Australasian Road Safety Research, Policing and Education Conference 2012
4 - 6 October 2012, Wellington, New Zealand

Enhancing road safety research and analysis using geospatial information tools

Renee Schuster, Michael Nieuwesteeg, Jodi Page-Smith

Transport Accident Commission, Victoria, Australia

email: renee_schuster@tac.vic.gov.au

Abstract

In 2011, the Victorian Transport Accident Commission (TAC) obtained new data and software designed to validate, correct and geographically code address information in large volumes. These new tools have enabled the TAC to enhance its datasets with more complete and accurate address information and allow the appending of geo-demographic information (such as Australian Bureau of Statistics (ABS) Census data) down to neighbourhood level.

These dataset enhancements have already had a significant impact on key areas of the TAC's road safety research program including an ongoing risk modelling project aimed at identifying key target markets, and key drivers of road trauma, injury severity and TAC compensation costs. The addition of geo-demographic information to input datasets have enabled conclusions to be drawn about crash risks associated with factors such as ancestry, language background, education, employment industry and income.

The dataset enhancements have also benefited TAC's social and market research survey program with improvements in sample representativeness, mail out address accuracy and phone number matching processes. The addition of geo-demographic data has opened up analytical opportunities for survey results with insight into behavioural and attitudinal differences by geo-demographic factors such as socio-economic disadvantage.

This paper discusses the TAC's new geospatial information tools and its applications to date within the road safety research program.

Keywords: road safety research and analysis, geospatial information tools, data enhancement, geocoding, address cleansing, geo-demographic data.

1. Introduction

The Transport Accident Commission (TAC) is a statutory no-fault compensation scheme that provides coverage for all persons injured in transport accidents in Victoria, Australia. The TAC is a "no-fault" insurance scheme, which means that medical benefits will be paid to an injured person regardless of who caused the accident. The TAC is funded by compulsory payments made by Victorian motorists as part of the vehicle registration and annual renewal process.

A key function of the TAC is "to promote the prevention of transport accidents and safety in use of transport" (Transport Accident Act 1986). This means that the TAC is also responsible for delivering

public education and road safety programs aimed at reducing road trauma. The TAC works in partnership with Victoria Police (Police), VicRoads[1] and the Department of Justice (DoJ) to deliver these objectives.

The TAC's accident prevention Strategy is a multifaceted program covering a number of road safety issues. The Strategy continues to evolve over time as the Victorian road trauma picture changes. Since 1989 there has been an emphasis on research in developing TAC road safety and marketing initiatives and on evaluating their effectiveness. Today, the TAC manages a comprehensive research program that is used to assess the merits of current programs as well as the basis for identifying effective and cost beneficial interventions for the future.

The TAC has a small internal research capacity with an externally focused approach to research and analysis. Since its inception, the research and development program has been primarily focused on research partnerships, performance monitoring and evaluation of marketing and road safety programs. Over recent years the research program has pursued the acquisition of analytical tools and data to enhance operations and analytics, including a number of geospatial information tools.

The research strategy has now begun to move beyond the acquisition and improvement of data towards the discovery of insights that provide clear direction to the road safety and marketing program.

2. Road Safety Research Program background

The TAC's Road Safety Research Program involves social and market research with the road user population, funding of external academic-based research, and internal analysis and performance measurement. This section presents background information on relevant aspects of the Program.

2.1 Data management and analysis

Data management and analysis is the central activity of the research team, and is a key source of information for strategy and program development as well as performance measurement. There are strong links and collaborative efforts with road safety partners and other data-collecting entities that serve to efficiently deliver the intelligence required to formulate good road safety strategy and policy.

The TAC is fortunate to manage and have access to a number of key road safety datasets. This includes the TAC's own client databases, Victoria Police Traffic Incident System, VicRoads Road Crash Information System and, more recently, the VicRoads Registration and Licencing (RandL) Database[2]. This wealth of data is also supplemented by other useful road safety information and datasets such as Ambulance Victoria and State Emergency Services road crash attendances.

2.3 Social survey and qualitative research program

Social surveys and qualitative research are key tools used for road safety performance monitoring and evaluation, and also form the basis for exploratory research. The TAC's road safety research program currently runs a series of qualitative focus groups as well as 5 to 6 quantitative surveys per year with the general motorist population or with targeted groups (such as motorcyclists, regional road users, youth, risk takers etc.) in order to collect information about road safety attitudes and behaviours, and recall of road safety advertising.

Prior to 2010, the TAC was reliant on using online panels and random digit dialling for research recruitment. Since 2010, the TAC has been able to receive an annual extract of the VicRoads RandL

[1] VicRoads is the road network, vehicle registration and vehicle license management authority in Victoria.

[2] The VicRoads Registration and Licencing (RandL) Database contains information about all persons who hold a drivers licence and/or have a registered vehicle in Victoria.

database, which contains name and address details. This acquisition opened up sampling and recruitment opportunities for both the quantitative and qualitative research avenues.

TAC survey methodology development 2001-2010

The TAC's Road Safety Monitor (RSM) survey has been conducted annually since 2001 and was initially run as a telephone survey. In recognition of falling response rates and diminishing community coverage, a parallel online survey was conducted from 2007 with the view that the internet version of the RSM would be able to form a new benchmark, enabling the CATI version of the RSM to be discontinued within a couple of years. While the internet pilot was largely satisfactory, it did not provide an acceptable benchmark for future studies.

With the arrival of the VicRoads RandL file in 2010, this list was able to be tested as a new survey frame, with the aim to improve the coverage and representativeness of the sample population. A mixed methodology involving mail invitation with the option for the participant to complete and return the survey either by mail or to complete the survey online was tested. The new methodology and frame base have succeeded in their aim, have continued to be implemented across a number of other survey instruments over the last 12 months and continue to improve the quality of survey outputs.

With some phone number information available, the new VicRoads list has now also been successfully tested for CATI based surveys, focus group research recruitment and conducting phone follow up / reminder activity in mail out based surveys.

2.2 Exploratory research

Usually conducted in-house, some research and analysis is directed towards exploration. This involves identification of key drivers of trauma and emerging issues. This aspect of research is a key input into strategy development and business planning.

To date, the TAC and other road safety agencies have relied heavily on police reported crash data to inform strategies and measure progress. The TAC has long maintained a link between data held on its claimants and data recorded by police about the crash. This enabled the construction of a linked dataset, thus providing a rich source of crash information supplemented with injury outcomes. However, until recently, TAC claims data has only been utilised to a limited extent, predominantly to supplement crash data with information on length of hospital stay and understand the broad cost implications of road trauma.

3. New Data and Information Acquisitions

In recent years the TAC has engaged widely with its stakeholders to increase its evidence base. Recent data acquisitions include annual snapshots of the VicRoads RandL database, more detailed information on injury classifications and severity, detailed vehicle specification data, estimates of lifetime cost of single claims and vehicle crashworthiness ratings.

In early 2011, the research team agreed it would also benefit from improved geospatial information in its datasets in order to enhance location intelligence; including improving address accuracy, allow more accurate spatial mapping and allow the appending of geographic based socio-demographic (geo-demographic) data. The geospatial tools purchased and implemented throughout 2011 are presented in Table 1.

Table 1: Geospatial tools purchased and implemented by TAC Road Safety Research, 2011

Intech IQ Standardiser	Intech's IQ Standardiser is a software package designed to validate and correct address data by matching it to an Address Reference File (ARF) of choice. During this process address data can also be enhanced through geocoding, which involves assigning geographic coordinates and other geographic codes to each address. This includes latitude/longitude, Australia Post Delivery Point Identifier (DPID) and ABS Geographic Codes down to neighbourhood level such as Census Collection District (CCD) and Statistical Area 1 (SA1).
Sensis MacroMatch	The Sensis MacroMatch service utilises White Pages[3] to append phone numbers to data containing name and address information. This is a bureau "pay per record" service.
ABS 2006 Census DataPack	The Australian Bureau of Statistics 2006 Census of Population and Housing DataPack is a product containing Basic Community Profile data for all of Australia to the Census Collection District (CCD) level and the matching digital boundary maps in generic Geographical Information System (GIS) format. The Basic Community Profile data contains basic demographic information for an area including age, ancestry, income, education, family type and much more.
MapInfo Professional	MapInfo Professional is a mapping and geographic analysis application designed to easily visualise the relationships between data and geography.

The TAC has now also acquired the ABS 2011 Census DataPack which was released in mid-July 2012.

4. Application of geospatial information tools

4.1 Social survey and qualitative research program

The acquisition of the VicRoads RandL snapshots and subsequent proof of concept of new methodologies in using this dataset as a survey frame was a significant step forward in improving population representativeness and response rates for the TAC survey program. The purchase of the new geospatial tools meant further refinements and enhancements to the new survey methodology were also possible.

The first step of this process involved undertaking a quality assessment of the key variables required for mail out and phone follow up activity; that is, address and phone number fields. Understanding completeness was a relatively simple task, revealing that address was available in more than 99.9% of cases and phone number was available in 56% of cases. Understanding the quality of the address and phone number fields was a more difficult task.

The main issue around address quality included currency and accuracy of address details for mail delivery purposes. The vehicle registration renewal process in Victoria involves an annual mail out to registered vehicle owners. The re-registration process is relatively costly and time consuming, which suggests that vehicle owners are unlikely to allow their vehicle registration to lapse unintentionally. It was therefore assumed that address information for records with a registration attached were highly likely to be kept up to date and accurate with VicRoads. On the other hand, a Victorian driver's licence can remain current

[3] The White Pages is the Australian residential telephone directory.

and valid for up to 10 years without any contact with VicRoads (depending on driver history, licence type, licence status and length of licence). Given that licence only records account for almost 50% of the file and are much less likely to be updated regularly, there were concerns about the quality of address information for these records in particular.

Basic phone number quality assessment and cleansing could be undertaken to a limited extent in house through simple text string assessment of factors such as field length and character types. The main issue with the phone number field was completeness.

The Intech IQ Standardiser in combination with the Sensis Macromatch service proved to be the solution to both assessing and improving the quality of address information and improving both completeness and quality of phone numbers.

With sufficient demographic and high level geographic information available in the RandL file (including age, sex an postcode) it was decided that in most cases, robust random samples could successfully be drawn from this raw file. Therefore, it was only necessary to run the Intech process for the resulting samples, which led to time and cost savings.

During the address cleansing process, the samples are also geocoded and geographic information such as longitude/latitude, CCD, SA1 and DPID are appended to the sample. The addition of CCD/SA1 in the sample file, and eventually in the respondent file, means that neighbourhood level data from the ABS Census can be used to analyse trends by a large range of socio-demographic factors.

Once the sample is final and has clean address information, the sample is then run through the Sensis MacroMatch bureau service to validate existing phone numbers and append new phone number information for given name and address information.

Table 2: Address cleansing, geocoding and phone number matching results

Address cleansing	Of the 8 samples that have now been drawn and geocoded in 2011 and 2012 (almost 30,000 sampling units): • 75% of address records were validated to be correct, • 15% were amended through the automated geocoding process (which may involve a correction of street name spelling, street type, postcode etc.) • 10% required further investigation through a manual checking process (often involving the assistance of public mapping sources such as Google maps). The final 10% marked for manual investigation are often those of recently subdivided land, new inner city apartments or rural in nature (e.g. roadside mail boxes or "care of post office"). To avoid unnecessary bias, such addresses which cannot be validated by the software or manually amended are not necessarily excluded from survey samples. For example, sampling units with a registration attached are generally left in the sample, as it is likely this address has been successfully used in the registration renewal process within the last 12 months. Overall, a mere 1-2% of units are discarded from the samples due to insufficient address information for mail delivery purposes.
Geocoding	The geocoder is generally able to append geographic codes at street address level to 80-90% of records.
Phone number	This process has been found to improve phone number availability rates from 55% to up to 75%. This means that at least one phone number is available for 75% of any given

matching	survey sample; with some sampling units having up to 3 or 4 phone numbers available.

As the capacity to append socio-demographic data at the neighbourhood level to respondent files is relatively new, no analysis or results are yet available to report on. Furthermore, given the only recent availability of 2011 Census results, no analysis has been undertaken on the appending of SA1 codes to date.

Case Study: Improving sample representativeness in the Motorcycle Monitor Survey

Motorcycle rider attitudes and behaviours have been surveyed on an ad-hoc basis over the last few years, generally as part of other surveys. In 2009, the TAC commissioned a continuous survey to specifically track motorcycle rider attitudes and behaviours in relation to road safety issues and to measure the prompted recall of motorcycle advertising campaigns when on air. In 2012, a further new survey was introduced to gather detailed information about motorcycle riders and their attitudes toward road safety; and their behaviour while riding their motorcycles. The intention was to gain a fully representative sample of the motorcycle rider population; with a particular focus on active riders.

On the VicRoads RandL database, there are approximately 300,000 motorcyclists with either a current licence or a registered motorcycle. Preliminary analysis revealed that a significant proportion of motorcyclists had a current motorcycle licence but no registration (62%). This suggested that a large proportion of the potential survey frame were not necessarily active riders and could potentially be removed prior to sampling. It was suspected however, that there may be a tendency for licence holders to share a registered motorcycle with others within their household.

To test this theory, the decision was made to geocode the entire motorcyclist population. Using DPID information, it was possible to aggregate individuals to a household level and flag each person record based on whether there was a motorcycle registered at their address.

One major limitation of this process was the fact that a large number of records could not confidently be aggregated to a household level due to missing or insufficient DPID information. Despite this limitation and the fact that only a further 4% of records were identified as potentially active riders, it provides a starting point to further refine and improve sampling representativeness for the benefit of future surveys.

The addition of the new geospatial tools has seen significant methodological improvements in the Road Safety Research Program over the last 12 months. This includes a 15-25% improvement in mail out address accuracy, a 20% improvement in phone number availability rates, as well as improvements in sample representativeness and ability to target specific populations of interest. The tools have also allowed enhanced analytical capability in the appending of neighbourhood level geo-demographic data to respondent files.

4.2 Road Safety Risk Models project

Analysing road safety risk factors in a one and two dimensional capacity can be potentially misleading. With access to so much data and information on road crashes, claims, licence holders and registered vehicles, and with ongoing data quality improvements and enhancements taking place, the next logical research question was “What are the significant drivers of crashes, injury outcomes and claim costs when controlling for all other factors”?

The core aims of the TAC Road Safety Risk Models project were to conduct sophisticated analyses of road safety and related data to identify key target markets, and key drivers of road trauma, injury severity

and TAC compensation costs. The project progressed throughout 2011 and involved building a suite of statistical models to identify significant factors when predicting crash frequency and crash severity.

The early phases of the project involved an intensive data build and exploration phase which included data sourcing, linking, cleaning, quality assessment and preparation of 5 years of data (2006-2010). During these early phases, geocoding was undertaken and socio-demographic data was subsequently appended to datasets where street address information was available. Where street address level geocoding could not be undertaken, socio-demographic data was appended at the postcode level.

During these early phases it was discovered that, for the frequency models, the modelling process required that the same information be available in both the predictor and exposure datasets. Predictor datasets contain data on the events the model is trying to predict (i.e. crashes/claims) and the exposure datasets contain data on those people/vehicles potentially exposed to the event; in this case licenced drivers and registered vehicles. This constraint significantly restricted the amount of variables available for modelling within the raw datasets. Through geocoding of selected datasets (including that of claimants, drivers and vehicle owners), and the subsequent addition of socio-demographic variables down to neighbourhood level, there were many more variables available to test in the modelling process that would otherwise not have been available.

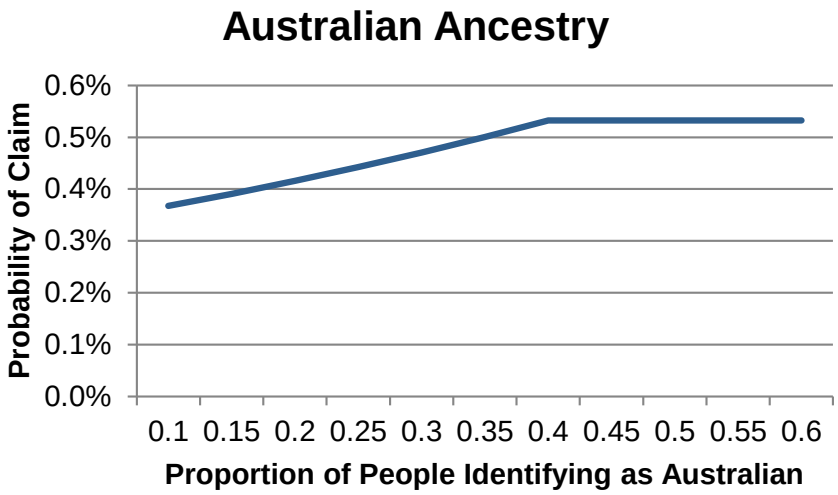
Table 3: Summary of TAC Risk Models and significant socio-demographic variables.

Model	Description	Significant socio-demographic variables
Claim Cost Severity	Estimates the no-fault cost of a claim.	Median household Income Language spoken at home No. of motor vehicles garaged at dwelling Level of education
Injury Severity	Estimates the probability of a claim being a minor / moderate / serious / severe injury.	Level of education
Vehicle Frequency - Single Vehicle - Multiple Vehicle	Estimates the probability of a vehicle being involved in an accident where a claim is subsequently made.	Median household Income Language spoken at home Ancestry No. of motor vehicles garaged at dwelling Level of education
Person Frequency - Single Vehicle - Multiple Vehicle	Estimates the probability of a driver being involved in an accident where a claim is subsequently made.	Median household Income Language spoken at home Ancestry No. of motor vehicles garaged at dwelling Level of education Employment industry

The following chart presents results from one of the Person Frequency models; the model for single vehicle crashes. This chart reveals that a person living in a neighbourhood with a relatively higher

proportion of persons identifying as Australian, has a higher probability of being involved in a single vehicle crash and consequently submitting a claim.

Figure 1: Probability of a person being involved in a single vehicle crash and consequently submitting a claim by the proportion of people identifying as Australian in their neighbourhood of residence, 2006-2010



4.3 Spatial mapping

The addition of MapInfo and GIS maps to our suite of geospatial tools has allowed the TAC to create custom maps on a needs basis to undertake geographical based analysis. This analytical capability has again been enhanced by the improved quality, detail and breadth of geospatial data across the datasets.

Appendix 1 presents a map that was prepared for the TAC Advertising Team to help inform the placement of targeted outdoor advertising.

6. Conclusion

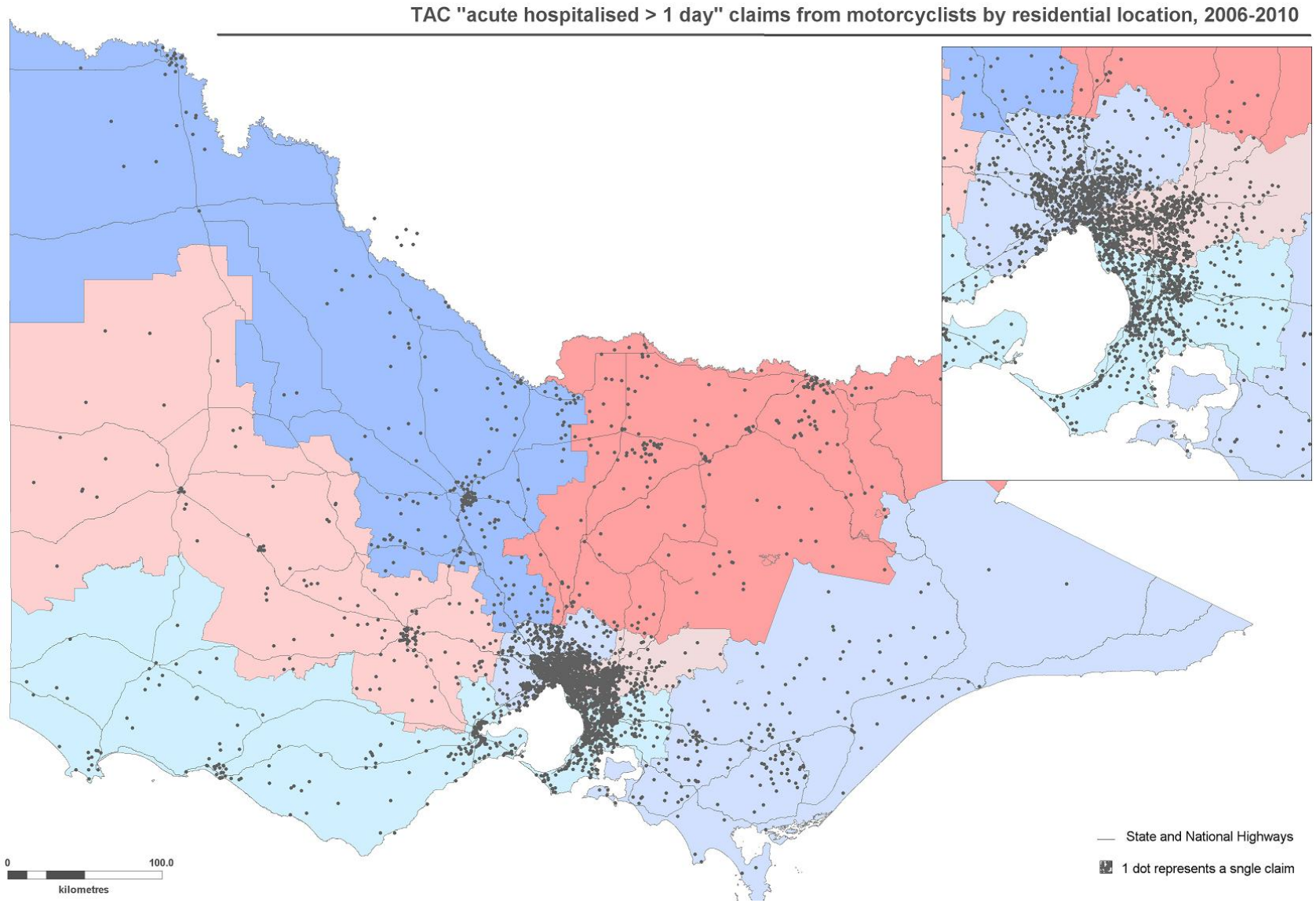
The successful implementation of the new geospatial tools has already made an impact within the Road Safety Research Program. This includes:

- Achieving significant improvements in mail out address quality, phone number quality and availability, and sample representativeness;
- Opening up analytical opportunities for survey results with insight into behavioural and attitudinal differences by geo-demographic factors;
- Enabling conclusions to be drawn in the Road Safety Risk Modelling Project about crash risks associated with a range of geo-demographic factors.

Over the next 12 months, the TAC Road Safety Research team will continue to work with the new tools to further enhance datasets and improve its processes. This includes transitioning to the new ABS geography structure and 2011 Census results, and working to better integrate the new software tools into existing data management software.

While efforts in recent years have been heavily focused on improvements to survey methodology and laying the foundations for future in-depth analytics, the proposed research plan ahead is very much concerned with deriving insights from this enhanced information that will guide and support the TAC Road Safety and Marketing strategy.

Appendix 1


TAC "acute hospitalised > 1 day" claims from motorcyclists by residential location, 2006-2010
0
100.0
kilometres
State and National Highways
1 dot represents a sngle claim

Appendix 2

Acronyms

ABS	Australian Bureau of Statistics
ARF	Address Reference File
CATI	Computer Assisted Telephone Interviewing
CCD	Census Collection District
DoJ	Department of Justice
DPID	Delivery Point Identifier
GIS	Geographic Information System
RandL	Registration and Licencing
RSM	Road Safety Monitor
SA1	Statistical Area 1
TAC	Transport Accident Commission

Definitions

ARF	An Address Reference File is a dataset containing a collection of addresses and other address related data. Examples of address reference files in Australia include the Australia Post Postal Address File (PAF) and the Geocoded National Address File (GNAF).
CCD	Census Collection District: collection unit and smallest reporting unit in the 2006 ABS Census.
DPID	Delivery Point Identifier: A unique number assigned to a postal delivery point (such as a street address) as recorded on the Australia Post Postal Address File.
Geocoding	Process of finding associated geographic coordinates and/or codes from other geographic data, such as street addresses or postal codes.
Geo-demographic	Demographic data for a given geographic area.
Geospatial	Pertaining to the geographic location and characteristics of natural or constructed features and boundaries; especially referring to data that is geographic and spatial in nature.
SA1	Statistical Area 1: smallest reporting unit in the 2011 ABS Census.
Survey Frame	The survey frame refers to the list of units (e.g. persons, households, businesses, etc.) in the survey population of interest.
VicRoads RandL Database	The VicRoads Registration and Licencing (RandL) Database contains information about all persons who hold a drivers licence and/or have a registered vehicle in Victoria.

References

Australian Bureau of Statistics	www.abs.gov.au
- Geography	www.abs.gov.au/geography
- Census	www.abs.gov.au/census
Department of Justice	www.justice.vic.gov.au
Intech Solutions	www.intechsolutions.com.au
- IQ Standardiser	http://www.intechsolutions.com.au/Products/IQStan.asp
MapInfo Professional	http://www.pbinsight.com.au/products/location-intelligence/applications/mapping-analytical/mapinfo-professional
Sensis MacroMatch	http://www.sensisdata.com.au/products/macromatch.html
Transport Accident Commission	http://www.tac.vic.gov.au
VicRoads	www.vicroads.vic.gov.au
Victoria Police	www.police.vic.gov.au