**PART B OF THE SUPPORTING STATEMENT**

# 1. QUESTIONNAIRE OBJECTIVES, KEY VARIABLES, AND OTHER PRELIMINARIES

## 1.1. Survey Objective

The objective of the survey is to collect information to inform decisions regarding how the nation's stormwater regulations should be strengthened and to support the technical and financial feasibility associated with such rulemaking, This Information Collection Request consists of two questionnaires for a single target population: owners and/or developers of new and redevelopment projects. The primary objectives of the owner/developer questionnaires are to: (1) characterize current building and real estate improvement projects including type, location, and size; (2) characterize the prevalence and type of stormwater controls implemented at new development and redevelopment sites to control long term stormwater discharges; and (3) characterize the operations and financial condition of owners and developers that could be subject to revised regulations.

## 1.2. Key Variables

The Owner/Developer Questionnaires request information on the following key variables:

- Project characteristics (e.g., type, location, duration, size, land cover, discharge location and method of stormwater conveyance)
- Long term stormwater best management practices and controls
- Stormwater permit and management requirements
- Firm, establishment, and project level financial information

See Part A, Section 4(b) of this ICR for detailed information on the data items for the questionnaires.

## 1.3. Terminology

In describing the statistical approach for the survey, EPA has used several statistical terms. They are:

- *Element or Sampling unit* is the individual reporting unit from which the questionnaire will collect information. For example, the MS4 questionnaire will be distributed to MS4s, and thus, each MS4 is an element or sampling unit.

- *Target population* consists of every element of interest for the questionnaire. EPA will use the questionnaire data to make statistical inferences about the target population. Statistical inferences include national estimates of certain characteristics, such as

estimated number of municipalities with MS4s. Because of the importance of the target population in reporting outcomes, a principal task in the development of a sample survey design is establishing a clear, concise description of the target population for each questionnaire. The definition should identify every element of the target population so that all non-population elements can be excluded.

- *Sample* is the collective term for the sampling unit selected to receive the questionnaire. There are three types:

  - *Census* is a complete enumeration of all elements in the target population. Information about the target population are obtained directly from the reported values. In other words, statistical estimates are not necessary because information is provided for every element in the population.

  - *Statistical sample* is a statistically selected subset of the target population. In a statistical sample, the elements have a known probability of selection. Statistical inferences about the entire target population can be made from the sample data.

  - *Judgment sample* is an arbitrarily selected subset of the target population. In a judgment sample, the elements are selected based upon particular characteristics of interest. For example, EPA might select several MS4s because of unique and innovative stormwater control programs. Because the data are not statistically representative of the target population, they are not used to make statistical inferences about the entire population. Instead, they are used to provide information about particular aspects of interest. For example, a judgment sample might be used as the basis for costs of innovative control programs. In its surveys, EPA typically receives a few unsolicited (voluntary) completed questionnaires and they are treated as judgment samples.

- *Sample frame* is a list or set of procedures for identifying all elements of a target population. Sample frames are essential to the quality of surveys because sample elements are drawn from them. Sample frames are typically created from one or more data sources. In addition to listing population elements, the resulting sample frame contains information that will be used to draw statistical samples. This information includes addresses and stratification variables used to draw the samples. Usually, each target population is associated with a single sample frame, although EPA is considering the statistically valid option of dual frames as explained later in the section.

# 2. TARGET POPULATIONS, SAMPLE FRAMES, AND GENERAL SAMPLE DESIGNS

This section describes the target populations, sample frames, and general sample designs for each of the questionnaires that EPA is considering in this Information Collection Request (ICR). Section 3 provides more information about the different statistical approaches that may be used in developing the sample design for each questionnaire.

EPA has developed two versions of the Owner/Developer Questionnaire: short and long. The short and the long versions include the same basic questions, with additional, more detailed questions appearing on the long version. The two versions have the same target population, sample frame, and general sample design, and they are discussed in the following sections.

### 2.1 Target Population

The *target population* is the same for both questionnaires. The target population is owners and developers of new and redevelopment projects. Owner is defined as the firm, individual, or institution for which the project is being built. Developer is defined as a person, business, or partnership that controls project design and/or land development activities associated with a project. In some cases, the developer may also be the owner.

### 2.2 Sample Frame

EPA is considering a statistical approach that will select statistical samples independently from two sample frames, or a *dual frame*. Traditionally, questionnaire surveys are conducted by taking a sample from a single sampling frame that lists all known members of a target population. In some cases, it may be necessary or useful to sample from multiple sample frames that, as a whole, cover the target population. This is the case when it is either difficult to create a single sample frame or when there are several different organizations that provide information about different subsets of the population.

Specifically, for the Owner/Developer Questionnaire, EPA is evaluating the use of two data sources either to: 1) merge and create a single sample frame; or 2) select samples from each frame (dual frame approach). EPA's preference is to combine the information into a single sample frame (#1), but the available information may not allow linkages of the same entities in each database to be easily and efficiently determined, making the second approach more practical. For example, a company may be listed with slightly different mailing addresses in each data source, and thus, appear to be separate entities. The advantage of the dual frame approach is that linkage only needs to occur for the members that are selected from each sample frame. This is the case because only those members that are selected will be assigned a survey weight for purposes of data analysis. The next two sections describe the two data sources.

#### 2.2.1 Dun and Bradstreet

"MarketPlace Pro," formerly known as the Duns Market Identifiers (DMI) register, is maintained by Dun & Bradstreet (D&B) and covers the entire United States economy. DMI is a file produced by D&B, Inc., that contains basic company data, executive names and titles, mailing and location addresses, corporate linkages, employment and sales data on over 10 million U.S. business establishment locations, including public, private, and government organizations. DMI is the only comprehensive publicly available database to provide coverage of business establishments. DMI is updated monthly and its coverage of the target population is relatively complete, but often contains out-of-date entries that can introduce inefficiencies in the sample design.

DMI provides the option of choosing alternative organizational levels. DMI defines a headquarters as a business establishment that has branches or divisions reporting to it, and is financially responsible for those branches or divisions. The headquarters record provides the total number of employees for the company, including the employees in the branches. Another corporate family linkage relationship provided by DMI is the subsidiary to parent linkage. A subsidiary is a corporation with more than 50 percent of its capital stock owned by another corporation and will have a different legal business name from its parent company.

Based on both primary and secondary NAICS codes (secondary NAICS codes for a specific entity may include up to 5 NAICS codes other than primary), EPA obtained approximately 740,000 records, screened for duplicates, from the D&B database. The following NAICS codes were used for selection:

- 236115 Single-family Builders, 236116 Multi-family Builders, 236117 Operative Builders, and any business with only a four digit NAICS code and listed as 2361 (i.e., all businesses in 2361 except 236118)
- 236210 Industrial Builders, 236220 (2362) Commercial and Institutional Builders
- 237210 (2372) Land Subdivision
- 237310 (2373) Highway, Street, and Bridge Construction
- 237990 (2379) Heavy and Civil Engineering Construction

### 2.2.3 Reed Connect

EPA also obtained detailed project and firm information from Reed Connect, a product of Reed Construction Data.[1] Reed Connect provides information about nonresidual and multifamily residential projects. Projects tend to be relatively large, requiring subcontractor support. Project data reported by Reed that are relevant to this analysis include project categorization, project location, company contact information, company categorization, project value, and several building characteristics (e.g., site size and constructed square footage).

## 2.3 General Sample Design

As explained in Section 2.2, EPA is considering a dual frame approach to the *statistical sample* design for the owner/developer questionnaire. In this type of design, samples are statistically selected independently from each of two frames (i.e., Reed and D&B). Because the project-level information in the Reed database can better identify which firms meet the target population definition, EPA is considering selecting firms in the Reed database with a higher probability for the long questionnaire than those in the D&B database. It is possible that firms in the Reed database will only receive a long questionnaire if selected, instead of a short questionnaire. Firms in the D&B database will receive either the short or the long questionnaire. EPA also will consider whether the statistical sample from either sample frame should be supplemented by a *judgment sample* to obtain information from owners/developers with unique characteristics.

Because some firms might have many projects, EPA is considering an approach to minimize the burden associated with reporting project-level information. Instead of reporting for an extended

[1] http://www.reedconstructiondata.com/construction-project-leads/reed-connect/

period or a specified number of projects, EPA is considering an alternative that will request each firm to report about its projects that were active on a specified date that EPA will randomly select for each questionnaire. In this manner, EPA would obtain statistically representative data about projects that could be extrapolated to the entire population.

# 3 CONSIDERATIONS IN SAMPLE SELECTION

This section provides more information about the general sample designs identified in the previous section. It describes the precision targets and statistical approaches to selecting the sample.

## 3.1 Precision

*Precision* is the sampling error (variability) associated with estimates calculated from the sample data and extrapolated to the larger target population. One measure of precision is the width of the confidence interval for the estimate. Confidence intervals provide a range of values for a particular estimate that would be likely if the study were repeated an infinite number of times (because, by statistical theory, our sample is only one of many possible samples that could have been selected). Thus, when using 95 percent confidence intervals, 95 percent of such intervals would include the true value, if we could take an infinite number of samples. The precision of the estimates depends on both the sample design and the sample size, that is, the number of elements in the sample.

Because EPA is developing a national rule, it is primarily concerned with the precision of the overall estimates. Consequently, in estimating the overall sample size, EPA intends to impose more stringent requirements for overall estimates than any subpopulation. First, EPA would assume that the sample (unadjusted for nonresponse) would be expected, with a certain confidence level (e.g., 90 or 95 percent), to yield sufficient data to estimate the value of an unknown proportion. EPA is considering a precision target of 90 percent which provides reasonable precision. If EPA were to use a more stringent precision target, such as 95 percent, it would need to collect more data which would increase the burden to the target populations.

Once it determines the overall precision target, EPA then allocates the sample among the different strata. EPA typically requires that each stratum meet a basic level of precision. For example, if the binomial distribution were used, a stratum sample might be selected so that it would be expected, with 90 percent confidence, to yield sufficient data to estimate the value of an unknown proportion to within ±0.15 of its true value for the target population.

## 3.2 Statistical Approaches

This section describes the statistical approaches that EPA is considering for selecting samples for each questionnaire. Depending on the target population characteristics, it is possible that EPA may use a different sample design for each questionnaire. For each design, EPA may use the following approaches, either individually or in combination.

In any sample design, if a stratum has a sample size of less than 10, EPA intends to sample all elements within the stratum.

### 3.2.1 Binomial Distribution

The binomial distribution is often used as the basis of sample designs, and can be used to estimate precision. The binomial distribution applies to situations where there are only two outcomes (yes or no) to a dichotomous question such as "Were stormwater post construction controls that retain stormwater onsite implemented on this project?" The presence or absence of the attribute for a particular project is a dichotomous, or binary, variable. The binomial distribution models these data, based on the notion of obtaining national estimates of the percentage or proportion of projects in the target population (or a subset of the target population) that have a particular attribute. The binomial distribution also provides estimates of the variance that is used to calculate the confidence intervals. Because a proportion of 0.5 (or 50 percent) results in the largest possible variance for the binomial distribution, if EPA uses this approach, it would assume that the probability of one outcome would be 0.5 (e.g., stormwater post-construction controls that retain stormwater onsite were implemented by 50 percent of the respondents). In other words, if the population value is any value other than 50 percent, the survey estimate will be more precise – in statistical expectation – than it would be if the population value is 50 percent.

### 3.2.2 Stratified Sampling

For a stratified sample design, *stratification* is performed by selecting one or more characteristics of interest provided in the sample frame and dividing the members of the population into the strata based on those characteristics. *Stratified sampling* consists of selecting a sample from within each stratum, then combining them to constitute the total sample. There are several benefits that result from stratifying the population, including:

- Ensuring that the sample contains representatives from every stratum;
- Improving the precision of parameter estimates;
- Allowing important parameters to be estimated at the stratum level; and
- Allowing certain subpopulations of particular interest to be sampled at a greater rate than others.

Classifying the target population by all of the stratification variables leads to a number of cells that increases quickly by multiples of the strata categories. For a simple example, see Table 2-1 with six total cells for two gender and three age classifications.

Table 2-1 Simple Example of Two Stratification Variables Producing Six Cells

| | | |
|---|---|---|
| Male Infant | Male Child | Male Adult |
| Female Infant | Female Child | Female Adult |

Cells with small population sizes are statistically inefficient, so EPA intends to classify each of the stratification variables into relatively few categories.

EPA is considering the following variables as stratification variables, but may consider others if appropriate:

- *Ecological regions*: EPA is considering the Level 1 ecological regions established for North America (**www.epa.gov/wed/pages/ecoregions/na_eco.htm#Level%20I**). EPA also considered evapotranspiration, precipitation, and other environmental factors, but considers the ecological regions to best categorize the United States for the purpose of the stormwater survey. To minimize the number of strata categories due to the number of regions, EPA intends to combine the smallest regions with another region in the same general location and/or climatic conditions. EPA also is considering a modification that would designate the Chesapeake Bay as a separate region to address Agency initiatives related to the Bay.

- *Industry Sector* would be determined by the primary business as identified by the primary NAICS code.

- *Size* would be determined by the primary business volume such as sales or revnue.

### 3.2.3 Probability Proportional to Size

The probability proportional to size (PPS) sample design uses size as a factor in selecting the sample. In this design, the sample would include most of the largest elements and fewer of the smaller ones. Size can be defined in a number of different ways, such as population or revenue.

### 3.2.4 Spatial Sampling

In simplest terms, spatial context is the information required to locate a sample point on the landscape, for example, latitude and longitude. For our purposes, there might be benefits to selecting elements in a manner that would cover the entire country and be selected from different watersheds. Although any random sample could accomplish this in a sense, there might be advantages to placing some spatial constraints on the sample so that the spatial distribution of the sample closely matches the spatial distribution of the population. Although EPA could approximate the latitude and longitude for element, other approaches such as systematic sampling within ecoregions may be simpler and easier to implement.

### 3.2.5 Systematic Sampling

Instead of stratifying by ecoregions, EPA is considering systematic sampling to incorporate some spatial context into the sample designs. Systematic sampling involves selecting every $k^{th}$ facility where $k$ is determined by the selection rate. For example, for firms within each ecoregion, EPA might sort the sample frame by state name, and then zip code within each state, and finally randomly sort the firm names within zip code. The next step would be to draw a systematic sample from the sorted list for each ecoregion. In this manner, the sample would be reasonably diverse from a geographical perspective.

# 4. SOURCES OF ERROR, INCLUDING NONRESPONSE

In developing the final sample design, EPA will consider the precision targets for data collected from the target populations. EPA also will consider potential error that could be associated with estimates calculated from the collected data, due to sources of error associated with sampling, such as response rates, as well as non-sampling sources of error, such as processing error. The following sections describe approaches that EPA may consider to minimize error in data collected using the final design.

## 4.1 Response Rates

In developing the sample design, EPA will consider both unit (questionnaire) and item (question) nonresponse. Response rates compare the number of completed questionnaires to the number that were distributed. EPA expects that the unit response rate would be 80 percent or better for this mandatory survey effort.

The survey would be conducted under the authority of the Clean Water Act. The cover letters and instructions would explain the legal authority, responsibility to respond, reasons for the questionnaire, and penalty for nonresponse. EPA would use reminder letters and/or telephone calls to remind respondents of the duty to respond under authority of the Clean Water Act. If possible, EPA would seek the endorsement of the major trade associations, which would be expected to increase the response rate from its members. EPA recognizes that some nonresponse is unavoidable, and in past survey efforts, EPA has waived the duty to respond in extreme and rare cases (e.g., natural disasters) which also might occur for this survey effort. To ensure that it receives enough completed questionnaires to meet its precision targets, EPA intends to adjust the sample sizes upward by 20 percent. In addition, EPA intends to further adjust the sample sizes upward to account for out-of-scope elements (including out-of-business). See Table 2-2. For example, a particular county may administer all of the stormwater programs in its county, and thus, all of the local governments (towns, etc.) located in that county would be considered “out-of-scope” for the questionnaire because they do not administer any stormwater programs.

Table 2-2 Sampled Populations: Out-of-Scope Assumptions

| Questionnaire | Out-of-Scope Assumptions |
|---|---|
| Owner/Developer – Short | 30% |
| Owner/Developer – Long | 15% |

Prior to distributing the detailed questionnaires (units), EPA would probably adjust the initial sample sizes to help ensure that the effective sample sizes (i.e., respondents) would be sufficient for precision requirements. For this reason, EPA would adjust the statistical sample size upwards to account for an estimated nonresponse rate of 20 percent. Nonresponse can result from a number of factors, including undeliverable addresses; out-of-business establishments for which nobody is available to respond; out-of-scope establishments that were incorrectly included in the

sample frame; and refusals. EPA typically evaluates each of these components and adjusts its statistical estimates accordingly.

In addition to increasing the initial sample size, EPA would strive to improve the response rate by sending reminder letters and/or telephone calls. Furthermore, after receiving the responses, EPA intends to adjust the questionnaire weights for any nonresponse.

If the nonresponse rate is greater than 20 percent for any questionnaire, EPA will evaluate whether non-respondents appear to have different characteristics than respondents. EPA would examine these characteristics both for the entire industry and for subgroups in the analyses as recommended in the OMB guidance of January 20, 2006 (www.whitehouse.gov/omb/inforeg/pmc_survey_guidance_2006.pdf). For any differences, EPA intends to determine the major causes, and to incorporate appropriate adjustments for bias. (Bias is the difference between the expected value of an estimate and the true value of a parameter or quantity being estimated. If the data collection process generates estimates that are consistently (or on average) above or below the true value, the data collection process is biased.)

To minimize item nonresponse, EPA's subject matter experts have worked closely with owners/developers to develop questions that would be easy to understand with clearly defined and familiar terms; are formatted in a logical sequence; and would request data that are readily available. In this manner, EPA expects to minimize inaccurate or incomplete response of the questions that can occur due to misunderstanding or misinterpretation of questions and the unintentional skipping of questions by respondents. Additionally, EPA would operate an e-mail helpline and website to assist respondents with the questionnaires. If necessary, EPA would impute responses to key questions in our analyses.

### 4.2 Processing Errors

Processing errors can occur when questionnaire responses are coded, edited, and entered into the database. The design and implementation of the questionnaire database would employ a number of quality assurance techniques to reduce the frequency of such errors. These techniques may include the following:

- Investigate whether web surveys are practical for any of the questionnaires, which would minimize transcription errors from paper copies
- Double-entry keypunch verification on critical questions
- Computerized comparison of selected responses to detect inconsistencies and illogical responses
- Computerized analyses to screen for out-of-range and inconsistent numerical values
- Computerized analyses to detect missing numerical data and missing units

## 5. PRETESTS AND PILOT TESTS

EPA does not intend to pre-test the questionnaire. For more than 30 years, EPA's Engineering and Analysis Division has conducted surveys of numerous sectors to collect information to support regulation development activities in the effluent guidelines program. In past years, EPA has relied predominantly on active participation by trade groups in reviewing the questionnaires. In EPA's experience, such collaboration generally tends to better reflect the sectors at large than pre-tests. For this reason, EPA considers additional review through the pre-test process to be unnecessary for this survey.

## 6. COLLECTION METHODS AND FOLLOWUP

*Please See Part A, Section 5(b) of this ICR for this information.*

## 7. DATA PREPARATION AND ANALYSIS

### 7.1 Data Processing Errors

As explained in Section 6 of Part A of this support statement, EPA may distribute the questionnaires in paper form, electronic PDF, through a letter with a link to the questionnaire for completion online, or some combination. Upon receipt of completed questionnaires, EPA would download the electronic or web responses directly to a database or prepare written responses for data entry. Concurrently, EPA and its contractors would review the questionnaires for completeness and accuracy. As necessary, EPA would perform follow-up calls to clarify inconsistencies in responses. Once the data are entered into a database, numerous manual and electronic QA activities would likely be performed and the results would be provided to engineering and economic staff for further resolution and documentation. This database would then be used to perform data analyses.

### 7.2 Analysis

The data collected through these questionnaires will provide EPA with information to characterize current building, transportation, and real estate improvement projects (i.e., new and redevelopment); long term stormwater controls and best management practices (BMPs) being installed at newly developed and redeveloped projects; state and local long term stormwater programs and requirements (including retrofit of existing development) and the areas covered by these requirements; the current capacity and expenditures by NPDES Permitting Authorities and local authorities to implement, enforce, and maintain long term stormwater programs and controls; and technical, financial, and environmental data needed to quantify the incremental pollutant removals, compliance costs, impacts, and benefits for various regulatory options that EPA might consider in this rulemaking. Ultimately, EPA would use the information to inform whether to expand its national stormwater program and how to best reduce long-term stormwater discharges from new and redevelopment and the built environment.

The objectives of the each questionnaire would be achieved by the statistically-designed sample survey because the resulting inferences and analyses would be as statistically unbiased and as precise as is practicable. EPA would apply sample weights derived from the statistical sample design and adjust for nonresponse to the data during statistical analysis. Weighting the data would allow inferences to be made about the entire target population, including those that did not respond to the questionnaires. Another advantage is that weighted estimates would have smaller variances than unweighted estimates. EPA would use accepted statistical methods for survey statistics, such as those described in *Sampling Techniques* (Cochran, 1977) and *Survey Sampling* (Kish, 1965). EPA would use the data from the judgment sample separately in a qualitative manner.

See Part A, Section 2(b) of this Information Collection Request for a detailed discussion of the technical and economic analyses.

# 8 REFERENCES


Cochran, W.G. (1977). *Sampling Techniques*. New York: Wiley.

Dillman, D. (2000). *Mail and Internet Surveys: The Tailored Design Method*. New York: Wiley.

Israel, G. (1992) "Sampling Issues: Nonresponse," University of Florida, IFAS Extension Electronic Document. Available at: http://edis.ifas.ufl.edu/PD006.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.