Intentional sampling by goal optimization with decoupling by stochastic perturbation

Marcelo De Souza Lauretto, Fábio Nakano, Carlos Alberto de Bragança Pereira, and Julio Michael Stern

Citation: AIP Conf. Proc. **1490**, 189 (2012); doi: 10.1063/1.4759603
View online: http://dx.doi.org/10.1063/1.4759603
View Table of Contents: http://proceedings.aip.org/dbt/dbt.jsp?KEY=APCPCS&Volume=1490&Issue=1
Published by the American Institute of Physics.

Intentional Sampling by Goal Optimization with Decoupling by Stochastic Perturbation

Marcelo de Souza Lauretto[+], Fabio Nakano[+],
Carlos Alberto de Bragança Pereira[*], Julio Michael Stern[*,**]

EACH-USP[+] and IME-USP[*], University of São Paulo
jstern@ime.usp.br[**]

Abstract.
Intentional sampling methods are non-probabilistic procedures that select a group of individuals for a sample with the purpose of meeting specific prescribed criteria. Intentional sampling methods are intended for exploratory research or pilot studies where tight budget constraints preclude the use of traditional randomized representative sampling. The possibility of subsequently generalize statistically from such deterministic samples to the general population has been the issue of long standing arguments and debates. Nevertheless, the intentional sampling techniques developed in this paper explore pragmatic strategies for overcoming some of the real or perceived shortcomings and limitations of intentional sampling in practical applications.

Keywords: Decoupling and randomization; Goal Optimization; Intentional sampling; Multiobjective Programming; Post-optimal and sensitivity analysis; Stochastic perturbations.
PACS: 01.30.Cc, 02.50.Ng, 02.50.-r

The conterquestion 'How can you justify purposive sampling?' has a lot of force in it. The choice of a purposive plan will make a scientist vulnerable to all kinds of open and veiled criticisms. A way out of the dilemma is to make the plan very purposive, but to leave a tiny bit of randomization in the plan.
Basu (1987,ch.XIV, p.257) - Why to Randomize?

1. INTRODUCTION

This extended abstract describes and comments some of the statistical models developed by the authors in 2002 for a project lead by *Datanexus*, a Brazilian survey institute for media audience. The main goal of this project was to provide new alternatives for measuring television audience in São Paulo, Brazil. The project's measurement technology was based on a new meter capable of monitoring and recoding, at 1 minute intervals, the time spent by each individual in the household watching each of the available open TV channels.

The goal of the statistical modeling was to help in the selection of the *Monitoring Sample*, a set of households where the equipment should be installed. The monitoring sample was subject to the budget constraint of $\beta = 250$ households. In a preliminary research, $m = 10.000$ households in a *Interview Sample* were asked to answer a detailed questionnaire. The monitoring sample of size β should be selected from the m house-

XI Brazilian Meeting on Bayesian Statistics
AIP Conf. Proc. 1490, 189-201 (2012); doi: 10.1063/1.4759603

holds in the interview sample.

Section 2 details the stratification and clustering techniques used in the selection procedures for the interview sample. Section 3 defines the matrix and other mathematical notations used in this paper. Section 4 explains the procedures used in the selection of the monitoring sample; these procedures are based on Goal Programming or Multiobjective Programming. Section 5 discusses the real or percieved dificulties in making statistical generalizations based on non-randomized samples. Section 6 presents stochastic perturbation techniques designed as succedanea or poor man's substitutes for randomized sampling. Section 7 presents some experimental results concerning the performance of the techniques and procedures developed in this paper. Section 8 presents our final remaks.

2. STRATIFICATION AND CLUSTERING

In 2002, IBGE, the Brazilian Institute of Geography and Statistics, had available data on household and populational socio economic features, aggregated by censitary units, CUs. The São Paulo metropolitan area was divided in 21.240 CUs. In this data bank, each unit is characterized by 556 variables. These variables constitute 4 major groups, namely: 87 describe the households themselves, 183 describe the households' responsible persons, 113 describe the level of instruction of the persons in the households, and 173 describe other characteristics of persons living in the households. Of the original CU's, 217 had to be eliminated from the study due to data bank inconsistencies.

Taking into consideration the correlation matrix of all these variables, as well as the analyses carried by the final users of the audience information, the 556 descriptive variables were aggregated into 30 artificial variables, where 7 variables relate to income, 7 relate to instruction, 8 relate to the households' sizes and facilities, and 8 variables relate to the sex and age of the persons living in the households.

The first step in the procedure for selecting the interview sample was to group the CUs into 10 strata (deciles) according to mean personal income. Next, using the 30 artificial variables, the CU's in each stratum were grouped into 10 clusters, using a Simulated Annealing algorithm and the L_1 metric, see Späth (1985, ch.6) and Pflug (2000). The main motivations for choosing the L_1 norm were:

- Unified treatment of Ordinal and Continuous variables.

- Robustness: Using the L_1 norm, cluster's centers are defined by median coordinates. Since median values depend only on the ranking of the individual points in the cluster, they are very robust to outliers,

After stratification and clustering, the CU's were divided into 100 (more or less) homogeneous groups. Finally, 10 households were selected in each cluster using a randomized sampling procedure, constituting the $m = 10,000$ household interview sample. Each household in the interview sample answered a detailed questionnaire that updated demographic and socio-economic data, and included new data focusing entertainment activities, like reading habits, internet access and, most importantly, the number and quality of the TV sets available at each household as well as individual TV watching habits. The questionnaires were used to update the 30 artificial variables on demographic and social economic data, and tabulate 30 more variables on entertainment habits.

Hence, with the information given by the questionnaires, each of the $m = 10.000$ households in the interview sample is described by Data Bank of $u + v = 30 + 30 = 60$ features, as detailed in the next section. From the 10,000 interview sample, 2,672 households were eliminated due to missing values.

3. MATRIX NOTATION AND DATA BANK STRUCTURE

The operator $r{:}s{:}t$, read - *from r to t* with *step s*, indicates the vector $[r, r+s, r+2s, \ldots t]$ or the corresponding domain of indices. $r{:}t$ is a short hand for $r{:}1{:}t$. $A(i, j)$ is the element in the i-th row and j-th column of matrix A. Index vectors can be used to build a sub-matrix by extracting a sub-set of rows and columns. For example, $A(1 : m/2, n/2 : n)$ is the north-east block, that is, the block of first rows and last columns, from matrix A. The notation $\text{rand}(m)$ designates a vector of independent uniform $]0,1[$ random numbers, while $\mathbf{1}$ and $\mathbf{0}$ designate matrices of ones's and zeros with the appropriate dimensions.

The pointwise arithmetic operators, $\oplus, \ominus, \odot, \oslash$, etc., act element by element on matrices of same dimension, for example,

$$C = A \odot B \Leftrightarrow C(i, j) = A(i, j)B(i, j) \ .$$

The p-norm of a vector is defined as

$$||x||_p = \left(\sum_j |x_j|^p\right)^{1/p} \ , \quad ||x||_\infty = \max_j |x_j| \ .$$

Each selected household is studied with respect to features of type $t \in \{1, 2, \ldots, u\}$. Each individual living in a selected household is studied with respect features of type $t \in \{u+1, u+2, \ldots, u+v\}$. A given feature type, t, entails a discrete, ordinal, $d(t)$-dimensional classification system, with classes $\{1, 2, \ldots, d(t)\}$. For example, in Brazil, it is common practice to use the A, B, C, D, E ordinal classification system, $d(t) = 5$, corresponding to percentile strata of interest for each feature, say top $3\%, 10\%, 30\%, 60\%$ and lower 40%. The auxiliary vector $c(t)$ gives cumulative class dimensions, that is, $c(t) = d(t) + c(t-1)$ where, by definition, $c(0) = 0$.

A single matrix A is used to tabulate all data from the exploratory research. The h-th line of matrix A, $A(h, :)$, contains all the data concerning household h and the individuals living in it. All entries of A are zero if not otherwise stated. For $1 \le t \le u$ and $c(t-1)+1 \le k \le c(t)$, $A(h,k) = 1$ iff household h is of class k for feature type t. For $u+1 \le t \le u+v$ and $c(t-1)+1 \le k \le c(t)$, $A(h,k)$ is the number of individuals of class k for feature type t living in household h. From the above definitions we have the following normalization conditions:

- For the household classification sub-matrix, i.e. for $1 \le t \le u$, and $1 \le k \le c(u)$, $A(h, c(t-1)+1 : c(t))\mathbf{1} = 1$. This is the 'partition' condition, stating that, for any feature type t, household h belongs to one and only one class. Also, for any feature type t, the total number of households of class k is given by $\mathbf{1}'A(1 : m, c(t-1)+k)$.

- For the individual classification sub-matrix, for $u+1 \le t \le u+v$ and $c(u)+1 \le k \le c(u+v)$, $A(h, c(t-1)+1 : c(t))\mathbf{1}$ gives the number of individuals in household h. Similarly, $\mathbf{1}'A(1 : m, c(t-1)+k)$ gives the number of individuals of class k for feature type t.

4. INTENTIONAL SAMPLING

The project's tight budget constraint of having only $\beta = 250$ households in the monitoring sample, precludes the use of traditional statistical randomized sampling techniques. For a thorough and humorous analysis of the limitations of traditional sampling techniques, see Brewer (2002). In order to make a pilot study viable, this project used non-probabilistic intentional (or purposive) sampling. Intentional sampling methods are non-randomized procedures that select a group of individuals for a sample with the purpose of meeting specific prescribed criteria.

4.1. Goal Optimization

The intentional sampling criteria discussed in the sequel seek the efficient selection of a monitoring sample where all pertinent pre-defined clusters or strata of the general population are well represented. However, the ideas motivating these intentional sampling criteria are compleatly different from the ideas motivating traditional forms of *representative sampling* in Statistical theory, see Brewer (2002).

In this sub-section, the intentional sampling criteria are formulated as Goal Optimization problems. The fulfillment of these intentional criteria gives, in turn, plausible arguments for generalizing (in proportion) some means (or first moments, or central values) of the monitoring sample to the general population.

The Goal Optimization problem for sample selection is formulated as a generalized *knapsack problem*, defined by the following entities:

- $g(1 : c(u+v))$, a *goal* or target vector for optimal panel representation;
- x, Boolean *decision variables*. x_h indicates if household h belongs (or not) to the selected monitoring sample;
- b, the monitoring cost and β, the budget. In this paper the monitoring cost is considered to be constant, with $b = \mathbf{1}$;
- w, a positive vector of *weights*. It may me convenient to write the weights as the ratio of an importance and a normalization vector, $w = wm \oslash wn$, see Romero (1991,Ch.7);
- r,s, non-negative *surplus and slack variables*. As usual in mathematical programming, these artificial variables measure departure from the problem's (idealized) constraints,

$$b'x \leq \beta \ , \ A'x - r + s = g \ .$$

Milan Zeleny (1982, p.156) enunciates the following *displaced ideal* criterion for optimal choice:

> *"Alternatives that are closed to the ideal are preferred to those that are farther. To be as close as possible to the perceived ideal is the rationale of human choice."*

Zeleny's displaced ideal criterion is translated into the following Goal Optimization problem:

$$\min \| \; w \odot (s + r) \; \|_p \; .$$

For $p = 1$ and $p = \infty$, the absolute and minimax norms, the Goal Optimization problem can be reduced to a Linear Programming problem that, in turn, can be solved very efficiently by the Simplex method or by interior point algorithms, see Martin (1998) and Murtagh (1981). The Simplex method, in its primal or dual forms, is specially well suited for this application, allowing fast and efficient reoptimization if constraints are added or modified, as required by several techniques used in this paper, like post-optimal and sensitivity analysis, see Gal (1979) and Gal, H.J.Greenberg (1997), or branch-and-bound algorithms for integer programming, see Kovács (1980) and Nemhauser and Wolsey (1988).

4.2. Multiobjective Programming

Vilfredo Pareto (1896) stated the following criterion of *dominance*:

> *In a Multiobjective Programming problem, a solution A dominates a solution B if and only if A is better than B with respect to at least one objective, and A is not worse than B with respect to the remaining objectives.*

For the statistical sampling problem at hand, consider two solutions, A and B. A is a solution that exactly achieves the goal, and B is an alternative solution that samples exactly the same number of households and individuals in each class of each feature type, but includes an extra individual at one of the households. From a statistical point of view, solution B dominates solution A, because B includes a sample equivalent to A as a sub-sample. Hence, in no circumstance sample B provides less information, and in some circumstances it possibly provides more information, than sample A.

In a Multiobjective Programming problem, a solution is *Pareto-efficient* iff it is not dominated by any other feasible solution or, as stated by Zelleny (1982,p.74),

> *A non-dominated [or efficient] solution is a feasible solution for which and increase in value of any criterion can be achieved only at the expense of a decrease in value of at least one other criterion.*
>
> *The non-dominated boundary [of the feasible region] is sometimes characterized as the trade-off curve [or efficient frontier].*

The Pareto efficiency criterion is slightly different from the displaced ideal criterion used in the last sub-section. Accordingly, Pareto efficiency leads an alternative optimization problem, namely, Multiobjective Programming. Zeleny (1982) points out that the Goal programming formulation may sometimes produce optimal solutions that are inefficient for an alternative, and better formulated, Multiobjective Programming problem. These considerations lead to an improved formulation for the sampling problem, where the surplus variables, r, are not explicitly penalized, only the slack variables, s. This is the Multiobjective Programming problem:

$$\min \| \; w \odot s \; \|_p \; .$$

Notwithstanding apparent benefits of Multiobjective Programming formulations, previous commitments assumed by the client concerning the performance and evaluation metrics to be adopted in the project, made Goal Optimization the formulation of choice. Similar reasons explain the addoption of a penaly function based on the absolute norm, $p = 1$, instead of the minimax norm, $p = \infty$, or even a more complex penaly function based on a convex combination of both norms.

5. THE ROLE OF RANDOMIZATION

As defined in the last section, intentional sampling methods are non-randomized procedures that select a group of individuals for a sample with the purpose of meeting specific prescribed criteria. The forth author of this paper is troubled about the validity of using statistical generalization arguments from such a non-randomized sample to the general population. Better said, he believes that one cannot do so without making very strong and often unrealistic assumptions about the (joint) probability distributions underlying the observed phenomena. For similar perspectives, see Hacking (1988, p.429-430), Kempthorne (1977, p.16), Peirce and Jastrow (1884) and Pearl (2000, p.340,348). Of course, that does not mean that, even using strict non-randomized intentional sampling methods, one can not construe generalization arguments under non-statistical formalisms, see Stern (2004) and Stern, Pereira (2012) and references herein. The forth author presents his opinions concerning the decouling principle, the role of randomization, and related issues in Stern (2008, 2011a).

Moreover, the forth author believes that only working under an appropriate epistemological framework and using appropriate methodologies, that most certainly include randomization methods, is it possible to achieve the goal of "objective" statistical inference, see Stern (2011b,c).

The above comments are not corroborated by the third author of this paper. For him, working under a "genuine" Bayesian methodology, any possible sample observation that could be possibly obtained by a randomization process, could be equality treated as if it was collected under the process. The reason for this opinion is that randomization is independent of the objective of the study, the inference about the state of nature. Hence, the randomization process cannot interfere in the statistical inference to be used. Whenever randomization is used, its role ends before the statistical inference work starts. For a clinical trial study conducted along these lines, see Fossaluza et al. (2010).

Furthermore, the third author does not corroborate any attempt to serach for "objective" inference procedures. In fact, for him, the subjective perspective is so essential to the "genuine" Bayesian framework that it can be used to define the scope of Bayesian statistics; As stated by I.G.Good (1972) in his Random Thoughts about Randomness:

- *A statistician who uses subjective probabilities is called a 'Bayesian'.*

The quotation of Debabrata Basu opening this paper inspired both of the contending authors of the current paper. However, taken this quotation in its proper context, it becomes clear that Prof. Basu would fully support the positions taken by the third author. The following paragraphs are reproduced from Basu (1988):

The object of planning a survey should be to end up with a good sample.

The term "representative sampling" has often been used in sample survey terminology. But no one has cared to give a precise definition of the term. It is implicitly taken for granted that the statistician with his biased mind is unable to select a representative sample. So a simplistic solution is sought by turning to an unbiased die (the random number tables). Thus, a deaf and dumb die is supposed to do the job of selecting a "representative sample" better than a trained statistician. Basu (1988,p.198).

(Why to randomize?) - The conterquestion 'How can you justify purposive sampling?' has a lot of force in it. It is only in transparently simple cases that one can give a clear-cut argument in favor of a particular purposive plan. In a true-to-life survey situation, it is very difficult to sell the idea of a fully purposive plan. The very purpose of a purposive plan is rooted in the scientific intuition and knowledge of a surveyor. No two surveyors are likely to agree on the choice of their survey plans. The choice of a purposive plan will make a scientist vulnerable to all kinds of open and veiled criticisms. A way out of the dilemma is to make the plan very purposive, but to leave a tiny bit of randomization in the plan; for example, draw a systematic sample with a random start or make a very extensive stratification of the population and then draw a sample of size 1 from each stratum! Basu (1988, p.257)

It is a clear imperative that the surveyor fully describe his survey plan and carefully explain all the considerations that led to the particular plan. And this inhibits the choice of a purposive plan. The possible criticism that the surveyor's chosen plan was not the optimum one (even with respect to his own background information) may not cast any doubt on his conclusions as long as the critic can analyze his (the surveyor's) data. No wonder, therefore, that all of us choose the path of least resistance and try to incorporate an element of randomness in the survey plan. Basu (1988, p.258).

Notwithstanding their different motivations, distinct theoretical backgrounds and divergent epistemological frameworks, both of the contending authors can agree on the convenience of reconciling randomization methods with intentional survey sampling. This is the objective of the methods developed in Section 6.

Before ending this section, it is important to emphasize that the statistical literature distinguishes between two intended uses of randomization, namely, *decoupling* (aka random design) and *model justification* (aka random sampling):

Decoupling: Randomization techniques aiming to eliminate experimental biases coming from systematic design problems, including several forms of uncontrolled influence, conscious or unconscious, received from or exerted by participating agents.

Model justification: Randomization techniques aiming to justify assumptions concerning the functional form of distributions and parametric constraints in the statistical model of the experiment.

A deeper probabilistic analysis of randomization, as it is commonly used is statistics, shows that, from a theoretical point of view, the two concepts can greatly overlap, see Gelman (2003), Pearl (2000) and Stern (2008, 2011a). Nevertheless, the methodology developed in the next section focus specifically on decoupling.

6. DECOUPLING BY STOCHASTIC PERTURBATIONS

This section discusses stochastic perturbation methods designed as succedanea or poor man's substitutes for the randomization decoupling techniques used in traditional statistical sampling. The decoupling methods developed in this section are based on celebrated analogies between stochastic processes and chaotic, turbulent or unstable phenomena, for general overviews, see Chaitin et al. (2011), Chazelle (2000), Claude and Chaitin (2008), Griffinths and Tenenbaum (2003), Schroeder (1991), Stern (2011) and Terwijn (2003). Accordingly, the methods developed in this section require, for the Goal or Multiobjective Programming problem formulated in Section 3, perturbations that are large enough to fully enter these mathematical programming problems' instability region. At the same time, in order to obtain solutions that may be slightly sub-optimal but still quite useful for the purposes stated in the previous sections, the same perturbations should be as small as possible.

Most of the work in post-optimal and sensitivity analysis for integer programming already publisher in the mathematical programming literature concerns linear programming, and focus sufficient stability conditions, like stability radii for cost or right hand side perturbations. These kind of stability analysis can only provide lower bounds to the perturbations required for our intended purpose.

Fortunately, there are available in the mathematical programming literature a few "negative results" concerning the instability of optimization problems that can be adapted to our purposes, see for example Blair (1997, 1998). Inspired by Brair's results, a perturbed problem is defined by a new generalized knapsack constraint that is obtained replacing β by $\widetilde{\beta} = \beta + 1$ and b by $\widetilde{b} = b + z$, where $z = \varepsilon(2/\beta)\text{rand}(m)$. The expected value $E(z(i)) = \varepsilon/\beta$. Hence, the expected increase in value for a boolean solution x with β non-zero elements is $E(z'x) = \varepsilon$. As discussed in the next section, a convenient choice of the perturbation parameter ε, positive and not much greater than 1, should produce appropriate perturbations.

7. EXPERIMENTAL RESULTS

The Goal Programming problem described in Section 4 was implemented using GAMS - the General Algebraic Modeling System compiler, and solved using the CPLEX optimization library. All pre and post processing routines, including the generation of random perturbation vectors, graphics, etc., were implemented in ANSI-C and R.

The features considered as targets for optimal panel representation and their respective categories are ($[a,b[$ represents the range $a \leq x < b$ for a real variable x, and $[a,b]$ represents the range $a \leq k \leq b$ for an integer variable k):

1. Household monthly income (minimum sallaries): $[0,1[$, $[1,2[$, $[2,3[$, $[3,5[$, $[5,10[$, $[10,20[$, $[20,\infty[$;
2. Individual social-economic class: A, B, C, D/E;
3. Sex: Male, Female;
4. Age range (years): $[0,4]$, $[5,9]$, $[10,14]$, $[15,19]$, $[20,24]$, $[25,29]$, $[30,34]$, $[35,39]$, $[40,44]$, $[45,49]$, $[50,59]$, $[60,\infty[$;
5. Scholarity level (years of study): $[0,1]$, $[1,4]$, $[5,8]$, $[9,12]$, $[12,\infty[$;

6. Individual daily TV attendance (hours/day): $[1,2]$, $[3,4]$, $[5,6]$, $[7,8]$, $[9,10]$, $[10,\infty[$.

The numerical experiments were based on 150 runs for each perturbation parameter $\varepsilon \in \{1, 1.5, 2, 2.5, 3\}$. Each run consists in generating a random vector z and solving the perturbed goal optimization problem, with norm $p = 1$ and homogeneous feature weight $w = \mathbf{1}$, generating an optimal portfolio of $\beta = 250$ households for the monitoring sample. Two criteria were used to analyse the performance of our Goal Optimization with Stochatic Perturbation method: *Decoupling* and *Optimality*.

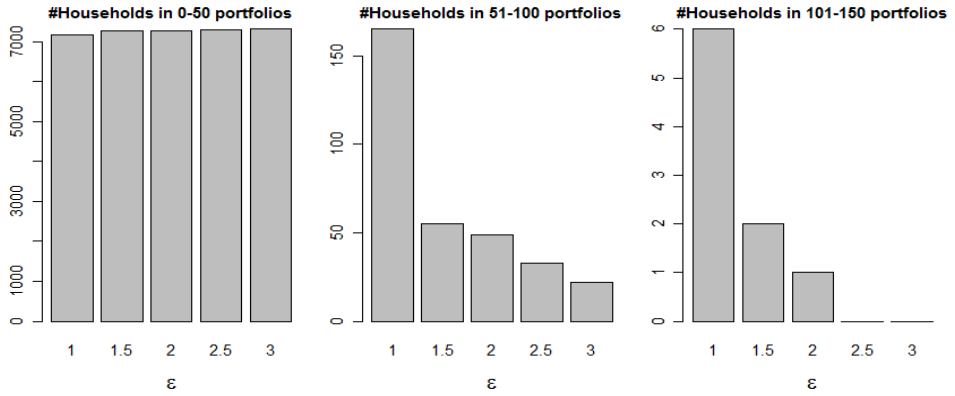

FIGURE 1. Household repetitions in (sub)optimal panels for $\varepsilon \in \{1, 1.5, 2, 2.5, 3\}$.

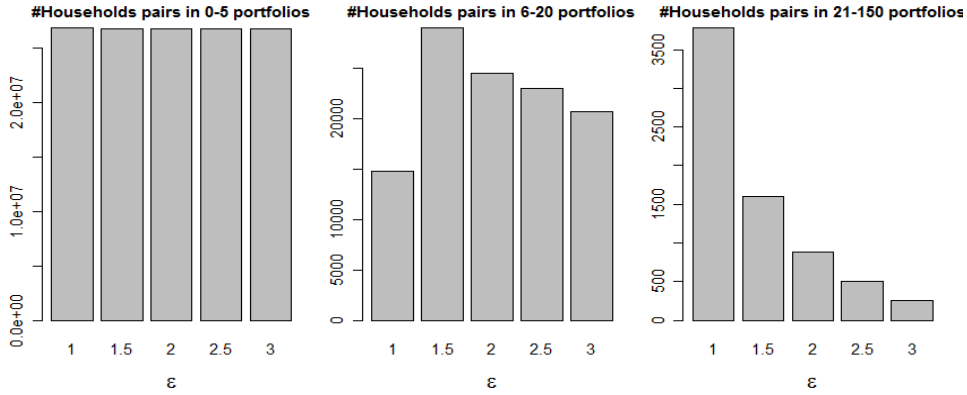

FIGURE 2. Pair of households repetitions in (sub)optimal panels for $\varepsilon \in \{1, 1.5, 2, 2.5, 3\}$.

The first criterion, *Decoupling*, concerns the absence of a tendency to select the same households or pairs of households in different runs. Figure 1 shows the repetition of selected households among the 150 runs for each $\varepsilon \in \{1, 1.5, 2, 2.5, 3\}$. The leftmost graph shows the number of households selected between zero and 50 times; the second shows the number of households selected $[51, 150]$ times; and the rightmost graph shows the number of households selected $[101, 150]$ times. Figure 2 contains analogous graphs considering pairs of households, using the ranges $[0,5]$, $[6,20]$ and $[21,150]$. We can observe that the tendency to choose the same households in different runs decreases rapidly with the parameter ε. For $\varepsilon = 3$, only 25 out of 7,328 candidate households are selected more than 50 times (33.3% of runs), and only 250 out of 2.7E+7 pairs are chosen in more than 20 times (13.3% of runs).

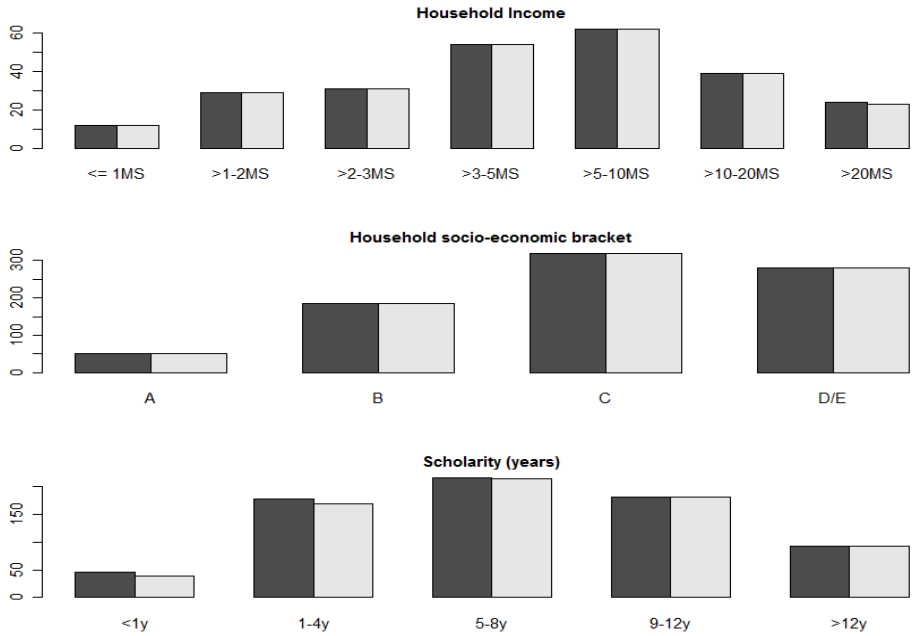

FIGURE 3. Desired (dark) and actual (light) sample frequencies, $\varepsilon = 3$.

The second criterion, *Optimality*, considers how close to the desired target is the optimal solution of the stochastically perturbed Goal Optimization problem. Naturally, larger values of the perturbation parameter induce strong decoupling, favouring the first criterion, while smaller values engender perturbed problems that are closer to the original deterministic problem, favouring the second criterion. From an empirical analysis by the two criteria, $\varepsilon = 3$ seems to provide a good compromise.

Figures 3 and 4 show a comparison between the desired (dark) and actual (light) sample frequencies, for a given run chosen randomly from the 150 available with $\varepsilon = 3$. This randomly chosen run is fairly typical for the batch, producing an optimal solution that is very close to the desired target.

Empirical evidence based on experiments like the one described above give us a strong indication that, once the single empirical perturbation parameter, ε, is adequately calibrated, the Goal Optimization with Stochatic Perturbation method performs admirably well in the following senses:

- Decoupling: Distinct perturbation vectors, $z = \varepsilon(4/\beta)\text{rand}(m)$, produce very different optimal solutions, that is, distinctively diverse monitoring samples;
- Optimality: The optimal solution of the stochastically perturbed Goal Optimization problem come very close to the desired target.

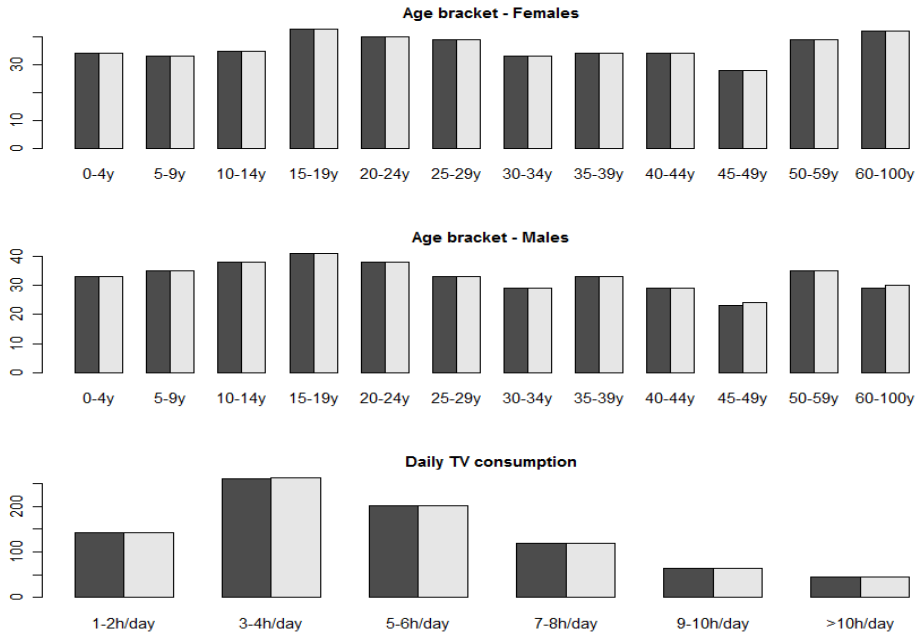

FIGURE 4. Desired (dark) and actual (light) sample frequencies, $\varepsilon = 3$.

8. FUTURE RESEARCH AND FINAL REMARKS

Future Research Concerning Optimization

At the present stage of development, the perturbation parameter ε is calibrated empirically by a trial and simulation procedure used in conjunction with numerical optimization by the bisection method. We hope to extend Blair's instability analysis to the generalized knapsack problem in order to obtain analytically a reasonable starting guess for the value of ε.

We also hope to extend this paper's approach to non-linear integer programming problems, allowing the use objective functions based on p-norms for $1 \leq p \leq \infty$, as preconized in Section 4.1. We believe that the results presented at Skorin-Kapov and Granot (1987) provide good insights and some useful hints for pursuing this path of research.

Future Research Concerning Inference

We hope to extend this paper's approach of integrating interntional sampling with decoupling by stochastic perturbation to sequential designs. Fossaluza et al. (2009) and references herein offer good examples for possible applications in this line of research.

Acknowledgements

The authors are grateful for stimulating conversations about the pragmatics of survey sampling they had with social scientist Carlos Novaes. The authors are grateful for the support of EACH, the School of Humanities Arts and Sciences, and IME, the Institute of Mathematics and Statistics, of USP, the University of São Paulo; FAPESP, Fundação de Amparo à Pesquisa do Estado de São Paulo; and CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico.

REFERENCES

D.Basu, J.K.Ghosh (ed.) (1988), Statistical Information and Likelihood, A Collection of Essays by Dr.Debabrata Basu, *Lecture Notes in Statistics,* 45, Springer. The following chapers are of special interest for the readed of this paper:
- Ch.XI, p.186-200, Likelihood Principle and Survey Sampling. Ch.XI is based on the article: D.Basu (1969), The Role of Sufficiency and Likelihood Principle in Sample Survey Theory, Sankhya 31, 441-454.
- Ch.XIV, p.245-270, Relevance of Randomization in Data Analysis. Ch.XIV is based on the article: D.Basu (1978), On the Relevance of Randomization in Data Analysis, with discussion, in: Survey Sampling and Measurements, 267-339, N.Nanboodiri ed. NY: Academic Press.
- Ch.XV, p.271-289, The Fisher Randomization Test and Ch.XVI, p.290-312, The Fisher Randomization Test: Discussions. Ch.XV is based on the article: D.Basu (1980). Randomization Analysis of Experimental data: The Fisher Randomization Test, with discussions, *JASA* 75, 575-595.

C.Blair (1997). Integer and Mixed-Integer Programming. p.9.1-9.25 in T.Gal, H.J.Greenberg (1997).

C.E.Blair (1998). Sensitivity analysis for knapsack problems: A negative result. *Discrete Applied Mathematics,* 81, 133-139.

K.R.W.Brewer (2002). *Combined Survey Sampling Inference: Weighing of Basu's Elephants.* Hodder Education Publishers.

G.Chaitin; F.Doria; N.Costa (2011). *Goedel's Way: Exploits into an undecidable world.* CRC Press.

C.Claude, G.J.Chaitin (2008). *Randomness and Complexity: From Leibniz to Chaitin.* World Scientific.

B.Chazelle (2000). *The Discrepancy Method: Randomness and Complexity.* Cambrige Univ.Press.

V.Fossaluza, J.B.Diniz, B.B.Pereira, E.C.Miguel, C.A.B.Pereira (2009). Sequential Allocation to Balance Prognostic Factors in a Psychiatric Clinical Trial. *Clinics,* 64, 511-518.

P.Gács (2010). *Lecture Notes on Descriptional Complexity and Randomness.* Tech.Rep., Boston University.

T.Gal (1979). *Postoptimal Analyses, Parametric Programming, and Related Topics.* NY: McGraw-Hill.

T.Gal, H.J.Greenberg (1997). *Advances in Sensitivity Analysis and parametric programming.* Dordrecht: Kluwer.

A.Gelman, J.B.Carlin, H.S.Stern, D.B.Rubin (2003). *Bayesian Data Analysis*, 2nd ed. NY: Chapman and Hall / CRC.

I.J.Good (ed.) (1962). *The Scientist Speculates. An Anthology of Partly-Baked Ideas.* NY: Basic Books.

T.L.Griffinths, J.B.Tenenbaum (2004). Probability, Algorithmic Complexity, and subjective Randomness. p.953-961 in S.Thrun, L.K.Saul, B.Schölkopf (2004). Advances in Neural Information. MIT Press.

I.Hacking (1988). Telepathy: Origins of randomization in experimental design. *Isis,* 79, 3, 427-451.

O.Kempthorne (1977). Why Randomize? *J. of Statistical Planning and Inference,* 1, 1-25

L.B.Kovács (1980). Combinatorial Methods of Discrete Programming. Budapest: Akadémiai Kiadó.

R.K.Martin (1998). *Large Scale Linear and Integer Optimization: A Unified Approach.* NY: Springer.

B.A.Murtagh (1981). *Advanced Linear Programming.* NY: McGraw Hill.

G.L.Nemhauser, L.A.Wolsey (1988). Integer and Combinatorial Optimization. Chichestr: John Wiley.

J.Pearl (2000). *Causality: Models, reasoning, and inference.* Cambridge University Press.

C.S.Peirce, J.Jastrow (1884). On small differences of sensation. *Memoirs of the National Academy of Sciences,* 3, 75-83.

G.C.Pflug (1996). *Optimization of Stochastic Models: The Interface Between Simulation and Optimization.* NY: Springer.

C.Romero (1991). *Handbook of Critical Issues in Goal Programming.* Oxford: Pergamon Press.

J.Skorin-Kapov, F.Granot (1987). Non-linear Integer Programming: Sensitivity Analysis for Branch and Bound. *Operations Research Letters,* 6, 269-274.

J.M.Stern (2004). Paraconsistent Sensitivity Analysis for Bayesian Significance Tests. SBIA'04, *LNAI*, 3171, 134–143.

J.M.Stern (2008). Decoupling, Sparsity, Randomization, and Objective Bayesian Inference. *Cybernetics and Human Knowing*, 15, 49-68.

J.M.Stern (2011a). Spencer-Brown vs. Probability and Statistics: Entropy's Testimony on Subjective and Objective Randomness. *Information,* 2, 2, 277-301.

J.M.Stern (2011b). Symmetry, Invariance and Ontology in Physics and Statistics. *Symmetry,* 2011, 3, 3, 611-635.

J.M.Stern (2011c). Constructive Verification, Empirical Induction, and Falibilist Deduction: A Threefold Contrast. *Information,* 2, 4, 635-650.

J.M.Stern, C.A.B.Pereira (2012). A Bayesian Possibilistic Epistemic Significance: Ev(H) - Focus on Surprise, Measure Probability. Submitted.

H. Späth (1985). *Cluster Dissection and Analysis. Theory, Fortran programs and examples.* Chichester: Ellis Horwood.

M.Schroeder (1991). *Fractals, Chaos and Power Laws: Minutes from an Infinite Paradise.* NY: W.H.Freeman.

S.A.Terwijn (2003). *Complexity and Randomness.* Tech.Rep. CDMTCS-212, Technical University of Vienna.

M.Zeleny (1982). *Multiple Criteria Decision Making.* NY: McGraw-Hill.