

Report on the dissertation of Mr. Wojciech Wójciak

*Title: Multi-Domain Optimum Sample Allocation with Controlled-Precision
under Upper-Bound Constraints*

Relevance and topicality of the topic:

Since the seminal work of Neyman (1934) and Tschuprov (1923), optimization has become popular in survey sampling. The aim was to allocate the given total sample within groups such that the variance of the respective estimator is minimized. The approach was further developed within 100 years to account for various aims, covering costs, different kinds of constraints, integrality, and others. Within the past years, more attention has been laid on numerical properties of algorithms for solving the different problems beyond the classical optimal allocation. Additionally, (numerical) algorithms were developed to calculate the optimal allocation.

Aim of the dissertation:

The aim of the dissertation was twofold. First, a novel sufficient condition for optimality in the multi-domain optimal allocation problem under upper-bound constraints. Second, the author should also develop an algorithm to compute the optimal allocation in the multi-domain context.

Contents and structure of the dissertation:

The dissertation is structured in seven chapters, the introduction, the problem formulation, the mathematical justification of the problem, three chapters on the optimization algorithm, and a summary with outlook. Additionally, seven chapters as appendix and a bibliography are added.

In Chapter one, the author briefly introduces optimum allocation in survey sampling with a focus on disjoint domains (each stratified independently). The aim is to allocate an overall given sample size to the domains such that the weighted (weight controlled, κ_d) domain-specific relative variances of the corresponding Horvitz-Thompson estimators are minimized simultaneously, where the stratum-specific sample sizes do not exceed the stratum-specific upper-bound constraints. Hence, the problem to be solved is an extension of the classical Neyman-Tschuprov problem on domain-level using single-sided constraints.

Within this chapter, a literature overview of the topic is given, which covers a wide range of related problems. However, very surprisingly, the author does not give Jerzey Neyman the credit as one of two originators of the underlying optimization problem.

In Equation (1.11), the author specifies the relationship between the domain specific variances and the overall variance. The relaxation of the sample sizes to continuous variables is justified via empirical studies from the literature. The final optimum allocation problem is formalized as the nonlinear Problem 1.1.1. By construction this is solvable, which is proven in Proposition 3.1.5. The two major contributions are (1) a new type of sufficient optimality conditions given in Theorem 3.2.8 and (2) the solving algorithm RDCA available via `rdca()` from the R package `stratallo` (CRAN).

The chapter concludes with related work and an extended motivation. One should point out that one major argument of emphasizing an exact solution is motivated by possible lacks in precision of numerical methods.

In Chapter two, the fundamental concepts and tools required to examine the input data for Problem 1.1.1 through an abstract model, which is illustrated with carefully selected examples, are introduced. This framework provides the foundation for a mathematically rigorous analysis of Problem 1.1.1 and consequently the RDCA. The notation is comprehensive, well structured, and is facilitating algebraic manipulations throughout the subsequent analysis.

The very formal way of the chapter seems, on the first sight, counterintuitive, at least for survey statisticians. However, finally it turns out that this is very straightforward for the rest of the thesis.

The controlled-precision domain allocation problem is presented in Chapter three. Independent from the input data, the Problem 1.1.1. is feasible (Proposition 3.1.4). This is followed by the existence of an optimal solution to the problem in Proposition 3.1.5.. The relaxed Problem 3.2.1 is introduced which is equivalent to Problem 1.1.1. in the sense that they yield the same optimal upper bound T and the same optimal sample size vector n . The relaxation is necessary because for Problem 3.2.1. Slater's condition naturally holds (proven in Proposition 3.2.6.). In general, the Karush-Kuhn-Tucker conditions are not guaranteed to be necessary and sufficient for nonlinear optimization problems. This allows for deriving sufficient optimality conditions for Problem 1.1.1. in Theorem 3.2.8., which represents one of the central results of the thesis. The conditions are derived in form of take-all strata and eigenpairs to the matrix $D_{\{\mathcal{A}\}p}$, efficiently computable from the model data.

In Chapter four, the Recursive Domain-Controlled Allocation (RDCA) algorithm is presented. The purpose of the algorithm is to find the optimal set of take-all strata, which uniquely determines the optimal solution to Problem 1.1.1. The algorithm recursively calls itself or, if the reduced domain set (J) is a singleton, the Domain-Controlled Allocation (DCA) algorithm. The later returns an "in some sense" optimal sample size vector for a specified set of take-all strata U . By this recursive structure the complexity of RDCA is exponential, which is a significant drawback as the number of strata may quickly exceed thousands and the number of domains multiple hundreds (E.g. earning survey from the German Federal Statistical office with approx. 1200 Domains und 7000 strata). The chapter contains a very nice visualisation of the RDCA for a problem of two and one of three domains.

Structural Properties of the RDCA algorithm are presented in Chapter five. Here, an advanced indexing scheme to address the states of the variables during the algorithm evaluation is introduced. This is very useful and facilitates not only the understanding of the algorithm, but also allows for explicitly addressing loops of iterative recursions for subsequent proofs. Very important is Proposition 5.2.2. showing the well-definedness of the algorithm and that i^* , the number of do-while loops within RDCA, is finite. A refined upper bound for i^* equals the number of strata in the current/active domain + 1.

Additionally, in this chapter the Last-Branch Recursion Path as well as the take-max strata sets V_i are introduced. Both are used to prove correctness of the algorithm in Chapter 6. The former describes the path inside a recursion tree that for i in $\{1, \dots, i^*\}$ in the most outer level chooses i^* for all inner recursion levels and is important to address the optimal take-max set at loop i at the most outer level.

Finally, in Chapter six, the correctness of the RDC Algorithm is examined. Figure 6.0.1 presents the interdependencies between Propositions, Lemmata, and Theorems in Chapters five and six, which is very helpful. Based on the partial correctness established in Theorem 6.4.2., the total correctness of the RDCA algorithm is shown in Theorem 6.4.3. Overall, Chapter six provides a mathematically concise chain of proofs establishing, finally, the correctness of the RDCA algorithm.

Chapter seven concludes the dissertation with a discussion of the results and gives an outlook to possible future work. Special emphasis is laid on introducing lower constraints (similar to Gabler, Ganninger, and Münnich, 2012), and achieving integrality of the solution, and using multivariate optimality.

The appendix covers seven chapters including abbreviations and symbols, which is very helpful though not complete, and further mathematical methods, which aim to provide additional mathematical methods that foster the understanding of the proofs. Finally, a bibliography with the main references is available.

Evaluation of the dissertation:

Overall, the dissertation is mathematically very sound and precise. The mathematical rigour within the dissertation is very convincing. Additionally, the thesis is well-structured and the (mathematically educated) reader is well-informed on the procedures and the connections between the different items.

To my view, the work would have benefited from short introductions in each chapter such that also practitioners would be able to follow the mathematical proofs. At least, this aim was emphasized in the introduction.

Nevertheless, few questions remain open to my view.

Despite the necessity to include the original reference of Neyman, one might add two more references. Kokan (1967) has presented an analytic algorithm that covers a subproblem of the thesis. Hohnhold (2009) introduced the multi-domain problem (based on our work in the project with DESTATIS, which was later published). Additionally, Münnich et al. (2013) showed that the box-constraint optimal allocation (with upper and lower bounds) can be solved directly with linear effort also for domains, but only for a quadratic norm over the domains.

As one major motivation, the author claims exactness of the solution. However, Gabler et al. (2013) introduce a necessary and sufficient condition, and Münnich et al. (2012) demonstrate linear algorithms to solve the problems. Together with Hohnhold (2009), one may raise the question if the given method is always preferable? In terms of exact solutions in the generalized problem it certainly has to be answered yes, if one claims the exactness as optimization calculus. When talking about solvability, one may think differently. In a test on German data (see above), the algorithm solved the problem in 74h, whereas our numerical method provided the same solution in seconds. It has to be pointed out that our method does not use general solvers, which, of course, may provide inferior solutions.

In terms of usefulness in practice, one may have wished to see a comparative simulation study with varying sizes of domains and strata. To my view, the algorithm performs quite convincing in small problems, whereas in large-scaled problems specified numerical algorithms seem preferable. As coming from the application side, one never shall forget that the final optimal solution by definition must be integer numbers.

Finally, the quality of the dissertation and the mathematical rigour is excellent. The author was able to prove the optimality, to develop a sufficient condition, and to turn the methods into an algorithm that is able to compute the exact solution in a very convincing manner. Limitations are related to practical purposes, which seem somehow lacking, and a slightly improved literature overview. The very sound demonstration throughout the thesis, however, leads to the conclusion that

I strongly support the acceptance of the PhD thesis

of Mr. Wojciech Wójciak.

Trier, 06. September 2026

(Prof. Dr. Ralf Münnich)