GENESYS Sampling System

METHODOLOGY

Epsem - Equal probability of selection method.

Table of Contents

1.0 INTRODUCTION

The Marketing Systems Group (MSG), founded in the Fall of 1987, was formed with the primary objective of developing a random digit dialing (RDD) sampling system capable of fulfilling virtually any methodological requirement. In addition, this system was to be made available as a standalone in-house installation.

Drawing on more than 25 years of varied sampling experience (ranging from governmental and social science research, to advertising and marketing applications), the design and programming of the initial system was accomplished within a few months. Christened GENESYS, the HP(Hewlett-Packard) version went into beta testing in February 1988. Over the next twelve months the family of GENESYS platforms was completed, including systems for the Digital Equipment Corporation VAX, IBM-compatible PCs and the Apple Macintosh.

Since 1989, a number of related GENESYS sampling products have been introduced, including GENESYS for Windows in 1995. Work continues on system enhancements, improved reporting options, expansion of database information, and new sampling modules. MSG continues to expand its products and services, providing access to a wide range of new technologies and integrated sampling solutions through its custom sampling operation.

Among commercial survey researchers, the pivotal nature of sampling methodology is often ignored and there remains widespread ignorance of telephone sampling procedures. It is not surprising that the phrase "rdd sampling" covers a lot of territory - it is used to describe a wide range of telephone sampling processes, most of them being less than rigorous. It is the standard "proposed sampling method," yet rarely is it defined or described in detail; and the actual methodology(ies) is typically production-cost driven. In summary, few commercial research companies have ever developed defensible proprietary RDD sampling systems and few practitioners have had the background, knowledge or inclination to examine telephone sampling processes they implement or propose.

The term "rdd sampling" has become a meaningless generic description covering a broad class of telephone sampling processes. In common parlance the term merely distinguishes the process from the use of a list sample.

> **For consistency, the term "RDD" will be used in this paper to describe only sampling processes which provide known, or directly determinable probabilities of selection. The generic "rdd" will be used to describe processes which are not tractable.**

Fifteen years ago, the first commercial sampling companies were founded as a response to the high cost of designing, building and maintaining an internal, proprietary RDD system. Even the largest such firms marketed rdd samples

based on non-tractable methodologies (i.e., probabilities of selection could at best be inferred, at worst, they were *unequal and indeterminate*). In hindsight, the marketing strategies were brilliant; by convoluting a simple, straightforward sampling process, and by overcomplicating explanations, they were able to foster a dependency based primarily on ignorance. It is not surprising that their clients, comprising much of the commercial survey research industry, have come to place near total reliance on a "branded rdd sample", rather than a methodology.

The problem continues to this day. Most commercial sampling suppliers are unwilling or, frankly, unable to detail their exact sampling methodologies; or to provide their database sources, standards and practices for compilation, maintenance procedures/schedules, etc. This has contributed to the general confusion, lack of knowledge, and the attitude that RDD processes are too complicated to understand. Consequently, a branded rdd sample provides confidence without having to bother with the details. What the industry really needs is truth in labeling.

It was into this atmosphere of "black boxes" and "rdd magic" that GENESYS was introduced. GENESYS, at first, may seem to be a complicated software and database system, but the basic tenets of strict frame definition, tractable sample selection processes, and an open architecture have been adhered to at every step. The result is that the user becomes familiar with the RDD sample design process and discovers that what once was a "black box" is now a clearly defined, straightforward, explicit decision making process.

From the beginning, our goal has been to make GENESYS the *de facto* RDD sampling standard for the survey research industry. Of course, for this to occur, every aspect of the sampling system must be clear, precise, and tractable. A "black box" standard is a contradiction in terms.

2.0 BACKGROUND

Roots of the GENESYS Sampling System go back to the late 1960's to the early telephone RDD methodologies utilized by Chilton Research Services and other commercial research organizations. Telephone-based survey research was in its relative infancy and sampling procedures had just begun to progress beyond telephone directory-based methods.

2.1 Early RDD Sampling Efforts.

In the late 1960's and the early 1970's, development of an efficient national telephone RDD sample was an expensive and time-consuming process. The equivalent of today's BELLCORE DDG file was available, but they were hard copy source documents. With this information, a single-stage unrestricted design was certainly feasible, but its implementation would have been as unproductive as it is today - the probability that a random number is assigned to a residence was, and still is, in the low 20% range.

To increase the productivity of RDD samples, it was necessary to identify the actual residential "working banks" within each NPA/NXX. [Note: the term "working bank" is equivalent to a "hundred series", and denotes the first two digits of the four-digit suffix.]

Even into the early 1980's most telephone exchanges were not digitally switched. Calls were handled by large arrays of bulky and relatively expensive electro-mechanical switching devices. Each individual array, or bank, handled 100 consecutive telephone numbers - thus the derivation of the term "working bank." Because of their cost, telephone companies limited these switches to what was absolutely necessary. This created large numbers of non-working banks, especially among the hundreds of small, local telephone companies. It was not unusual to find exchanges with just two or three residential working banks, providing a pool of 200-300 residential numbers.

Procuring working bank information was time consuming, labor intensive and expensive as the residential working bank data was only available directly from the business office or traffic center associated with each NPA/NXX. This meant that: (1), the business office and traffic center needed to be identified for each NPA/NXX; (2), skilled personnel were required to call the designated offices and elicit the information as to which banks had residential number assignments; and (3), this data needed to be transferred into machine readable format and entered into an appropriate database for access by the RDD program.

So, the development of an RDD system was, for the vast majority of companies, a prohibitively expensive undertaking. Only the largest research organizations could afford the talent, cost, and staff time required to develop and maintain their

own proprietary RDD systems. As a result, these firms had a virtual monopoly on RDD sampling for many years.

Because of the cost involved in obtaining working bank information, these first generation RDD systems relied primarily on a two-stage sampling procedure:

Stage I: NPA/NXX combinations, or PSUs (Primary Sample Units), were selected with equal probabilities of selection.

Stage II: An equal number of random four-digit suffixes were then generated within each of the 1st stage PSUs.

The result was clearly an *epsem* sample. The 1st stage PSUs were of equal size, with each containing 10,000 unique elements (i.e, the suffixes 0000-9999) and, each of these individual elements were given an equal probability of selection in the 2nd stage. The actual number of interviews obtained from each PSU was a function of the number, or density of household assignments. In other words, the 2nd stage resulted in a proportional sample of households across PSU.

Following the generation of random numbers in the 2nd stage, the first two digits of each suffix were checked against the working banks for the associated NPA/NXX. If it corresponded to a working bank the 10-digit number was output, otherwise, it was discarded. Assuming the working bank data was accurate, this process did not affect the statistical characteristics of the sample - the discarded suffixes *were not replaced*, since they corresponded to banks where there were no residences assigned.

In summary, maintaining a large national two-stage RDD sample was time consuming, expensive, and could at times be unreliable - it was not unusual to get conflicting reports on the exact banks containing residential number assignments. Moreover, the updating process was equally difficult and almost as expensive as the original development. However, the result was an RDD sample which theoretically provided an *epsem* sample of all residential telephone numbers, albeit for most companies, at a prohibitive cost.

2.2 The Methodological Schism.

The level of effort required to develop and maintain RDD sampling capabilities severely restricted the number of research companies willing or able to invest the required capital. Moreover, the demand for telephone research and the utilization of an RDD methodology was increasing from both the governmental and commercial sectors.

The result was a classic case of demand exceeding supply. And, as one would have liked to predict, ingenuity on the one hand and the profit motive on the other, worked to increase the supply:

1) Joe Waksberg developed what has come to be known as the Mitofsky-Waksberg method of RDD sampling; and,

2) separately, the first commercial rdd sampling company was founded.

Occurring within a year of each other in the mid-1970's, these proved to be watershed events for both commercial and social science telephone sampling. In their own way, each made RDD sampling widely available. At the same time, they signaled a divergence in methodological paths which has remained into the 1990's.

The Mitofsky-Waksberg methodology was an elegant solution to the problem faced by research organizations with limited resources. By reversing the process to effect a PPS (probability proportion to size) selection of 1st stage units and an equal allocation across PSUs in the 2nd stage, one maintains an *epsem* sample of residential telephone numbers. The procedure eliminated the cost liability associated with obtaining working bank information and, in fact, minimizes the cost implications relating to inclusion of all NPA/NXXs in the sample frame - even those exchanges one can be 99.999% certain contain no residences. The result is a sample frame with theoretic 100% coverage.

It is important to recognize that both early RDD alternatives utilized two-stage designs, and though theoretically equivalent, both had practical difficulties and limitations:

- As mentioned, working bank information was sometimes unreliable and incomplete. Moreover, the sample quickly aged as new NXXs were established and non-working banks were opened to residential assignment. And, with the update process being almost as costly as the original specification, maintenance often waned, decreasing frame coverage.

- On the other hand, the Mitofsky-Waksberg procedure requires near perfect execution of the 1st stage process to achieve its goals. Anyone familiar with the literature or who has implemented a Mitofsky-Waksberg design knows that this is an elusive goal in more ways than one. Moreover, even the relatively successful completion of the 1st stage process leaves the practitioner with implementation difficulties in the second: (1) PSUs which contain too few households; and (2), sample control problems relating to hundreds of small PSUs.

Many of the operational difficulties with the Mitofsky-Waksberg methodology have been mitigated by revised procedures, but sample control remains cumbersome and labor intensive. It remains the preferred RDD sampling procedure for the vast majority of social science and governmental sponsored survey research efforts and is clearly the standard RDD procedure in these sectors.

2.3 Impact of the Commercial Sample Companies.

In contrast, the survey research industry came to rely increasingly on commercial sampling companies, Survey Sampling, Inc. in particular, as their source for rdd samples. In reality, it was the growth of the direct mail industry and especially the large consumer database compilers, such as Donnelley and Metro Mail, that allowed the commercial sampling industry to become established. These White Page derived national databases provided an alternative source for identifying residential NPA/NXXs and most importantly, their associated working banks.

Although not perfect, it did provide a viable alternative to the increasingly difficult, and at times impossible, task of obtaining working bank information from telephone company business offices. The availability of these databases did not however, have any significant impact on the development of new proprietary RDD systems by research organizations. In fact, these databases may have hastened the demise of some RDD systems, while the commercial sampling industry had a positive impact on the establishment of new research companies.

These results were due to a number of factors, some of which are important in understanding the changes that have occurred in the commercial survey research industry over the past fifteen years; the role and composition of the commercial sample suppliers over this time; and, why a system such as GENESYS became feasible.

- Prior to establishment of commercial sampling companies, RDD sampling was primarily the domain of the larger, well-established research companies. RDD capabilities served as an entrance barrier to smaller companies. The availability of working bank information from database compilers did not remove this barrier, but in actuality, heightened it, since now there was an added pre-requisite of computer facilities to process the large data files.

- The commercial sampling companies, who by their very nature had access to the computer facilities necessary, utilized these same databases to make RDD samples widely available.

- The demand for RDD sampling was steadily increasing from research buyers, who were becoming convinced that the biases associated with other methods created risks they were unwilling to accept. At the same time, the availability of commercial rdd sample suppliers removed an entrance barrier into the research supplier market.

- The result was an explosion of new research companies, and new life for many which were having a difficult time converting their organizations from primarily face-to-face to telephone interviewing.

> Many of these new companies were low budget affairs, little more than a president/sales person and a small staff - nearly all work was contracted out: sampling, field work, coding, keypunch and tabulation. In turn, this generated a number of new research related business opportunities (i.e., WATS House, tabulation companies, etc.).

Implied by the above, and obvious to an economist, is that the commercial research business became increasingly cost competitive. At the same time, many of the larger companies lost their competitive advantage because of the widespread availability of rdd samples and other computer-intensive research-related services.

These new smaller companies were deficient in most technical areas, especially sampling. While the larger firms could afford sampling statisticians, the smaller companies, even after they grew, relied upon outside commercial sample suppliers. Sampling was typically not the principal's area of expertise; more than likely he was unfamiliar with constructing an RDD sample and he generally had more important things to concern himself with. Anyway, why should he be concerned when there was an "expert" company out there he could rely on.

The dilemma facing the sampling company was primarily related to interviewer productivity. Under constant competitive cost pressures, research companies could not absorb the vagaries of RDD sampling. They did not understand the RDD sampling process, and consequently were unable to compensate for differences from one area to another in their estimation processes. This lack of knowledge also meant that they did not know what questions to ask.

As a defensive posture, it appears that the sampling companies:

- obfuscated their exact procedures
- made methodological changes which at face value made sense to the layman, but had serious statistical implications
- offered alternative sample selection methodologies that improved productivity by varying sampling rates in uncontrolled, undocumentable, and unjustified ways.

The mystification and reliance on "black box sampling" was aggravated by the lack of knowledge and expertise as to how their product impacted their clients' business - namely the data collection phase of survey research. From experience, we know that researchers' first reaction is to blame the sample, whether due to unexpected survey results, or lower than budgeted interviewer productivity. Without the expertise to evaluate the situation from a production standpoint and consult with their clients on that common ground, the only backstop is what sounds like statistical double-talk. Making matters worse, was the transformation of what should have been relatively simple, straightforward

procedures into a series of complicated processes and "balancing" procedures which no one, neither the client nor the sample supplier, really understood.

Since there were few good alternatives to using a commercial sampling company, the actual sampling procedures, database sources, etc., became of lessor concern. An RDD sample was proposed, the specified sample supplier shipped the "branded rdd sample" (in most cases a non-*epsem* sample), the survey was conducted, results presented and everyone came away fairly happy: It was black box rdd sampling at its best.

There were notable exceptions to the above pattern. Many of the larger commercial research companies, especially those involved in government and social science research maintained their own proprietary RDD samples. Firms such as Chilton Research Services, Market Facts Inc., Opinion Research Corporation and others had built reputations on their sampling capabilities and took pride in their RDD systems. They had the computer resources to take advantage of the information now available via the database compilers. A few other companies, even through the mid-1980's, developed legitimate RDD systems, usually as a result of key personnel demanding an internal system.

3.0 THE GENESYS CONCEPT AND SYSTEM OVERVIEW

The founders of MSG had been able to avoid the problems of dealing with commercial sampling companies. By 1987, they had personally designed, redesigned and/or consulted on the development of six proprietary RDD systems. In fact, discussions regarding the possibility of developing a comprehensive RDD system which could be marketed to different research companies went back as far as 1981. At that time, the major stumbling block was potential clients' lack of requisite computer facilities to store and run such a system. In other words, there was a product, but no means of delivery.

With the advent of the personal computer (PC) and its continual growth and acceptance during the mid-1980's, the situation began to change. Research companies began to adopt PCs, and by early 1987 the question had become whether or not the required software and databases would operate on the new 286 PCs. It should be noted that even at this time relatively few research companies possessed mainframe or mini-computers; for most, PCs were their first in-house computer equipment.

Following the formal establishment of MSG, it was decided to initially develop GENESYS for Hewlett-Packard (HP) mini-computers. This decision was conditioned by the prevalence of CFMC's computer assisted telephone interviewing (CATI) software, which ran on HPs, and the availability of an HP system on which development could proceed. Moreover, it was determined that the basic computer-intensive software and database development would proceed much more efficiently on a mini-computer than on a PC.

The GENESYS Sampling System was conceptualized to incorporate:

- the most comprehensive array of geodemographic variables available for each NPA/NXX,
- a flexible system for sample frame definition and evaluation of alternative stratification strategies, and
- a choice of tractable sampling methodologies, ranging from a single stage *epsem* process to more cost effective, non-*epsem* designs to replicate the more common commercial sampling procedures
- a real time facility for evaluating cost parameters associated with the interplay of stratification, allocation, and alternative RDD generation methods.

The bottom-line is that GENESYS was envisioned as a second generation RDD system, one capable of meeting a wide variety of methodological requirements, offering meaningful sample design facilities and real-time feedback for evaluating cost and design alternatives.

In order for GENESYS to conform with these basic design objectives, it was decided that the following elements must incorporated into its plan:

1) A standard database, which included all residential NPA/NXXs.

2) For each NPA/NXX, a comprehensive series of commonly used geographic codes including Census Division, FIPS State & County Codes, MSAs, ADIs and DMAs.

3) For each NPA/NXX, the Zip Code service distribution, as determined from the actual listed telephone households served.

4) NPA/NXX-level household and population demographic estimates, providing the ability to:

   - create universe estimates for exchange-defined sample frames, rather than using geographic approximations;
   - create and evaluate geodemographic-based stratification alternatives;
   - designate measure-of-size (MOS) variables based on estimated demographic characteristics of individual NPA/NXXs for use with alternative, non-*epsem* RDD methods.

5) The industry's only true single-stage *epsem* RDD method.

6) Alternative non-*epsem* generation RDD methods, which utilize user-defined MOS variables which are retained as part of the sample records for post-survey weighting.

7) Facilities for real-time evaluation of incidence and cost implications of alternative design strategies, including stratification and RDD generation method tradeoffs.

As a result, GENESYS is a comprehensive RDD sample design and generation system which is continually evolving through the suggestions and feedback from an increasingly large base of users.

GENESYS has an installed base of over 100 systems. Some users have multiple systems, one of which is typically used for production, the other(s) being used strictly for sample design evaluation at the proposal stage of the process.

The GENESYS-PC, GENESYS-VAX and GENESYS for Windows systems are used by MSG on a day-in-day-out basis for the production of RDD samples for clients which do not have an in-house installation. This use provides immediate and ongoing feedback for testing new system capabilities, real time emulation of

licensee problems and, the ability to replicate a licensee's design for back-up, consultation, or problem resolution.

Although most research companies have licensed the standard GENESYS database, a number of clients have requested modifications to it. These modifications have ranged from altering the demographic variable classes, and inclusion of special geographic code systems, to changes in the standard working bank cut-off. [Note: the standard cut-off for classification of a working bank as residential is two or more listed telephone households, but this can be modified higher or lower.] The GENESYS Database is completely updated twice per year. This twice-annual update is included in the standard license fees.

The recent availability of GENESYS, a sample design system coupled with true second generation RDD sampling methodology, should raise theoretical as well as practical questions among users of alternative methods. The major advantage in this second generation methodology is a true single stage *epsem* sampling procedure. The theoretical advantages to the survey practitioner should be obvious: residential numbers are included in the sample frame, with known and equal probabilities of selection.

The practical advantages should also be apparent to those involved in the implementation of the Mitofsky-Waksberg procedure or its variants. And, as far as frame coverage issues are concerned, the limitations are both quantifiable, bounded and minimized by the twice-annual update process which incorporates both new exchanges and new working banks.

GENESYS also incorporates a fine implicit stratification, as well as a true random selection process within equal-sized intervals defined sequentially through the sample frame - this is not an *nth* selection. Moreover, all sample generation methodologies, both *epsem* and non-*epsem*, are 100% tractable and replicable [See Section 5.0 for detailed descriptions of sampling methodologies.]

4.0 THE GENESYS DATABASE

The GENESYS Database contains over 42,000 residential telephone exchanges, their associated two digit "working banks", various descriptors of the individual NPA/NXXs geographic service area, and demographic estimates of the population served by each exchange.

The entire database is reconstructed semi-annually. This minimizes potential non-coverage biases which of course are a function of the currency of the NPA/NXX and working bank information - an environment which is continually changing.

4.1 The Ten-Digit Telephone Number.

Telephone numbers contain ten (10) digits, which GENESYS breaks down into four (4) distinct components:

215	525	XX	XX
Area Code (NPA)	Exchange (NXX)	Two Digit Bank	Two Digit Suffix

The Area Code and Exchange - The first 3 digits of the phone number make up the Area Code (NPA). As Administrator of the North American Numbering Plan, Bell Communications Research (BELLCORE) assigns Area Codes.

Following the Area Code is the 3 digit Exchange (NXX). Exchanges are assigned by the local telephone companies, within existing Area Codes.

Two (2) Digit Bank and Two (2) Digit Suffix - The last four digits of the telephone number are assigned by the local telephone companies. Historically, telephone numbers were made available for assignment in banks or blocks of 100 numbers, and were commonly referred to as two-digit banks (i.e., 215-525-00xx through 215-525-99xx).

When a two-digit bank is open, the last two-digits of the number are randomly assigned (i.e., 215-525-0000 through 215-525-0099).

Originally, this method of telephone number assignment was required because of the switching equipment. Advancements in switching equipment technology no longer require phone numbers to be assigned within designated two-digit banks, but it is still a commonly used method

of telephone number assignment, as it allows for convenient segregation of business or other special numbers within an NPA/NXX.

In summary, for every NPA/NXX combination there are:

- 100 possible two-digit bank combinations (00xx through 99xx); and,
- within each 2 digit bank, 100 individual numbers can be assigned (xx00 through xx99).

So, for every NPA/NXX combination, there are 10,000 unique ten-digit telephone numbers (NPA/NXX-0000 through NPA/NXX-9999).

4.2 GENESYS Database Sources.

A variety of sources are required to construct the GENESYS database. Prior to each semi-annual update, the most current version of each is obtained. The following listing details the individual informational source, the critical data contained, as well as the functional application of that data for the GENESYS database.

1) Bellcore Tape

Content: Current valid (i.e, dialable) NPA/NXX combinations

Used For: Identification of valid NPA/NXX combinations.

2) Donnelley Quality Index^2 Database (100% Phone File)

Content: Every directory listed residential telephone number in the United States., and associated FIPS State and County Codes, Zip Code, Zip+4, and Tract/BG (NOTE: both 1980 and 1990 Census Tract definitions are included.).

Used For: Identification of NPA/NXXs; creation of two-digit working bank residential listing counts; and, determination of geographic distribution of NPA/NXX service areas.

3) Claritas/NPDC Update File (National Planning Data Corporation)

Content: Current Zip Code-level household and population demographic estimates.

Used For: Development of estimates of total households served by each NPA/NXX; development of exchange-level demographic estimates.

4) United States Postal Service Tape

Content: Geographic correspondence of valid five-digit Zip Codes.

Used For: Resolution of Zip Code mapping difficulties.

5) Rand McNally Atlas & Claritas/NPDC

Content: Current Metropolitan Statistical Area (MSA/PMSA) definitions, and NECMA definitions for the New England Census Division.

Used For: Assigning NPA/NXX combinations to appropriate MSA.

6) A.C. Nielsen Co. Television Market Report

Content: Designated Market Area (DMA) codes, ranks, and composition; Nielsen County Size Designations

Used For: Assigning NPA/NXX combinations to appropriate Nielson Television Markets.

4.3 The Database Compilation Process.

The process of creating the GENESYS Database is a result of combining and summarizing data from the list of sources outlined in the prior section. The entire process spans four to five weeks and requires the full-time efforts of three to four individuals.

The following sections provide the significant details regarding the compilation and specification process, including standards and quality control routines. As the reader will notice, significant care is taken at each stage of the compilation process to insure that data irregularities and inaccuracies are identified and corrected. However, it would be wrong to assume that 100% of such occurrences, most of which can be characterized as random errors, can be reliably trapped,

identified and corrected. It is standard practice to immediately inform our users of any significant problem identified following issuance of an update.

4.3.1 Creation of the Master Matrix. The Master Matrix serves as the starting point for the entire database compilation process. Each subsequent phase in the process incorporates information from the Master Matrix, and ultimately results in the Master Matrix being updated with new/additional data items.

The initial establishment of the Master Matrix utilizes:

Input Data Source - Donnelley 100% Phone File(65M+ residential telephone number records);

Data Record Information - A ten-digit telephone number and its associated:

FIPS Code - State and County in which the household is located.

Zip Code - Zip Code in which the household is located.

OSLO Indicator (Outside Scheme Limits):

- Y (Yes) for OSLO records
- N (No) for non-OSLO records

OSLO - An OSLO is a telephone number appearing in a white page telephone directory without a complete address - most often because the household requested that their address be omitted entirely. Donnelley imputes the FIPS and Zip Code for that household, utilizing the primary community served by the NPA/NXX. When making these assignments, the FIPS Code for the community is assigned to the OSLO record, while the lowest sequence Zip Code for that same community is assigned.

The steps in the compilation process are as follows:

[1] The 65 Million+ telephone records contained on the Donnelley 100% Phone File are sorted, in ascending order, by ten-digit telephone number. Duplicate telephone numbers are then flagged and;

- For those duplicates with the same FIPS and Zip Code, only one record is retained (i.e., multiple listings of the same phone numbers).

- For duplicates with different FIPS/Zip Codes, a manual check is done to ensure that they are an actual duplication. These situations are most often a result of (1) multiple listings of the same telephone number, with one of the records being an OSLO and/or (2) the same telephone number being used at two different addresses.

[2] OSLO records are then modified by replacing the Donnelley assigned Zip Code with zeros, due to the inaccuracy of the Zip Code assignment method.

[3] Telephone numbers with the same NPA/NXX, FIPS and Zip Codes are combined to form one record (OSLO records are combined based on NPA/NXX and FIPS Code).

[4] For each NPA/NXX-FIPS-ZIP summary record, frequencies are calculated in total and for each of the two-digit bank combinations (i.e., 00-99).

[5] The summary data file is then loaded into the Master Matrix file, and resorted in ascending order by NPA/NXX-FIPS-ZIP. The NPA/NXX-FIPS-ZIP summary records include:

- A total listed household count.
- An array from 00 to 99, containing the listed household frequencies for each of the two-digit banks.

4.3.2 Identification of Valid NPA/NXXs and Two-Digit Residential Working Banks. This phase of the compilation process is designed to: (1) validate NPA/NXX records obtained from the Donnelley file; and, (2), insure that NXX duplications across NPA within the BELLCORE file are identified and resolved.

This phase utilizes the following data sources:

Primary Input Data Source: Master Matrix

Data Record Information: NPA/NXX-FIPS-ZIP record combinations and associated frequencies for each two-digit bank.

Secondary Input Data Source: BELLCORE Tape

Data Record Information: Six-digit NPA/NXX combinations corresponding to potential residential exchanges - cellular, Telco-only, and future additions are eliminated.

The steps in the identification process are as follows:

[1] NPA/NXX-FIPS-ZIP records in the Master Matrix are collapsed to NPA/NXX, along with the listed household frequencies for the two-digit banks (00-99).

[2] The NPA/NXX combinations are reviewed to ensure that all NXXs affected by NPA Splits have been incorporated into records derived from the Donnelley Database.

NPA Splits. When an existing NPA becomes saturated, a new NPA is created by BELLCORE, with specific NXXs from the original NPA reassigned to the new NPA. Subsequently, the telephone numbers in the affected NXXs will have a new NPA. During the defined "grace period", one can access the affected telephone numbers using both the original and new NPA. Following the grace period, only the new NPA will connect to the desired telephone number.

NXXs that have been affected by past or upcoming NPA splits are checked, based on the Bellcore data. This is necessary because the incorporation and coding of NXXs affected by NPA Splits is handled quite differently by Donnelley and BELLCORE:

- During the "grace period", BELLCORE will list the affected NXXs in two NPAs - the original NPA and the newly created NPA. Once the "grace period" is over, the BELLCORE File will list the affected NXXs only with the new NPA.

- There is no standard as to when Donnelley incorporates the NPA Splits, and there are cases where most of the NXXs affected have been incorporated but some have not.

The first problem that is dealt with in this phase is the identification of those NXXs which are double-listed in the BELLCORE File. These are identified by matching the NXX listings for the "old" and "new" NPAs involved in the Split. The matching NXXs are assumed to be valid only for the "new" NPA.

Cautionary Note: Although the local telephone company(ies) involved in an NPA split submits to BELLCORE the NXXs affected by the split (i.e., which NXXs will remain with the "old" NPA and which NXXs will move to the "new" NPA), this list is subject to change throughout the defined Grace Period.

This process is necessary to maintain a one-to-one correspondence between a residential telephone and the ten-digit number assigned. If the NXXs affected by the "split" are ignored, and Donnelley has not completed their conversion, the GENESYS Database could include two NPA/NXX combinations which serve the same residences. By implication, these household telephone numbers would have twice the probability of selection throughout the Grace Period.

Second, the NPA/NXX combinations in the Master Matrix are searched for NXXs affected by NPA splits. If "old" NPA/NXXs are found, they are changed to correspond to the "new" NPA. Duplicate NPA/NXX combinations and their associated working bank counts are collapsed to single records.

[3] The NPA/NXX Master Matrix file is processed against the valid BELLCORE file produced earlier in this phase with NPA/NXXs not found on the BELLCORE file being flagged. These flagged records are then manually reviewed to determine whether the discrepancy is due to:

- an entry error by Donnelley; or,
- the NPA/NXX is functional, but is not recognized by Bellcore. [*Note: This situation occurs with NXXs near or across NPA borders; in these situations, telephone companies may allow local calls across NPA, and the white page directory listings may not be NPA specific. Consequently, Donnelley will at times show an invalid NPA/NXX - these situations can only be resolved by a manual check.]*

If a flagged NPA/NXX combination can be rectified, it is retained; if not, it is purged. Again, the Matrix File is resorted and duplicate NPA/NXX combinations and their associated working bank counts are collapsed.

[4] For each NPA/NXX, the listed household frequencies for the two-digit banks are reviewed. If a two-digit bank has at least one listed household, that bank is classified as a "working bank". Then,

- NPA/NXXs without any working banks are deleted.
- NPA/NXXs with one or more working banks are retained.

[5] For each NPA/NXX retained, a "working bank array" is created. This new array contains each of the possible two-digit banks (00-99), a binary working or non-working designation for each combination, and the corresponding number of listed households in each bank.

Note: A bank is deemed working in GENESYS if it contains at least 1 listed household (LHH). At the sample generation phase of GENESYS, that threshold can be raised (e.g., to 1, 2, 3...). While most

social science applications utilize the 1 LHH threshold, 2 listings is the preferred threshold among commercial survey researchers.[1]

[6] This working bank array is merged into the Master Matrix, replacing the old array which contained the listed household frequencies. Finally, for each retained NPA/NXX, the total number of working banks is calculated, and is placed in the Master Matrix.

[7] The Master Matrix now contains only those NPA/NXXs that contain at least one working bank. And, each NPA/NXX-FIPS-ZIP combination contains:

- Total listed households for that combination
- Working bank array for that entire NPA/NXX
- Number of working banks for that entire NPA/NXX

4.3.3 Development of NPA/NXX Demographic Estimates. One of the most often used portions of the GENESYS Database are the NPA/NXX-level demographic estimates. During this phase of the process, estimates for both total households, total population, and various demographic variables are produced. These estimates are utilized in a number of ways, from basic stratification and sample allocation to establishing MOS values for use in modified RDD procedures.

This phase utilizes the following data sources

Primary Input Data Source: Master Matrix

Data Record Information: NPA/NXX-FIPS-ZIP record combinations and associated Total Listed Household frequency for that combination.

Secondary Input Data Source: Claritas/NPDC (National Planning Data Corporation) Update File

Data Record Information: Zip Codes and associated County and Household/Population demographics (current year estimates)

---

[1]For details on coverage bias and efficiency gains see:
Kulp, Dale W. "Dynamics of List-Assisted RDD Frame Coverage", Proceedings of the American Statistical Association, 1994.
Brick, J. Michael , et al. "Bias in List-Assisted Telephone Samples"; Public Opinion Quarterly, Summer 1995.
Giesbrecht, Lee H., Dale W. Kulp, and Amy W. Starer. "Estimating Coverage Bias in RDD Samples with Current Population Survey (CPS) Data"; 1996 AAPOR Conference.

The steps in the development process are as follows:

[1] The NPDC file is expanded to reflect those Zips Codes that serve more than one county. [Note: NPDC assigns each Zip Code to only one county, based on the plurality of residence.] For each of these Zip Codes, the proportional service by county is estimated based on the associated listed household frequencies in the Master Matrix. These percents are then applied to each demographic variable for that Zip Code, creating Zip-County demographic estimates.

[2] Listed Households for each NPA/NXX record serving the same ZIP-FIPS combinations are summed together to form Total Listed Households counts for each ZIP-FIPS. A Listed Household frequency is then calculated for each NPA/NXX serving that ZIP-FIPS [i.e., Percent of Listed Households for an NPA/NXX-ZIP-FIPS out of the Total Listed Households for that ZIP-FIPS]. This proportion is referred to as the "NPA/NXX-ZIP percent".

[3] The Household/Population frequencies for each demographic variable on the NPDC file can now be allocated across NPA/NXX records based on the "NPA/NXX Zip percent".

[4] NPA/NXX-ZIP-FIPS records (and associated counts by each demographic variable) are combined together based on NPA/NXX, to form one NPA/NXX record. In addition, the NPA/NXX Zip Codes (and associated listed household counts for each Zip Code) are retained.

The Master Matrix now contains unique NPA/NXX records containing:

- Estimated Total Households
- Actual Listed Households
- Estimated Demographic Variable Counts (and percents)
- Working Bank Array
- Total Number of Working Banks
- Zip Codes served (and associated Listed Household Counts)

4.3.4 Final Addition of Geographic Information. Independent of the Master Matrix is the GENESYS Master County File. A variety of information is maintained and updated semi-annually for each county in the United States, .

Each exchange in the GENESYS database is nominally assigned to one county, based on the plurality of listed telephone households as determined by data from the Master Matrix.

Note: Although GENESYS assigns each exchange to one county, all of the information used to determine an exchange's assignment is retained for possible use in projects that require a full roster of counties served by any given exchange. This information allows review of exchanges not completely contained in a single county (and the

associated incidence/coverage) for possible inclusion in a frame's definition.

Once the county definition has been determined for a given NPA/NXX, the following geographic information is appended to that record:

County/Independent City Name: As defined by the Federal Government.

State: State in which county/independent city is located. GENESYS carries the standard two character USPS abbreviation.

FIPS State/County Code: Every county/independent city has a unique FIPS Code. The FIPS Code is always 5 digits long; the first two digits signifies the state and the last three digits the county.

MSA Code: An MSA (Metropolitan Statistical Area) defines a specific geographic area and is assigned by the Federal Office of Management and Budget. MSAs are defined as heavily populated and economically integrated geographic areas. An area's population is the primary determination as to whether it is recognized as an MSA.

MSAs are defined on a county/independent city basis, except in New England where the boundaries are based on minor civil division (MCD). For New England, GENESYS utilizes the alternative NECMA Statistical Areas which are county based; generally these are aggregates of the regular MSA definitions, but they incorporate geographic areas which are not part of a defined MSA.

- If a county is a part of an MSA/NECMA, the standard four-digit MSA Code is attached.

- If a county is not a part of an MSA, the field is blank.

CMSA Code: Like the MSA, a CMSA (Consolidated Metropolitan Statistical Area) is assigned by the Federal Office of Management and Budget. They represent combined MSAs that are economically integrated.

Arbitron ADI Code: The Area of Dominant Influence (ADI) Code is assigned by the Arbitron Company and defines their television markets based on measured viewing patterns. Unlike MSAs, every county in the United States is assigned to an ADI. In the vast majority of cases, an entire county is assigned to only one ADI market, but there are instances where a county is split between more than one television market.

ADI Rank: Each ADI is given a rank, based on Arbitron's estimated number of television households, and ranked in descending order from 001.

DMA Code: The Designated Market Area (DMA) Code is assigned by the A.C. Nielson Company. Like an ADI, DMAs are television market areas. Differences between ADI and DMA geographic definitions are typically limited to counties which define the fringe or boundary of a market. Again, an entire county is usually assigned to only one DMA market, but there are cases where counties have split assignments.

DMA Rank: Each DMA is assigned a rank, based on television households. DMA markets are ranked in descending order beginning with 001.

Nielson County Size: Assigned by A.C. Nielson, the code designates counties as A, B, C, or D:

Code	Description
A	Counties belonging to the top 21 metro areas (based on 1990 Census household counts)
B	Counties not included in (A), that are in metropolitan areas with more than 85,000 households.
C	Counties not classified as (A) or (B), that have more than 20,000 households, or are in metropolitan areas with more than 20,000 households.
D	All Remaining Counties.

Census Division: Census Divisions are comprised of states, with each state assigned to one Census Division. There are ten (10) Census Divisions, coded 1 through 9 and 0.

Census Region: Census Regions are comprised of Census Divisions. There are four (4) Census Regions; Northeast, North Central, South and West.

4.4 Geo-Metro Sequencing of the Database.

The final step in the preparation of the GENESYS Database is the imposition of a strict geographic hierarchy to the 38,000+ NPA/NXXs contained therein. The model for this hierarchy and sequencing is the stratification and ordering usually imposed prior to selection of an area probability sample. The resultant implicit stratification is imprinted on every NPA/NXX File extracted from the Database.

Explicit stratification, by Census Region or Division, is always user-defined at the sample design phase. However, unless the user specifically exports an NPA/NXX file extracted from the database, and alters the sequence by resorting the records prior to sample generation the imprint of the hierarchy will remain.

The underlying structural hierarchy creates twenty implicit strata - a combination of the ten (10) Census Divisions and a metropolitan/non-metropolitan split within each:

Within each of the ten (10) regional metro strata, counties and their associated NPA/NXXs are ordered from those serving the largest MSA/PMSA to those serving the smallest;

- within each MSA/PMSA, exchanges are then ordered by those serving the county (or counties) containing the central city, followed by those serving the remaining non-central city county(ies);
  - and within each county, exchanges are ordered numerically - lowest to highest.

For the ten (10) non-metro strata, states are ordered in a serpentine fashion - North to South/East to West within division;

- within each state, non-metro counties and their associated exchanges are ordered in a serpentine fashion, North to South/East to West;
  - and within each non-metro county exchanges are ordered numerically - from lowest to highest.

Finally, the twenty (20) files are sequenced in Census Division order from 1-10. Thus the database begins with Division 1 Metro, followed by Division 1 Non-Metro, Division 2 Metro and so on through to Division 10 Non-Metro.

The purpose for ordering the GENESYS Database with such a strict geo-metro hierarchy is to insure strict geographic representation, especially within larger geographic sample frames. Moreover, the imposition of even this implicit stratification on the RDD sampling process will tend to reduce the expected sampling variation relative to that of a simple random sample (srs) of the same size.

5.0 RDD SAMPLING METHODOLOGIES

The basic generation methods supported within GENESYS have been designed to provide the first and only set of tractable commercial RDD sampling procedures. The default RDD methodology provides single stage *epsem* samples of telephone numbers regardless of the defined sample frame. By their very nature, these samples are self-weighting in terms of residential number assignments within the set NPA/NXXs comprising the sample frame.

MOD1 and MOD2, which relate to modified RDD processes, are extensions of the *epsem* RDD process. These modified processes utilize user-defined Measure of Size (MOS) variables for each NPA/NXX.

The MOS variable provides the basis for directly varying sampling rates across NPA/NXXs. Although these processes are non-*epsem*, the defined MOS for each NPA/NXX is retained in the sample records for use in post-survey weighting adjustments.

5.1 Single Stage *EPSEM* RDD.

This option represents the default methodology supported by the GENESYS Sampling System. To the best of our knowledge it was the first such procedure available either commercially or even among the few proprietary sampling systems remaining in use.

By its very nature, this methodology provides a known and equal probability of selection to every possible telephone number in the defined NPA/NXX sample frame; the process produces a single stage sample, in that every possible number is implicitly included in the selection array for every sample draw. This implicit inclusion is, however, just a practical convenience which makes the generation process more efficient.

It is not necessary to explicitly show or list every possible telephone number within the defined frame to provide a single stage selection process. Rather, one needs to be able to uniquely relate the relative position within a theoretical sample frame, or selection interval, to one and only one telephone number in the actual sample frame. Any conceptual difficulties should be overcome as the procedure is detailed below.

5.1.1 Conceptualizing the Sample Frame. Prior to initiation of sample generation, the user has defined a sample frame comprised of a set of NPA/NXX combinations, K; has determined the total number of sample pieces required, n'; and, specified the number of replicates, R.

The operative sample frame is the variable number of residential working banks associated with each of the NPA/NXX combinations. The maximum number of

telephone numbers, N^K, is equal to the number of two-digit residential banks, in the defined set of NPA/NXXs times 100 - since there are by definition 100 unique two-digit combinations, (i.e., XX00,XX01, ... XX99) in each working bank, or hundred series:

$$N^K = WB^K \times 100 = \sum_k (wb_k) \times 100 \text{ where,}$$

N^K = total possible 10-digit telephone numbers in the defined sample frame, k;

WB^K = total working banks in the defined sample frame, k;

wb_k = the number of residential working banks in the k^{th} NPA/NXX; and,

k = the number of NPA/NXXs which comprise the defined sample frame, k.

As mentioned previously, the NPA/NXX combinations are in a specified geo-metro hierarchy, and this sequence is maintained in the file of NPA/NXXs:

$$K = K_1, K_2 \ldots K_k$$

Within each of these k NPA/NXXs, we have a specified variable number of working residential banks, wb_K:

$$WB^K = wb_{11}, wb_{12} \ldots wb_{1k}, wb_{21} \ldots wb_{km}$$

And finally, within each of the wb_{km} banks there are exactly 100 two digit suffixes:

$$wb_{11}00, wb_{11}01, wb_{11}02 \ldots wb_{11}99, wb_{12}00 \ldots wb_{km}99$$

By specifying the order of the NPA/NXXs, one has literally specified the location of every potential telephone number relative to all others. By extension, one can envision a 50 State National NPA/NXX sample frame with an exhaustive listing of all possible telephone numbers: from the first available number in the first NPA/NXX in Boston, MA, to the last available number in the last non-metro county in Alaska; a frame containing approximately 207,527,300 distinct elements - 2,075,273 working banks, or hundred series, with 100 two-digits suffixes for each.

The notational expansion clearly shows that the NPA/NXXs comprising the sample frame can be conceptualized as one long string of 10-digit telephone numbers - each unique, and each in a known and replicable position relative to all others. In other words, if it was necessary to identify the ith element, or telephone number in this string, the process would be as follows:

1) Unique Position $= 1 + (i / 100.0)$

2) The whole number portion of the result will always designate the sequential working bank, and by association the NPA/NXX - in other words, the first eight digits of the telephone number;

3) While the two-digit fraction directly indicates the suffix $(00 - 99)$.

This concept is important for understanding various operational aspects of the GENESYS sampling process as well as the single stage *epsem* RDD process.

5.1.2 *EPSEM* Sample Selection Process. The defined sample frame, K, is in actuality an extended string of elements, each element being a 10-digit telephone number occupying a unique and identifiable position. The length of the string is N^K. At this point the user has already determined the sample size, n', to be selected from the defined sample frame. The sample selection process first determines an initial selection interval size, H^K, by dividing the total number of elements, N^K by the desired sample size, n',

$H^K = N^K / n'$ where, $H^K =$ length of the selection interval;

$N^K =$ number of elements in the defined sample frame; and,

$n' =$ the desired sample size.

This operation creates n' equal size selection intervals, or implicit strata, of size H^K. However, since the interval H^K will not necessarily be integral, the implied boundaries will cause telephone numbers to be divided between neighboring intervals. Consequently, the GENESYS system provides three selection options:

1)The interval H^K is modified by truncating any fractional portion, with the new interval, $H^{k'}$, being integral. However, this will increase the expected sample size by a factor r (where $r = H^{k'}/H^k$).The desired sample size is then attained by use of an external double sampling routine.

2) If the interval remains as originally defined, telephone numbers straddling interval boundaries effectively have their probabilities parsed between two neighboring strata, meaning that they have a probability of being selected twice. Consequently, the resultant sample size, will be slightly smaller than desired, due to these dual selections.

3) Alternatively, a telephone number selected a second time, from a neighboring interval can be replaced with another random selection from the same interval.

Although last procedure is not strictly epsem, as are the prior two options, it does always achieve the objective of obtaining the exact desired sample size, n'. Since RDD sampling rates are typically very small, any potential bias resulting from ignoring resultant variations in probabilities of selection will also be small. And, will be offset somewhat, by insuring a selection within each of the n' sampling intervals.

The actual RDD selection process is identical for all the above options:

1) A random number greater than 0, and less than or equal to .0 is then generated, this number is multiplied by the interval size, H^K or $H^{K'}$, providing a pointer to a designated element, n'_1, in the first interval. Dividing this result by 00.0 provides the sequential wb_i bank in which the number is located, while the fractional portion, truncated to two digits, provides the suffix.

2) For the second interval, a new random number is generated, and again multiplied by the interval size. Adding H^K or $H^{K'}$ to the result, and dividing by 00.0, the number now points to a unique element in the second interval, n'_2.

3) The process in step three is repeated until the string is exhausted and exactly n' ten-digit suffixes are generated.

The process effectively segments the string of N^K elements into n' equal selection strata. From each stratum a single element is selected.

[Again, please note that selection interval boundaries will often "split" an element, since the stratum size H^K, is typically not an integral value. Consequently, there is a probability that a "split" element may be selected twice from neighboring strata. Following selection of a number in stratum h_i, the number is checked against the telephone number selected in h_{i-1}; if it is duplicated, the number is discarded and the selection in h_i is repeated until a unique element is selected.]

5.2 Non-*EPSEM* RDD Methods.

The use of non-*epsem* rdd methods is commonplace in the commercial research industry. In fact, the most widely used, most well-known sampling method

marketed by the largest commercial sample supplier is a non-*epsem* rdd sample. The non-*epsem* procedures supported by GENESYS can be employed to replicate many of the "cost-effective" commercial rdd methods. The difference being that the GENESYS procedures are tractable, *employing known but unequal probabilities of selection.*

Although a primary motivation for supporting such methods is the continuing need to replicate other commercial sampling procedures, these methods are eminently useful in their own right. In combination with the NPA/NXX-level demographic estimates, these procedures provide an efficient and tractable means of oversampling NXXs serving households with selected geo-demographic characteristics.

GENESYS provides the user with the ability to explicitly redefine the MOS variable associated with an NPA/NXX. As detailed in Section 4.3, the GENESYS Database contains an estimate of total households and total population, the exact count of directory listed telephone households, as well as a number of demographic variables for each NPA/NXX. These variables can be utilized individually, or in combination, to explicitly define the MOS variable associated with the defined set of NPA/NXXs.

The most popular non-*epsem* commercial rdd methods increase the probability that a particular number will result in a household contact by use of non-*epsem* sampling procedures. This is accomplished by varying sampling rates based on various implicit MOS variables that are correlated with the density of residential number household assignment.

Typically, such methods use the number of directory listed telephone households. In practice, this may not be the most highly correlated variable with density, but it is the most easily obtained. Sampling in proportion to the number of directory listed telephone households, will result in over sampling NPA/NXXs with higher densities and undersampling those with lower densities. What is usually not recognized however, is that the resultant implicit sampling rates are impacted by actual variations in residential unlisted rates:

1) Listed rates vary from 40 to over 80% depending upon geographic location - central cities as well as high income suburban areas typically have higher than average unpublished rates.

2) High growth areas and the NPA/NXXs serving them, typically have high effective unlisted rates because telephone directories are more out-dated. This situation is exacerbated when new NPA/NXXs are created.

In short, uncontrolled sampling processes utilizing listed households or other variables may result in significant sample biases, both on a geographic or demographic basis. Although one achieves data collection cost reductions, by increasing the average likelihood of reaching a household by 10% or more, the

issue of potential sample bias cannot be resolved since neither actual nor relative probabilities of selection are known or reported.

GENESYS supports two non-*epsem*, or modified RDD sampling methods. These alternatives have been appropriately named MOD1 and MOD2.

Although these modified RDD methods are patterned after widely used methodologies, their significance is that they represent the industry's first tractable sampling applications. By assigning an explicit measure of size (MOS) to the individual NPA/NXXs comprising the sample frame, or within a particular stratum, MOD1 produces a single-stage PPS (probability proportional to size) sample of residential telephone numbers. By contrast, MOD2 produces sampling rates proportional to MOS^2.

In both cases, the explicit MOS values are retained in each sample record for use in constructing weighting factors to compensate for the disproportionate sampling.

The following sections detail the methodological procedures employed in these non-*epsem* RDD processes.

5.2.1 MOD1. Operationally, the MOD1 procedure parallels the *epsem* RDD process. However, where the *epsem* RDD process assumes equal measures of size (MOS) across NPA/NXXs, this assumption is now relaxed, allowing for unequal MOS assignments. Whether one examines individual NPA/NXXs or working banks, each has an equal number of elements, 10,000 or 100, respectively. In other words, the *epsem* RDD process assumes a constant, implicit MOS variable associated with each sampling unit.

The GENESYS MOD1 procedure does not alter the operative sample frame. The frame remains as detailed in Section 5.1.1. It comprises the variable number of residential working banks associated with each of the NPA/NXX combinations. However, the sampling rate applicable to working banks comprising each NPA/NXX is a variable determined explicitly by the MOS_k assigned to each respective combination.

For any defined sample frame K, the sum of the k Measures of Size, $MOS1_T$, is defined as

$MOS1_T = \sum_K MOS_k$ where, MOS_k = the user defined Measure of Size associated with the kth NPA/NXX.

The expected sample take from the kth NPA/NXX can then be expressed as

$E[n_k] = n' * (MOS_k / MOS1_T)$ where n' = the desired overall sample size.

And, the sampling fraction, f_k, within the kth NPA/NXX is equal to,

$$f_k = E[n_k] / (WB_k * 100)$$

The previous equations can also be expressed in terms of the relative sampling fraction and a constant.

$f_k = (n' / MOS1_T) * (MOS_k / (WB_k * 100) = c * f'_k$ where, c = a constant based on the desired sample size and the defined sample frame's total MOS1.

f_k = the relative sampling fraction associated with the *kth* NPA/NXX

This is an important result as the inverse of the relative sampling fraction, $f_k'^{-1}$, is included in the output record of each sample telephone number generated, and is intended for use in developing probability-based selection weights.

As with the *epsem* RDD process, there is a determinable maximum sample file yield for the MOD1 process, which is in general less than that for the *epsem* process. The maximum yield is a function of the relative sampling fractions and the constraint that the maximum sampling rate must be less than or equal to one, $f_{max} \leq 1.0$. First, the NPA/NXX with the largest ratio of MOS_k to available numbers is identified as this will provide the limiting sampling rate.

$$f_{max} = \max_k [MOS_k / (wb_k * 100)]$$

The maximum sample file yield can then be determined directly as

$N1_k = MOS1_T * f_{max}^{-1}$ where, $N1_k$ = the maximum sample file yield from the sample frame, K, using a MOD1 process based on the defined MOS1.

Again, the ultimate sampling fractions, f_k, are a function of the desired sample yield, constrained by the maximum sampling rate, $f_{max} \leq 1.0$. As the sample frame's defined MOS1 approaches uniformity in distribution, the maximum sample yield approaches that of the *epsem* process as a limit.

If we define c as the desired sample size, n', f_k reflects the actual sampling rate within each NPA/NXX. And, as indicated previously, the expected yield from each NPA/NXX is

$$E[n_k] = f_k * wb_k * 100$$

The underlying order of the NPA/NXX combinations comprising the defined sample frame are in the identical geo-metro hierarchy, described in the *epsem* RDD process.

$$K = K_1, K_2 \ldots K_k$$

And again, within each of these k NPA/NXXs, we have a specified variable number of working residential banks, wb_k:

$$WB_k = wb_{11}, wb_{12} \ldots wb_{1k}, wb_{21} \ldots wb_{km}$$

And finally, within each of the wb_{km} banks there are exactly 100 two-digit suffixes:

$$wb_{11^{00}}, wb_{11^{01}}, wb_{11^{02}} \ldots wb_{11^{99}}, wb_{12^{00}} \ldots wb_{km^{99}}$$

The specified order of the NPA/NXXs, still uniquely identifies the location of every potential telephone number relative to all others.

However, the MOD1 process is non-*epsem*, and where the implicit selection intervals were conceptualized in terms of equal segments of 10-digit telephone numbers, the MOD1 selection interval is now defined as equal *MOS*1 segments. In other words, the selection interval, $H1^K$ is measured in *MOS*1 units rather than individual telephone numbers.

$H1^K = MOS1_T / n'$ where, $H1^K$ = length of the selection interval;

$MOS1_T$ = sum of MOS_k in the defined sample frame;

n' = desired sample size, $(\leq N1_K)$

The result is n' equal size *MOS*1-based intervals of size $H1^K$, comprised of a variable number of ten-digit telephone numbers. The actual selection interval definition and selection process is sequential, beginning with the first NPA/NXX in the ordered sample frame.

1) A random number RN greater than 0, and less than or equal to 1.0 is generated. This number is multiplied by the interval size, $H1^K$, providing a pointer that is then mapped to an individual element, n'_1, in the first interval. This sample element is identified by accumulating, sequentially, the element-based MOS measures until the indicated total is reached. In other words,

$n'_1 = n_{kj}' \{RN * H1^k = \sum_k \sum_j MOS_k / (WB_k * 100)$ where, n'_1 = sample element selected from the first interval;

j = jth element, or four-digit suffix in the kth NPA/NXX.

2) For the second interval, a new random number is generated, and again multiplied by the interval size. Adding $H1^K$ to the result, the accumulation is continued until n'_2 is identified.

$$n'_2 = n_{kj}' \{H1^K + (RN * H1^K) = \sum_k \sum_j MOS_k / (wb_k * 100)\}$$

3) The process is repeated until the string is exhausted and exactly n' ten-digit suffixes are generated.

$$n'_i = n_{kj}' \{H1^K (i-1) + (RN * H1^K) = \sum_k \sum_j MOS_k / wb_k * 100)$$

The process segments the string of N^K elements into n' equal selection strata. From each stratum a single element is randomly selected. Please note, that selection interval boundaries will often "split" an element, since the stratum size $H1^K$, is typically not an integral value. Consequently, there is a probability that a "split" element may be selected twice from neighboring strata. Following selection of a number in stratum h_i, the number is checked against the telephone number selected in h_{i-1}, if it is duplicated, the number is discarded and the selection in h_i is repeated until a unique element is selected (i.e, the sampling is accomplished without replacement).

5.2.2 MOD2. The MOD2 procedure parallels exactly the MOD1 methodology with the primary difference being the formulation of the standard measure of size variable assigned to each NPA/NXX. The MOD2 methodology results in sampling rates which are the square of those found in the MOD1 methodology.

Again, this procedure does not alter the operative sample frame in any way. The frame remains as detailed in Section 5.1.1. However, the sampling rate applicable to individual working banks comprising each NPA/NXX is determined explicitly by the $MOS2_k$ assigned to each respective combination:

$MOS2_k = MOS_k^2 / (WB_k * 100)$ where, MOS_k = user defined Measure of Size associated with the kth NPA/NXX.

For any defined sample frame K, the sum of the k Measures of Size, MOS2T, is again defined as

$$MOS2_T = \Sigma MOS2_k$$

The expected sample take from the kth NPA/NXX can then be expressed as

$E[n_k] = n' * (MOS2_k / MOS2_T)$ where, n' = desired overall sample size.

And, the sampling fraction, f_k, within the kth NPA/NXX is equal to,

$f_k = E[n_k] / (WB_k * 100)$

The above equations can again be expressed in a form comprised of a relative sampling fraction and a constant.

$$\begin{aligned} f_k &= (n' / MOS2_T) * (MOS2_k / (WB_k * 100)) \\ &= (n' / MOS2_T) * (MOS_k / (WB_k * 100))^2 \\ &= c * f_k' \end{aligned}$$

where c = the constant based on the desired sample size and the defined sample frame's total MOS2.

f_k' = the relative sampling fraction associated with the kth NPA/NXX

The inverse of the relative sampling fraction, $f_k'^{-1}$, is included in the output record of each telephone number generated, and is intended for use in the developing probability-based selection weights.

Again, there is a determinable maximum sample file yield for the MOD2 process, which is, in general, less than that for both the MOD1 and the *epsem* RDD processes. The maximum yield is a function of the relative sampling fractions and the constraint that the maximum sampling rate must be less than or equal to one, $f_{max} \leq 1.0$. First, the NPA/NXX with the largest ratio of $MOS2_k$ to available numbers is identified as this will provide the limiting condition.

$f_{max} = \max_k [MOS2_k / (wb_k * 100)]$

The maximum sample file yield can then be determined directly as

$N2_K = MOS2_T * f_{max}^{-1}$ where, $N2_K$ = the maximum sample file yield from the sample frame, K, using a MOD2 process based on the defined MOS2.

Again, the ultimate sampling fractions, f_k, are a function of the desired sample yield, constrained by the maximum sampling rate, $f_{max} \leq 1.0$. As the sample frame's defined $MOS2$ approaches uniformity in distribution, the maximum sample yield approaches that of the *epsem* process.

If c is defined as the desired sample size, n, f_k reflects the actual sampling rate within each NPA/NXX. And as indicated previously, the expected yield from each NPA/NXX is

$$E[n_k] = f_k * wb_k * 100$$

The underlying order of the NPA/NXX combinations comprising the defined sample frame are in the identical geo-metro hierarchy described previously. In other words, the specified order of the NPA/NXXs, still uniquely identifies the location of every potential telephone number relative to all others.

The MOD2 process is non-*epsem*, and where the implicit selection intervals were conceptualized in terms of equal intervals of 10-digit telephone numbers, the MOD2 selection interval is defined as equal $MOS2$ segments. In other words, the selection interval, $H2^k$ is measured in $MOS2$ units rather than individual telephone numbers.

$H1^K = MOS2_T / n'$ where, $H2^k$ = the length of the selection interval;
$MOS2_T$ = the sum of $MOS2_k$ in the defined sample frame; and,
n' = the desired sample size, $(\leq N2_K)$

The result is n' equal size $MOS2$-based intervals of size $H2^k$, comprised of a *variable number of ten-digit telephone numbers*. The actual selection interval definition and selection process is sequential, beginning with the first NPA/NXX in the ordered sample frame.

1) A random number RN greater than 0, and less than or equal to 1.0 is generated. This number is multiplied by the interval size, $H2^k$, providing a pointer that is then mapped to an individual element, n_1', in the first interval. This sample element is identified by accumulating sequentially element-based MOS measures until the indicated total is reached. In other words,

$n_1' = n_{kj}'\{RN * H2^K = \sum_k \sum_j (MOS_k / wb_k * 100)^2$ where, n_1' = the sample element selected from the first interval;
j = jth element, four-digit suffix in the kth NPA/NXX.

2) For the second interval, a new random number is generated, and again multiplied by the interval size. Adding $H2^k$ to the result, the accumulation is continued until n'2 is identified.

$n_2' = n_{kj}'\{H2^K + (RN * H2^K) = \sum_k \sum_j (MOS_k / wb_k * 100)^2$

3) The process is repeated until the string is exhausted and exactly n' ten-digit suffixes are generated.

$$n_i' = n_{kj}'\{H2^{K_{(i-1)}} + (RN * H2^K) = \sum_k \sum_j (MOS_k / wb_k * 100)^2\}$$

The process segments the string of N^K elements into n' equal selection strata. From each stratum a single element is selected. Please note that selection interval boundaries will often "split" an element, since the stratum size $H2^k$, is typically not an integral value. Consequently, there is a probability that a "split" element may be selected twice from neighboring strata. Following selection of a number in stratum h_i, the number is checked against the telephone number selected in h_{i-1}, if it is duplicated, the number is discarded and the selection in h_i is repeated until a unique element is selected (i.e, the sampling is accomplished without replacement).

5.2.3 Measure of Size (MOS) Manipulation. As detailed in prior sections, the Measure of Size (MOS) is an explicit, user specified, combination of household or population estimates derived for each telephone exchange. And, the MOS explicitly controls the allocation of sample and the sampling rates across the exchanges comprising each sample frame.

In most cases the user will find that the specification of the desired MOS is closely approximated by the demographic categories included in the GENESYS database. However, the MOS definition field does allow for both arithmetic and logical operators. Providing the following types of manipulations:

- Combinations of age and/or income categories;
- Creating MOS values for categories which are not explicitly defined in the database; one example would be estimating the number of households with $60,000+ by apportioning, say 45% of the of the $50-75,000, and adding it to the $75,000+ category, to approximate a $60,000+ MOS;
- Inserting an MOS of zero, for exchanges which meet a specified criteria, or increasing the MOS for certain exchanges; although operations such as this may be better handled through specific stratification and sample allocation procedures, the capabilities are there.

Finally, it is also possible to alter the MOS field to include a specific MOS. This can be accomplished through an arithmetic equation or identity, with or without logical operators.

5.3 Sample File Replication Process.

As detailed in Section 4.4, the GENESYS Database is ordered in a very strict metro-geographic hierarchy. This ordering provides benefits in so far as control

and reduction of sampling variance, and the sample files generated will retain the imprint of this same hierarchy.

If a straight *nth* replication process is employed, it is a certainty that the resultant replicates will be geographically biased to some extent. For example, lower numbered replicates will over represent larger Northeast metro areas in a national sample. This situation is attenuated as sample sizes decrease and/or the number of replicates increases. To eliminate this potential problem, the replication software routine included in GENESYS was designed to provide randomized yet balanced replicates regardless of the size of the sample file or the number of replicates desired.

The user may specify anywhere from 1 to 999 replicates for any sample file. The starting, or lowest numbered replicate can also be designated as any integral value, not necessarily 001. This has a number of applications:

1) Replicates for tracking studies can be "keyed" to week/month.

2) Studies with multiple phases or overlapping geographic areas can be replicate-keyed.

3) If an initial sample contained 25 replicates, and additional sample were needed, replicate numbering for the supplement could begin at 26.

The replicate sequencing facility also has advantages when CATI sample control routines are utilized, as replicate sequencing can be utilized to designate stratum as well as within stratum replications.

The replication procedure utilizes an internal fair random string of all integral values between 1 and 999. Upon the initiation of a sample generation session, a random number between 1 and 999 is generated, as a "location pointer" into this string. Beginning at this location, GENESYS creates a replicate numbering array from the string by selecting all values less than or equal to the user-defined number of replicates, r. When the end of the internal string is reached, the search continues from the beginning of the string, creating a circular search from a random start. The resultant replicate numbers provide an array R, of randomized numbers, from 1 to the total number of replicates, r.

The first random telephone number generated is assigned the first replicate number from the array, R_1; the second telephone number is assigned the second replicate number in the string R_2, and so on until the last number in the array R_r is assigned to the *rth* telephone number generated.

At this point, the replicate numbering array, R, is shifted one place to the "right." In other words, the replicate number stored in R_2 is shifted to R_1; R_3 to R_2; ... R_1 to R_r. The generation process then proceeds until the next r telephone numbers

are generated, whereupon the array elements are again shifted. This process is repeated after every r numbers are generated, until the session is completed.