as a completely randomized design (Wooding, 1973; Box, 1996; Simpson et al., 2004). The main consequence to not account for the two error terms when dealing with a split-plot design is an inappropriate inference about the significance level of both the WP and SP factors. Letsinger et al. (1996) introduced the two types of bi-randomized designs: namely, non-crossed and crossed designs, the latter to which takes part the split-plot design. Moreover, Letsinger et al. (1996) was the first to also describe which are the consequences when a split-plot design is incorrectly analyzed as a completely randomized design. As the authors demonstrated, in that case, certain WP effects could be erroneously considered as significant when in fact they are not; at the same time, certain SP or SP by WP intercations could be considered as no-significant when in fact they are significant. The authors went further by also comparing different estimation methods for the two types of bi-randomized designs, namely Ordinary Least Squares (OLS), Generalized Least Squares (GLS), Iterative Reweighted Least Qquares (IRLS) and Restricted Maximum Likelihood (REML) estimation methods. They concluded that the REML estimation method is preferred towards the IRLS one, given its better asymptotical properties. Both Draper and John (1998) and Trinca and Gilmour (2001) confirmed these findings.

Subsequently, Vining et al. (2005) established the conditions under which the OLS-GLS equivalence estimation is achieved, by considering two standard Response Surface designs, namely the Central Composite Design (CCD) (Box and Wilson, 1951) and the Box-Behnken design (Box and Behnken, 1960), in a split-plot structure. The OLS-GLS equivalent estimation was further extended to unbalanced split-plot designs (Parker et al., 2007). Moreover, Vining and Kowalski (2008) provided the conditions for exact tests for the model coefficients based on residual estimates of the variance components; the authors provided the appropriate error terms for testing WP and SP effects, by also demonstrating how to calculate the WP and SP residual values to check model assumptions. Lastly, Wang et al. (2009) extended the OLS-GLS equivalence estimation to orthogonally blocked split-plot CCDs.

The inclusion of the split-plot design in the RSM context further contributed to its theoretical developments by also considering optimal design theory. Generally, the criterion of D-optimality has been the dominant one (Goos and Vandebroek, 2001a, 2001b, 2003; Jones and Goos, 2007). More recently, other design optimality criteria have been applied, by also considering the Bayesian design framework (Mylona et al., 2014), as well as more complex design algorithms (Jones and Goos, 2012; Sambo et al., 2015). Moreover, the OLS-GLS equivalent estimation have been also largely explored by considering optimal split-plot designs (Goos, 2006; Macharia and Goos, 2010; Mylona et al., 2013).

By considering other relevant issues related to the split-plot design, it must be noted that further developments have been also achieved by considering the analysis of split-plot designs with non-normal response(s). To this end, Robinson et al. (2004) illustrated in details the use of Generalized Linear Mixed Models (GLMMs) (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) to this purpose. More recently, Goos and Gilmour (2012) outlined a general strategy for analyzing data from split-plot and multistratum designs through GLMMs, by concentrating specifically on the determination of the unit structure and the definition of the random effects to be used in the analysis. Arnouts and Goos (2017) extended these results to split-plot designs with ordinal responses by applying the cumulative logit model in order to study the adhesion between steel tire cords and rubber. More recent developments to analyze data from split-plot experiments with non-normal responses make also use of the Bayesian methodology (Robinson et al., 2012, Tan and Wu, 2013).

Moreover, it must be also mentioned the possibility to define the split-plot design as a particular fractional factorial design. To this end, a first research path maked used of the aberration criterion (Fries and Hunter, 1980) to obtain fractional factorial split-plot designs (Huang et al., 1998; Bingham and Sitter, 1999a; Bingham and Sitter, 1999b; Bingham and Sitter, 2001). A second research path is primarely represented by the contribution of Bisgaard (2000), who defined the split-plot design as a product of two fractional factorial designs through the use of the partial Resolution criterion. Subsequently, the Bisgaard's approach was further developed by several other authors, especially for early stage factor screening (Kulahci and Bisgaard, 2005; Tyssedal et al., 2011; Kulahci and Tyssedal, 2017).

To conclude, it must be noted that nowadays there are many situations of industrial experiments that involve further restrictions on randomization, primarely due to production process requirements. In cases like this, the split-plot design still prove to be a valid experimental plan, as long as a proper detection and analysis is carried out in order to better understand the implications of this further lacking of randomization. The third project of the thesis aims to address this issue. More precisely, it is based on a real case-study in which further restrictions on randomization in a split-plot design are imposed following stringent

requirements related to the production process. To this end, we examine how to detect lacking of randomization in a split-plot design and which are the implications for the WP and SP estimates.

Chapter 2

Optimal Approximate Choice Designs with Correlated Preferences Through a Compound Criterion

2.1 Introduction to the project

The project in this Chapter deals with an innovative approach for constructing heterogeneous choice designs with correlated choice preferences based on an approximate design theory and a compound design criterion. To this end, in this Section we briefly introduce the real case-study on which our proposal is applied. For a detailed description of the case-study, we refer to Section 2.6.

The real case-study concerns the analysis of the consumers' preferences for coffee consumption by integrating choice experiments with extra preference information based on a chemical analysis and consumer sensory tests. More precisely, the study consists of three main steps, each related to a specific evaluation of the coffee being studied. The first relates to the chemical analysis of the caffeine contained in the coffee, and performed by a High Performance Liquid Chromatography (HPLC) evaluation method. The second concerns the scores of the sensory assessments obtained through a guided tasting session planned in order to analyse the role of taste in the consumers' preferences. The last part relates to the choice experiment that has to be physically administered to the respondents. Therefore, the design matrix also evaluates: i) the HPLC measurement results related to the quantities of caffeine in the coffee, and ii) the scores of the sensory assessments of the coffee obtained from the guided tasting session. We administer the same choice experiment twice: before (Choice 1) and after (Choice 2) the guided tasting session. To this end, we develop our proposal

under a compound design criterion (Wynn, 1970; Atkinson and Bogacka, 1997; Atkinson et al., 2007) in order to address the following two main issues: i) an efficient estimation of the attributes of the choice experiment, and ii) detection of the effects related to the HPLC results (in Choice 1) and the scores obtained through the consumer sensory tests (in Choice 2). The same compound design criterion was applied in the choice experiment literature by Henderson and Liu (2016) where the authors, assuming a selective choice process consisting of active and inactive attributes, apply this criterion in order to incorporate prior information for the joint purpose of an efficient estimation of the coefficients for the active attributes and detection of the effects of the inactive attributes.

Following, in the next Section we briefly describe the approximate design framework for choice experiments and the applied compound design criterion, as well as the algorithm used to compute the optimal choice designs.

2.2 Optimal designs for choice experiments

2.2.1 Approximate design framework in the choice experiment context

An approximate design is represented by a probability measure ξ over a compact design space χ. In our context, when considering choice experiments, an approximate design ξ can be expressed as (Kiefer and Wolfowitz, 1960; Liu and Tang, 2015; Tian and Yang, 2017):

$$\xi = \begin{Bmatrix} C_1 & C_2 & \dots & C_q & \dots & C_Q \\ w_1 & w_2 & \dots & w_q & \dots & w_Q \end{Bmatrix} \tag{2.1}$$

where C_q is a choice-set that belongs to the space of all possible choice-sets Q, and w_q, $q = 1, ..., Q$, is its corresponding weight with $0 \leq w_q \leq 1$ and $\sum_{q=1}^{Q} w_q = 1$.

Let y_q be the response to the choice-set C_q and assume the response is a random variable with probability density function (pdf) $f(y_q; C_q, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the vector of coefficients to be estimated. Then the Fisher Information matrix (FIM) at C_q is $I(C_q) = \frac{\partial^2 \log f(y_q; C_q, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}$. Therefore, for an approximate design, the FIM is defined as follows:

$$I(\xi) = \sum_{q=1}^{Q} w_q I(C_q) \tag{2.2}$$

where $I(C_q)$ is the FIM for a single choice-set C_q.

In the choice experiment context, each w_q in formula (2.1) indicates to which proportion of respondents the choice-set C_q should be supplied. It must be also noted that the weights in

formula (2.1) are continous. Consequently, the choice design is defined in an approximate design framework that is different from an exact one, usually employed to find optimal choice designs. More precisely, when considering an exact design framework, the search for the design points on the experimental region is discrete in nature, and therefore there are no mathematical tools for ensuring that the final design is globally optimal unless a global check is made of each possible design belonging to the entire design space. Conversely, when considering an approximate design framework, the General Equivalence Theorem (GET) (Kiefer and Wolfowitz, 1960) provides the necessary and sufficient condition for checking optimality. The advantage of building choice designs based on the approximate design theory is that their optimality can be checked via the GET searching only over the space of all possible choice-sets Q, which is much smaller than the space of all possible choice designs. As an example, consider the case in which we have a choice experiment with five attributes, each at two levels, and with two alternatives in each choice-set. This means $2^5 = 32$ possible combinations of the attributes and $\binom{32}{2} = 496$ possible choice-sets. If we want to compute an optimal exact choice design with eight choice-sets, we have to search over all possible choice designs of size 8, that is equal to:

$$\binom{496}{8} = 4326356 \times 10^{26}$$

Conversely, if we want to construct an optimal approximate choice design, we have to compute the information matrix just for each of the 496 possible choice-sets.

In the context of choice designs, the GET could be briefly expressed as follows.

General Equivalence Theorem (Whittle, 1973): Let Φ be a concave criterion function with the usual conditions. The directional derivative of Φ at ξ in the direction of ξ' is usually defined in optimal experimental design theory as:

$$\partial\Phi(\xi,\xi') = \lim_{\lambda\to 0^+} \frac{\Phi((1-\lambda)\xi + \lambda\xi') - \Phi(\xi)}{\lambda}$$

A design ξ is Φ-optimal if and only if:

$$d(C_q,\xi) \leq 0 \tag{2.3}$$

for any choice-set C_q that belongs to the space of all possible choice-sets Q, where $d(C_q,\xi)$ are the directional derivatives of Φ at the information matrix $I(C_q)$ in the direction of C_q.

2.2.2 A compound design criterion

In this Subsection, we briefly describe the compound design criterion under which we build the optimal heterogeneous choice designs. For the moment, assume that the vector β is of dimension $K \times 1$. In order to address the main objective of our real case study, described in detail in Section 2.6, we apply the following compound D-optimality criterion, $\Phi_C(\xi)$, (Wynn, 1970; Atkinson and Bogacka, 1997; Atkinson et al., 2007):

$$\Phi_C(\xi) = \frac{\alpha}{K_1}(\log|I_{11}(\xi)|) + \frac{(1-\alpha)}{K_2}(\log|I(\xi)| - \log|I_{11}(\xi)|) \qquad \alpha \in [0;1] \qquad (2.4)$$

where $K_1 + K_2 = K$.

Through the criterion expressed in formula (2.4), we balance two objectives, namely: i) efficient estimation of the effects of the attributes of the choice experiment ($D-$optimality), and ii) detection of the effect of the HPLC results and the scores of the sensory assessment (D_s-optimality). More precisely, in formula (2.4), $I_{11}(\xi)$ is the FIM containing the K_1 coefficients related to the attributes of the choice experiment, while $I(\xi)$ is the FIM that contains both the K_1 coefficients of the choice experiment and the K_2 coefficients related to i) the HPLC results in Choice 1, and ii) the scores obtained through the guided tasting in Choice 2.

The coefficient α ($0 \leq \alpha \leq 1$) reflects the relative interest in both objectives in the criterion in formula (2.4). When $\alpha = 1$ we obtain a D-optimal choice design for the K_1 model coefficients, while for $\alpha = 0$ we obtain a D_s-optimal design for the K_2 model coefficients. In order to determine the best value of α, the efficiencies for a series of values of α can be calculated, and a plot of them against α offers a practical tool for choosing a design with an optimal balance for the efficiencies for both aspects (Atkinson and Bogacka, 1997).

When considering the D-compound criterion in formula (2.4), the derivative function, $d(C_q, \xi)$, in formula (2.3) is expressed as follows (Atkinson and Bogacka, 1997; Atkinson et al., 2007):

$$\begin{aligned} d(\xi, C_q) = {} & \frac{\alpha}{K_1}\left(\mathrm{Tr}\{[I_{11}(C_q) - I_{11}(\xi)]I_{11}(\xi)^{-1}\}\right) + \\ & + \frac{(1-\alpha)}{K_2}\left(\mathrm{Tr}\{[I(C_q) - I(\xi)]I(\xi)^{-1}\} - \mathrm{Tr}\{[I_{11}(C_q) - I_{11}(\xi)]I_{11}(\xi)^{-1}\}\right) \end{aligned} \qquad (2.5)$$

For the design ξ^* that maximizes the criterion reported in formula (2.4) the maximum value of (2.5) is equal to one, providing the GET conditions for the compound D-optimum design.

2.2.3 Algorithms for the construction of optimal choice designs

Various algorithms have been used in the choice experiment research in order to obtain optimal choice designs. Some of them have been developed especially for the costruction of optimal choice designs, such as the Relabeling and Swapping algorithm of Huber and Zwerina (1996) and its modified version Relabeling-Swapping-Cycling of Sandor and Wedel (2002, 2005). Other algorithms, such as the modified Wynn-Fedorov algorithm (Wynn, 1970; Fedorov, 1972) and the coordinate-exchange algorithm (Meyer and Nachtsheim, 1995) have also been adapted for this purpose (Kessels et al, 2006; Kessels et al., 2009, Yu et al., 2009).

Recently, the optimal weight exchange (OWE) algorithm has been proposed by Yang, Biedermann and Tang (2013) for deriving optimal designs also for non linear models. This algorithm, making use of the Newton-Raphson optimisation method, substantially outperforms existing algorithms in terms of computation time. It has also been extended to the computation of optimal choice designs by Liu and Tang (2015) where the authors modified the finer grid approach of the original OWE algorithm slightly by iteratively updating the choice-sets with their corresponding weights until they converged into a globally optimal approximate design, and where the necessary and sufficient condition of the GET is satisfied. The four main steps of the modified-OWE (mOWE) algorithm could be briefly summarized as follows (Liu and Tang, 2015):

1) initial design ξ_0: to obtain the initial design ξ_0, randomly choose K choice-sets from all possible choice-sets, Q, and assign an equal weight for each selected choice-set;

2) updated design ξ_t: update the equal weights of the selected choice-sets in the initial design ξ_0, with the optimal weights obtained through the Newton iteration method;

3) for the updated design ξ_t, and for each possible choice-set C_q, $q = 1, ..., Q$, calculate $d(\xi_t, C_q)$, formula (2.5), in order to find the choice-set C_q^* which maximizes $d(\xi_t, C_q^*)$. If the necessary and sufficient condition of the GET is satisfied (formula 2.5), then ξ_t is the optimal approximate design;

4) otherwise, obtain a new initial design, by adding to the design ξ_t, the choice-set C_q^* with an weight assigned to zero; repeat steps 2) and 3).

For further details see Liu and Tang (2015). This modified version of the algorithm has been used to compute single-objective optimal choice designs. We have used this mOWE algorithm to compute the optimal heterogeneous choice designs, by adapting it to the compound design criterion in formula (2.4).

2.3 Random Utility framework for choice experiments

The model-dependent nature of the optimal design theory requires the selection a priori of the statistical model(s) for building the design. To this end, in this Section we describe the statistical models applied to analyze the consumers' preferences collected through choice experiments; these models belong to the Random Utility (RU) class of models (Thurstone, 1927; McFadden, 1974). The fundamental assumption in the RU theory framework for consumers' preferences is that for each choice-set the respondent chooses the alternative that maximizes his/her utility. Assume that a respondent n ($n = 1,...,N$) receives s_n choice-sets belonging to the set S ($S \in Q$); for the sake of simplicity, in the rest of the theory we leave out the suffix n for s_n without any loss of generality. Each choice-set C_s ($s = 1,...,S$) contains J alternatives, where each alternative j ($j = 1,...,J$) is defined as a combination of different attribute levels. Therefore, according to the RU theory, a respondent n has a random utility function, U_{nsj}, expressed as:

$$U_{nsj} = \boldsymbol{x}_{nsj}^{'}\boldsymbol{\beta} + \varepsilon_{nsj} \tag{2.6}$$

where $\boldsymbol{x}_{nsj}^{'}$ is the vector of the attribute levels, $\boldsymbol{\beta}$ is the vector of unknown coefficients and ε_{nsj} is the random component. Moreover, suppose that the responses for the respondent n are given by the vector $\boldsymbol{y}_n^{'} = (y_{n11},...,y_{nsj},...,y_{nSJ})$. Therefore, in a RU theory framework, we assume that the respondent n chooses the alternative j from choice-set C_s by maximising the following utility fucntion:

$$y_{nsj} = \begin{cases} 1 & \text{if } \max(U_{nsj}) \text{ for one } j \in C_s \\ 0 & \text{otherwise} \end{cases}$$

The random component, ε_{nsj}, formula (2.6), is generally supposed to be independent and also Gumbel or type I extreme value distributed.

2.3.1 Multinomial Logit model

The Multinomial Logit (MNL) model is the simplest model belonging to the RU class of models. Accordint to the MNL model for the analysis of consumers' choices, the probability of a respondent n choosing the alternative j from choice-set C_s is expressed as follows (McFadden, 1974):

$$p_{nsj} = \frac{\exp(\boldsymbol{x}_{nsj}^{'}\boldsymbol{\beta})}{\sum\limits_{\substack{j=1 \\ j\in C_s}}^{J} \exp(\boldsymbol{x}_{nsj}^{\prime}\boldsymbol{\beta})} \tag{2.7}$$

where the the vector $\boldsymbol{\beta}$ is assumed as the vector of fixed and unknown coefficients. Although very simple, this model assumes the limiting property of the Independence from Irrelevant Alternatives (IIA property) that means that the probability in one choice-set is independent of the presence of other attribute values or any other alternative. By considering formula (2.7), it is easy to observe the practical limit of the IIA property. More precisely, if we compare two alternatives, j and l, we can see how the ratio is only expressed on the attribute values included in these two alternatives, without evaluating any other alternatives, as shown in the following formula:

$$\frac{p_{nsj}}{p_{nsl}} = \frac{\exp(\boldsymbol{x}_{nsj}^{'}\boldsymbol{\beta})}{\exp(\boldsymbol{x}_{nsl}^{'}\boldsymbol{\beta})} = \exp(\boldsymbol{x}_{nsj}^{'}\boldsymbol{\beta} - \boldsymbol{x}_{nsl}^{'}\boldsymbol{\beta}) \tag{2.8}$$

In fact, the IIA property implies an equal proportional substitution between alternatives. Moreover, the MNL model does not take care of: i) the presence of the preference heterogeneity among respondents; ii) the correlation that can arise given that a respondent faces a sequence of choice-sets. In order to relax these limiting assumptions, the Mixed Logit model has been devoloped (McFadden and Train, 2000).

2.3.2 Mixed Logit model

Unlike the MNL model, formula (2.7), the coefficients in the vector $\boldsymbol{\beta}$ in the Mixed Logit model are assumed as random variables (McFadden and Train, 2000). In our setting, we assume $\boldsymbol{\beta} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where MVN stands for the multivariate normal distribution; $\boldsymbol{\mu}$ is a vector of the means and $\boldsymbol{\Sigma}$ is a symmetric and positive definite diagonal matrix with $\sigma_1^2, ..., \sigma_K^2$ on its main diagonal (Sándor and Wedel, 2005). Then $\boldsymbol{\beta} = \boldsymbol{\mu} + \boldsymbol{Z}\boldsymbol{\sigma}$, where $\boldsymbol{Z}$ is a $K \times K$ diagonal matrix with the K elements $\boldsymbol{z} = (z_1, ..., z_K)$ on its main diagonal supposedly i.i.d. Standard Normal distributed. Therefore, in a Mixed Logit model, the probability that a respondent n chooses the alternative j from choice-set C_s is defined as in the following:

$$\pi_{nsj} = \int_{\mathbb{R}^K} p_{nsj}(\boldsymbol{\beta}) f(\boldsymbol{\beta}) d\boldsymbol{\beta} = \int_{\mathbb{R}^K} \frac{\exp(\boldsymbol{x}_{nsj}^{'}(\boldsymbol{\mu} + \boldsymbol{Z}\boldsymbol{\sigma}))}{\sum\limits_{\substack{j=1 \\ j \in C_s}}^{J} \exp(\boldsymbol{x}_{nsj}^{\prime}(\boldsymbol{\mu} + \boldsymbol{Z}\boldsymbol{\sigma}))} \phi(z_1)...\phi(z_K) d\boldsymbol{z} \tag{2.9}$$

As already stated (Chapter 1, Section 1.2), the Mixed Logit model (formula 2.9) allows to account for: i) the preference heterogeneity across the respondents, e.g. the Cross-Sectional Mixed Logit (C-MIXL) model (McFadden and Train, 2000), and ii) the correlation between the responses given by the same respondent on a sequence of choice-sets, by also allowing to

evaluate the preference heterogeneity, e.g. the Panel-Mixed Logit (P-MIXL) model (Revelt and Train, 1998). Following, we describe both types of Mixed Logit models, together with their corresponding Information matrices that do not have a closed-form expression and have to be approximated numerically.

C-MIXL model: FIM

The C-MIXL model assumes that the consumer's responses on a sequence of choice-sets are independent. Therefore, the derivation of the FIM for this type of RU model is based on the following log-likelihood function, expressed for all the respondents:

$$\log L(\boldsymbol{y}|\boldsymbol{\beta}) = \sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{j=1}^{J} y_{nsj}\log\left(\int_{\mathbb{R}^K}\left(\frac{\exp(\boldsymbol{x}_{nsj}^{'}(\boldsymbol{\mu}+\boldsymbol{Z}\boldsymbol{\sigma}))}{\sum\limits_{\substack{j=1\\ j\in C_s}}^{J}\exp(\boldsymbol{x}'_{nsj}(\boldsymbol{\mu}+\boldsymbol{Z}\boldsymbol{\sigma}))}\right)d\Phi(\boldsymbol{z})\right) \tag{2.10}$$

Therefore, the C-MIXL FIM is defined as follows:

$$I(\boldsymbol{X},\boldsymbol{\beta}) = \sum_{n=1}^{N}\sum_{s=1}^{S} I(\boldsymbol{X}_{ns},\boldsymbol{\beta}) = \sum_{n=1}^{N}\sum_{s=1}^{S}\left\{\begin{array}{cc} \Pi'_{ns}\boldsymbol{D}_{ns}^{-1}\Pi_{ns} & \Pi'_{ns}\boldsymbol{D}_{ns}^{-1}\boldsymbol{R}_{ns} \\ \boldsymbol{R}'_{ns}\boldsymbol{D}_{ns}^{-1}\Pi_{ns} & \boldsymbol{R}'_{ns}\boldsymbol{D}_{ns}^{-1}\boldsymbol{R}_{ns}\end{array}\right\} \tag{2.11}$$

where:

$$\Pi_{ns} = \textstyle\int_{\mathbb{R}^K}[\mathrm{P}_{ns}(z) - \mathrm{p}_{ns}(z)\mathrm{p}'_{ns}(z)]\boldsymbol{X}_{ns}\phi(z_1)...\phi(z_K)d\boldsymbol{z};$$

$$\boldsymbol{R}_{ns} = \textstyle\int_{\mathbb{R}^K}[\mathrm{P}_{ns}(z) - \mathrm{p}_{ns}(z)\mathrm{p}'_{ns}(z)]\boldsymbol{X}_{ns}\boldsymbol{Z}\phi(z_1)...\phi(z_K)d\boldsymbol{z};$$

$$\mathrm{p}_{ns}(z) = (p_{ns1}(z),\ \ldots,\ p_{nsJ}(z))';$$

$$\mathrm{P}_{ns}(z) = \mathrm{diag}(p_{ns1}(z),\ \ldots,\ p_{nsJ}(z))\ ;$$

$$\boldsymbol{D}_{ns} = \mathrm{diag}(\pi_{ns1}\ldots\pi_{nsJ});$$

$\boldsymbol{X}_{ns}$ is the design matrix for choice-set C_s evaluated by respondent n;
$\boldsymbol{Z}$ is a $K\times K$ diagonal matrix that contains the standard normal distributed random draws. The probabilities in the vector $\mathrm{p}_{ns}(z)$ are calculated according to formula (2.7) and the probabilities $\pi_{ns1},...,\pi_{nsJ}$ in the matrix $\boldsymbol{D}_{ns}$ are calculated according to formula (2.9). For the detailed derivation of the C-MIXL FIM, see Sándor and Wedel (2002).

P-MIXL model: FIM

Differently from the C-MIXL model, according to the P-MIXL specification, the responses expressed by the same respondent are no longer assumed indipendent. Therefore, the P-MIXL log-likelihood expressed for all the respondents is given by:

$$\log L(\boldsymbol{y}|\boldsymbol{\beta}) = \sum_{n=1}^{N} \log \left(\int_{\mathbb{R}^K} \prod_{s=1}^{S} \prod_{j=1}^{J} \left(\frac{\exp(\boldsymbol{x}_{nsj}^{'}(\boldsymbol{\mu} + \boldsymbol{Z}\boldsymbol{\sigma}))}{\sum_{\substack{j=1 \\ j \in C_s}}^{J} \exp(\boldsymbol{x}_{nsj}^{'}(\boldsymbol{\mu} + \boldsymbol{Z}\boldsymbol{\sigma}))} \right)^{y_{nsj}} d\Phi(\boldsymbol{z}) \right) \tag{2.12}$$

that reflects the fact that the responses given by the same respondent are correlated, while those between different respondents are assumed to be independent. The full expression of the P-MIXL FIM is more complex with respect to that of the C-MIXL, formula (2.11), due to the further complexity of dealing with products of logit probabilities in the log-likelihood function, formula (2.12), compared to the summation, formula (2.10). Based on the detailed derivations of Sándor (2013) and Zhang et al. (2017), we give the following expression for the P-MIXL FIM, by considering that a respondent is asked to express his/her preferences for S choice-sets:

$$\tilde{I}(\boldsymbol{X}, \boldsymbol{\beta}) = \sum_{n=1}^{N} \tilde{I}(\boldsymbol{X}_n, \boldsymbol{\beta}) = \sum_{n=1}^{N} E_{y_n} \begin{pmatrix} \boldsymbol{V}_n^{'} \Delta_n^{-1} \boldsymbol{V}_n & \boldsymbol{V}_n^{'} \Delta_n^{-1} \Lambda_n \\ \Lambda_n^{'} \Delta_n^{-1} \boldsymbol{V}_n & \Lambda_n^{'} \Delta_n^{-1} \Lambda_n \end{pmatrix} \tag{2.13}$$

where:

$$\boldsymbol{V}_n = \int_{\mathbb{R}^K} \left(\boldsymbol{P}_n - \boldsymbol{p}_n \boldsymbol{p}_n^{'} \right) \boldsymbol{X}_n d\Phi(\boldsymbol{z});$$

$$\Delta_n = \begin{pmatrix} \text{diag}(\pi_{n1}) & \pi_{n1}\pi_{n2}^{'} & \dots & \pi_{n1}\pi_{nS}^{'} \\ \pi_{n2}\pi_{n1}^{'} & \text{diag}(\pi_{n1}) & \dots & \pi_{n2}\pi_{nS}^{'} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{nS}\pi_{n1}^{'} & \pi_{nS}\pi_{n2}^{'} & \dots & \text{diag}(\pi_{nS}) \end{pmatrix};$$

$$\Lambda_n = \int_{\mathbb{R}^K} \left(\boldsymbol{P}_n - \boldsymbol{p}_n \boldsymbol{p}_n^{'} \right) \boldsymbol{X}_n \boldsymbol{Z}_n d\Phi(\boldsymbol{z});$$

$$\boldsymbol{P}_n = \begin{pmatrix} \text{diag}(\boldsymbol{p}_{n1}) & \boldsymbol{p}_{n1}\boldsymbol{p}_{n2}^{'} & \dots & \boldsymbol{p}_{n1}\boldsymbol{p}_{nS}^{'} \\ \boldsymbol{p}_{n2}\boldsymbol{p}_{n1}^{'} & \text{diag}(\boldsymbol{p}_{n1}) & \dots & \boldsymbol{p}_{n2}\boldsymbol{p}_{nS}^{'} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{p}_{nS}\boldsymbol{p}_{n1}^{'} & \boldsymbol{p}_{nS}\boldsymbol{p}_{n2}^{'} & \dots & \text{diag}(\boldsymbol{p}_{nS}) \end{pmatrix};$$

$$\boldsymbol{Z}_n = (\boldsymbol{Z}_1, ..., \boldsymbol{Z}_S);$$

$$\boldsymbol{P}_{\text{n}} = \text{diag}(\boldsymbol{p}_{\text{n1}},\ \dots,\ \boldsymbol{p}_{\text{nS}})\ ;$$

$$\boldsymbol{p}_n = (\boldsymbol{p}_{n1}, ..., \boldsymbol{p}_{ns}, ..., \boldsymbol{p}_{nS})' \text{ where each } \boldsymbol{p}_{ns} = (p_{ns1}, ..., p_{nsJ})';$$

$\boldsymbol{X}_n$ is the design matrix containing the S choice-sets evaluated by a respondent n;
$\boldsymbol{Z}_n = (\boldsymbol{Z}_1, ..., \boldsymbol{Z}_S)$ are diagonal matrices that contain the standard normal random draws for respondent n. As before, the probabilities in the vectors $\boldsymbol{p}_n$ and π_n are calculated through formulas (2.7) and (2.9) respectively. It must be noted that for both the C-MIXL and P-MIXL models, the dimension of the coefficient vector $\boldsymbol{\beta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ is equal to $2K$ (e.g. K coefficients for the vector $\boldsymbol{\mu}$ and K coefficients for the vector $\boldsymbol{\sigma}$).

For both the C-MIXL and the P-MIXL models, the computation of the FIM involves multidimensional integrals that have to be approximated numerically through the commonly used methods such as quasi-Monte Carlo or Gaussian quadratures. The approximation with the quasi-Monte Carlo methods with Halton sequences is usually applied for the Mixed Logit model (Train, 2000; Train, 2003; Yu et al., 2010), and this is also the case in this project. Moreover, the further complexity in the derivation of the P-MIXL FIM is the calculation of $E_{\boldsymbol{y}_n}$ in formula (2.13) that involves the sum over all possible realisations of $\boldsymbol{y}_n$, and equal to J^S. To this end, a further simulation is needed in order to evaluate the expression in formula (2.13). In our case we approximate this expectation with Monte Carlo simulations (Sándor, 2013).

2.4 A new proposal for the construction of optimal heterogeneous choice designs

2.4.1 Approximate Heterogeneous Choice Design with correlated preferences

In what follows we describe the approach proposed on how to efficiently build heterogeneous choice designs for correlated preferences based on an approximate design theory. Our aim is to build subdesigns composed of choice-sets that have to be administered to the respondents

according to optimal weights. We call these subdesigns "groups" of choice-sets and indicate them by $\varsigma_1, \varsigma_2, ..., \varsigma_g, ..., \varsigma_G$. Furthermore, we assume that:

- each group ς_g $(g = 1, ..., G)$ contains m choice-sets;
- no choice-set can appear twice in the same group ς_g, but the same choice-set could appear more than once in different groups;
- observations (i.e. the respondent's choices) within the same group are correlated while observations between groups are independent.

Hence, we focus on an approximate design framework and generalise formula (2.1) in order to consider as design points the groups of choice-sets with fixed size m as follows (Atkinson and Woods, 2015):

$$\xi_G = \begin{Bmatrix} \varsigma_1 & \varsigma_2 & \dots & \varsigma_g & \dots & \varsigma_G \\ w_1 & w_2 & \dots & w_g & \dots & w_G \end{Bmatrix} \tag{2.14}$$

under the constraints of $0 \leq w_g \leq 1$ and $\sum_{g=1}^{G} w_g = 1$. It must be noted that according to the design in formula (2.14), each weight w_g indicates the proportion of respondents to which the group of choice-sets ς_g should be supplied.

Therefore, the FIM for the design ξ_G is given by the following:

$$\tilde{I}(\xi_G) = \sum_{g=1}^{G} w_g \tilde{I}(\boldsymbol{X}_g, \boldsymbol{\beta}) \tag{2.15}$$

where $\tilde{I}(\boldsymbol{X}_g, \boldsymbol{\beta})$ is the FIM for the g-th group composed of m choice-sets, and $\boldsymbol{X}_g$ is the design matrix containing the m choice-sets in group g. The optimality of the design ξ_G could be assessed through the multivariate General Equivalence theorem (Fedorov, 1972; Atkinson, 2008; Atkinson and Woods, 2015).

2.4.2 Construction of the $D_{comp}-$optimal heterogeneous choice designs

In an approximate design framework, when considering just one choice-set as a design point, the design space is equal to the number of all possible choice-sets, Q. That is, if L is the number of all possible experimental combinations, then $Q = \binom{L}{J}$. Conversely, when the design points are the groups of choice-sets, as in formula (2.14), the design space, G, equal to $G = \binom{Q}{m}$, becomes prohibitively huge and in very practical situations (as in for our case study), the search over G cannot be computed. As an example of this fact, suppose that

we are dealing with a choice experiment with four attributes each at three levels, and three alternatives in each choice-set. Then $L = 3^4 = 81$ and $Q = \binom{81}{3} = 85320$. Suppose that we want to build an optimal heterogeneous choice design with groups of choice-sets of size eight. Then $G = \binom{85320}{8} = 6.9 \times 10^{34}$. In fact, this highlights how the search is impossible over this very large number of design points even when considering approximate designs.

To this end, in order to deal with this computational issue, we propose the following approach for obtaining optimal heterogeneous choice designs that can be summarised briefly in the following three main steps:

1) we consider the C-MIXL model specification (formula (2.10)) that is equivalent to the situation in which each respondent receives just one choice-set, for N respondents. Note that under the C-MIXL model we have to ensure the estimation of a minimum of $2K$ coefficients. Therefore, we compute the optimal choice design for the C-MIXL model according to formula (2.1), where the design point is a choice-set, and obtain the optimal design points $C_1, ..., C_{2K}$ where each one is a choice-set to be administered to a total of N respondents.

2) we extend the heterogeneous choice design to the situation in which a respondent receives a subset of m choice-sets. In order to obtain a new design space for the search, say $\tilde{G}$, with a size in which it is possible to carry out the search, we consider the optimal design points $C_1, ..., C_{2K}$ obtained at Step 1 as the ones that will compose the groups of choice-sets of size m for the search. Therefore, we compute the optimal heterogeneous choice design consisting of groups of choice-sets of size m, where the design space for the search is equal to $\tilde{G} = \binom{2K}{m}$, under the P-MIXL model that is equivalent to the situation in which each respondent receives a subset of the choice-sets by accounting for the correlation between the responses given by the same individual;

3) once the optimal groups of choice-sets are obtained through the search on $\tilde{G}$, we administer each one to a proportion of respondents according to the optimal weights.

Through our proposal, we are able to obtain optimal groups of choice-sets by searching over $\tilde{G}$ groups of choice-sets, where $\tilde{G} \ll G$, in this way overcoming the computational issue regarding the impossibility of searching over the entire design space G. Thanks the use of an approximate design theory, the optimality of the optimal groups of choice-sets is checked and guaranteed through the GET. Moreover, our approach allows for making use of both the Mixed Logit model specifications, e.g. the C-MIXL and the P-MIXL models, and consequently of obtaining optimal heterogeneous choice designs that are fully consistent with the underlying RU theory related to both models.

It must be noted that in order to verify the validity of our proposal, e.g. whether the optimal choice-sets obtained in Step 1 are similar to those that compose the groups of choice-sets in Step 2, we also carried out simulations by considering a simple choice experiment in which the search for the groups in Step 2 could be performed over the entire design space G. We anticipate that according to these simulation results (presented in detail in the next Section), the optimal groups of choice-sets obtained in Step 2 contain the same optimal choice-sets as those obtained in Step 1, a result that further confirms that our innovative proposal is likely to work.

2.5 Simulation results for the proposal of construction of approximate heterogeneous choice designs

In what follows, we present the simulation results on a choice experiment in order to test the validity of our proposal (Subsection 2.4.2), e.g. whether the optimal choice-sets obtained in Step 1 are similar to those that compose the optimal groups of choice-sets obtained in Step 2. These simulations are necessary because by considering this simple choice experiment it is possible to search over the space G for all possible groups of choice-sets.

For the simulated choice experiment, we assume that we have just two attributes related to the choice experiment: 1) "Type of Coffee" at two levels: "100% Arabica" and "Arabica-Robusta", and 2) "Price" at two levels "€4.50" and "€6.00". Furthermore, we also assume a third attribute related to the scores of the sensory assessment, "Tasting", at two levels "Score for Coffee no.1" and "Score for Coffee no.2". We consider two alternatives for each choice-set (e.g. $J = 2$). Given the dependence of the FIM on the unknown coefficient values, for all the optimal designs computed here, we use the following nominal values: $\boldsymbol{\mu} = (\mu_1 = -1.5, \mu_2 = 1.5, \mu_3 = -1.5)$ and $\boldsymbol{\sigma} = (\sigma_1 = \sqrt{1.5}, \sigma_2 = \sqrt{1.5}, \sigma_3 = \sqrt{1.5})$ chosen in terms of medium response accuracy and medium respondent heterogeneity (Arora and Huber, 2001; Toubia et al., 2004, Zhang et al., 2017). All the optimal choice designs presented here are computed under the compound D-criterion, formula (2.4), through the mOWE algorithm (Interactive Matrix Language-IML; SAS, Windows Platform vs.9.4.). Moreover, the design optimality is checked and verified through the GET.

One choice-set

Following the three Steps of our proposal, in Step 1 we first build an optimal choice design, ξ_1, under the C-MIXL model. The total number of experimental combinations, L, is equal to $L = 2^3 = 8$, and the size of all possible choice-sets Q is equal to $Q = \binom{L}{J} = \binom{8}{2} = 28$. We

Table 2.1 Optimal design ξ_1

Design points	*Weights*
C_4	0.13
C_7	0.17
C_{10}	0.12
C_{17}	0.21
C_{20}	0.18
C_{21}	0.19

have to ensure the estimation of six coefficients (e.g. $2K = 6$): three coefficients related to the vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$, and three coefficients related to the vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$. The optimal design, ξ_1, we have obtained is supported on six design points and reported in Table 2.1. If the choice-sets of the optimal design ξ_1 are supposed to be administered to the respondents, this means that each respondent will receive just one of these choice-sets according to the optimal weight of the specific choice-set (Table 2.1). More precisely, the choice-set no.4 should be administered to 13% of the respondents, the choice-set no.7 to 17% of the respondents, and similarly for the rest of the choice-sets (Table 2.1).

In order to determine the value of α for the compound design criterion, Figure (2.1) shows the plot of the efficiencies against different values of α. According to this plot, we choose a value of $\alpha = 0.75$ for which we obtain 56% efficiency with respect to a D-optimal design for the attributes of the choice experiment, and 56% efficiency with respect to a D_s-optimal design for the attribute related to the "Tasting".

Two, three, four and five choice-sets

Once we have obtained the optimal choice design ξ_1 at Step 1, we proceed to compute the optimal heterogeneous choice designs composed of groups of two, three, four and five choice-sets (Step 2). When considering the groups of choice-sets as design points, the size of the design space increases exponentially when the number of choice-sets in each group increases. In fact, the design space G rises from $\binom{Q}{m} = \binom{28}{2} = 378$ for groups with two choice-sets, to $\binom{Q}{m} = \binom{28}{5} = 98280$ for groups consisting of five choice-sets.

Table 2.2 contains the optimal heterogeneous choice designs obtained, all of which were checked via GET for groups of 2-5 choice sets (labelled ξ_2, ξ_3, ξ_4 and ξ_5 respectively). It must be noted that the numbering of the design points, e.g. the groups composed of choice-sets, is obtained computationally through the function "allcomb" in SAS, Interactive Matrix Language-IML (SAS, Windows Platform vs.9.4.).

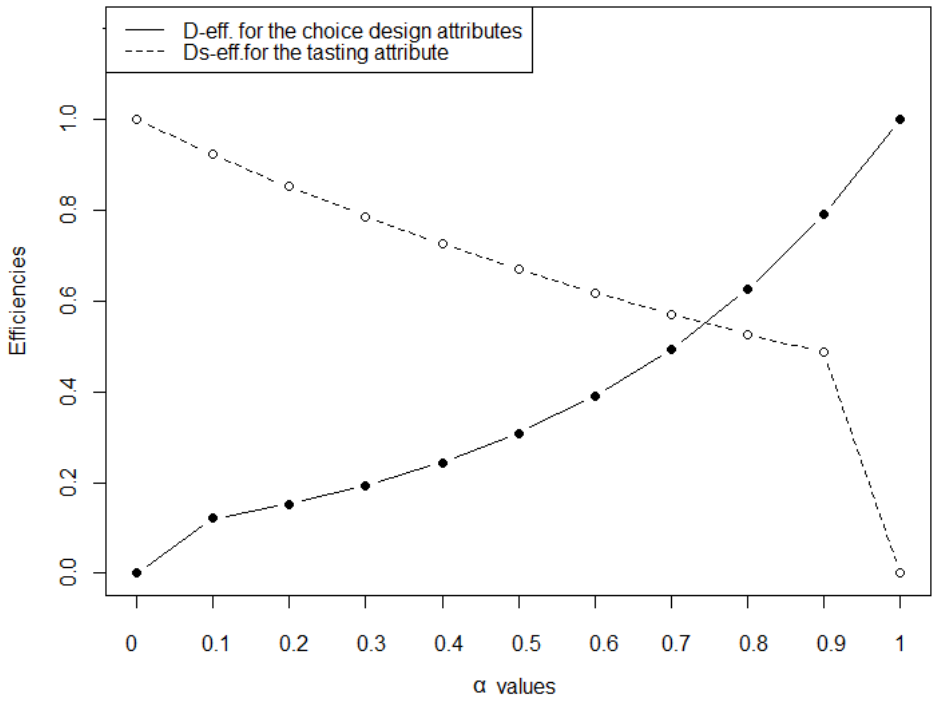

Fig. 2.1 Plot of the efficiencies for different values of α for design ξ_1

Table 2.2 Optimal heterogeneous choice designs for groups of two, three, four and five choice-sets

Design	*Design points*	*Weights*
ξ_2		
	(C_7, C_{10})	0.34
	(C_4, C_{20})	0.30
	(C_{17}, C_{20})	0.04
	(C_{17}, C_{21})	0.29
	(C_4, C_{17})	0.03
ξ_3		
	(C_4, C_7, C_{10})	0.17
	(C_7, C_{10}, C_{17})	0.33
	(C_4, C_{20}, C_{21})	0.32
	(C_{17}, C_{20}, C_{21})	0.18
ξ_4		
	$(C_4, C_7, C_{10}, C_{20})$	0.35
	$(C_4, C_{10}, C_{17}, C_{21})$	0.36
	$(C_7, C_{17}, C_{20}, C_{21})$	0.29
ξ_5		
	$(C_4, C_7, C_{10}, C_{17}, C_{21})$	0.37
	$(C_4, C_{10}, C_{17}, C_{20}, C_{21})$	0.34
	$(C_4, C_7, C_{10}, C_{17}, C_{20})$	0.29

As can be observed in Table 2.2, all the optimal designs consisting of groups of two, three, four and five choice-sets contain exclusively the optimal choice-sets selected in the design ξ_1. This important result confirms that our proposal is likely to work.

In Figures 2.2 and 2.3 we report the efficiencies of the optimal choice designs (ξ_2, ξ_3, ξ_4 and ξ_5) against different values of α. More precisely, in order to select the best value of α, we plot against α, the D and D_s efficiencies of both the attributes of the choice experiment and the attribute related to the tasting. As can be observed from the plots in Figures 2.2 and 2.3, there is an increasing efficiency when the number of choice-sets per group increases. In fact, the efficiency increases from 64% for groups with two choice-sets to 75% for groups composed of five choice-sets.

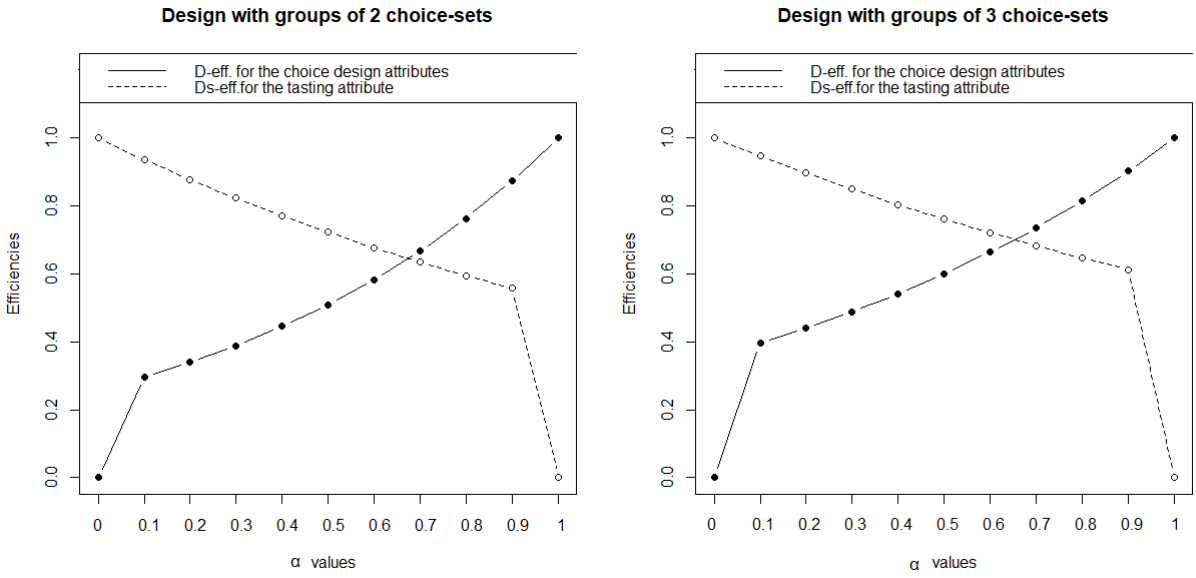

Fig. 2.2 Plot of the efficiencies for different values of α for designs ξ_2 and ξ_3

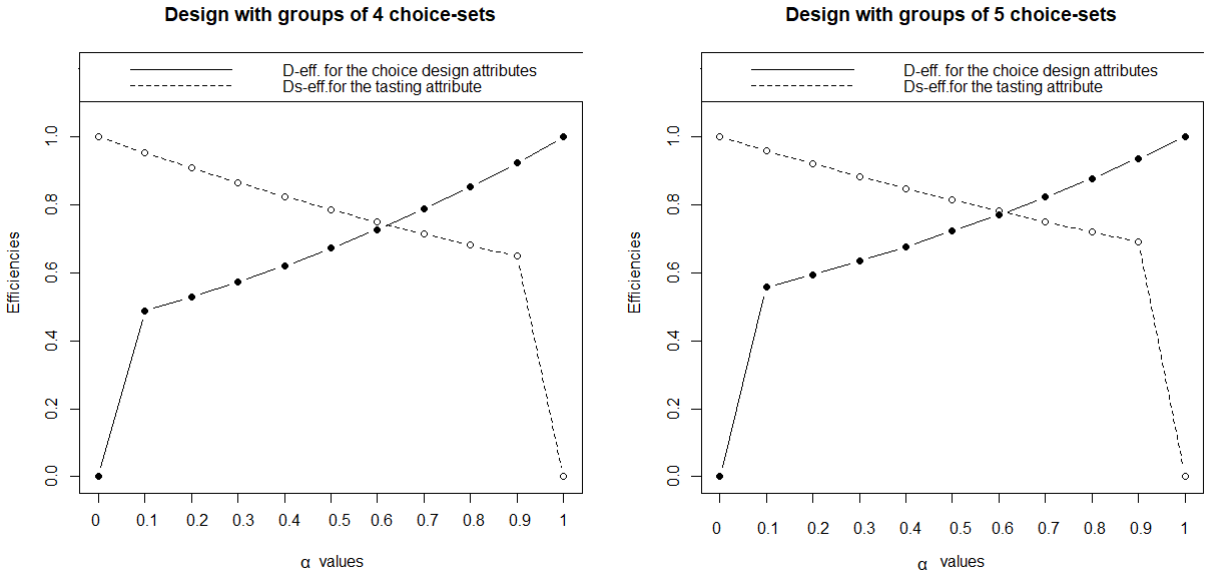

Fig. 2.3 Plot of the efficiencies for different values of α for designs ξ_4 and ξ_5

2.6 A real case study

2.6.1 Integrating a choice experiment with consumer sensory test and chemical analysis

In what follows we describe the real case study from which our proposal has been motivated. As already mentioned, it regards the analysis of consumers' preferences for coffee integrating a choice experiment with consumer sensory test and chemical analysis. Therefore, we build an optimal heterogeneous choice design under the compound design criterion, formula (2.4), in order to obtain: i) an efficient estimation of the attributes of the choice experiment; and ii) detection of the effect related to the HPLC measurement results for the caffeine (Choice 1 before tasting) and the scores obtained from the consumer sensory test (Choice 2 after tasting).

HPLC analysis

Firstly, two types of coffee with different organoleptic characteristics were chosen: an intense, soft and aromatic blend (100% Arabica) and a round blend with a high aftertaste intensity (Arabica and Robusta varieties). The two types of coffee selected were analysed with respect to their caffeine content with an HPLC method for obtaining the quantity of caffeine contained in each one.

Guided tasting: consumer sensory tests

Once the HPLC results are obtained, a consumer sensory test is also planned. More precisely, two scoring cards, one for each type of coffee, are developed for the organoleptic evaluation and the consumers have to give a score for the organoleptic descriptors of each coffee (Masi et al., 2013). The ten organoleptic descriptors are as follows: the colour related to sight, the intensity and the quality of the aroma related to olfaction; four different descriptors related to taste, namely, bitter, acidic, sweet and aroma; the tactile sensation related to body, aftertaste and the general equilibrium of the coffee. For each type of coffee, respondents gave a score expressed on a scale of seven points. Subsequently, we obtained two overall evaluation scores for each type of coffee, properly normalised and standardised.

Attributes and levels for the choice experiment

We identified six attributes of the choice experiment reported in Table 1. The first is the type of coffee, labelled by "Coffee Type", at two levels (a blend of Arabica and Robusta and a

Table 2.3 Attributes and levels for the choice experiment

Attribute	**Levels (coding in brackets)**
Coffee Type	1: Blend of Arabica and Robusta (-1) 2: Blend of 100% Arabica (1)
Packaging	1: Soft bag in a modified atmosphere (-1) 2: Jar in a modified atmosphere (1)
Label Indication	1: Geographical origin (-1) 2: Certification of sustainability (1)
Intense Aromatic Taste	1: Fairly present (-1) 2: Highly present (1)
Soft Velvety Taste	1: Fairly present (-1) 2: Highly present (1)
Price	1: €4.50 2: €6.00 3: €7.50

blend of 100% Arabica); the second is the packaging (labelled by "Packaging") at two levels: a soft bag in a modified atmosphere and a jar in a modified atmosphere. We established two attributes related to the taste: a Soft and Velvety taste ("Soft Velvety Taste") and an Intense and Aromatic taste ("Intense Aromatic Taste"), both at two levels, e.g. fairly present and highly present. Moreover, in order to evaluate the consumers' preferences regarding the sustainability and geographical origin, we included the attribute "Label Indication" at two levels: "Geografical Origin", that is, the indication of the geographical origin of the coffee, and "Certification of Sustainability", that is, the presence of any type of certification of sustainability (economic, social and/or environmental). The last attribute we identified was the price ("Price") for a quantity of 250 grams of coffee at three levels: €4.50, €6.00 and €7.50. The coded levels for each attribute are reported in brackets in Table 3.1.

Procedure for administering the Choice Experiment

In order to administer the choice experiment we consider a similar procedure performed by Lombardi et al. (2017) for evaluating the consumers' preferences for milk referring to several organoleptic characteristics. They administered the same choice experiment twice, before and after an informative video projection, and highlighting the important role played by the information in defining the respondents' preferences. Therefore, according to our procedure of administering, a questionnaire containing items related to baseline variables (e.g. age, marital status, consumption and purchasing of the coffee) is first administered to each respondent. Subsequently, the choice-sets are administered to the respondents (Choice 1

before tasting). Once the first choice experiment session is completed, the consumer sensory test is conducted by an expert. During this step, the expert also provides information about each type of coffee, explaining its organoleptic properties. In the meantime, the respondents taste both types of coffee and mark their scores on the sensory assessment score cards. Once the guided tasting is completed, we once again administer the same choice-sets (Choice 2 after tasting) to find out if there are any differences in the consumers' preferences collected from Choice 1 and Choice 2 sessions due to the information step and the consumer sensory test conducted through the guided tasting.

2.6.2 Optimal heterogeneous choice designs

In this Subsection we describe the heterogeneous choice design for our real case study obtained according to our proposal described in Subsection 2.4.2 through the mOWE algorithm (Liu and Tang, 2015). When considering our real case study, we have six attributes related to the choice experiment: five attributes at two levels and one attribute at three levels (Table 3.1). Moreover, we also have one attribute related to the caffeine results from the HPLC analysis (evaluated in Choice 1), and the sensory scores (evaluated in Choice 2) at two levels; namely, the low level (-1) related to the caffeine results and the scores for the blend Arabica and Robusta, and the high level (1) related to those obtained for the blend 100% Arabica. It must be noted that when we estimate Choice 1, we include the attribute "Caffeine" at two levels related to the quantity of caffeine in the model. Instead, when we estimate Choice 2 (after tasting), we include the attribute "Taste Score" related to the scores obtained through the guided tasting.

Optimal choice design when each respondent receives one choice-set

By following our proposal in Subsection 2.4.2, firstly (Step 1) we build the optimal heterogeneous choice design under the C-MIXL model. The total number of experimental combinations L is equal to: $L = 2^6 \cdot 3 = 192$, and by considering binary alternatives, the total number of possible choice-sets, Q, is equal to: $Q = \binom{L}{J} = \binom{192}{2} = 18336$. We have to ensure the estimation of sixteen coefficients: eight related to the vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_8)$ and eight related to the vector $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_8)$. As before, we used the following nominal values, chosen in terms of medium response accuracy and medium respondent heterogeneity: $\boldsymbol{\mu} = (\mu_1 = -1.5, \mu_2 = 1.5, \mu_3 = -1.5, \mu_4 = 1.5, \mu_5 = -1.5, \mu_6 = 1.5, \mu_7 = 0, \mu_8 = 1.5)$ and $\boldsymbol{\sigma} = (\sigma_1 = \sqrt{1.5}, \sigma_2 = \sqrt{1.5}, \sigma_3 = \sqrt{1.5}, \sigma_4 = \sqrt{1.5}, \sigma_5 = \sqrt{1.5}, \sigma_6 = \sqrt{1.5}, \sigma_7 = \sqrt{1.5}, \sigma_8 = \sqrt{1.5})$. The optimal choice design we obtained under the C-MIXL model is supported on 16 points as reported in Table 2.4 and the design optimality was checked via GET.

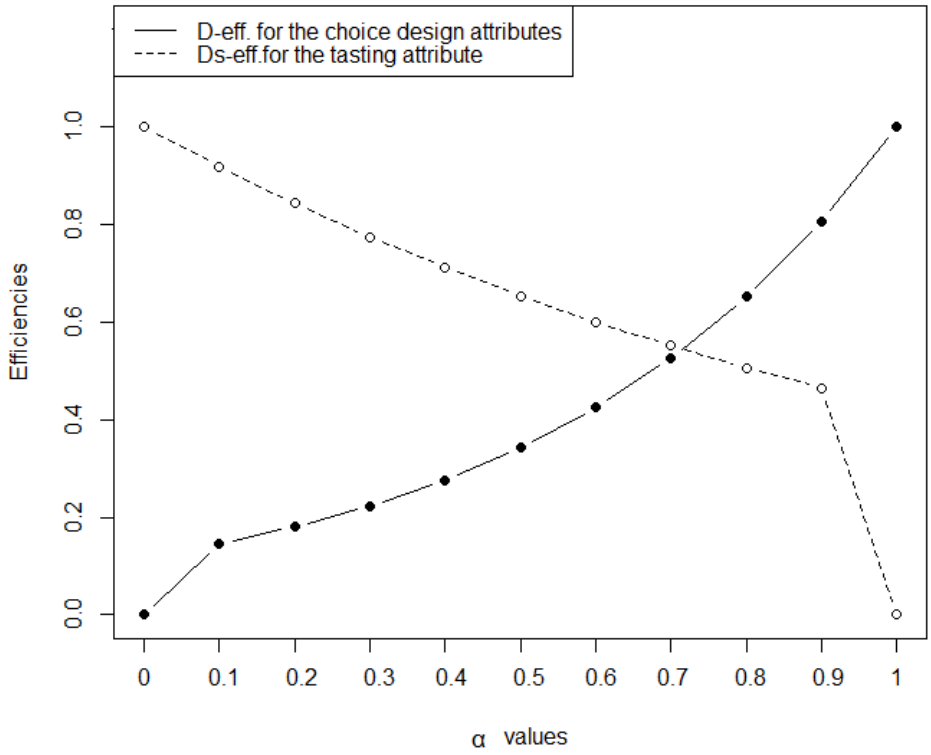

Fig. 2.4 Plot of the efficiencies for different values of α

Table 2.4 Optimal design ξ^*

Design points	*Weights*
C_{1423}	0.06
C_{3337}	0.01
C_{3494}	0.04
C_{3955}	0.05
C_{6903}	0.06
C_{8328}	0.07
C_{8857}	0.09
C_{11050}	0.09
C_{11647}	0.13
C_{12388}	0.05
C_{12939}	0.04
C_{13601}	0.14
C_{16646}	0.03
C_{16768}	0.04
C_{16852}	0.06
C_{17122}	0.04

By plotting the efficiencies for different values of α in Figure 3.3, the best value of α is equal to 0.72 for which we obtain 57% efficiency with respect to a D-optimal design for the attributes of the choice experiment, and 57% efficiency with respect to a D_s-optimal design related to the caffeine and the scores from the guided tasting.

Optimal heterogeneous choice design when each respondent receives a group of eight choice-sets

Once the optimal design ξ^* has been obtained (Step 1, Section 2.4.2), we compute the optimal heterogeneous choice design composed of groups of eight choice-sets (Step 2, Section 2.4.2). Note that in this case we cannot search over the entire design space G that is extremely vast and equal to: $G = \binom{Q}{m} = \binom{18336}{8} = 1417526 \times 10^{39}$. In order to address this issue, we apply our proposal by combining the optimal choice-sets obtained in the design ξ^* in groups of size $m = 8$. Therefore, the new design space $\tilde{G}$ for the search of the optimal groups of choice-sets of size eight is equal to: $\tilde{G} = \binom{2K}{m} = \binom{16}{8} = 18872$, where $2K$ are the optimal choice-sets in the design ξ^*.

The optimal heterogeneous choice design consisting of groups of eight choice-sets is supported on six design points reported in Table 2.5 as well as the corresponding optimal weights. Therefore, according to our innovative proposal (Subsection 2.4.2, Step 3), each group of choice-sets of the design $\xi^*_{\tilde{G}}$ is subsequently supplied to a proportion of respondents according

Table 2.5 Optimal design $\xi^*_{\tilde{G}}$

Design points	*Weights*
$\{C_{12388}, C_{13601}, C_{17122}, C_{8328}, C_{3494}, C_{1423}, C_{18852}, C_{16768}\}$	0.10
$\{C_{3955}, C_{8857}, C_{3494}, C_{6903}, C_{11647}, C_{11050}, C_{16768}, C_{3337}\}$	0.15
$\{C_{3955}, C_{8857}, C_{13601}, C_{8328}, C_{3494}, C_{11647}, C_{12939}, C_{3337}\}$	0.13
$\{C_{17122}, C_{3494}, C_{1423}, C_{16646}, C_{11647}, C_{11050}, C_{12939}, C_{16768}\}$	0.14
$\{C_{3955}, C_{12388}, C_{13601}, C_{1423}, C_{16646}, C_{11647}, C_{16852}, C_{12939}\}$	0.25
$\{C_{12388}, C_{8857}, C_{17122}, C_{8328}, C_{6903}, C_{16646}, C_{11050}, C_{3337}\}$	0.23

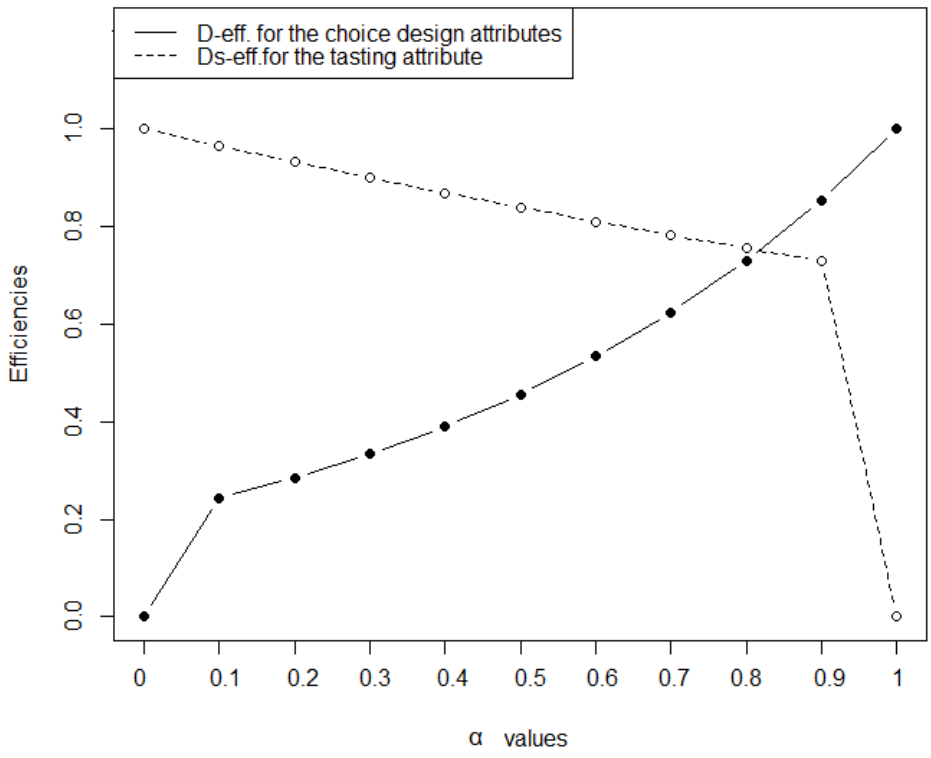

Fig. 2.5 Plot of the efficiencies for different values of α

to the corresponding weight of the specific group of choice-sets. For example, the first group composed of the following choice-sets : $\{C_{12388}, C_{13601}, C_{17122}, C_{8328}, C_{3494}, C_{1423}, C_{18852}, C_{16768}\}$ will be supplied to 10% of the respondents; the second group composed of the choice-sets $\{C_{3955}, C_{8857}, C_{3494}, C_{6903}, C_{11647}, C_{11050}, C_{16768}, C_{3337}\}$ will be supplied to 15% of the respondents, and similarly for the rest of the groups of choice-sets (Table 2.5).

From plotting the efficiencies for different values of α in Figure (2.5), the best value of α is equal to 0.8 for which we obtain 75% efficiency with respect to a D-optimal design for the attributes of the choice experiment, and 75% efficiency with respect to a D_s-optimal design related the attribute of "Caffeine" (Choice 1) and "Taste Score" (Choice 2).

2.6.3 P-MIXL model estimates

In what follows, we describe the estimation results obtained for the P-MIXL model related to the data collected for a total of 35 respondents by considering: i) Choice 1, that is, the choice experiment administered before the guided tasting session (Table 2.6), and ii) Choice 2, that is, the same choice experiment administered after the guided tasting session (Table 2.7). For each attribute k ($K = 8$) we estimate its main effect labelled by $\hat{\mu}_k$, and its heterogeneity effect labelled by $\hat{\sigma}_k$ in Tables 2.6 and 2.7. Moreover, the $\hat{\sigma}_k$ estimates obtained express the effect related to the respondents' heterogeneity, e.g. for a given attribute k, the higher the value $\hat{\sigma}_k$ estimates, the greater the consumers' heterogeneity for the k-th attribute.

Firstly, from the estimated results related to the P-MIXL models reported in Tables 2.6 and 2.7, we can see that by applying our proposal we have obtained excellent estimates for the attributes involved in the choice experiment. More specifically, for both P-MIXL models the convergence is perfectly achieved, and we have also obtained highly significant p-values (p-value (Choice 1)< 0.001) and p-value (Choice 2)< 0.001) related to the Likelihood-Ratio test: a result indicating that both models are statistically significant. Moreover, almost all the estimated coefficients for Choice 1 and Choice 2 are statistically significant with small standard errors. All these results demonstrate the efficiency of the underlying choice design, and consequently the efficiency of our innovative approach.

When observing the results for Choice 1 (Table 2.6) and Choice 2 (Table 2.7), we can note that the guided tasting session, together with the information provided on each type of coffee, play a relevant role in unequivocally determining the respondents' preferences. As a matter of fact, when facing Choice 1, the consumers' preferences are collected in a situation of complete "misinformation", e.g. each respondent chooses only according to their previous knowledge about the coffee and own lifestyle. Instead, when facing Choice 2, the respondents were exposed to: i) information given by an expert about the taste properties and organoleptic characteristics of the two types of coffee, and ii) a guided tasting of the

Table 2.6 Panel Mixed Logit Model results before tasting: Choice 1

	Estimate	*Std.Error*	*Z value*	*p-value*
$\hat{\mu}$ ***(main effect)***				
Coffee Type	1.4179	0.5551	2.55	0.011
Packaging	-1.1127	0.4332	-2.57	0.010
Intense Aromatic Taste	-1.5353	0.4686	-3.28	0.001
Soft Velvety Taste	3.0697	0.8528	3.60	0.000
Label Indication	1.3485	0.5067	2.66	0.008
Caffeine	-2.5869	0.9792	-2.64	0.008
Price 2 (level 2)	1.4828	0.8618	1.72	0.085
Price 3 (level 3)	-3.0143	1.0749	-2.80	0.005
$\hat{\sigma}$ ***(heterogeneity effect)***				
Coffee Type	-1.3214	0.4783	-2.76	0.006
Packaging	1.9310	0.5946	3.25	0.001
Intense Aromatic Taste	0.0610	0.2669	0.23	0.819
Soft Velvety Taste	3.3874	0.9527	3.56	0.000
Label Indication	0.7743	0.3266	2.37	0.018
Caffeine	4.7288	1.6911	2.80	0.005
Price 2 (level 2)	-4.1939	1.5510	-2.70	0.007
Price 3 (level 3)	1.5805	0.6963	2.27	0.023

two types of coffee carried out by the same expert. Therefore, in Choice 2 the respondents' preferences are better defined with respect to Choice 1. More precisely, in Choice 1 (Table 2.6) and by also considering the attributes at their coded levels (Table 3.1), the positive sign of the estimated coefficient related to the "Coffee Type" indicates a preference for the 100% Arabica blend. This result is also confirmed when considering the estimated coefficients related to the Soft and Velvety taste: their positive sign indicates a preference for a high presence of the Soft and Velvety taste that is typical of the 100% Arabica blend. Moreover, for the Intense and Aromatic taste, its negative sign indicates a preference for a fair presence of the Intense and Aromatic taste; a result that is in line with the 100% Arabica blend chosen. However, in Choice 1, when considering the estimated coefficient of the "Caffeine" related to the HPLC results, its negative sign indicates a preference for the Arabica and Robusta blend. This controversial result is probably due to the respondents' "misinformation" about the two types of coffee investigated in Choice 1, by also considering the high heterogeneity effect related to the caffeine with a value of $\hat{\sigma} = 4.7288$.

Instead, when considering Choice 2, we can observe a relevant change in the consumers' preferences with respect to Choice 1. More specifically, the negative sign related to the "Coffee Type" indicates a preference for the Arabica and Robusta blend. Coherently with this

Table 2.7 Panel Mixed Logit Model results after tasting: Choice 2

	Estimate	*Std.Error*	*Z value*	*p-value*
$\hat{\mu}$ ***(main effect)***				
Coffee Type	-1.9681	0.6276	-3.14	0.002
Packaging	1.6923	0.5969	2.84	0.005
Intense Aromatic Taste	1.3233	0.4626	2.86	0.004
Soft Velvety Taste	-1.6290	0.4339	-3.75	0.000
Label Indication	0.7979	0.4179	1.91	0.056
Taste Score	-2.4714	0.7479	-2.79	0.005
Price 2 (level 2)	-3.2336	1.1098	-2.91	0.004
Price 3 (level 3)	-2.2330	1.0092	-2.21	0.027
$\hat{\sigma}$ ***(heterogeneity effect)***				
Coffee Type	2.7946	0.8318	3.36	0.001
Packaging	1.8176	0.5247	3.46	0.001
Intense Aromatic Taste	0.9589	0.6162	1.56	0.120
Soft Velvety Taste	0.2041	0.2659	0.77	0.443
Label Indication	0.5492	0.3077	1.78	0.074
Taste Score	-2.9991	0.9419	-3.18	0.001
Price 2 (level 2)	2.0336	1.0140	2.01	0.045
Price 3 (level 3)	-0.1162	0.4839	-0.24	0.810

result, the respondents prefer a high presence of Intense and Aromatic taste (typical for the Arabica and Robusta blend), and, in line with the coffee type preference, a fair presence of Soft and Velvety taste. Moreover, in Choice 2, the negative sign related to the "Taste Score", in accordance with the preference expressed for the "Coffee Type", indicates a preference for the Arabica and Robusta blend by also considering the lower heterogeneity in the consumers' preferences for this estimated coefficient with respect to the caffeine's heterogeneity in Choice 1. This result confirms the fact that the guided tasting, together with the information step, makes the consumers more capable of better differentiating between the two types of coffee and also of discriminating between different tastes, e.g. a Soft and Velvety taste and an Intense and Aromatic taste.

A change in the consumers' preferences also concerns the packaging: in Choice 1 the negative sign of the "Packaging" indicates a preference for the soft bag with a modified atmosphere. On the contrary, in Choice 2 consumers prefer the jar in a modified atmosphere. When considering the Label Indication, there is no change in the respondents' preferences between Choice 1 and Choice 2: on both occasions the consumers choose the certification of product sustainability with respect to the indication of geographical origin. Moreover, in Choice 2 (Table 2.7) the price coefficients are both negative and statistically significant

indicating that the consumers' willingness-to-pay decreases when the price increases (as can usually be expected). The same result is also confirmed for Choice 1 even though the intermediate level for the price coefficient (Price 2) in Choice 1 is positive.

Lastly, it must be noted that when considering the estimates related to the heterogeneity vector σ, it can be observed how most of them are statistically significant, and in general the heterogeneity is smaller for Choice 2 with respect to Choice 1. More precisely, in Choice 1, except for the intense and aromatic taste, all the estimated coefficients related to the heterogeneity effect are statistically significant indicating a large preference variation among individuals. Moreover, the largest effect of heterogeneity is related to the "Caffeine" coefficient, which also confirms the respondents' "misinformation" in Choice 1. Conversely, in Choice 2 the coefficient related to the "Taste Score" has a lower heterogeneity effect with respect to the caffeine's heterogeneity in Choice 1, confirming the fact that after the guided tasting session the consumers' preferences are better defined.

2.7 Conclusions

In this project we have proposed an innovative approach for the construction of optimal heterogeneous choice designs for correlated preferences. More precisely, our approach exploits an approximate design theory that allows us to i) verify that the final choice designs are really optimal, and ii) obtain heterogeneous choice designs consisting of groups of choice-sets to be administered to the respondents according to the optimal weight by also taking into account the fact that the responses faced by the same respondent are correlated. To achieve this, our proposal employs the Mixed Logit model that allows us to evaluate the respondents heterogeneity and to account for the correlation between the responses given by the same respondent. We have tested the validity of our innovative proposal under a compound design criterion through simulations on a simplified choice experiment, and through an application to a real case study. The estimation results for the P-MIXL models are very satisfactory and further confirm the validity of our innovative approach.

Despite the numerous benefits discussed above, our proposal is not without limitations. More precisely, our proposal takes explicitly care of the correlation among the respondent's responses only. We are fully aware that in the choice experiment context not only these responses could be correlated, but also the attributes themselves. This last issue could be addressed through the use of the Heteroscedastic Extreme Value model (HEV) (Bhat, 1995; Hensher, 1999). In fact, the HEV model assumes independent but not identical error terms, so as allowing to evaluate the heteroscedasticity across alternatives. Nevertheless, it does not take care neither of the heterogeneity across respondents nor of the correlation between the

respondent's responses. For this line of reasoning, our study explicitly addresses these last two issues in the construction of optimal choice designs, and an excellent further development of our proposal could be to also takes account of the correlation between the attributes.

Lastly, further analyses should be carried out to jointly investigate the results of the design efficiency, number of choice-sets by group, and optimal weights, also by considering the proposed methodology in terms of different design criteria. Moreover, a further development could be the inclusion in the modelling step of the socioeconomic characteristics collected through the initial questionnaire.

Chapter 3

Latin Hypercube Designs based on Strong Ortoghonal Arrays and Kriging Modelling to Improve the Payload Distribution of Trains

3.1 Introduction to the project

In this project we apply computer experiments and Kriging modelling to the railway field in order to improve the payload distribution of freight trains, a topic that is particularly relevant for Railway Undertakings. In this regard and by especially considering Europe, there are several codes that regulate international freight traffic, and which establish stringent limits on hauled mass of freight trains (UIC 421, 2012). According to the operational experience of the Railway Undertakings, it has been determined that depending on the freight train set specific arrangement, there can be significant differences in terms of in-train forces. According to a Code (UIC 421, 2012), Railway Undertakings have to statistically simulate freight train sets in order to prove the safety of a new family of trains (e.g. characterized by a new type of braking technology). Moreover, in Europe, there is a new research attempt towards longer and heavier trains with distributed power/braking. The assessment of such a type of trains requires suitable statistical methods in order to evaluate the effect of the payload distribution on the in-train forces. For instance, an optimal payload distribution is able to reduce the in-train forces (i.e. forces exchanged by consecutive vehicles), which limit the maximum hauled mass of "typical" freight trains. In this context, typical freight trains refer to trains with one or more locomotives in front of the train, and freight wagons equipped with block

brake devices. Undoubtedly, when the hauled mass of the train is increased, the efficiency of freight transportation is increased as well. The reason of high in-train forces is given by the intrinsic characteristics of the traditional *Union Internationale des Chemins de fer* (UIC) braking system (part of the equipment of most freight trains circulating in Europe) that does not allow a synchronous braking application.

An important issue that should be pointed out is that there is not an "unique" optimal payload distribution able to minimize the in-train forces for each specific train set arrangement. Rather, it is strictly connected and depends on the specific train set arrangement considered. To this end, the main aim of this project is to develop a general approach consisting in proper statistical methods allowing to assess which is the best payload distribution for the specific train composition. In order to address this issue, we propose a novel approach to improve the payload distribution through a suitable experimental design for the computer experiment, and Kriging modelling in order to assess which is the best payload distribution according to the specific train arrangement considered. One of the main advantages of our proposal is that it could be successfully applied to solve similar problems.

In the following Section, the engineering problems and the issues related to the braking performance of freight trains are described in details.

3.2 The engineering problem: braking performance issues on freight trains

Longitudinal Train Dynamics (LTD), that is the relative motion of adjacent railway vehicles running in track direction, has received great attention in the research literature, see for example Cole et al. (2017). This is mainly due to the fact that LTD is a key element for determining the safety of freight train sets. In fact, high in-train compressive forces can cause train derailments, while high in-train tensile forces can produce train disruption (because of draw gears failure), so as causing a freight traffic inefficiency. Various LTD simulators are employed by research centres, universities and Railway Undertakings in order to develop longer and heavier (still safe) freight train sets (Wu et al., 2018). Among these simulators, only few of them are capable to simulate simultaneously the air pneumatics and the mechanical behaviour of a train. TrainDy is one of such simulators and it has been used in this study to compute the LTD of the simulated freight train sets (Cantone, 2011). TrainDy is currently owned by the International Union of Railways (UIC) and it has been internationally certified against more than thirty experimental test campaigns.

The main source of in-train forces during braking is the non-synchronous activation of brake cylinders hence, of the braking force of consecutive wagons. This source of in-train

forces is inherent to the UIC braking scheme. Although there are some freight trains that use electro-pneumatic braking (a typical braking scheme of passengers trains), this UIC braking scheme equips the majority of freight trains in Europe. In this regard, both in-train compression and tensile forces need to be reduced, since very high compression forces can lead to train derailment, if the wagons are on a curve (of short radius) and/or if the height of consecutive buffers vary because of different payload. Moreover, high in-train tensile forces may cause "train disruption", that is the division of the train in two (or more) sections due to the breaking of the draw gears. Both circumstances should be strongly avoided in order to keep freight transport safe for people, infrastructure and goods. In order to weaken such source of in-train forces and avoid very high compression forces during a braking, several strategies could be adopted as follows:

1. applying an appropriate brake position, that is switching from "passenger train" brake position to "freight train" brake position: as different temporal application of consecutive wagons braking forces causes in-train forces, in-train compression forces are reduced by reducing the amplitude of the braking force during the initial braking phase. The disadvantage is the increase of the stopping distances, and this is feasible only up to a certain extent.

2. Reducing the hauled mass, so as the inertial effects are reduced as well as the in-train forces exchanged by consecutive vehicles. The main drawback is the inefficiency of freight transportation, since trains haul a "low" mass.

3. Placing more locomotives along the train in order to reduce the pneumatic length of the train; since more points spill air, the effect of a finite speed of sound in the air is mitigated. Recently, the European founded project Marathon has addressed this issue (Marathon, 2011).

4. Reducing the differences in maximum braking force of consecutive wagons: of course, if the time to reach the maximum braking force is different from wagon to wagon, and if such force is different, it is clear that in-train forces occur. Controlling the percentage of braked weight achieves such effect (UIC 544-1, 2014). This issue is indeed connected to the problem of train mass distribution that this project aims to address. It is worthwhile to remember that the wagon percentage of braked mass is the ratio between braked mass and wagon mass, where braked mass is a measure of wagon braking efficiency (UIC 544-1, 2014).

By considering the above strategies, and given that international regulations (UIC 544-1, 2014) do not restrict mass distribution, but only the overall hauled mass, one valid way to

increase freight train efficiency is to place the payload (and therefore the percentage of braked weight) along the train so as to reduce the in-train forces. In freight train transportation, trains can be divided in two families, according to specific needs: trains with fixed composition (e.g. shuttle trains) and trains with variable composition. The latter can also change composition along the way, since some wagons in the tail of the train can be removed or just unloaded. Since trains with variable composition cannot be generalized with few parameters, in this project the train mass distribution optimization for trains with fixed composition is considered: an overall fixed mass is assumed, but which allow to freely distribute the payload (though respecting the maximum load per axle), as in scrap material transport.

3.3 Latin Hypercube based on strong orthogonal arrays: theoretical issues

3.3.1 Latin Hypercube designs

As already stated (Chapter 1, Section 1.3), Latin Hypercube (LH) designs are the most commonly used class of space-filling designs for computer experiments. The seminal paper of McKay et al. (1979) introduced the LH sampling for selecting the input variables for computer experiments. More precisely, the LH sampling, as proposed by McKay et al. (1979), consists in dividing the domain of each input variable in n strata of equal marginal probability $(1/n)$, and in sampling once from each stratum. The authors also demonstrated that the LH sampling outperforms the random and the stratified sampling methods in terms of variance reduction of the corresponding estimates. Following the pioneering contribution of McKay et al. (1979), several other methods to build LH designs for computer experiments have been developed. In this project, we focus on LH designs built through strong orthogonal arrays (He and Tang, 2013). To this end, in what follows we define the LH design in general, as presented in Tang (1993) and in Lin and Tang (2015).

Let's define a LH with n runs and d factors as a matrix of dimension $n \times d$, in which each column j ($j = 1,...,d$) is a random permutation of n equally spaced levels. Let's indicate such a matrix by $\mathbf{A}$. Moreover, the n equally spaced levels are taken to be: $\{-\frac{(n-1)}{2}, -\frac{(n-3)}{2}, ..., \frac{(n-3)}{2}, \frac{(n-1)}{2}\}$, without loss of generality (Tang, 1993; Lin and Tang, 2015). Given the matrix $\mathbf{A}$, the LH design matrix $\mathbf{D}$ over the design space $[0,1)^d$ is obtained as follows:

$$\mathbf{D} = \frac{\mathbf{A} + \frac{(n-1)}{2}\mathbf{J} + \mathbf{U}}{n} \tag{3.1}$$

where **U** is an $n \times d$ matrix that contains the independent random numbers from the Uniform distribution $(0,1)$; **J** is an $n \times d$ matrix of all ones; therefore the LH design matrix **D** is of dimension $n \times d$. Furthermore, a LH design has exactly one point in every one of the n equally spaced intervals: $\{[0,\frac{1}{n}),[\frac{1}{n},\frac{2}{n}),...,[\frac{(n-1)}{n},1)\}$, the latter known as its one-dimensional uniform property. However, on one side given that usually there are several input variables, it is desirable to attain the uniform property in lower- and/or multi-dimensional projections; on the other hand, there is no guarantee that a random LH design will attain the maximum stratification in these projections. To this end, we focus on LH designs based on strong orthogonal arrays (SOA-based-LH) that achieves excellent low-dimensional space-filling properties.

3.3.2 LH designs based on strong orthogonal arrays: theoretical issues

As already stated (Chapter 1, Section 1.3), strong orthogonal arrays and the associated LH designs have been developed by He and Tang (2013). To define a Strong Orthogonal Array (SOA), we consider the basic theory on orthogonal arrays.

Definition of an Orthogonal Array (OA) (Tang, 1993; Hedayat et al., 1999):
An OA of *strength t* is an $n \times d$ matrix where the j-column has s_j levels $(1,2,...,s_j;\ j = 1,...,d)$, and it is such that for any $n \times t$ submatrix, each possible level combination occurs with the same frequency. If $s_1 = ... = s_j = ... = s_d = s$, then the OA is symmetric and it is denoted by:

$$\mathrm{OA}(n,d,s,t) \tag{3.2}$$

Definition of a SOA (He and Tang, 2013):
A SOA of *strength t* is an $n \times d$ matrix with entries $(1,...,s^t)$, such that any subarray of g columns, for any g with $1 \leq g \leq t$, can be collapsed into an OA $(n,g,s^{u_1} \times ... \times s^{u_g},g)$ for any positive integer $u_1,...,u_g$ with $u_1 + ... + u_g = t$. Such an array is denoted by:

$$\mathrm{SOA}(n,d,s^t,t) \tag{3.3}$$

Thus, differently from an OA in which each factor has s levels (formula (3.2)), in a SOA each factor has s^t levels (formula (3.3)). It is exactly this major number of factor levels of a SOA with respect to an OA that allows to obtain LH designs with very good space-filling properties. More precisely, a SOA-based-LH design with *strength t* achieves the same uniformity as an OA-based-LH design in all t-dimensional projections. Moreover, in all g-dimensional projections, where $2 \leq g \leq t$, a SOA-based-LH design is more space-filling than an OA-based-LH design.

The construction of a SOA-based-LH design could be briefly summarized in the following four main steps:

1. begin with an OA with n rows, d columns, s levels for each column and *strength* t;

2. from the OA obtained at the previous step, construct the SOA with n rows, d columns, s^t levels and *strength* t, by expanding the s levels of the OA into s^t levels. Once the SOA is obtained, the SOA-based-LH could be generated. To this end, let λ be the index of the SOA equal to:
$$\lambda = \frac{n}{s^t} \tag{3.4}$$

3. for each column of the SOA, replace the λ entries for the level c ($c = 1, ..., s^t$) by any permutation of $c\lambda, c\lambda + 1, ..., (c+1)\lambda - 1$. Denote such a matrix by $\mathbf{A^*}$ and obtain the matrix $\mathbf{A}$ as follows:
$$\mathbf{A} = \mathbf{A^*} - \frac{(n+1)}{2}\mathbf{J} \tag{3.5}$$

4. generate the associate SOA-based-LH design through formula (3.1), which design matrix $\mathbf{D}$ is of dimension $n \times d$.

The SOA-based-LH design promises a better stratification in low-dimensional projections with respect to an OA-based-LH design. Nevertheless, two issues should be noted. The first one relates to the fact that the existence of a SOA-based-LH design depends on the existence of the corresponding OA. This is also true when considering the OA-based-LH designs. This issue seems to be a minor drawback, given that it has been largely addressed by considering the general developments achieved in the theory for orthogonal arrays (Hedayat, 1999; Mukerjee and Wu, 2006).

The second issue refers to the fact that, while the existence of a SOA (n, d, s^t, t) implies the existence of an OA (n, d, s, t), the converse is not necessarely true. To this end, He and Tang (2013) established how a SOA could be built from an OA, through the use of a Generalized Orthogonal Array (GOA) (Lawrence, 1996). Following, we give the definition of such a type of array.

Definition of a GOA (He and Tang, 2013):
A matrix $\mathbf{B}$ of dimension $n \times (td)$, in which the td columns $\mathbf{b_{ij'}}$ are arranged into d groups of t columns each (e.g. $\mathbf{B} = (\mathbf{B_1}, ..., \mathbf{B_d})$ where $\mathbf{B_i} = (\mathbf{b_{i1}}, ..., \mathbf{b_{it}})$), is called a GOA of size n, d constraints, s levels and *strength* t, if the matrix $\mathbf{B}^*$ consisting of t columns, ($i = i_1, ..., i_g$; $j' = 1, ..., u_i$) is an OA of strength t for any $1 \leq g \leq t$, any $1 \leq i_1 < ... < i_g$, and any positive integer $u_1, ..., u_g$ with $u_1 + ... + u_g = t$. Such an array is denoted by:
$$\text{GOA}(n, d, s, t) \tag{3.6}$$

Through the use of GOAs (formula (3.6)), He and Tang (2013) established an excellent method to obtain a SOA from the corresponding OA. To do so, they make use of the fundamental results established by Lawrence (1996), according to which if an OA(n,d,s,t) exists, then a GOA(n,d',s,t) can be built following the rules:

$$d' = \begin{cases} \frac{d}{e} & \text{if } t = 2e \text{ is even;} \\ \frac{(d-1)}{e} & \text{if } t = 2e+1 \text{ is odd.} \end{cases} \tag{3.7}$$

By considering this fundamental connection between OAs and GOAs, He and Tang (2013) demonstrated the fundamental results according to which, once a GOA is obtained from an OA, the corresponding SOA can be built as follows.

SOA construction from a GOA (He and Tang, 2013):
Let $\mathbf{B} = [(\mathbf{b_{11}},...,\mathbf{b_{1t}});...;(\mathbf{b_{d1}},...,\mathbf{b_{dt}})]$ be a GOA(n,d,s,t). Then, the matrix $\mathbf{E} = (\mathbf{e_1},...,\mathbf{e_j},...,\mathbf{e_d})$ is a SOA(n,d,s^t,t) , where each column $\mathbf{e_j}$ $(j = 1,...,d)$ is obtained as follows:

$$\mathbf{e_j} = \sum_{j'=1}^{t} \mathbf{b_{ij'}} s^{t-j'} \quad \forall j = 1,...,d \tag{3.8}$$

Therefore, starting from formulas (3.7) and (3.8), it follows the following result (Theorem no.1 in He and Tang, 2013).

Given an OA (n,d,s,t), then a SOA (n,d',s^t,t) could be built following the rules:

$$d' = \begin{cases} \frac{d}{e} & \text{if } t = 2e \text{ is even;} \\ \frac{(d-1)}{e} & \text{if } t = 2e+1 \text{ is odd.} \end{cases} \tag{3.9}$$

According to formula (3.9), for a *strength* greater than or equal to three, one or more columns of the OA will be lost when obtaining the corresponding SOA. Moreover, in order to build the SOA through formula (3.8), firstly corresponding GOA should be built. To this end, He and Tang (2013) described the construction of a GOA for different array *strengths*. In what follows, we describe in details the construction of a GOA only for a *strength* equal to three, that is used to build the SOA for the case-study here reported.

Let $\mathbf{O} = (\mathbf{o_1},...,\mathbf{o_d})$ be an OA$(n,d,s,3)$, where $(\mathbf{o_1},...,\mathbf{o_d})$ refer to the columns of the array. Therefore, a GOA(n,d',s,t) exists with $d' = (d-1)$ (formula (3.7)). The GOA matrix $\mathbf{B}$ of dimension $(n \times d't)$, is expressed as follows:

$$\mathbf{B} = [(\mathbf{b_{11}},\mathbf{b_{12}},\mathbf{b_{13}}),...,(\mathbf{b_{d'1}},\mathbf{b_{d'2}},\mathbf{b_{d'3}})], \tag{3.10}$$

where each column $\mathbf{b_{ij'}}$ of the matrix $\mathbf{B}$ is obtained as follows:

$$(\mathbf{b_{11}},...,\mathbf{b_{d'1}}) = (\mathbf{o_1},...,\mathbf{o_{d'}}) \tag{3.11}$$

$$(\mathbf{b_{12}},...,\mathbf{b_{d'2}}) = (\mathbf{o_d},...,\mathbf{o_d}) \tag{3.12}$$

$$(\mathbf{b_{13}},...,\mathbf{b_{d'3}}) = (\mathbf{o_2},...,\mathbf{o_{d'}},\mathbf{o_1}) \tag{3.13}$$

Once the matrix $\mathbf{B}$ of the GOA is obtained, the SOA(n, d', s^t, t) is built through formula (3.8). More precisely, for $t = 3$, each column $\mathbf{e_j}$ (matrix $\mathbf{E}$ of the SOA) is obtained as follows:

$$\begin{cases} \mathbf{e_1} = 4\mathbf{a_1} + 2\mathbf{a_d} + \mathbf{a_2} \\ \mathbf{e_2} = 4\mathbf{a_2} + 2\mathbf{a_d} + \mathbf{a_3} \\ \vdots \\ \mathbf{e_{d'-1}} = 4\mathbf{a_{d'-1}} + 2\mathbf{a_d} + \mathbf{a_{d'}} \\ \mathbf{e_{d'}} = 4\mathbf{a_{d'}} + 2\mathbf{a_d} + \mathbf{a_1} \end{cases} \tag{3.14}$$

Therefore, through this construction of a SOA from an OA *via* a GOA, the s levels of the OA are expanded in s^t levels of the SOA. This allows to obtain a SOA-based-LH design with better space-filling properties with respect to an OA-based-LH design.

3.4 LH designs planned for the railway field

3.4.1 Trains Assembling

In this project we consider a freight train that transports scrap material. In order to find the payload distribution able to reduce the in-train forces, we focus on the distribution of braked mass within the train (Subsection 3.2). Regarding the application here reported and differently from Arcidiacono et al. (2017), we consider that the train is unloaded in five different sections, representing five different destinations of the transported goods; within each section the distribution of percentage of braked mass is freely distributed. Therefore, we define the distribution of percentage of braked mass (e.g. ratio between braked mass and wagon mass) within each train section, and not considering the whole train as Arcidiacono et al. (2017). This means that we assume that each train section has its payload and its percentage of braked mass. Consequently, the wagon position within a given train section could change in order to have a desired distribution of braked mass percentage. This strategy turns out to be particularly useful to further improve the payload distribution. More precisely,

it allows to better define which is the best payload distribution for the specific train set arrangement by considering that the whole train could not be limited to only one payload. Rather, different payload distributions can be found as better ones at the beginning, in the middle or at the end of the train.

Without any loss of generality, in the rest of this Subsection we do not consider a suffix indicating each train section. In order to represent the percentage of braked mass distribution within each train section, we use the following two variables as defined in Arcidiacono et al. (2017):

- B denotes the length of the train section (in terms of number of wagons);
- Q denotes the area of each distribution, corresponding to the sum of percentages of braked mass of the wagons.

In Figure 3.1 we report the different geometries in terms of the variables B and Q (Arcidiacono et al., 2017). According to these two variables, within each train section we have a distribution that could be:

1. uniform;
2. trinagular;
3. trapezoidal.

Furthermore, in order to univocally represent the percentage of braked mass distribution for each train section, two input variables, h and x are defined. The input variable h defines the shape of the percentage of braked mass distribution for a given train section. This shape can gradually change from uniform to triangular one. The input variable x identifies the position of the maximum load along the length of the section. In Figure 3.2 we report the different shapes, according to the different values of the variables x and h for a given train section. For example, whatever is the value of x, if h is equal to 0, the shape of percentage of braked mass distribution is uniform. Conversely, if h is equal to 1, this shape is triangular, with position of the vertex defined by the variable x. For other values of h, the shape is a trapeze, with minor basis becoming smaller as h increases from 0 to 1.

3.4.2 Building of the SOA-based-LH design for our case-study

As noted in the previous Subsection, the train was divided in five different sections, where each section has its own payload distribution represented by its shape and the position of maximum load. Therefore, we have a total of ten experimental factors: i) five factors related

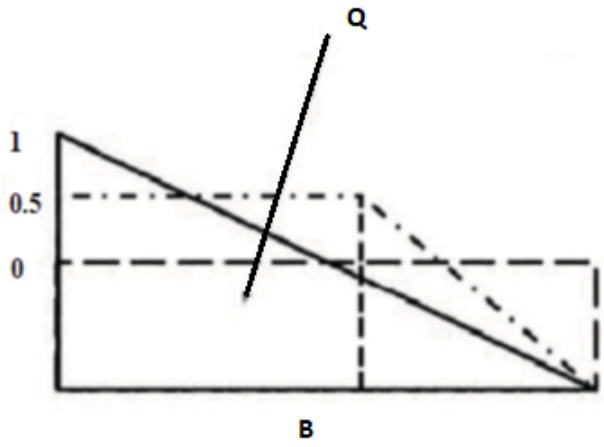

Fig. 3.1 Different geometries with same area Q and base B for a given train section (source: Arcidiacono et al., 2017)

.

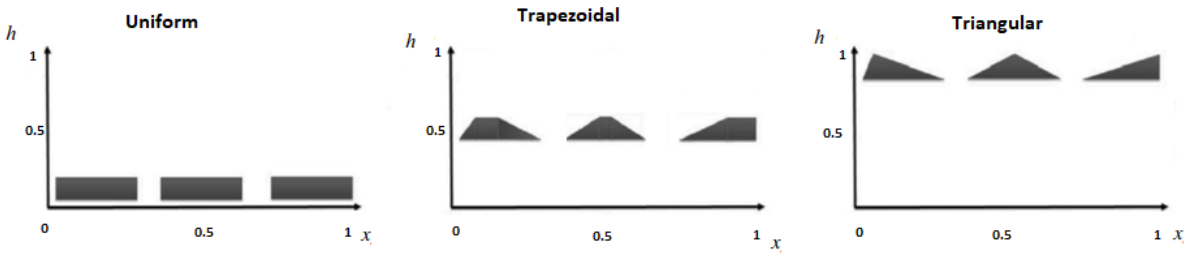

Fig. 3.2 Different payload distributions for a given train section according to the values of the variables h and x.

to the shape of the payload distribution, e.g. $h = \{h_1, h_2, h_3, h_4, h_5\}$; and ii) five factors related to the position of maximum load, e.g. $x = \{x_1, x_2, x_3, x_4, x_5\}$. In order to build the design, we have to construct a SOA with ten columns, and to choose *a priori* the *strength* of the array. In this case, we only consider a SOA with $t = 3$. This choice is performed because, by this way, we are able to obtain:

- a design that attains better space-filling properties with respect to those based on an OA with $t = 3$; otherwise, if $t = 2$, this is not the case;
- a design with a relatively low number of experimental runs; otherwise, in general, if $t > 3$, the design will contains more experimental runs.

As pointed out in Subsection 3.3.2, a SOA with $d = 10$ and $t = 3$ could be constructed starting from an OA loosing one column (formula (3.9)). Therefore, we begin with a symmetric OA with $n = 64$ runs, and eleven columns, where each column has two levels and *strength* three:

$$\mathrm{OA}(64, 11, 2, 3) \tag{3.15}$$

In order to choose the OA described in formula (3.15), we refer to the fundamental property according to which a regular fractional factorial design of Resolution R is an orthogonal array of *strength* $t = \mathrm{R} - 1$ (Hedayat et al., 1999; Dey and Mukerjee, 1999). Therefore, for obtaining the array in formula (3.15), we have used the catalog of 2^{11-5}_{IV} fractional factorial designs, provided in the **R** package *FrF2* (Grömping, 2014).

Subsequently, through the $\mathrm{OA}(64, 11, 2, 3)$ reported in formula (3.15), the following SOA is constructed:

$$\mathrm{SOA}(64, 10, 2^3, 3) \tag{3.16}$$

Thus, by considering formula (3.9), one column of the OA (formula (3.15)) is lost when the SOA is built (formula (3.16)). Moreover, the two levels of the OA are expanded in eight levels of the SOA through the following steps:

a) the construction of a $\mathrm{GOA}(64, 10, 2, 3)$ from an $\mathrm{OA}(64, 11, 2, 3)$ (formulas (3.11), (3.12) and (3.13));

b) the construction of the $\mathrm{SOA}(64, 10, 2^3, 3)$ from the $\mathrm{GOA}(64, 10, 2, 3)$ (formulas (3.8) and (3.14)).

Finally, the SOA-based-LH design is obtained through the last two steps described in Subsection 3.3.2, where the array index is equal to $\lambda = 8$. The space-filling properties of the design are as follows: the LH design based on the SOA in formula (3.16) achieves a

stratification on a $2 \times 2 \times 2$ grid in any three-dimensional projection, and, in addition, the stratifications on the finer grids of $2^2 \times 2$ and 2×2^2 in any two-dimensional projection.

By considering our case-study, each entry of the generated SOA-based-LH design should be converted in one and only one of the following five values:

$$\{0; 0.25; 0.50; 0.75; 1\}$$

These five values, equally spaced in the interval [0,1], turn out to be the most suitable choice for the train assembling necessary for the subsequent train simulations. In order to do so, we adopt the following strategy:

1. firstly, the interval $[0,1]$ is divided in five subintervals of equal dimension as follows:

$$\{[0, 0.20), [0.20, 0.40), [0.40, 0.60), [0.60, 0.80), [0.80, 1]\}$$

2. let a_{ij} be the entry for the i-th row and the j-th column of the SOA-based-LH design. Therefore:

$$\text{if } a_{ij} \in [0.20, 0.40) \text{ then } a_{ij} = 0.25 \ \forall i = 1, ..., n \text{ and } \forall j = 1, ..., d;$$

$$\text{if } a_{ij} \in [0.40, 0.60) \text{ then } a_{ij} = 0.50 \ \forall i = 1, ..., n \text{ and } \forall j = 1, ..., d;$$

$$\text{if } a_{ij} \in [0.60, 0.80) \text{ then } a_{ij} = 0.75 \ \forall i = 1, ..., n \text{ and } \forall j = 1, ..., d;$$

Regarding the extreme values, we assign zero to the values lower than 0.2 and one to the values greater than 0.8 because of engineering issues. This strategy is particularly useful for our case-study, given that it allows us to obtain a perfect level of balance for all the factor levels of the SOA-based-LH design when converting its entries in the original values.

3.5 Kriging: outlined theory and specific issues

3.5.1 The Kriging method

As already stated (Chapter 1, Section 1.3), the starting point for the Kriging methodology dates back to the geosciences (Krige, 1951; Matheron, 1971). Subsequently, the seminal paper of Sacks et al. (1989) introduced the Kriging methodology as a valid metamodel to approximate the deterministic relation between input and output variables, simulated through computer experiments. Since then, the Kriging has been widely used and developed as a valid statistical interpolator with a high prediction accuracy.

Let us consider a set Ξ of n experimental points $x_i \in \boldsymbol{X}$ over a d-dimensional design space χ_d, and the corresponding response values (output) y_i, $i = (1,...,n)$, e.g.: $\Xi = (x_1,y_1),....,(x_i,y_i),....,(x_n,y_n)$. Therefore, the set of trials $\boldsymbol{X}$ is selected within the d-dimensional experimental region χ_d and each y_i is the realization of a random dependent variable $Y(\boldsymbol{X})$ estimated through a physical experimental design.

The Kriging method is subsequently applied to explore the experimental region χ_d in order to predict new simulated observations on the basis of the information gained through Ξ. The final aim is an optimal prediction of Y through a statistical model involving a deterministic part, also named trend function, and a stochastic part, the latter replacing the error component for a standard statistical model.

The initial expression for Kriging is named Simple or Ordinary Kriging, and it is formulated as in the following, for a single point $\boldsymbol{x}$:

$$y_{\boldsymbol{x}} = \mu + Z(\boldsymbol{x}) \tag{3.17}$$

In formula (3.17), the deterministic part μ is supposed to be a non-random constant, so that, in this case the trend does not vary along time or space; the second term, $Z(\boldsymbol{x})$, identifies a spatial stochastic process. In simple Kriging, $Z(\boldsymbol{x})$ is usually assumed as a second order (weakly) stationary process. Therefore $Z(\boldsymbol{x})$ reduces to the covariance between any two points, $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{h}$:

$$Z(\boldsymbol{x}) = Cov(Z(\boldsymbol{x}),Z(\boldsymbol{x}+\boldsymbol{h})) = \sigma_y^2 R(\boldsymbol{h};\boldsymbol{\omega}) \tag{3.18}$$

where $\boldsymbol{h}$ is the distance between the two points belonging to χ_d, σ_y^2 is the process variance, e.g. the variance of Y, and $\boldsymbol{\omega}$ is the vector of parameters defining the stationary stochastic process for the correlation function R. It must be noted that, in simple Kriging, given conditions on $Z(\boldsymbol{x})$, the stochastic process reduces to a Gaussian random field, in which the predicted values for Y are conditional on the experimental set Ξ. Furthermore, the process variance is independent with respect to Y which implies the homoscesdasticity assumption.

The main difference between Simple and Universal Kriging is related to the expression for the deterministic part of formula (3.17). Universal Kriging assumes a non-constant trend, thus formula (3.17) becomes:

$$\begin{aligned} y_{\boldsymbol{x}} &= \mu(\boldsymbol{x}) + Z(\boldsymbol{x}) \\ \mu(\boldsymbol{x}) &= \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta} \end{aligned} \tag{3.19}$$

where $\boldsymbol{f}'(\boldsymbol{x}) = (f'_1(\boldsymbol{x}), ..., f'_m(\boldsymbol{x}))$ is a set of trend functions defined for each new point $\boldsymbol{x}$. By considering n selected points, $\boldsymbol{F}$ is defined as the model regression matrix of dimension $[n \times m]$ formed by the n independent functions $f(\boldsymbol{x})$:

$$\boldsymbol{F} = (f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n)) \tag{3.20}$$

The vector $\boldsymbol{\beta}$ is the column vector $[m \times 1]$ of unknown coefficients. In this situation, each point $\boldsymbol{x}$ is a new trial simulated by the experimenter: $\boldsymbol{x} \in \chi_d$.

Furthermore, by assuming

$$E(Z(\boldsymbol{x})) = 0$$

and

$$Cov(Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})) = Cov(\boldsymbol{x} - (\boldsymbol{x}+\boldsymbol{h})) \tag{3.21}$$

implicitly we assume that, by performing simulations, we move along a path depending only on the allocation of vector $\boldsymbol{h}$, e.g. the vector of the differences between any two points belonging to χ_d. In fact, formula (3.21) is equal to σ_y^2 for $\boldsymbol{h} = \boldsymbol{0}$, therefore an equivalent expression is as follows:

$$\sigma_y^2 R(\boldsymbol{x} - (\boldsymbol{x}+\boldsymbol{h})) \tag{3.22}$$

highlighting that the path depends only on $\boldsymbol{h}$.

3.5.2 The Matérn covariance function

A fundamental issue to consider when dealing with Kriging modelling is the choice of the covariance kernel $K(.)$ that has to be symmetric and positive semi-definite. In general, the mostly applied covariance kernels are the Matérn, the Exponential, the Power-Exponential and the Gaussian ones. For the case study illustrated in this project, the Matérn covariance function is applied (Rasmussen and Williams, 2006).

The Matérn class of covariance functions $K_M(r)$ is expressed as follows:

$$K_M(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{\sqrt{2}\nu r}{l} \Big)^{\nu} K_\nu \Big(\frac{\sqrt{2}\nu r}{l} \Big) \tag{3.23}$$

where l and ν are positive parameters and K_ν is the Bessel function. The l parameter is the scale coefficient while the ν parameter controls the smoothness of the process. For this function, the process is k-times mean square differentiable if and only if $\nu > k$. Therefore, a subset of functions for this class is specified when $\nu = p + (1/2)$ and $p \in I^+$. In this case, the covariance structure reduces to a product of a polynomial of order p and an exponential

function, and the general expression becomes:

$$K_{\nu=p+(1/2)}(r) = \exp\Big(\frac{\sqrt{2}\nu r}{l}\Big)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\Big(\frac{\sqrt{8}\nu r}{l}\Big)^{p-i} \tag{3.24}$$

We can evaluate formula (3.24) by choosing $p = 1$ and $\nu = 3/2$; therefore, we obtain:

$$K_{\nu=1+(1/2)}(r) = \Big(1+\frac{\sqrt{3}r}{l}\Big)\exp\Big(-\frac{\sqrt{3}r}{l}\Big) \tag{3.25}$$

Furthermore, the scale coefficient l is replaced by the vector ϕ of scale parameters (strictly positive), and r is replaced by the module $\|h_j\|$, $j = 1,...,d$. Therefore, we may define the specific applied Matérn function, with $\nu = 3/2$ and the vector ω composed by (ϕ, ν), as in the following formula:

$$R(h;\omega) = \prod_{j=1}^{d}\Big(1+\frac{\sqrt{3}\|h_j\|}{\phi_j}\Big)\exp\Big(-\frac{\sqrt{3}\|h_j\|}{\phi_j}\Big); j = 1,...,d \tag{3.26}$$

In formula (3.26) if all ϕ_j are assumed to be equal, e.g. $\phi_1 = \phi_2 = ... = \phi_j = ... = \phi_d = \phi$, then the kernel process is isotropic; otherwise, as assumed in our case study, the kernel process is anisotropic (fomula (3.26)). The characteristic length scale coefficients in the vector ϕ regulates the speed of decay of the correlation between any two points in the corresponding dimension (Rasmussen and Williams, 2006). That is, they determines the distance over which the response is expected to vary significantly for the corresponding dimension. A very large length scale ϕ_j indicates that the response is expected to be essentially a constant function of the corresponding input variable j. Therefore, the estimated values of these coefficients are useful to also determine the relevance of each input variable: the smaller the value ϕ_j, the more the input x_j is relevant for the process under study, and viceversa (MacKay, 1998; Rasmussen and Williams, 2006).

The Matérn function with $\nu = 3/2$ is differentiable of order one; if $\nu = 1/2$ the Matérn function reduces to the exponential covariance function. It must be noted that function (3.26) corresponds to a Gaussian process with a different level of differentiability; therefore the continuity is always guaranteed, which is relevant to prevent either discontinuities or numerical instability during simulations, such as jumps in the response or problems in the computation of the inverse of the covariance matrix.

Nevertheless, computational problems and jumps may be encountered during simulations and, to this end, a nugget parameter has been introduced. The inclusion of this coefficient allows to avoid instability, e.g. it may be viewed as a noise added to the process variance. A fundamental feature is that Kriging still preserves its interpolating property although with the

inclusion of the nugget term. Moreover, the estimation of a nugget parameter allows to also account for potentially possible deviations of inaccurate assumptions on the stationarity of the process and on the chosen correlation function (Stein, 1999; Gramacy and Lee, 2012).

3.6 Kriging modelling results

As previously stated, the main aim of this project is to develop a general approach for improving the payload distribution of freight trains. Therefore, we can consider that the optimal solution searched through the Kriging models is the best payload distribution along the train that guarantees an emergency braking with the minimum in-train forces among wagons, both in compression and in tensile. To this end, the output variables are the three aforementioned in-train forces:

i) compression forces computed at 2m;

ii) compression forces computed at 10m;

iii) tensile forces computed at 2m.

By considering the SOA-based-LH design built with $n = 64$ runs (Subsection 3.4.2), the true values of compression and tensile forces for the specific train set arrangement here considered are calculated through the TrainDy software (Cantone, 2011), internationally certified for the computation of in-train forces of freight trains. It must be also noted that the three response variables have a different measurement scale. For this reason, in order to make the results comparable, we properly standardize them through the min-max standardization method (OECD, 2008).

For each response variable, a Kriging model is estimated through the **R** package *DiceKriging* (Roustant et al., 2012). More precisely, the Universal Kriging is applied, formula (3.19), by assuming a non constant unknown trend function. The Kriging models are estimated by considering the anisotropic Matérn covariance function with $\nu = 3/2$, formula (3.26). The Matérn class of covariance functions has very good properties. More precisely, it is highly flexibile in representing the smoothness and the differentiability of the process according to the choice of the value for the parameter ν (Stein, 1999). In this regard, larger is the chosen value for ν, smoother is the corresponding process. For $\nu = 3/2$ the Matérn covariance function is once differentiable, while for $\nu = 5/2$ it is twice differentiable. Moreover, the Matérn covariance function includes the Exponential and the Gaussian ones as a special case when $\nu = 1/2$ and $\nu = \infty$ respectively. Stein (1999) investigated in details the Matérn properties and behaviour recommending it for the use in many practical situations. By considering these

issues, this type of covariance function has been chosen for the process under study, allowing us to overcome the strong assumptions of smoothness and differentiability of other (possible) covariance functions, like the Power-Exponential and the Gaussian ones. Moreover, in the absense of explicit prior knowledge about the existence of second or higher order derivatives of the process under study, the parameter $\nu = 3/2$ is chosen, assuming that the process is differentiable only of order one. Lastly, a nugget term δ is also estimated.

When considering the specification of the trend function, it must be noted that, at the beginning, we estimated both sets of models:

1) the Kriging models with a first order polynomial trend;

2) the Kriging models with a full second order polynomial trend function.

For the first group of models, the fitting was not satisfactory, while a perfect fitting has been obtained for the second group of models. Nevertheless, from an engineering point of view, some of the quadratic and first interaction terms in the full second order polynomial trend function do not make sense for the process under study. This is the case for the interactions between variables of different train sections, and for the quadratic effects related to the shape of the percentage of braked mass distribution. Therefore, the final trend functions include the first order polynomial trend, plus the quadratic effects related to the position of maximum load, and the first interaction terms on the two variables, h and x, strictly related to the same train section.

For each response variable, the estimated trend coefficients are reported in Table 3.1. Except for the intercept term, the estimated trend coefficients related to the tensile forces at 2m are smaller in magnitude with respect to those related to the compression forces at 2m and 10m. The estimated coefficients related to the characteristic length scales, the process variance and the nugget are reported in Table 3.2. The largest process variance, ($\hat{\sigma}_y^2$), is obtained for the compression forces at 2m, while the smallest one is related to the tensile forces at 2m. For each response, the nugget estimates are very low and close to zero.

The estimated length scale coefficients, (Table 3.2), give us some insights on the relevance of each input variables for each in-train force. More precisely, for the compression forces computed at 2m, the most relevant input variables are the shape of the percentage of braked mass distribution at the beginning of the train (e.g. first train section), and the position of maximum load at the second and last train sections. When considering the compression forces at 10m, the shape of the percentage of braked mass distribution at the beginning, in the middle and at the end of the train turns out to be most relevant; instead, the most influent position of maximum load appears to be in the middle of the train. Lastly, for the tensile forces at 2m, the most important results are the position of maximum load at the second train

Table 3.1 Estimated trend coefficients for the three Universal Kriging models

Coefficient	Estimates for 2m Compression forces	Estimates for 10m Compression forces	Estimates for 2m Tensile forces
β_0	-0.9016	-0.9991	0.9886
β_{h_1}	-0.0198	-0.0078	0.0008
β_{x_1}	-0.2108	0.1455	0.0014
β_{h_2}	0.0274	0.0374	-0.0006
β_{x_2}	0.0506	0.1148	-0.0004
β_{h_3}	0.0993	0.0072	-0.0013
β_{x_3}	0.3159	-0.0208	-0.0052
β_{h_4}	0.0238	0.0088	0.0003
β_{x_4}	-0.1315	-0.0083	0.0047
β_{h_5}	0.0171	-0.0049	0.0002
β_{x_5}	0.0069	-0.0714	-0.0045
$\beta_{x_1^2}$	0.1935	-0.1446	-0.0011
$\beta_{x_2^2}$	-0.0385	-0.0888	0.0002
$\beta_{x_3^2}$	-0.2194	0.0297	0.0025
$\beta_{x_4^2}$	0.0844	0.0045	-0.0036
$\beta_{x_5^2}$	0.0576	0.0287	0.0038
$\beta_{x_1h_1}$	-0.0603	0.1425	0.0006
$\beta_{x_2h_2}$	-0.1827	-0.0140	0.0114
$\beta_{x_3h_3}$	-0.1478	-0.0231	0.0023
$\beta_{x_4h_4}$	0.0351	-0.0362	-0.0025
$\beta_{x_5h_5}$	-0.0541	0.0224	0.0078

Table 3.2 Estimated covariance and nugget coefficients for the three Universal Kriging models

Coefficient	Estimates for 2m Compression forces	Estimates for 10m Compression forces	Estimates for 2m Tensile forces
$\phi_1(h_1)$	0.1333	0.3516	1.9612
$\phi_2(x_1)$	1.9651	1.9651	1.9651
$\phi_3(h_2)$	1.9682	1.9682	0.0684
$\phi_4(x_2)$	0.6190	1.9811	0.3902
$\phi_5(h_3)$	1.9850	1.4348	0.4033
$\phi_6(x_3)$	1.9758	0.1511	1.9758
$\phi_7(h_4)$	1.9574	0.5810	1.9574
$\phi_8(x_4)$	1.3766	1.9851	1.9851
$\phi_9(h_5)$	1.9762	0.5926	1.9762
$\phi_{10}(x_5)$	0.2058	1.9761	1.9761
σ_y^2	$7.5e^{-4}$	$6.2e^{-4}$	$1.5e^{-14}$
δ	$7.5e^{-12}$	$6.2e^{-12}$	$1.5e^{-14}$

section, and the shape of the percentage of braked mass distribution at the second and third train sections. These results show that it is not easy to find a distribution of percentage of braked mass that optimize all types of forces simultaneously. Nevertheless, they confirm the operational experience with respect to 10m compressive forces and 2m tensile forces.

For each response variable, we report the Kriging surfaces for each train section (Figures 3.3, 3.4 and 3.5). Nevertheless, some of these Kriging surfaces cannot be easily interpreted. To this end, in order to better highlight the behavioir for each estimated Kriging model, we also provide the plots of the estimated Kriging surfaces for each input variable (Figures 3.6, 3.7 and 3.8). At the beginning of the train (first and second train sections), the best payload distribution which minimizes the 2m compression forces, gradually changes from a trapezoidal (with a position of maximum percentage of braked weight in the middle of the section), to a triangular one; where the maximum is carried out at the end of the train section (Figures 3.6 and 3.3). Nevertheless, when considering the 10m compression forces, the best percentage of braked mass distribution at the beginning of the train seems to be the uniform one (Figure 3.7); the same result is obtained when considering the tensile forces at 2m for the second train section (Figure 3.8). This result seems to be in contraddiction with what has been found before, and this depend on the non linearity of the considered problem; in fact, in-train forces depend not only on the distribution of percentage of braked mass, but also on the braked region, and on mechanical characteristics of buffers and draw gears. In what follows we discussed this further issue. When considering the payload distribution in the middle of the train for the three in-train forces, no clear pattern for the best payload distribution is highlighted.

At the end of the train, a pattern for the fourth train section is outlined for both the compression forces at 2m and 10m: for these two in-train forces the triangular distribution is the best one, with the position of maximum load at the beginning and at the end of the train section (Figures 3.7, 3.8 and 3.4). Lastly, for the fifth train section, the uniform payload distribution is the best one, by minimizing the 2m tensile forces (Figure 3.8).

The validation of the Kriging models is performed through the leave-one-out cross validation method. To this end, in Table 3.3 we illustrate three goodness-of-fit measures: the Q^2 predictivity coefficient, the standard error of the leave-one-out residuals (labelled "SE-LOO") and the root mean square error of the leave-one-out residuals (labelled "RMSE-LOO"). Q^2 values and SE-LOO values related to the predictivity coefficient and the standard errors of the leave-one-out residuals are close to one; this result indicates that the model prediction is perfectly accurate. As can be seen in Table 3.3, the values of the goodness-of-fit measures for the estimated Kriging models are very satisfactory. More precisely, the best fitting according to the Q^2 predictivity coefficient is obtained for the tensile forces at 2m.

Table 3.3 Goodness-of-fit measures for each estimated Kriging model

	Compression forces at 2m	Compression forces at 10m	Tensile forces at 2m
$Q2$	0.8408	0.8189	0.9018
SE-LOO	0.8080	0.8032	0.7706
RMSE-LOO	0.0207	0.0192	0.0009

The same result is confirmed when considering the RMSE-LOO and the SE-LOO measures, even though the standard errors of the leave-one-out residuals are slightly larger for the 2m compression forces.

A graphical tool to analyse the goodness-of-fit of the models is provided by the **R** package *DiceKriging*. To this end, for each response variable three plots are provided: the residual analysis, the standardized variance of the residuals and the quantile-quantile (QQ) plot (Figures 3.9, 3.10 and 3.11). Overall, the goodness-of-fit is very satisfactory for all the three Kriging models. The residual plots seem to suggest that the Kriging model with the tensile forces at 2m is the best one. The QQ plot results do not indicate any visible deviation from the gaussinity.

Lastly, it must be noted that we have also investigated the Kriging models with other types of covariance functions, e.g. the Exponential, the Power Exponential and the Gaussian ones. The reason for these further investigations is that in general the Kriging is strongly affected by the chosen covariance function. For the sake of brevity we do not report these results. We only point out that after considering this issue, the Matérn covariance function seems to be the most appropriate for the specifical case study here analyzed. According to these results, this type of covariance function outperforms the other three ones by especially considering the results related to the leave-one-out cross validation method (residual analysis and QQ plots). More precisely, deviations from normality are obtained for the Exponential, the Gaussian and the Power Exponential covariance functions for the compression forces at 2m and at 10m. For the tensile forces at 2m, the deviations from normality of the Gaussian covariance function are even more pronounced. The results related to the leave-one-out residuals confirm the goodness-of-fit of the Matérn for all the estimated Kriging models.

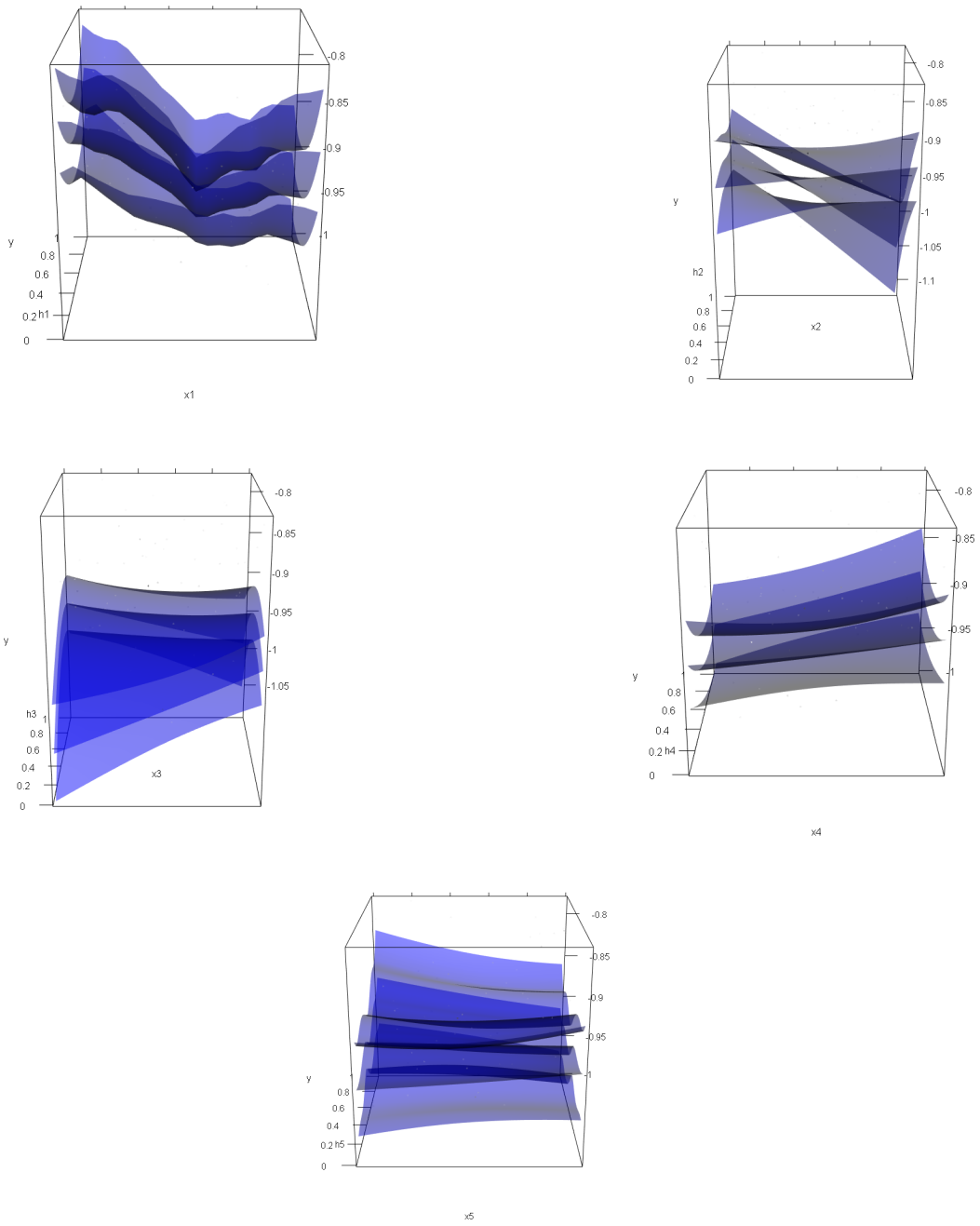

Fig. 3.3 Compression forces at 2m: Kriging surfaces for each train section

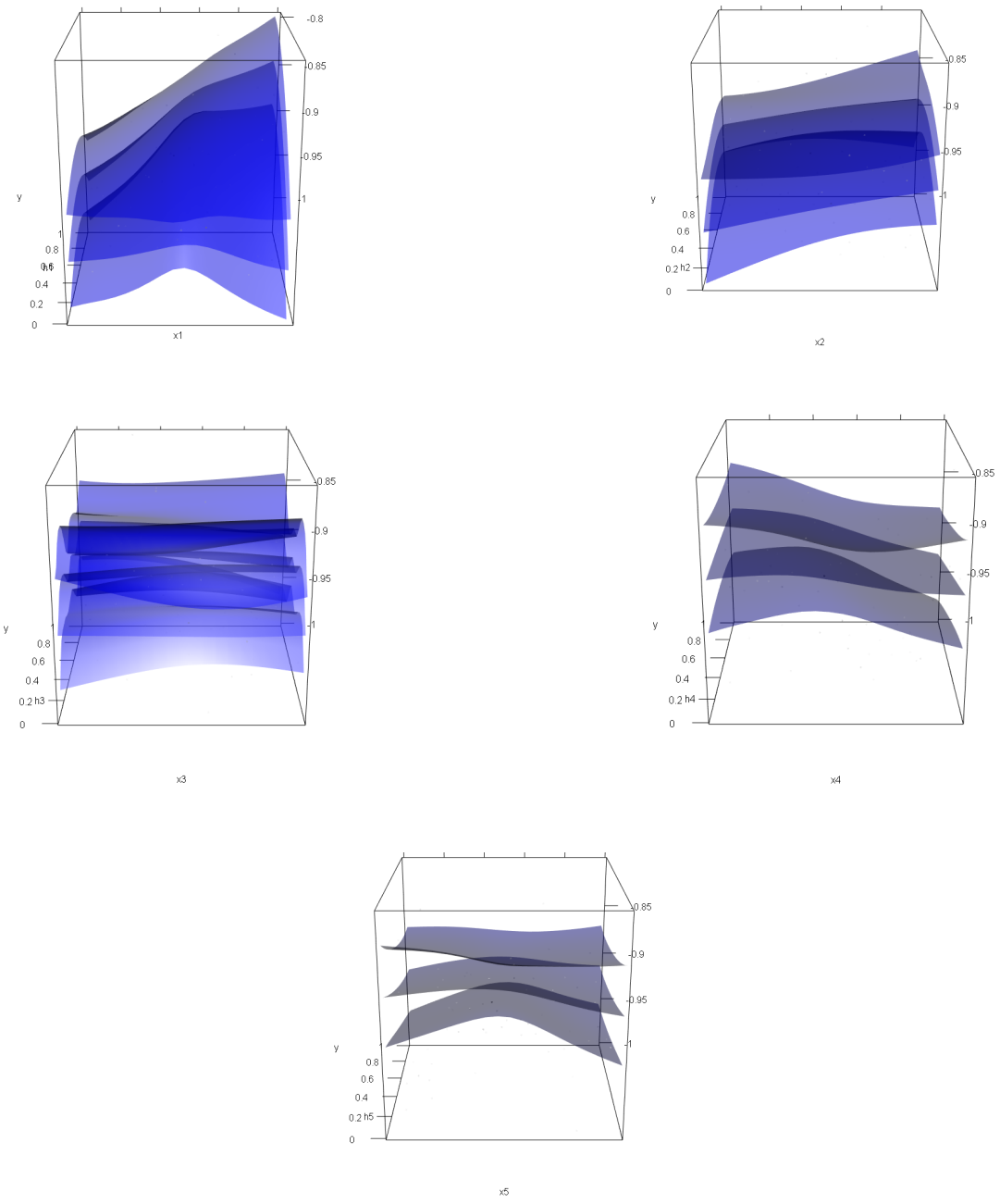

Fig. 3.4 Compression forces at 10m: Kriging surfaces for each train section

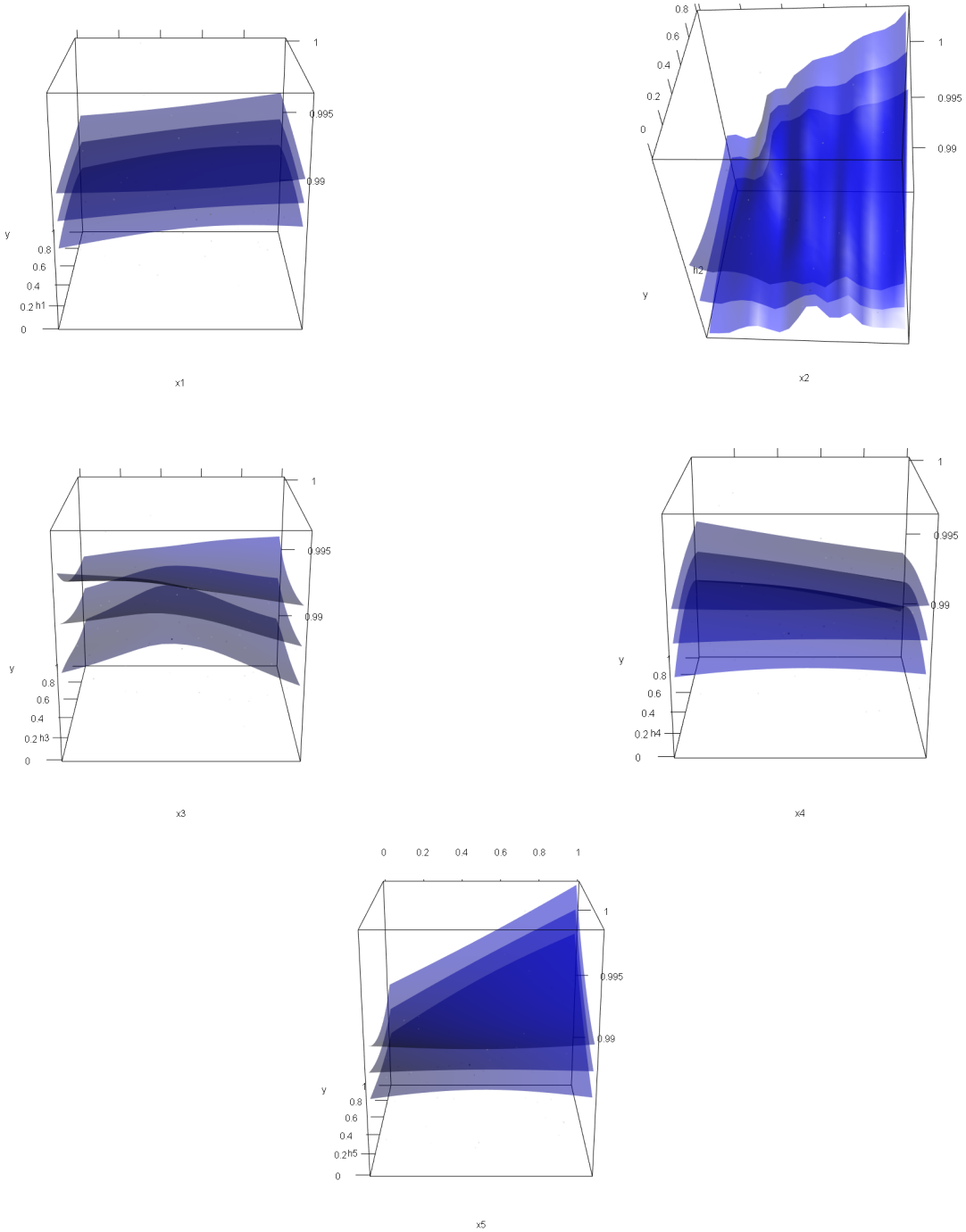

Fig. 3.5 Tensile forces at 10m: Kriging surfaces for each train section

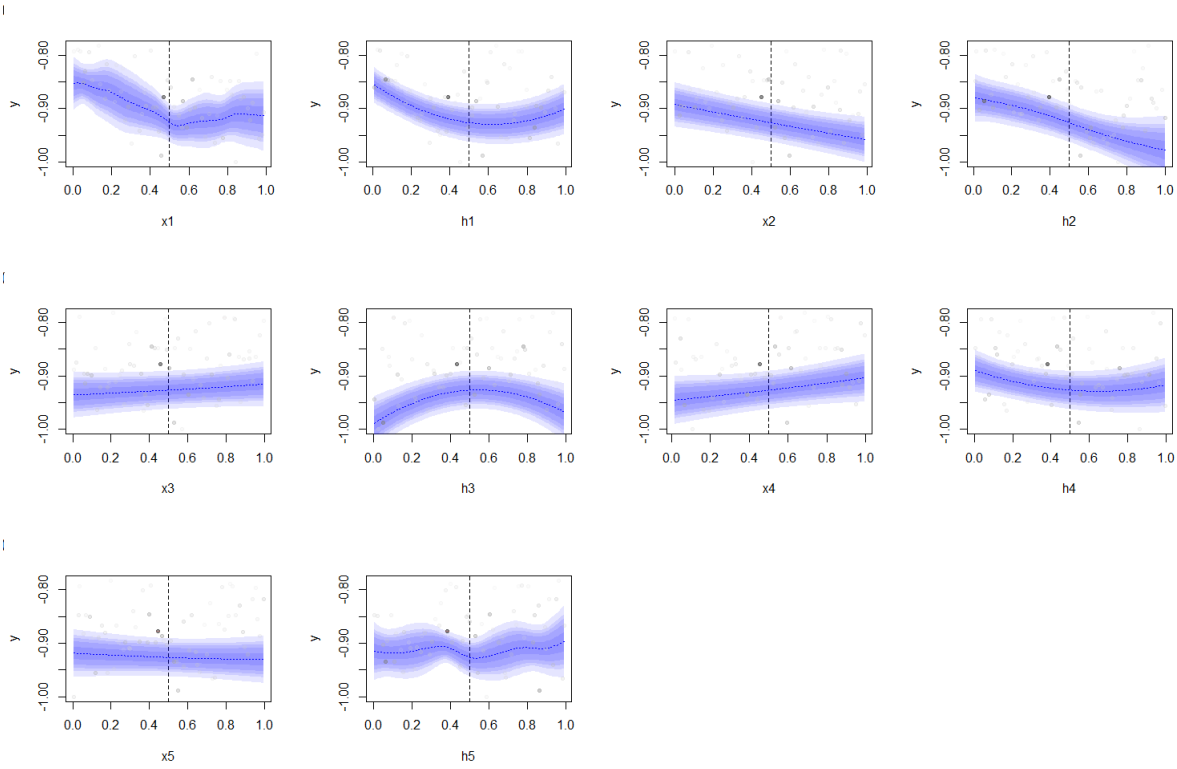

Fig. 3.6 Compression forces at 2m: Kriging surfaces for each input variable

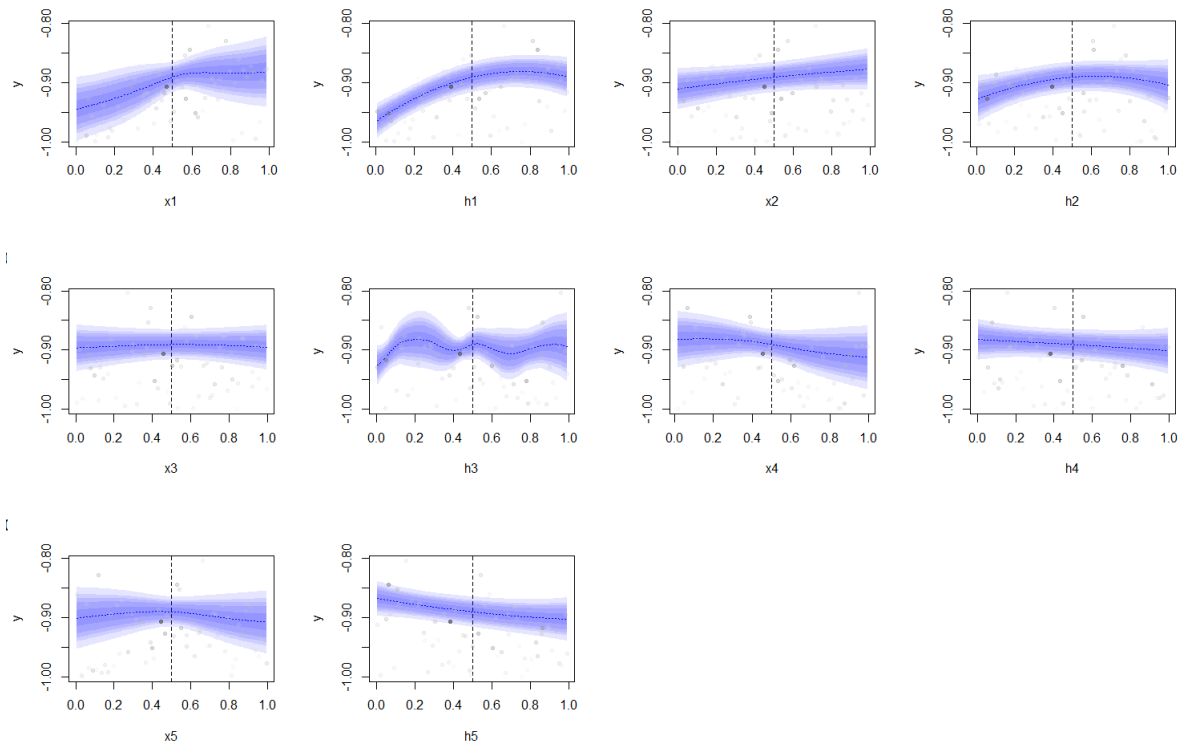

Fig. 3.7 Compression forces at 10m: Kriging surfaces for each input variable

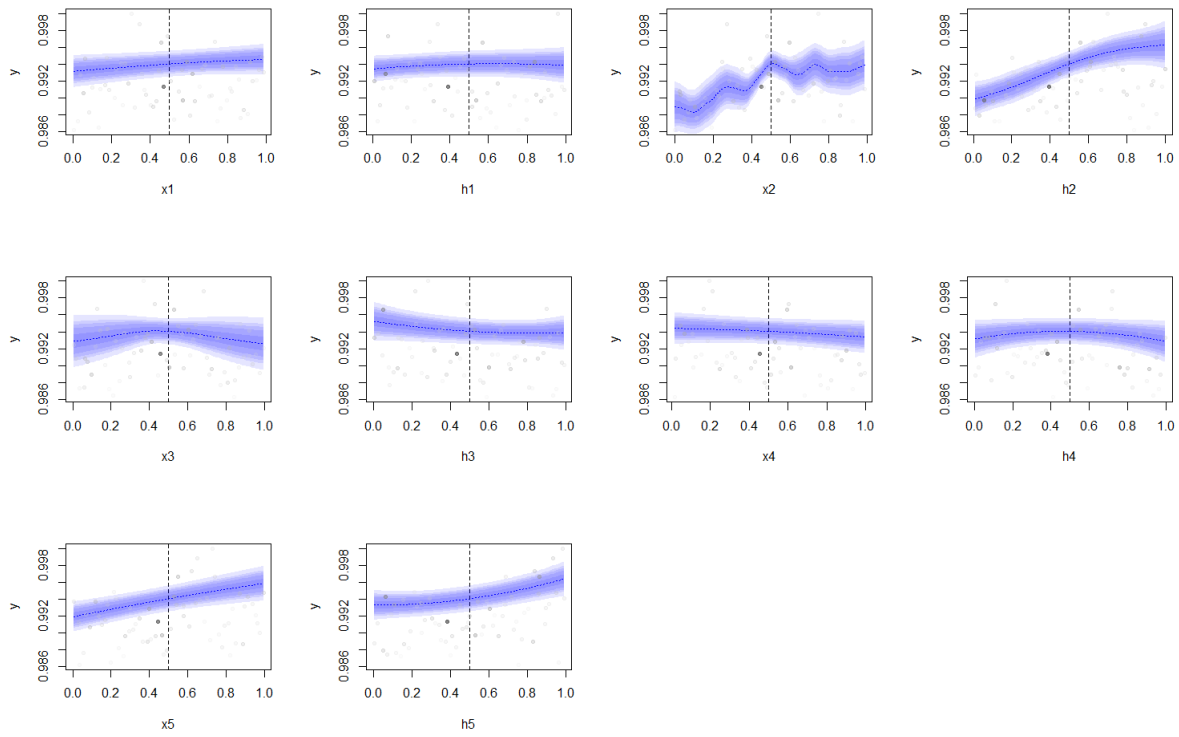

Fig. 3.8 Tensile forces at 10m: Kriging surfaces for each input variable

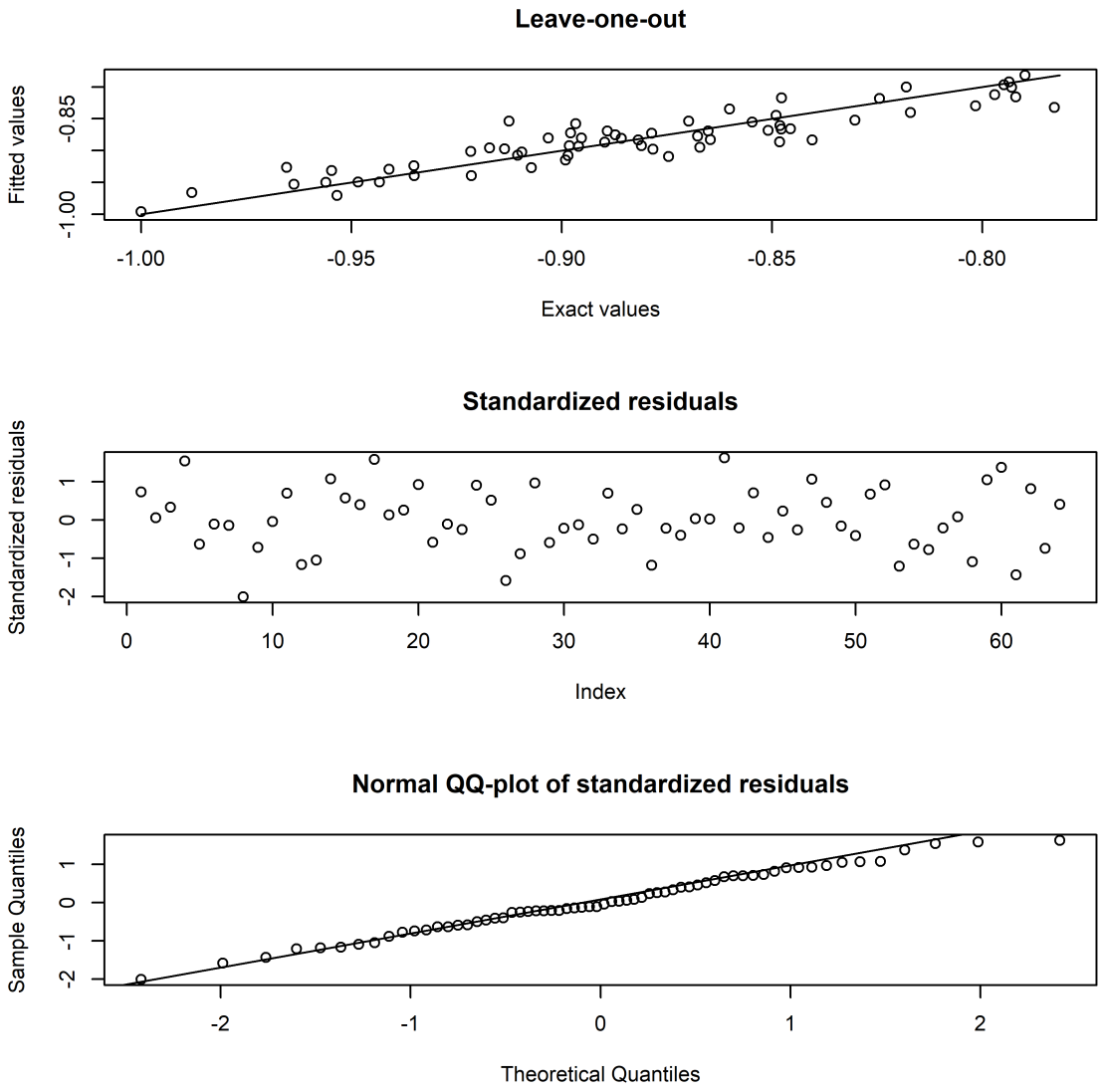

Fig. 3.9 Compression Forces at 2m: goodness-of-fit with leave-one-out method. The three plots presented are as follows: the residuals (top), the standardized variance of residuals (middle) and the Normal QQ plot of the residuals (bottom).

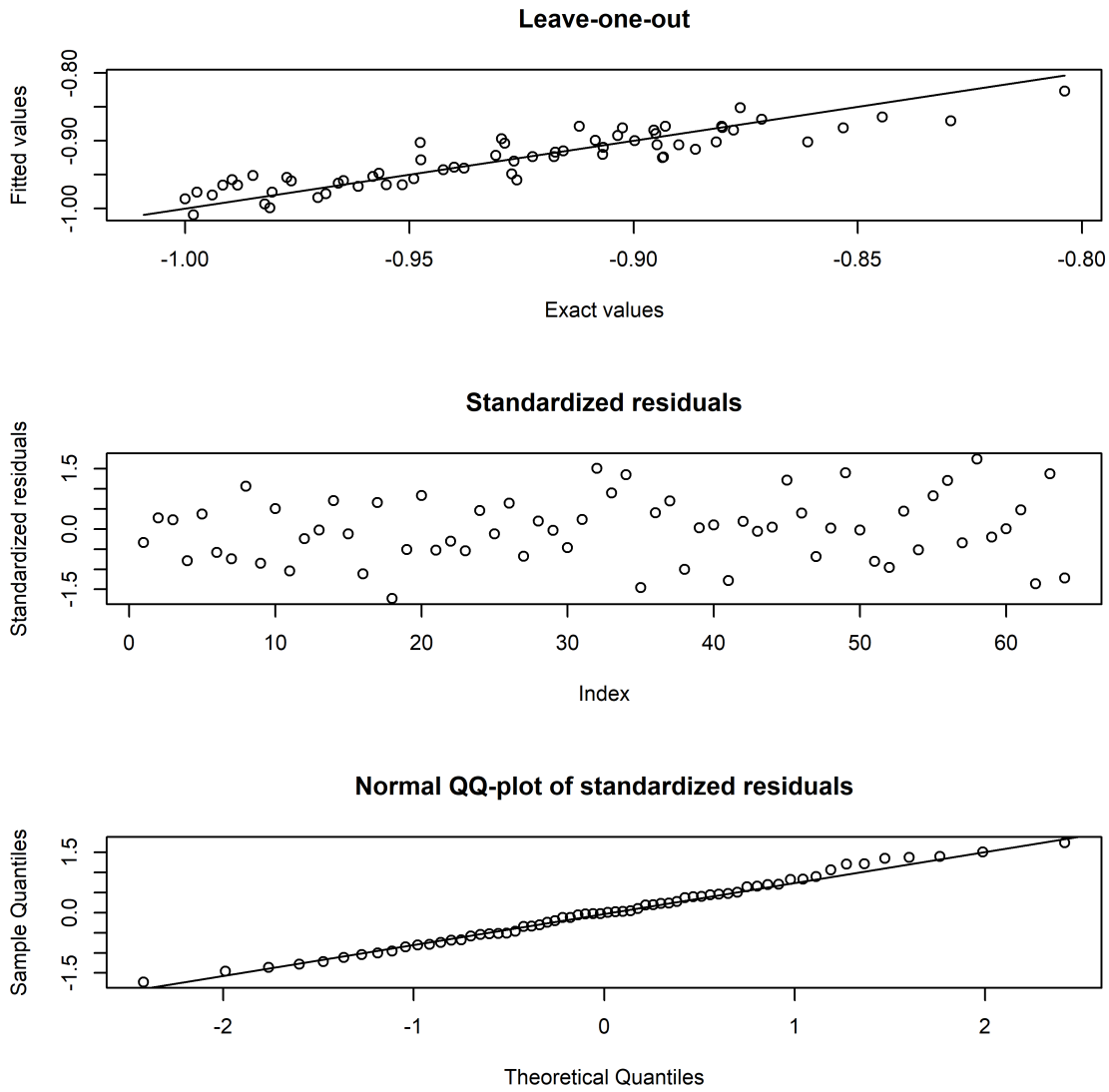

Fig. 3.10 Compression Forces at 10m: goodness-of-fit with leave-one-out method. The three plots presented are as follows: the residuals (top), the standardized variance of residuals (middle) and the Normal QQ plot of the residuals (bottom).

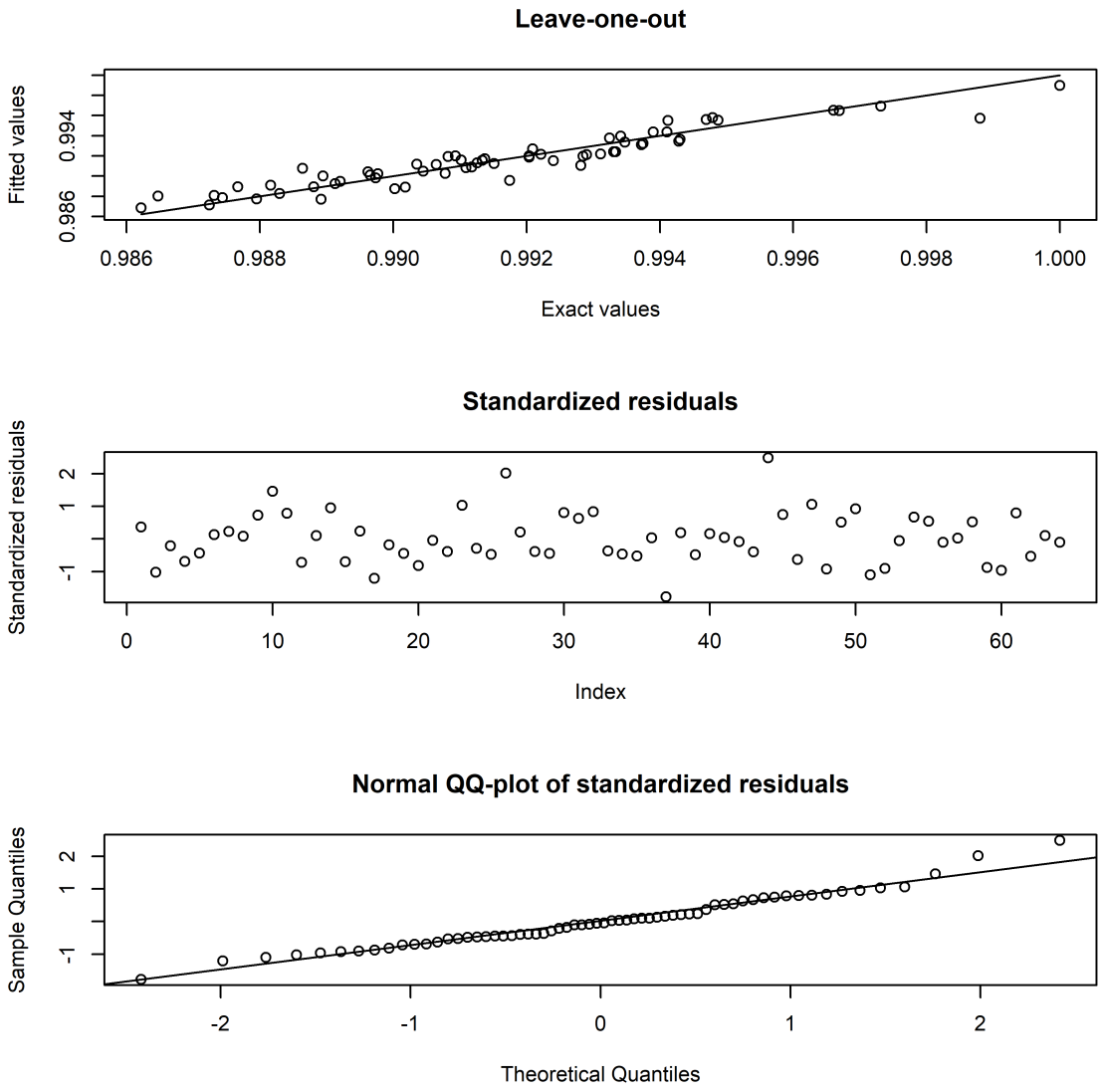

Fig. 3.11 Tensile Forces at 2m: goodness-of-fit with leave-one-out method. The three plots presented are as follows: the residuals (top), the standardized variance of residuals (middle) and the Normal QQ plot of the residuals (bottom).

3.7 Conclusions

In this project we illustrated an innovative approach to improve the payload distribution of freight trains through a suitable design for the computer experiment and Kriging modelling as subsequent model analysis. To this end, we considered a train divided in five different sections, where each train section is supposed to have its own payload distribution. The chosen train configuration is also one of the most critical, and it is generally used for the computation of in-train forces, allowing us to address more general results through the investigation of a common railway scenario. Given that the in-train forces can cause a derailment or a train disruption during an emergency braking manoeuvre, the goal was to find the best payload distribution that minimizes the in-train forces.

A SOA-based-LH design with a relatively low number of experimental runs and very good space-filling properties has been built for the computer experiment; subsequently, three Universal Kriging models with anisotropic Matérn covariance function have been estimated to find the best payload distribution able to reduce the in-train forces. A nugget parameter has been also estimated in order to prevent numerical instabilities. The data related to the output of the in-train forces have been obtained by using the TrainDy simulator (Cantone, 2011), internationally certificated by UIC. The Kriging models have been validated through the leave-one-out cross validation method. The diagnostic results are very satisfactory by confirming that the estimated Kriging models are highly predictive.

The results discussed in the previous Section showed that it is not possible to generalize an optimal payload distribution: looking at the different estimated parameters we found different optimal distributions. In fact, as already stated in Section 3.1, there is not an universally optimal payload distribution; rather it could change according to the specific train set arrangement considered. Anyway, as it has been proved also in Arcidiacono et al. (2018), the best benefit of the SOA-based-LH design, here applied, is that it is able to accurately describe the behaviour of the various payload distributions with a relatively small computational effort for the train simulations. This point is crucial not only for safety reasons, but especially for the optimization of train maintenance in terms of fatigue analyses of train couplers (central coupler or buffers/draw gears). As a consequence, this type of design can be used, together with TrainDy software, to investigate the best payload distribution for each running train. Of course, a desiderable strategy is the integration of TrainDy with this design technique in the "productive flow" for interested Railway Undertakings.

Further developments of this work mainly rely on the application of this methodology to container traffic, where the payloads (along with the percentage of braked mass for each wagon) cannot be varied continously as in this approach. In-train forces optimization of

container traffic is important for the new European trains with distributed power/braking, that will allow the circulation of trains having length up to 1500 m.

Chapter 4

The Impact of Not Randomizing a Split-Plot Experiment and How to Detect Its Effect

4.1 Introduction to the project

This project considers the impact when the experimenter either cannot or chooses not to randomize the application of the experimental factors to their appropriate experimental units for a split-plot experiment. It is based on a real case-study related to the production process of an ultrasound transducer devoted to medical imaging. The transducer is composed of hundreds of elements fabricated in several complex production phases. The compliance on the stringent process requirements as well as the high production costs require a suitable experimental planning in order to avoid as much as possible the number of failures of the transducer. To this end, a split-plot design has been planned that results particularly suitable to solve complex technological problems by also considering the well-known differentiation among hard-to-change and easy-to-change factors. Nevertheless, the technology underlying the transducer production process does not allow complete randomization of some of the experimental factors involved in the design. Randomization is a fundamental principle underlying the statistical planning of experiments, and it protects the experiment from bias induced by any systematic effect that could be present during the conduct of the experiment. This bias can have a profound negative effect on the experimental results that produce misleading conlcusions during the analysis.

The purpose of this project is to illustrate the potential consequences of lacking of randomization in a split-plot design when there is a systematic effect over the course of the experiment. To this end, we use a simulation study based on the real case-study in which the systematic effect is represented by a linear trend over time. The main aim is to show through

the simulation study the real implications of this systematic effect on both the randomized and non-randomized situations. More precisely, the simulation study considers the situations where there is no linear trend as well as when a linear trend is present at the Whole-Plot (WP) level and at the Sub-Plot (SP) level for both the not randomized and the randomized designs. It therefore demonstrates the fundamental role of randomization by especially considering the split-plot experiment in which the WP and the SP factors have a different importance for the experimenter. In what follows, we briefly review the basic principles of the experimental design theory.

4.2 Experimental design theory: basic principles

The main principles for experimental designs may be divided by considering the general (basic) theory, and specific properties, e.g. according to the peculiar features for each single design. In general, an experimental design is characterized by a set of fundamental elements, such as: the experimental unit; the experimental trial, also named run or treatment, the latter for specific cases not technological; the replicates, and the concept of factor (with the corresponding levels). A key-point in the experimental design theory is the allocation of trials to experimental units in order to estimate the differences among runs in their effects on the response measurements, where the response measurements are explicitly identified through one (or several) response variable (Cox and Reid, 2000). The response can be differently chosen and defined according to the field of application for the experimental design. In a technological context, the response variable must be improved, and optimized; moreover, in this situation the underlying interpretation of the technical process is a further aim with respect to the detecting of significant effects.

Nevertheless, a general main aim for an experimental design is the reduction of random and systematic errors. The random error is the residual error, e.g. the (residual) not-explained variability, and it should be minimized if the experimental planning has been performed involving all the source of variabilities that are considered relevant for the defined response variable. The systematic error, instead, must be avoided. To this end, the principle of randomization plays a relevant role and it constitutes the kernel point for the classical experimental design (Cox and Reid, 2000). Furthermore, randomization can be achieved through: i) the random allocation between an experimental run and an experimental unit; ii) properties of the design matrix. The latter is a crucial element for planning an experimental design and to account for its properties. For example, in a full factorial design balance and orthogonality are guaranteed by the algebraic property of the design matrix for orthogonality, while balancing is a sufficient condition when the replicates are present. Moreover, in more

complex designs, such as optimal designs, orthogonality is achieved through the Fisher Information matrix. Furthermore in a linear model, when considering the difference between two treatments (groups) with a random allocation, the residual error (ε) captures all the random differences external to the systematic effect. In this case, randomization allows for guaranteeing estimation properties according to the 2nd moment assumptions for ε, given that the 1st moment is always assumed equal to zero. In particular:

1. assuming homoscedasticity and uncorrelated residual errors implies to have a Minimum Variance Unbiased Linear Estimator (MVULE) for the mean and variance difference between the two treatments;

2. assuming that residual errors are Normally distributed with constant (homogeneous) variance implies Minimum Variance Unbiased Estimator (MVUE) for the mean and variance difference between the two treatments.

In general randomization ranges from a crucial to a less relevant role, while the main concern is related to the allocation between experimental units and trials. However, even though the avoidance of bias is not so strictly urgent when patients are not involved, the tools for avoiding it are: i) the use of randomization, and ii) the inclusion of a-priori information through intermediate and/or baseline variables (Cox and Reid, 2000).

In a technological situation, and especially when considering a split-plot design, the inclusion of intermediate variables often means that we are also evaluating: i) additional information of the technological process (variables not included in the experimental design), and, particularly, ii) noises and classification factors.

Furthermore, specific arrangements and/or inclusion of strategic variables may help to avoid or to reduce bias; for example, by including a block variable or by a constrained randomization. When considering specific allocation of trials (arrangements) balancing is a relevant condition. In fact, perversely, if all the units receive the same treatment, no effect can be estimated. Therefore, a guarantee for a good randomization could be a randomization constrained to balancing (Cox and Reid, 2000).

Essentially, randomization provides protection against any systematic effect that may occur with the experimental units during the conduct of the experiment. If such a linear trend is present, it produces biased estimates of the model coefficients if the specific factor settings are not randomly assigned to the corresponding experimental units. Conversely, the estimates of the model coefficients are unbiased if the design is randomized; however, it does so at a price. Essentially, randomization moves the impact of the linear trend into the experimental error, making it larger. As a result, randomization allows also to detect the linear trend through the residual plots. Instead, the non-randomized design cannot identify

the presence of the linear trend through the residul plots because the impact of the trend is to bias the estimated coefficients. As a consequence, the presence of the systematic effect reduces the power of the tests on the treatment effects, i.e, the probability of detecting an important treatment difference is reduced. In the following Section, these randomization issues are discussed specifically for the split-plot experiment. To do se, we briefly introduce the basic theory for split-plot designs.

4.3 Basic split-plot theory

The split-plot design has received a great attention as a valid plan in the technological field. As already stated (Section 1.4, Chapter 1), in a split-plot design there are two sets of factors, namely the WP factors and the SP factors, involved in a bi-randomization procedure. The set of trials relating to the SP factors are nested and randomized within each WP experimental unit, e.g. Whole-Unit (WU); the WUs are formed by the combinations of the WP factor levels. Each experimental unit for the SP factors, e.g. Sub-Unit (SU), is an observational unit for the WP factors. The bi-randomization strategy creates two different error terms: one at the WP level and one at the SP level. These two error terms lead to two different error variances with different degrees of freedom for performing inference. Following, these concepts are better highlighted through the definition of the general form of the split-plot model.

4.3.1 General form of the split-plot model

Let's define the general form of the split-plot model as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon} \tag{4.1}$$

where $\mathbf{y}$ is the $N \times 1$ vector of the responses and $\boldsymbol{X}$ is the $N \times p$ model matrix; $\boldsymbol{\beta}$ is the $p \times 1$ vector of WP and SP unknown coefficients; $\boldsymbol{\delta}$ is the $N \times 1$ vector of WP errors, while $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of SP errors. We assume that $(\boldsymbol{\delta} + \boldsymbol{\epsilon})$ follows a Normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ equals to:

$$\boldsymbol{\Sigma} = \sigma_{\varepsilon}^2 \mathbf{I} + \sigma_{\delta}^2 \mathbf{J} \tag{4.2}$$

In formula (4.2), σ_δ^2 and σ_ε^2 are the WP and SP variances respectively; $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{J}$ is a block-diagonal matrix which general form is expressed as follows:

$$\mathbf{J} = \begin{bmatrix} \mathbf{1}_{n_1}\mathbf{1}_{n_1}^{'} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n_2}\mathbf{1}_{n2}^{'} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_{n_m}\mathbf{1}_{n_m}^{'} \end{bmatrix} \tag{4.3}$$

where m is the total number of whole-plots, n_i is the number of SP runs in the i-th WP ($i = 1, ..., m$). In a balanced split-plot design, each WP contains the same number n of SP runs (e.g. $n_i = n$), and the total number of runs in the design is equal to $N = mn$.

Moreover, it must be noted that the WP and SP error variances (σ_δ^2 and σ_ε^2) have different degrees of freedom due to the differences in information from the different number of experimental units. More precisely, there are much less experimental units for the WP factors than for the SP factors. Therefore, the WP error variance has fewer degrees of freedom (often much fewer) than the SP error variance. As a result, the split-plot experiment can detect smaller differences in treatment effects at the SP level than at the WP level. We thus have much more information on the SP effects than the WP effects.

Lastly, it must be also noted that the two error terms (σ_δ^2 and σ_ε^2) in the split-plot design identify two different sets of residuals: one at the WP level and the other at the SP level. The SP residuals are the traditional residuals. Instead, the WP residuals are obtained as the difference between the mean of each WP and the predicted value of the WP, obtained using only the part of the model matrix $\boldsymbol{X}$ corresponding to the WPs (Vining and Kowalski, 2008; Jensen and Kowalski, 2012). In the next Subsection, we briefly discuss the randomization issues for a split-plot design.

4.3.2 Randomization issues in a split-plot design

First, let's define the ratio of the WP to SP error variances as follows (Myers et al., 2002):

$$d = \frac{\sigma_\delta^2}{\sigma_\varepsilon^2} \tag{4.4}$$

Suppose there is a systematic effect. More precisely, assume that there is a linear effect over time within each WU, and defined as follows:

$$v(t_i) = \eta_0 + \eta_1 t_i \quad \forall i = 1, ..., m \tag{4.5}$$

where t_i ($i = 1, ..., m$) is an index variable for the corresponding WU.

Similarly, assume a linear effect over time at the SP level. Let $\boldsymbol{t}_t^*$ be an $(mn \times 1)$ vector for the "times". In general:

$$\boldsymbol{t}_t^* = \begin{bmatrix} \boldsymbol{t}_1^* \\ \boldsymbol{t}_2^* \\ \vdots \\ \boldsymbol{t}_i^* \\ \vdots \\ \boldsymbol{t}_m^* \end{bmatrix} \tag{4.6}$$

where $\boldsymbol{t}_i^*$ are the times within the i^{th} WP. In our simulations, we assumed that

$$\boldsymbol{t}_1^* = \boldsymbol{t}_2^* = \ldots = \boldsymbol{t}_i^* = \ldots = \boldsymbol{t}_m^* = \boldsymbol{t}^*. \tag{4.7}$$

Therefore, for each SU k-th ($k = 1, ..., N$), the systematic linear trend over time is defined in general as follows:

$$\alpha(t_k^*) = \gamma_0 + \gamma_1 t_k^* \quad \forall k = 1, ..., N \tag{4.8}$$

where t_k^* $(k = 1, ..., N)$ is an index variable for the corresponding SU.

If we perform a proper randomization, we expect to see no bias in the estimates of the treatment effects. However, we expect to see inflated estimated variances for the error terms. This is because randomization moves the systematic trends (formulas (4.5) and (4.8)) to the corresponding error terms. On the other hand, if we do not randomize the experiment either at the WP or at the SP levels, then we should expect to see biased estimates of the treatment effects. Finally, we should not expect to see an impact at the WP level if the systematic effect is purely at the SP level (formula (4.8)). Similarly, we do not expect to see an impact at the SP level if the systematic trend is purely at the WP level (formula (4.5)).

The consequences outlined above assume that we have a large enough number of WPs and SPs to see the effects. We should be able to see the impacts at the SP level much better than at the WP level since we have more (often much more) experimental units at the SP level. It is entirely possible that for a small enough experiment one will not be able to see the trend at the WP level (formula (4.5)). In fact, often the number of WUs is sufficiently small that randomization has very little benefit over not randomizing.

An important question is how could we detect the presence of the linear trend. One should expect that the appropriate residuals would reflect the presence of the linear trend, providing a basis for estimating the linear trend in an effort to minimize its bias on the estimated treatment effects.

Following, after a brief description of the motivation study and the split-plot planning, the issues discussed above are demonstrated and explained in details through the simulation study results.

4.4 The motivation study

4.4.1 The technical problem

There exists an ever-increasing demand for piezoelectric and electrostrictive sensors and actuators fabricated with high levels of geometrical precision and with the minimum of material damage. The applications, which include miniature high-frequency ultrasonic arrays, infrared imaging devices and print-head actuators, usually require either thin electroactive layers, or finepitch multi-element arrays, the former being created by grinding and lapping and the latter by dicing. At the same time, the bonding between materials with different thermal expansion coefficient becomes critical, when the fundamental requirement of strength adhesion must be satisfied. The general structure of the transducer devoted to medical imaging based ultrasounds is shown in Figure 4.1.

Ultrasonic diagnostic apparatus is known which radiate ultrasonic pulse beams into an object to be examined, receiving the echoes which are reflected by the boundary of the structures of organs in accordance with a difference in acoustic impedance, displaying them on a monitor, so that the structure of the organs can be observed from the displayed image. Since they enable the interior of the body to be diagnosed from the exterior, they are widely used. The study takes the linear phased array probe head (Figure 4.2) and their potential failures during the manufacturing process into account.

This probe is used for medical application and, in particular, to monitor the cardiac muscle. In order to focus the ultrasonic beam properly on the heart, passing through the ribs of the thoracic cage, the probe is characterized by an acoustic stack composed by 128 PZT array element (sources and receiver for ultrasound waves), in which electrical connection and matching layer for each element are provided. The support of the stack is the backing, fundamental component that provides not only mechanical robustness, but also the mitigation of acoustic noise effect. The compliance (100%) of each array element is mandatory and, so, dicing and bonding process are very critical.

Even though the study is based on the real technical problem described above, we point out that our results are obtained only through simulations.

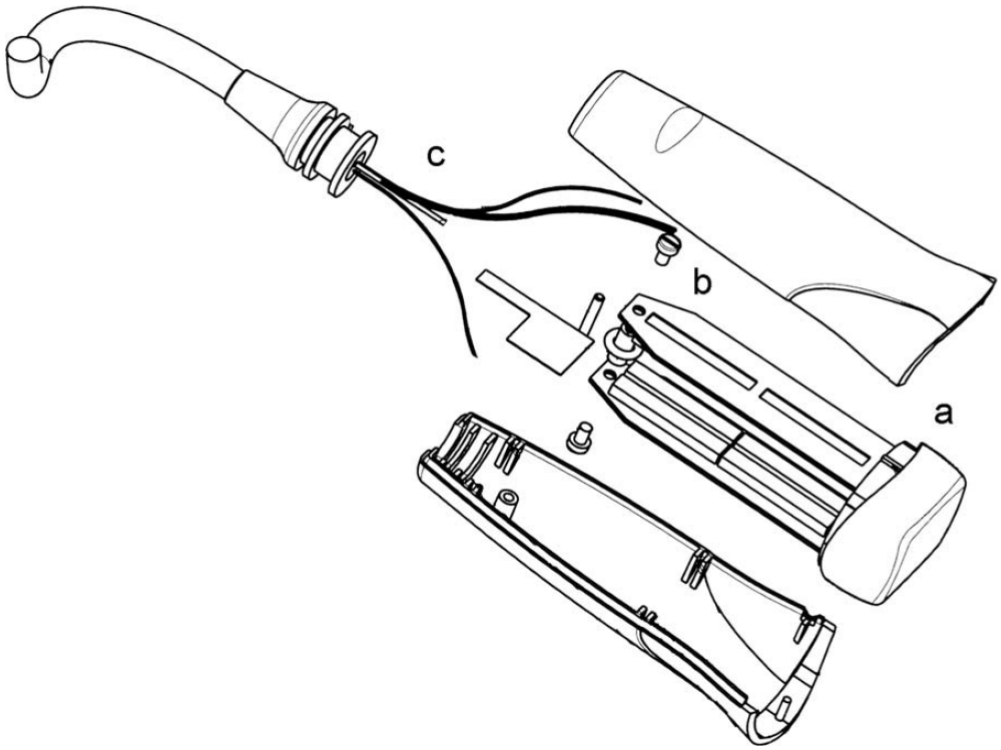

Fig. 4.1 Probe system: a) Probe head group; b) Connection groups; c) Cable (source: Catelani et al., 2012).

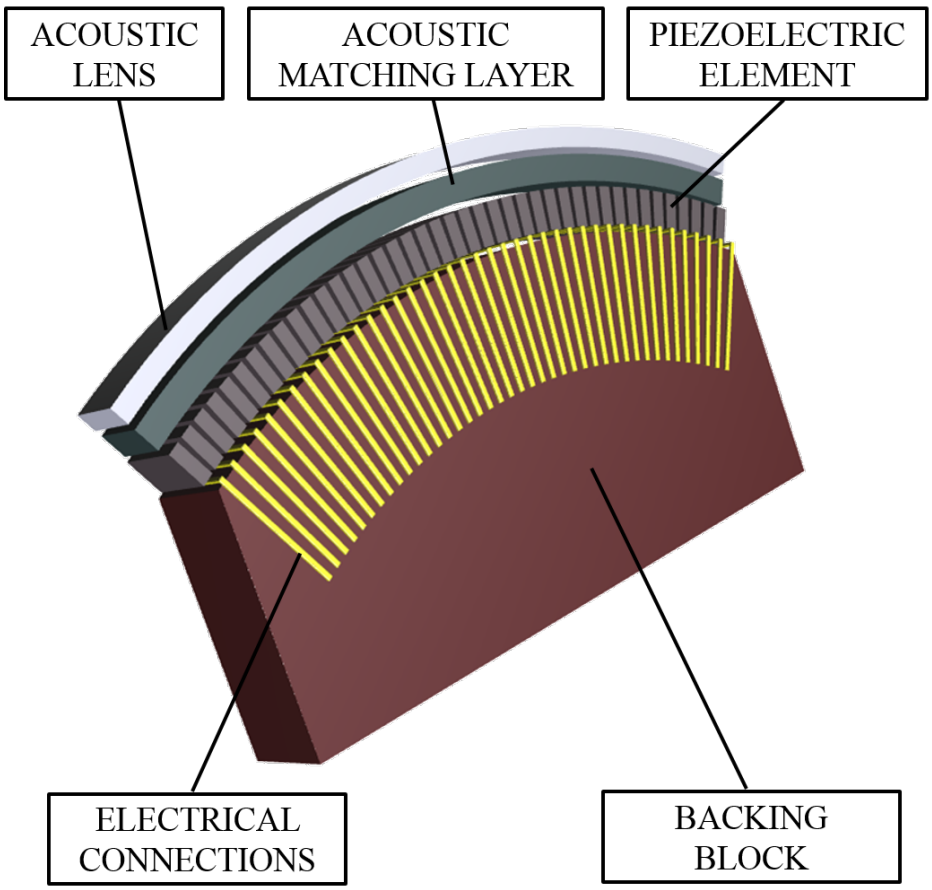

Fig. 4.2 Phased array probe: acoustic stack (source: Spicci, 2012).

4.4.2 The split-plot planning

The response variable is the total array capacitance after repoling. The split-plot design is planned by considering all the sources of variabilities involved in the transducer process during phases no.4 and no.5. It must be noted that, in addition to the control and process variables involved in the design, some other variables should be included as external variables, such as: operators, environmental variables (humidity and temperature of the laboratory), times (hour) spent between phases, and so on. Nevertheless, at this stage of the analysis, these external variables have still not been considered, because at this stage we are dealing with simulations for detecting the impact of no-randomization versus randomization. The plan is to conduct the experiment over four nights in a single week. During each night two dicing-machines are employed and then they are settled to the same levels of blade-spindle-revolution (rotation/min) and blade-feed-rate (mm/sec). For each machine, the transducers involved in the dicing process (during each night) are four. Then, the four transducers are built through a processing layer bonding, which use two types of adhesives and two levels of bonding time (minutes).

Blade-spindle-revolution and blade-feed-rate are hard-to-change or WP factors. The dicing machine, adhesive, and bonding time are easy-to-change or SP factors. The experiment uses two levels for each factor. Therefore, the total number of WP experimental units is four, e.g. $m = 2^2 = 4$. The total number of SP experimental units within each WU is eight, e.g. $n = 2^3 = 8$, while the total number of SP experimental units in the experiment is $N = mn = 32$. Furthermore, at this stage of the analysis, all the design factors are considered quantitative, and may be treated as continuous in the range [-1,1].

4.5 The simulation study

Randomization is one of the three basic principles for conducting of experimental design. Nevertheless, as usual in the technological field, randomization cannot be completely achieved due to production process requirements. This is the case for our motivating study in which a split-plot design is planned for the study of ultrasound transducer process. In fact, as already stated in the previous Section, two of the SP factors, e.g. adhesive and bonding time, are not randomized. To this end, we carry out a simulation study to illustrate the possible consequences by not randomizing the design completely. Even though the motivating example only does not randomize some SP factors, this study investigates the implication of not randomizing at both the WP and SP level. Therefore, we consider the following two different cases:

A) a split-plot design in which only the WP factors are not randomized;

B) a split-plot design in which only the SP factors are not randomized.

For both cases, we assess the effect of randomizing/not randomizing on: i) the estimated model coefficients, and ii) the WP and SP residuals respectively.

In order to show the effect from not randomizing the design, we simulate with and without adding a linear trend at the WP level only (Case A, formula (4.5)), and at the SP level only (Case B, formula (4.8)). To this end, for each case we consider the following four different arrangements:

1. Not Randomized Design Without Trend (NRDWT): that is the split-plot design in which, depending on the case, the WP or the SP factors are not randomized;
2. Not Randomized Design With Trend (NRDCT): the not randomized split-plot design with the inclusion of the corresponding linear trend in the data;
3. Randomized Design Without Trend (RDWT): the split-plot design with all the factors completely randomized and without trend;
4. Randomized Design With Trend (RDCT): the randomized split-plot design in which we have also added the corresponding linear trend in the data.

For each case (A and B), and for each split-plot arrangement (NRDWT, NRDCT, RDWT and RDCT), the study uses the R software to generate ten thousand datasets of $N = 32$ observations. Instead, the study uses SAS PROC MIXED (Windows Platform 9.4) to estimate the models.

We report box-plots for the WP and SP factors for each split-plot arrangement to illustrate the impacts of not fully randomizing the designs. For the sake of brevity, we report the box-plots for the blade-spindle-revolution WP factor and for the dicing-machine SP factor only. The WP factor blade-feed-rate exhibits the same basic behavior as the blade-spindle-revolution. Similarly, the other two SP factors exhibit the same basic behavior as the dicing-machine. We report the WP and SP residual plots for one and for a hundred datasets (SP residuals), and for five and a thousand datasets (WP residuals), again pointing out that the other plots show the same pattern.

4.6 Case A): Not randomizing the WP factors

First, we investigate the effect of not randomizing the WP factors by considering the implications on: i) the estimated model coefficients, and ii) the WP and SP residuals respectively. To this end, we add to the NRDCT and the RDCT datasets, the linear trend at the WP level

(formula (4.5)). Given that the $\hat{\beta}$ vector of estimated coefficients should be unbiased, we have fixed $\eta_0 = 0$ and $\eta_1 = 1.4$. The η_1 value has been chosen according to the variance ratio d in formula (4.4). Therefore, the trend in formula (4.5) is computed as follows:

$$v(t_i) = \eta[t_i - \bar{t}] \quad \forall i = 1, ..., m \tag{4.9}$$

where t_i ($i = 1, ..., m$) is the index variable for the corresponding WP experimental unit and $\bar{t} = \frac{1}{m}\sum_{i=1}^{m} t_i$; moreover, without any loss of generality we omit the suffix for η from now on.

4.6.1 The implications on the estimated model coefficients

We now illustrate the implications on the estimated model coefficients as the result of not randomizing the WP factors. The differences of not randomizing vs randomizing are highlighted when the linear trend at the WP level is included in both the randomized and the not randomized designs (formula (4.9)).

We first examine the effect on the WP factor estimates when the WP factors are not randomized (e.g. blade-spindle-revolution and blade-feed-rate). As expected, when the WP factors are not randomized, the impact is clearly highlighted only in the WP factor estimates (Figure 4.3). As a matter of fact, we refer to the blade-spindle-revolution WP factor for which the true coefficient value is equal to 0.2. More precisely, the WP estimate distribution of the NRDCT dataset has a different location (median) value with respect to the other three ones (Figure 4.3). This result confirms that the estimated coefficient for the NRDCT dataset is biased while the other three are not. Furthermore, when the WP factors are randomized (RDCT dataset), we observe only the inflation in the variance of the estimated coefficient due to the presence of the linear trend, but unbiased estimates.

When considering the results for the SP factors, we can conclude that there is no effect on the SP factor estimates when the WP factor are not randomized. To illustrate this, we refer to the estimated coefficients box-plots for the dicing machine SP factor (Figure 4.4), pointing out that the other two SP factors show the same behavior. We clearly see that there are no differences in the SP estimate distributions for all four arrangements. This means that not randomizing at the WP level has no implications on the SP level: whether the WP factors are randomized or not randomized, the SP coefficients are always unbiased. This result is consistent with what we would expect: the linear trend is purely at the WP level, and as a result, its effect is purely at the WP level.

4.6.2 The implications on the WP and SP estimated residuals

Next, we examine the behaviour of the WP and SP residuals respectively. To this end, in Figures 4.5 and 4.6 we report the residual plots for the WPs for five and a hundred consecutive simulation runs respectively. We see that the residuals from the randomized experiment in the presence of the linear trend are different from the other three plots. However, one cannot see a clear pattern revealing the trend. The basic reason is that we did not have large enough WP trials. There is only one degree of freedom for the WP error term after fitting the intercept and the two main effects. Therefore, we can conclude that the experiment is too small to allow us to see the linear trend. The small size of the degrees of freedom does not allow the randomization to distribute the linear trend effectively. Figure 4.7 displays the box-plot of the mean squared residuals for the WPs. It clearly demonstrates the inflation in the estimate of the WP error variance as a result of the linear trend at the WP level.

When considering the SP residuals, there are no differences between each of the four split-plot arrangements, and therefore between not randomizing versus randomizing the design (Figures 4.8 and 4.9). The mean squared residuals for the SPs also confirm this result (Figure 4.10) that is completely consistent with the previous findings for the SP estimates. Namely, we do not observe an effect on the SP estimates due to not randomizing the WP factors, and consequently, we do not observe it on the SP residuals.

4.7 Case B): Not randomizing the SP factors

We now consider the case in which the SP factors are not randomized. As previously for case A), we investigate the effect of not randomizing the SP factors by considering the implications on: i) the estimated model coefficients, and ii) the WP and SP residuals respectively. To this end, we add to the NRDCT and the RDCT datasets, a linear trend at the SP level. Similarly to case A), we fixed $\gamma_0 = 0$ and $\gamma_1 = 1.4$. Without any loss of generality we omit the suffix for γ_1 from now on. Therefore, the linear trend at the SP level in formula 4.8 is computed as follows:

$$\alpha(t_k^*) = \gamma[t_k^* - \bar{t}] \quad \forall k = 1, ..., N \tag{4.10}$$

where t_k^* $(k = 1, ..., N)$ is the index variable for the corresponding SP experimental unit. The linear trend at the SP level in formula (4.10) is reported in Figure 4.11.

4.7.1 The implications on the estimated model coefficients

Consistently with case A), the differences of not randomizing versus randomizing are highlighted when the linear trend at the SP level is included in both the randomized and the not randomized designs (formula (4.10)).

First, we examine the effect on the WP factors estimates due to not randomizing the SP factors. To this end, we consider the box-plots for the blade-spindle-revolution WP factor for which the true coefficient value is equal to 0.2 (Figure 4.12). As expected, there is no effect on the WP factor estimates when the SP factors are not randomized. That is, whether the SP factors are randomized or not randomized, the WP estimates are always unbiased with median and mean values of the estimates equal to the true ones (Figure 4.12). This result is completely consistent: the linear trend is purely at the SP level, and therefore, its effect is purely at the SP level. To illustrate this issue, we consider the dicing machine SP factor for which the true coefficient value is equal to -0.77. More precisely, the SP estimate distribution for the NRDCT data has different median value with respect to the other three ones (Figure 4.13). This result indicates that not randomizing at the SP level biases the SP estimates by the linear trend. Moreover, even though the SP estimates for the RDCT dataset show a greater (inflated) variability with respect to the corresponding estimates of the NRDCT data, the SP estimates of the RDCT resulted unbiased with an expected value of the beta coefficient equal to the true coefficient value. The inflated variability can be explained through the trend effect that randomization moves to the SP error variability, but preserving unbiased estimates.

4.7.2 The implications on the WP and SP residuals

Next, we examine the implication on the WP and SP residuals as a result of not randomizing the SP factors. The effect on the WP residuals reflects the results obtained for the WP estimates. More precisely, we do not observe an impact on the WP estimates when the SP factors are not randomized, and consequently, we do not observe it also on the WP residuals for the four different design arrangements (Figure 4.14). The results is also confirmed when considering the overall pattern of the WP residuals reported for the first hundred simulations: the residual pattern for the four different scenarios are all identical (Figure 4.15).

The results are quite different for the SP residuals when there is a linear trend at the SP level. Figure 4.16 illustrates the SP residuals for five consecutive simulation runs. As expected, there is no difference between the randomized and non-randomized designs if there is no linear trend (NRDWT and RDWT datasets). On the other hand, we see a pronounced difference between the NRDCT and the RDCT datasets when the linear trend is present in

the data. More precisely, the non-randomized case is exactly the same as the plots when there is no linear trend. However, the plot of the residuals for the randomized design clearly identifies the trend. The results is also confirmed when considering the general pattern for the SP residuals for the first hundred simulations (Figure 4.17). In fact, the SP error variance for the RDCT data is inflated by the linear trend as illustrated by the mean square box-plots for the SP residuals (Figure 4.18). These findings confirm that randomization allows to effectively detect any systematic effect. As a consequence, if the residual plot clearly show the presence of the trend, the analyst will be also able to correct for it through the inclusion of the linear trend as a covariate in the statistical model.

Finally, it must be noted that we simulated also a third case in which both the WP and the SP factors are not randomized. The plots show the same impacts at the WP level as for Case A), and the same results at the SP level as for Case B). For this reason, the results are not reported.

4.8 The Impact of Different Trend Effects

In this Section we carried out a further analysis in order to study the impact of different magnitudes for the trend effects, η and γ, on the results. The trend effects η and γ control the impact of the linear trend at the WP and SP levels respectively. Therefore, an important issue to consider is to what extent is still possible to see the impact of the systematic effect. To illustrate this issue, we use the following values for the trend effects at the WP and the SP levels respectively:

- $\eta = 1.4$ (Case A) and $\gamma = 1.4$ (Case B);
- $\eta = 0.7$ and $\gamma = 0.7$;
- $\eta = 0.35$ and $\gamma = 0.35$ (considered small for technical knowledge).

First, we examine the impact of the linear trend at the WP level (Case A) by considering the three different values for η. Figure 4.19 displays the box-plots for the blade-spindle-revolution WP estimates in terms of the three values for η. By specifically considering the NRDCT datasets, more the trend effect η decreases, slighter is the impact in terms of bias on the WP estimates (Figure 4.19). Furthermore, whatever is the magnitude of the trend effect, the RDCT dataset still perfectly preserves unbiased coefficient estimates. The only impact of the magnitude of the trend effect on the RDCT data is that smaller is its value, smaller is the inflated variability in the WP estimates distribution.

When considering case B), the results are even more pronounced. The linear trend at the SP level in terms of different values for γ is reported in Figure 4.20. It is straightforward to show that more the trend effect decreases, slighter is the linear trend at the SP level (Figure 4.20). The impact of the trend effects on the SP estimates is consistent with what obtained for the WP estimates for case A). More precisely, for the dicing machine SP factor, as γ decreases, so does the variability in the coefficient estimates for the randomized experiment, and so does the bias in the estimated coefficient for the non-randomized experiment (Figure 4.21).

Last, we report the SP residuals for the randomized design in terms of the three values for the trend effect γ (Figures 4.22 and 4.23). The SP residual plots for five consecutive simulation runs clearly highlights the linear trend (Figure 4.22). More precisely, for large values of γ (e.g. $\gamma = 1.4$), the linear trend in the SP residuals is very strong. As the γ value decreases, it becomes ever more slighter. Moreover, for small values of γ (e.g. $\gamma = 0.35$), the SP residuals still show the trend. The overall decreasing pattern for the SP residuals when the trend effect decreases is clearly highlighted when considering the residual plots for hundred simulations (Figure 4.23). Again, we see the decrease in variability as γ decreases in the box-plots of the mean squared residuals for the randomized design (Figure 4.24).

4.9 Conclusions

Randomization is a fundamental principle for the design and analysis of experiments. It plays a key role, and it primarily allows us to protect the factor estimates by the presence of bias, which could derives from any systematic effect when conducting the experiment. Through the simulation study carried out in this project we showed which is the real impact of lacking of randomization. Although our study investigates the implications in a split-plot design, the main consequences due to lacking of randomization discussed in the previous Sections are valid also in general. More precisely, by not randomizing the application of the experimental factors to the experimental units, any systematic effect that could be present during the conduct of the experiment will lead to biased coefficient estimates. This bias created by the linear trend if we do not randomize is a very serious issue. More precisely, coefficients that are not practically significant from an engineering perspective can become statistically significant. Conversely, coefficients that are of practical significance can become statistically insignificant. The bias clearly can lead to very inappropriate conclusions from the statistical analysis. Moreover, the results related to the different trend effects discussed in the previous Subsection confirm that also a small systematic effect will still produce biased estimates. Consequently, the analyst will draw unmeaningful conclusions about the importance of the

factors in the analysis. On the other hand, randomization provides a protection against the presence of any systematic effect. A fundamental issue to stress is that randomization does not remove the systematic effect. Rather, it moves it to the error term making it larger so as preserving unbiased coefficient estimates.

The split-plot experiment on which we have assessed the effect of lacking of randomization is nowadays a basic design for industrial experimentation. Its bi-randomization procedure due to the presence of two types of the experimental units should be carefully considered when planning and analyzing the design. To this end, if the WP factors could not be randomized we expect to obtain biased WP estimates but unbiased SP estimates and this is what we have demonstrated through the simulation results. Conversely, if the SP factors could not be randomized we obtain biased SP estimates but unbiased WP ones. Moreover, differently from the linear trend at the WP level, the trend at the SP level was large enough to allow randomization to effectively distribute the trend. The simulation study establishes that if one can randomize the SP factor settings but not the WP, then there is no effect on the SP inferences, which is very important. Almost always, the SP effects are of most interest, especially in split-plot designs with a small number of WP experimental units. It also shows that if only the SP factors are not randomized, then all of the impact of the non-randomization are at the SP level. Once again, the SP effects typically are very important. As a result, one must be very careful when the design is not randomized at the SP level.

In this simulation study, we have assumed that the systematic effect consists in a linear trend. We are fully aware that in practice other possible forms could occur for the systematic effect, and further investigations could also consider this issue. Nevertheless, the basic linear form of the trend allows to fully address the main aim of this project; that is, the presence of the linear trend illustrates excellently the main implications of not randomizing the experiment.

A last important issue to note is that our simulation study is based on a real case study for which some of the SP factors was not randomized due to production process requirements. In practical industrial experimentation this situation is very frequent than expected. Engineers and analyst therefore should to carefully keep in mind the fundamental role of randomization in the design and analysis of experiment by especially considering that by not randomizing there is no way to remove the effect of any systematic effect that almost surely occurs during the experimentation.

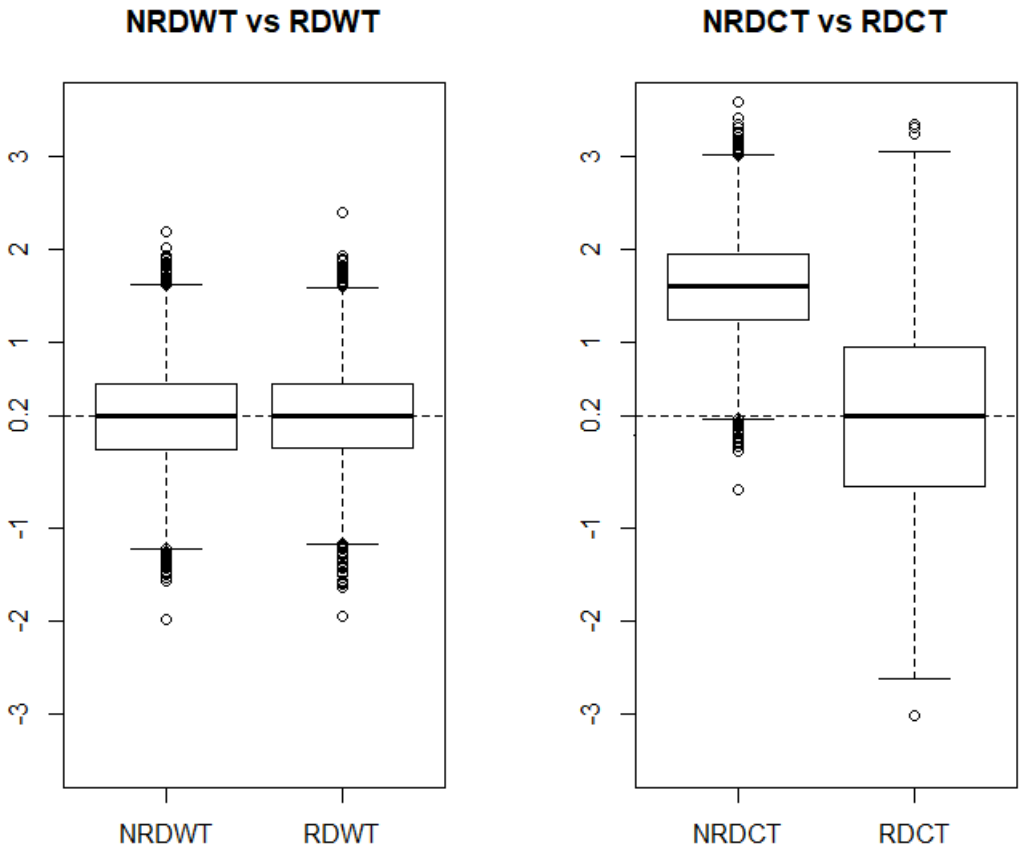

Fig. 4.3 Case A: estimated coefficients box-plots for the blade-spindle-revolution WP factor.

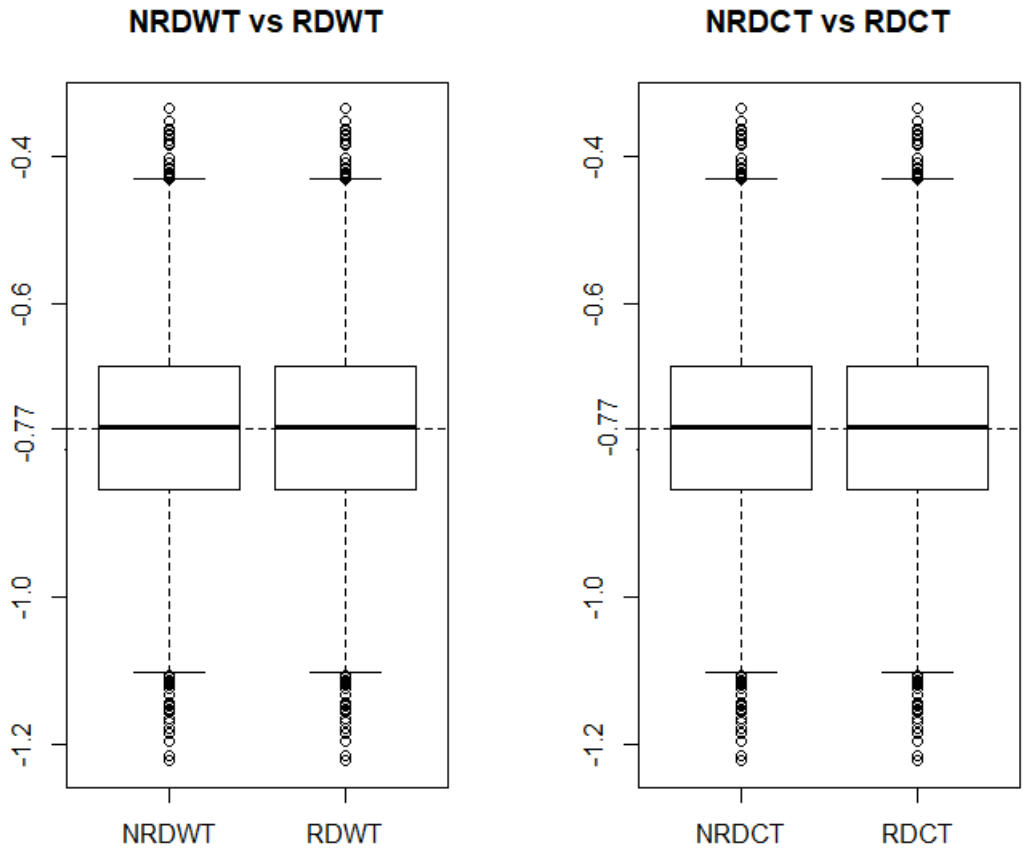

Fig. 4.4 Case A: estimated coefficients box-plots for the dicing machine SP factor.

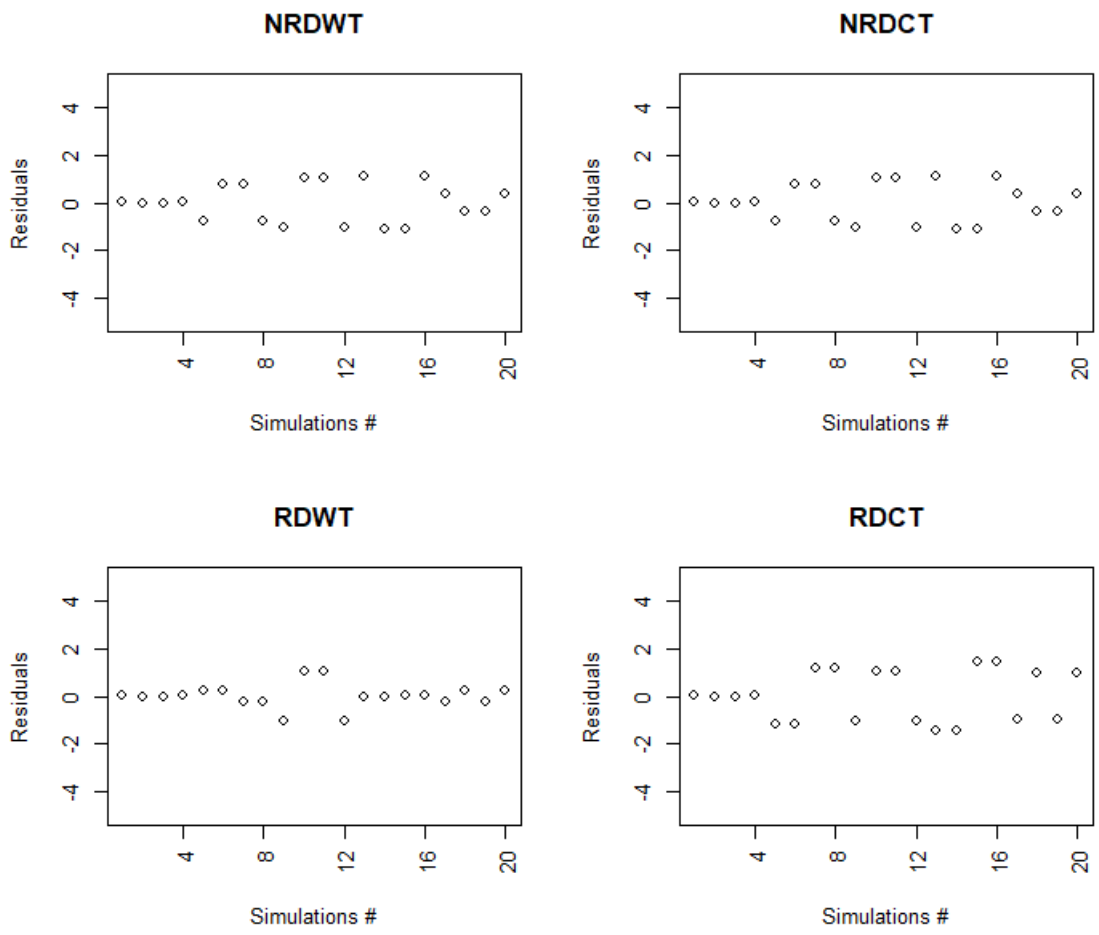

Fig. 4.5 Case A: WP residual plots for simulations no.3-7.

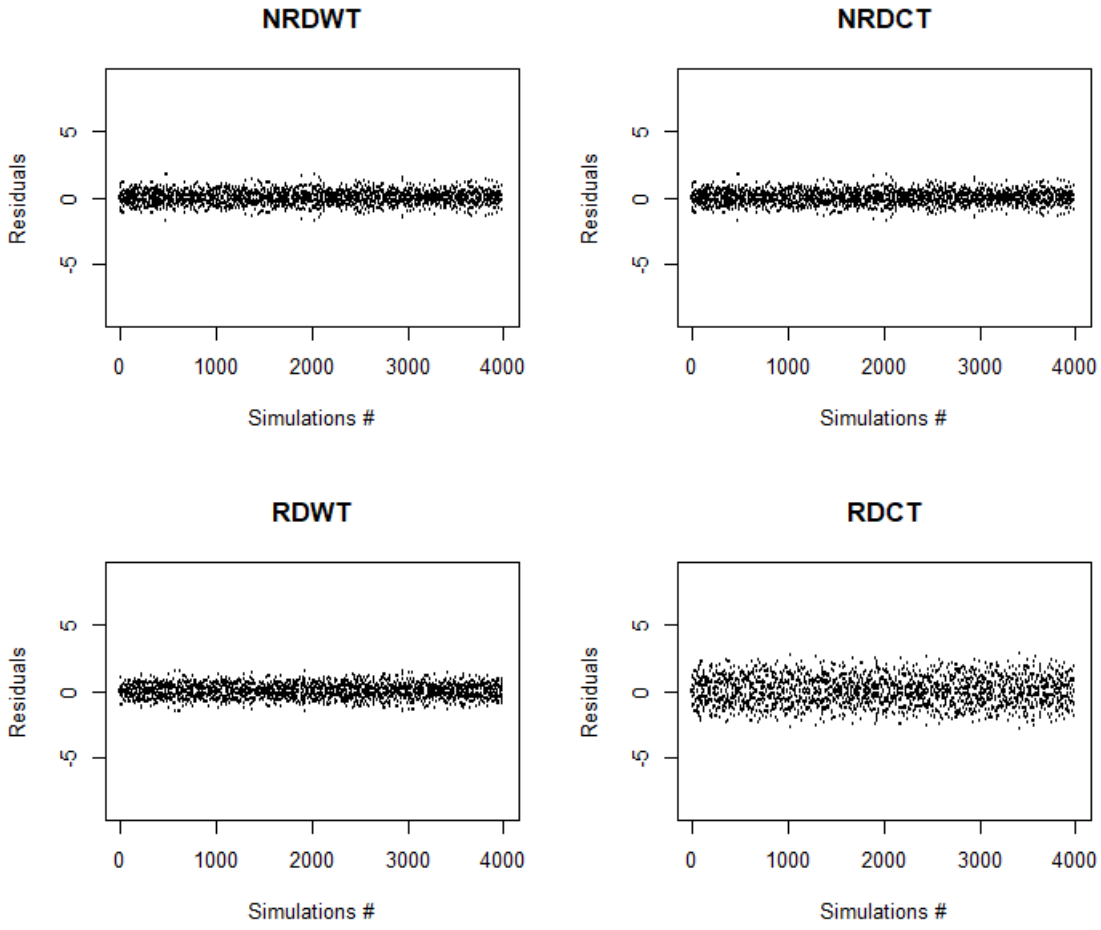

Fig. 4.6 Case A: WP residual plots for simulations no.1-1000.

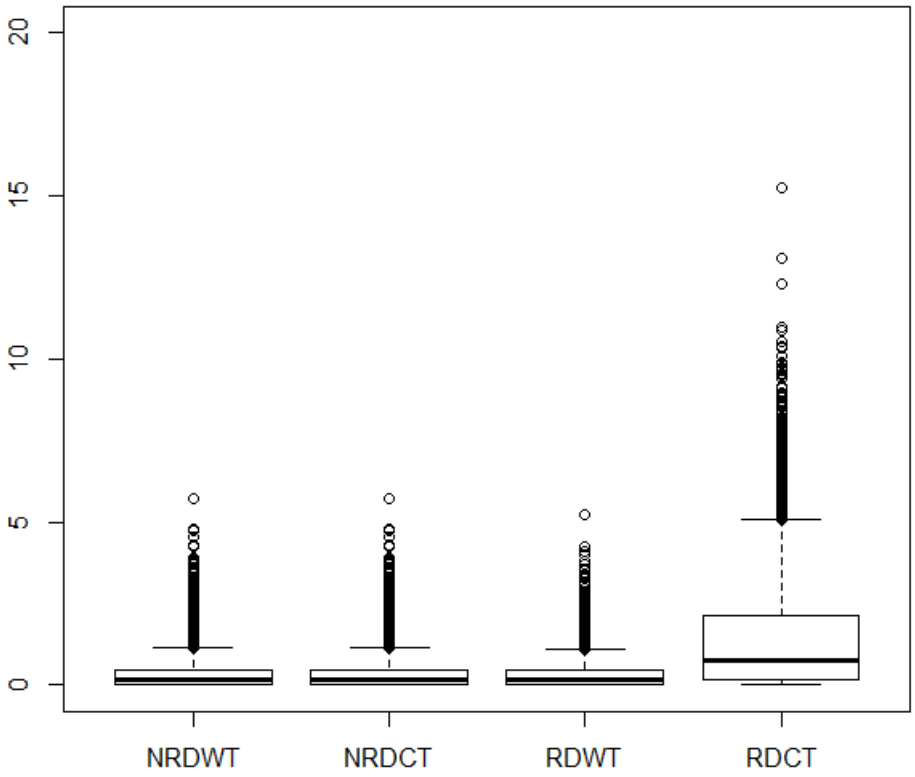

Fig. 4.7 Case A: Mean square box-plots of the WP residuals.

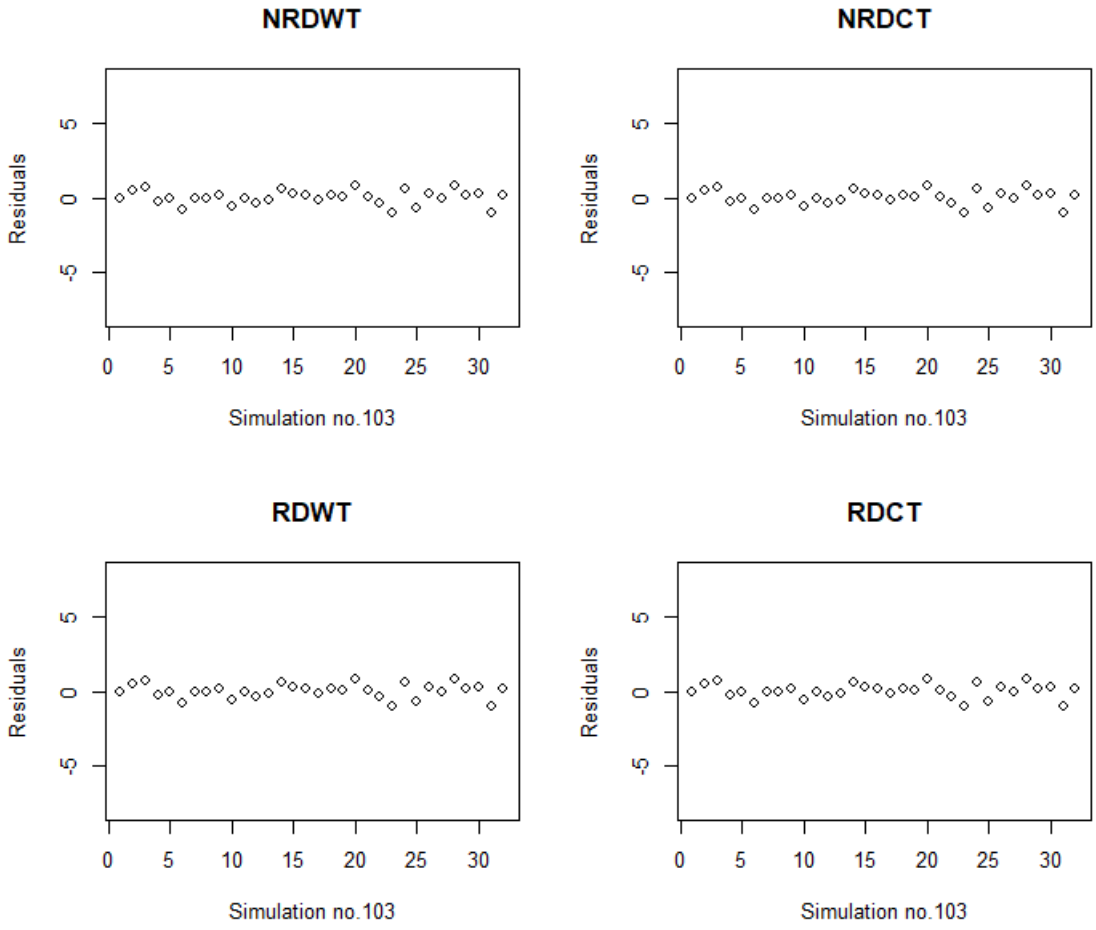

Fig. 4.8 Case A: SP residual plots for simulation no.103.

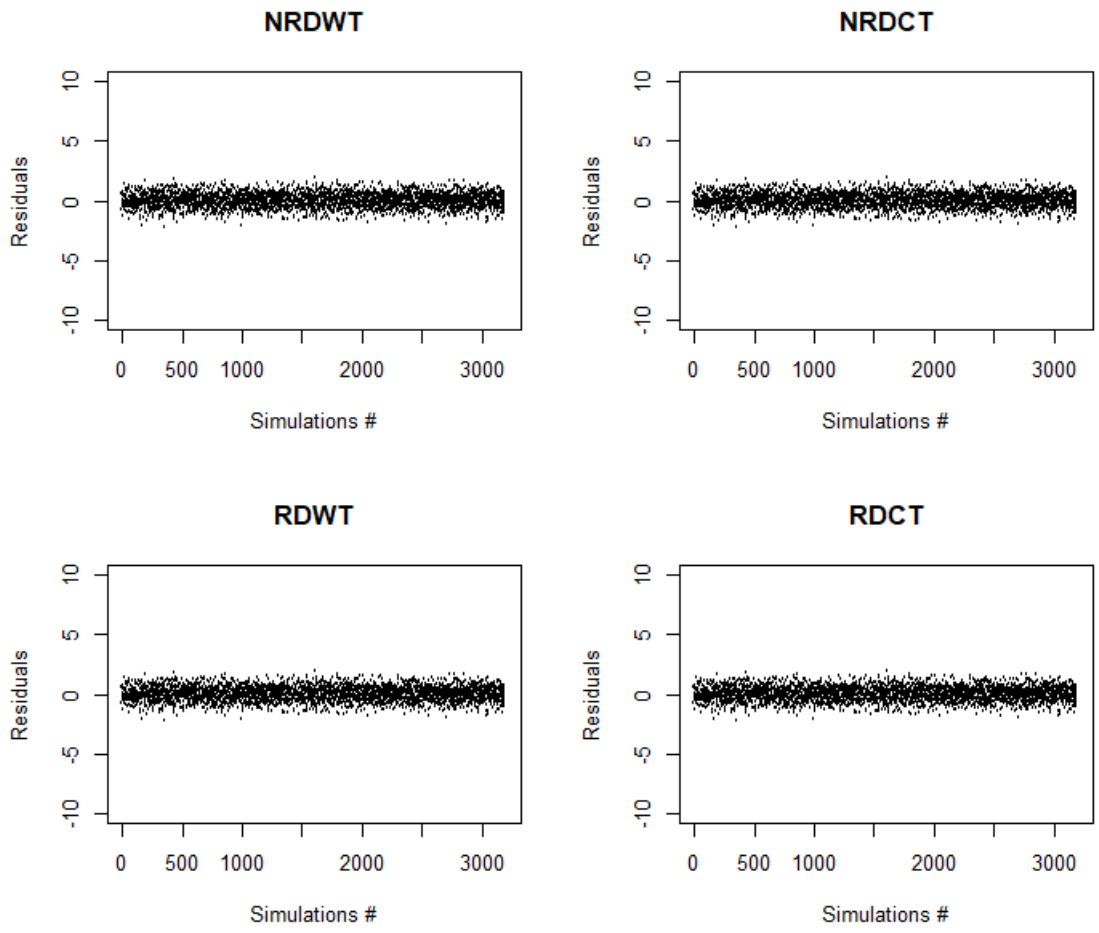

Fig. 4.9 Case A: SP residual plots for simulations no.1-100.

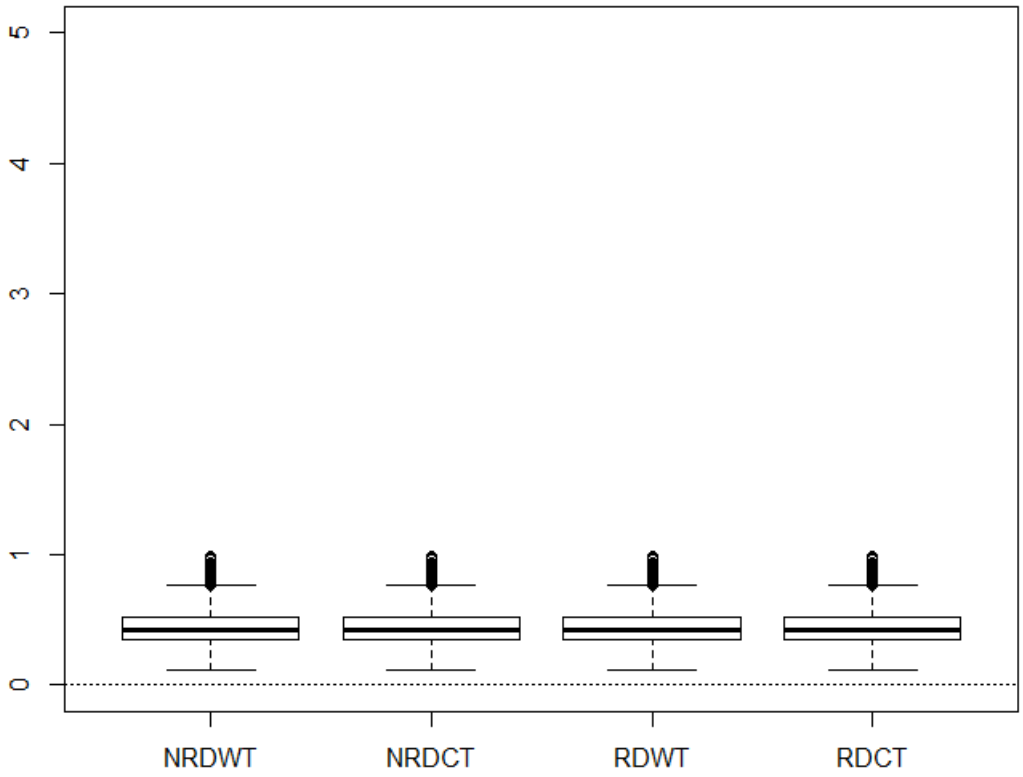

Fig. 4.10 Case A: mean square box-plots for the SP residuals.

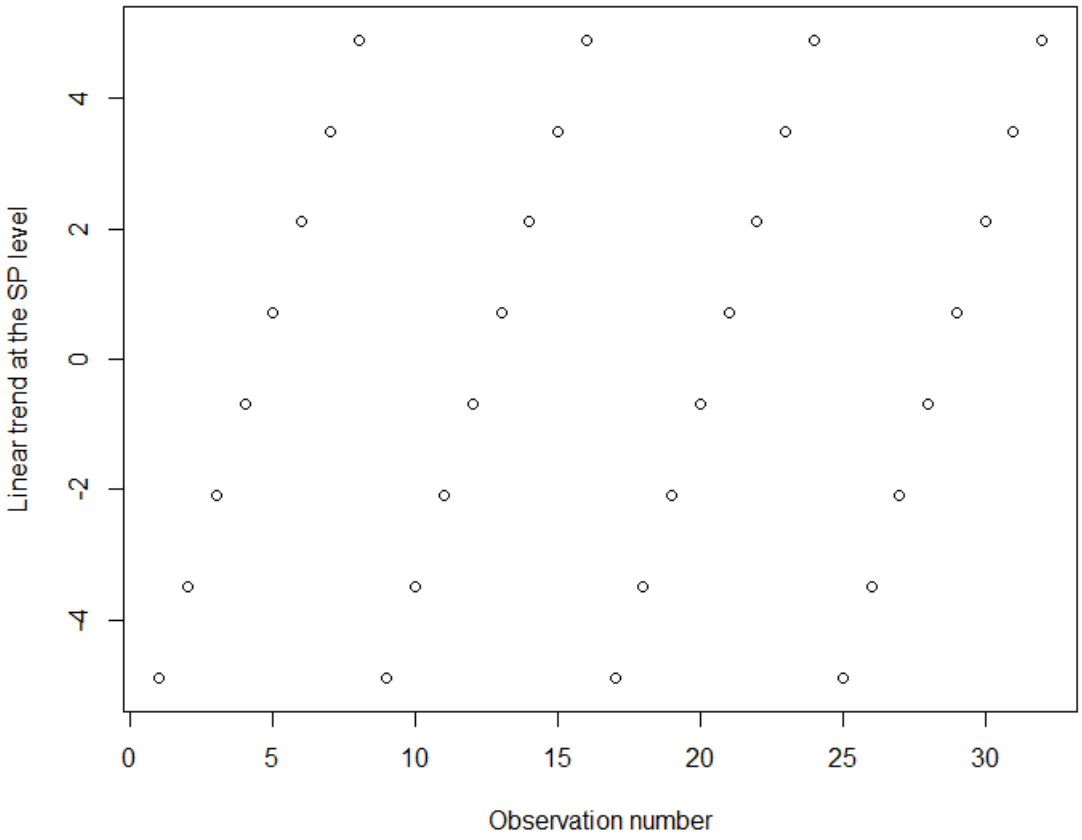

Fig. 4.11 Case B: linear trend at the SP level.

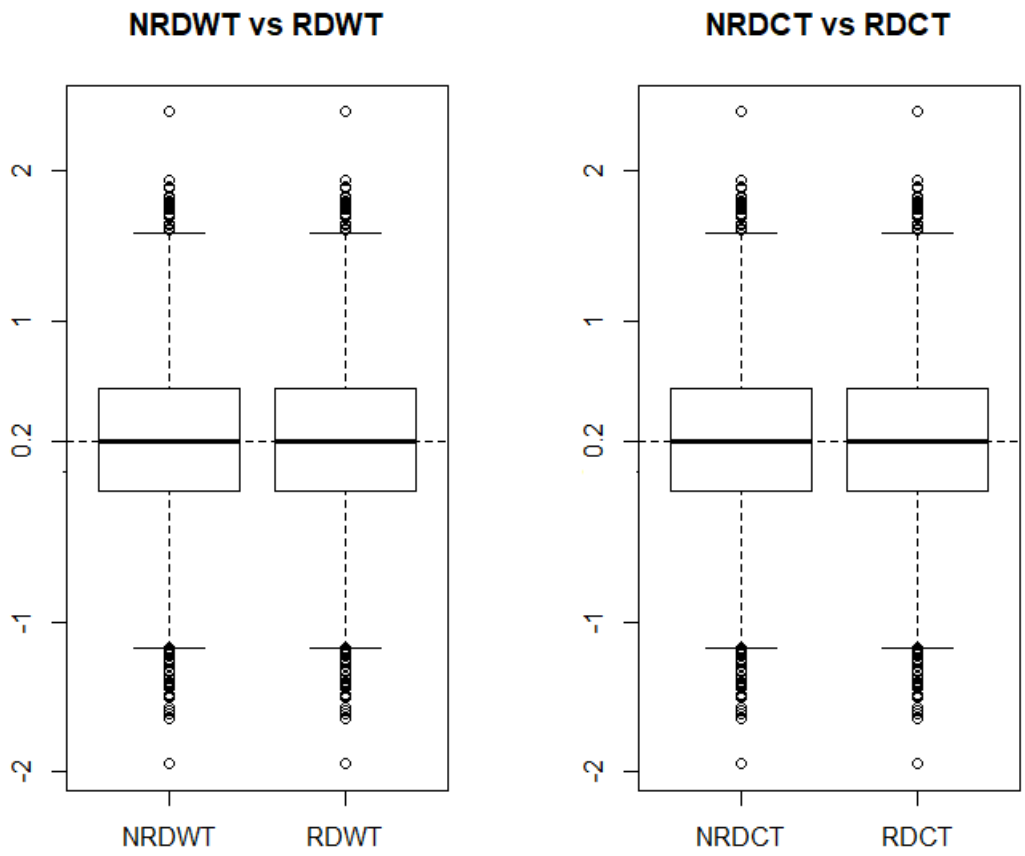

Fig. 4.12 Case B: estimated coefficients box-plots for the blade-spindle-revolution WP factor.

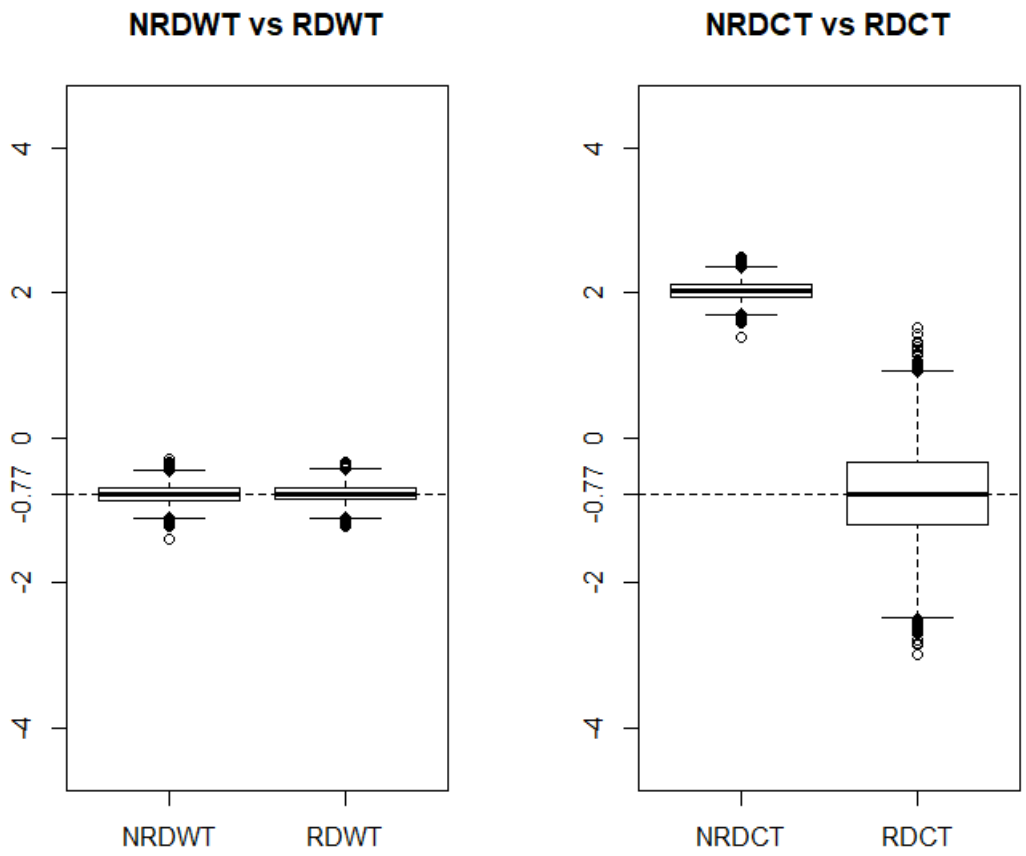

Fig. 4.13 Case B: estimated coefficients box-plots for the dicing machine SP factor.

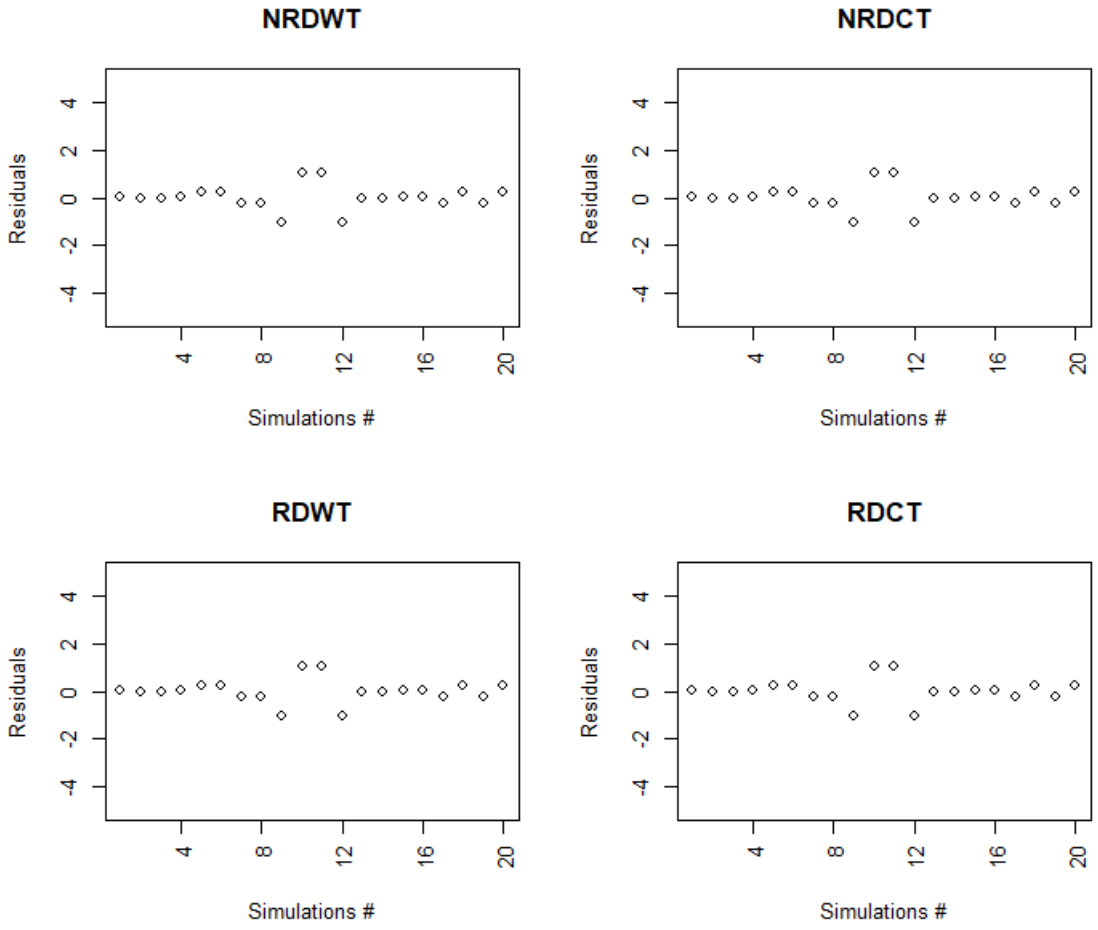

Fig. 4.14 Case B: WP residual plots for simulations no.3-7.

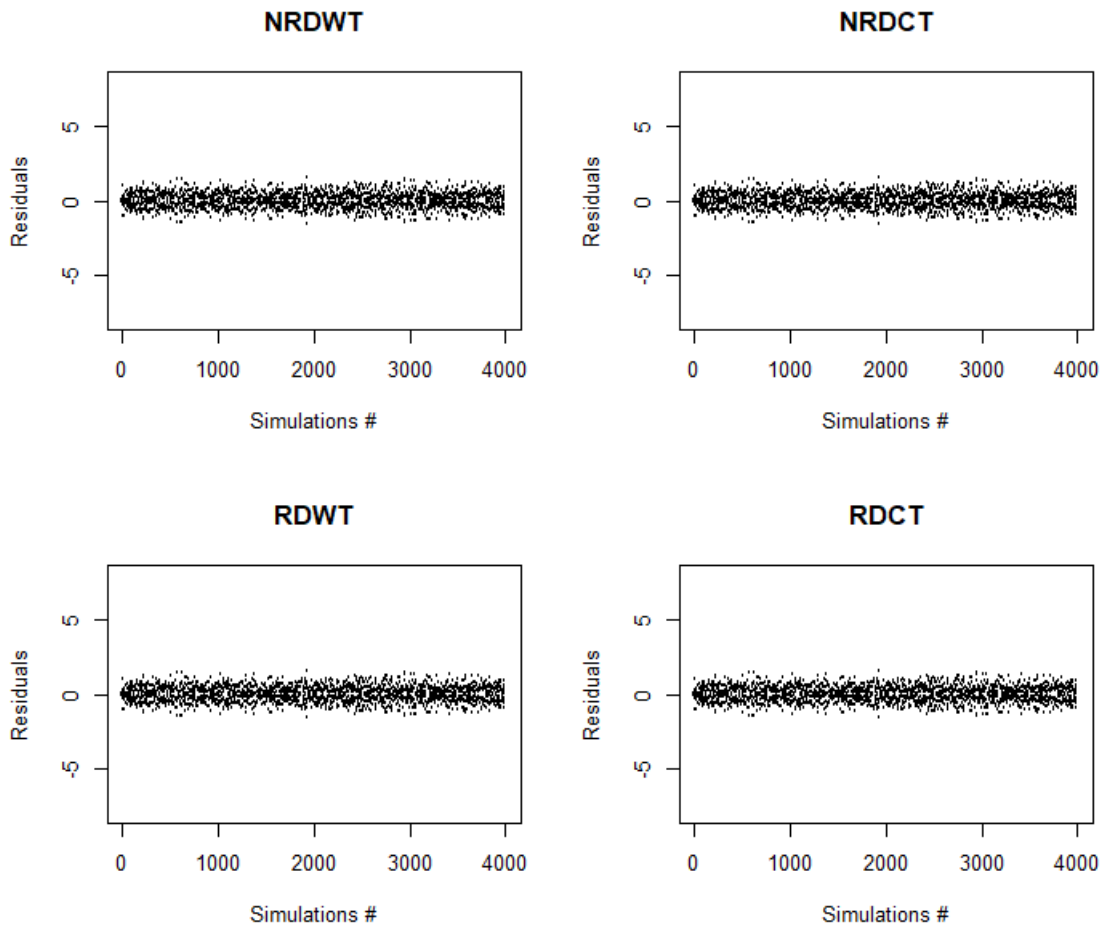

Fig. 4.15 Case B: WP residual plots for simulations no.1-1000.

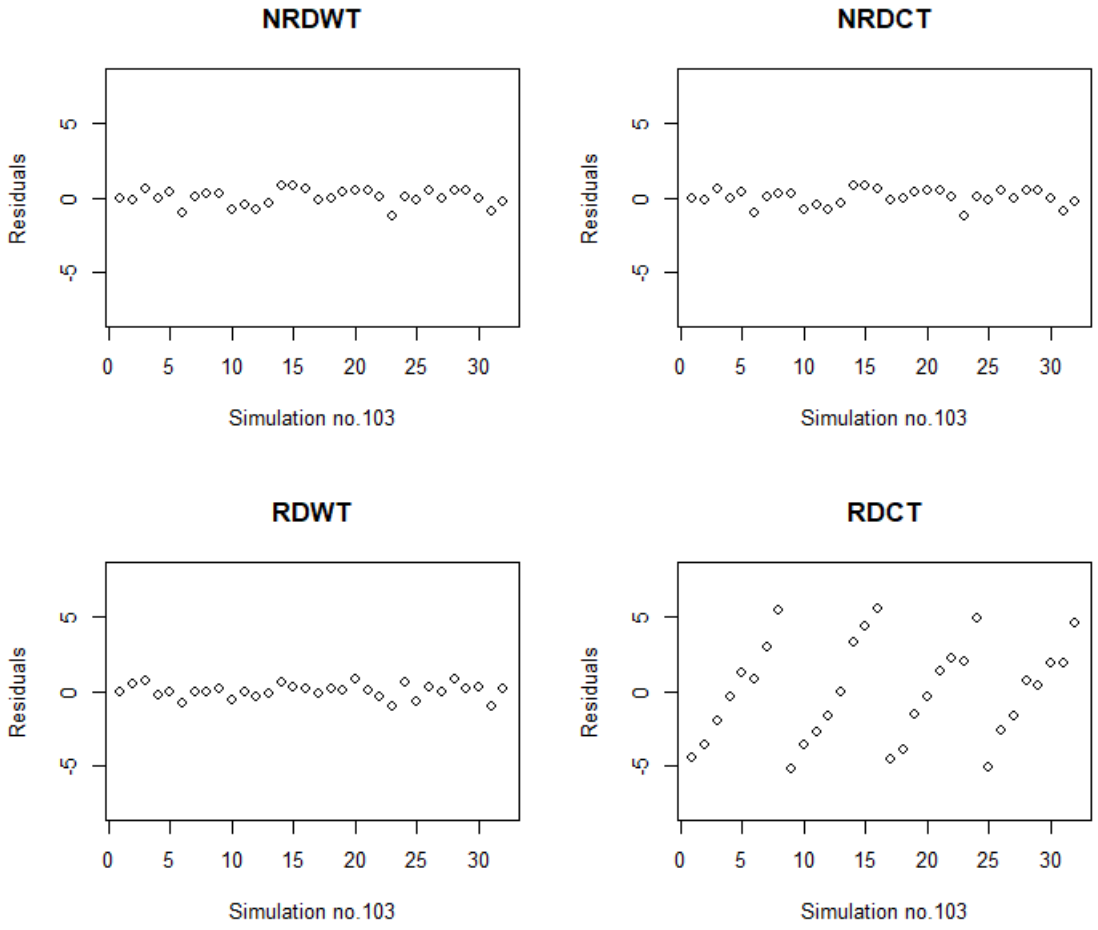

Fig. 4.16 Case B: SP residual plots for simulations no.103.

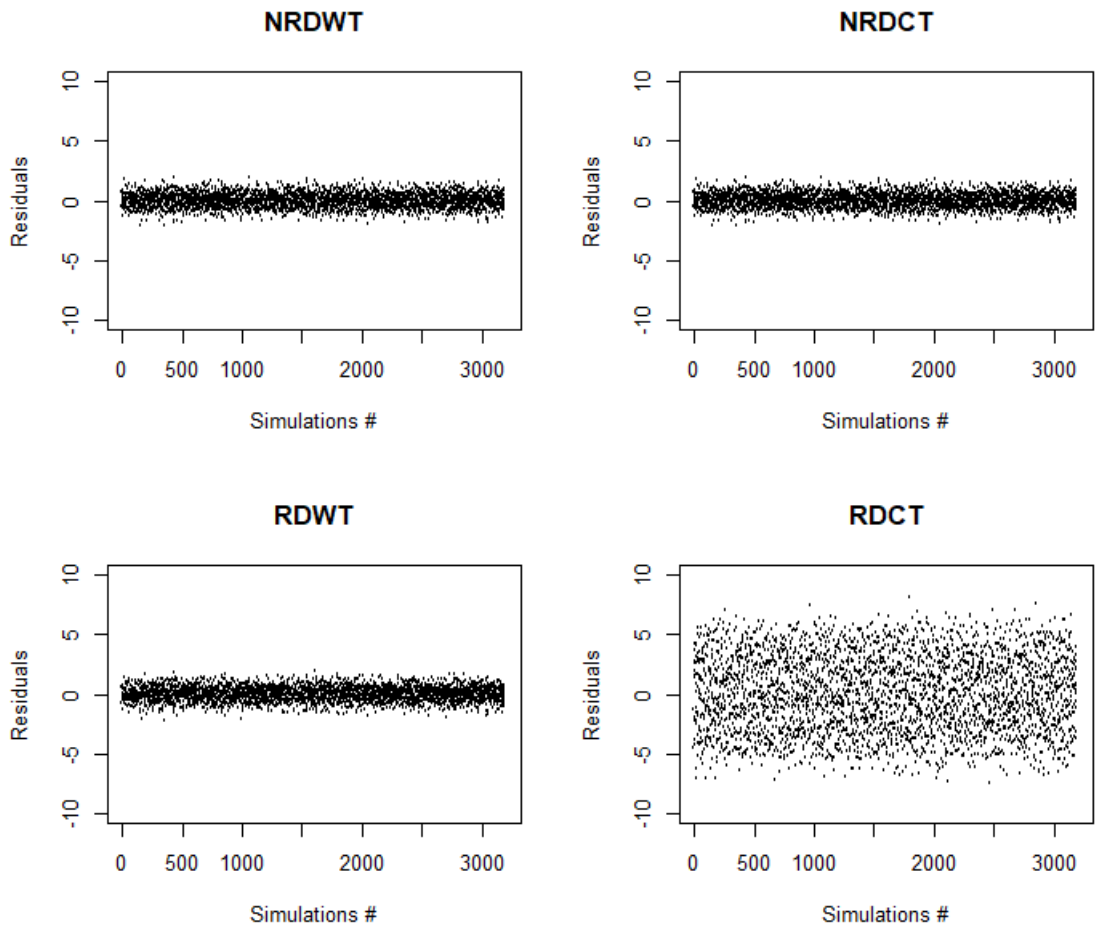

Fig. 4.17 Case B: SP residual plots for simulations no.1-100.

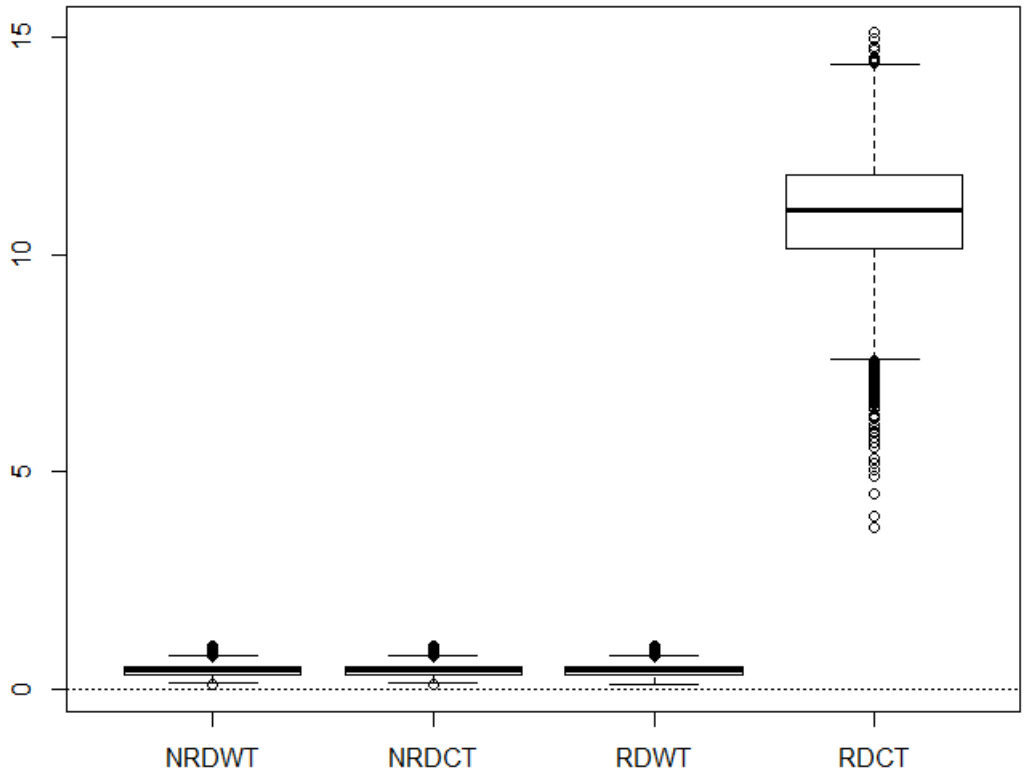

Fig. 4.18 Case B: mean square box-plots for the SP residuals.

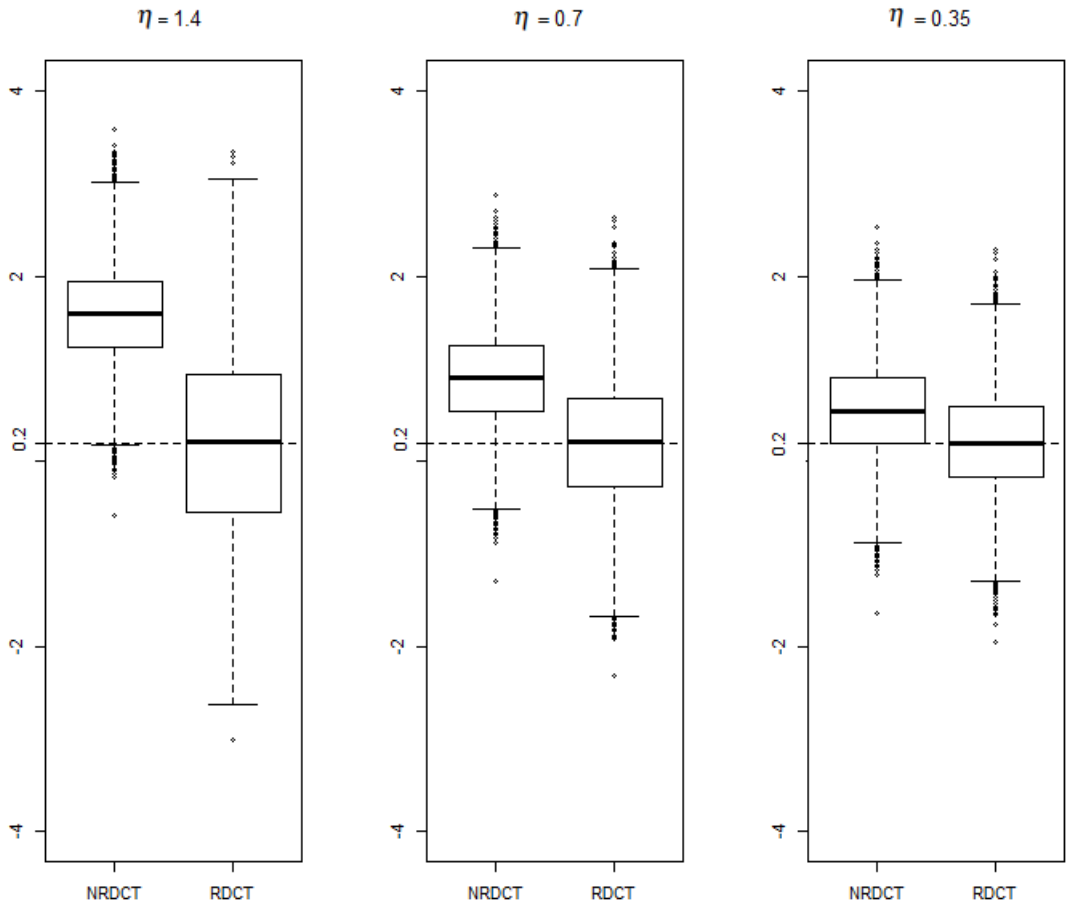

Fig. 4.19 Case A: estimated coefficients box-plots for the blade-spindle-revolution WP factor in terms of different values for η.

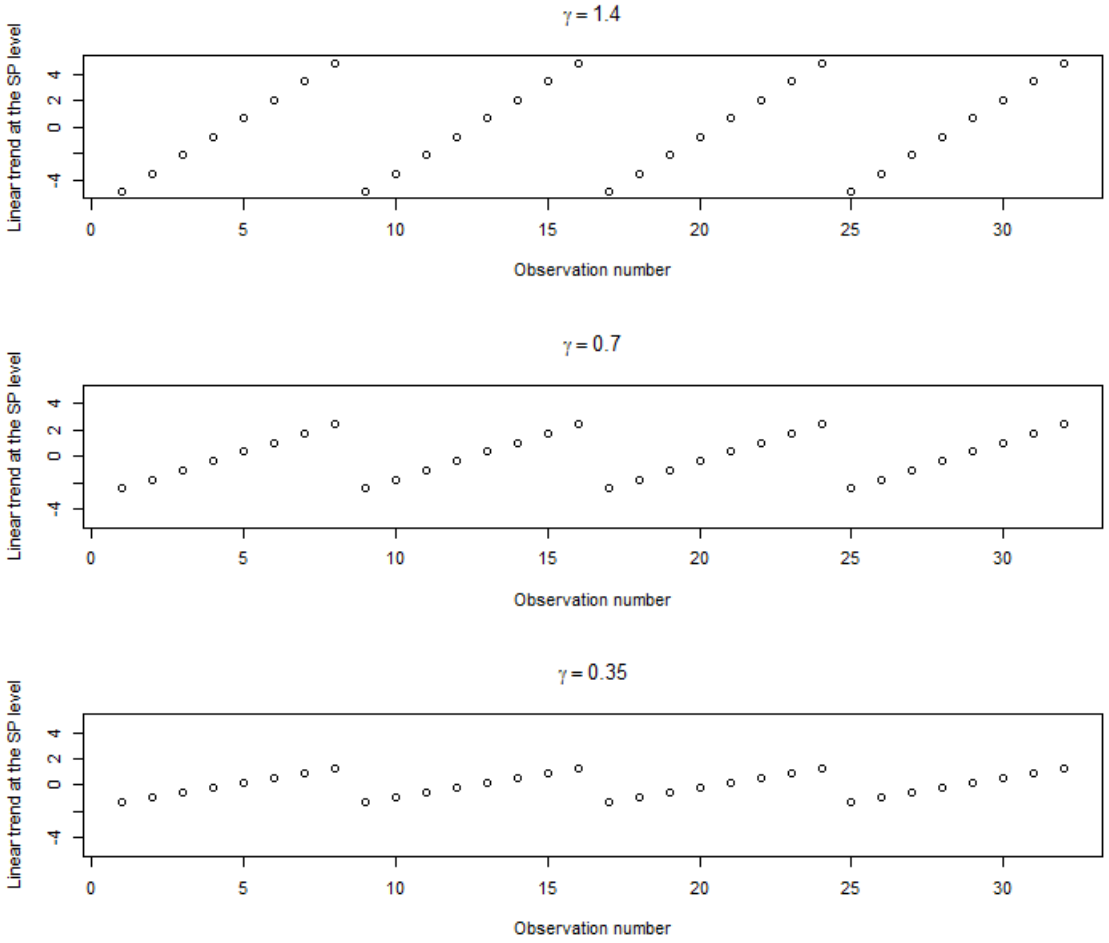

Fig. 4.20 Case B: linear trend at the SP level in terms of different values for γ.

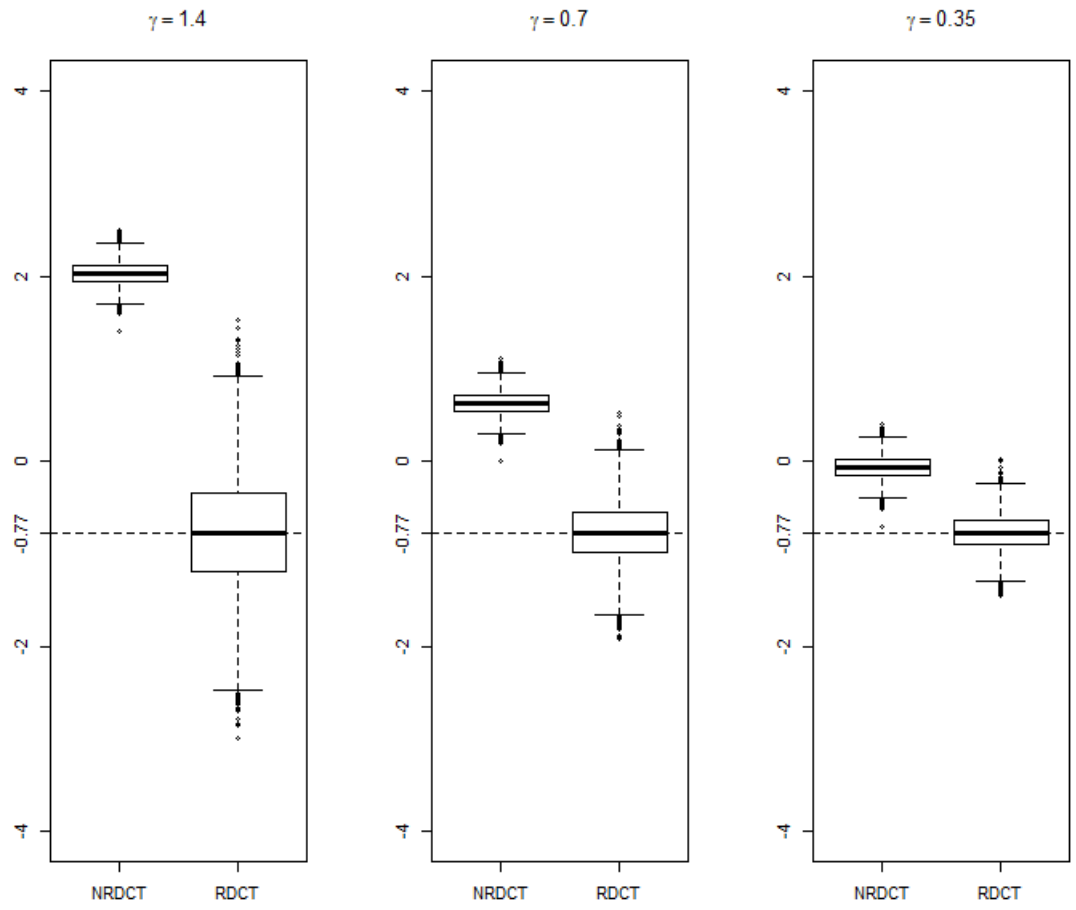

Fig. 4.21 Case B: estimated coefficients box-plots for the dicing machine SP factor in terms of different values for γ.

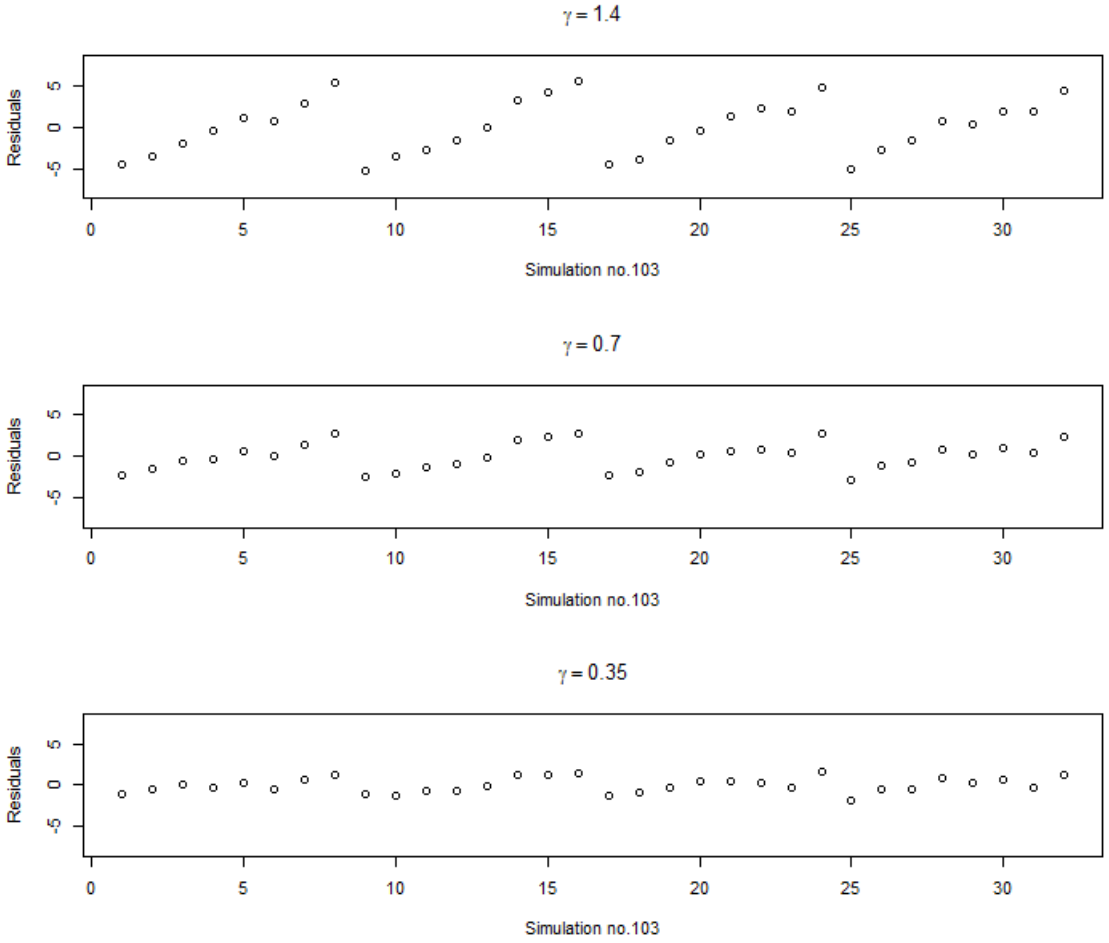

Fig. 4.22 Case B: SP residual plots for simulation no.103 - RDCT data.

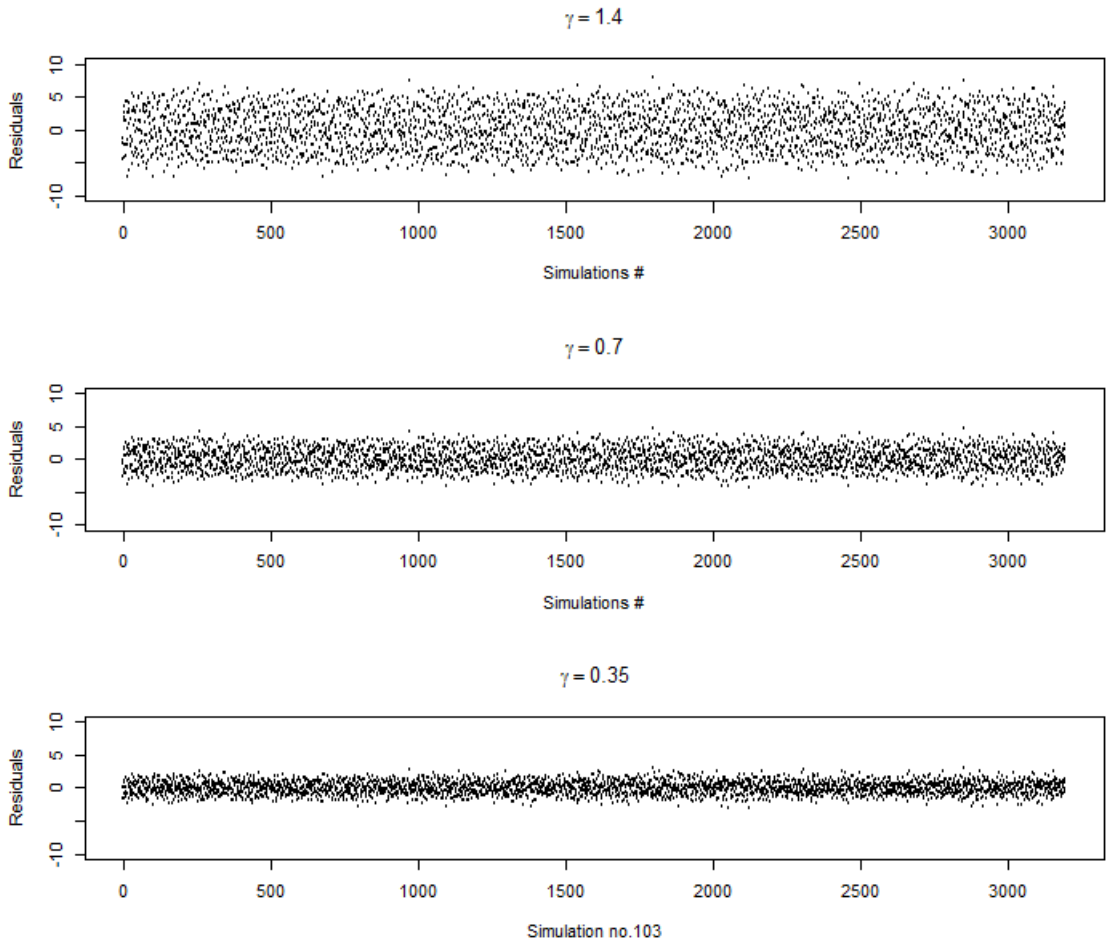

Fig. 4.23 Case B: SP residual plots for 1000 simulations - RDCT data.

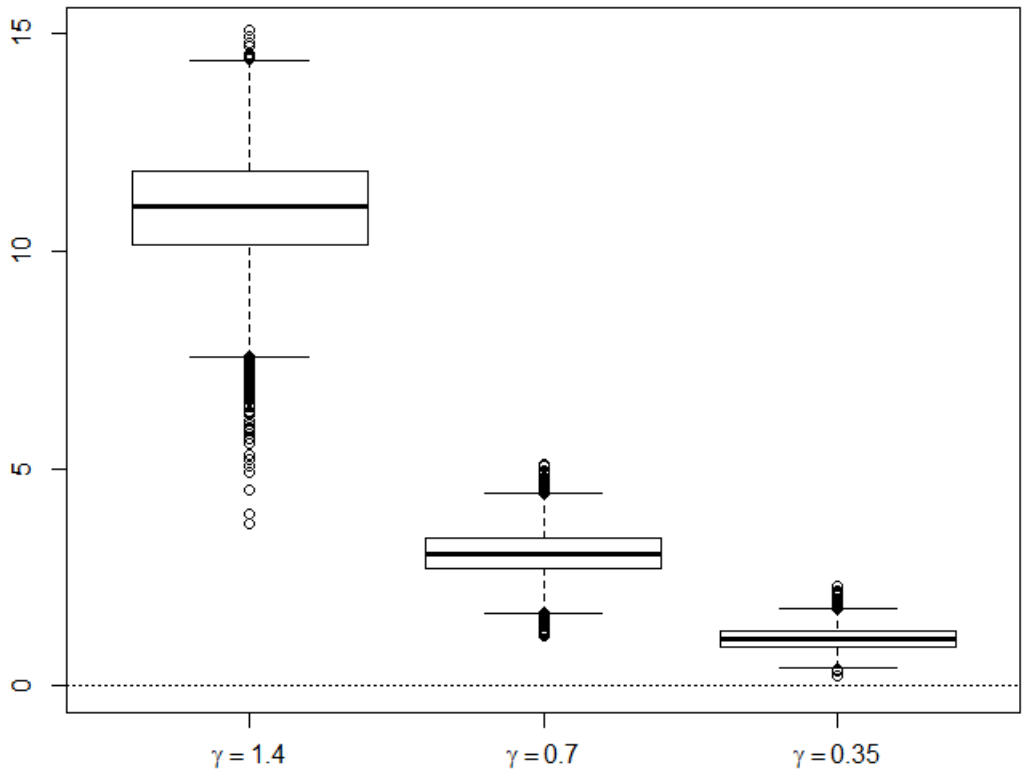

Fig. 4.24 Case B: mean square box-plots for the SP residuals in terms of different slope values - RDCT.

References

[1] Arcidiacono G., Berni R., Cantone L. and Placidoli P. (2017). Kriging models for payload distribution optimization of freight trains. *International Journal Of Production Research*, 55(17): 4878-4890.

[2] Arcidiacono G., Berni R., Cantone L., Nikiforova N.D. and Placidoli P. (2018). Fast method to evaluate payload effect on in-Train forces of freight trains. *The Open Transportation Journal*, 12: 77-87.

[3] Arnouts H. and Goos P. (2017). Analyzing ordinal data from a split-plot design in the presence of a random block effect. Quality Engineering, 29(4): 553-562.

[4] Arora N. and Huber J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments. *Journal of Consumer Research*, 28(2): 273-283.

[5] Atkinson A.C. (2008). Examples of the use of an equivalence theorem in constructing optimum experimental designs for random-effects nonlinear regression models. *Journal of Statistical Planning and Inference*, 138(9): 2595-2606.

[6] Atkinson A.C., Donev A.N. and Tobias R.D. (2007). Optimum Experimental Designs, with SAS. New York: Oxford University Press.

[7] Atkinson A.C. and Woods D.C. (2015). Designs for Generalized Linear Models. In *Handbook of Design and Analysis of Experiments* (eds A. Dean, M. Morris, J. Stufken, D. Bingham), chapter 13, pp. 471-548, Boca Raton: Chapman & Hall.

[8] Bhat C.R. (1995). A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research Part B - Methodological*, 29: 471-483.

[9] Bhat C.R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B*, 35(7): 677-693.

[10] Bingham D. and Sitter R.S. (1999a). Minimum-aberration two-tevel fractional factorial split-plot designs. *Technometrics*, 41(1): 62-70.

[11] Bingham D. and Sitter R.S. (1999b). Some theoretical results for fractional factorial split-plot designs. *The Annals of Statistics*, 27(4): 1240-1255.

[12] Bingham D. and Sitter R.S. (2001). Design issues in fractional factorial split-plot experiments. *Journal of Quality Technology*, 33(1): 2-15.

[13] Bingham D., Sitter R.R. and Tang B. (2009). Orthogonal and nearly orthogonal designs for computer experiments. *Biometrika*, 96(1): 51-65.

[14] Bisgaard S. (2000). The design and analysis of $2^{k-p} \times 2^{q-r}$ split-plot experiments. *Journal of Quality Technology*, 32(1): 39-56.

[15] Box G.E.P. (1999). Statistics as a catalyst to learning by scientific method part II - a discussion. *Journal of Quality Technology*, 31(1): 16–29.

[16] Box G.E.P. and Behnken D.W. (1960). Some new three-level designs for the study of quantitative variables. *Technometrics*, 2(4): 455-475.

[17] Box G.E.P. and Jones S. (1992). Split-plot designs for robust product experimentation. *Journal of Applied Statistics*, 19(1): 3-26.

[18] Box G.E.P. and Jones S. (2001). Split-plots for robust product and process experimentation. *Quality Engineering*, 13(1): 127-134.

[19] Box G.E.P. and Wilson K.B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society Serie B*, 13(1): 195-241.

[20] Boxall P.C. and Adamowicz W.L. (2002). Understanding heterogeneous preferences in random utility models: a latent class approach. *Environmental & Resource Economics*, 23(4): 421–446.

[21] Breslow N.R. and Clayton D.G. (1993). Approximate inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(3): 9-25.

[22] Burgess L. and Street D.J. (2003). Optimal designs for 2^k choice experiments. *Communications in Statistics: Theory and Methods*, 32(11): 2185-2206.

[23] Burgess L. and Street D.J. (2005). Optimal designs for choice experiments with asymmetric attributes. *Journal of Statistical Planning and Inference*, 134(1): 288-301.

[24] Cantone L. (2011). TrainDy: the new Union Internationale des Chemins de Fer software for freight train interoperability. *Journal of Rail and Rapid Transit*, 225(1): 57–70.

[25] Carlsson F. and Martinsson P. (2002). Design techniques for stated preference methods in health economics. *Health Economics*, 12(4): 281-294.

[26] Catelani M., Scarano V.L. and Bertocci F. (2012). Experimental Stress Characterization of a Biomedical Ultrasound Probe Soldered With Innovative Silver Isotropically Conductive Adhesive. *IEEE Transactions on Instrumentation and Measurement*, 61(3): 719-728.

[27] Chaloner K. and Verdinelli I. (1995). Bayesian experimental design: a review. *Statistical Science*, 10(3): 273-304.

[28] Chen R., Hsieh D., Hung Y. and Wang W. (2013). Optimizing Latin hypercube designs by particle swarm. *Statistics and Computing*, 23(5): 664–676.

[29] Cioppa T.M. and Lucas T.W. (2007). Efficient nearly orthogonal and space-filling Latin hypercubes. *Technometrics*, 49(1): 45-55.

[30] Cole C., Spiryagin M. and Wu Q. (2017). Modelling, simulation and applications of longitudinal train dynamics. *Vehicle System Dynamycs*, 55(10): 1498-1571.

[31] Cox D.R. (1958). Planning of Experiments. New York: John Wiley & Sons.

[32] Cox D.R. and Reid N. (2000). The Theory of the Design of Experiments. Boca Raton: Chapman & Hall.

[33] Cuervo P.D., Kessels R., Goos P. and Sörensen K. (2016). An integrated algorithm for the optimal design of stated choice experiments with partial profiles. *Transportation Research Part B*, 93: 648–69.

[34] Currin C., Mitchell T., Morris M. and Ylvisaker D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(12): 953-963.

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

[36] Deng H., Shao W., Ma Y. and Wei Z. (2012). Bayesian metamodelling for computer experiments using Gaussian kriging models. *Quality and Reliability Engineering International Journal*, 28(4): 455-466.

[37] Dette H. and Pepelyshev A. (2010). Generalized latin hypercube design for computer experiments. *Technometrics*, 52(4): 421–429.

[38] Dey A. and Mukerjee R. (1999). Fractional Factorial Plans. New York, NY: John Wiley & Sons.

[39] Draper N.R. and John J.A. (1998). Response surface designs where levels of some factors are difficult to change. *Australian and New Zealand Journal of Statistics*, 40(4): 487-495.

[40] Durrande N., Ginsbourger N. and Roustant O. (2012). Additive covariance kernels for high-dimensional Gaussian process modeling. *Annales de la faculte des Sciences de Toulouse*, 21(3): 481-499.

[41] Fang K.T., Lin D.K.J., Winker P. and Zhang Y. (2000). Uniform design: theory and application. *Technometrics*, 42(3): 237–248.

[42] Federer W.T. and King F. (2007). Variations on Split Plot and Split Block Experiment Designs. Hoboken, NJ: John Wiley & Sons.

[43] Fedorov V.V. (1972). Theory of Optimal Experiments, New York: Academic Press.

[44] Fisher R.A. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd.

[45] Fries A. and Hunter W.G. (1980). Minimum aberration 2^{k-p} designs. *Technometrics*, 22(4): 601-608.

[46] Georgiou S.D. (2009). Orthogonal Latin hypercube designs from generalized orthogonal designs. *Journal of Statistical Planning and Inference*, 139(4): 1530–1540.

[47] Ginsbourger D., Dupuy D., Badea A., Carraro L. and Roustant O. (2012). A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments. *Applied Stochastic Models in Business and Industry*, 25(2): 115-131.

[48] Goos P. (2006). Optimal versus orthogonal and equivalent estimation design of blocked and split-plot experiments. *Statistica Neerlandica*, 60(3): 361-378.

[49] Goos P. and Gilmour S.G. (2012). A general strategy for analyzing data from split-plot and multistratum experimental designs. *Technometrics*, 54(4): 340-354.

[50] Goos P. and Vandebroek M. (2001a). D-optimal response surface designs in the presence of random block effects. *Computational Statistics and Data Analysis*, 37(4): 433-453.

[51] Goos P. and Vandebroek M. (2001b). Optimal split-plot designs. *Journal of Quality Technology*, 33(4): 436-450.

[52] Goos P. and Vandebroek M. (2003). D-optimal split-plot designs with given numbers and sizes of whole plots. *Technometrics*, 45(3): 235-245.

[53] Gramacy R. and Lee H. (2012). Cases for the nugget in modelling computer experiments. Statistics and Computing, 22(3): 713-22.

[54] Grömping U. (2014). R Package FrF2 for creating and analyzing fractional factorial 2-Level designs. *Journal of Statistical Software*, 56(1): 1-56.

[55] Grossmann H., Holling H., Grasshoff U. and Schwabe R. (2006). Optimal designs for asymmetric linear paired comparisons with a profile strength constraint. *Metrika*, 64(1): 109–119.

[56] Grossmann H., Grasshoff U. and Schwabe R. (2013). A catalogue of designs for partial profiles in paired comparison experiments with three groups of factors. *Statistics*, 48(6): 1268-1281.

[57] Johnson M.E., Moore L.M. and Ylvisaker D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2): 131–148.

[58] Joseph V.R., Gul E. and Ba S. (2015). Maximum projection designs for computer experiments. *Biometrika*, 102(2): 371–380.

[59] Han G., Santner T.J., Notz W.I. and Bartel D.L. (2009). Prediction for computer experiments having quantitative and qualitative input variables. *Tecnometrics*, 51(3): 278-288.

[60] He X. and Qian P.Z.G. (2011). Nested orthogonal array-based Latin hypercube designs. *Biometrika*, 98(3): 721—731.

[61] He Y. and Tang B. (2013). Strong orthogonal arrays and associated Latin hypercubes for computer experiments. *Biometrika*, 100(1): 254–260.

[62] He Y. and Tang B. (2014). A characterization of strong orthogonal arrays of strength three. *The Annals of Statistics*, 42(4): 1347-1360.

[63] He Y., Cheng C-S. and Tang B. (2018). Strong orthogonal arrays of strength two plus. *The Annals of Statistics*, 46(2): 457-486.

[64] Hedayat A.S., Sloane N.J.A. and Stufken J. (1999). Orthogonal Arrays: Theory And Applications. New York, NY: Springer.

[65] Henderson T. and Liu Q. (2016). Efficient design and analysis for a selective choice process. *Journal of Marketing Research*, 54(3): 430-446.

[66] Hensher D.A. (1999). HEV choice models as a search engine for the specification of nested logit tree structures. *Marketing Letters*, 10: 339-349.

[67] Hensher D. and Greene W. (2003). A latent class model for discrete choice analysis: contrasts with Mixed Logit. *Transportation Research-Part B*, 37(8): 681–98.

[68] Huang P., Chen D. and Voelkel J.O. (1998). Minimum-aberration two-level split-plot designs. *Technometrics*, 40(4): 314-326.

[69] Huber J. and Zwerina K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, 33(3): 307-317.

[70] Hung Y. (2011). Penalized blind kriging in computer experiments. *Statistica Sinica*, 21(3): 1171-1190.

[71] Hung Y., Joseph V.R. and Melkote S.N. (2015). Analysis of computer experiments with functional response. *Technometrics*, 57(1): 35-44.

[72] Iman R.L. and Conover W.J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics Part B-Simulation and Computation*, 11(3):311–334.

[73] Jensen W.A. and Kowalski S.M. (2012). Response surfaces, blocking and split-plots: an industrial experiment case study. *Quality Engineering*, 24(4): 531-542.

[74] Jones B., Silvestrini R.T., Montgomery D.C. and Steinberg, D.M. (2015). Bridge designs for modeling systems with low noise. *Technometrics*, 57(2): 155-163.

[75] Jones B. and Goos P. (2007). A Candidate-Set-Free algorithm for generating D-optimal split-plot designs. *Journal of the Royal Statistical Society, Series C*, 56(3): 347–364.

[76] Jones, B. and Goos P. (2012). I-optimal versus D-optimal split-plot Response-Surface designs. *Journal of Quality Technology*, 44(2): 85–101.

[77] Jourdan A. and Franco J. (2010). Optimal Latin hypercube designs for the Kullback-Leibler criterion. *Advances in Statistical Analysis*, 94(4): 341–351.

[78] Kessels R., Goos P. and Vanderbroek M. (2006). A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research*, 43(3): 409-419.

[79] Kessels R., Jones B., Goos P. and Vandebroek M. (2008). Recommendations on the use of Bayesian optimal designs for choice experiments. *Quality and Reliability Engineering International*, 24(6): 737–744.

[80] Kessels R., Jones B., Goos P. and Vandebroek M. (2009). An efficient algorithm for constructing Bayesian optimal choice designs. *Journal of Business and Economic Statistics*, 27(2): 279-291.

[81] Kessels R., Jones B. and Goos P. (2011). Bayesian optimal designs for discrete choice experiments with partial profiles. *Journal of Choice Modelling*, 4(3): 52–74.

[82] Kessels R., Jones B. and Goos P. (2014). An improved two-stage variance balance approach for constructing partial profile designs for discrete choice experiments. *Applied Stochastic Models in Business and Industry*, 31(5): 626-648.

[83] Kiefer J. (1959). Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society Serie B*, 21: 272–319.

[84] Kiefer J. and Wolfowitz J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12(3): 363-366.

[85] Krige D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52(9): 119–139.

[86] Kuhfeld W.F., Tobias R.D. and Garratt M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, 31(4): 545-557.

[87] Kulahci M. and Bisgaard S. (2005). The use of Placket-Burman designs to construct split-plot designs. *Technometrics*, 47(4): 495-501.

[88] Kulahci M. and Tyssedal J. (2017). Split-plot designs for multistage experimentation. *Journal of Applied Statistics*, 44(3): 493-510.

[89] Lawrence K.M. (1996). A combinatorial characterization of (t, m,s)-nets in base b. *Journal of Combinatorial Design*, 4(4): 275–93.

[90] Lazari A.G. and Anderson D.A. (1994). Designs of discrete choice experiments for estimating both attribute and availability cross effects. *Journal of Marketing Research*, 31(3): 375-383.

[91] Letsinger J.D., Myers R.H. and Lentner M. (1996). Response surface methods for bi-randomization structures. *Journal of Quality Technology*, 28(4): 381-397.

[92] Li R. and Sudjianto A. (2005). Analysis of computer experiments using penalized likelihhod in Gaussian kriging models. *Technometrics*, 47(2): 111-120.

[93] Li J. and Qian P.Z.G. (2013). Construction of nested (nearly) orthogonal designs for computer experiments. *Statistica Sinica*, 23(1): 451-466.

[94] Lin C.D., Mukerjee R. and Tang B. (2009). Construction of orthogonal and nearly and orthogonal Latin hypercube designs. *Biometrika*, 96(1): 243-247.

[95] Lin C.D and Tang B. (2015). Latin hypercubes and space-filling designs. In Handbook of Design and Analysis of Experiments, Dean A, Morris M, Stufken J, Bingham D (eds). Chapman & Hall/ CRC Press: Boca Raton, USA, pp.: 593-625.

[96] Liu H. and Liu M-Q. (2015). Column-orthogonal strong orthogonal arrays and sliced strong orthogonal arrays. *Statistica Sinica*, 25(4): 1713-1734.

[97] Liu Q. and Tang Y. (2015). Construction of heterogeneous conjoint choice designs: a new approach. *Marketing Science*, 34(3): 346-366.

[98] Lombardi G.V., Berni R. and Rocchi B. (2017). Environmental friendly food. Choice Experiment to assess consumer's attitude toward climate neutral milk: the role of communication. *Journal of Cleaner Production*, 142(1): 257-262.

[99] Louviere J.J. and Woodworth G. (1983). Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *Journal of Marketing Research*, 20(4): 350-367.

[100] Macharia H. and Goos P. (2010). D-optimal and D-efficient equivalent-estimation second-order split-plot designs. *Journal of Quality Technology*, 42(4): 358–372.

[101] MacKay D.J.C. (1998). Introduction to Gaussian Processes. In Bishop CM, editor, Neural Networks and Machine Learning. Springer-Verlag.

[102] Marathon (2011). Accessed October 30, 2015, http://www.marathon-project.eu/.

[103] Masi C., Dinnella C., Barnabà M., Navarini L. and Monteleone E. (2013). Sensory properties of under-roasted coffee beverages. *Journal of Food Science*, 78(8): 1290-1300.

[104] Matheron G. (1962). Traité de géostatistique appliquée, Mém BRGM 14. éditions Technip, Paris.

[105] Matheron G. (1967). Kriging or polynomial interpolation procedures? *Canadian Institute of Mining and Metallurgy Transactions*, 70: 240–242.

[106] Matheron G. (1971). The theory of regionalized variables and its applications. Ecole Nationale Superieure des Mines.

[107] McFadden D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.) Frontiers in Econometrics. New York: Academic Press, pp: 105-142.

[108] McFadden D. and Train K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5): 447-470.

[109] McKay M., Beckman R.J. and Conover W.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2): 239–245.

[110] Meyer R.K. and Nachtsheim C.J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37(1): 60-69.

[111] Miller A. (1997). Strip-plot configurations of fractional-factorials. *Technometrics*, 39(2): 153-161.

[112] Montgomery D.C. (2013). Design and Analysis of Experiments. John Wiley & Sons. 8th edition, USA.

[113] Moon H., Dean A.M. and Santner T.J. (2011). Algorithms for generating maximin orthogonal and Latin hypercube designs. *Journal of Statistical Theory and Practice*, 5(1): 81–98.

[114] Morris M.D. and Mitchell T.J. (1995). Exploratory designs for computer experiments. *Journal of Statistical Planning and Inference*, 43(3): 381–402.

[115] Mukerjee R. and Wu C.F.J. (2006) A Modern Theory of Factorial Designs, Springer Verlag, New York.

[116] Myers R.H. and Montgomery D.C. (2002). Process and Product Optimization Using Designed of Experiments, 2nd edition, New York, NY: John Wiley & Sons.

[117] Myers R.H., Montgomery D.C. and Anderson-Cook C.M. (2009). Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley: New York, USA.

[118] Myers R.H., Khuri A.I. and Vining G.G. (1992). Response Surface alternatives to the Taguchi robust parameter design approach. *The American Statistician*, 46(2): 131-139.

[119] Mylona K., Goos P. and Macharia H. (2013). Three-level equivalent-estimation split-plot designs based on subset and supplementary difference set designs. *IIE Transactions*, 45(11): 1153–1165.

[120] Mylona K., Goos P. and Jones B. (2014). Optimal design of blocked and split-plot experiments for fixed effects and variance component estimation. *Technometrics*, 56(2): 132–144.

[121] Niederreiter H. (1992). Random number generation and Quasi-Monte Carlo methods, SIAM CBMSNSF Regional Conference Series in Applied Mathematics. Philadelphia, PA.

[122] Niederreiter H. (2008). Nets,(t, s)-sequences, and codes. *Monte Carlo and Quasi-Monte Carlo Methods*, 2006(1): 83–100.

[123] OECD (2008). *Handbook on Constructing Composite Indicators. Methodology and user guide*. Paris: OECD Publications.

[124] Owen A.B. (1992). A central limit theorem for Latin hypercube sampling. *Journal of the Royal Statistical Society Series B*, 54(2): 541–551.

[125] Pang F., Liu M.Q. and Lin, D.K.J. (2009). A construction method for orthogonal Latin hypercube designs with prime power levels. *Statistica Sinica*, 19(4): 1721-1728.

[126] Park J-S. (1994). Optimal Latin-hypercube designs for computer experiments. *Journal of Statistical Planning and Inference*, 39(1): 95-111.

[127] Parker P.A., Kowalski S.M. and Vining G.G. (2007). Construction of balanced equivalent estimation second-order split-plot designs. *Technometrics*, 49(1): 56-65.

[128] Peng C.-Y. and Wu C.F.J. (2014). On the choice of the nugget in Kriging modeling for deterministic computer experiments. *Journal of Computational and Graphical Statistics*, 23(1): 151-168.

[129] Pistone G. and Vicario G. (2010). Comparing and generating Latin Hypercube designs in Kriging models. *Advances in Statistical Analysis*, 94(4): 353–366.

[130] Pistone G. and Vicario G. (2013). Kriging prediction from a circular grid: application to wafer diffusion. *Applied Stochastic Models in Business and Industry*, 29(4): 350-361.

[131] Pronzato L. and Müller W. (2012). Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3): 681-701.

[132] Qian P.Z.G., Wu H. and Wu J.C.F. (2008). Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics*, 50(3): 383-396.

[133] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[134] Rasmussen C.E. and Williams C.K.I. (2006) Gaussian Processes for Machine Learning. Massachusetts Institute of Technology: the MIT Press.

[135] Revelt D. and Train K. (1998). Mixed Logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics*, 80 (4): 647-657.

[136] Robinson T.J., Myers R.H. and Montgomery D.C. (2004). Analysis considerations in industrial split-plot experiments with non-normal responses. *Journal of Quality Technology*, 36(2): 180-192.

[137] Robinson T.J., Pintar A.L., Anderson-Cook C.M. and Hamada M.S. (2012). A Bayesian approach to the analysis of split-plot combined and product arrays and optimization in robust parameter design. *Journal of Quality Technology*, 44(4): 304-320.

[138] Roustant O., Joucla J. and Probst P. (2010). Kriging as an alternative for a more precise analysis of output parameters in nuclear safetyLarge break LOCA calculation. *Applied Stochastic Models in Business and Industry*, 26(5): 565-576.

[139] Roustant O., Ginsbourger D. and Deville Y. (2012). DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by Kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1): 1–55.

[140] Sacks J., Welch W.J., Mitchell T.J. and Wynn H.P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4): 409–423.

[141] Sambo F., Borrotti M. and Mylona K. (2014). A coordinate exchange two-phase local search algorithm for the D- and I-optimal designs of split-plot experiments. *Computational Statistics & Data Analysis*, 71(3): 1193–1207.

[142] Sándor Z. and Wedel M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38(4): 430-444.

[143] Sándor Z. and Wedel M. (2002). Profile construction in experimental choice designs for mixed logit models. *Marketing Science*, 21(4): 455-475.

[144] Sándor Z. and Wedel M. (2005). Heterogeneous conjoint choice designs. *Journal of Marketing Research*, 42(2): 210-218.

[145] Sándor Z. (2013). Monte Carlo simulation in Random Coefficient Logit models involving large sums. *Acta Universitatis Sapientiae, Economics and Business*, 1(1): 85-108.

[146] Sang H. and Huang J.Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 74(1): 111-132.

[147] SAS Institute Inc. (2013). SAS Version 9.4. SAS Institute Inc.: Cary, NC.

[148] Shewry M.C. and Wynn H.P. (1987). Maximum Entropy Sampling. *Journal of Applied Statistics*, 14(2): 165-170.

[149] Simpson J.R., Kowalski S.M. and Landman D. (2004). Experimentation with randomization restrictions: targeting practical implementation. *Quality and Reliability Engineering International*, 20(5): 481-49.

[150] Spicci L. (2012). FEM simulation for "pulse-echo" performances of an ultrasound imaging linear probe. Proceedings of the 2012 COMSOL Conference.

[151] Stein M.L. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2): 143–151.

[152] Stein M.L. (1999). Interpolation of Spatial Data. New York, NY: Springer.

[153] Steinberg D.M. and Lin D.K.J. (2006). A construction method for orthogonal Latin hypercube designs. *Biometrika*, 93(2): 279-288.

[154] Street D.J., Bunch D.S. and Moore B.J. (2001). Optimal designs for 2^k paired comparison experiments. *Communications in Statistics: Theory and Methods*, 30(10): 2149-2171.

[155] Tan M.H.Y. and Wu J.C.F. (2013). A Bayesian approach for model selection in fractionated split-plot experiments with applications in robust parameter design. *Technometrics*, 55(3): 359-372.

[156] Tang B. (1993). Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association*, 88(12): 1392—1397.

[157] Thurstone L.L. (1927). A law of comparative judgment. *Psychological Review*, 34: 273-286.

[158] Tian T. and Yang M. (2017) Efficiency of the coordinate-exchange algorithm in constructing exact optimal discrete choice experiments. *Journal of Statistical Theory and Practice*, 11(2), 254-268.

[159] Toubia O., Hauser J.R. and Simester D.I. (2004). Polyhedral methods for adaptive choice-based conjoint analysis. *Journal of Marketing Research*, 41(1): 116-131.

[160] Toubia O. and Hauser J.R. (2007). On managerially efficient experimental designs. *Marketing Science*, 26(6): 851–858.

[161] Train K. (2000) Halton sequences for mixed logit. Economics Working Papers, E00278, University of California, Berkeley.

[162] Train K. (2003). Discrete Choice Methods with Simulation. New York: Cambridge University Press.

[163] Trinca L.A. and Gilmour S.G. (2001). Multistratum Response Surface designs. *Technometrics*, 43(1): 25-33.

[164] Tyssedal J., Kulahci M. and Bisgaard S. (2011). Split-plot designs with mirror image pairs as sub-plots. *Journal of Statistical Planning and Infererence*, 41(12): 3686-3696.

[165] UIC 421 OR. (2012). *Rules of the consist and braking of international freight trains*, 9th edition.

[166] UIC Leaflet 544-1. (2014). *Brakes Braking performance*. 6th edition.

[167] van Dam E.R., Husslage B., den Hertog D. and Melissen H. (2007). Maximin Latin hypercube designs in two dimensions. *Operations Research*, 55(1): 158–169.

[168] Vicario G., Craparotta G. and Pistone G. (2016). Meta-models in computer experiments: Kriging versus artificial neural networks. *Quality and Reliability Engineering International*, 32(6):2055-2065.

[169] Vicario G., Giraudo M.T and Calí V. (2018). Kriging modelization in predicting metal sheet elongation. *Quality and Reliability Engineering International*, 34(7): 1390-1399.

[170] Vining G.G. and Myers R.H. (1990). Combining Taguchi and Response Surface philosophies: a Dual Response Approach. *Journal of Quality Technology*, 22(1): 38-45.

[171] Vining G.G., Kowalski S.M. and Montgomery D.C. (2005). Response surface designs within a split-plot structure. *Journal of Quality Technology*, 37(2): 115-129.

[172] Vining G.G. and Kowalski S.M. (2008). Exact inference for response surface designs within a split-plot structure. *Journal of Quality Technology*, 40(4): 394-406.

[173] Wang L., Kowalski S.M. and Vining G.G. (2009). Orthogonal blocking of response surface split-plot designs. *Journal of Applied Statistics*, 36(3): 303-321.

[174] Wen C. and Koppelman F. (2001). A Generalized Nested Logit model. *Transportation Research Part B: Methodological*, 35(7): 627–641.

[175] Whittle P. (1973). Some general points in the theory and construction of D-optimum experimental designs. *Journal of the Royal Statistical Society - Series B*, 35(2): 123-130.

[176] Wolfinger R. and O'Connell M. (1993). Generalized Linear Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation*, 48(3-4):233-243.

[177] Wooding W.M. (1973). The Split-Plot Design. *Journal of Quality Technology*, 5(1): 16–33.

[178] Wu J.C.F. and Hamada M.S. (2009). Experiments: Planning, Analysis, and Optimization. John Wiley & Sons. New York, USA.

[179] Wu Q., Spiryagin M. and Cole C. (2018). International benchmarking of longitudinal train dynamics simulators: results. *Vehicle System Dynamycs*, 56(3): 343-365.

[180] Wynn H.P. (1970). The sequential generation of D-optimal experimental designs. *Annals of Mathematical Statistics*, 41(5): 1055-1064.

[181] Yang J.Y., Liu M.Q. and Lin D.K.J. (2014). Construction of nested orthogonal Latin hypercube designs. *Statistica Sinica*, 24(1): 211-219.

[182] Yang X., Yang J.F., Lin D.K.J. and Liu M.Q. (2016). A new class of nested orthogonal Latin hypercube designs. *Statistica Sinica*, 26(3): 1249-1267.

[183] Yang J.F., Lin C.D., Qian P.Z.G. and Lin D.K.J. (2013). Construction of sliced orthogonal Latin hypercube designs. *Statistica Sinica*, 23(3); 1117-1130.

[184] Yang M., Biedermann S. and Tang E. (2013). On optimal designs for nonlinear models: a general and efficient algorithm. *Journal of the American Statistical Association*, 108(6): 1411-1420.

[185] Yang J.Y., Chen H., Lin D.K.J. and Liu M.Q. (2016). Construction of sliced maximin-orthogonal Latin hypercube designs. *Statistica Sinica*, 26: 589–603

[186] Ye K. Q., Li W. and Sudjianto A. (2000). Algorithmic construction of optimal symmetric Latin hypercube designs. *Journal of Statistical Planning and Inference*, 90(1): 145–159.

[187] Ye K.Q. (1998). Orthogonal column Latin hypercubes and their application in computer experiments. *Journal of the American Statistical Association*, 93(12): 1430-1439.

[188] Yin Y., Lin D.K.J. and Liu M.Q. (2014). Sliced Latin Hypercube designs via orthogonal arrays. *Journal of Statistical Planning and Inference*, 149(6): 162–171.

[189] Yu J., Goos P. and Vanderbroek M. (2009). Efficient conjoint choice designs in the presence of respondent heterogeneity. *Marketing Science*, 28(1):122-135.

[190] Yu J., Goos P. and Vanderbroek M. (2011). Individually adapted sequential Bayesian designs for conjoint choice experiments. *International Journal of Research in Marketing*, 28(4): 378-388.

[191] Zhang W., Mandal A. and Stufken J. (2017). Approximations of the information matrix for a panel mixed logit model. *Journal of Statistical Theory and Practice*, 11(2): 269-295.

[192] Zhou Q., Qian P.Z.G. and Zhou S. (2011). A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics*, 53(3): 266-273.