European Journal of Operational Research 241 (2015) 880–887

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Information weighted sampling for detecting rare items in finite populations with a focus on security ☆

André J. Hoogstrate [a,b,*], Chris A. J. Klaassen [c]

[a] Knowledge and Expertise Centre for Intelligent Data Analysis, Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB The Hague, The Netherlands
[b] Centre for Terrorism and Counterterrorism, Campus The Hague, Leiden University, Koningin Julianaplein 10, P.O.Box 13228, 2501 EE The Hague, The Netherlands
[c] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 107, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 9 December 2013
Accepted 15 September 2014
Available online 13 October 2014

Keywords:
Applied probability
Probability sampling
Rare events
Profiling

ABSTRACT

Frequently one has to search within a finite population for a single particular individual or item with a rare characteristic. Whether an item possesses the characteristic can only be determined by close inspection. The availability of additional information about the items in the population opens the way to a more effective search strategy than just random sampling or complete inspection of the population. We will assume that the available information allows for the assignment to all items within the population of a prior probability on whether or not it possesses the rare characteristic. This is consistent with the practice of using profiling to select high risk items for inspection. The objective is to find the specific item with the minimum number of inspections. We will determine the optimal search strategies for several models according to the average number of inspections needed to find the specific item. Using these respective optimal strategies we show that we can order the numbers of inspections needed for the different models partially with respect to the usual stochastic ordering. This entails also a partial ordering of the averages of the number of inspections.

Finally, the use, some discussion, extensions, and examples of these results, and conclusions about them are presented.

1. Introduction

This research is motivated by several problems relevant to security applications. Examples thereof are the search for a terrorist among a group of passengers, for a container carrying illicit material on a vessel entering a port, for a murderer that has left his DNA profile at a crime scene in a small community, etc. In general, one has to search within a finite population for a particular item with a rare characteristic. Only close inspection will reveal if an item possesses the characteristic or not. Based on profiling, a relatively quick assessment is obtained on the probability that an individual item has the rare characteristic. Subsequently, the possibly expensive or intrusive inspection of the high probability individuals or items is started. The underlying idea is that this is an economically desirable, logistically possible, and hopefully socially acceptable way of improving security in contrast to purely random checks or inspection of all relevant individuals.

The search for an optimal and practical solution received a big boost after the events of 9/11. A survey of operation research models used in Homeland Security (Wright, Liberatore, & Nydick, 2006) references numerous examples of research related to the above mentioned idea, most prominently put forward in airline and airport security. Examples of more recent research centered around information based risk selection to improve airport and airline security include Press (2009), Press (2010), Meng (2012), Babu, Batta, and Lin (2006), Bagchi and Paul (2014), Cavusoglu, Koh, and Raghunathan (2010) and Nie, Batta, Drury, and Lin (2009). Although airport security has drawn the most attention, other topics have received research attention as well, like port security, cf. Bier and Haphuriwat (2011) and a host of other subjects, cf. Wright et al. (2006).

In this research we limit ourselves to the situation where it is certain that exactly one individual or item with the rare characteristic belongs to the population. This situation was studied earlier by Press (2009). He considered a subset of the models that we have studied in Hoogstrate and Klaassen (2011) and that we study in this article. Press' results and our results in Hoogstrate and Klaassen (2011) are limited to the average number of inspections, while we extend these results here to the stochastic ordering of the numbers of inspections themselves. Meng (2012) and Press (2010) extended the results of Press (2009) from a population at one checkpoint to a population flowing through a network of airports with multiple checkpoints. In the present study we use Press (2009) as a starting point, but we apply an axiomatic approach, thus specifying our assumptions clearly.

☆ Part of the research was funded by the program VIA: Veiligheidsverbetering door Information Awareness (Improving Security by Information Awareness).

* Corresponding author. Tel.: +31 70 8886399.
E-mail addresses: a.hoogstrate@nfi.minvenj.nl, a.j.hoogstrate@cdh.leidenuniv.nl (A. J. Hoogstrate), c.a.j.klaassen@uva.nl (C. A. J. Klaassen).

http://dx.doi.org/10.1016/j.ejor.2014.09.035
0377-2217/

After introducing our models and assumptions we discuss in Section 1.3 similarities to and differences with Meng (2012), and Press (2009, 2010).

1.1. Assumptions

For the population the following assumptions hold.

1. **Finite population.** The population consists of a finite number N of items, numbered $i = 1, 2, \dots, N$.
2. **Uniqueness.** One and only one of the items in the population possesses the characteristic Γ.
3. **Prior probabilities.** Each item i can be assigned a known probability $p_i > 0$ of possessing the characteristic Γ we are searching for, and these probabilities add up to 1.

The index of the Γ-item may be viewed as the result of one draw from the set $\{1, 2, \dots, N\}$ with sampling probabilities $(p_1, p_2, \dots, p_N)$. We know $(p_1, p_2, \dots, p_N)$, but not the result of the draw, which follows a multinomial distribution with parameters 1 and $(p_1, p_2, \dots, p_N)$.

For the procedures of inspection we vary the following assumptions.

4. **Enumeration.** Whether or not it is possible to enumerate and order the items according to their associated prior probability of possessing characteristic Γ. This translates into the issue whether or not one can deterministically control the order in which items will be inspected.
5. **Recognition.** Whether or not recognition of characteristic Γ is perfect. We introduce the parameter $s_i, 0 < s_i \le 1, \ i = 1, \dots, N$, as the probability of recognizing characteristic Γ when item i is inspected and actually has the characteristic.
6. **Replacement.** Whether or not it is possible to apply sampling without replacement.
7. **Memory.** Whether or not it is possible to use the information that an item has been selected before, and to use the outcome of this inspection.

Assumptions 4–7 result in 16 different models as listed in Table 1. Procedures are allowed only if they stop searching once the Γ-item has been found. Formally we put the following two conditions on the search procedures.

8. **Stopping rule.** Once the Γ-item has been found or when no items remain for inspection, no further inspections take place.
9. **Finiteness.** The search procedure terminates after a finite number of inspections.

Next we introduce the inspection probabilities. These are the probabilities within the models I–P that govern the process that selects items for inspection. We note that these probabilities are called public profile probabilities by Press (2009).

10. **Inspection probabilities.** If an inspection takes place, the probability that item i will be inspected, is q_i. We require $\sum_{i=1}^{N} q_i = 1$ and $q_i > 0, \ i = 1, \dots, N$.

To enable a more detailed analysis we define the following probabilities.

11. **Attention.** The sampling probability that item i comes to the attention of the inspector, is denoted by λ_i. We require $\sum_{i=1}^{N} \lambda_i = 1$ and $\lambda_i > 0, \ i = 1, \dots, N$.
12. **Conditional inspection.** Given item i has come to the attention of the inspector, it has probability $\pi_i > 0$ of being inspected.

Note that $q_i, \ i = 1, \dots, N$, result from the two processes described in Assumptions 10 and 11, and that the probabilities concerned are related by

$$q_i = \frac{\lambda_i \pi_i}{\sum_{j=1}^{N} \lambda_j \pi_j}, \quad i = 1, \dots, N. \tag{1.1}$$

1.2. Discussion of the assumptions

In Assumption 3 the probabilities p_i are assumed to be given without error. Of course, in practice this will often not be the case. We will not assess the effects of uncertainty in these probabilities caused by estimation here, as it is our objective to find optimal strategies first.

Assumption 5 does not allow for false positives. We could enhance the models by introducing a parameter representing the probability that an item is incorrectly classified as possessing the specific characteristic Γ while this is actually not the case. Such an addition is left for further research.

Note that each item i with p_i positive could be the Γ-item, and hence should not be excluded from inspection under any procedure. Exclusion would be in conflict with Assumption 9. This implies that procedures using a positive threshold to $p_i, \ i = 1, \dots, N$, are excluded from our study.

Assumption 10 introduces the probabilities that govern the process for selecting the individuals to be inspected in case enumeration is not possible. When it is possible to enumerate the items, one can decide in which order the items have to be inspected. In the models without enumeration the order in which items are inspected, is random and depends on two processes. First it depends on the stochastic mechanism that determines in which order items come to the point of inspection (Assumption 11), secondly it depends on the probability with which the item is inspected, once the item has come to the point of inspection (Assumption 12). If some properties or characteristics of the individuals or items in the population are known, the resulting profiles may be used in determining the conditional inspection probabilities π_i or even the sampling probabilities λ_i. Obtaining an estimate for π_i is commonly associated with the term profiling. The items will be inspected in an orderly sequential fashion but the order in which items are to be inspected, might be determined beforehand. Finally, we point out explicitly that we assume the probabilities p_i, q_i, s_i, λ_i, and π_i to be constant over time and to be the same in repeated trials and for all inspections. In practice, this assumption will often only hold by approximation. In Section 4.4 brief comments will be made on the possible relaxation of this assumption.

1.3. Comparison between approaches

The most important question Press raises in Press (2009) is whether actuarial methods will, from a mathematical or probabilistic point of view, deliver the security levels as expected by government. Subsequently he studies a stylized model of reality and obtains both expected and surprising results. The analyzed models however are formulated mathematically sloppily, what makes determining their practical relevance rather difficult.

Press (2010) analyzes the same kind of model and the same optimization criterion, the average number of inspections, called secondary checks, necessary to catch the malfeasor, but for a network of checkpoints and under the extra constraint of allowing for only M secondary checks. As Meng (2012) shows in his formula (4) Press' formula (4) is valid only in a limiting sense as $N \to \infty$.

In this setting of a maximum of M secondary checks, as several researchers, notably Meng (2012), think, the optimization criterion of minimizing the average number of checks makes hardly any sense anymore. There is always a positive probability that the terrorist, or malfeasor, will go through undetected. So, they propose to optimize the probabilities q_i of being selected for inspection such as to minimize the probability of a terrorist going through undetected. Meng (2012) analyzes this new optimization criterion under the constraint on the number of inspections and obtains some surprising results.

In our research we consider all models presented in Table 1 without a maximum of M secondary checks and with the mean of the number of secondary checks as the optimization criterion. However,