1. INTRODUCTION

Opus International Consultants Ltd (Opus) were commissioned to undertake the Household Interview Survey (HIS) to support the development of the Waikato Regional Transport Model.

A first version of the "final" survey database was received on 17 November 2008, followed by a subsequent delivery on the 11th December with the final, complete version received on 15th January 2009.

The following issues arose regarding the deliverables for that contract:

- Problems in drawing the sample;
- Failure to complete the project within the timeframe specified by Opus;
- Inability to calculate expansion factors;
- Incorrect geocoding;
- Delivery of survey database with missing data and in a format incompatible with the RFT requirements;
- Delivery of survey database containing a multitude of errors; and
- Contents and quality of the Final Report.

The issues associated with each of these items are summarised in the following sections.

2. DRAWING THE SAMPLE

In the RFT for the Survey, a detailed sample frame was included as Attachment B, which specified the number of households to be sampled within each Sample Area. This was designed to ensure an appropriate sampling rate in all regions of the study area for the purposes of model build.

It was evident that Opus had some difficulty in interpreting the sampling requirements as specified, and they proposed an alternate sampling frame on the 16th April 2008, which randomly sampled meshblocks within the study area to meet the required 2000 households regardless of location. Following a review of this data it was established that following the Opus guidelines, some regions of the study area were grossly undersampled.

The peer reviewer appointed by LASS (Ian Clark of Flow Transportation Services) was consulted and in a letter to James Bevan on 19th May 2008, expressed concerns that

the survey would need to collect data from a higher number of zones than the Opus sample frame selected.

On 22nd May James Bevan and Grant Smith met with Bill Pitt and Darren Walton in Wellington to work through the issue of the sample frame and to agree a course of action. As a result, a more detailed sample frame was created and forwarded on the 16th of June 2008. This sample frame established HIS survey areas by aggregating Statistics New Zealand Area Units into regions and randomly selected meshblocks within each for Opus to use as a sample frame. The number of individual meshblocks to be sampled was chosen such that low sample sizes would be avoided in each respective area.

Despite having a second sample frame, which was agreed to be suitable by both parties, many of the meshblocks were not surveyed, alternate meshblocks were selected instead (some of which were located outside of the contract study area) and a number of HIS survey areas had no households sampled at all. This caused additional work upon delivery of the HIS data from Opus in the expansion process.

3. DELAY IN SURVEY PROJECT COMPLETION

According to Opus's work programme (dated 5 May 2008), completion of the fieldwork was programmed for 20th September 2008, with the project completed by the end of October.

The fieldwork had to be extended until 15th November to collect the required sample, with the final survey database not delivered until 15th January 2009.

The delay in the completion of the HIS led to a delay in the model build project, which had consequences in terms of additional costs.

4. EXPANSION FACTORS

Expansion is the process of applying factors so that the sample represents the total population of the study area. Clause 1.2 in section D of the RFT for the survey required the calculation of expansion factors to ensure the sample was representative.

The Opus methodology offered during the tender negotiation was to sample *'…each region proportionally and weight the data after the survey is completed in accordance with the NZ Travel Survey method'*

The Contractors for the Christchurch survey (TUTI) use the words 'weight' and 'expansion' interchangeably, which we consider to be industry-standard in the field of collecting interview data for the purpose of building transportation models. So we therefore assumed (erroneously as it turned out) that Opus did as well. What we did not realise at the time is that the NZ Travel Survey does not appear to be expanded – the weighting referred to by Opus is simply correcting sample means to population means –

instead of deriving the factor that enables estimation of the population from the sample. This did not become clear until we were provided with the documentation on 12th November in response to the expansion procedure that we asked them to use.

No expansion of the survey sample was carried out by Opus, who made it clear in correspondence that they were not prepared to follow the required procedure. Instead, the consultant team undertook the calculation of the expansion factors so that the sampled number of households summed to the number in the study area defined in the 2006 Census.

As the revised sample frame presented to Opus was not strictly adhered to, some households sampled were situated outside of the study area whilst a number of the defined HIS survey areas were undersampled (i.e. less than 10 households interviewed) and others were not sampled at all (i.e. no households interviewed).

Subsequently, it was necessary to aggregate the HIS Survey Areas into HIS Expansion Areas to ensure that at least 10 households were sampled in each area and that the level of expansion was not excessive. The primary concern with large expansion factors is that a very small number of households will represent a relatively large population, and in doing so any peculiarities become grossly exaggerated. This aggregation process was an iterative and time-consuming exercise.

This process has been exacerbated by problems with the survey database, which led to removal of certain households, which in turn required recalculation of the expansion factors. These issues, which are elaborated on in Section 7 of this Technical Note, included households surveyed outside of the study area, households with incomplete survey data, and duplicate households.

As noted in the comments later on geocoding, the check on meshblock allocation to Household sample numbers and addresses could not be reproduced in about one third of cases. This is currently being checked as it is an integral requirement of data expansion.

5. GEOCODING

Geocoding is the process of assigning an x and y coordinate to each location.

The first delivery of the survey database on 17th November was also the first opportunity to review the geocodes for each trip leg. The geocodes provided by Opus were mapped, which identified a notable number of stops geocoded to outside of the study area. These were checked in more detail by comparing the geocoded location with the address in the survey database, and as a result, many of the geocodes outside of the study area were found to be incorrect. For example, trips to Mary Street in Thames were geocoded to Mary Street in Invercargill. Given the extent of the discrepancies, a more comprehensive assessment was undertaken by applying the TDG geocoding procedure used for the Waikato Roadside Interview Surveys (RSI). From this, a quantified comparison could be made between the geocodes provided by Opus and those output by the TDG process. A memorandum was circulated summarising the

findings (HIS Checking geocodes.2008 12 17.doc). It was concluded that the geocoding had likely been carried out considering the street name and not the suburb or town.

Opus were notified of the problem and agreed to re-geocode the data taking into account the town as well as the street name.

The revised geocodes were included in the second delivery of the HIS database. On receipt of this, the TDG geocoding procedure was again applied to check the geocoding. A comparison of the TDG and Opus results found that 70% of the stop records were geocoded to within 100 metres and 80% to within 600 metres. There were several locations where the geocodes were clearly incorrect (in the ocean), but the relatively small number of errors was determined (at the time) to be a reasonable degree of error. The comparison was summarised in a memorandum (HIS Checking geocodes.2008 12 22.doc), which recommended that the second version of geocoding be accepted.

The third version of the HIS database was received in January 2009. This was in Access format (for the first time, a truly relational database was delivered as required by the RFT) and was principally supplied to resolve the problem of missing data. No geocodes were attached to this database (although it was possible to extract the geocodes from the second version of the database). Also for the first time, a unique list of all addresses was provided although there were no geocodes appended. Review and comparison of the addresses in the second and third databases revealed that some trips had not been geocoded at all and that some addresses had different geocodes for the last location. The TDG geocoding process was reapplied. Travel speeds were calculated using reported travel times and crow-fly distances based on both the Opus and TDG geocodes to identify outliers. Manual geocoding was undertaken for locations where no geocode was provided by Opus (and no match was produced in the TDG procedure). Manual geocoding was also used to correct anomalies in the database, and to resolve partial address data that could not be geocoded using an automated approach.

6. MISSING INFORMATION IN THE SURVEY DATABASE

The two versions of the HIS database delivered in December 2008 were incompatible with the RFT stipulations of “relational databases for the household, vehicle, person, and trip data”.

Instead, an Excel table was provided containing the required data for households, persons and stops, but omitted data on all household vehicles and the location of each person at the start of the travel day. This information was collected but was not delivered because of the format of the Excel file.

The complete survey information was received in the January 2009 delivery, in which tables for households, persons, vehicles and stops (more consistent with the requirements in the RFT) were provided.

The missing data meant that progress on the model build project was stalled until mid January, at which point, work could continue. Again, this had cost implications for the main model build contract.

7. ERRORS IN THE SURVEY DATABASE

In interrogating the survey database, a number of serious discrepancies were uncovered. These problems had to be rectified by the consultant team so as not to undermine the platform for the model development.

Examples of the more serious errors in the data are:

- Households outside of the study area were included within the survey database. These had to be removed.
- Data for two households was included twice, but with different household identifiers. Effectively, where it appeared four houses had been surveyed, there were only two. The duplicate data had to be removed from the database.
- One household had an incomplete travel diary and was removed as non responding, which was in accordance with the survey RFT and the definition of a "responding household".
- The lookup table to convert the trip activity from a number to a text string was inconsistent with the MOT survey software provided during the Pilot Survey. It was assumed that the lookup table was incorrect and the table was therefore modified by the consultant team. This error could have resulted in incorrect allocation of travel to model trip purposes.
- There were records with essential data missing. In most cases, responses were infilled by the consultant team to complete the interview so that the household did not have to be rejected, which would reduce the sample size.
- Some records were so incomplete that it seemed likely that the record was invalid. For example, there were 8 records for vehicles that contained no other information aside from a vehicle counter created by the software. In these cases, it was concluded as unlikely that the record represented a vehicle and was therefore deleted from the database.
- One household was reported with a home address of Thames Pak N Save (102 Mary Street). This was corrected based on information in the Address database, where a valid address was available.
- Households interviewed during the Pilot Survey were included within the database. Some of these were interviewed during school holidays, in direct conflict with the survey requirements. All of these households had to be retained since following removal of duplicate households, incomplete surveys, and households outside of the study area, the remaining number of households

would have been less than the required minimum sample size of 2000 households (weekday).

- Some fields contained a significant number of unlikely responses. It is possible that the interviewers were instructed not to complete these survey questions, but as the data was supplied, it was necessary to review the fields and assess what data looked correct and what appeared to be incorrect. For example, the field "off-road" is a flag for an off-road trip. The survey report states that these trips were not collected, but the survey database includes both on and off-road trips (i.e. the field was used). However, as a high proportion of off-road trips occurred (e.g.: a shopping trip within Hamilton in a Celica was supposedly "off-road"), this field was assumed to contain incorrect data.

- There were a massive number of inconsistent responses to questions on the same trip, some of which are likely to be mis-typed by the survey enumerators. An example of a likely typo is a trip to a farm which was listed with a Destination Type of "Airport". The correct response was an "Agricultural Establishment", adjacent to Airport on the drop-down list in the survey software. An example of inconsistent responses is a trip where the person was a passenger in a vehicle but the vehicle only contained one person (i.e. there was no vehicle driver). There were many hundreds of inconsistent responses such as these.

- One household with incomplete survey data was included in the January version of the database but excluded from the December deliverables. This household was removed from the database.

In summary, the consultant team invested a month cross-tabulating the responses from the HIS to detect and repair ("clean") errors such as these.

8. QUALITY OF THE FINAL SURVEY REPORT

According to the RFT, the Final Survey Report was to include "a summary of analysis of travel patterns" as well as details of the sample rates, all data coding frames, reporting on the range and logic checks, and details of the data expansion process. The Survey Design Report, which was submitted prior to the fieldwork, was to be incorporated in the Final Survey Report as the initial sections.

The Final Survey Report dated January 2009, was delivered in hard copy only. An electronic copy, specified in the RFT, was not provided (although a copy of the draft report was provided electronically).

The main concerns with the Final Survey Report are:

- Although the Survey Design Report was incorporated as required, no information on the travel patterns collected was summarised.

- There is limited reference to the objectives of the survey.

- The study area is not defined, although there is a reference to Franklin District being excluded and some area units in Western Bay of Plenty District being included.

- Reference is made to 149 "survey areas" to draw the sample. No diagrams or tables are provided to illustrate the survey areas, which were available from Gabites Porter.

- There is no reference to conducting a pilot survey, which is crucial for testing and verifying all field and data handling procedures. A pilot survey was conducted but is not reported.

- The report states the total response rate was 2,586 households (section 13.1). The breakdown of households by travel day, however, sums to 2,551 households although the total is sill shown as 2,586 in section 13.3 of the report. The database included 2,551 households (although there was one extra in the January 2009 version). It is considered that Opus delivered 2,551 "responding households", although it was subsequently discovered that some of these were outside the study area, were incomplete or duplicated.

- The report states that the survey was conducted from 21 July to 15 November 2009. The survey database, however, contains 50 households surveyed during the pilot survey and with a travel day prior to the 21 July 2009. There is a mismatch between the survey database and the final report in terms of the survey dates.

- There is no indication of the accuracy of geocoding achieved, aside from stipulating that 3% of the data had to be manually geocoded. The final database does not indicate the level of accuracy for each geocode.

- The complete survey questionnaire is not provided. In fact, there is no listing of all of the questions.

- A copy of the pre-contact letter referred to in the flow diagram in section 3 of the report is not included.

- The coding frames for the survey database are not included within the report. In section 14 of the report, it is stated that the "coding frames" are included within the Excel workbook, although these were not comprehensive.

- The "Range of Logic Checks" (should have been range AND logic checks) listed in section 12 of the report do not appear to have been comprehensively applied to the database. Data was delivered that failed certain checks. For example, the database included one house with a person number zero, which fails the first Person check "invalid person number (0).

- There is no indication in the "Range of Logic Checks" which denoted a warning message (i.e. please consider more carefully) and which were fatal flaws.

In summary, the Final Survey Report contains errors, omits key information about the survey and does not conform to the specifications in the RFT.