CHICAGO REGIONAL HOUSEHOLD TRAVEL INVENTORY

White Paper: Sampling Considerations

Primary Authors:
Sudeshna Sen, NuStats
Carlos Arce, NuStats
Keith Lawton, Consultant

Secondary Authors:
Johanna Zmud, NuStats
Stacey Bricka, NuStats

NuStats

3006 Bee Caves Rd., Suite A-300. Austin, Texas 78746
(512) 306-9065 . fax (512) 306-9077 . *www.nustats.com*

Contact: Stacey Bricka, Research Director

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

INTRODUCTION

The Chicago Regional Household Travel Inventory is a comprehensive study of the demographic and travel behavior characteristics of residents in the greater Chicago area. Sponsored by the Chicago Metropolitan Agency for Planning (CMAP) and the Illinois Department of Transportation (IDOT), the study universe is defined as households residing in the Illinois counties of Cook, DuPage, Grundy, Kane, Kendall, Lake, McHenry, and Will. The project has two phases: Design and Data Collection. The design phase took place in the fall of 2006. The full data collection effort will take place in 2007.

The purpose of the design phase of the study was to identify (through research and primary data collection) the most appropriate design and methodological aspects that maximize the quality and validity of the inventory data for modeling purposes. The three main objectives of the design phase were: (1) to validate existing budgetary assumptions regarding data collection efforts anticipated for the full study (and establish new assumptions as necessary), (2) to ensure that the inventory design elements and methods provide for a data set that supports the development of a valid model, and (3) to vet the inventory design recommendations through a series of white papers, supported by both primary and secondary research, using a peer review panel of both topical and regional experts. This document is one of the four white papers developed as part of the study's design phase.

The purpose of the white papers prepared under this design phase is to address specific issues pertaining to the design of the data inventory and supporting data collection effort. Because the data will be used to both update the current regional travel demand model as well as for developing new models, the actual elements contained in the inventory need to meet the needs of both efforts. These white papers serve to delineate those elements that are critical to both efforts. Ultimately, the cost trade-offs, respondent reactions, white paper recommendations, and input from the expert and local peer review panels will be used by CMAP staff to finalize the actual inventory contents.

Each white paper has a primary author team and a secondary author. The primary author team was responsible for ensuring that the document addressed the necessary elements and provided actionable recommendations for the data collection phase. To facilitate this, the primary authors provided the project manager with a list of key questions or design elements for the pilot test (these are discussed below). The secondary author's role was as reviewer, with the specific intent being to balance the paper, to ensure that it was well-rounded and practical in approach and recommendations.

The white papers combine secondary research with primary data collection (through the study pilot) in order to make recommendations on key issues that impact inventory design. These issues were identified at the project kick-off meeting[1], held Tuesday, May 23, 2006 in Austin, Texas and include: (1) inventory content, (2) sampling considerations, (3) maximizing participation, and (4) efficient data collection. Each of these is discussed in a separate document. This paper, focusing on sampling issues, addresses the following issues:

- Frame/frames
- Bias associated with cellular-only households
- Population coverage
- Differences in politics and respondents, and other related issues resulting from the size of the Chicago metro region

[1] This kick-off meeting included the project team members as well as members of the project's expert and peer review panels.

- Definition of household, including treatment of unrelated, student, and multi-family HH
- Newcomers to the region – immigrants
- Stratification (transit riders? Income?)
- Continuous sample/seasonality effects
- Choice-based sample or fully random sample?
- Language needs
- Panel hooks

The team charged with writing this white paper had no specific data needs from the pilot. Rather, these authors felt that a research effort into the utility of a mode-density leadership or alternative sampling approach that accounted for income and mode options as evidenced from other regions where these types of sampling approaches have been used was warranted.

PAPER OBJECTIVES

The objective of the Chicago Regional Household Travel Inventory is to provide data for the continuing development and refinement of the Chicago regional travel demand forecast models. Thus, from a modeling standpoint, it is important that the data reflects the full diversity of the behavioral determinants of travel activity and provide for a statistically valid model. This paper discusses the sampling considerations towards collecting such a rich source of information for use in conventional or new generation modeling efforts.

Sampling is a consideration made in every survey to draw inferences about the population based upon the inferences from the sample. Sampling a population, rather than conducting a census on the study population, saves time and money and hence results in an effective use of the resources. Also, it is a more cost effective approach compared to data collected from a full population census.

Ideally, developing a statistically reliable sample includes identification of the survey population or the universe, identification of the sampling frame, selection of sampling methods, determination of necessary level of precision for one or more data items collected, calculation of sample size and estimation of necessary resources. In addition, sampling includes an assessment of the sample quality, in terms of accuracy and precision of the sample. In particular, from a sample quality standpoint, the goal of sampling is two-fold: (1) to reduce sampling errors that cause the parameter estimates and other measures to be imprecise and second, to reduce non-sampling errors or survey biases that can cause the measurements to be inaccurate[2] and (2) consideration of survey-related factors such as issues related to participation of the respondents, and future extensions to the survey.

To direct the content of the Chicago Regional Household Travel Inventory, we propose a sampling plan by taking into consideration all the aforementioned factors. For clarity, we have grouped these factors into five categories. They are: (1) Population definition and sampling frame issues, (2) Sample design that includes selection of sampling method, selection of sample stratification plan and calculation of sample size, (3) Sample quality assessment in terms of measurement of nonresponse and sampling errors, (4) Participation issues that addresses the language needs of respondents, influence of politics on participation of communities, and participation of new-comers to the region, and (5) Extensions of the survey. The next section discusses each of these issues in detail.

[2] Cambridge Systematics, Inc., Travel Survey Manual. Publication No. FHWA-PL-96-029, prepared for the U.S. Department of Transportation and the U.S. Environmental Protection Agency, July 1996.

SAMPLING ISSUES

POPULATION DEFINITION AND SAMPLING FRAME ISSUES

The population or survey universe represents the entire group of households that is the focus of the study[3]. Ideally, the survey universe for this travel behavior inventory is defined as all households living in the Illinois counties of Cook, DuPage, Grundy, Kane, Kendall, Lake, McHenry, and Will.[4] However, as a practical matter, it is impossible to enumerate all the households in the study area. For instance, though the state records indicate complete listings of persons with active driver's licenses or state Ids, it is impossible to track households that have moved to a new location and have not updated their current addresses on the state records.

Table 1 shows the total household population for the study area as defined above. Note that just two of the 8 counties (i.e., Cook and DuPage) comprise three-fourth of all households in the study area. To summarize, the study universe or the population is comprised of 2,940,007 households.

TABLE 1: COUNTIES IN THE STUDY AREA

County	Total Households	% of Total Households in Study Area
Cook	1,974,181	57.5%
DuPage	325,601	9.5%
Lake	216,297	6.3%
Will	167,542	4.9%
Kane	133,901	3.9%
McHenry	89,403	2.6%
Kendall	18,789	0.5%
Grundy	14,293	0.4%

Source: Census 2000.

Following the identification of the population, it is important to choose a sampling frame that is representative of the population. A sampling frame can be defined as a body of information about the population being investigated that is used as the basis for selecting samples and in subsequent estimation procedures[5]. In the context of household travel surveys, a sampling frame is an up-to-date listing of every household in the population, with identification information such as telephone numbers or addresses. There are three types of sampling frames that can be used for this travel behavior inventory. They are: (1) Random Digit Dial (RDD), (2) Directory/Address-based, and (3)

[3] Handbook of Household Surveys, Revised Edition, Studies in Methods, Series F, No. 31, United Nations, New York, 1984, para. 4.5.

[4] The current Chicago Metropolitan Agency for Planning study boundary excludes Grundy County. It is included in this inventory as it is assigned to the CMAP modeling area, based on an Illinois requirement that all counties be included in a regional travel demand model.

[5] Handbook of Household Surveys, Revised Edition, Studies in Methods, Series F, No. 31, United Nations, New York, 1984, para. 4.6.

Dual Frame (combination of RDD and Address-based)[6]. In order to choose an appropriate sampling frame, it is important to evaluate the advantages and disadvantages of each of the aforementioned sampling frames. In this study, we evaluated the sampling frames based on the following indicators.

1) Coverage – This measure indicates the extent to which the frame includes and/or excludes members of the survey universe. Technically, the RDD frame provides near 100% coverage of listed (urban and rural) and unlisted households with land-based telephones. However, RDD frame cannot provide any coverage of the cell-phone only households. Also, the RDD frame may over-cover households with multiple landlines. The Address-based list, on the other hand, covers non-telephone or cell-phone only households[7]. However, Address-based frames can lead to under-coverage if the addresses are not up-to-date. The dual frame sampling approach utilizes both RDD and Address-based sampling procedures, and thus, provides a comprehensive coverage of the study area with much lower coverage bias than if either frame were used exclusively.

2) Accuracy – This measure indicates the level of precision to which the frame can be used to locate members of the survey universe (for instance, how up-to-date the list is). Both the RDD and the Address-based frame require an up-to-date list of the telephone numbers and addresses respectively. The dual frame, on the other hand, can utilize the MSG's ADVO database to attach address information to the RDD sample, thus providing much more accuracy of households than addresses based on telephone directories. Addresses not matched with a telephone number can safely be assumed to be either unlisted or non-telephone households[8.]

3) Efficiency – This measure indicates the amount of effort required to make contact with members of the survey universe (for instance, the number of sampled elements that must be screened to find a household eligible to be interviewed). Both the RDD frame and the address-based frame sampling approaches are efficient ways of contacting the members of the survey universe. Thus, the dual frame approach, which builds on the strengths of each of these methods, is an efficient sampling approach. Also, as noted above, the accuracy of locating households with specific characteristics using dual sampling frame translates to an efficient method for in-person data collection since the amount of effort needed to screen for these households is reduced.

Therefore, based on our evaluation of the advantages and disadvantages of the sampling frames, and a careful consideration of the sampling objectives, we propose to use a Dual-Frame sample. Dual Frame sampling combines the strengths of the RDD and Address-based methods, and provides considerable savings in cost compared to a single frame with similar precision. Thus, two random sample sets are generated, one from a list of addresses with phone numbers attached and another from a list of addresses without phone numbers using any of the various random generation methods.

[6] The availability of a sampling frame from which to draw a probability sample for Internet surveys is virtually non-existent, so initial contact with a selected household via email is not a consideration for the Chicago Regional Household Travel Inventory.

[7] Data from the Current Population Survey (CPS), October 2003 indicates that more than 10% of households in the CMAP study area counties do not have a landline telephone in the household but do have a cell phone available.

[8] Census 2000 data indicate that the average rate of telephone ownership (i.e., landline or cell phone) in the eight counties to be surveyed in this study is 98% of households, with a low of 96% in Cook County.

SAMPLE DESIGN

Sample design is a vital component in sampling. The key issues to be considered in sample design are selection of the sampling method, selection of sample stratification plan, and calculation of the sample size. In the following sections, we discuss each of these key issues in detail. In particular, we discuss the stratified sampling method, the different types of stratification and calculation of an appropriate sample size that we recommend for this study.

Sampling Method

The selection of a sampling method is interrelated with the broad objectives of the survey; the study population, and the corresponding appropriate sampling frame, and sampling unit; and the desired level of precision[9]. The sampling methods can be broadly classified into (1) Non-probability sampling method[10], and (2) Probability sampling method.

The non-probability sampling method involves the selection of sampling units on the basis of their availability (for instance, willingness to volunteer) or because of the researcher's personal judgment that they are representative. This method of sample selection results in an exclusion of unknown portion of the population (for instance, those who did not volunteer). One of the most commonly used non-probability sampling methods is *convenience sampling*, where the researcher uses all the individuals that are available for survey rather than selecting from the entire population. This sampling method is used when individuals that are of interest to the researcher are otherwise not represented in the sample. In the context of CMAP study, it is very likely that certain community groups such as Latinos would be grossly underreported. Hence, we recommend the non-probability sampling method to survey all the individuals who belong to these communities and volunteer to participate in the survey. The data obtained from this non-probability sample is important from a modeling standpoint to understand the travel behavior of these community groups.

Contrary to the non-probability sampling methods, the probability sampling method is a sampling technique where every sampling unit has some non-zero probability of being selected into the sample. This sampling method allows for statistically valid estimates of population characteristics. In addition, probability sampling methods ensures high levels of coverage, accuracy, and efficiency compared to non-probability sampling methods. Hence, our approach is to select a probability sample of households for this study.

The probability sampling method can be further sub-classified into the following methods of sampling[11]:

- *Simple random sampling* where each population element has the same probability of being chosen.
- *Systematic sampling* where sample items are chosen in a systematic manner (e.g., every 10th name in a telephone directory)
- *Stratified sampling* where the population is divided into smaller groups and a random sample is chosen within each group
- *Cluster sampling* where a sample of groups is selected and every member of the group is selected

[9] G.A. Churchill. Marketing Research: Methodological Foundations, The Dryden Press, 1984.

[10] Non-probability sampling methods are considered less accurate and rigorous compared to probability sampling methods. Hence, the focus of this white paper is on probability sampling methods.

[11] A.J. Richardson, E.S. Ampt, and A.H. Meyburg. Survey Methods for Transport Planning, Eucalyptus Press, 1995.

- *Choice-based sampling*, a special case of stratified sampling where groups are formed based on an endogenous variable.

Among the aforementioned probability sampling methods, the simple random sampling method is considered the most straightforward and commonly used method of sample selection. However, a simple random sample under-represents certain market segments of particular interest such as transit users. For such cases, a stratified sampling method is employed. Further, there are cases where market segments of particular interest are difficult to reach due to its low incidence in the population such as households with transit access by park-and-ride. This necessitates a choice-based sampling method.

In this study, it might be tempting to suggest a simple random sample of households in the study area, as there are over 2.9 million households in the survey universe. Also, due to the wide variety of travel behavior in the study area, a simple random sample would generate sufficient information about most of the data elements in the travel behavior inventory. However, the distribution of the population according to some critical dependent variables would not adequately be captured in a simple random sample. This would be problematic because the development of both conventional travel models as well as more advanced micro-simulation models (if undertaken at some future point) would suffer or not be feasible without adequate sample representation across the full distribution of these variables in the sample. Clearly, a more appropriate stratified sampling method is warranted in this study. The following section discusses the different types of stratification that can be considered in the stratified sampling method.

Sample Stratification

Sample stratification consists of dividing the study population into subsets of market segments (called strata) within each of which an independent sample is selected. The stratification ensures adequate representation of market segments that are of particular interest in the study population with greater degree of precision. In many cases, the strata are homogenous groups of respondents such as respondents with similar socio-economic or travel behavior characteristics. The two most commonly used types of stratification in household travel surveys are:

1) *Geographic Stratification* – This form of stratification is mostly based on political boundaries or may include land-use or transportation-based measures to define the stratification areas. For instance, stratification by counties or by areas defined by transit availability or residential/commercial density.

2) *Demographic Stratification* – This form of stratification is based on the socio-economic characteristics of the sampling unit such as household income categories (upper, middle and lower income groups), automobile ownership categories (zero-vehicle households, one vehicle households and multiple vehicle household), gender, and ethnicity.

In this study, sample stratification is necessary to provide adequate coverage of residential density variation as well as to enable the capture of significant but hard-to-find sub-populations (e.g., transit users, zero-vehicle households, etc.). In particular, sample stratification is required to ensure coverage of:

- Area type measures: Certain area types, such as suburban mixed use or recent transit-oriented development, are relatively rare, and the number of reported trips in such areas may be low without stratification.
- Travel mode: Sufficient representation of the choice variable for modes that are not widely used must be included. While all modes cannot be defined, some main modes, which are relatively rare (walk, bike) could be underrepresented. Transit usage in the Chicago Transit

Authority (CTA) service area is relatively high and thus, is not as difficult to capture in the travel survey as in some other large metropolitan areas. CTA bus has about 296.4 million annual trips, Pace bus has 34.4 million, CTA rail has 147.9 million trips per year and Metra commuter rail has 77.6 million trips. However, transit trips would be harder to capture in some suburban outlying areas without stratification.

- Transit access modes: It is likely that for many transit modes, the reported transit trips will be from areas with good access to transit. The effects of longer access time on transit mode choice, therefore, will be largely unobserved in a sample random sample. This issue is particularly critical in suburban areas where transit shares are much lower. So it is important to account for transit access by park-and-ride and kiss-and-ride.

To ensure that all of the aforementioned issues are accounted for in the travel behavior inventory sample, we propose stratification by composite measures of density and/or mixed use factors. Specifically, we recommend developing composite measures of transit accessibility and walk accessibility, where transit accessibility can be defined as the number of total jobs, retail jobs, and service jobs accessible by transit by 30, 45 and 60 minute time bands, and walk accessibility can be defined as the number of total jobs, retail jobs, and service jobs accessible by walking by 10, 20 and 30 minute time bands. In addition, we propose stratification with minimum sample sizes by county to ensure adequate levels of precision for survey estimates.

Sample Size

The sample size can be defined as a threshold that is required to measure the socioeconomic characteristics and travel behavior of the study population in a precise and accurate manner, and to provide statistically robust inputs to modeling[12]. This definition of sample size depends on the definition of a complete household because a complete household definition determines when the sample size specified for this survey is met. In this study, a household is considered to be complete, when every member of the household has completed all of the travel information and personal details. Following the definition of a complete household, the sample size is determined in the following way.

Assuming a simple random sample, the specification of a "confidence interval of 95% with a precision factor of +/- 5 percent" translates into a sample size of 384 completed interviews. If we were to target this number of interviews per county in the Chicago region, this would call for a total sample size of 384 x 8= 3,072 households. In the event that weighting effects are present, the minimum sample would need to increase from 384 to roughly 500. This latter scenario would call for a total sample size of 500 X 8 = 4,000.

However, as we discussed above, it is important this survey adequately represent travel mode, transit access, and area type variables at the county level. For instance, we expect that some counties, like Cook County, will contain many sub-areas of interest with different transit and walk accessibility, requiring much larger sample sizes. Other counties, such as Grundy County, will contain very few of these categories and thus, require the minimum sample size (N=500). This warrants the design of a detailed stratification plan to specify the exact number of sampled households per county. But, using assumptions of estimates on key variables, we can provide an outlook of the total sample size and the sample size by county (see Table 2).

[12] Cambridge Systematics, Inc., Travel Survey Manual. Publication No. FHWA-PL-96-029, prepared for the U.S. Department of Transportation and the U.S. Environmental Protection Agency, July 1996.

TABLE 2: MINIMUM SAMPLE SIZES BY COUNTY (ASSUMING A ONE-DAY TRAVEL DIARY)

County	Total Sampled Households
Cook	6,000
DuPage	1,500
Lake	1,000
Will	1,000
Kane	600
McHenry	600
Kendall	500
Grundy	500
	11,700

SAMPLE QUALITY

The primary objective of this household travel and activity survey is to collect data to build the Chicago Regional Household Travel Inventory that will enable analysts to accurately estimate the parameters in conventional travel models as well as more advanced micro-simulation models. Ideally, the parameter estimates of these models should reflect the true value of the parameter in the population. However, in practice, the parameter estimates might be biased and inconsistent. The accuracy of these estimates depends upon the sample quality. The sample quality is usually measured by the bias introduced by two factors: non-response and sampling error. This section discusses each of these factors and the amelioration strategies used by NuStats in detail.

Non-Response

Nonresponse is one of the primary concerns in household travel surveys. It leads to inconsistent and biased survey estimates. Nonresponse can be primarily because of two reasons: (1) failure to obtain a specific piece of data from a responding member of the sample[13], also called item nonresponse, and (2) absence of information from some part of the target population of the survey sample[14], also called unit nonresponse, caused by refusals and non-contacts. Item nonresponse, which occurs when data is missing or incorrect, can be minimized by good design and execution of the survey. Unit nonresponse, on the other hand, can be reduced by the use of pre-survey monetary incentives, a pre-notification letter and reminders, training of the interviewers, and increasing efforts to contact households that are difficult to contact, amongst other things.

NuStats has studied and researched non-response issues[15] and has developed time-tested strategies for increasing the likelihood that we will contact an actual person and obtain their consent to participate.

[13] Zimowski, M., R. Tourangeau and R. Ghadialy, *"An Introduction to Panel surveys in Transportation Studies"*, prepared for Federal Highway Administration, 1997.

[14] Black, T., and A. Safir, Assessing Nonresponse bias in the National Survey of America's families, 2000; Harpunder, B.E. and J.A. Stec, Achieving an Optimum Number of Callback Attempts: Cost savings Versus Nonresponse Error Due to Non-Contacts in RDD Surveys, 1999.

[15] Zmud,. "*Designing Instruments to Improve Response*" in Transport Survey Quality and Innovation, Pergammon Press, 2003. Zmud and Arce, "*Item Non-response in Travel Surveys: Causes and Solutions*."Published in Conference Proceedings, Transport Surveys: Raising the Standard, International Conference on Transport Survey Quality and Innovation, Grainau, Germany, May 1997.

We always include public awareness activities in our projects. These activities include producing a study brochure that explains the project, its importance and relevance, how the results benefit the population, and what participation entails. And, also includes endorsements from local community groups. These brochures are not only mailed to some sampled households but also to the media, law enforcement offices, city and county officials, schools, and major employers. When we send the brochure to sampled households, we also send a relevant prenotification letter that is personalized to the household.

We monitor participation during the course of data collection by important geographic, demographic, and socio-economic variables so that we can focus on those segments of the population in any given survey that are not participating fully. We reserve an "incentive fund" to use judiciously to ensure a representative sample. We test, re-test, and revise our survey materials with every project to ensure that our materials are simple, easy, respondent-friendly, and customized to each locale. We make multiple calls to each sampled number. We call back on different days and at different times of day to increase the likelihood that we will reach a person. Our interviewers are trained to be polite, efficient, and helpful during the interviewing process. They have time-tested responses to respondents' questions as well as attempts to evade participating in the survey.

Importantly, dual-frame sampling provides the additional benefit to the CMAP travel behavior inventory of enabling the mailing of an advance letter to most sampled households. Advance letters have been shown to significantly increase participation rates in surveys. NuStats proposes to use the advance letter to not only provide information about the survey to increase participation, but also to invite households to provide their information via modes other than telephone – namely, mail and Internet. However, given the findings of the pilot report, we may use the advance letters only to specific target populations, as most pilot participants cited the recruitment call as the primary reason they participated (25%), with only 14% indicating that the advance mailing swayed their decision. Methods research conducted by NuStats on survey outcomes for several large-scale household travel surveys (e.g., Southern California, Atlanta) have indicated that the use of multiple modes of data collection mitigates nonresponse bias associated with specific types of households (i.e., higher income, larger household size, young adults, two-worker families with children).

The aforementioned amelioration strategies are crucial measures that can reduce the bias due to nonresponse. Following the implementation of these strategies, response rates can be calculated to assess the quality of the sample. In household travel surveys, response rates are calculated at two stages - the *recruitment stage* and at the *travel information retrieval stage* and then multiplied together for an overall response rate. The typical response rates for household travel surveys range from lows of 20% to highs of 46%. In recent household travel surveys, the biggest deterrent to achieving higher response rates has been in making contact with an actual person at a home telephone number. Most call attempts to reach a person end in answering machines, no answers or busy signals.

In this study, based on our experience and taking into consideration the amelioration strategies to be employed by NuStats, we expect to achieve a response rate of 40 percent during the recruitment interview and then to retrieve travel information from 65 percent of all recruited households for an overall response of 26%. With these rates to collect completed one-day travel diaries from 11,700 households, we would need to contact approximately 45,000 households and recruit 18,000 of them to complete trip logs.

Sampling Error

Sampling errors are random errors introduced into the sample because not all the members of the population are included in the sample. It reflects the deviation of the estimation of the population in the sample from its true value in the population[16]. In order to reduce this sample bias, weighting of the data is employed. Weighting is the process of assigning weights to the observations in the sample so that the weighted sample accurately represents the population. From a finite population sampling theory perspective, analytic weights are needed to develop estimates of population parameters and more generally to draw inferences about the population that was sampled. Without the use of analytic weights, population estimates are subject to biases of unknown (possibly large) magnitude. It would be inappropriate, for instance, to treat the survey data as a simple random sample of households in the greater Chicago region, since unequal probability sampling (via the stratification) must be reflected in the construction of estimators, in the evaluation of statistical precision and in other statistical inferences. Weighting compensates for these "departures" from simple random sampling. Consequently, analytic weights will be developed. The common components and features of these analytic weights are as follows:

- Sampling weights – adjusting for probabilities of selection,
- Nonresponse weight adjustments – compensating for differential response rates across adjustment cells, and
- Post stratification adjustments – aligning the weighted sample to known population distributions from census or other reliable data.

Post stratification variables will be specified at a later stage in building the Chicago Regional Household Travel Inventory, but are likely to reflect (at the household level) such factors as:

- Household size;
- Number of vehicles;
- County;
- Household income; and
- Household race/ethnicity

The final analytic weight FW(j) is simply the product of the selection probability, the nonresponse adjustment and the post stratification weight. Finally, if there is a desire to analyze subsets of the database separately, then analytic weights would need to be developed for these subsets of the data.

PARTICIPATION

The participation of the respondents is crucial for collecting data of high quality. This participation can be maximized by addressing the language needs of the respondents, understanding the influence of politics on the participation of certain communities, and facilitating cooperation of newcomers to the region. This section discusses in detail how NuStats will address each of these issues for the Chicago region.

[16] Cambridge Systematics, Inc., Travel Survey Manual. Publication No. FHWA-PL-96-029, prepared for the U.S. Department of Transportation and the U.S. Environmental Protection Agency, July 1996.

Language needs

According to Census 2000 Summary File 3 (STF 3), 6% - 31% of households in the study area counties are linguistically isolated (*i.e.* non-English speaking). The largest percentage of linguistically isolated households is in Cook County (31%). Of these, half of these linguistically isolated households are Spanish speaking, and 3% speak Serbo-Croatian. The remaining 11% of linguistically isolated households speak 16 different languages, including Italian, German, Russian, Polish, Armenian, and Hindi. A detailed review of the language data from all other counties indicates that the majority of the linguistically isolated households are Spanish speaking. For no other language (other than the pocket of Serbo-Croatian households identified in Cook County) does the linguistically isolated percentage reach one percent (1%) of the county's population. Given this data, we propose to conduct the CMAP survey in English and Spanish. The small percentages that the other language populations comprise mean that the chances of randomly hitting one will be rare. So we do not think it would be cost-efficient to translate the survey instruments into these other languages. However, we will consider other languages, as we know that sometimes, political realities can affect the language of surveys.

Participation of communities and influence of politics

The CMAP study area is comprised of communities that are typical of non-responding households (i.e., non-English speaking, very urban, lower income). In order to solicit the participation of these communities, NuStats conducted a series of community meetings in Chicago and surrounding areas. Each meeting targeted a unique demographic known to have under-reported in previous, similar travel and activity surveys. These demographics included African-Americans, predominantly Spanish-speaking Hispanics, predominantly English-speaking Hispanics and Youth. The findings from the community meetings highlight the influence of politics on participation of the following communities:

- *Latino:* Latinos, particularly predominately Spanish-speaking Hispanics, had an intimate sense of community. Meeting participants made it very clear that, given the current political state, it was very likely that Hispanics would grossly under-report. To increase Latino participation, both Spanish and English speaking Hispanics recommended holding "community survey days" where Hispanic community leaders recruited other Hispanics to attend group sessions to learn about the survey and complete the survey on site. Many of the participants volunteered their time and effort in setting up these events.
- *African-Americans:* Their community meeting painted a picture of an African American community characterized by strong ties to family, and, simultaneously, a very independent and civic-minded group. Meeting participants indicated that a key factor in determining the success of the survey would be survey endorsements by civic organizations and civic leaders in the African-American community.

Based on the findings of these community meetings, we would consider using non-traditional, non-probability survey methods, primarily for the Latino community.

Participation of newcomers to the region

The participation of newcomers to the region, primarily, illegal immigrants is difficult to obtain, because the respondents are at considerable risk legally if information they divulge should get into the hand of the authorities. This was reaffirmed in the community meeting of Latinos in the CMAP study area. Given the current politics surrounding immigration, meeting participants communicated the assurance of complete confidentiality with reported data and the necessity for anonymity with survey participation. Clearly, these considerations need to be made while surveying the illegal immigrants.

EXTENSION/BEYOND THE BASE SURVEY

STATED PREFERENCE SURVEY

Stated Preference (SP) Surveys are being increasingly used to understand the behavior of respondents under hypothetical conditions. These SP surveys can answer questions that cannot be asked in traditional Revealed Preference (RP) surveys such as the impact of road pricing on travel behavior. The CMAP travel behavior inventory provides an ideal foundation for links to SP surveys so that “choice” responses could be placed in their actual context. Thus, adding SP extensions to the RP survey can accommodate for the new generation of applied travel demand models. The potential topics that could be considered for SP surveys include congestion pricing, impact of transportation infrastructure and land use on travel choices, vehicle ownership and use, residential location choice, impacts of telecommunications technology on travel choices, and parking location choice, among others.

PANEL SURVEY

Panel surveys are surveys where the same respondents are surveyed on consecutive occasions. This type of survey is ideal for capturing the long-term dynamics in travel demand. Capturing these dynamics is fundamental to understanding how travelers adapt and change in response to their environment, interact with numerous in-home and out-of-home agents, and make decisions. Day-to-day dynamics could be captured by considering a two-day travel log, which would enable examination of differential patterns of inter-household interaction in terms of destination choice or mode choice. However, longer-term dynamics could be addressed only by implementing a panel design. Thus, the current iteration of the CMAP travel behavior inventory could be the baseline, with specific design elements prescribed for converting a portion of the baseline sample to a rotating panel.

RECOMMENDATIONS

The previous section provides a detailed insight into the issues to be considered in sampling for the CMAP study. Based on the analysis of these issues, we have the following recommendations:

POPULATION

The population will represent all the households residing in the CMAP modeling area as defined by eight Illinois counties. Thus, the population or the study universe will be comprised of over 2.9 million households.

SAMPLING FRAME

We propose a Dual Frame sampling for this study. Dual Frame sampling combines the strengths of Random Digit Dialing (RDD) and Directory/Address-based samples. Specifically, Dual Frame sampling combines the 100% coverage provided by RDD frame of the listed and unlisted households with landline telephones, and the coverage of households with no telephones or cell-phone only households provided by address-based frame. Thus, a Dual Frame sample provides a comprehensive coverage of the study area, more accuracy in locating the survey universe and high efficiency in contacting the households in the survey universe.

SAMPLING METHOD:

In this study, our approach is to select a stratified probability sample of households, primarily because a probability sample ensures high levels of coverage, accuracy, and efficiency compared to non-probability samples. In particular, we use a stratified sampling method as opposed to the commonly used random sampling method because the latter under-represents certain market segments of particular interest in this study such as transit users. The stratified sampling method over-samples some strata to ensure that we capture the diversity of the population according to specific geographic and behavioral factors affecting travel activity in the CMAP study area. Thus, within strata and frame, households will be selected with equal probabilities but the combined sample (across strata and frames) will comprise an unequal probability sample of households.

SAMPLE STRATIFICATION:

As activity- and tour-based models are considered, the usefulness of on-board transit surveys diminishes as the tour context cannot be captured. It therefore becomes more important to capture the behaviors of interest in the household survey. A sampling strategy to maximize the capture of behaviors of interest is therefore needed. The use of choice-based sampling gives problems in terms of biased model estimation, and should be minimized. Using screening techniques can give serious problems in both cost and, more importantly, biasing towards low-activity households. The following is a description of a recommended strategy that should yield unbiased results, with an adequate representation of the behaviors of interest by market segment desired for modeling and policy analysis.

In reviewing sampling schemes employed for other similar studies, the problem is usually in adequate representation in the lesser-used modes such as walk access transit, rail transit, walk and bicycle. Transit and walk can be accommodated with the following strategy, which will also maximize bicycle. But Bicycle will probably need special survey techniques (if desired), as will auto-access transit.

We proposed the following stratification scheme for each zone:

1. Calculate retail jobs in the production zone (1/4 section) plus adjacent zones. This will be the pedestrian access number (PAN).

2. Calculate total jobs accessible by transit in (say) 45 minutes. This by creating a matrix of total transit travel time (in-vehicle+walking+waiting) and summing the employment accessible within 45 minutes – becomes a production zone characteristic. For network assignment zones that are comprised of more than one production/attraction zone (if any) assign the travel times to the parent zone to the attraction zones included. This will be the transit access number (TAN).

3. It will be desirable to get a single factor – combined transit and pedestrian accessibility for sampling guidance. A scheme of equal weighting is needed, so that the average value (or possibly median) for transit access and pedestrian access should be calculated and the ratio between these be used to create a value for non-auto accessibility. For example:
   i. Non-Auto Accessibility (NAA)=TAN+PAN*AVGE(TAN)/AVGE(PAN)

4. The households in each production zone can be assigned the NAA value of the zone and then ranked. This list can be used to provide quantile ranges (probably deciles) for all of the households in the region. The households that fall within each decile can then be used to identify the zones that should be "oversampled".

5. A probable starting strategy might be equal samples from: decile1, decile2, deciles 3, 4 and5, and deciles 5 through 10. This would need to be amended if county minima were to be required (that would not be a model-estimation based approach).

6. The strategy would be to track the samples by purpose and mode and age (children and adults) to ensure at least 30 instances of desired market segment behaviors, but preferably 100 instances where the segments are important. If a desired mode-purpose-age group is running low – the sample rates in the higher deciles would need to be increased.

7. For hard to sample groups (auto-access transit, for example) an intercept, choice based approach will need to be utilized (sampling households of users at park and ride lots).

8. The final design of the sampling structure will be dependant on the market segment-mode choice behaviors desired for policy analysis,

In sum, we recommend that sample stratification will be done by composite measures of density and/or mixed-use factors. Specifically, we propose developing composite measures of transit accessibility and walk accessibility. Stratification of the sample by accessibility via transit and walk ensures the coverage of ensures the coverage of households using certain modes of transportation such as walk and bike that are relatively rarely used and hence, are more likely to be underrepresented. In addition, this stratification plan also ensures the coverage of certain area types such as suburban mixed use or recent transit-oriented development that are relatively rare and hence may be underreported.

SAMPLE SIZE

We propose a total sample size of 11,700 for the CMAP study area. The sample size varies by county in order to adequately represent travel mode, transit access and area type variables by county. For instance, we expect that some counties, like Cook County, will contain many sub-areas of interest with different transit and walk accessibility measures, requiring much larger sample sizes. Other counties, such as Grundy County, will contain very few of these categories and thus, require the minimum sample size (N=500).

SAMPLING QUALITY

The sample quality will be ensured by our amelioration strategies to nonresponse. NuStats will use its time-tested strategies for increasing the likelihood of contacting the actual person and obtaining their consent to participate. Importantly, dual-frame sampling will provide additional benefit to the CMAP travel behavior inventory of enabling the mailing of an advance letter to most sampled households that have been shown to significantly increase participation rates in surveys. Based on our experience and taking into consideration the amelioration strategies to be employed by NuStats, we expect to achieve a response rate of 40 percent during the recruitment interview and then to retrieve travel information from 65 percent of all recruited households for an overall response of 26%. With these rates to collect completed two-day travel diaries from 11,700 households, we would need to contact approximately 45,000 households and recruit 18,000 of them to complete trip logs.

In addition to minimizing nonresponse to ensure a good quality sample, we will also ensure that the sampling error – another key factor that affects the sample quality - is minimized. In order to reduce sample bias due to sampling error, we will employ weighting of the sample. Specifically, we will use analytic weights with three components: (1) Sampling weights to adjust for probabilities of selection, (2) Nonresponse weight adjustments to compensate for differential response rates across adjustment cells, and (3) Post stratification adjustments to align the weighted sample to known population distributions from census or other reliable data.

PARTICIPATION

We intend to maximizing the participation of the respondents by addressing the language needs of the respondents, understanding the influence of politics on the participation of certain communities, and facilitating cooperation of newcomers to the region, in the following way:

- Language needs: The Census 2000 statistics indicates that 6% - 31% of households in the study area counties are linguistically isolated, with the majority being Spanish-speaking households. Given this data, we propose to conduct the CMAP survey in English and Spanish.
- Participation of communities and influence of politics: The CMAP study area is comprised of communities that are typical of non-responding households (*i.e.,* non-English speaking, very urban, lower income). In order to solicit the participation of these communities, we conducted a series of four community meetings in Chicago and surrounding areas. These communities included African Americans, Predominantly Spanish speaking Hispanics, Predominantly English speaking Hispanics and Youth. Based on the findings of these community meetings, we would consider using non-traditional, non-probability survey methods primarily for the Latino community.
- Participation of newcomers to region: The participation of newcomers to the region such as illegal immigrants is difficult to obtain. This was reaffirmed in the community meeting of

Latinos in the CMAP study area. Given the current politics surrounding immigration, meeting participants communicated the assurance of complete confidentiality with reported data and the necessity for anonymity with survey participation. These considerations will be made while surveying the illegal immigrants.

EXTENSION/BEYOND THE SURVEY

We recommend adding Stated Preference (SP) survey extension to this survey to accommodate for the new generation of applied travel demand models. The potential topics that could be considered for SP surveys include congestion pricing, impact of transportation infrastructure and land use on travel choices, vehicle ownership and use, residential location choice, impacts of telecommunications technology on travel choices, and parking location choice, among others. Furthermore, we recommend extending this survey to a panel survey to capture the long-term dynamics in travel demand. Thus, the current iteration of the Chicago Regional Household Travel Inventory could be the baseline, with specific design elements prescribed for converting a portion of the baseline sample to a rotating panel.