Abstract. While probability sampling has the advantage of permitting unbiased population estimates, many past and existing monitoring schemes do not employ probability sampling. We describe and demonstrate a general procedure for augmenting an existing probability sample with data from nonprobability-based surveys ('found' data). The procedure, first proposed by Overton (1990), uses sampling frame attributes to group the probability and found samples into similar subsets. Subsequently, this similarity is assumed to reflect the representativeness of the found sample for the matching subpopulation. Two methods of establishing similarity and producing estimates are described: pseudo-random and calibration. The pseudo-random method is used when the found sample can contribute additional information on variables already measured for the probability sample, thus increasing the effective sample size. The calibration method is used when the found sample contributes information that is unique to the found observations. For either approach, the found sample data yield observations that are treated as a probability sample, and population estimates are made according to a probability estimation protocol. To demonstrate these approaches, we applied them to found and probability samples of stream discharge data for the southeastern US.

1. Introduction

Water quality surveys are not commonly based on statistical sampling designs. These surveys tend to target specific problems where impact is limited in extent to reasonably small, well-defined areas, and an assumption is made, whether explicitly or implicitly, that purposive samples may be substituted for randomly selected ones. The lack of a statistical sampling design, however, may not be detrimental to the objectives of the survey, if the objectives do not depend on unbiased population parameter estimation for the water quality variables of interest.

A variety of reasons may exist for the paucity of statistically-based sampling plans employed for water quality survey and monitoring work, but little discussion of design alternatives characteristically accompanies reports on the results of survey activities. However, the properties of estimates available from a probability sample (p-sample) make it the protocol of choice when possible. Consequently, interest exists in the development of a procedure to combine data from a nonprobability-
based survey or monitoring program ('found sites') with observations from a true statistical sample and retain the advantages for estimation inherent in the statistical sample. The main benefits that may be provided by such a procedure include: (1) augmenting the sample size of the p-sample; and (2) permitting attributes measured only at the found sites to be estimated for the target population.

Overton (1990) used variable probability sampling theory to analyze the problem of combining two distinct p-samples from the same universe and extended this model to the problem of combining found sites with a p-sample. In addressing found site inclusion, he cautioned that the use of found sites involves an unavoidable replacement of a protocol that assures site representativeness with an unprovable assumption of representativeness. He suggested, however, steps that could be taken to make the assumption as acceptable as possible if the decision to use found site data seems warranted.

1.1. OBJECTIVES

Starting with the principles set forth by Overton (1990), we investigated the use of found site information to augment probability samples. Our purpose was to illustrate this process and develop a general procedure for its use. We demonstrated the process by augmenting a regional p-sample developed for the stream survey of the USEPA's National Surface Water Survey (NSWS) with data from found sites contained in WATSTORE, the vase hydrological and water quality database of the USGS. In demonstrating the use of found site data to complement a p-sample we wanted to:

(1) develop the steps involved in implementing the procedure, and the decisions required at each step,

(2) indicate the trade-offs and costs inherent in these decisions,

(3) note the limitations on extrapolation of found site information to the p-sample,

(4) determine steps that improve the effectiveness of the procedure, and

(5) assess the potential usefulness of the procedure in other applications of interest.

2. Procedure

Here we present a general procedure for enhancing a p-sample with found data. For each step, a general explanation of the procedure is followed by examples from our demonstration. This procedure is schematically illustrated in Figure 1.

2.1. OVERVIEW

The general procedure begins with selection of a found sample and a valid p-sample. Found sites are then chosen from the overall found sample that conform to the population characteristics of the p-sample. These found data are used to