INTERNATIONAL EXPERIENCE IN JOURNEY-TO-WORK DATA FROM NATIONAL CENSUSES

Ram M. Pendyala
Department of Civil and Environmental Engineering
University of South Florida, ENB118
4202 East Fowler Ave
Tampa, FL 33620
(813) 974-1084
Fax: (813) 974-2957
Email: pendyala@eng.usf.edu

Amlan Banerjee
Department of Civil and Environmental Engineering
University of South Florida, ENB118
4202 East Fowler Ave
Tampa, FL 33620
(813) 974-8727
Fax: (813) 974-2957
Email: abanerj2@eng.usf.edu

Prepared for the *Census Data for Transportation Planning* Conference
May 11-13, 2005, Irvine, California

Draft – May 9, 2005

ABSTRACT

This paper provides an overview of international experience in journey-to-work data from national censuses. In the United States, the census is moving towards a continuous data collection process through the implementation of the American Community Survey (ACS). The move to a continuous survey format raises several issues in the context of the collection and reporting of journey-to-work data. These issues include sampling, sample size, data reliability (particularly for small geography) and accuracy, and disclosure limitations. In this paper, international experience with journey-to-work data, collected as part of national censuses, is reviewed with a view towards identifying lessons, techniques, methods, and solutions that might be useful in the U.S. context.

INTRODUCTION

The United Nations defines a census as: "the complete process of collecting, composing, evaluating, analyzing, and publishing demographic, economic and social data of all the people in a country at specific point in time". It is the scientifically designed official count of the population of a nation usually undertaken once every five or ten years. Today, virtually every country conducts a periodic census to obtain population statistics that can help assess and track the social and economic condition of people. This unique source of information allows central and regional government agencies to formulate policies, undertake planning activities, and allocate resources effectively.

In the 210 year history of the U.S. decennial census, the latest and 22nd Census was taken in 2000. Census 2000, conducted on April 1, 2000, counted 281 million people and 115.9 million households in the 50 states and the District of Columbia.

The Changing U.S. Census

The U.S. Census Bureau is responsible for conducting the decennial census in the United States. The decennial census has two parts: 1) the short form, which counts the population; and 2) the long form, which obtains demographic, housing, social, and economic information. The "short form" information is collected for every individual (i.e., sampling rate of 1:1) in the population while the "long form" information is collected for a 1:6 sample of households in the population. The "short form" includes seven questions for each household: name, sex, age, relationship, Hispanic origin, race, and whether the housing unit was owned or rented. The long form includes questions about ancestry, education, employment, income, place of work, journey-to-work characteristics, and size and nature of the housing unit (U.S. Census Bureau Website).The U.S. census has generally been conducted in the first year of each decade for the past 100+ years.

Until the year 2000, as the census was conducted once every 10 years, the long-form information generally became obsolete within a few years of the census year. The data were near-obsolete by the time the information was released to agencies and public/private enterprises that used it. To keep up with rapidly changing community characteristics throughout the nation, a new strategy that involves conducting a continuous measurement alternative to the traditional decennial U.S. census has been adopted. This alternative continuous measurement survey is known as the

"American Community Survey (ACS)". The ACS provides long-form information every year instead of once every ten years, but with smaller samples in each year. This ongoing monthly survey replaces the traditional long form component of the decennial census.

The distinguishing features of the continuous measurement plan are intended to serve several key purposes (Rust, 1994):

- Virtually continuous data collection operations (instead of starting and stopping every 10 years) ensure benefits for data quality through the maintenance of a permanent enumeration staff and through continuous experience;
- Access to more current census data throughout the decade, except for small geographic units, for which updates of census data might be based on a multi-year (say, 5-10 year) moving average of sample data;
- Reduce the cost and burden of the decade enumeration by removing the need to collect small-area sample data as part of the decennial census; and
- Improve the frequency, timeliness, and quality of small-area sample data

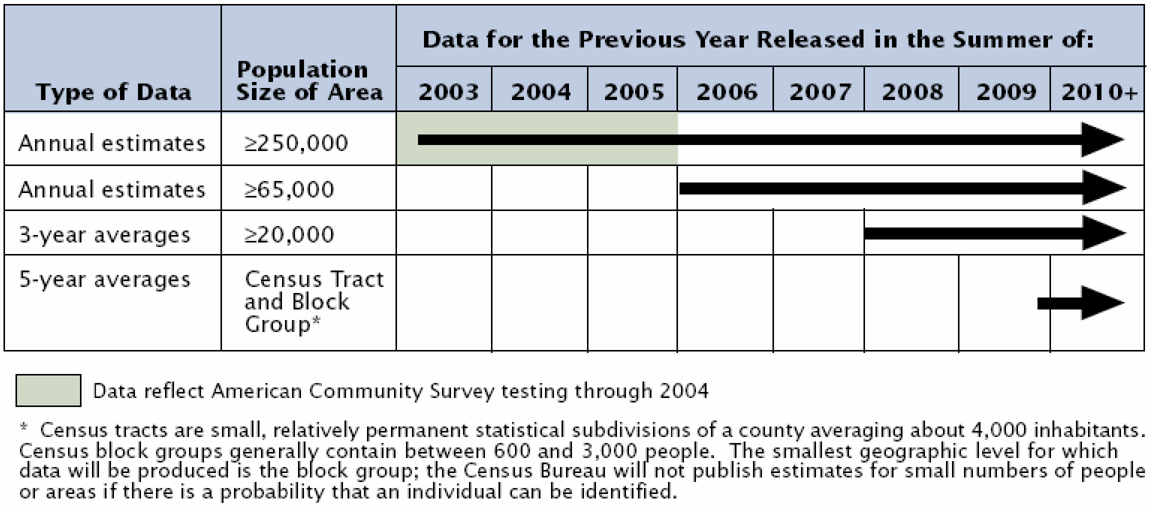
American Community Survey (ACS)

The American Community Survey is a nationwide survey based on continuous measurement approach. It is designed to replace the long form in future decennial censuses by sampling small geographic units throughout the year.

The U.S. Census Bureau selects a random sample from its file of housing unit addresses or Master address file (MAF). An address has about 1 chance in 480 of being selected in any month to participate in the survey. No address is selected more often than once every five years. The questionnaire content itself is similar to that of the decennial census long form (American Community Survey Brochure).

As the Census Bureau is planning to eliminate the long form for 2010 Census, they are working towards the implementation of the ACS as a small, monthly, continuous sample. Starting its full-scale implementation in 2005, the ACS will be mailed to about 250,000 addresses at the start of each month, selected from an updated MAF (Master Address File). Thus the sample adds up to 3,000,000 addresses each year, and to about 15,000,000 over a five year period. The data will be weighted to obtain annual average estimates. The population totals by age, race, sex, and Hispanic origin will be controlled to be consistent with the official intercensal population estimates (Butani, 1999).

Figure 1 shows the American Community Survey timeline by data type. The ACS sample size is potentially sufficient to obtain useful annual estimates for areas down to about 65,000 population. For smaller areas, several years of data may be cumulated to obtain accurate estimates. For very small areas, such as census tracts, which average about 4,000 population each, cumulating five years of data would be typically required. This would give precision close to that of the census long form sample (Butani, 1999).



Type of Data	Population Size of Area	Data for the Previous Year Released in the Summer of:							
		2003	2004	2005	2006	2007	2008	2009	2010+
Annual estimates	≥250,000	🟩 ━	🟩 ━	🟩 ━	━	━	━	━	━▶
Annual estimates	≥65,000				━	━	━	━	━▶
3-year averages	≥20,000						━	━	━▶
5-year averages	Census Tract and Block Group*								━▶

🟩 Data reflect American Community Survey testing through 2004

* Census tracts are small, relatively permanent statistical subdivisions of a county averaging about 4,000 inhabitants. Census block groups generally contain between 600 and 3,000 people. The smallest geographic level for which data will be produced is the block group; the Census Bureau will not publish estimates for small numbers of people or areas if there is a probability that an individual can be identified.

Source: American Community Survey Brochure

Figure 1. American Community Survey Timeline by Data Type

Journey-to-Work Tabulation in the United States

The journey-to-work (JTW) questions asked on the census long form and ACS have a long history of use by the transportation community. In 1960, the Office of Management and Budget (OMB) first sponsored the tabulation and the packaging of the census journey-to-work data at geographies defined by the Census Bureau. Subsequently, the transportation community became integrally involved in the preparation of special tabulations for geographies conducive to transportation planning and modeling (CTPP Historical Perspective Web link).

In the year 1990, a multi-organizational team developed a transportation data package called the Census Transportation Planning Package (CTPP) based on journey-to-work data from the census long form. The objective of CTPP was to provide the transportation planners a comprehensive source of data set for the development of travel demand models and analyzing demographic and travel trends.

CTPP is a set of special tabulations derived from data collected on the long form of the decennial census based on geographies defined by transportation planning agencies. The CTPP summarizes data by home location, work location, and worker flows between home and work locations. Traffic Analysis Zone (TAZ) is the smallest level geography at which the CTPP data are tabulated. The TAZs are similar in size to block groups but are specifically designed to fit local transportation planning and modeling needs (CTPP Historical Perspective Web link).

Three sets of standard tabulations are maintained for CTPP 2000 (http://www.trbcensus.com/ctpp.html).

- Part 1: At Residence (Home End)
    - Content: Characteristics of Persons, All Households, All Workers, Workers by Residence Type, Housing Units, Computed Tables

- Part 2: At Workplace (Work End)
    - Content: Characteristics of All Workers, Workers in Households, Computed Tables
- Part 3: Worker Flows
    - Content: Flows and Times, Computed Tables

The key journey-to-work variables that the transportation community has been using are the following (CTPP Historical Perspective Web link):

- Household: Income, vehicle availability, household size, number of workers, home location
- Person: worker status, age, sex, race, Hispanic origin, disability
- Journey-to-work: work location, mode to work, departure time and arrival time, travel time to work

Impact on Census Transportation Planning Package

In general, the idea of continuous measurement is based on the collection of small-area sample data throughout the decade, rather than once every ten years (Kish 1981, 1990). Although this procedure holds promise over the decennial single-point-in-time data collection system, some issues remain in relation to the potential implications of continuous measurement data products. The motivation of this paper is to review and synthesize international experience with national censuses with a view to address these issues in relation to the transportation element of the census.

The transportation planning community has been a census journey-to-work data user since 1960. Decennial census data on journey-to-work characteristics has been a strong component of transportation planning and programming processes at the federal, state, regional, and local levels. With the major changes in data structure, sampling, and format brought about by the implementation of the ACS, the transportation planning and modeling community has been concerned with possible implications on data accuracy and usage. In response to these concerns, the USDOT conducted a study to address the possible implications of continuous measurement on the use of journey-to-work data in transportation planning. The issues addressed in this study are summarized here (BTS and USDOT, 1996).

Workplace geocoding and small area OD flow tabulation are common to both the ACS format and the traditional long form decennial census format. However, the ACS raises new challenges associated with the reporting, tabulation, accuracy and confidentiality issues of journey-to-work data due to the smaller sampling rate.

Impact on CTPP Part 2: Geocoding of Place-of-Work

The accuracy of place-of-work geocoding has been a major concern in any census (whether or not continuous measurement is implemented). The problem is that the Master Address File continuously corrects and updates residence addresses, but not businesses. Therefore, job locations are often impacted by geocoding errors and allocation inaccuracies, which lead to incorrect commute trip destinations and result in inter-censal job location trends being distorted. In addition, incomplete responses on the census questionnaire also contribute to difficulties in geocoding workplace data from the decennial census (BTS and USDOT, 1996).

The Census Transportation Planning Package (CTPP) 2000 status report documents a new procedure for the allocation of missing place-of-work data in decennial censuses (a brief introduction to the history of CTPP and related confidentiality issues are discussed in the later sections of the paper). In the processing of census data, the work location is based upon both the workplace address and employer name given by respondents on the long-form questionnaires. These responses are then geocoded with a two-phase operation. The first phase is an automated or computer-match operation, which is also called "standard allocation". Standard allocation codes work locations, at a minimum, to a State, County and Place geocode. Many records are allocated down to the Block Group and Traffic Analysis Zone (TAZ) level during the standard allocation process. This information is compared against the address ranges in **T**opologically **I**ntegrated **G**eographic **E**ncoding and **R**eferencing (TIGER) to determine the street side (block) of each respondent's workplace. TIGER is a digital database developed by US Census Bureau to support its mapping needs for the decennial censuses. Records not resolved during this phase move on to a computer-assisted clerical phase, also called "extended allocation". The extended allocation procedure developed for use in CTPP 2000 is targeted at assigning workplace tract and block codes to workers whose work place could not be coded during the standard allocation process. If State or County codes were assigned in the standard allocation process, those codes are not changed in the extended allocation process. However, even with the two-phase operation, not all workers can be coded to all geographic levels (Limoges, 2004).

Impact on CTPP Part 3: Confidentiality of OD Tabulation

The Census Bureau's Disclosure Review Board (DRB) controls the release and usage of specific geographic levels of census data to ensure strict confidentiality of respondents. The CTPP 2000 tabulations were also subject to a number of disclosure avoidance rules imposed by DRB. Two specific disclosure avoidance rules that are identified as having the most adverse implications on the transportation community are 'rounding' and 'thresholding'.

Rounding

All the CTPP2000 tables were applied the following rounding rules except those containing means, medians and standard deviations (Christopher and Srinivasan, 2005).

- Values of zero would remain zero
- Values between 1 and 7 would be rounded to 4
- Values of 8 or more would be rounded to the nearest multiple of 5

Christopher and Srinivasan (2005) have analyzed potential effects of these rounding rules on CTPP2000. Inconsistencies among CTPP table values and systematic undercount of workers are the most critical among them. They also have found in their analysis that rounding to 5 instead of 4 could potentially minimize systematic undercount bias.

Thresholding

Thresholding is the second disclosure suppression technique that was applied to the CTPP2000 OD tabulation. The threshold stated that no data would be provided for any OD pair that had 3 or less records (worker flows) before weighting (Christopher and Srinivasan, 2005).

Several adverse implications are reported (Christopher and Srinivasan, 2005) in the context of applying threshold technique. First of all, threshold value of 3 trips strictly applied on un-weighted OD flows cause a high degree of data suppression even at the medium level of geographic aggregation eliminating most OD pairs for worker tables. Undoubtedly, the rule severely undermines the small area data. In addition to that, as the ACS tests indicate (NCHRP 8-48 test data set tables) lower response rates compared to decennial Census long form (the actual number of unweighted survey records of ACS is 50-60 percent that of the long form), it is apparent that the compounding data loss for small areas due to negative effect of thresholds will be substantial after accumulation of 5 years of ACS data. It is expected that about 40-50 percent of tract-to-tract flows would be suppressed (Christopher and Srinivasan, 2005).

Public Use Microdata Samples & Public Use Microdata Areas

Public Use Microdata Sample (PUMS) data are based on individual census records and is available only at large geographic areas. The U.S. Census Bureau provides two sets of Public Use Microdata Sample (PUMS) files: a 1 percent (1-in-100 Census responses) national characteristics file that provides a full range of detail in population characteristics and 5 percent (1-in-20 Census responses) state files that provide greater geographic detail but less detail in population characteristics. Both of these files are based on Census long form data. These files provide the greatest possible disaggregate resolution, while protecting the confidential nature of the data.

The PUMS data is released for geographic areas called **Public Use Microdata Areas (PUMAs)**. The minimum size threshold for the 5-percent PUMAs is 100,000 persons, and the minimum size threshold for the 1-percent "super-PUMAs" is 400,000 persons (http://www.trbcensus.com/pums.html).

Some of the disclosure-limitation techniques adopted for the public use microdata are (Mackun, 2001) described below.

Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. Swapping is applied to individual records and, therefore, also protects microdata (Mackun, 2001).

Top-coding is a method of disclosure limitation in which all cases at or above a certain percentage of the distribution are placed into a single category. For example, in the case of implementing income top-coding, the recorded income is rounded on a graduated scale and independently top-coded by variable type. The value inserted for observations at and above the top-code will be the state mean of all cases at and above the top-code minimum value. Incomes will then be summed across household members to obtain household totals, without any additional top-coding. The bottom-coding for all income types that can have negative dollar values is set at a maximum negative value of $10,000 (Mackun, 2001).

Geographic population thresholds prohibit the disclosure of data for individuals or households in geographic units with population counts below a specified level.

Age perturbation involves modifying the age of household members for large households (households containing ten people or more) due to concerns about confidentiality.

Details for categorical variables are collapsed if the categories do not meet a specified national minimum population threshold (Mackun, 2001).

Recent trends in transportation planning and modeling indicate a growing emphasis on conducting travel demand analysis at increasingly greater levels of geographical detail. Social and community impact assessment, environmental justice assessment, and social equity analysis place greater demands on the ability to analyze small market segments and geographies. Transportation planning analysis is being conducted at levels of ZIP codes, TAZs, census tracts and blocks, and even down to XY coordinates. However, with the launch of the ACS, the DRB of the U.S. Census Bureau is paying more attention to confidentiality issues and is enhancing disclosure avoidance standards to protect the identity of small geographical units. Consequently, transportation planners, who need data tabulated at fine levels of detail for accuracy in small area studies are losing access to valuable census data because of stringent confidentiality protection measures that suppress such information. For instance, confidentiality issues could limit the use of journey-to-work Census data in transportation planning applications.

From the international experiences, it appears that some nations around the world have been experiencing growing needs for timely Census data at smaller geographical scales. France is one of the first nations in the world in promoting a continuous measurement approach as a replacement for all or part of the traditional census. Other nations (e.g. Australia, New Zealand, Canada) conduct a census every 5 years providing even more timely data compared to the U.S. decennial census. However, some of the same issues regarding tabulation of small area flows and disclosure avoidance are shared across national contexts. Similarities also exist with regard to data post-processing issues such as workplace geocoding, data imputation, and allocation.

This paper is intended to document, within the transportation planning context, national census experiences from around the world with a view to summarize issues that are both common and specific to national censuses. It is envisaged that such a synthesis will provide the basis for learning from experiences of and techniques used in other countries.

INTERNATIONAL EXPERIENCES

In light of the challenges and questions raised by the implementation of the ACS in the context of reporting, tabulation, and accuracy of journey-to-work data, it was considered beneficial to explore experiences of other countries' national censuses with particular focus on journey-to-work data collection, tabulation, and reporting. This section of the paper summarizes national census experiences for selected nations around the world. The countries covered in this paper were included for the following reasons:

- They generally comprise developed economies similar to the United States
- Their national census includes some level of journey-to-work data collection
- Information regarding their national census and journey-to-work data collection, tabulation, and reporting could be obtained.

The intent of the discussion in this paper is to shed light on the following issues:

- For each country, describe:
    - Sampling
    - Frequency
    - Duration and period covered
    - Information content
    - Data published/released
    - Transportation applications/uses of data
    - Challenges encountered/experienced by professionals in those contexts

- For each country, attempt to answer following questions:
    - How are workplace addresses geocoded, and how are origin-destination matrices tabulated? How are non-geocoded work locations imputed or allocated?
    - What small area flow geography is available?
    - How are the small area flow tables used in combination with household travel/activity surveys which lack the sample size for small geographic area reporting?
    - Has data accessibility been impacted by disclosure avoidance rules?
    - For who is the data produced? Is there an additional cost to obtain any part of data? What data are in the public domain?
    - Has a continuous measurement approach been discussed as an alternative? What do they see as potential advantages or disadvantages to an ACS approach for journey-to-work O/D matrices?

Table 1 shows a comparison of geographic divisions across the countries whose national censuses have been reviewed in this paper. Exact definitions for all geographic scales are not readily available; therefore, this table should be interpreted with caution. In general, it appears that the block or its equivalent comprises the smallest level of geography for which census data is reported in any nation.

Table 1. Comparison of International Census Geographic Divisions

U.S.	U.K.	Canada	Australia/N Zealand	France
Nation	Nation	Nation	Nation	Nation
State	Region	Province	State	Province
County	County	Census Division (CD)	Statistical Div. (SD)	Metropolitan Area
		Cen. Sub Div. (CSD)		Urban Unit
Place(City)	District	Cen. Agglomeration (CA)	Statistical Sub Div. (SSD)	
Census Tract		Census Tract		Municipality/ Commune
Block-Group	Ward	Dissemination Area	Stat. Local Area	
TAZ	Output Area (OA)	Block	Cen. Collection Dist.	
Block	Post Code	Block-face	Mesh Blocks	

AUSTRALIA

Census 2001 Background

Australia's 14th national Census of Population and Housing was held on August 7, 2001. The Census of Population and Housing is the largest statistical collection undertaken by the Australian Bureau of Statistics (ABS). The frequency of Australian Census is five years. The census includes all people including visitors in Australia on census night, with the exception of foreign diplomats and their families (Trewin, 2000).

Census Content

The Census 2001 questionnaire was designed to provide information about (Trewin, 2000):

- Respondents personal data: name, address, age, sex
- Household composition
- Heritage, language, religion
- Owning personal computer and internet access
- Participation in education and educational qualification
- Income and employment status
- Journey-to-work
- Person temporarily absent in the home, houses, homes and dwellers

Note: The census planning process led to the inclusion of two new topics in the 2001 Census: ancestry and use of personal computers and the Internet (Trewin, 2000).

Journey-to-Work Data

The employer's address was used to find out what journeys people make to get to work. Employer's names and addresses were destroyed after statistical processing was completed. This information was then combined with the information on work mode choice and car availability and used for the planning purposes of roads and public transport. Daytime population was also estimated using this information because it was considered that many services need to be located where people will be located during the day, rather than where they live (Australia Bureau of Statistics, 2005).

The address of each employed person's usual workplace was used to code the work destination zone. These destination zones were designed by the transport authorities in each state and territory and were used to analyze data on urban transport patterns and plan public transport systems. Destination Zones aggregate to Statistical Local Areas (SLAs), where the SLA is a standard geographic area used in Census products, such as the Working Population Profile. The relevant State/Territory transport authorities (STAs) designed the zones using their own geographic databases; therefore, the indexes and boundaries used to code this data are the property of the STAs. Data at the Destination Zone level were not fully validated by the Australian Bureau of Statistics (ABS) (Australia Bureau of Statistics, 2005).

Journey-to-Work data are primarily used by State Transport Authorities for the analysis of travel patterns within major metropolitan areas, the modeling of fuel usage, the forecasting of public transport patronage, and the analysis of catchment areas for transport routes. The data also assist policy makers in the planning of transport systems, industrial development, and the release of residential and industrial land (Australia Bureau of Statistics, 2005).

Improvements in 2001 Journey-To-Work Data Quality

Reference Period Differences

The variables most commonly cross-classified with 'workplace address' are SLA of Usual Residence - Census Night and Method of Travel to Work. However, the Census questions for these variables related to different reference periods. SLA of Usual Residence - Census Night and Method of Travel to Work both referred to the specific census day. 'Workplace address' referred to the main job held last week, i.e., the week before the census night. The Workplace address question referred to last week rather than the specific census day/night to improve comparability of Census labor force data with other ABS labor force data. Thus, the different reference periods for these questions could produce outliers in the data who are:

- people who changed their place of work between last week and Census Night; or
- people who changed their place of usual residence during the week prior to Census Night; or
- people employed in the week prior to the Census but who were no longer employed on Census day; or
- people who were not at their usual residence on Census Night (Australia Bureau of Statistics, 2005).

Person Workplace Address

One issue identified in the 1996 Census journey-to-work data was that some people answered the question about their employer's workplace address by providing the head office address of their employer rather than the address of the actual location where they worked. For 2001, the same question was modified; rather than asking for 'the employer's workplace address", the question asks for ‘the person's workplace address’. This was done to minimize the number of respondents reporting the address of a head office rather than their actual workplace destination (Australia Bureau of Statistics, 2005).

No Fixed Place of Work

The instructions for the address of workplace question were also changed with a view to improving the quality of the data. For the 2001 Census the instructions included: “For persons with no fixed place of work: if the person usually travels to a depot to start work, provide depot address”. This was done to capture the maximum possible journey-to-work information by accurately coding the journeys to work of those with no fixed workplace address, but who usually journeyed to a specific address in order to begin work (for the main job held last week) (Australia Bureau of Statistics, 2005).

Destination Zone Coding Procedure in 1996

This section describes in detail the process of coding destination zones adopted in Census 1996. The discussion is derived from a report on 1996 Census Data Quality that focused on journey-to-work data (Robertson, 2000).

Note: Journey-to-work Destination Zones aggregate to Statistical Local Areas (SLA) which are then comparable to the Australian Standard Geographic Classification. Finer level of geography than this can not be obtained from the Census database.

The most important issue in evaluating the journey-to-work information in the census is the quality of Destination Zone (DZN) data. DZN information provides the endpoint of respondents’ journeys to work. In order to appropriately appraise the quality of journey-to-work data, it is important to understand the coding procedures used to ascribe a DZN code to a particular workplace address.

Each journey-to-work study area comprised a number of four-digit codes in the range of 0001 to 9999 called DZNs. These DZNs are geographical units designed to represent areas with working populations of at least 100 persons. Although DZNs aggregate to Statistical Local Area (SLA) boundaries, they have no relationship to Collection Districts (CDs). DZNs are based upon an area’s working population, while CDs are designed according to the number of residential dwellings in the area. Thus, an area like a Central Business District (CBD) may contain many DZNs, because the working population is great, but a small number of CDs because few people live there (Robertson, 2000).

Coding of DZN information took place through Computer Assisted Coding in two stages. Coders were instructed to code DZNs based primarily on the workplace address provided by the respondent. Only if a match could not be made using workplace address information then the DZN coding of that respondent would be based upon the provided building or business name if supplied in the index.

The first stage of coding in 1996 used a locality index. Locality, in this case, indicated a city, town, suburb, large employer name, or shopping centre. When a coder sought to allocate a DZN code for the stated workplace address, the locality and State provided by the respondent was entered first. If the entered locality was entirely contained within a DZN, the coder was prompted to assign that DZN code to this response. If the locality overlapped two or more destination zones, the coder was instructed to proceed to Level 2 (street level) coding and to match the street name and number of the workplace address to a DZN code. In CBDs, the entry of a street name alone was frequently insufficient to assign a DZN, as many streets overlapped two or more DZNs. Thus, it was critical that respondents provided accurate street number information (Robertson, 2000).

When circumstances arose in which the provided information was ambiguous or required close inspection before a correct DZN could be assigned, coders sent the response to a Query Resolution Team. Such a referral might take place if a respondent provided insufficient workplace address information to be successfully coded, but included an employer's business name. The query resolution team then searched telephone books for the address of that particular business and used the additional information to get the right index and to assign a DZN. The comprehensiveness of journey-to-work data was therefore increased because partially complete responses were assigned to a DZN as if they contained complete information.

The use of Query Resolution also allowed for frequent updating of indexes. Common examples of changes made during processing involve the updating of indexes to include an acronym or colloquialism for a particular locality, street, building or employer. For example, in 1996 the DZN index entry for the Civil Aviation Authority was updated to include the acronym CAA (Robertson, 2000).

Coding of Insufficient Responses

Responses to the question, 'For the main job held last week, what was the employer's workplace address?' were frequently of insufficient detail to assign a DZN code with complete accuracy. For example, the NSW Transport Study Group reported from a 1991 survey that 23 percent of respondents answered in insufficient detail to be coded (Robertson, 2000). One quarter was widely accepted as a reasonable estimate of imprecise responses, although as many as 45 percent of respondents in a 1989 Census test failed to provide street number details. Several possible factors are identified behind this problem. First, it is highly likely that the specifics of an employer's workplace address (particularly street number) may not be at the forefront of many respondents' minds. Secondly, imprecise responses in 1996 journey-to-work data may have been more frequent as the workplace address question was located towards the end of the census form and because of that, the ability of fatigued or uninterested respondents to answer was diminished. This issue became serious, particularly in CBDs and areas of high density employment where address specifics are highly required. Given the frequency of imprecise responses, the ability of the indexes to cope with the range and quality of employer addresses provided by respondents is the great determinant of the quality of journey-to-work data (Robertson, 2000).

Incomplete Address Responses

The imprecise workplace address responses were dealt with two different ways. The first is to assign a mappable DZN to every incomplete response based upon a decision rule during the index building process. Such an action gives the illusion of uniformly high quality data, which may be more acceptable to users. Alternatively, imprecise responses may be left as stated and assigned to various non-mappable 'dump' codes. Thus the data might be incomplete, but accurately reflects the responses provided by respondents. The following example explains the methodologies applied to New South Wales (NSW) in 1996 (Robertson, 2000).

In NSW, if a respondent worked in a CBD region and provided insufficient address information, that person would be assigned to a non-mappable dump zone. The DZN codes for CBD dump zones fell in the range 6000-7999. Outside CBD areas, respondents who provided insufficient information were assigned to a mappable DZN based upon an imputation rule devised by the Transport Data Centre of the NSW Department of Transport. For example, a respondent might have provided his/her workplace address as 'Hurstville Rd, Hurstville' (which overlaps three DZNs) with no street number information. The NSW Department of Transport built an imputation rule into the indexes to automatically assign 'no number' responses to one of the three mappable zones. ABS relies upon the major data users (the STAs) to use their local knowledge to appropriately allocate these codes (Robertson, 2000).

When a response was allocated to a DZN based on these definitions, it was indistinguishable from a complete response; therefore it gave a false impression of the completeness of the respondent's answer. Secondly, since these imputations became part of the normal DZN data, it was not possible to know how many responses were treated in this way (Robertson, 2000).

Non-Responses to Workplace Address Question

The improvements in questionnaire design helped a lot to reduce the non-response rate of workplace address questions in Census 1996. In the 1991 Census, some respondents answered that they had a full or part-time job, but ignored the appropriate sequencing instructions. They would respond to a question that they had not actively looked for work in the previous four weeks and then jump several questions completely missing the workplace address question. The likelihood of this error occurring was compounded by the fact that these questions were at the bottom of a page on the census form and the fact that the sequencing instructions required the respondents to skip the subsequent page entirely. Therefore respondents would not read the workplace address question at all, and would not realize that the address question may apply to them. To reduce this confusion, the 'Looking for Work' and 'Hours Worked' questions were moved to the end of the employment related questions for the 1996 Census (Robertson, 2000).

A further change was made to the processing of census forms, which may also have helped to reduce the non-response rates. Specifically, prior to 1996, responses that were not interpretable to coders were coded as 'Not Stated'. In 1996, such cases were more frequently referred to the query resolution team. This increased the likelihood that these answers would be treated as a stated response and some level of DZN allocated (Robertson, 2000).

Coding Strategies for Census 2001

In Census 2001, the State Transport Authorities expressed a new coding strategy as an alternative to the strategies adopted in the previous censuses (referring the methodologies discussed above). Rather than using the name and number of the street of respondent workplace, it was suggested that the assignment of DZNs should increasingly use a facilities index, or index of business names. Thus, in 2001, if no street name or number (where required) was entered, coders would be prompted each time to enter a business name and to allocate a DZN based upon this list of business names. This strategy also reduced the likelihood of widespread miscodings of an institution to an incorrect DZN. For example, a specific hospital would be listed on the facilities index and would be coded to the correct DZN. The increased use of facility-based coding, as well as the use of Intelligent Character Recognition technology to scan completed census forms was also expected to reduce the workload of census processors. This is expected to have important implications in saving time and increasing the accuracy of journey-to-work data coding (Robertson, 2000).

Geocoding Approach: Census 2006

Geocoded National Address Database

In association with key partners, the Australian Electoral Commission (AEC), CentreLink and Australia Post, Public Sector Mapping Agencies (PSMA) Australia Limited developed the Geocoded National Address File (G-NAF). This data were released on March 2, 2004. G-NAF is the first authoritative database of addresses for all of Australia. The data are updated quarterly and is very quickly establishing itself as the definitive geocoding reference file for Australia (Lowe, 2005).

Mesh Blocks

Across its statistical collections, the ABS uses a common standard for geographical areas known as the Australian Standard Geographical Classification (ASGC). Local Government Area (LGA) is a key spatial unit of the ASGC. Within an LGA, the smallest building block of the ASGC is the Census Collection District (CD) or enumeration area, typically comprising 100 to 220 households. CDs are commonly used for detailed spatial analysis of Census data at the local level (Lowe, 2005).

However, CDs are very heterogeneous in nature with respect to shape and size and also, with areas of interest such as neighborhoods, electoral boundaries and various catchment area boundaries. The ABS proposed to overcome this shortcoming by developing a new micro-level geographical unit known as a Mesh Block. Mesh Blocks would contain a minimum of 20 to 50 households (about one-fifth the size of a CD) and would align with a wide range of administrative and natural boundaries. After the 2006 Census, very basic census data will be available at the Mesh Block level such as number of dwellings and population counts; however, a range of census data will be available for combinations of Mesh Blocks to meet individual user needs (Lowe, 2005).

Tabulation of O-D Matrices

The Journey-to-work field is cross-classified with a Usual Residence field. Both fields are hierarchical and contain a common level called Statistical Local Area (SLA), which enable the identification of origin and destination at the same level of geography (Viner, 2005).

Confidentiality

Similar to any other nation, Australia also takes many measures to protect direct or indirect disclosure of respondent's identity from the Census data. There are three major confidentiality issues that the Australian Bureau of Statistics (ABS) is mostly concerned with:

- Security arrangements
- Retention of name-identified information
- Confidentiality of tabular data

A discussion of all of these issues is beyond the scope of this paper. Therefore, only confidentiality issues related to the release of tabular data are discussed here (Australia Bureau of Statistics, 2002).

Confidentiality of Tabular Data

As discussed earlier, tables containing cells with very small counts may potentially provide enough information to identify respondents. To avoid releasing identifiable information, all tables are subjected to two confidentiality protection processes before release:

- Assessing the size of the table; and
- Introducing random error.

These steps are taken to avoid releasing information that may identify particular individuals, families, households, or dwellings without impairing the usefulness of the tables (Australia Bureau of Statistics, 2002).

Assessing the Size of the Table

This process involves comparing the total number of cells in a table to the total population for that table. The total number of cells may include categories such as 'Not stated', 'Not applicable', and 'Inadequately described'. If the number of cells is the same as, close to, or exceeds the population size, then the table will not be released. This practice avoids the release of tables containing a large proportion of small cells containing identifiable data (Australia Bureau of Statistics, 2002).

For example, consider a table in which three variables are cross-classified as follows:

- Sex: 2 categories

- Labor force status: 5 categories
- SLAs in Sydney: 46 categories

This table would result in 460 cells (i.e., 2 × 5 × 46). The population for this table is 2,942,389 persons. Therefore this table is suitable for release.

Consider another example where the cross-classification is as follows:

- Sex: 2 categories
- Labor force status: 455 categories
- SLAs in Sydney: 6542 categories

This table would result in 5,822,380 cells (i.e., 2 × 445 × 6,542). The population for this table is 2,942,389 persons. Therefore this table would not be suitable for release (Australia Bureau of Statistics, 2002).

Introducing Random Error

The technique of introducing random errors has been developed to help avoid identification of individuals. The confidentiality protection technique applied by the ABS is to randomly adjust cells with very small values. These adjustments do not impair the value of the table as a whole. The technique allows very large tables having a strong client demand to be produced even though they contain numerous cells with very small numbers. However, it is ABS policy not to release the detailed methodology employed by the ABS to adjust the data (Australia Bureau of Statistics, 2002).

Tables that have been randomly adjusted are internally consistent; however, comparisons with other tables containing similar data may show minor discrepancies. These small variances are normally ignored. Care is taken when specifying tables to minimize the number of small cells. Generally, no statistical reliance is placed on cells with small values. Aside from the effects of introduced random error, possible respondent and processing errors have the greatest relative impact on small cells (Australia Bureau of Statistics, 2002).

Miscellaneous

Australian Bureau of Statistics has decided to make some major changes in Census 2006. The use of electronic Census questionnaire (via Internet) and field communication will be one of the major implementation. The use of electronic Census questionnaire has been under investigation since March 2002. A facility to be used in the 2005 Dress Rehearsal is currently being developed in conjunction with IBM and a recent testing has shown that only around 5-10 percent of the population may choose to use this option (Lowe, 2005).

CANADA

Census Background

On May 15, 2001, Statistics Canada conducted the 19th Census of Population and the Census of Agriculture to develop a statistical portrait of Canada and its people. The census is a reliable source for describing the characteristics of Canada's people, dwellings, and agricultural operations. The Canadian Census is conducted every 5 years in May in years ending in digits 1 or 6. In Census 2001, between May 1 and May 12, 11.8 million households received a Census of Population questionnaire. An adult in each household was asked to fill in the questionnaire and mail it back to Statistics Canada. Similar to the United States, two kinds of questionnaires were designed: 1) short form and 2) long form. The short questionnaire contained seven questions and was completed by 80 percent of households. The long questionnaire contained the same questions as the short form plus 52 additional questions. The long form was completed by the remaining 20 percent of the population. For 2 percent of the population, who live in remote areas and on Indian reserves, a census representative completed the questionnaire during a household interview (Canada Census Website).

Questionnaire Content

The questionnaires obtained the following information for all respondents:

- Characteristics of the population: Information about age, sex, household composition, etc.
- Activity limitations: Questions asked about any sort of activity limitations at home, school, work or in any other aspects of their lives, such as travel or recreation. This information was intended to identify respondents for a post-censal Participation and Activity Limitation Survey (PALS).
- Language: The long questionnaire contained questions on the first language learned in childhood, languages understood and spoken at home, as well as knowledge of official and non-official languages in the various regions of Canada (Statistics Canada, 2004).

A few new questions were introduced in the 2001 Canadian Census. These include (Statistics Canada, 2004):

- Language at work
- Birthplace of parents
- Religion
- Languages spoken at home
- Common-law couples

Journey-to-Work Data

Place-of-Work Questions

All non-institutional residents aged 15 years or older who had worked at some time since January 1, 2000 were asked to respond to the Place-of-work question. The Place-of-work question has two objectives. The first is to identify the general workplace location of the respondent through the check-off categories and the second, to identify a precise workplace location using the specified address. The questions were designed to capture information about work status, i.e. part time or full time, and detailed address of the job location whether it is home or employer's address. If the employer's address was unknown, then the respondent was asked to indicate any landmark or street intersection nearest to the job location (Statistics Canada, 2004).

Several variables were created related to the place-of-work questions. The variables are listed below.

Place-of-work status

The Place-of-work Status variable is created directly from the responses to the Place-of-work question. The variable has four categories (Statistics Canada, 2004):

- worked at home
- worked outside Canada
- no fixed workplace address
- worked at the address

Workplace location

The second objective of the Place-of-work question was to identify the specific location of a workplace by assigning it a precise geographic location. This was accomplished through the workplace location coding system. Workplace location data were available for most standard geographic areas either directly from the database or by aggregating other standard areas. Outside census metropolitan areas (CMA) and census agglomerations (CA), the data were coded to the census subdivision level. Within census metropolitan areas and census agglomerations, the coding was much more detailed. The following geographic areas are available only within census metropolitan areas and census agglomerations: census tracts, urban areas, dissemination areas, and blocks. However, postal codes and designated places are not supported. By using the workplace location and place of residence information, origin-destination commuting flows are produced for a variety of levels of geography (Statistics Canada, 2004).

Commuting distance

Commuting distance was derived by calculating the straight-line distance in kilometers between a respondent's residential block and the workplace location representative point.

Mode of Transportation Questions

All persons responding to the Place-of-work question who marked "No fixed workplace address" or "Worked at the address specified" were asked to answer the Mode of Transportation question.

The Mode of Transportation question had eight categories: Car, truck or van – driver; Car, truck or van – passenger; Public transit (e.g., bus, streetcar, subway, light-rail transit, commuter train, ferry); Walk; Bicycle; Motorcycle; Taxicab; Other method. Persons who use more than one mode of transportation were asked to identify the single mode they used for most of the travel distance to capture the primary mode. The question did not measure multiple modes of transportation, seasonal variation in mode of transportation, or trips made for purposes other than commuting from home to work (Statistics Canada, 2004).

Workplace Coding

Workplace location coding in the Canadian Census 2001 shared many of the same challenges as other national censuses. In addition to abbreviated, incomplete, invalid and ambiguous responses, in some cases, the responses consisted of a building name or nearby street intersection instead of a street address. However, the coding system was designed to handle various types of responses through the use of a series of reference files. The reference files used in coding were derived from the June 2001 version of Statistics Canada's National Geographic Base (Statistics Canada, 2004).

All questionnaires providing a written response to the Industry or Place-of-work questions were processed through the workplace location coding system. For Census 2001, this represented approximately 3.5 million responses over an eight-month period. Workplace locations were coded to a point location represented by a latitude-longitude coordinate. When the workplace is located outside a census metropolitan area (CMA) or census agglomeration (CA), the point location represents a census subdivision (CSD). For CMAs and CAs, the point location usually represents a blockface or a block; however, it can also represent a dissemination area or a census tract (for areas that are census tracted). When the combination of response and reference materials was insufficient to code to these detailed levels, the record was either coded to a census subdivision, or was considered 'uncodeable'.

The workplace location coding system consisted of two distinct components: the automated component and the interactive component (Statistics Canada, 2004).

Automated System

An automated coding system matched responses to a series of reference files until a match was found. Responses that could not be coded automatically were sent to the interactive coding system.

The first step in automated coding system involved filtering out responses such as "Worked at home", "Worked outside Canada", "No fixed workplace address", and not applicable responses such as "didn't work", "full-time student", "retired", "on disability", etc.

Responses that were not filtered out were sent to the Place Name module. The province and city/town response were matched to a list of province/territory names and place names within the Place Name Reference File. When a match was found and the reference file record was located outside CMA- and CA-covered areas, then the response was coded at the census subdivision level. When a matching record was located within a CMA- or CA-covered area, a preliminary CSD (census subdivision) code was assigned and the response was sent through subsequent coding system modules for more detailed coding. The preliminary code was used to control the matching process.

Additional coding modules were applied in the following order until a match was found:

- The Postal Code module matched the response to the Postal Code Reference File;
- The Business/Building module matched responses to a list of large employment locations, business names, and addresses;
- The Street Address module matched the response to the Street Address File;
- The Street Intersection module matched the response to the Street Intersection Reference File. (Statistics Canada, 2004)

Interactive System

Responses that were not coded by the automated system were coded through the interactive coding system. The system was similar to the Computer-assisted Clerical Coding. Responses were separated into 172 regional databases based on the preliminary geography assigned in automated coding. The most valuable file is the electronic map of Canada, which allows coders to search for addresses and postal codes, zoom in on specific streets, see civic address ranges, and code responses to a specific block. Quality Control (QC) coding was applied to all cases, where a coder had coded to create groupings or clusters of similar responses. QC coders checked to ensure that all responses in the group represented the same workplace location and then recoded the group. The original and QC codes were compared and, if there was a difference, the correct code was determined by an adjudication coder. Responses that were coded by the automated system were not included in QC coding, but were subject to further analysis (Statistics Canada, 2004).

Edit and Imputation of Journey-To-Work Data

The Edit and Imputation process (E&I) was accomplished through a number of steps: 1) identifying Place-of-work Universe; 2) resolving inconsistent responses applying donor imputation technique; 3) deriving a number of variables. Throughout the E&I process, a series of tabulations were used to carefully review and monitor changes to the various variables and ensure that the processing worked correctly. The Journey-to-work E & I process was divided into four separate modules. The first module finalized the Place-of-work Status variable, the second and third modules finalized the Workplace Location variables, while the fourth module finalized the Mode of Transportation variable.

Uncoded workplaces were allocated using a “hot-deck” (donor) imputation procedure. Two different automated systems were used to carry out this processing. The first one was the

Nearest-neighbor **I**mputation **M**ethod (NIM), which was developed for the 1996 Census to perform Edit and Imputation for basic demographic characteristics such as age, sex, marital status, common-law status and relationship to Person 1. Then it was expanded for 2001 and implemented in a system called CANCEIS (**CAN**adian **C**ensus **E**dit and **I**mputation **S**ystem) to include Edit and Imputation for such variables as industry, place of work, mode of transportation, and mobility. CANCEIS allowed more extensive and exact edits to be applied to the response data, while preserving responses through minimum-change hot-deck imputation. Along with CANCEIS, SPIDER (**S**ystem for **P**rocessing **I**nstructions from **D**irectly **E**ntered **R**equirements) was used to process the remaining census variables such as mother tongue, dwelling, income, etc. SPIDER performed both deterministic and hot-deck imputation (Statistics Canada, 2004).

Imputation of Workplace Location

The second imputation module focused on assigning workplace locations at the CMA, CA and CSD levels of geography to persons with a usual place-of-work status. Records requiring imputation were either non-responses or responses that were incomplete or too ambiguous to be geocoded to the CSD. The variables used in the matching process were Industry Code, Occupation Code, Age and Sex. Stratification was not used in this imputation module.

The third module focused on assigning detailed workplace locations (i.e. census tract, dissemination areas, block and representative point) for persons with a usual Place-of-work status and who worked within a CMA or CA. The records imputed in this module fall into two groups: records that had no workplace location and were imputed for the CMA, CA and CSD levels in the previous imputation module, and records that could only be coded to the CSD level because of the limited information provided. This module used the same matching variables as the previous module. For imputation purposes, the records were stratified (i.e. separated into mutually exclusive groups) based on the CSD of work. Therefore, the donor record was restricted to donors working in the same CSD as that of the failed record.

After imputation, the workplace locations were finalized by assigning the residence locations as the workplace locations for all persons who worked at home. The commuting distance was derived by calculating the straight line distance between the latitude and longitude coordinates between residence block and workplace location (Statistics Canada, 2004).

Confidentiality and Disclosure Protection

Similar to other national censuses, confidentiality issues are of major concern to Statistics Canada. At the same time, they are also facing user complaints about data being suppressed. The data suppression mechanism is based on the population living or working in a given geographic area. If the population of the area is below a threshold value, then the data are suppressed. For standard areas such as CSDs or Census tracts, the threshold is 40 (weighted); for user-defined areas, it is 100; and for all areas, a threshold of 250 is applied if the tabulation includes income data. Tabulations are randomly rounded to the nearest 5, except for counts below 10 which are rounded to 0 or 10 (Murakami[b], 2004).

Users of O-D Flow Tabulations/Matrices

The users of journey-to-work O-D matrices are not very different between the U.S. and Canadian contexts. Most users are involved in transportation or urban planning for municipal and provincial governments. One major difference between the two contexts is that the use of Census journey-to-work data are not incorporated into any standard federal or provincial programs. Canada does not have an equivalent to the CTPP package that is prepared in the United States; however, Statistics Canada has agreements with provinces and municipal governments to provide a set of tabulations at the CSD (city) and Census tract levels of geographies. This includes residence based, workplace based, and home-work based tabulations. Users are also able to commission custom tabulations for their own user-defined areas such as traffic analysis zones (Murakami[b], 2004).

FRANCE

Census Background

France has been conducting census since 1801 but historically, no periodicity was found in census taking. Recently, French Census taking procedure has been through a major reform. The French National Institute of Statistics and Economic Studies has introduced a large program to redesign the traditional census based on the “rolling census concept” with a goal to alleviate and spread the burden by conducting census over a longer period and also to meet demand of fresher data. This procedure was implemented after the last traditional census taken in France in 1999.

Methodology of French Rolling Census

The key principle of French Rolling Census methodology is to collect information over a five years period cycle to produce every year significant data for the medium year. For example, in a given year ‘Y’, data are produced with the data collected in the years “Y-4”, “Y-3”, “Y-2”, “Y-1”, and “Y” and are significant on average for “Y-2”.

Considering budget constraint, a second principle is adopted such that the budget allocation in each year would be one seventh of a general census assuming a seven year period is sufficient between two traditional censuses. As a result, one seventh of the population was planned to be enumerated in each year.

Sampling Strategy

‘Commune’ is the key geographic unit in French Census. Due to heterogeneous nature of 37000 French communes, two different sampling strategies are adopted for small (inhabitant threshold under 10,000) and large communes (inhabitant threshold over 10,000). Smaller communes are visited once in every five year or in other words, sampled at the average rate of 1/5 (20 percent) and all larger communes are visited every year or sampled at the rate of 8 percent. As each type of the communes has 50 percent share of total population in France, therefore every year half of the population at 20 percent plus half of the population at 8 percent provides a sample of 14 percent or 1/7th of the population.

Within the domain of 22 smaller communes in France, five homogeneous (in terms of socio-demographic characteristics) groups are created to minimize year to year variation. Each of the groups is surveyed once in every five years in rotation.

The larger commune samples are based on a "building register", which is a list of buildings (residential, commercial or institutional) uniquely identified and located so as to create a set of digitized maps. The "building register" is appended with the socio-demographic statistics taken from 1999 census and in this way it serves as the sampling frame for the large communes. The sampling strategy considered for each large commune is a stratified 2-stage sample of dwellings. Firstly, each commune is divided into five rotating groups similar to the smaller commune sample mechanism, and then a random sample of 40 percent dwellings is drawn from a group in each year in rotation, which in turn represents 8 percent of the total number of dwellings in the commune (40 percent × 15).

Estimation and Dissemination

Two different procedures were adopted for annual estimation of population for larger and smaller communes. The procedure for a larger commune is apparently straightforward and is shown in the table below.

The table shows that detailed population estimates are made available for all groups of a large commune in the current year 'Y' and year 'Y-2' either from a census or sample survey or from synthetic estimation. A synthetic estimation is done using the relationship between the observed and administrative data for a given area at a given point of time.

Table 2. Population Estimation for a Large Commune

ROTATION GROUP					Dissemination Reference Date		Current Year
	Y-6	Y-5	Y-4	Y-3	Y-2	Y-1	Y
GROUP 1					C	S	S
GROUP 2	C				S	C	S
GROUP 3		C			S		C
GROUP 4			C		S		S
GROUP 5				C	S		S

C = census taken
S = synthetic estimate

Source: Dumais et al.

For a given small commune, the estimation procedure is similar to the large commune. The only difference here is an additional synthetic estimation based on the information obtained during the most recent census year 'Y' and backcast the intercensal period. An anticipated problem of this procedure is the mismatching of the forecasting and backcasting series of synthetic estimates. However, Dumais et al. has proposed a procedure to blend the series into a "composite" series anchored at both ends to the census counts. For more details, the readers are suggested to refer the above mentioned paper.

Table 3. Population Estimation for a Small Commune

ROTATION GROUP					Dissemination Reference Year		Current Year
	Y-6	Y-5	Y-4	Y-3	Y-2	Y-1	Y
GROUP 1					C	→S	→S
GROUP 2	C				→S ←S	C	→S
GROUP 3		C	→S ←S	→S ←S	→S ←S	→S ←S	C
GROUP 4			C		→S		→S
GROUP 5				C	→S		→S

C = census taken

Source: Dumais et al.

→S = synthetic estimate by forecasting

←S = synthetic estimate by backcasting

GERMANY

Census Background

The last traditional census was carried out in Germany in 1987. The uprise of political controversy in eighties surrounding the privacy issues of census data prompted the German authority to move from traditional census to a register-based census. The new method of German population census is the combination of administrative registers and surveys as data sources. The main data sources are as follows (Szenzenstein, 2004):

- Population registers
- Employees registers
- Housing census (postal survey)
- Sample survey

Population register (PR) is considered as the back-bone of the new German register-based census. A PR maintains records of every legal resident living in a municipality. Municipality is the smallest geographical unit used for the German Census.

The employee registers are maintained by Federal Employment Agency (FEA). Employee register is the main source of providing occupational and work place information. However, the self-employed persons are not covered in this register. Separate survey is conducted to cover this population segment.

Test Surveys

Test surveys were conducted on December 5th, 2001 to prepare a register-based census in Germany. There were two kinds of test surveys conducted to measure the quality and efficiency of the new system: 1) Register Check; 2) Double Entry Check.

Register Check

- Objective: estimate the number of under and over counts.
- Sampling method: about 38,000 addresses in 550 municipalities were sampled.
- Procedure: PR records and household interview data were compared.
- Result: Great differences in PR quality were found between large cities and small municipalities (Szenzenstein, 2004).

Double Entry Check

- Objective: estimate the number of persons with double entries in PR
- Sampling method: all persons born on 1st January, 15th May, 15th September or with an incomplete date of birth (21.2 percent of population) were sampled
- Procedure: PR and sampled data were cross checked with six different matching techniques
- Result: double count consist only 1/5th of total over counts and the rates of over counts caused by double entries do not vary municipality size classes (Szenzenstein, 2004).

Advantages and Disadvantages of the New Method

Advantages

- Substantial cost saving: new method - € 336 million; traditional census - € 1020 million.
- Smaller response burden: new method - 27 million respondents; traditional census - 82 million respondents.

Disadvantages

- The method does not guarantee full census information for small municipalities and reliable census results for lower unit below the municipality level (Szenzenstein, 2004).

Journey-To-Work Questions

Journey-to-work questions that were included in the census are (Federal Statistical Office, Germany, 2005):

- Name and address of place of work or school/university
- Means of transport to work or school

- Time needed for way out to work or school

Confidentiality Protection

The following points are noted with respect to confidentiality of data (Federal Statistical Office, Germany, 2005):

- Data collected in German Census are subject to strict confidentiality
- Data including individual name and address are strictly restricted for public release, though individual data exclusive of names and addresses may be transmitted to the municipalities
- Some individual data relating to non-agricultural local units may be published as part of tabulations, but only down to the level of the "parts of municipalities".
- Immediately upon the termination of the population census work at the local survey offices, the survey forms and the data on the inhabitants were transmitted to the survey offices for organizing the census and then the data were forwarded to the land statistical office. All other personal data of the respondents existing at the survey office were deleted.

German Microcensus (After 1987)

The microcensus is the official representative statistics of the population and the labor market, involving every year 1 percent of all households in Germany (continuous household sample survey). This has been done in Germany after 1987, the year in which the last full population census was conducted. The total number of households participating in the microcensus is about 370,000 (820,000 persons). The organizational and technical preparation of the microcensus is done at the Federal Statistical Office (German Microcensus Website).

All households have the same probability of selection for the microcensus (random sample). A one-stage stratified area sampling scheme is adopted. The sampled areas are called "sampling district". Every year, a quarter of all households (or sampling districts) included in the sample are exchanged (rotated off). This means that every household stays in the sample for four years (partial rotation procedure).

The purpose of the microcensus is to provide statistical information on the economic and social situation of the population as well as on employment, the labor market, and education. The annual standard program of the microcensus provides data on person and household characteristics including main or secondary place of residence, employment status, age, student status, educational qualification, sources of subsistence, insurance, and income. The annual supplementary program includes additional questions on employment and training. Data collected as part of the four year additional programs include information on commuting related to occupation or training, the housing situation, and health insurance in addition to health and disability status (German Microcensus Website).

NEW ZEALAND

Census Background

New Zealand (NZ) conducts a census once every five years. The 2001 Census was the latest census year and it was held on Tuesday, March 6, 2001. It was a snapshot of the population on that day/night. It covered all dwellings and every person alive in New Zealand or on a vessel in New Zealand waters. Some questions relate to a longer period of time, e.g., income (12 months) and employment (4 weeks) (Jonasen, 2005).

Tabulation of Workplace Address

The Census questionnaire asked for a person's physical workplace address. Each address was taken and matched to a census area – Mesh block, Area Unit, Territorial Authority, or Regional Council. Then a table was prepared by cross-tabulating the usual residence area and the workplace address area. These tabulations provide basic journey-to-work flows by geographic area pair. If a person does not fill in enough detail to record a workplace address, then it is coded as a NFD (not further defined) area. The Mesh block is the smallest level of geography for which the census information is released (Jonasen, 2005).

Journey-to-work information is generally used by the Territorial Authorities (Local Government) in NZ. Transport planners and consultants also tend to use the information. Statistics NZ operates on a cost recovery basis by charging the users for the time it takes to complete the information request. However, some basic population tabulations have been made available to the public (about 250 tables) free of cost (http://xtabs.stats.govt.nz/eng/tablefinder/index.asp) (Jonasen, 2005).

Confidentiality Issues

Rounding Procedures

Random rounding is a standard disclosure avoidance procedure applied on census data in order to reduce the amount of data loss which occurs with suppression. In random rounding, cell values are rounded making a random decision to whether they will be rounded up or down. The additive nature of the table is generally destroyed by this process.

Over the last several decades Statistics New Zealand has been applying random rounding technique to census statistics to the nearest multiple of three. This enables the greatest possible amount of census data to be published, without compromising the privacy of individual responses (Jonasen, 2005).

Under the random rounding process, all table cell values including row and column totals are rounded as follows (Jonasen, 2005):

- zero counts and counts that are already multiples of three are left unchanged;
- other counts are rounded to one of the nearest multiples of three.

All rounding, including separate rounding of totals and subtotals, is carried out on the recorded results. The probabilities of rounding up or down are set so that the long run expected value is equal to the original count. The effect of this rounding on the accuracy of census statistics for practically any proposed use is insignificant. Furthermore, on occasion, figures or percentages have been rounded off to the nearest unit or decimal point (Jonasen, 2005).

THE NETHERLANDS

Census Background

The last traditional census based on a complete enumeration of the whole population was held in 1971. Declining willingness of the population in census participation and sharply increasing cost of conducting traditional census caused Statistics Netherlands to find an alternative register-based method, which is called "Virtual Census". The methodology of this latest Dutch census program is based on compilation of population registers data with Labor Force and housing surveys. The census rounds in 1981 and 1991 were focused on the development of population registers and surveys. The latest Dutch Census in 2001 was the integration of microdata from all the registers and surveys (Nordholt, 2004).

The data sources for the Dutch census program can be distinguished into three types (Nordholt, 2004):

- Registers
    - Population register (PR): 16 million records (Demographic variables: gender, age, household status etc.)
    - Jobs files: employees (6.5 million records), self-employed persons (790 thousand records), dates of job, branch of economic activity
    - Fiscal administration (FIBASE): jobs (7.2 million records) and pensions and life insurance benefits (2.7 million records)
    - Social Security administrations: 2 million records,
- Surveys
    - Survey on Employment and Earnings (SEE): working hours, place of work (3 million records)
    - Labor Force Survey (LFS): education, occupation, economic activity (230, 000 records)

Combining Data Sources

The central population register is the backbone of the Netherlands Census. It is combination of all municipal population registers. The procedures for matching administrative registers and household survey data are briefly described here (Nordholt, 2004).

Record linkage

Statistics Netherlands has developed a linkage strategy subject to the requirements that the number of matched records should be maximized and the number of mismatched records should be minimized. To match the microdata sources, the following variables are available as linkage variables (keys) (Nordholt, 2004):

- Linkage key:
    - Registers: Social security and Fiscal number (SoFi)
    - Surveys: Sex, Date of Birth, address (postal code and house number)

SoFi-number represents unique personal identifier from the register and the combination of sex, date of birth, and address makes a unique identification in the survey data. The Population Register is the pivot in this linking process. Subject to the availability of keys in the data sources to be matched, the following linkage strategy was followed (Laan, 2000):

Microdata Integration

After the creation of the input files with all administrative and survey data on persons, families, households, jobs, benefits, and living quarters exactly matched, the integration process was started by data editing and imputation. During the process of microdata integration the following steps were taken (Nordholt, 2004):

- Collecting data from several sources
- Compare sources
    - coverage
    - conflicting information (reliability of sources)
- Integration rules
    - checks
    - adjustments
    - imputations
- Optimal use of information

Social Statistical Database

Social Statistical Database (SSD) is one of the vital developments of Dutch Census Program. It is the integration of a set of micro-data files with coherent and detailed demographic and socio-economic data on persons, household, jobs and benefits.

Census Journey-To-Work Questions

All members of a household were asked to report their trips during one specific day of the year. Every day, a new sample of households is asked to report over a particular day (for instance, Monday, June 1st, 2002). Every day of the year is included in the survey (January 1 until December 31) including Christmas, Easter, and other national holidays. Vacation trips are excluded from the survey (Franssen, 2005).

People are asked to fill in the address (street, house number and postal area code) of departure and the address of destination. The Dutch postal area code consists of four figures and two syllables (e.g., 6414HA). Only the first four digits are coded. There is no specific question in the survey about people's actual working place. But if people go to work on the day they report their trips, work trip information can be derived based on the purpose of the trip that is reported by the respondent. If trips are not reported, people are contacted and asked if it is possible that they have forgotten to report certain trips. Forgotten trips for those people that can not be contacted are imputed (Franssen, 2005).

Confidentiality

As the 2001 Census was compiled partially on the basis of survey data, some confidentiality techniques were applied on published tables. For example, tables obtained from Dutch Labor Force Survey (LFS) were applied the following rules (Nordholt, 2004):

- Table cells based on less than 10 persons were always suppressed.
- Table cells based on 25 or more persons were always published.
- Table cells based on 10-24 persons were only published if they form a part of a breakdown (e.g. age or sex), in which case no cells contain less than 10 entries. In addition, 50 percent of the cells in breakdown should have more than 25 persons to get published. The threshold of 25 persons corresponds to an estimated relative inaccuracy of at most 20 percent.
- The same rules were applied to the Survey of Housing Conditions (SHC), only higher threshold values were applied as the smaller sample size of SHC was smaller compared to LFS.

UNITED KINGDOM (UK)

Census Background

Since 1801, the UK Office of National Statistics (ONS) has been conducting decennial censuses in England and Wales. The General Register Office for Scotland and the Northern Ireland Statistics and Research Agency are responsible for conducting censuses elsewhere in the UK. The latest Census was taken on Sunday, April 29, 2001. The UK census costs about £ 255 million and it targets data collection in areas such as population, health, housing, employment, transport, and ethnic group (UK National Statistics Website).

A census form is delivered to every household, establishment, or to people living anywhere else, by a field force set up throughout the country. The forms are designed for self-completion by form-fillers to provide information for the census day (e.g., April 29, 2001). Most forms are then mailed back to temporary local offices and the remainder collected by the field force. Summary figures for all census topics from national to local authority level were released for the 2001 census on February 13, 2003. The National Report for England and Wales was released on June 30, 2003, followed by a release of detailed results in July and September 2003 (UK National Statistics Website).

Travel-to-Work Data

The UK Census 2001 was designed to provide comprehensive statistics on the origin and destination of migration and travel to work or place of study. journey-to-work questions were included in all the Census questionnaires and asked to every respondent aged between 16-74 years. This information was intended for users to analyze flows of people moving from one area to another and to analyze flows of people commuting between where they live and where they work.

The Census questions relevant to journey-to-work statistics are (UK National Statistics Website):

- home address one year ago
- commuting destination
- means of travel to work or study.

Workplace Data in the 2001 Census

The Census 2001 was planned to capture a wide-range of workplace data. There were some key differences between the Census workplace data collected in the year 2001 and 1991. In the 1991 Census of Population, data on workplaces was held in the Special Workplace Statistics (SWS). The SWS were based on 10 percent of the forms returned. But the publication of workplace data from the Census 2001 went through the following changes:

- In 2001, all workplace data are based on 100 percent of the forms returned.
- The 2001 SWS contain only 'flows' (i.e. origin–destination data). Three sets of output are being produced: local authorities (SWS1), wards (SWS2) and output areas (SWS3).
- In 2001, 'stock' counts on the nature of the workplace population (aged 16-74) are available in the standard area statistics (Standard Tables/Census Area Statistics) (Murakami[a], 2004).

Standard Tables and Theme Tables

Standard Tables (S) and Theme Tables (T) based on workplace population were released at a local authority level by ONS in February 2004.

A Standard Table (S) is a set of detailed cross-tabulations that provides the base of pre-planned outputs for England and Wales. The tables are available at a geography ranging from the national to the ward level. Standard tables are used to gain insight into the spatial division of labor at the micro-area level.

Theme tables are particularly valuable sources of information as those contain data on a range of topics for various commuting types. The variables included in the theme tables include age, family type, industry, National Statistics Socio-Economic Classification (NS-SEC), distance traveled to work, mode of transport to work, and hours of work (Murakami[a], 2004).

Census Area Statistics (CAS)

CAS consists of a number of tables based on daytime/workplace population. However, the amount of workplace information available from the CAS tables is more limited than that available from standard tables – but the CAS tables are available at a finer level of geographical disaggregation (output area or higher). CAS provides output for a wide variety of geographies from the national to Output Area (containing an average of around 125 households) levels (UK National Statistics, 2004).

The 2001 SWS

The 2001 Special Workplace Statistics (SWS) contain a set of tables based on employment and journey-to-work. SWS tables provide detailed matrices at the ward or higher geographic level. A matrix of journey-to-work data containing mode of travel information was released in May 2004 (Murakami[a], 2004).

Samples of Anonymized Records (SARs)

In 1991, for the first time in a UK Census, a product known as the Samples of Anonymized Records (SARs) was released. The SARs differ from traditional Census outputs in that rather than aggregated data, they are abstracts of individual census returns (i.e., the data relates to individuals rather than areas). In 2001, SAR covered a three percent sample of individuals, geographically disaggregated to 278 SAR areas, and a one percent sample of households. The release of SARs has no geographical disaggregation at a finer level than Government Office Region (Murakami[a], 2004).

Quality of Workplace Data Capture and Coding

The task of setting up the processing operation to extract data from the 2001 census forms and code text responses were outsourced to the contractor named Lockheed Martin (LM). There were just over 22 million records with responses on workplace address. LM coded 13.5 million and the remaining 8.5 million (38.6 percent) had valid postcodes that were captured from the form. The contract with LM allowed them to provide partial postcodes for this question, and the samples of data checked included full and partial postcodes along with records marked as uncodeable. The postcodes held in the address fields on the Census database were used to create origin-destination zones. For instance, at the lower level, postcodes were grouped into the same areas as those to be used for Census Area Statistics – Output Areas (OAs) (UK National Statistics Website).

Imputation of Migrant Origin, Workplace and Study Address

This section describes the research on imputation methodology of migrant origin, workplace and study address conducted by the General Register Office for Scotland on behalf of the Office of National Statistics. The methodology that is proposed in this study is a form of donor imputation (Mortimer, 2000).

The components of this work package were as follows:

a) Identification of matching variables - To identify the optimum combination of variables on which a potential donor must match an intended recipient. Satisfying this criterion maximizes the accuracy of the imputation while preserving the joint and marginal distributions of the data.
b) Assessment of the accuracy of imputation - To establish a suitable measure of accuracy for imputed data and report these results for a number of simulated imputations. Figures of accuracy have been produced for wholly missing and partially complete addresses.
c) Proof of concept - Research has been conducted to determine whether the imputation carries the potential to reduce any bias in the Census data arising from item non-response to the workplace and migration data.
d) Specification of imputation process for these variables (Mortimer, 2000)

Types of missing data

There are three basic types of output that can be generated by the data capture and coding system (DCCS) for these variables:

- A full postcode entered by the respondents or derived by the DCCS
- A partial postcode produced due to imprecise description of the location or actually a partial postcode given by the respondent
- A wholly missing postcode (Mortimer, 2000)

Scale of the Problem in Census 1991

Workplace: In the 1991 Census, 3.41 percent of postcodes were missing or could not be assigned in Scotland

Migrant origin: In the 1991 Census, 2.29 percent of postcodes were missing or could not be assigned in Scotland.

Methodology

The two primary goals of this research were to identify a stable set of matching variables for the donor imputation of postcodes and to assess the accuracy of the imputation itself. Clean and consistent data were used throughout to look at the original and imputed postcodes for each record - essentially a 'before and after' comparison. The following comparisons were used (Mortimer, 2000):

a) determine the number of 'perfect' imputations, i.e., where the imputed postcode is an exact match to the original postcode.
b) using the coordinates of each postcode, the distance between the original and imputed postcodes can be calculated, essentially another measure of the accuracy of each imputed postcode.

c) using the postcode coordinates, the net change in the distance traveled that is caused by the imputation can be calculated.

To determine the optimum combination of variables that a potential donor must match upon, complete and consistent data were used to examine the effect that each variable has in minimizing the above three measures. Using the matching variables identified in the previous stage, an extract of complete and consistent data were used again to examine the accuracy of imputed data in practice, using a prototype imputation system.

For the sake of brevity, only a broad overview of the general methodology of imputation applied to migrant origin and workplace and study addresses is provided above. However, the full report describes the imputation procedures in detail separately for each of the cases (Mortimer, 2000).

Analysis

Two basic analyses performed on the 'imputed' data are reported here (Mortimer, 2000):

1. Analysis of the accuracy of the imputed postcodes. The percentage of imputations where the imputed postcode is an exact match for the original postcode is used as a base measure of accuracy. This measure is an examination of the distance between the original and imputed postcodes where the imputed postcode does not equal the original. This second measure shows the percentage of imputed postcodes falling within successively larger radii around the original postcode.

2. Comparison of original and imputed commuting/migration distance. One of the outputs from the SWS data is the distance traveled when migrating or commuting. It was considered important that the imputation of postcodes should not distort these statistics to any great, or rather significant, degree.

Disclosure Control Protection

The following measures were applied to all 2001 Census output for England and Wales to prevent the inadvertent disclosure of information about identifiable individuals (UK National Statistics, 2005).

Small-Cell Adjustment

- A small count appearing in a table cell was adjusted.
- Totals and subtotals in tables were calculated as the sum of the adjusted data so that all tables were internally additive; within tables, totals and subtotals were the sum of the adjusted constituent counts.
- Tables were independently adjusted; this means that counts of the same population in two different tables may not necessarily be the same.
- Tables for higher geographical levels were independently adjusted, and, therefore, would not necessarily be the sum of the lower geographical component units.

- Output was produced from one database and adjusted for estimated undercount; the tables from this particular database provide consistent pictures of the population (UK National Statistics, 2005).

Record Swapping

The individual records on the output database were slightly modified by record swapping in which a sample of records was 'swapped' with similar records in other geographical areas. The proportion of records swapped is confidential (UK National Statistics, 2005).

Thresholds

Two pairs of thresholds apply for reporting data:

- For the release of Standard Tables an area must contain at least 1,000 residents and 400 resident households.
- For the release of Census Area Statistics (CAS), an area must contain at least 100 residents and 40 resident households (UK National Statistics, 2005).

All thresholds applied to populations on the census day of April 2001. Where civil parishes (England) or communities (Wales) or wards fell below the CAS threshold but contained more than 50 people and 20 households, profiles with summary statistics were released. The 1991 census equivalent threshold was 50 people and 16 households. The increase in the household component to 20 reflected the change in average household size between 1991-2001. Where those areas had less than 50 people and less than 20 households, counts of the total numbers of residents, males, females, and resident households were released after performing small cell adjustment procedures. However, where parishes, communities, or wards fell below thresholds, they were unified with contiguous areas, in consultation with the local authorities concerned, into areas with sufficient population for the release of Standard Tables or CAS (UK National Statistics, 2005).

Design of Table

A general principle of making the average cell count in a table greater than or equal to one was applied to the design of all 2001 Census Output (UK National Statistics, 2005).

Disclosure Control and Census 2001 Origin-Destination Flow Data

Census counts of the flows of people migrating, or traveling to work, from area to area were published in the Origin-Destination Statistics for Output Areas in May 2004, for wards in July 2004 and for local authorities in October 2004. As with all Census outputs, small counts in the tables were adjusted before release in order to protect confidentiality. A greater effect on the variability of the published Origin-Destination data is reported due to these adjustments because of the large number of small counts in the Origin-Destination Statistics compared to the standard Census tables. The following section provides the guidance on the levels of variability, which are expected in the O-D statistics and discusses how this variability was minimized (UK National Statistics, 2004).

Count Adjustments

Small counts were adjusted or 'perturbed' to protect confidentiality. According to the perturbation scheme, the cells with small values were adjusted independently upwards or downwards based on prescribed probabilities. The scheme was designed so that perturbations should have an expected mean of zero (that is, the adjustment does not introduce systematic biases into the counts) and a variance (a measure of how much the perturbed counts vary from the true value) that is proportional to the number of small cells that were adjusted. The more cells adjusted, the larger the possible variations from the true values (UK National Statistics, 2004).

The probability mechanism that was devised for carrying out the perturbations is described in a paper written by Shlomo (2001). The paper explains that the method of small-cell adjustment is internally consistent within a table, since totals are obtained by summing the component cells of the table. The small-cell adjustment was repeated randomly. However, the method may introduce some variability in output and it tends to be largest for small areas that are likely to contain small counts. There are other sources of variation in the data, which arise from coverage error, respondent error, other forms of processing error, and record swapping. (Shlomo, 2001).

Possible Bias in the Adjusted Data

Several matrices relating to migration flows were checked for any evidence of bias. Statistical tests suggest that in some cases, there is a very small amount of bias because of the nature of the random processes used in the method. The reasons relate to the adjustments in the Origin-Destination Matrices but do not apply to the main sets of output such as Standard Tables and Census Area Statistics (CAS) (UK National Statistics, 2004).

Variability in the Adjusted Area

The adjustments made to the data to protect confidentiality mean that many counts in the Origin-Destination Matrices would differ from the underlying 'true' counts. Therefore, the counts produced by aggregating several entries in an Origin-Destination Matrix are expected to be different from corresponding counts in the Census Area Statistics tables (UK National Statistics, 2004).

Advice for Users

Some possible solutions were suggested to tackle the problems noted above. The Origin-Destination Statistics are reported to be more affected than other tables by adjustments to the data to protect confidentiality; however, they are more reliable than similar results published in 1991 which were based only on 10 percent of the total population. The variability in the aggregated totals can be minimized by using the highest level geography possible - for example, deriving results for a Government Office Region by aggregating counts for local authorities rather than Output Areas.

In addition, the most accurate count of the overall flow between two areas is most likely to be contained in the table with the fewest cells. For sub-groups of the population, it is suggested to choose the table that has the fewest possible cells that need to be aggregated to obtain the overall flow (UK National Statistics, 2004).

Miscellaneous

There are also some other issues that the transportation community is facing in UK due to stringent data suppression by Office for National Statistics. Some of them are indicated below:

- Initially, ONS decided to apply a common rounding procedure to all output tabulations irrespective of the cell sizes. They simply rounded all outputs to multiples of 3. But after much negotiation the Census community persuaded ONS upon applying their rounding rule only for small cells. However, the problem remained for the Output Area matrices, where most of the cells are very small and are affected by disclosure protection. It becomes a real challenge to aggregate the small area data correctly to larger zonal level for transport modeling and it seems that the data are almost useless for transportation community.
- The second issue is concerned about the suppression of data on industry at the ward level or below. This problem becomes critical for the transportation planners, who use this data for modeling trip attraction.

Data Access and Condition of Use

The Census 2001 data are free, but "a condition of use included in all end-user licenses is that the Census material shall not be used to attempt to derive information relating to an identified person or household nor shall a claim be made that such information has been obtained or derived". The Census statistics are available at the website http://www.neighbourhood.statistics.gov.uk .

CONCLUSIONS

This paper provides an overview of international experience with national censuses with particular emphasis on journey-to-work data in light of the changes occurring in the U.S. context, where journey-to-work data will hereafter be collected continuously (annually) through the American Community Survey (ACS). Additional information is expected from several additional country contacts to whom requests for information have been sent. As soon as the information becomes available, it will be incorporated into the paper.

From the review of international experience conducted so far, it appears that there are some common issues of interest and concern that have implications for the application of ACS journey-to-work flow data for transportation planning. In addition, there are a variety of techniques that are being used to address these concerns and issues. The following is a brief summary of these concepts:

Population Register

Declining willingness of the population in census participation and sharply increasing cost of conducting traditional census have caused many countries to find an alternative procedure, which is primarily based on population register and housing surveys. The Netherlands and Germany are the best example in this regard. After gathering all information from different sources, record linkage and microdata integration are accomplished to develop a complete population database. However, U.S. does not have any population register similar to the countries mentioned.

Mid-decade Census

Mid-decade Census is considered as a viable alternative of the traditional decennial census for many countries, for example, Australia, Canada and New Zealand. In the United States, decision on the implementation of American Community Survey as a substitute of U.S. decennial Census has come through after a prolonged planning process continued over last two decades. However, in 1980s a proposal was placed in support of mid-decade census as an alternative, though the proposal was not extensively considered in the ACS development after the failure of a mid-decade census to be funded for 1985 or 1995. In this context, Charles H. Alexander of U.S. Census Bureau documented (Alexander, USCB) that "For the purpose of updating census profiles for small areas, a mid-decade sample census is arguably as effective as the five-year averages proposed for the ACS. However, a quinquennial (a period of five year) snapshot is not effective for monitoring year-to-year changes, the second major use of the ACS. This new use seems to have made the difference in obtaining support for the ACS."

Sampling Rates and Sample Sizes

Sampling rate and sample size issues are critical to the U.S. Census because of its typical questionnaire (short form + long form) structure. Traditionally, the sampling rate for the short form has been 1:1 (covering every individual) while for the long form, it has been 1:6 (17 percent). From a review of international experiences, it is found that different countries follow different sampling schemes, but the schemes do not necessarily differ very much. For example, the Canadian census is rather similar to U.S. format; their short form covers 80 percent of the households and the long form, which contains all the short form questions plus an additional 52 questions, is administered to the remaining 20 percent of the households. With the introduction of the ACS strategy by the U.S. Census Bureau, sample sizes have dropped by about 50-60 percent when compared with the decennial census long form. Smaller sample sizes for small geographic areas are subject to greater data suppression to comply with confidentiality issues. As a result, this could have severe implications for any flow tabulation and it is possible that about 40-50 percent of tract-to-tract flows could be suppressed.

Protecting Confidentiality

Geographical Scale

The geographical scale at which census data are reported and tabulated remains a key issue for transportation planning professionals. The international experience varies in this regard, but once again some common threads are found. Starting with the smallest geographic unit, several levels of aggregation are produced to meet the requirements of various planning agencies in a country. For example, 'blocks' in the U.S., 'output areas' in U.K., and 'mesh block' in Australia/NZ are all small geographic units for which selected census data are reported. However, a common issue is that census data reported at the smallest geographic level have to be often suppressed to protect confidentiality. This paper reports on several statistical and rounding procedures that are being applied around the world to tackle this problem. The situation is even more critical in the U.S. According to CTPP 2000, a TAZ is defined as a geographic unit with 400 residents and 200 workers; a table with 40 cells tabulating worker flows would have an average cell size of 5 (http://www.trbcensus.com/notes/content.html). With the implementation of the ACS, new thresholds are needed to define TAZ-based tabulations; geographic areas with small population often fall below the existing threshold values and TAZ tabulations suffer adverse consequences due to data suppression.

Accuracy of Journey-to-Work Flows/Tabulations

One of the key features of the CTPP is the home-to-work flow matrix. It is identified that several issues regarding tabulations of O-D flows as a result of the implementation of ACS. When the tabulation is done for very small geographical units, about the size of a "block group", small O-D flows are subject to a threshold of 3 unweighted records; this is a source of substantial problems for travel demand modelers, as nearly one-half of the worker flows are lost. From a review of international experiences, it is found that other nations have been facing similar kinds of problems with small area O-D data suppressed by disclosure control policies. Notably, the U.K. has adopted a unique mechanism of small-cell adjustment to publish O-D tables that retain as much information as possible while protecting individual confidentiality.

Geocoding Workplace Data

The accuracy of place-of-work geocoding has been a major concern in any census. Job locations are highly prone to geocoding errors and geographical allocation inaccuracies depending on the quality of data collected and completeness of response provided by the individual.

User-defined Geographic Tabulation Areas

To enhance the accuracy of geocoding residential and place of work locations, Australia has used different zonal structure for residential and CBD areas. Areas having predominantly residential population are divided into small geographic units called 'Collection District' and area like CBD is divided into units called 'Destinations Zones' having mostly working population. Interestingly, this mechanism has benefited Australia to capture most of the OD flows of journey-to-work tabulations even under confidentiality protection. This could be a good lesson for United States as we know that OD flows of small TAZs are expected to be lost substantially due to implementation of ACS with strict disclosure protection for small geographic units. Because, it is highly likely that TAZs belong to the residential areas have very small working population and on the other hand, TAZs belong to CBD areas have low residential units. This would cause OD flows between these TAZs fall below cell threshold value and as a consequence data are suppressed. However, some alternatives have already been proposed in the U.S. context. Purpose specific user-defined geographic areas like redefining TAZs with respect to residence, workplace or flow tabulation could be a solution to this problem. Also, multi-site businesses, government employment, and other challenges associated with defining workplace geography are still being researched. There are probably issues on the residence geography as well, e.g. areas with ‘transient’ populations such as students, snowbirds, and migrant laborers.

Geocoding Technique

It is found that different countries have different ways of coding workplace locations and work destination zones and adopt numerous techniques for handling insufficient responses. In the U.S. and Canadian context, the workplace location coding system consists of two distinct components: 1) automated component; 2) interactive clerical coding component. U.K. and Canada have reported on imputation methodologies that are based on donor imputation procedures. Both of these countries have developed their own in-house imputation systems that are unique and customized to their particular contexts. A detailed description of the Australian imputation procedure is also discussed in this paper. Notably, Australia is increasingly using facility indices instead of street name or addresses to geocode workplaces. This mechanism helps them to accurately code a workplace address to its corresponding destination zone where as street addresses often create problem of overlapping multiple zones resulting widespread miscoding of institutions to incorrect zone. From a review of international experiences, it appears that there are well established methods for geocoding work locations that handle two potential issues: 1) impute missing and incomplete responses, and 2) protect confidentiality of workers.

The accuracy of place-of-work geocoding has been major concern in any national census. The problem is that some sort of Master Address File or Population Registers in any nation continuously updates and corrects residential addresses, but not the business addresses. Therefore, job locations are often impacted by geocoding errors and allocation inaccuracies. Fortunately, TIGER developed by US Census Bureau provides a complete source residential and business addresses that has substantially alleviated the workplace geocoding problem. Similarly, Australia has developed a comprehensive database called Geocoded National Address File (G-NAF) containing all addresses for the whole nation. G-NAF is expected to be a most important source of geocoding reference in 2006 Australian Census.

Disclosure Avoidance Techniques

Many disclosure avoidance techniques are being used on census data to protect individual confidentiality.

Rounding

Rounding has been a very commonly applied disclosure avoidance techniques around the world. This technique is based on protecting small counts in tabular data against disclosure. The basic idea behind this disclosure control method is to round each count up and down either deterministically or probabilistically to the nearest integer multiple of an integer rounding base. Random rounding to nearest multiple of 3 (UK, Australia, New Zealand) is the most common practice, but little deviations are found for some nations including U.S. For example, Canada is traditionally using random rounding to the nearest multiple of 5, except for counts below 10 which are rounded to 0 or 10. In the U.S., "Rules of Four-Seven" are applied to CTPP 2000 tables. Under this rule, values between 1 and 7 are rounded to 4 and values of 8 or more are rounded to the nearest multiple of 5.

Small-Cell Adjustment

Small-Cell Adjustment is another disclosure control protection technique evident from U.K. Under this procedure a small count appearing in a tabulated data is adjusted. Totals and subtotals are calculated as the sum of the adjusted data so that all tables are internally additive. Each table is adjusted independently and therefore, tables for higher geographic levels are not necessarily be the sum of the lower geographical component units.

Data Swapping

From the international experience, it is found that U.K. and U.S. have been using this technique for the release of census microdata. The individual records on the output database are slightly modified by record swapping in which a sample of records was swapped with similar records in other geographical areas. This transformation technique guarantees the maintenance of table statistics. However, the proportion of data swapped is never disclosed.

Thresholding

Threshold is a specified number compared to the cell values of a tabular data. With the threshold rule, a cell of table of frequencies is defined to be sensitive or decided not to be released if the cell value is less than the threshold value. On CTPP OD tabulation, a threshold of 3 is applied for any unweighted OD pair. Often threshold rule is also applied for releasing data of geographic units with small population or number of households. For example, for the release of Census Area Statistics (CAS) in UK Census, an area must contain a threshold of population of 100 residents or 40 households.

Random Data Perturbation

Random Perturbation is a Perturbation-based method used to falsify the data before publication by introducing an element of noise purposefully for confidentiality reasons. U.K. and Australia use this technique to randomly adjust table cells with very small values. These adjustments do not impair the value of the table as a whole. The advantage of this technique is that it allows releasing tables containing numerous cells with very small counts.

Impact of ICT

Over the past few decades, advances in information and telecommunication technology (ICT) and accessibility to broadband Internet at home have prompted a remarkable growth in home-based business, teleworking, e-shopping, e-recreation, etc. Americans working from home three days a week or more rose 23 percent over the decade ending in 2000. Traditionally, the U.S. Census has focused on Americans who work three days a week or more at home, thus overlooking the many who might be telecommuting less often. The number of telecommuters working at home during business hours at least once a month rose to 24 million in 2004 (http://www.motiontemps.com/white_papers.html). Considering the dynamics of telecommuting arrangements (both formal and informal), it may be necessary to make changes to census questionnaires to capture the work engagement patterns of full time and part time home workers in addition to occasional informal telecommuters. Realizing the importance of these issues and their potential implications for telecommunications-travel behavior interactions, a few countries have initiated the restructuring of their census questionnaires by including ICT and telework related questions (e.g. Australia).

REFERENCES

Alecxander, C. H. The American Community Survey Issues and Initial Test Results. United States Census Bureau. Available from <http://www.census.gov/acs/www/Downloads/Bibliography/SYMP97R1.doc>

American Community Survey Brochure. (http://www.census.gov/acs/www/Special/brochure.htm)

American Community Survey Website. Available from <http://www.census.gov/acs/www/>

Australian Bureau of Statistics, 2002. Confidentiality of Census Output: 2001 Census of Population and Housing. Australian Bureau of Statistics Website. August 21. Available from <http://www.abs.gov.au/Websitedbs/D3110124.NSF/0/062180402B6368D4CA256AB800816C0C?Open>

Australian Bureau of Statistics, 2005. 2001 Census of Population and Housing - 2001 Census Working Paper - Fact Sheet: Journey-to-Work. Australia Bureau of Statistics Website. Available from <http://www.abs.gov.au/websitedbs/D3110124.NSF/0/fc5b4d8473bf32ddca256c78007a52cf?OpenDocument>

BTS and USDOT, 1996. Implications of Continuous Measurement for the Uses of Census Data in Transportation Planning. USDOT (April), Washington D.C.

Butani, S., Alexander, C., and Esposito, J., 1999. Using the American Community Survey to Enhance the Current Population Survey: Opportunities and Issues. Paper presented at the Federal

Committee of Statistical Methodology Conference. Available from <http://www.fcsm.gov/99papers/>

Canada Census Website. Available from <http://www12.statcan.ca/english/census01/home/index.cfm>

Christopher, E. and Srinivasan, N., 2005. Disclosure and Utility of Census Journey-to-Work Flow Data from the American Community Survey: Is There a Right Balance? Paper prepared for presentation in the conference on Census Data for Transportation Planning: Preparing for the Future, organized by Transportation Research Board, Irvine, California, May 11-13, 2005.

CTPP Historical Perspective Web Link. Available from <http://www.trbcensus.com/articles/ctpphistory.pdf>

CTPP Status Report, 1998. Available from <http://www.trbcensus.com/ctpp17.html>

Dumais, J., Eghbal, S., Isnard, M., Jacod, M., and Vinot, F., 1999. An Alternative to Traditional Census Taking: Plans for France. Paper presented at the Federal Committee of Statistical Methodology Conference. Available from <http://www.fcsm.gov/99papers/>

Durr, J. M., 2004. The New French Rolling Census. Presented at the UNECE Seminar on New Methods for Population censuses organized in cooperation with UNFPA, Geneva, November 22, 2004.

Franssen, F., 2005. Email Communication with Centraal Bureau voor de Statistiek, The Netherlands. University of South Florida. April 6th.

Federal Statistical Office, 2005. German Census 1987 Information Brochure. Mailing correspondence with Federal Statistical Office, Germany. University of South Florida.

German Microcensus Website. Available from <http://www.destatis.de/micro/e/micro_c1.htm>

Jonasen, E., 2005. Email Correspondence with Statistics New Zealand. University of South Florida, April 8th.

Kish, L., 1981. Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau. U.S. Government Printing Office, Washington, D.C.

Kish, L., 1990. Rolling Samples and Censuses. Survey Methodology (16), pp. 63-79.

Laan, P. V. N., 2000. The 2001 Census in the Netherlands Integration of Registers and Surveys. Paper presented in the conference on “The Census of Population: 2000 and Beyond”, organized by Cathie Marsh Center for Census and Survey Research, Faculty of Economics and Social Studies, University of Manchester, Manchester, UK June 22-23.

Limoges, E., 2004. Allocation of Missing Place-of-work Data in Decennial Censuses and CTPP 2000. Published in CTPP 2000 Status Report, USDOT, January, Washington D.C.

Lowe, P., 2005. 2006 Census of Population and Housing – Australia: Innovation to Meet the Challenges. Paper Presented at 22nd Population Census Conference, Seattle, March 7-9.

Mackun, P., 2001. The U.S. Census Bureau's Plans for the Census 2000 Public Use Microdata Sample Files. Population Division, U.S. Census Bureau. December 2001.
Available from <http://www.census.gov/population/www/cen2000/pums.html>

Mortimer, A., 2000. Imputation of Migrant Origin, Workplace and Study Address. General Register Office for Scotland, March.

Murakami[a], E. and Noble, B., 2004. Email Communication. October 8.

Murakami[b], E. and Singbeil, B., 2004. Email Communication. October 8.

Nordholt, E. S., 2004. The Dutch Virtual Census 2001: A New Approach by Combining Different Sources. Presented at the UNECE Seminar on New Methods for Population censuses organized in cooperation with UNFPA, Geneva, November 22, 2004.

Robertson, E., 2000. 1996 Census Data Quality: Journey-to-Work. Census Working Paper. Australian Bureau of Statistics. Available from <http://www.abs.gov.au/websitedbs/D3110122.NSF/0/6F1905F5D033E5D2CA25691800020293?Open>

Rust, K. F., 1994. Counting People in the Information Age. Published in the Report "Counting People in the Information Age". The National Academics Press.

Sanchez, F., 2005. Email Communication. University of South Florida.

Shlomo, N. Analysis of Statistical Disclosure Control in Census 2001 Origin-Destination Tables. SDC Center of the Methodology Group, Office of National Statistics. Available from <http://www.statistics.gov.uk/census2001/pdfs/od_d_paper.pdf>

Singapore Census 2000 Website. Available from <http://www.singstat.gov.sg/stats/methods/census2000.html>

Singapore Department of Statistics, 2002. Singapore's New Approach to Census Taking. Conference on Chinese population and Socioeconomic Studies: Utilizing the 2000/2001 Round Census Data. Hong Kong University of Science and Technology, June 19-21, Hong Kong SAR.

Statistics Canada, 2004. Journey-to-Work: 2001 Census Technical Report. Statistics Canada Website. Available from <http://www12.statcan.ca/english/census01/Products/Reference/tech_rep/journey/index.cfm>

Szenzenstein, J., 2004. The New Method of the Next German population Census. Presented at the UNECE Seminar on New Methods for Population censuses organized in cooperation with UNFPA, Geneva, November 22, 2004.

Trewin, D., 2000. 2001 Census of Population and Housing: How Australia Takes a Census. Australian Bureau of Statistics, October 3rd.

UK National Statistics Website. Available from <http://www.statistics.gov.uk/census2001/default.asp>

UK National Statistics, 2005. Disclosure Protection Measures. UK National Statistics Website. Available from <http://www.statistics.gov.uk/census2001/discloseprotect.asp)

U.S. Census Bureau Website. Available from <http://www.census.gov/>

UK National Statistics, 2004. Variability in the Census Origin-Destination Counts. UK National Statistics Website. Available from <http://www.statistics.gov.uk/census2001/od_d_paper.asp>

Viner, E., 2005. Email correspondence with Census Data and Technical Services. Australian Bureau of Statistics. University of South Florida, April 4th.