Methodology Branch

Direction de la méthodologie

Household Survey
Methods Division

Division des méthodes
d'enquêtes auprès des ménages

Statistics Canada Statistique Canada

Canada

WORKING PAPER
METHODOLOGY BRANCH

A STUDY OF METHODS FOR OBTAINING US ADDRESS DATA, ASSIGNING A STATE OF RESIDENCE AND ESTIMATING STATE PROPORTIONS FOR US AIR TRAVELLERS IN THE INTERNATIONAL TRAVEL SURVEY

HSMD–2012-007-E

Jack Singleton

Household Survey Methods Division
Statistics Canada

July 2012

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada.

A STUDY OF METHODS FOR OBTAINING US ADDRESS DATA, ASSIGNING A STATE OF RESIDENCE AND ESTIMATING STATE PROPORTIONS FOR US AIR TRAVELLERS IN THE INTERNATIONAL TRAVEL SURVEY

Jack Singleton*

ABSTRACT

Weighting adjustment in classes defined by state of residence for US air travellers is being introduced as an International Travel Survey (ITS) improvement initiative. This report describes development of a methodology for producing counts of travellers by state of residence using ITS E311 data.

After background research, it was decided to assign state of residence based on E311 state and zip code data, linked by a concordance file produced by the United States Census Bureau for the 2000 US Census. Statistics Canada operations personnel agreed that keying US address data would be possible within the current production system, although the set-up costs would be high. Arrangements were made to key state, zip code and other fields, outside of ITS E311 processing, for samples of US E311 cards. Three procedures for estimating state proportions based on extracted data were evaluated in terms of coverage and precision, by comparison to benchmarks calculated based on keyed data. The study recommended a procedure that combines extracted and keyed data, and also identified risks associated with procedures that use extracted data alone, especially those that exclude significant portions of US E311 cards.

Key words: Weighting adjustment; frontier counts; E311; state-zip code concordance; extraction; keying; Intelligent Character Recognition.

* Jack Singleton, Household Survey Methods Division, Statistics Canada, 16th floor R.H. Coats Bldg., Ottawa, Ontario, K1A 0T6.

ÉTUDE DE MÉTHODES POUR OBTENIR DES DONNÉES SUR LES ADRESSES AMÉRICAINES, ASSIGNER UN ÉTAT DE RÉSIDENCE ET ESTIMER LES PROPORTIONS PAR ÉTAT POUR LES VOYAGEURS AÉRIENS DES ÉTATS-UNIS DANS L'ENQUÊTE SUR LES VOYAGES INTERNATIONAUX

Jack Singleton[†]

RÉSUMÉ

L'ajustement de la pondération dans les classes définies en fonction de l'état de résidence pour les voyageurs aériens des États-Unis est en train d'être introduit comme initiative d'amélioration de l'Enquête sur les voyages internationaux (EVI). Le présent rapport décrit le processus d'élaboration d'une méthodologie pour dénombrer les voyageurs par état de résidence au moyen des données de l'EVI provenant des cartes E311.

À la suite d'une recherche documentaire, la décision a été prise d'assigner l'état de résidence en se fondant sur les données sur l'état et le code postal provenant des cartes E311, couplées à un fichier de concordance produit par le United States Census Bureau pour le recensement américain de 2000. Selon le personnel des opérations de Statistique Canada, le système actuel de production permettrait de saisir les données sur les adresses américaines, mais les coûts de mise en place seraient élevés. Des dispositions ont été prises pour saisir l'état, le code postal et d'autres champs pour des échantillons de cartes E311 des voyageurs des États-Unis, en marge du traitement des cartes E311 de l'EVI. On a évalué, sur le plan de la couverture et de la précision, trois procédures pour estimer les proportions par état en se fondant sur les données extraites en comparaison aux valeurs repères calculées en se fondant sur les données saisies. Cette étude recommande une procédure fondée sur une combinaison de données extraites et de données saisies. Elle cerne par ailleurs les risques associés aux procédures qui se fondent uniquement sur des données extraites, surtout celles qui excluent des portions significatives des cartes E311 des États-Unis.

Mots clés : Ajustement de la pondération; dénombrement à la frontière; E311; concordance état–code postal; extraction; saisie; reconnaissance intelligente de caractères.

[†] Jack Singleton, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, immeuble R.-H.-Coats, 16^e étage, Ottawa (Ontario), K1A 0T6.

EXECUTIVE SUMMARY

The Air Exit Survey - US visitors (AES US) is being introduced as an International Travel Survey (ITS) improvement initiative. Adjustment of ITS weights in classes defined by state of residence is being implemented to improve the quality of estimates for US air travellers. This weighting adjustment requires counts of travellers by state of residence, which cannot be produced currently since US address information is not keyed in ITS E311 processing. Statistics Canada studied development of a methodology for estimating state proportions, which would be applied to US frontier counts to give the required totals.

Three principal focal points of the study are obtaining US address information, assigning state of residence and estimating state proportions. After background research, it was decided to include the E311 fields Province/State (state) and Postal/Zip (zip) and set aside other fields, including Town/City and Country of Residence. A concordance file linking state and zip code was selected after comparing alternatives. Statistics Canada operations personnel agreed that keying US address data would be possible within the current production system. Arrangements were made to key state, zip and other fields, outside of ITS E311 processing, for three airports – Toronto, Montreal and Vancouver. Improving efficiency through use of US address data extracted in E311 processing was investigated. Three procedures for estimating state proportions based on extracted data were elaborated, along with a recommended procedure that combines extracted and keyed data. These procedures were evaluated in terms of coverage and precision, by comparison to benchmarks calculated based on keyed data. To study behaviour over time, a full year of estimates of state proportions was calculated for each of the three airports, using the ITS volume sample.

Highlights of the study include:

- Few high-quality, current state-zip code concordance files are available free of charge. The study selected a file produced by the United States Census Bureau for the 2000 Census.
- Most US travellers use the standard two-letter postal abbreviation when reporting state, although other common short forms were identified.
- It is operationally feasible to key state and zip data, although the set-up costs would be high. Also, if cards beyond the current ITS sample were needed, keying of fields other than US address components would be unavoidable.
- For over 60% of US E311 cards, state of residence can be assigned from extracted data with error rates less than 1%. For the balance, strategies that combine extracted and keyed data are recommended.
- Estimates of state proportions were generally within two percentage points for all procedures studied. However the study identified risks associated with procedures that use extracted data alone, especially those that exclude significant portions of US E311 cards. Such risks include unexplained spikes or dips, anomalies that impact states disproportionately and estimates based on sample sizes too small to represent all states.

TABLE OF CONTENTS

1. BACKGROUND

The Air Exit Survey – US visitors (AES US) is being introduced as an International Travel Survey (ITS) improvement initiative. An ITS requirement is reliable estimates for US regions or other subpopulations defined by state of traveller residence. Some control over the composition of the AES US sample is achieved by including flight destination in the selection of time stints for interviewing. However the AES US is vulnerable to samples that are not geographically representative, for two primary reasons. First, for many travellers the flight departing Canada does not land in their home state. This is particularly true of travellers taking connecting flights home and of flights to "hub" airports. Secondly, large numbers of US-bound flights at major airports lead to travellers of multiple flights waiting together. Although flight destination does influence the placement of AES US interviewers, they are instructed to proceed with any US resident, and do not consider flight destination or respondent residence. Therefore, adjustment of survey weights in classes defined by state of residence will be implemented to counter this vulnerability.

Counts of numbers of travellers by state of residence are required for this weighting adjustment. Such state counts will be produced from US frontier counts, tabulated from Canada Border Services Agency (CBSA) Declaration Cards. CBSA Declaration Cards, which are commonly known as E311 cards, are completed by commercial air travellers entering Canada on international flights. E311 cards collect traveller data including residency and trip characteristics. Although the primary purpose of E311 cards is Customs clearance of travellers entering Canada, they are provided to Statistics Canada through a Memorandum of Understanding (MoU) between the two agencies and are used for tabulation of frontier counts. The current E311 card allows up to four travellers of the same residence to report using the same card. An image of a blank E311 card is given in Appendix A. State counts cannot be produced in regular ITS production at this point in time, since keying (capture) of address information is not part of ITS E311 processing for US travellers.

Development of a methodology for estimating state proportions is the subject of the study described in this report. State proportions, indicating the share of US air travellers from that state, would be applied to frontier counts to give state counts. The study seeks to answer the following questions:

- What E311 address fields are needed to assign a state of residence? What concordance files or other background information is required?
- Is keying of US address information operationally feasible in ITS E311 processing? What would be the costs, both of development and on-going? What would be the impact on ITS operations?
- Could US address data extracted in E311 processing be used to improve efficiency in estimating state proportions? As this extracted data has never been studied, what is its quality?
- What would be procedures for estimating state proportions? Would the current ITS sample be adequate for supporting such estimation? What would be the cost and impact on ITS operations?

Future work will investigate different strategies for weighting adjustment, based on the characteristics of state counts and needs of the ITS. However elaboration of a weighting adjustment strategy[1] is beyond the scope of this report.

The document is organized as follows. Section 2 describes the methodology of the study, the principal steps of which are obtaining US address information, assigning state of residence and estimating state proportions. Options are presented, including their operational impact and cost, and evaluation of them is discussed. Production of a full year of estimated state proportions using ITS production data is presented in Section 3. Conclusions are provided in Section 4, and Section 5 rounds out the report with recommendations and plans for future work.

2. METHODOLOGY

Because US address information is not part of current ITS processing, new methods are required to both assign state of residence and produce the counts required for weighting adjustment. Statistics Canada undertook a study to determine the feasibility of doing this, the methodology of which is described here. The first steps, reported in Section 2.1, were investigating obtaining US address information from E311 data through extraction and keying and, subsequently, selecting and pre-processing fields for assigning state of residence. In Section 2.2 we research available reference data on US addresses and define a set of rules for assigning state of residence based on E311 and reference data. The rules were evaluated through comparison to keyed data, and quality statistics were calculated. In Section 2.3, several approaches for estimating state proportions based on these rules were proposed. Estimated proportions and related statistics, including confidence intervals, were calculated for several states, and feasible approaches for subsequent study were identified based on the results and operational constraints. Before outlining the details of the methodology, we note a couple of points for clarification.

We focus on assigning state of residence and estimating state proportions, although weighting adjustment may be at regions or other groups of states. We estimate 51 proportions, as the ITS considers the 50 states and the District of Columbia as valid US states of residence. We thus exclude US protectorates such as Puerto Rico. Because we are estimating state proportions, we are not required to assign state of residence to all E311 cards.

For purposes of estimating state proportions we define US air travellers by the ITS variables Traveller Type Final (TTF) and Mode of Transport (MoT)[2]. We limit our interest to assigning US state, and do not attempt to assign other similar-level geographies, including Canadian provinces or jurisdictions within overseas countries, that

[1] As ITS and AES questionnaires are combined in weighting, such weighting adjustment would be applied to all questionnaires of US air travellers.

[2] In particular, TTF=15 (US residents) and MoT=1 (commercial air travellers).

are found in the address fields of US cards. As well, our study excludes those cards with US addresses that are not classified as US air travellers using TTF and MoT.

2.1 Obtaining US address information

After examining E311 US address data, it was decided that our procedures for assigning state of residence would be based on the fields Province/State and Postal/Zip. Other E311 data fields that could provide relevant information for assigning state of residence were considered by the study. The primary examples are Town/City and Country of Residence. Town/City was excluded due to lack of a comprehensive US city-state or city-zip code concordance file, the construction of which is complicated by city names used in multiple states. Country of Residence was not included since we do not study assigning non-US geographies. The next two sections describe how E311 and reference data can be used to assign state of residence from each of the two chosen fields.

2.1.1 Province/State and pre-processing of state data

The space to enter Province/State on an E311 card is limited (see Appendix A). Most US residents use the standard two-letter postal abbreviation, such as "CA" for California. However analysis of E311 data revealed other nomenclatures are also employed, such as full state names when space permits (e.g., "OHIO") or other common short forms (e.g., "PENN" rather than "PA" for Pennsylvania). As the first two letters of some such nomenclatures coincide with the standard abbreviation of another state (e.g., "MINN" for Minnesota, with "MI" the standard abbreviation for Michigan), the first two characters may not be adequate. A set of rules for pre-processing E311 Province/State data was created. See Appendix C1 for a full specification of these rules.

For the balance of this document we use the term State to refer to data obtained from the Province/State field. Because our interest lies in estimating state proportions, we create a variable State(State) that takes on 52 values – the standard two-letter postal abbreviations of the 51 ITS states of residence and a null or missing value that indicates state of residence is undetermined. We speak of State(State) being "valid" or indicating a state of residence when it specifies one of the 51 ITS states of residence, and being "not valid" otherwise. Thus State(State) is valid if Province/State is equal to one of the 51 standard two-letter postal abbreviations after applying pre-processing, and is not valid otherwise.

2.1.2 Postal/Zip and state-zip code concordance

Zip codes are defined by the United States Postal Service (USPS) to facilitate mail delivery. Our interest in zip code lies in its relationship to state. Prior to beginning the study we investigated the possibilities for obtaining concordance files that would link zip codes to state and sub-state geographies. Our research revealed that the candidates could, generally speaking, be divided into two groups. The first give high-quality and detailed neighbourhood and socio-demographic information for each zip code. These are integral in market research, and are available for purchase from the USPS or private companies. We found that few such files are available free of charge, and that those available are generally dated. The second group give a zip code range (based on the first two or three digits) for each state, and ignore exceptions. Many such files are posted on the internet and slight differences exist among them.

The concordance file selected for the study was produced by the US Census Bureau (USBC) as a reference material accompanying release of data from the 2000 US Census. It contained approximately 33,000 zip codes (five digits[3]) with state and population data for each. On it zip codes generally respect state borders, although some exceptions exist (most are remote areas served by a USPS installation located in a neighbouring state). Generally a range of prefixes (first three digits) is assigned to each state. However some zip codes are assigned to locations outside the state range, for historical or other reasons. Our investigation revealed that for most multiple-state zip codes or prefixes, the vast majority of population (often in excess of 95%) belonged to one (primary) state. We constructed four concordance files for use in the study, listed below with the number of members in parentheses. The first two contain only zip codes or prefixes that correspond to a unique state, and are used in early portions of the study. The final pair of concordance files also contains the multiple-state zip codes or prefixes, and is used in latter portions of the study and will be used in future work. Lists of the five- and three-digit zip codes that correspond to multiple states are given in Appendices B1 and B2. The four concordance files are:

- ***StateZip5*** – five-digit zip codes that correspond to a single state (31,871)
- ***StateZip3*** – three-digit values for which all zip codes thus beginning correspond to a single state (846)
- ***StateZip5a*** – StateZip5 augmented by multiple-state zip codes. For the 42 multiple-state zip codes, a primary state was selected based on population (31,871+42=31,913)
- ***StateZip3a*** – constructed analogously to StateZip5a, but based on StateZip3 (846+37=883).

The Census 2000 file was the most recent available at the time of the study, as the USBC does not update files inter-Census. We could not quantify the impact of being a decade out of date. However we did explore the potential gains from an expanded source of state-zip code concordance, such as those found on the internet. We found these yielded few matches beyond those found using the chosen concordance file. They were not pursued further.

We use the term Zip to refer to data obtained from the Postal/Zip field. Because our interest lies in assigning state of residence, we create a variable State(Zip5) to express the state of residence obtained from the full (five-digit) zip code through a concordance file StateZip5 or StateZip5a. As noted above, assigning state of residence may not require obtaining a full zip code, and we thus create a variable State(Zip3) to express the state of

[3] We observed a small subset of US travellers report a nine-digit zip code. As the last four digits are not beneficial in assigning state of residence, they are ignored. However we did recommend extracting an additional character in the Postal/Zip field, in order to not truncate zip codes of the form XXXXX-XXXX.

residence obtained from the partial (three-digit) zip code, through a concordance file StateZip3 or StateZip3a. Some notes:

- Both of these created variables take on 52 values and we speak of them being valid or not valid.
- We use State(Zip3) to assign state of residence only if State(Zip5) is not valid.

2.2 Assigning state of residence

The focus of this section is developing a methodology to assign state of residence based on E311 and reference data. We first discuss realizations of the variables State and Zip, both those available currently from extraction in ITS production and keyed values that were obtained specifically for the study. The extracted and keyed data were used to specify rules for assigning state of residence, and to define a framework for evaluating their quality. These quality measures are a useful tool in selecting E311 cards from which estimates of state proportions will be calculated.

2.2.1 Extracted data

As noted earlier, the only realizations of State and Zip currently available are from extraction of E311 data in ITS processing. In this section, we first give a brief summary of ITS E311 processing up to the point of extraction, followed by description of the extractions of these two data items.

As per the MoU between CBSA and Statistics Canada, E311 cards are sent by CBSA from major airports to Statistics Canada at regular intervals. Received card are batched and sorted in preparation for processing. In the first step of processing, an electronic copy or image of each card is made in a process known as scanning or imaging. Next, data is obtained from card images in extraction. A template, corresponding to the E311 card layout, is placed over each card image. The template delineates an area on the card image for each data item; these areas correspond to the sections filled in by travellers (see Appendix A). Extraction software obtains data from the specified areas of the template. We use the terms extracted state and extracted zip in reference to the extractions of Province/State and Postal/Zip, respectively. Each consists of a string of characters obtained through Intelligent Character Recognition (ICR). The quality of extracted state and zip has never been studied.

As fully-automated processes, imaging and extraction are subject to errors for a variety of reasons. First, if the correct template is not identified or the template is not aligned properly, the application will not be able to extract the address information. This also occurs when the address information is displaced, as often happens when an initial response is crossed out and corrected by a CBSA Officer. Errant markings on the card, by CBSA Officers or otherwise, may lead to extraction of misleading data.

The programming of the ICR application may also lead to error. The set of characters that extracted values are permitted to contain is a parameter of the ICR application, and varies by field depending on the characteristics of that data item. For example, Postal/Zip extractions allow both letters and numbers, as the field must accommodate both US zip codes and Canadian postal codes. In contrast, extractions of Province/State are limited to letters. The programming of the ICR application influences the data extracted. For example, in August 2010 the programming of the Postal/Zip extraction was modified to no longer permit punctuation marks, including the slash (/). This led to greater recognition of the digit "1", by including values written at an angle that were previously interpreted as "/".[4] Pre-processing of State and Zip data to work in harmony with the ICR application will be discussed in Section 2.2.2.2.

The lack of keyed data limits the scope for our study in that i) we have no benchmark by which to measure the quality of the extracted values or of state proportions estimated from them and ii) we have no indication of the improvement in state proportions that could be achieved by using both extracted data and keyed data. Due to these limitations, it was consequently decided to arrange keying of E311 address data for US travellers.

2.2.2 Keyed data

Although not carried out for US address data, the ITS E311 processing system is designed to permit capture (keying) of any specified data field of cards selected for the ITS sample. For alphanumeric fields like State and Zip, keyed values are obtained by entering a string of characters, based on viewing the card image. To save keystrokes, a proposed value may be presented to the keyer, who has the option of confirming this value or entering another value[5]. Based on ongoing ITS experience, the data keying operation yields data of superior quality to extracted values.

Two phases took place – an initial phase of "desktop capture" by a single individual and a subsequent phase of "offline production capture" by the production team which captures data for the ITS sample. The first phase verified that State and Zip data can be interpreted correctly by human viewing and helped develop the procedures for the second phase. The second phase proceeded to capture substantive data for statistical analysis. In both phases, data on the US E311 universe file were stratified based on characteristics of extracted state and zip, and a random sample was selected in each study stratum. It is important to note this stratification, for study purposes, is different than that employed in current ITS production. In both phases keyers were instructed to enter what was viewed on the card image, for each field individually. Even though it was clear US address data were being entered, keyers were not given background information such as a list of state

[4] Prior to August 2010, punctuation marks including /,&$)*-.:~ were permitted in the extraction of Postal/Zip. From August 2010 onward, extraction was limited to letters, numbers and selected punctuation marks including the dash. Also, the extractions of both fields allow for spaces and convert all lower-case letters to capital letters.

[5] Confirming a proposed value leads to savings as the keyer does not need to enter the value. However correcting a proposed value requires more keystrokes than if no value were proposed, as the keyer needs to reject the proposed value and then enter the corrected value.

abbreviations and did not see the extracted values. The two phases of capture are described below.

2.2.2.1 Desktop capture

A Statistics Canada employee with experience sorting E311 cards, but not currently working on the ITS, was available to the project team. A system was set up for desktop capture, consisting of three main components:

- Selection of card images
- Desktop access to images of selected cards
- Excel spreadsheet for entry of data keyed from the card image.

Work was organized in batches, with the keying instructions, including the fields to be keyed, differing by batch. Controls were put in place to allow a batch to be completed over several days, and to ensure the captured data matched the correct card image. There were two steps of desktop capture. The first yielded estimates of the throughput rate for various operations, gave a glimpse of the quality of extracted data and helped identify which fields would be captured subsequently. The second studied gains in efficiency from stratifying based on characteristics of US address information.

The first step of desktop capture consisted of 500 cards selected at random from US travellers arriving at Pearson (Toronto) International Airport in April 2010[6]. To further assess the quality of extracted data and to evaluate the quality of captured data, images where extracted values disagreed with desktop capture were sent for parallel capture (verification capture) by another individual. In contrast to desktop capture, the entire address was utilized in verification capture, along with the individual's knowledge of US address data.

Results of desktop and verification capture for the 500 observations of Postal/Zip and Province/State are presented in Table A below. As an initial attempt, the focus was on the quality of the two fields as data items, rather than as address components. However we did incorporate the goals of our study by evaluating only the first two characters of Province/State and the first five of Postal/Zip.

Table A
Desktop capture, Toronto April 2010 data

Variable	*Exclusions*	*Extracted= Captured*	*Extraction error*	*Capture error*
Province/State (2 char)	*27*	*377*	*88*	*8*
Postal/Zip (5 char)	*22*	*299*	*172*	*7*

[6] Toronto was selected as one of the first airports at which the AES US was introduced. The most recent month for which final E311 data were available was selected at each step of the study.

Reasons for exclusion included non-US cards, an incorrect card image was specified to the keyer, and data were in an incorrect area of a card. An extraction error occurred when the extracted and captured values differed, and verification showed the captured value was correct. In contrast, for a capture error, verification revealed the captured value was incorrect. Two reasons for the higher error counts for Postal/Zip were the need to match on five characters rather than two and the presence of letters and numbers. Because many similar zip codes are associated with the same state, this may be an overstatement of the relevant error count when assigning state of residence.

Verification capture was extended to a sample of cards where the extracted and captured values agreed. Verification capture confirmed that, in all instances, the common value was correct. Analysis of capture errors revealed that many could have been averted if knowledge of US addresses were incorporated into keying. However, as such specialized keying is not possible in the ITS operational framework, it will not be considered for implementation.

To study the capacity to assign state of residence and to evaluate the quality of extracted data in greater depth, the second step was to select further samples of Toronto E311 data. Four study strata were created, based on State(State), State(Zip5) and State(Zip3) as calculated using the concordance files StateZip5 and StateZip3. The four study strata are defined in Table B. Random samples were selected sequentially from the four study strata (May 2010 E311 data) and sent for desktop keying of State and Zip. The samples contained just over 1,500 cards. The results of keying State and Zip were compared to extracted state and zip.

We calculate disagreement rates for the first two characters of State (as most travellers use two-letter abbreviations) and the first three characters of Zip (as these generally are sufficient to determine state of residence). Due to clustering of zip codes, extraction errors in the second or third character of Zip may not lead to an incorrect state of residence. We therefore calculate a third disagreement rate, which does not count differences in the first three characters of Zip that do not lead to an incorrect state of residence. Results of the second step of desktop capture are provided in Table B below.

Table B
Desktop capture, Toronto May 2010 data

Study stratum	*Sample*	*Population*	*Population (%)*	*Disagreement rate (%)*		
				State (2 char)	*Zip (3 char)*	*State from Zip*
State(State)=State(Zip)	*695*	*52,252*	*56.3*	*0.0*	*8.7*	*0.3*
State(Zip5) valid	*108*	*10,810*	*11.6*	-	*13.0*	*11.1*
State(Zip3) valid	*253*	*7,508*	*8.1*	-	*67.6*	*60.9*
State(State) valid	*462*	*9,618*	*10.4*	*6.3*	-	-

We note that these four study strata, collectively, cover 86% of the Toronto May 2010 population. The remaining 14% was excluded from sampling, as no disagreement rate could be calculated since neither extracted value indicated a state of residence. For 56% of the population, the extracted zip code and extracted state both indicate the same state of residence, for which captured values indicate a very low error rate. The full zip code shows promise as an indicator of state of residence, as does the state field[7]. The partial zip code shows less promise as an indicator of state of residence. However refinement of the criteria for inclusion could lead to better results as this study stratum contains non-US cards and cards where the partial zip code and the state indicate different states of residence. These results also show looking at the first three digits of the zip code overestimates the error rate for assigning state of residence.

Desktop capture was able to produce high-quality keyed data, offered flexibility in exploring options and gave insight regarding the extent to which E311 data could be used to assign state of residence. However its capacity in terms of volume of cards was limited, with only a single individual available on a part-time basis for a specified time period. Moreover, initial investigation also revealed that large samples would be required to make inferences on a population basis with a high degree of confidence, as error rates were low. Also, several months of keyed data might be required to accumulate sufficient sample for rare populations. It thus became clear that a larger-scale alternative to desktop capture was required.

2.2.2.2 Offline production capture

Arrangements were made with the area of Statistics Canada responsible for capture of the ITS sample. In order to not disturb ITS E311 processing operations and given that keying of US address data is not currently undertaken, special runs with new specifications were arranged and new samples of E311 cards were selected. Operational considerations implied the same data were captured for all selected cards:

- Postal/Zip
- Province/State
- Town/City
- Number of travellers (traveller count).

For the fourth, keyers were instructed to count the number of travellers by viewing the card image. This method of determining the number of travellers is different than actual E311 production, where the count is determined based on a combination of electronic and human interpretation of the card. Extracted versions of the three address components above were returned, in addition to keyed values. Because these card images were keyed as part of special runs, these extracted values may not coincide with those on the universe file.

[7] Lower error rates may be attainable in these two study strata. The disagreement rate in the "State(Zip5) valid" study stratum is overstated by the presence of cases where State(State), rather than State(Zip5), would be used to assign state of residence even though both are valid. Also no pre-processing of state data was incorporated in the definition of the "State(State) valid" study stratum, merely matching the first two characters to the 51 standard abbreviations.

The quality of offline production capture was not studied specifically. However some informal feedback was provided. Capture error rates were within acceptable guidelines. Common capture errors included:

- Adjacent keys on the keyboard, such as "K" and "L".
- Extra digits keyed, especially repeated digits. This often led to zip codes longer than five digits being keyed. For example, "21134" keyed as "211134".
- Similar letters, such as "D" and "O".

Similarly to the analysis of desktop capture data described in Section 2.2.2.1, data from offline production capture were used to study the assignment of state of residence based on extracted and keyed values. Although the procedures for defining study strata (henceforth called "state assignment classes"), assigning state of residence and analyzing results were similar to those used in desktop capture, several refinements were introduced. In particular:

- The operator State(State) described in Section 2.1.1 was enhanced to compensate for cases where numeric zip codes are extracted as letters, especially 1 and 0 as L, I or O. These additional rules are specified in Appendix C2. In particular this addressed observed underestimation for New York state, whose zip codes begin with "0" or "1".
- Separate state assignment classes were introduced for cases where extracted zip (full and partial) and extracted state were "in conflict" or indicated different states of residence. State assignment rules for these classes are described in the next section.
- The analysis framework was enhanced to provide multiple measures of the accuracy of state assignment.

Six stratified samples were selected for offline production capture. First, a sample of May 2010 Toronto data was used to confirm the results of desktop capture, refine state assignment rules and to investigate the potential for assigning state based on keyed values when extracted values cannot. This was followed by a sample of June 2010 Toronto data. We next selected samples, based on July and August 2010 data, for each of the two airports at which the AES US would be next introduced – Trudeau (Montreal) International Airport and Vancouver International Airport. The samples contained approximately 46,400 cards, sent in batches of approximately 500 cards each.

In the balance of Section 2 we describe methods for assigning state of residence and evaluating their quality and present results of analysis based on data from offline production capture.

2.2.3 State assignment rules

To group cards where a common rule can be used to assign state of residence, state assignment classes were defined based on State(State) and State(Zip). State assignment rules are based on extracted values and the concordance files StateZip5a and StateZip3a, and they always lead to assigning a valid state of residence. For state assignment classes where both extracted values indicate the same state or only one is valid, the rule is clear.

For the two conflict state assignment classes, extracted and keyed values were analyzed to see which extracted value is a better predictor of state of residence. For conflicts between State(Zip5) and State(State), State(Zip5) more often agreed with the keyed values than did State(State). However State(State) more often agreed with the keyed values than did State(Zip3), in case of conflict between these. For each of the two conflict state assignment classes, a set of exceptions was identified. That is, for conflicts between State(Zip5) and State(State), the exceptions give cases where State(State) is a better predictor. Similarly, the exceptions give cases where State(Zip3) is a better predictor than State(State). See Appendix C3 for a list of these exceptions. State assignment class descriptions and state assignment rules are given below. The classes are defined sequentially, so that each class includes only cases not assigned previously.

State assignment class (#)	***State assignment class (description)***	***Rule for assigning State of residence***
1	*State(State)=State(Zip)*	*State(State)*
2	*State(Zip5) valid*	*State(Zip5)*
3	*State(State) valid*	*State(State)*
4	*State(Zip3) valid*	*State(Zip3)*
5	*State(State)≠State(Zip3)*	*State(State) + exceptions*
6	*State(State)≠State(Zip5)*	*State(Zip5) + exceptions*

2.2.4 Evaluation framework for state assignment rules

Because the rules above assign state based on extracted values, keyed zip and state values can be used for evaluation. As a first step in formalizing the evaluation framework, we identify four mutually-exclusive outcomes representing the result of state assignment:

- ***Good***: the assigned state of residence equals the keyed value(s)
- ***Bad***: the assigned state of residence differs from the keyed value(s)
- ***Questionable***: the assigned state of residence equals one keyed value and the other keyed value indicates another state
- ***Inconclusive***: neither keyed value is valid, or the keyed values indicate two states, both of which are different from the assigned state of residence.

Possible outcomes are listed below. The value A indicates the assigned state of residence, while values B and C indicate other valid states. Blank indicates the state of residence of a keyed value is undetermined.

State Assigned	***Statekeyed***	***Zipkeyed***	***Result***
A	*A*	*Blank*	*Good*
A	*Blank*	*A*	*Good*
A	*A*	*A*	*Good*
A	*B*	*Blank*	*Bad*
A	*Blank*	*B*	*Bad*
A	*B*	*B*	*Bad*
A	*B*	*A*	*Questionable*
A	*A*	*B*	*Questionable*

A	*B*	*C*	*Inconclusive*
A	*Blank*	*Blank*	*Inconclusive*

The state assignment classes above cover all cards where at least one of extracted state or extracted zip is valid. To see if keyed values could be used to assign a state of residence and what characteristics of extracted data are correlated with success in doing so, we also selected a sample of cards where neither extracted state nor extracted zip is valid. For cards among these for which extracted state or extracted zip was "non-empty"[8], generally one or both of the keyed values was valid. However, if both extracted values were "empty", generally neither keyed field was valid. We thus define two additional state assignment classes for non-valid extracted values, based on the characteristics of State and Zip. In one, we attempt to assign state of residence based on keyed values while in the other we do not attempt to assign state of residence. These classes will be described subsequently.

With keyed data used to assign state of residence, the scope to analyze the quality of the state assigned is limited, and thus the definitions of the four outcomes given above are not relevant. However, by redefining the four outcomes we can quantify the success in obtaining a state of residence from keyed zip and keyed state when neither extracted value is valid. Good, Bad and Questionable correspond to the state assignment class in which we key state and zip code and assign a state of residence based on them. Inconclusive corresponds to the state assignment class in which we do not key values and thus do not assign a state of residence. The rules for state assignment and the definitions of the four outcomes for non-valid extracted values are given below. Note that here State Assigned is based on keyed values. In contrast, earlier State Assigned was based on extracted values and was compared to keyed values. Also, unlike other state assignment classes, there are multiple state assignment rules in class 7.

State assignment class (#)	***State assignment class (description)***	***Zipkeyed***	***Statekeyed***	***Result***	***State Assigned***
7	*State or Zip non-empty*	*Blank*	*A*	*Good*	*A*
7	*State or Zip non-empty*	*A*	*Blank*	*Good*	*A*
7	*State or Zip non-empty*	*A*	*A*	*Good*	*A*
7	*State or Zip non-empty*	*A*	*B*	*Questionable*	*Blank*
7	*State or Zip non-empty*	*Blank*	*Blank*	*Bad*	*Blank*
8	*State and Zip empty*	*Blank*	*Blank*	*Inconclusive*	*Blank*

2.2.5 Quality measures

To quantify the accuracy of state assignment in each class based on the evaluation framework described above, we define three measures *E1*, *E2* and *E3*, from the numbers

[8] In the terminology of this report, an extracted value is "empty" when the ICR application extracted a null value, i.e., a character string composed entirely of blanks. In contrast, an extracted value is "non-empty" when the ICR application extracted a non-null character string.

of Good, Bad and Questionable cards[9]. These measures differ in their consideration of the Questionable cards, and are refinements of the disagreement rates calculated in Section 2.2.2.1. In the context of assigning a state of residence based on extracted values, these measures represent error rates, where an error is assigning an incorrect state of residence. In the context of attempting to assign a state of residence based on keyed values (when extracted values are not valid), these measures represent failure rates, where a failure is not being able to assign a state of residence. The three quality measures are expressed mathematically as:

$$E1 = \frac{Bad}{Good + Bad + Questionable} * 100$$

$$E2 = \frac{Bad}{Good + Bad} * 100$$

$$E3 = \frac{Bad + Questionable}{Good + Bad + Questionable} * 100$$

In general, the three quality measures convey the following assumptions:

- ***E1***: consider Questionable as Good (no error/no failure)
- ***E2***: consider Questionable as Inconclusive (exclude from calculation)
- ***E3***: consider Questionable as Bad (error/failure).

It can be shown that $E1 \leq E2 \leq E3$. As the number of Questionable cards is generally small in relation to Good and Bad, the three measures should be close in value. However they do provide insight, especially for processes with small error rates.

2.2.6 Results

Results of applying the state assignment rules to the six samples described in Section 2.2.2.2 are presented in Table C below. For each state assignment class we give[10]:

- ***Pop'n***: the number of cards on the US E311 universe file
- ***Pct.***: *Pop'n* expressed as a percentage of the population
- ***Cum.***: *Pop'n* expressed as a cumulative percentage of the population
- ***Samp.***: the number of cards selected in the offline production sample
- ***Good, Bad, Quest., Incon.***: estimates of the numbers of Good, Bad, Questionable and Inconclusive cards, calculated by weighting the numbers in the offline production sample using the weights of the study stratification
- ***E1, E2, E3***: values of the three quality measures calculated from these estimates.

[9] The calculations exclude cards classified as Inconclusive, as the quality of state assignment cannot be assessed for these.

[10] For completeness we include class 8, even though no attempt assign a state of residence was made for cards in it. Note that no values of E1, E2 and E3 are calculated for this state assignment class.

Table C
Analysis of state assignment rules

Sample and State assignment class	*Pop'n*	*Pct.*	*Cum.*	*Samp.*	*Good*	*Bad*	*Quest.*	*Incon.*	*E1*	*E2*	*E3*
May 2010 Toronto	***92,866***	***100.0***		***12,324***	***85,157***	***6,782***	***1,631***	***1,228***			
State(State)=State(Zip)	*52,218*	*56.2*	*56.2*	*3,898*	*51,898*	*0*	*320*	*0*	*0.0*	*0.0*	*0.6*
State(Zip5) valid	*8,328*	*9.0*	*65.2*	*1,694*	*7,630*	*296*	*200*	*114*	*3.6*	*3.7*	*6.1*
State(State) valid	*15,141*	*16.3*	*81.5*	*2,125*	*14,172*	*796*	*340*	*139*	*5.2*	*5.3*	*7.4*
State(Zip3) valid	*3,462*	*3.7*	*85.2*	*1,103*	*2,119*	*968*	*88*	*239*	*30.5*	*31.4*	*33.3*
State(State)≠State(Zip3)	*3,967*	*4.3*	*89.5*	*992*	*3,175*	*540*	*212*	*40*	*13.7*	*14.5*	*19.1*
State(State)≠State(Zip5)	*2,849*	*3.1*	*92.6*	*950*	*1,395*	*1,047*	*342*	*66*	*37.6*	*42.9*	*49.9*
State or Zip non-empty	*6,271*	*6.8*	*99.3*	*1,356*	*4,769*	*3,134*	*129*	*0*	*39.0*	*39.7*	*40.6*
State and Zip empty	*630*	*0.7*	*100.0*	*206*	*0*	*0*	*0*	*630*			
June 2010 Toronto	***117,445***	***100.0***		***11,040***	***108,212***	***5,926***	***1,979***	***1,328***			
State(State)=State(Zip)	*71,999*	*61.3*	*61.3*	*1,714*	*71,452*	*3*	*544*	*0*	*0.0*	*0.0*	*0.8*
State(Zip5) valid	*10,770*	*9.2*	*70.5*	*1,489*	*10,054*	*376*	*208*	*132*	*3.5*	*3.6*	*5.5*
State(State) valid	*14,250*	*12.1*	*82.6*	*2,403*	*13,162*	*712*	*286*	*90*	*5.0*	*5.1*	*7.0*
State(Zip3) valid	*4,475*	*3.8*	*86.4*	*1,119*	*2,847*	*1,164*	*144*	*320*	*28.0*	*29.0*	*31.5*
State(State)≠State(Zip3)	*4,335*	*3.7*	*90.1*	*1,072*	*3,651*	*430*	*238*	*16*	*10.0*	*10.5*	*15.5*
State(State)≠State(Zip5)	*3,175*	*2.7*	*92.8*	*1,048*	*1,779*	*1,022*	*332*	*42*	*32.6*	*36.5*	*43.2*
State or Zip non-empty	*7,713*	*6.6*	*99.4*	*2,013*	*5,266*	*2,219*	*228*	*0*	*28.8*	*29.6*	*31.7*
State and Zip empty	*728*	*0.6*	*100.0*	*182*	*0*	*0*	*0*	*728*			
July 2010 Montreal	***45,642***	***100.0***		***7,358***	***42,408***	***1,952***	***787***	***495***			
State(State)=State(Zip)	*29,222*	*64.0*	*64.0*	*3,145*	*29,002*	*0*	*220*	*0*	*0.0*	*0.0*	*0.8*
State(Zip5) valid	*4,197*	*9.2*	*73.2*	*423*	*3,907*	*120*	*110*	*60*	*2.9*	*3.0*	*5.6*
State(State) valid	*4,847*	*10.6*	*83.8*	*986*	*4,530*	*205*	*97*	*15*	*4.2*	*4.3*	*6.2*
State(Zip3) valid	*1,624*	*3.6*	*87.4*	*553*	*1,073*	*442*	*53*	*56*	*28.2*	*29.2*	*31.6*
State(State)≠State(Zip3)	*1,585*	*3.5*	*90.9*	*519*	*1,320*	*148*	*97*	*20*	*9.4*	*10.1*	*15.6*
State(State)≠State(Zip5)	*1,247*	*2.7*	*93.6*	*416*	*692*	*396*	*138*	*21*	*32.3*	*36.4*	*43.5*
State or Zip non-empty	*2,597*	*5.7*	*99.3*	*1,299*	*1,883*	*642*	*72*	*0*	*24.7*	*25.4*	*27.5*
State and Zip empty	*323*	*0.7*	*100.0*	*17*	*0*	*0*	*0*	*323*			
July 2010 Vancouver	***75,624***	***100.0***		***6,682***	***69,483***	***4,015***	***1,056***	***1,071***			
State(State)=State(Zip)	*48,940*	*64.7*	*64.7*	*2,236*	*48,648*	*40*	*252*	*0*	*0.1*	*0.1*	*0.6*
State(Zip5) valid	*7,697*	*10.2*	*74.9*	*771*	*7,087*	*380*	*190*	*40*	*5.0*	*5.1*	*7.4*
State(State) valid	*6,202*	*8.2*	*83.1*	*634*	*5,720*	*304*	*108*	*70*	*5.0*	*5.0*	*6.7*
State(Zip3) valid	*2,674*	*3.5*	*86.6*	*669*	*1,551*	*871*	*60*	*192*	*35.1*	*36.0*	*37.5*
State(State)≠State(Zip3)	*3,276*	*4.3*	*91.0*	*807*	*2,744*	*346*	*162*	*24*	*10.6*	*11.2*	*15.6*
State(State)≠State(Zip5)	*2,662*	*3.5*	*94.5*	*663*	*1,481*	*921*	*224*	*36*	*35.1*	*38.4*	*43.6*
State or Zip non-empty	*3,464*	*4.6*	*99.1*	*866*	*2,252*	*1,152*	*60*	*0*	*33.3*	*33.8*	*35.0*
State and Zip empty	*709*	*0.9*	*100.0*	*36*	*0*	*0*	*0*	*709*			
August 2010 Montreal	***46,633***	***100.0***		***4,371***	***43,192***	***2,348***	***533***	***560***			
State(State)=State(Zip)	*30,419*	*65.2*	*65.2*	*326*	*30,406*	*3*	*10*	*0*	*0.0*	*0.0*	*0.0*
State(Zip5) valid	*5,105*	*10.9*	*76.2*	*512*	*4,656*	*210*	*150*	*90*	*4.2*	*4.3*	*7.2*
State(State) valid	*3,353*	*7.2*	*83.4*	*681*	*3,115*	*174*	*45*	*20*	*5.2*	*5.3*	*6.6*
State(Zip3) valid	*1,914*	*4.1*	*87.5*	*653*	*1,159*	*581*	*55*	*119*	*32.4*	*33.4*	*35.4*
State(State)≠State(Zip3)	*1,981*	*4.2*	*91.7*	*649*	*1,692*	*181*	*91*	*17*	*9.2*	*9.7*	*13.8*
State(State)≠State(Zip5)	*1,471*	*3.2*	*94.9*	*490*	*812*	*518*	*129*	*12*	*35.5*	*38.9*	*44.3*
State or Zip non-empty	*2,087*	*4.5*	*99.4*	*1,044*	*1,351*	*682*	*54*	*0*	*32.7*	*33.5*	*35.2*
State and Zip empty	*302*	*0.6*	*100.0*	*16*	*0*	*0*	*0*	*302*			
Aug. 2010 Vancouver	***65,759***	***100.0***		***4,636***	***60,554***	***3,518***	***794***	***892***			
State(State)=State(Zip)	*42,481*	*64.6*	*64.6*	*369*	*42,449*	*0*	*32*	*0*	*0.0*	*0.0*	*0.1*
State(Zip5) valid	*6,906*	*10.5*	*75.1*	*692*	*6,446*	*280*	*150*	*30*	*4.1*	*4.2*	*6.2*
State(State) valid	*4,830*	*7.3*	*82.4*	*611*	*4,511*	*216*	*80*	*24*	*4.5*	*4.6*	*6.2*
State(Zip3) valid	*2,564*	*3.9*	*86.3*	*848*	*1,421*	*879*	*77*	*187*	*37.0*	*38.2*	*40.2*
State(State)≠State(Zip3)	*2,950*	*4.5*	*90.8*	*729*	*2,486*	*284*	*164*	*16*	*9.7*	*10.3*	*15.3*
State(State)≠State(Zip5)	*2,420*	*3.7*	*94.5*	*604*	*1,319*	*853*	*212*	*36*	*35.8*	*39.3*	*44.7*
State or Zip non-empty	*3,009*	*4.6*	*99.1*	*753*	*1,922*	*1,007*	*80*	*0*	*33.5*	*34.4*	*36.1*
State and Zip empty	*599*	*0.9*	*100.0*	*30*	*0*	*0*	*0*	*599*			

2.2.7 Observations

Study of the results in Table C yields several observations on application of the rules for state assignment. In this section we discuss these observations and pose some underlying explanations.

- May 2010 Toronto results are not comparable to the other samples, in part because they did not contain the methodological refinements applied to subsequent samples.
    - The absence of pre-processing led to a lower population coverage of class 1.
    - Higher error rates were observed for the conflict classes. In part this could be attributed to not incorporating the exceptions described in Section 2.2.3.
    - Oversampling of rare populations for study purposes contributed to a higher failure rate in class 7.
- Population coverage by class is consistent over the six samples.
    - Class 1: over 60% except for May
    - Classes 1-3: 80% to 85%
    - Classes 1-6: 90% to 95%
    - Classes 7 and 8: 5% to 10%
    - Class 8: less than 1%.
- Very low error rates were observed when both state and zip are valid and indicate the same state (class 1). These will be denoted "very high confidence".
- High quality results were observed if either the state or the full zip code is valid (classes 2 and 3) – "high confidence". For these, state can be assigned using the extracted values. However a mechanism for monitoring, such as quality control sampling, is recommended.
- If the full zip code and the state indicate different states (class 6), keying values to obtain state of residence is recommended. The high error rate for zip (the better of the two predictors) is not a reflection of its quality, but of the choice between two high-confidence options.
- Higher error rates were observed when only a partial zip code is valid (class 4). This class contains many non-US cards. Extractions of Canadian postal codes or overseas jurisdiction indicators often begin with three digits[11]. An example is a Canadian postal code beginning with "L3B" that is extracted as "138". With 883 three-digit combinations on the concordance file, matches are not unexpected. The error rate is too high to assign state of residence based on extracted zip. Thus keyed values would be required to assign state of residence for this class. As described in Section 2.2.2, a state could be proposed to the keyer based on the partial zip code. Although specifying operational procedures is beyond the scope of this report, it is noted that a linkage mechanism to allow the keyer to associate a proposed state (expressed as a name or abbreviation) with the zip code (sequence of digits) would be needed.
- When the partial zip code and the state indicate different states (class 5), we have a choice between a high-confidence option and one for which extracted values are not adequate to assign state of residence. It is therefore not surprising that state can be

[11] In fact, actual values of some overseas jurisdiction indicators begin with three digits. For example, French and German postal addresses contain a five-digit numeric postal code.

used to assign state of residence with an error rate around 10%. However this error rate is too high to assign state of residence based on extracted values, and keying values would be required. As with class 4, the extracted values could possibly lead to keystroke savings through proposing a state of residence.

- The classes where neither extracted value is valid (classes 7 and 8) cannot be ignored when studying assigning state of residence. However, for all but a small portion (less than 1% of the population) there is reason to believe a state of residence can be assigned based on keyed values. We therefore recommend keying state and zip for class 7. Our study shows that doing so leads to a state in about two-thirds of cases.
- The subset of the population where we cannot assign a state of residence (class 8 plus class 7 Bad and Questionable) was shown to be less than 3% in our study[12]. We recommend monitoring this population, as action would be required if its share were to grow. As well, advances in ICR technology could lead to changes in the composition and size of this population.
- There is limited scope to compare month-to-month behaviour. Improvement from May to June observed in Toronto samples was due primarily to refinements in state assignment methodology, as described in the first point of this section. July and August results were similar, and thus there was no indication that changes to Zip extraction (see Section 2.2.1) led to improvement between July and August. Additional samples from subsequent months are required to draw conclusions.
- Although population sizes and the mix of states of residence differed among the airports, results were similar.

Although the methods and analysis of this section address the capacity to assign state of residence on a card-by-card basis, our interest lies in the quality of aggregate estimates of state proportions derived from E311 cards. As such estimation does not require assigning a state of residence to all E311 cards, an important aspect is evaluating trade-offs between the stability of including more state assignment classes and the gains in precision by restricting to those of highest quality.

2.3 Estimating state proportions

In this section we define and evaluate five approaches to produce estimated state proportions. The approaches differ in terms of the state assignment classes included and the use of extracted and keyed data to assign state of residence. By being more selective in terms of which classes are included we are able to restrict to cards for which we assign state of residence with high confidence. However the trade-off is increased variability and potential bias, as our estimated proportions are based on a smaller, and not necessarily representative, subset of E311 cards. For each approach, we calculate estimated proportions based on offline production data. To quantify variability, we also calculate confidence intervals based on the number of cards for which state of residence is assigned.

[12] The inclusion of rare populations in the May 2010 Toronto sample led to a slightly higher rate (just over 4%).

2.3.1 Approaches

The first two approaches assume keyed values are available.

- ***Approach 0*:** This approach requires extracted and keyed values of state and zip for all cards. State of residence is assigned from keyed values, using extracted values only if keyed state and zip are both blank or in conflict[13]. This approach attempts to assign a state of residence to all cards, but undetermined state of residence is one possible outcome. Although this approach may not be attainable in production, it provides a benchmark in analysis of offline production data and represents a "gold standard".
- ***Approach 1*:** This approach corresponds to the methodology recommended in the last section. For the three high-confidence classes, state of residence is assigned based on extracted values. The acceptance of extracted values implies that they would bypass the keying operations, thereby reducing keying costs. For the classes in which proposing an extracted value is recommended, state of residence is assigned using the keyed version of the proposed extracted value. Operationally, keying costs are reduced for the cases where the recommended value is accepted. Keyed values are used in class 7. State of residence is undetermined in class 8, where no attempt is made to assign a value, and also when keyed values are not valid in classes 4-7.

Even though our recommended approach includes keyed data, we define three approaches that assign state of residence based on extracted values only. These approaches are included because, if adopted, they could be implemented without incurring additional processing cost to the ITS, as compared to current E311 data capture operations. They offer a trade-off of population coverage and heterogeneity versus confidence in assigning state of residence.

- ***Approach 2.1*:** State assignment classes 1-6. Contains the most cards of the three approaches (over 92%) and a broad mix of state assignment classes, but includes classes in which assigning state of residence based on extracted values is not recommended.
- ***Approach 2.2*:** State assignment classes 1-3. Contains a mix of state assignment classes yet includes only high-confidence cards and still covers over 80% of the population.
- ***Approach 2.3*:** State assignment class 1 only. Contains only very-high-confidence cards and thus does not include a mix of state assignment classes and excludes 35% or more of the population.

A summary of the procedure for assigning state of residence in each state assignment class, for each of the five approaches, is given below.

[13] Because of the extremely low error rates observed in extraction, keying data in class 1 would be difficult to justify. We therefore substitute the state of residence obtained from extracted values in the results presented here.

State assignment class	*Approach*				
	0	*1*	*2.1*	*2.2*	*2.3*
1	*ES=EZ*	*ES=EZ*	*ES=EZ*	*ES=EZ*	*ES=EZ*
2	*K*	*EZ5*	*EZ5*	*EZ5*	-
3	*K*	*ES*	*ES*	*ES*	-
4	*K*	*KZ*	*EZ3*	-	-
5	*K*	*KS*	*ES*	-	-
6	*K*	*KZ*	*EZ5*	-	-
7	*K*	*K*	-	-	-
8	*K*	-	-	-	-

ES=EZ: State(Zip)=State(State) - extracted state and extracted zip indicate the same state
EZ5: State(Zip5) – state from the extracted full zip code
ES: State(State) – state from the extracted state
EZ3: State(Zip3) – state from the extracted partial zip code
KS: State(keyed state) – state from the keyed state
KZ: State(keyed zip) – state from the keyed zip code, full or partial
K: keyed state or zip code, using state assignment rules of class 7 (see previous section)
-: do not attempt to assign a state of residence (undetermined).

We estimate the percentage of travellers assigned to each state of residence and the percentage for which state of residence is undetermined, under each of the five approaches. It has been shown that the number of travellers per card varies seasonally and by state of residence. We therefore estimate percentages based on the number of travellers, rather than the number of cards, using traveller count from offline production capture (see Section 2.2.2.2). We present estimated percentages based on June data for Toronto and based on July data for Montreal and Vancouver[14].

Comparison of actual percentages among approaches is complicated by the fact that the subset for which state of residence is undetermined (Undet) varies by approach - lowest for Approach 0 to highest for Approach 2.3. Because our goal is estimating state proportions, we assume that the unassigned cases represent random non-response and we redistribute them proportionally, based on assigned cases, to calculate percentage distributions. For future purposes of using the state proportions, we also estimate the number of travellers that would be assigned to each state in each approach, including undetermined, in a typical ITS sample. For comparability we express our estimates relative to an ITS sample size of 5,000 travellers[15] for each of the three airports, although it is important to note that the actual sample sizes vary by airport and by month. We note this is a random sample from the population, in contrast to the offline capture samples. We also calculate 95% confidence intervals for each state proportion. In these calculations the sample size is the number of travellers for which a state of residence is assigned, which is equal to 5,000 minus the number with state of residence undetermined.

[14] Estimates based on the other months were also calculated but are not presented.

[15] Although E311 sample sizes are typically expressed as numbers of cards, they are expressed here as numbers of travellers.

2.3.2 Results

Actual percentages for Approach 0 and percentage distributions for all five approaches were calculated for each state of residence. As it is likely only certain states would form weighting adjustment classes, we isolate six states, and combine the rest as "Other". These six states, listed below with their abbreviations, figure among the top ten states by traveller volume for all three airports. They are:

- California (CA)
- New York (NY)
- Texas (TX)
- Florida (FL)
- Washington (WA)
- Illinois (IL).

In Table D we present the percentages described above for the selected states. We also present, for four approaches, numbers of travellers (sample sizes) in an ITS sample of size 5,000 travellers, the width of 95% confidence intervals for each state and the average confidence interval width for the six states.

Table D
State proportions and other statistics for five approaches

June 2010 Toronto

	Actual %	*Percentage distribution*					*Sample size (of 5,000)*				*Width of 95% Confidence Interval*			
State	*App. 0*	*App. 0*	*App. 1*	*App. 2.1*	*App. 2.2*	*App. 2.3*	*App. 1*	*App. 2.1*	*App. 2.2*	*App. 2.3*	*App. 1*	*App. 2.1*	*App. 2.2*	*App. 2.3*
NY	9.2	9.5	9.5	8.9	9.1	9.3	459	412	378	284	1.6	1.6	1.8	2.1
CA	15.4	16.0	15.8	15.5	15.8	15.6	766	719	651	479	2.1	2.1	2.2	2.6
IL	6.7	6.9	6.9	6.8	7.3	7.3	336	315	300	224	1.4	1.4	1.6	1.8
WA	1.5	1.6	1.6	1.5	1.5	1.4	76	69	61	42	0.7	0.7	0.7	0.8
FL	8.6	8.9	8.8	8.7	8.7	9.1	427	401	361	278	1.6	1.6	1.7	2.0
TX	6.9	7.1	7.0	6.9	7.0	7.5	341	320	289	230	1.4	1.5	1.6	1.9
Other	48.3	50.0	50.4	51.8	50.6	49.8	2,446	2,401	2,091	1,523	2.8	2.9	3.0	3.5
Undet	3.4						149	362	870	1,940				
Avg.(6 states)											1.5	1.5	1.6	1.9

July 2010 Montreal

	Actual %	*Percentage distribution*					*Sample size (of 5,000)*				*Width of 95% Confidence Interval*			
State	*App. 0*	*App. 0*	*App. 1*	*App. 2.1*	*App. 2.2*	*App. 2.3*	*App. 1*	*App. 2.1*	*App. 2.2*	*App. 2.3*	*App. 1*	*App. 2.1*	*App. 2.2*	*App. 2.3*
NY	10.5	10.9	10.8	10.2	10.6	10.9	524	476	445	349	1.7	1.7	1.9	2.2
CA	16.0	16.5	16.3	16.1	16.0	16.1	797	754	673	516	2.1	2.1	2.2	2.5
IL	5.3	5.5	5.5	5.5	5.8	5.7	270	258	244	183	1.3	1.3	1.4	1.6
WA	1.3	1.3	1.3	1.3	1.3	1.2	63	59	53	40	0.6	0.6	0.7	0.8
FL	10.2	10.5	10.4	10.1	10.3	10.7	508	474	433	344	1.7	1.7	1.8	2.1

TX	*7.0*	*7.2*	*7.2*	*7.1*	*7.3*	*7.9*	*350*	*332*	*306*	*252*	*1.4*	*1.5*	*1.6*	*1.9*
Other	*46.6*	*48.1*	*48.5*	*49.7*	*48.7*	*47.4*	*2,365*	*2,323*	*2,043*	*1,522*	*2.8*	*2.9*	*3.0*	*3.5*
Undet	*3.1*						*123*	*323*	*804*	*1,793*				
Avg.(6 states)											*1.5*	*1.5*	*1.6*	*1.8*

July 2010 Vancouver

	Actual %	***Percentage distribution***					***Sample size (of 5,000)***				***Width of 95% Confidence Interval***			
State	***App. 0***	***App. 0***	***App. 1***	***App. 2.1***	***App. 2.2***	***App. 2.3***	***App. 1***	***App. 2.1***	***App. 2.2***	***App. 2.3***	***App. 1***	***App. 2.1***	***App. 2.2***	***App. 2.3***
NY	*5.0*	*5.2*	*5.1*	*5.1*	*5.1*	*5.2*	*251*	*244*	*216*	*171*	*1.2*	*1.3*	*1.3*	*1.5*
CA	*28.2*	*29.0*	*28.8*	*28.0*	*28.6*	*29.6*	*1,402*	*1,328*	*1,200*	*978*	*2.5*	*2.6*	*2.7*	*3.1*
IL	*3.6*	*3.7*	*3.7*	*3.6*	*3.8*	*3.9*	*180*	*172*	*161*	*128*	*1.1*	*1.1*	*1.2*	*1.3*
WA	*4.6*	*4.8*	*4.8*	*4.7*	*4.7*	*4.7*	*232*	*221*	*197*	*155*	*1.2*	*1.2*	*1.3*	*1.4*
FL	*4.3*	*4.5*	*4.4*	*4.3*	*4.5*	*4.7*	*216*	*206*	*190*	*157*	*1.2*	*1.2*	*1.3*	*1.4*
TX	*8.4*	*8.7*	*8.7*	*8.6*	*8.8*	*9.1*	*422*	*406*	*370*	*302*	*1.6*	*1.6*	*1.7*	*2.0*
Other	*42.8*	*44.1*	*44.5*	*45.7*	*44.4*	*42.9*	*2,169*	*2,169*	*1,865*	*1,418*	*2.8*	*2.8*	*3.0*	*3.4*
Undet	*3.0*						*128*	*254*	*802*	*1,692*				
Avg.(6 states)											*1.5*	*1.5*	*1.6*	*1.8*

State proportions and other statistics were also calculated for groups of states. These groups correspond to US regions defined by the Canadian Tourism Commission (CTC) and used by the ITS. See Appendix D for the CTC assignment of states to US regions.

2.3.3 Observations

Analysis of estimated state proportions based on offline production capture data sheds some light on differences among the approaches, but is not definitive. Additional observations and other analytical techniques are needed, especially those that show behaviour over time. Some specific observations include:

- State proportions of all five approaches are within two percentage points, with few exceptions.
- For most states the value from Approach 0 was contained in the 95% confidence interval of all other approaches.
- Approach 1 had the fewest undetermined (Undet) cards and produced state proportions that are closest to those from Approach 0.
- The study results do not point to a clear choice between Approaches 2.1 and 2.2. The proportions of the six states are generally within half a percentage point. Although Approach 2.1 has fewer undetermined cards, it has the highest percentage of "Other" states.
- The results identify some potential risks associated with Approach 2.3.
    - There was a considerable portion of cards with state undetermined - 39% in Toronto, 36% in Montreal and 34% in Vancouver.
    - There were considerably more cards with state undetermined, and thus wider confidence intervals for the same ITS volume sample size, as compared to the two other approaches.

    - Approach 2.3, with the greatest number of undetermined cards, is particularly vulnerable to smaller sample sizes that exclude a number of smaller states. Such exclusion leads to overestimation of state proportions of larger states.
- California is the top state for all three airports, although it is much more dominant in Vancouver. State proportions are similar for Montreal and Toronto, two airports in close geographic proximity, save a higher contribution from Florida in Montreal. However differences in traveller volumes and patterns between June and July, before and during the summer vacation season, may have contributed to this difference.
- Many results were similar for all three airports.
    - The percentage of cards in "Other" states was between 40% and 50%. It was slightly smaller for Vancouver, possibly reflecting the inclusion of dominant states Washington and California among its six states.
    - The percentage of undetermined cards was around 3% in Approach 0.
    - The average confidence interval widths are essentially identical.

We suspect, for each approach, state proportions and state rankings would vary by season and would be influenced by exogenous factors. With only two months of data per airport, we cannot analyze state proportions over time and cannot observe the impact of the holiday season and special events like the Vancouver Olympics. In particular we do not see behaviour in low-volume months and cannot analyze quarterly proportions.

In order to study behaviour over time, our next step was to generate and study a full year of state proportion estimates. However, as will be described in Section 2.4, replication of the offline production data study procedures was not possible, for cost and operational reasons.

2.4 Operational considerations

The operational implications and cost of implementing the approaches above were discussed with ITS operations personnel. It was confirmed that the technical requirements of any of the five approaches could be implemented within ITS production. Regarding cost, two important points were noted:

- If the current ITS sample is adequate (i.e., no increase in sample size is needed), development and set-up costs of procedures for keying US address information would be much greater, on a per-card basis, than the operating cost of keying US address information.
- If an increase to the ITS sample size were required, costs beyond keying US address fields would be incurred for the additional cards. That is, additional fields would be necessarily captured even though only US address information might be required.

Although it was desired to study state proportions based on a full year of data, it was decided, based on the results from offline production data and a lack of resources, that further study would examine only approaches without keyed data and that no additional capture of US address data would be done. This decision is also consistent with the realization that the ITS would not have the resources to implement keying of state or zip code in the ITS production system. We refer to these as simulated ITS production state proportions, as they use data available in current operations but they are not part of current E311 processing.

3. SIMULATED ITS PRODUCTION STATE PROPORTIONS

As indicated earlier, study of behaviour over time is necessary to uncover seasonal or other factors not revealed by analyzing samples of a limited number of months. As no additional keyed US address data was available, alternatives to the methods of the previous section are needed. In this section we propose three approaches, analogous to Section 2, for estimating state proportions based on extracted data available in current ITS production. The three approaches, and the modifications to implement them in the ITS production system rather than based on offline production data, are described in Section 3.1. For each approach, we study behaviour over time by producing a full year (12 consecutive months) of estimated state proportions. Analysis of these estimates is discussed in Sections 3.2 and 3.3. Without keyed data as "true" values, error rates cannot be calculated and a modified analysis framework is required, as compared to offline production data. In addition to studying behaviour over a full year, these simulations will help evaluate further the adequacy of the current ITS sample for estimating state proportions.

3.1 Approaches

The three approaches for which we will simulate ITS production state proportions use the same state assignment rules as Approaches 2.1, 2.2 and 2.3 of the previous section (see Section 2.3.1). For clarity we renamed them to Approaches A, B and C, and also reordered them to go from fewest to most cards with a state of residence assigned, as the definitions are cumulative (i.e., states assigned by A are included in B and similarly those of B in C). The naming convention and specification of the state assignment rules for the three approaches are given below.

State assignment class		***Approach***	
	A	***B***	***C***
1	*ES=EZ*	*ES=EZ*	*ES=EZ*
2	-	*EZ5*	*EZ5*
3	-	*ES*	*ES*
4	-	-	*EZ3*
5	-	-	*ES*
6	-	-	*EZ5*

ES=EZ: State(Zip)=State(State) - extracted state and extracted zip indicate the same state
EZ5: State(Zip5) – state from the extracted full zip code
ES: State(State) – state from the extracted state
EZ3: State(Zip3) – state from the extracted partial zip code
-:do not assign state of residence (undetermined).

In order to calculate estimates based on the number of travellers on each card (see Section 2.3.1), it was decided to use the ITS volume sample, rather than the entire E311 universe, for simulated ITS production state proportions to study behaviour over time. The primary purpose of the volume sample is estimating frontier counts. Consequently, the traveller count is verified by a keyer as part of ITS processing operations of cards selected for the volume sample. In contrast, the only value of traveller count available for cards not selected for the volume sample is based on automated methods alone. Study has shown this value of traveller count is unreliable, in part due to errant markings on E311 cards, as mentioned in Section 2.2.1. An additional benefit is the scrutiny given to the ITS volume sample, in particular in detecting problems with CBSA submission or completion of E311 cards.

Our study used the most recent 12-month period with final E311 data, which was October 2009 to September 2010. Although we do not have a calendar year of data, we do have complete quarters. Having a full year of data allows us to see holiday and other seasonal trends and to observe the month-to-month impact of significant events like the Vancouver Olympics.

For each approach, estimates of monthly proportions by state and CTC US region were produced for the three airports. In addition to the estimates using the ITS volume sample and based on traveller count, two other sets were produced. The first was based on the ITS volume sample but used card count, rather than traveller count, to estimate state proportions. The second used the entire US E311 universe file (not just the ITS volume sample), for which only card-count based estimates were tabulated due to the unreliability of its traveller count. Analysis of these three sets confirmed the hypothesis that different estimates of state proportions result if traveller count is not used. In particular, overestimation for states with large percentages of business travellers, such as New York and Washington, DC, results if estimates are based on card count[16]. These observations confirm the importance of using the ITS volume sample rather than the entire US E311 universe file.

The statistics calculated are similar to those of offline production capture data, although with three rather than five approaches. We estimate state proportions and numbers of travellers that would be assigned to each state of residence (including undetermined) in an ITS sample of size 5,000 travellers, and calculate 95% confidence intervals. Analysis confirmed some observations from offline production capture.

- The percentage with state of residence undetermined is 30% to 40% with Approach A, 15% to 20% with Approach B and approximately 5% with Approach C.
- With few exceptions, state proportions are within two percentage points of each other.
- The state proportions produced by Approaches B and C are generally very close and Approach A displays greater risk by yielding state proportions that are sometimes separated from the others.

[16] Analysis of E311 data has shown that i) the average traveller count per card is lower and ii) the percentage of single-traveller cards is higher for business travellers as compared to non-business travellers.

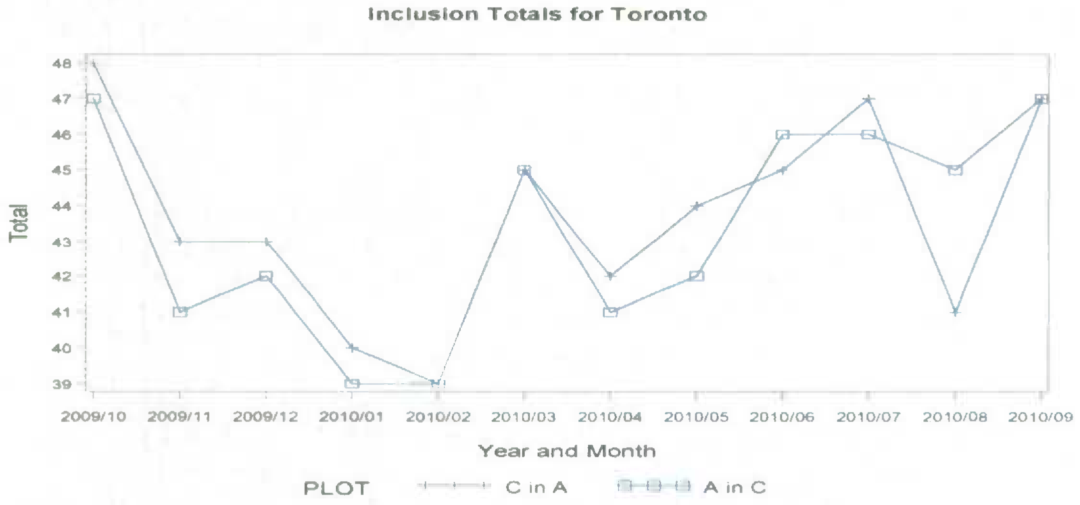
With a full year of data available for each airport, rather than two months, analysis beyond that of offline production capture data is possible. In particular time series analyses of confidence intervals and point estimates are described in the following two sections.

3.2 Confidence interval analysis

To evaluate the quality of results obtained under the three approaches and to differentiate among them, we calculated confidence intervals for estimates of the 51 state proportions under each of the three approaches. The width of the confidence interval is a function of both the estimated state proportion and the sample size. Smaller estimated proportions[17] and larger sample sizes lead to narrower confidence intervals. As the sample size is the number of travellers with a valid state of residence[18], on average, we expect Approach A to have the widest confidence intervals of the three approaches, and Approach C the narrowest. However, higher estimated proportions may counter the impact of a larger sample size and lead to wider confidence intervals. To quantify both the difference between approaches and the impact of undetermined cards, we computed two inclusion totals.

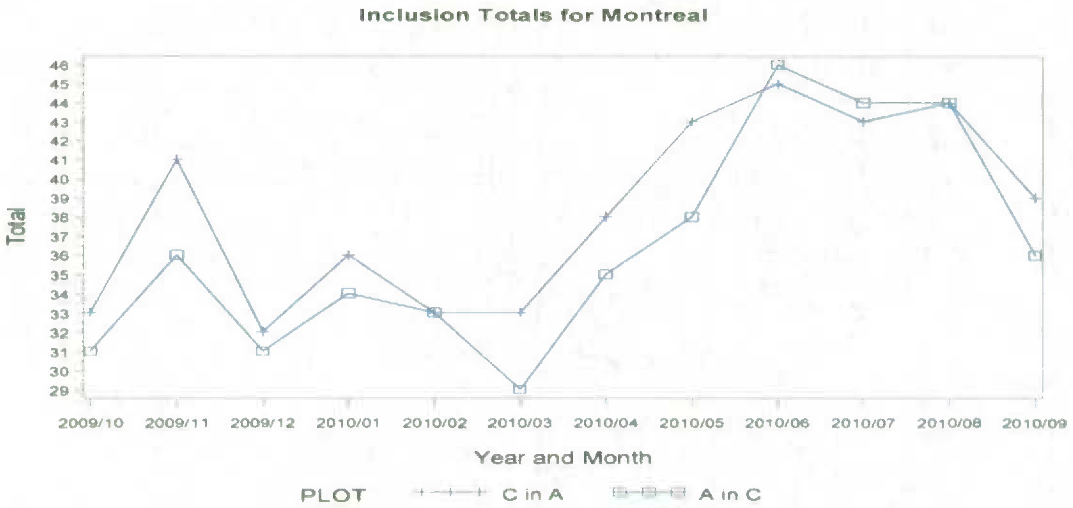
- ***C In A***: The number of states for which the 95% confidence interval of Approach A contains the estimate of Approach C.
- ***A In C***: The number of states for which the 95% confidence interval of Approach C contains the estimate of Approach A.

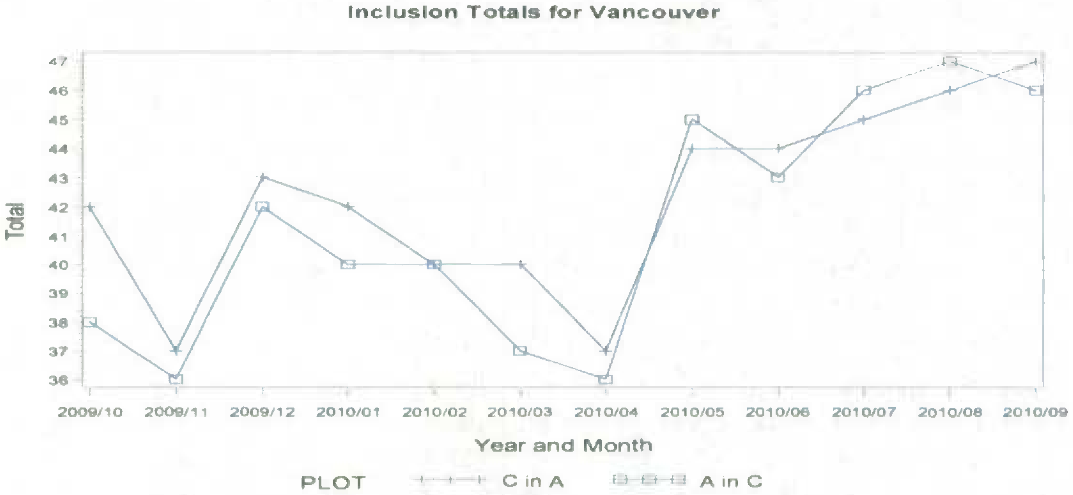
As the number of cards for which state of residence is undetermined is greatest with Approach A and least with Approach C, we expect the first inclusion total to be larger than the second. The two inclusion totals are plotted below, for each of the three airports.



[17] This assertion is true for estimated proportions below 0.5, and thus for all proportions observed in the study. For estimated proportions greater than 0.5, larger estimated proportions lead to narrower confidence intervals.

[18] The number with a valid state of residence is calculated relative to an ITS sample of 5,000 travellers.



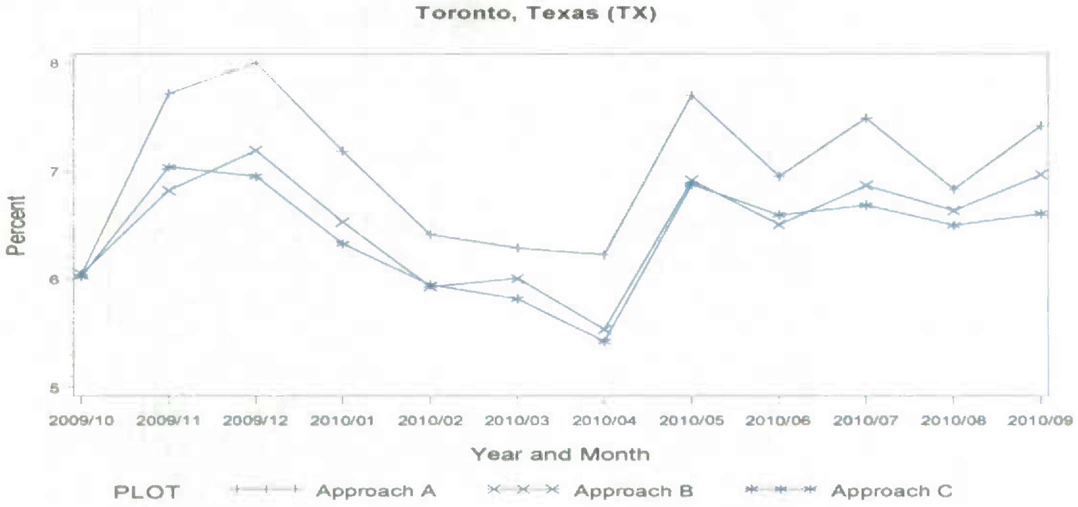


More undetermined cards and thus wider confidence intervals in Approach A lead to *C In A* generally above *A In C*. This observation conveys the greater risk associated with Approach A, in which estimates of state proportions are based on the smallest number of cards. Although the inter-dependence of the two inclusion totals means they move up and down together, the range in values over time is somewhat surprising. We first looked at the percentage of undetermined cards, and found it to remain relatively stable over time in both approaches. Additional study revealed that high inclusion totals are generally associated with months with larger ITS volume sample sizes. That is, when the state proportions are based on more cards, there is more chance of estimates based on two approaches being close together. The presence of months when both inclusion totals are low (below 40 states) is of concern and conveys risk in estimating state proportions under all approaches with smaller ITS sample sizes.

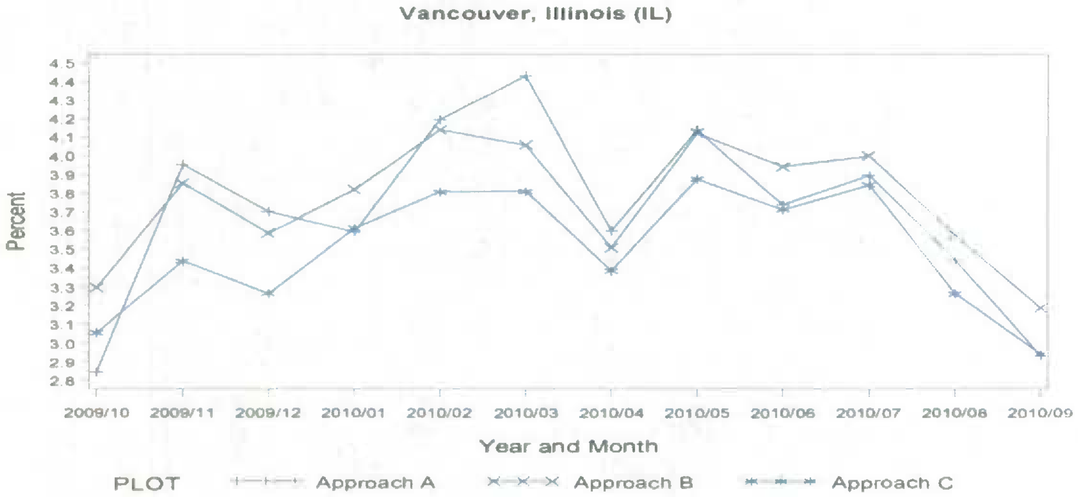
3.3 Time series analysis of estimated state proportions

As with previous analysis, for discussion we isolate six states and group the balance as "Other". We graph the estimated state proportions over 12 months for each of the three approaches. A full set of time series for the six states is included in Appendix E, and we present some highlights here.

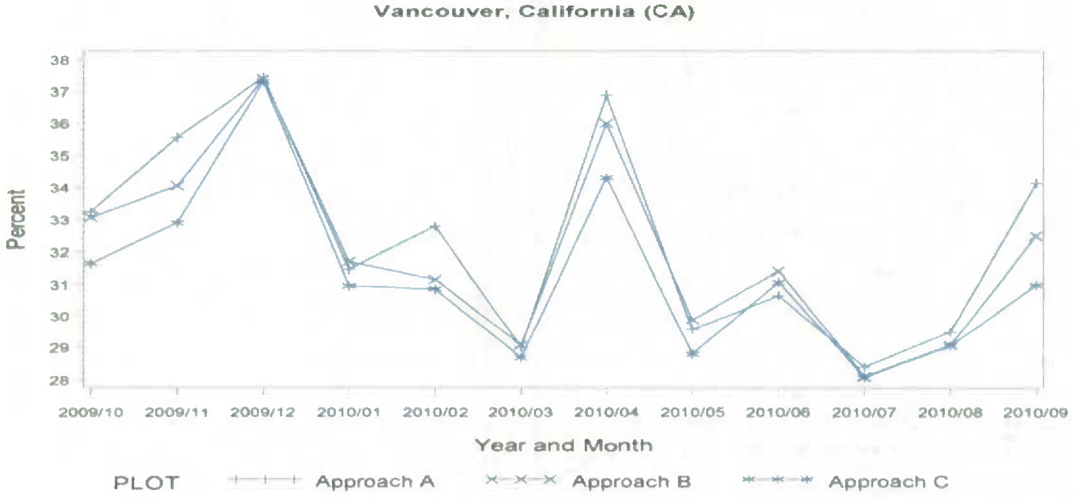
Analysis of the time series of estimated state proportions confirms our earlier concerns about the riskiness of Approach A. Many graphs revealed that the state proportions of Approaches B and C are close to each other, with Approach A somewhat separated. In the example below, we see the estimated percentage of Texas travellers entering at Toronto is consistently higher with Approach A.

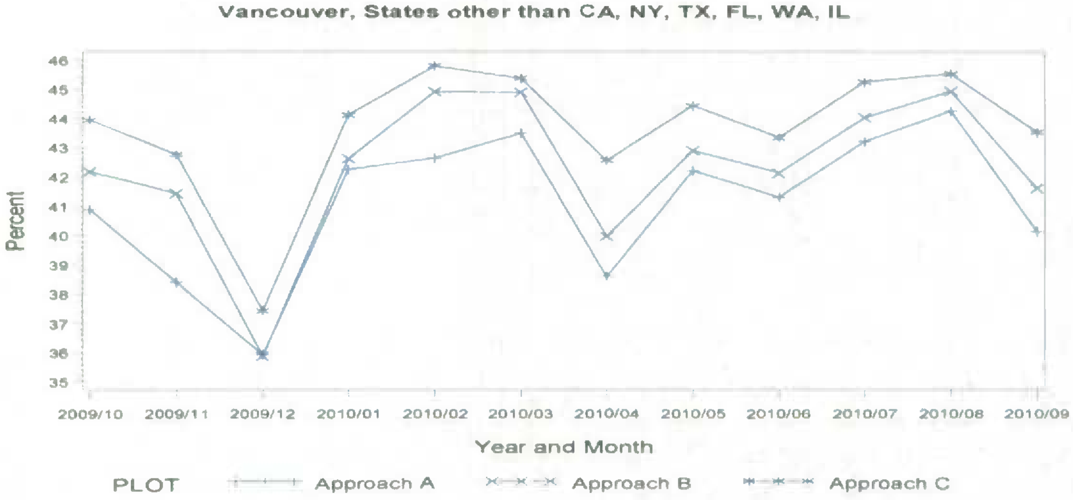


In other states, the graphs of Approaches B and C move in parallel, with Approach A following a slightly different path over time. In the example of Illinois travellers entering at Vancouver, shown below, the estimate of Approach A lies outside of the range of the estimates of Approaches B and C for the first seven months, and between for the final five months. Also, for seven months the estimated percentage of Illinois travellers of Approach A is closer to that of Approach B, and is closer to that of Approach C for the other five months.



Analysis also reveals months in which we have unexplained results. An example is the spike (by about seven percentage points using Approach A) in the percentage of California travellers to Vancouver in April 2010 and the accompanying drop (by about five percentage points) in the percentage of travellers from "Other" states. No exogenous factors that would lead to the spike were identified by subject matter experts, and no similar spikes of California travellers were observed at the other two airports. Investigation revealed that the number of California travellers to Vancouver increased by about 2,000 between March and April 2010, although the overall number of US travellers to Vancouver dropped by about 2,000. With only one year of data it was not possible to determine if this behaviour repeats annually. We also note this spike was most prominent with Approach A, which has the greatest number of cards with state of residence undetermined.

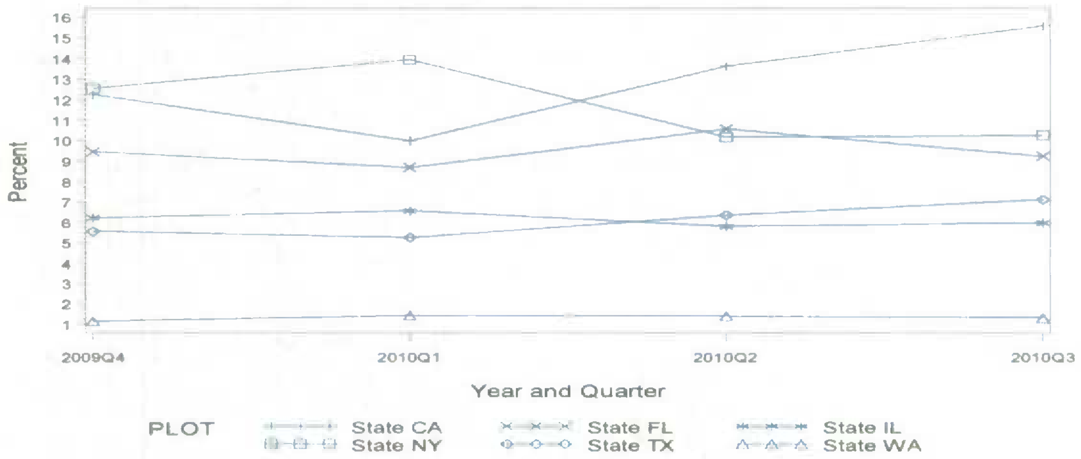




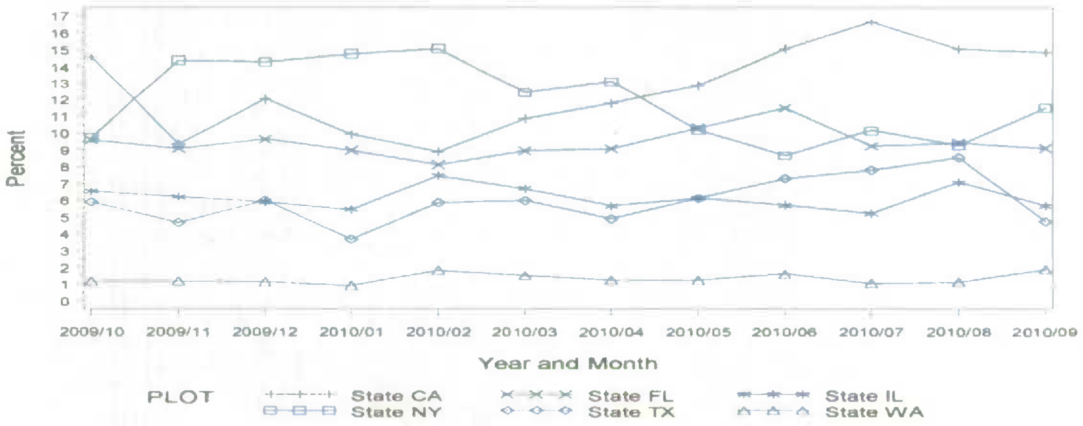
Although state proportions will be produced on a monthly basis in ITS E311 operations, they will be required with a quarterly frame of reference for ITS weighting adjustment.

Simulated ITS production quarterly state proportions were also produced. These absorb some of the up-and-down variation within quarters, leading to smoother time series often within a lesser range of values. Quarterly time series may also uncover trends not apparent by looking at monthly values. Graphs of quarterly and monthly time series of state proportions are given below, using Approach C and travellers entering at Montreal. The Texas quarterly proportions were within a single percentage point, while the range of the monthly proportions was about five percentage points.

Montreal, Approach C, quarterly



Montreal, Approach C, Monthly



4. CONCLUSIONS

After background research on US address concepts and evaluation of E311 data, the study recommends a strategy for assigning state of residence based on the E311 fields Province/State and Postal/Zip. Pre-processing of the Province/State data is necessary, in particular identifying standard abbreviations and common short forms used by travellers. Background research on the relationship between zip code and state revealed that, save a few exceptions, state can be determined from zip code. As well, the study identified when a partial zip code is adequate to determine state. Since a current, comprehensive state-zip code concordance file was not available, the study selected a file produced by the US Census Bureau for the 2000 Census. Other fields, including Town/City, could also be used to determine state of residence. However they were excluded from the study, primarily due to the absence of a concordance file.

Consultation with ITS subject matter and operations personnel confirmed keying of US address information is feasible in ITS E311 processing. Offline keying of new samples revealed that data of sufficient quality could be obtained. However integration of keying US address data into ITS E311 processing would involve considerable set-up costs. If cards beyond the current ITS sample were required, keying fields in addition to US address components would be necessary.

US address data obtained in extraction in E311 processing could be used to improve efficiency in estimating state proportions, although both extracted and keyed data are required in the recommended procedure. The study revealed that, for over 60% of US E311 cards, extracted values of Province/State and Postal/Zip could be used to indicate a state of residence with very small error rates. However comparison to keyed data revealed risks in estimating state proportions from this subpopulation alone. The study also identified a further 15% to 20% of the E31 population for which extracted values indicate a state of residence with acceptable error rates, and that the resulting estimates of state proportions are closer to those based on keyed data. The study recommends a procedure that uses extracted values for these cards and keyed values when extracted values do not indicate a state of residence. It also identifies a small sub-population (less than 1%) where keying is not recommended.

Simulated ITS production state proportions revealed that the current ITS volume sample would be adequate for estimating state proportions. Confidence intervals of less than two percentage points generally resulted. However the study revealed that over-estimation of larger states may result if small sample sizes lead to smaller states not being represented.

5. RECOMMENDATIONS AND FUTURE WORK

Continue to encourage the ITS to capture traveller address data for US E311 cards, in particular the State and Zip fields. If keying of US address information is not included in regular ITS E311 operations, periodically draw samples for which US address data would be keyed. Such data would be used to monitor the quality of extracted data and validate state proportions obtained from them, as well as to update and improve state assignment rules.

Continue to calculate state proportions. The study included only one year of data, and therefore could not quantify year-over-year trends. The 12-month period of the study also included significant events like the Vancouver Olympics and other observations that might not occur annually, such as the April 2010 spike in the percentage of California residents entering at Vancouver. Analysis over time would permit development of methods for outlier detection and would detect changes in the state mix. Looking at additional data would also permit finalizing the recommended methodology for estimating state proportions.

Fine-tune state assignment rules. Multiple sets of rules may exist, such as a stringent set for assigning state of residence and a more relaxed set when proposing a value to a keyer.

Formulate a methodology for calculating counts of US air travellers by state of residence, and integrate into ITS E311 processing operations.

Develop a revised strategy for weighting of ITS US air questionnaires. A methodology for adjustment by state of residence should be included. Important aspects include:

- In conjunction with ITS, identify a set of "targeted states" for each airport. These targeted states will be used to define weighting adjustment classes.
- Integrate state adjustment with other weighting adjustment, including adjustment by duration and purpose.
- Study methods of weighting adjustment. Exact adjustment or approximation methods are possibilities, depending on the number of weighting adjustment categories, number of questionnaires and quality of state benchmarks.

Refinements to the state assignment methodology should be studied. Examples include obtaining an updated state-zip code concordance file (possibly from the 2010 US Census) and a city-zip code concordance file.

Explore obtaining concordance information at the USBC-Statistics Canada Interchange and other interaction with American colleagues.

The software AnyDoc, used for ITS data capture, includes a module AccuZip for processing of US address data. The module would replace some of the pre-processing employed in the study methodology, and might provide a more up-to-date source of state-zip code concordance. We recommend that AccuZip be studied and that the ITS consider its purchase.

Other uses of E311 US address data should be identified and studied. One example is using valid State and Zip data to identify cards not classified as US travellers in ITS E311 processing. Another is looking for Canadian postal codes or other geographic indicators in the State and Zip fields, to identify Canadian or overseas cards classified as US.

Acknowledgements

The author would like to note the contribution of the colleagues indicated below to the completion of the study and preparation of this report.

Alex Diaz-Papkovich – programming of state assignment rules and framework for calculating error rates; analysis of desktop capture and offline production capture data.

Mark Stinner – analysis of estimated state proportions based on ITS production data; production of graphs.

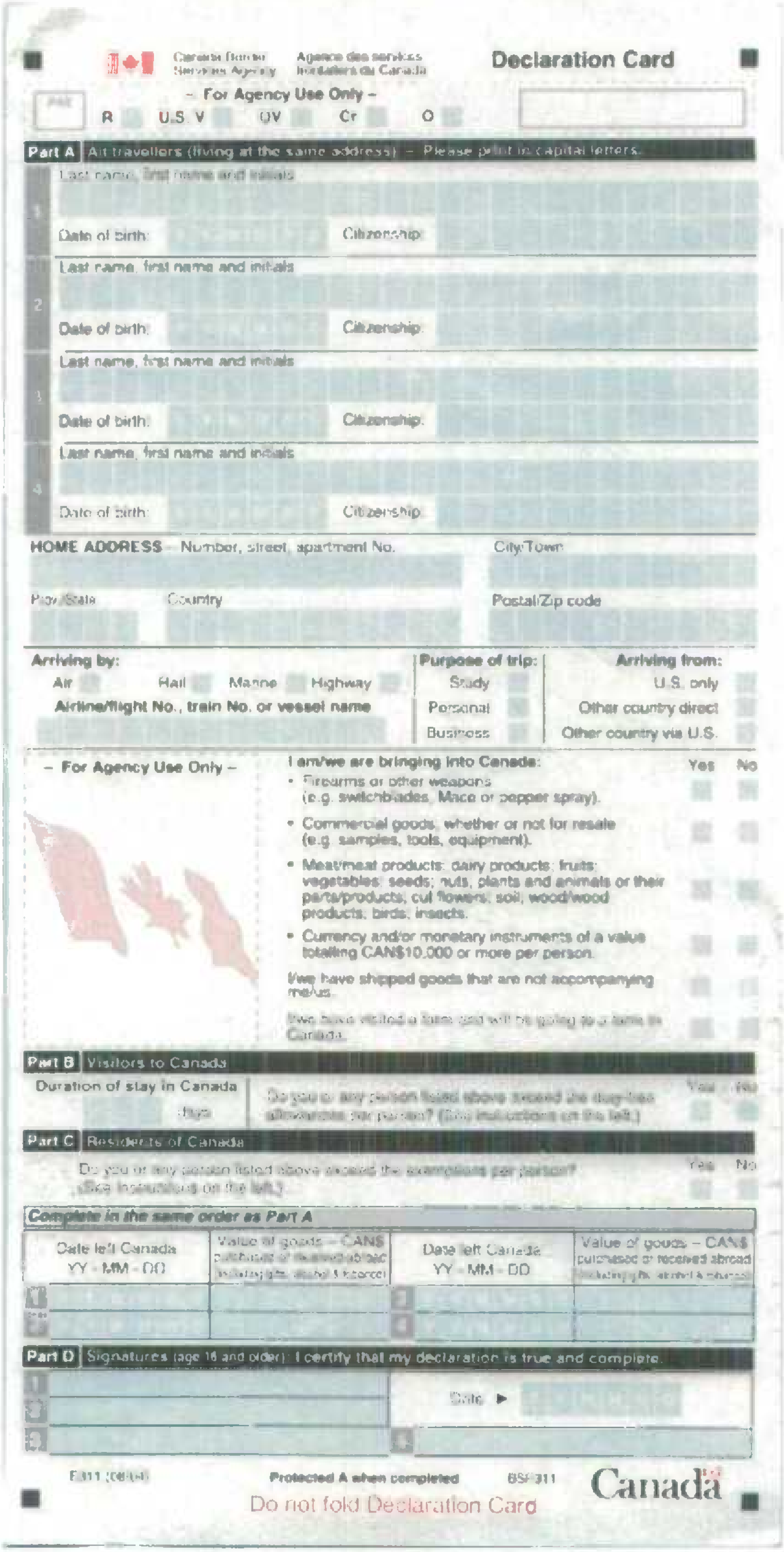
Simon Cheung and Sylvain Perron – methodological consultation; review of report; expertise regarding the implementation of methods in the ITS production system.

Joseph Duggan – peer review of report for submission as a working paper.

APPENDICES

APPENDIX A

CBSA Declaration Card (E311)



Canada Border Services Agency　Agence des services frontaliers du Canada

Declaration Card

– For Agency Use Only –

R □　U.S. V □　UV □　Cr □　O □

Part A Air travellers (living at the same address) – Please print in capital letters.

1 Last name, first name and initials

Date of birth:　Citizenship:

2 Last name, first name and initials

Date of birth:　Citizenship:

3 Last name, first name and initials

Date of birth:　Citizenship:

4 Last name, first name and initials

Date of birth:　Citizenship:

HOME ADDRESS – Number, street, apartment No.　City/Town

Prov./State　Country　Postal/Zip code

Arriving by:

Air □　Rail □　Marine □　Highway □

Airline/flight No., train No. or vessel name

Purpose of trip:

Study □

Personal □

Business □

Arriving from:

U.S. only □

Other country direct □

Other country via U.S. □

– For Agency Use Only –

		Yes	No
I am/we are bringing into Canada:			
• Firearms or other weapons (e.g. switchblades, Mace or pepper spray).		□	□
• Commercial goods, whether or not for resale (e.g. samples, tools, equipment).		□	□
• Meat/meat products; dairy products; fruits; vegetables; seeds; nuts; plants and animals or their parts/products; cut flowers; soil; wood/wood products; birds; insects.		□	□
• Currency and/or monetary instruments of a value totalling CAN$10,000 or more per person.		□	□
I/we have shipped goods that are not accompanying me/us.		□	□
I/we have visited a farm and will be going to a farm in Canada.		□	□

Part B Visitors to Canada

Duration of stay in Canada　days

Do you or any person listed above exceed the duty-free allowances per person? (See instructions on the left.)　Yes □　No □

Part C Residents of Canada

Do you or any person listed above exceed the exemptions per person? (See instructions on the left.)　Yes □　No □

Complete in the same order as Part A

	Date left Canada YY - MM - DD	Value of goods – CAN$ purchased or received abroad (including gifts, alcohol & tobacco)		Date left Canada YY - MM - DD	Value of goods – CAN$ purchased or received abroad (including gifts, alcohol & tobacco)
1			3		
2			4		

Part D Signatures (age 16 and older): I certify that my declaration is true and complete.

1

2　Date ►

3　4

E311 (08/04)　**Protected A when completed**　BSF311　Canada

Do not fold Declaration Card

APPENDIX B1

Full (five-digit) zip codes that correspond to multiple states

Zip (full)	*Primary State*	*Secondary State*	*Zip (full)*	*Primary State*	*Secondary State(s)*
10004	*NY*	*NJ*	*71749*	*AR*	*LA*
37642	*TN*	*VA*	*72644*	*AR*	*MO*
38041	*TN*	*AR*	*73949*	*OK*	*TX*
38063	*TN*	*AR*	*79922*	*TX*	*NM*
38079	*TN*	*KY*	*82063*	*WY*	*CO*
38852	*MS*	*AL*	*82082*	*WY*	*NE*
42223	*KY*	*TN*	*83120*	*WY*	*ID*
51630	*IA*	*MO*	*84536*	*UT*	*AZ*
51640	*IA*	*MO*	*85534*	*AZ*	*NM*
52626	*IA*	*MO*	*86044*	*AZ*	*UT*
52761	*IA*	*IL*	*86504*	*AZ*	*NM*
55954	*MN*	*IA*	*86508*	*AZ*	*NM*
56027	*MN*	*IA*	*86514*	*AZ*	*NM, UT*
56129	*MN*	*IA*	*86515*	*AZ*	*NM*
56219	*MN*	*SD*	*87328*	*NM*	*AZ*
57638	*SD*	*ND*	*89439*	*NV*	*CA*
59221	*MT*	*ND*	*97635*	*CA*	*OR*
65733	*MO*	*AR*	*97910*	*OR*	*ID*
65761	*MO*	*AR*	*97913*	*OR*	*ID*
67950	*KS*	*OK*	*99128*	*WA*	*ID*
69337	*NE*	*SD*	*99362*	*WA*	*OR*

Primary State: included on the concordance file StateZip5a.

APPENDIX B2

Zip code prefixes (three-digit) that include multiple states

Zip	*Primary*	*Secondary*	*Zip*	*Primary*	*Secondary*
(partial)	*State*	*State(s)*	*(partial)*	*State*	*State(s)*
063	*CT*	*NY*	*717*	*AR*	*LA*
100	*NY*	*NJ*	*726*	*AR*	*MO*
376	*TN*	*VA*	*739*	*OK*	*TX*
380	*TN*	*AR, KY*	*790*	*TX*	*OK*
388	*MS*	*AL*	*799*	*TX*	*NM*
422	*KY*	*TN*	*820*	*WY*	*CO, NE*
516	*IA*	*MO*	*831*	*WY*	*ID*
526	*IA*	*MO*	*845*	*UT*	*AZ*
527	*IA*	*IL*	*855*	*AZ*	*NM*
559	*MN*	*IA*	*860*	*AZ*	*UT*
560	*MN*	*IA*	*865*	*AZ*	*NM, UT*
561	*MN*	*IA*	*873*	*NM*	*AZ*
562	*MN*	*SD*	*884*	*NM*	*TX*
576	*SD*	*ND*	*890*	*NV*	*AZ*
582	*ND*	*MN*	*894*	*NV*	*CA*
592	*MT*	*ND*	*976*	*OR*	*CA*
657	*MO*	*AR*	*979*	*OR*	*ID*
679	*KS*	*OK*	*991*	*WA*	*ID*
693	*NE*	*SD*	*993*	*WA*	*OR*

Primary State: included on the concordance file StateZip3a.

APPENDIX C1

Pre-Processing of Province/State

Province/State	*State Assigned*	*State Name*
ALASKA	*AK*	*Alaska*
ARIZONA	*AZ*	*Arizona*
CALIF	*CA*	*California*
s/w CONN	*CT*	*Connecticut*
COLO	*CO*	*Colorado*
s/w DIST	*DC*	*Washington D.C.*
FLA,FI	*FL*	*Florida*
s/w GEOR	*GA*	*Georgia*
s/w HAWA	*HI*	*Hawaii*
ILL	*IL*	*Illinois*
IOWA	*IA*	*Iowa*
s/w KANS	*KS*	*Kansas*
s/w KENT	*KY*	*Kentucky*
s/w LOUI	*LA*	*Louisiana*
s/w MARY	*MD*	*Maryland*
MASS	*MA*	*Massachusetts*
MAINE	*ME*	*Maine*
MICH	*MI*	*Michigan*
s/w MINN	*MN*	*Minnesota*
s/w MISSISS	*MS*	*Mississippi*
MISSOURI	*MO*	*Missouri*
NEBR	*NE*	*Nebraska*
NEVADA	*NV*	*Nevada*
s/w NEW JER, N J	*NJ*	*New Jersey*
s/w NEW MEX	*NM*	*New Mexico*
NEW YORK, s/w N Y	*NY*	*New York*
NORTH CAROLINA	*NC*	*North Carolina*
NORTH DAKOTA	*ND*	*North Dakota*
OHIO	*OH*	*Ohio*
OKLA	*OK*	*Oklahoma*
s/w PENN	*PA*	*Pennsylvania*
s/w RHOD	*RI*	*Rhode Island*
s/w SOUTH CARO	*SC*	*South Carolina*
s/w TENN	*TN*	*Tennessee*
TEXAS	*TX*	*Texas*
UTAH	*UT*	*Utah*
s/w VERM	*VT*	*Vermont*
s/w VIRG	*VA*	*Virginia*
WASH	*WA*	*Washington*
WISC	*WI*	*Wisconsin*
s/w WEST VIR	*WV*	*West Virginia*
Otherwise	*Province/State*	

s/w: starts with

APPENDIX C2

Additional rules for pre-processing of Province/State and Postal/Zip

Province/State	*Postal/Zip*	*State Assigned*	*State Name*
s/w ARIZ	-	*AZ*	*Arizona*
s/w CAL	-	*CA*	*California*
s/w EL, s/w RL	*s/w 3*	*FL*	*Florida*
LL, II	*s/w 6*	*IL*	*Illinois*
OMO	-	*OH*	*Ohio*
IX	-	*TX*	*Texas*
NY	*First char I;L;/; or ,*	*NY*	*New York*
NY	*Second char O;I;L;,;/;or Z*	*NY*	*New York*
NY	*Third char O;I;L;,;/; or Z*	*NY*	*New York*
MN	*s/w SS*	*MN*	*Minnesota*

s/w: starts with

APPENDIX C3

Exceptions to state assignment rules for conflicts between Zip and State

a) State(State) and State(Zip5) indicate different states of residence.
State Assigned=State(Zip5) except for:

State(State)	***State(Zip5)***	***State Assigned***	***State Name***
CA	*OH*	*CA*	*California*
CA	*KY*	*CA*	*California*
TX	*PA*	*TX*	*Texas*
NJ	*MA*	*NJ*	*New Jersey*
FL	*TN*	*FL*	*Florida*
PA	*NY*	*PA*	*Pennsylvania*

b) State(State) and State(Zip3) indicate different states of residence.
State Assigned=State(State) except for:

State(State)	***State(Zip3)***	***State Assigned***	***State Name***
AL	*AZ*	*AZ*	*Arizona*
IL	*FL*	*FL*	*Florida*
FL	*IL*	*IL*	*Illinois*
NE	*NC*	*NC*	*North Carolina*
LA	*CA*	*CA*	*California*
MD	*MO*	*MO*	*Missouri*
MO	*MD*	*MD*	*Maryland*
NV	*NY*	*NY*	*New York*
PA	*CA*	*CA*	*California*

APPENDIX D

Canadian Tourism Commission assignment of states to US regions

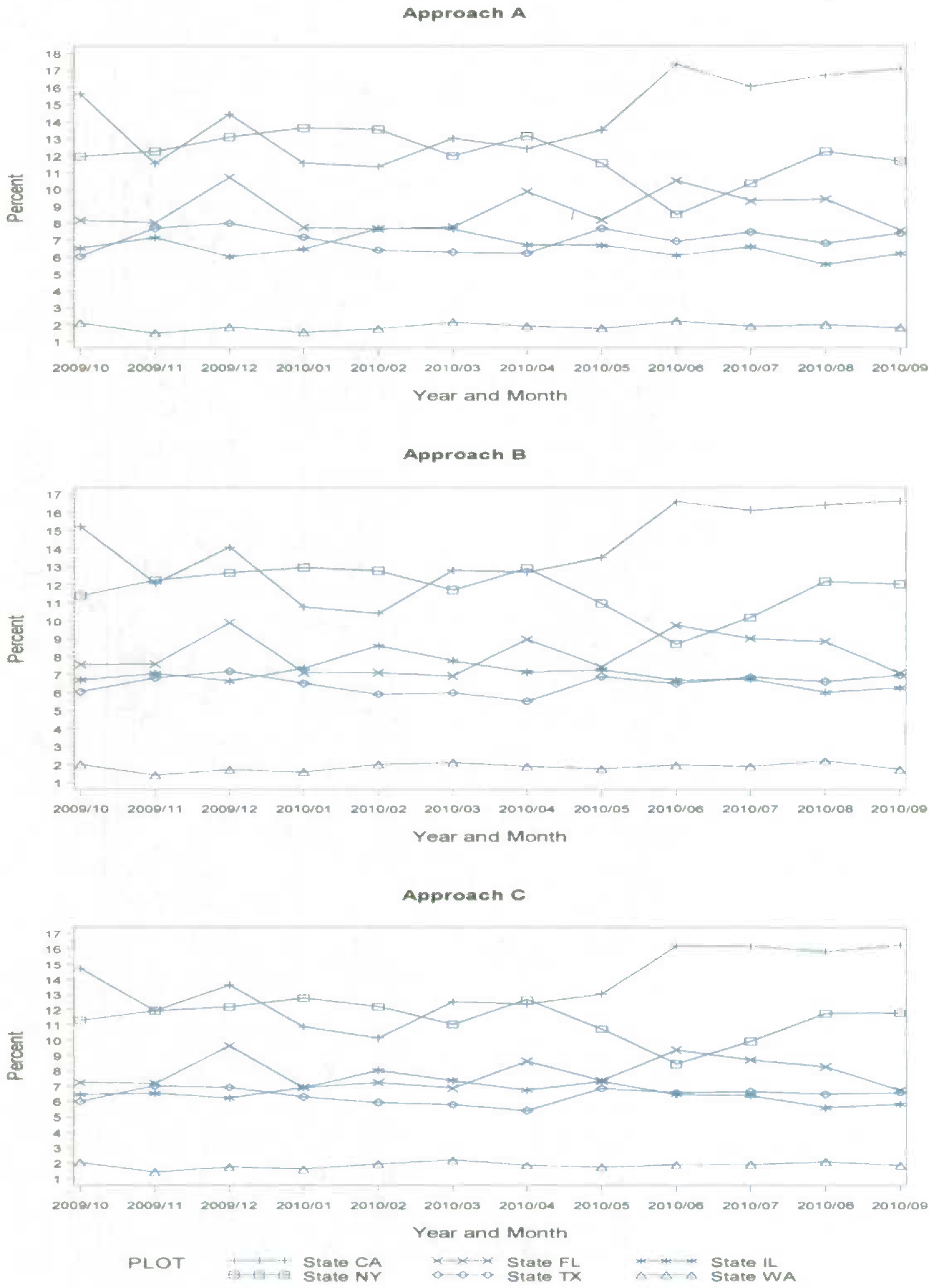
CTC US Region	*State Name*	*AB*	*CTC US Region*	*State Name*	*AB*
North East	*Maine*	*ME*	*Central*	*Illinois*	*IL*
North East	*New Hampshire*	*NH*	*Central*	*Iowa*	*IA*
North East	*Vermont*	*VT*	*Central*	*Kansas*	*KS*
North East	*New York*	*NY*	*Central*	*Nebraska*	*NE*
North East	*Pennsylvania*	*PA*	*Central*	*South Dakota*	*SD*
North East	*Ohio*	*OH*	*Central*	*Colorado*	*CO*
Central East	*Connecticut*	*CT*	*Central*	*Wyoming*	*WY*
Central East	*Massachusetts*	*MA*	*South Central*	*Missouri*	*MO*
Central East	*Rhode Island*	*RI*	*South Central*	*Arkansas*	*AR*
Central East	*New Jersey*	*NJ*	*South Central*	*Louisiana*	*LA*
Central East	*Indiana*	*IN*	*South Central*	*Oklahoma*	*OK*
Central East	*Kentucky*	*KY*	*South Central*	*Texas*	*TX*
Central East	*Tennessee*	*TN*	*South Central*	*New Mexico*	*NM*
Central East	*Delaware*	*DE*	*North West*	*Washington*	*WA*
Central East	*District of Columbia*	*DC*	*North West*	*Idaho*	*ID*
Central East	*Maryland*	*MD*	*North West*	*Montana*	*MT*
Central East	*North Carolina*	*NC*	*Central West*	*Oregon*	*OR*
Central East	*Virginia*	*VA*	*Central West*	*Nevada*	*NV*
Central East	*West Virginia*	*WV*	*Central West*	*Utah*	*UT*
South East	*Alabama*	*AL*	*South West*	*California*	*CA*
South East	*Mississippi*	*MS*	*South West*	*Arizona*	*AZ*
South East	*Florida*	*FL*	*Others*	*Alaska*	*AK*
South East	*Georgia*	*GA*	*Others*	*Hawaii*	*HI*
South East	*South Carolina*	*SC*			
North Central	*Michigan*	*MI*			
North Central	*Wisconsin*	*WI*			
North Central	*Minnesota*	*MN*			
North Central	*North Dakota*	*ND*			

AB: standard two-letter postal abbreviation

APPENDIX E

Time series of estimated state proportions

Toronto



Approach A

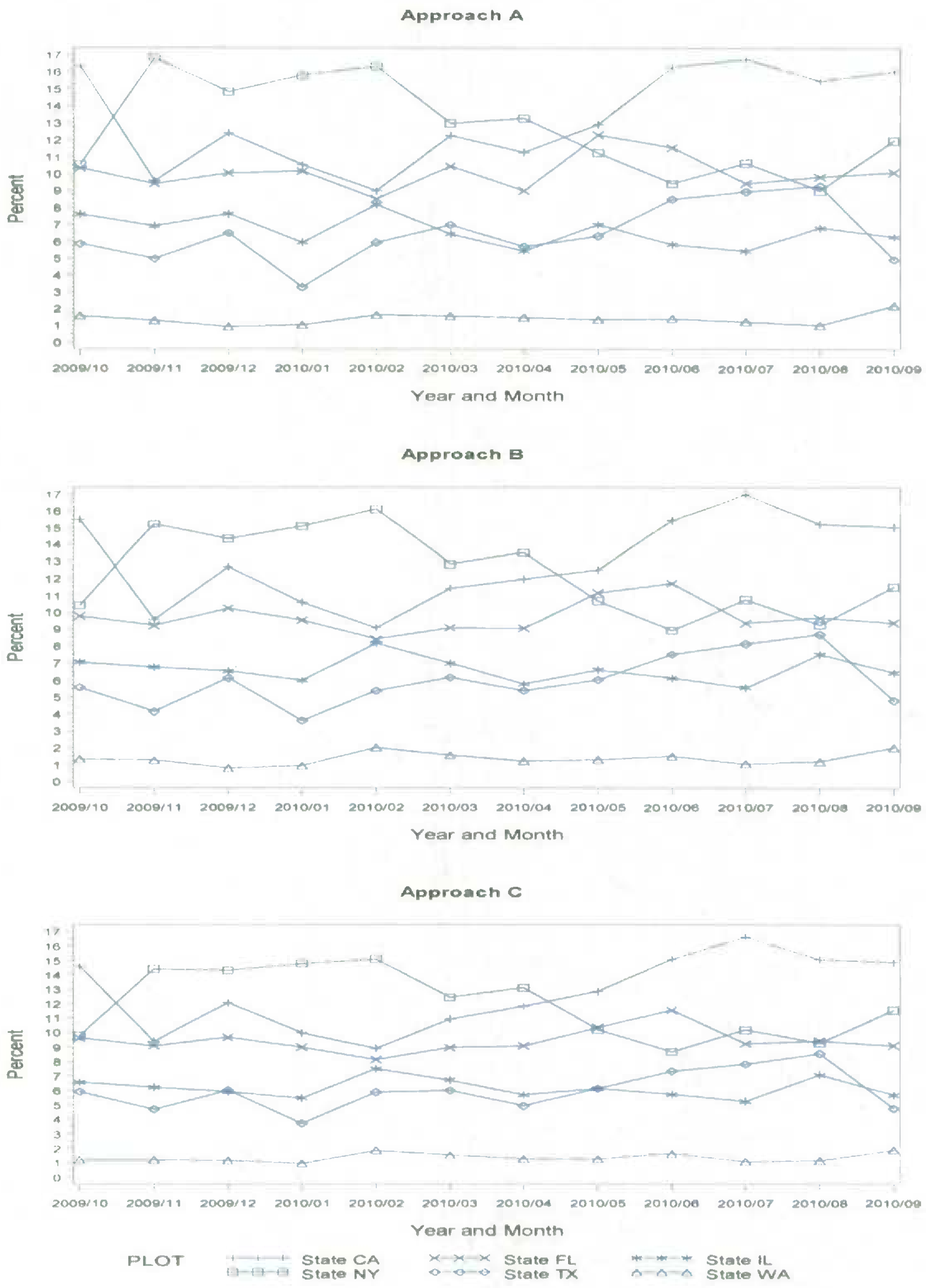
Approach B

Approach C

APPENDIX E

Time series of estimated state proportions

Montreal



Approach A

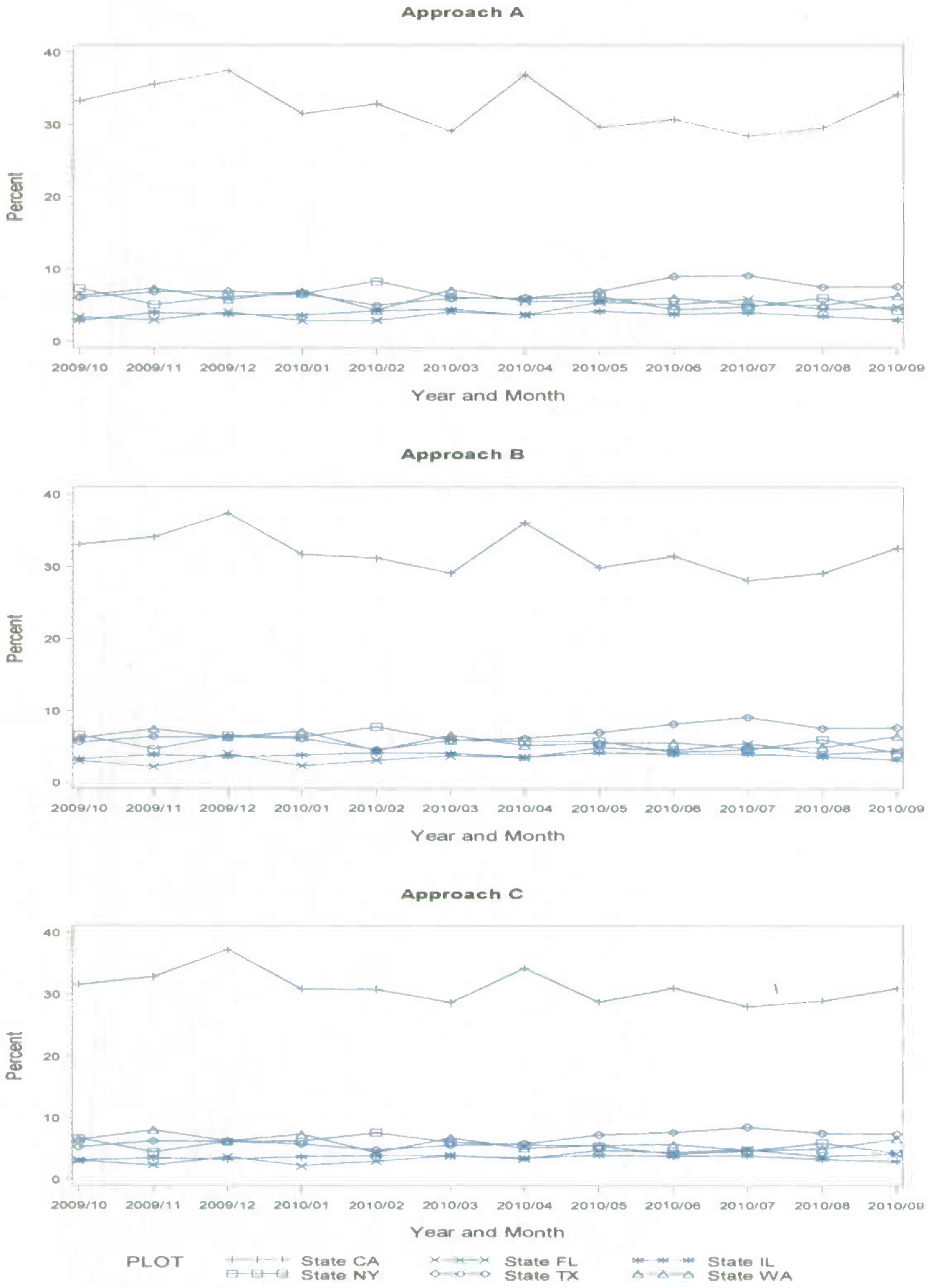
Approach B

Approach C

APPENDIX E

Time series of estimated state proportions

Vancouver



Approach A

Approach B

Approach C