JOHN REID
Managing Director
Austraffic & Global Counting Systems

John.Reid@austraffic.com.au
John@globalcountingsystems.com

Affiliations
Life Member AITPM
Past National President AITPM

Co-Author Chapter 31 “Traffic Surveys” “Traffic Engineering Methods” Text Book,
Monash University 2017 ISBN 978-0-6481898-0-0

Big Data: Informed wisdom or fuzzy logic?

Big data is often talked about in the abstract. We can now collect huge numbers and apply clever algorithms, so we think the answers must always be better and can be taken straight from the new technology. But the devil is in the detail.

With data collection and algorithms running on auto pilot, they can end up averaging a huge quantity of measurements down to a few numbers. Errors and unwanted information can be buried within a mass of records. Big data sets, make it easier to force a correlation that isn’t necessarily accurate.

Modern survey data collection techniques, such as video recording and Number Plate Recognition software, has given us the illusion of a 100% perfect sample. This is not true and adjustment factors and factoring of results are still a significant part of the final result.

The theme of this paper is to take an origin destination survey data set and apply a range of fuzzy logic matching techniques that might be applied to produce a trip table with a greater number of matches and then test the results for their credibility. The pressure to drive more numbers into the matched tables creates distortions.

By correlating with travel time results and managing the impaired entries in a data file of vehicle registration plates, debunks the usage of a common technique of just substituting any two characters of a plate record to force a greater number of matched events.

There can be a genuine application of fuzzy logic to enhance a match rate, but this has to be done in the context of traffic engineering skills, knowledge and wisdom in producing credible data outcomes.

Proper data collection, management and processing is a critical part of our profession which is too often not taught or over looked.

1. Automatic algorithms cannot always replace traffic engineering judgement

At the previous AITPM National Conference I noted the concerns across the world of how algorithms can lead to immense misunderstandings of the real situation (Reid 2017a).

The danger is that algorithms can be buried in the computing black box and that the answers they produced are not subject to credibility checks.

In the sample origin/destination we consider below, we review a number of traditional character substitution techniques that should increase the number of matched trips.

These character substitution techniques, which look at matched trips and travel time statistics, logically should, reflect the trends of the raw data.

When they don't and with a deeper look at the data we see that automatic fuzzy matching of number plates, needs to be applied in a limited way and with great understanding and it does not replace the need for the core elements of traffic management intelligence in data processing.

2. Origin/Destination Surveys

2.1 Background - The old and the new of collection techniques.

An excellent demonstration of the dangers of striving to "optimise" survey numbers via data filtering and massaging is the Origin Destination (O/D) survey.

O/D surveys determine trips across and through a defined area. Number plates are recorded when a vehicle crosses a boundary or a particular point within the study area. If a plate is recorded at two or more locations, then it is defined as a specific trip or multiple trips. If a plate is only recorded once, then it is assumed to have an internal origin and/or destination.

In the past, in order to keep the manual recording process to an achievable level, a sample of number plates (it could have been white and/or red vehicles, trucks, buses etc.) were recorded, usually with one surveyor "spotting" and calling the plates and one writing down the results. The whole number plate may not have been recorded. A partial number plate is reasonably accurate for matching (and with a reduced work load there is an assumed improvement in the accuracy of what is identified and recorded). Information was recorded in 15-minute intervals (any reasonable volume of traffic exceeds a person's ability to write number plate and the exact time in a HH:MM:SS format). Results then had to be factored up to represent 100% of vehicle flows.

2.2 New technology

Video technology is now used to record the plates and transcribers can review the tapes or NPR systems can be applied. The advantages are obvious:

- Video technology facilitates the recording of complete (whole) number plates, enhancing the accuracy of plate matching for trip identification.
- With the ability to record all events there is an opportunity to account for most of the vehicles passing through a station.

- Where number plates are not clear, the vehicle can at least be counted, possibly classified and or partially recorded, increasing match rates and data quality.
- Video technology allows data to be reviewed in slow time, rewound for checking or at speed giving control to the operator to maximise outcomes.
- Partnering O/D and overview cameras technology introduces another level of information that can be used to assess various traffic conditions such as curfews or freight movements.
- The video is time stamped and so the time of the event can be identified to the second.

2.3 General problems in control sampling

But while the system may appear to get a 100% sample and therefore seem foolproof, there is room for error.

- Videos can be affected by site conditions such as the angle of the sun, storm or wind conditions. You have to know when this is a problem rather than try and apply a correction factor across the entire period.
- Number plates can be obscured or missing.
- The O/D camera is designed for a narrow focal point targeted at the likely location of a registration plate. O/D cameras require infrared attributes to record registration details in challenging light or night time conditions. O/D cameras consequently do not capture a good overview of an intersection and may be unsuitable for distinguishing vehicle classifications.
- The pressure to reduce costs can mean that transcribers are expected to review the videos at higher speed which decreases the accuracy.
- Problems arise in differentiating between digits; is it an “0”, “O” and “Q” for example. Should you differentiate or assume they are the same with the small risk that two number plates are almost exactly the same? The alternative is that you assume you can record it accurately every time. If there is any inaccuracy in transcribing a plate, then errors tend to compound in the one direction, making through trips look like two trips with an internal origin or destination in the survey area rather than one through trip.

2.4 The need for validation processes

Across the board, aggregation of data is the great masker of erroneous or missing event records.

The bigger the data file, the bigger the challenge. Every data event should have a place and purpose within the context of the analysis the data purports to represent, or more correctly the analyst purports to deploy as being relevant.

Credible validation processes can peel back the layers, search for gaps in the data and test the data consistency and variations, looking beyond simple data banding techniques to know and understand what lies within.

Of course, individuals look to be credible and ethical in their approach, but the true test of this is the institutionalised care that is given to the processes.

There are many aspects to identifying errors that can creep in, including:

- Partial record event/time phase shift - where a vehicle starts or finishes outside the recording time.
- Missed record event.
- Directionality assignment in field or post survey.
- Poor data file management.

2.5 Algorithms and dependency on condition tables

This paper, will now focus on the effect of algorithms that try to match number plates where the recording of the plate may have a small error due to the difficulty of recognising similar number and/or letter characters. Character recognition errors occur with both manual human observation and ANPR/LPR systems. Automated systems have the additional problem of multiple plate record generations per vehicle. Extraneous records that may not be identical. The focus of the following review is toward the condition table parameters and less about the algorithm.

2.6 Automatically adjusting for letters and numbers that might look the same

It is not hard to imagine a computer program or a transcriber looking at a video, being ambivalent between what is the letter “O” or the number “0”. But it is not always as simple as that.

Table 1 shows the character alternatives that have been recognised in the past. In some cases, there is only one alternative such as “Z” and “2”, but for the letter “O” there might be eight other different numbers or letters that may be incorrectly recorded as “0”.

Character Alernates			
Digit	Alternates	Digit	Alternates
0	OQ8PU9DC	I	17Y
1	I4L79EGJT	J	51K
2	57RZ	K	JR
3	657	L	1
4	A17UY	M	NHW
5	S283J6F	N	MH
6	G3958	O	Q0CDPU89
7	4231TI	P	O0F8RBDT
8	B5O06RP	Q	O0DC
9	6O01	R	H8P2KB
A	4	S	5F
B	8PR	T	E17P
C	GO0Q	U	VYO04W
D	O0QPE	V	UY
E	T1FGD	W	UM
F	EHP5S	X	Y
G	C6E1	Y	UV4XI
H	RFMN	Z	2
?	[?0123456789ABCDEFGHIJKLMNOPQRST		

Table 1: Control table – All numbers and letters and corresponding digits for which they MAY be considered for substitution.

NPR doesn't fix it all

The technology of number plate recognition is improving, yet errors still occur. To assess the accuracy and remove poor recordings, a visual inspection of each record would also be required.

Figure 1: A Victorian number plate

Figure 1 shows a Victorian number plate that an NPR system (in a car park used by the owner a number of times) recorded incorrectly in one direction. The erroneous recordings listed below are not just a simple mistake such as confusing an "0" with the letter "O".

- M001K
- M001GK
- T001GK

The system would assume that the start of the trip was logged but mysteriously did not finish or the trip mysteriously started in the parking area having never entered in the first place. The owner spent about an hour on each occasion trying to clear up the misunderstanding before he could get his car out. In a road survey, each time this happened the system would record two trips with an internal origin or destination and not completed trips. I have given other examples in another paper (Reid 2018b)

Changes in the trip table with various fuzzy matching techniques

There are many ways and means to affect a rework to enhance the results, which might be said to have some logic, but if you do find additional matches it is reasonable to expect that overall, they are similar in pattern to the verified raw numbers.

For the purposes of this paper data sets were worked up using 5 derivations:

1. Raw Exact data match - No fuzzy logic applied.
2. Residue matches using control table to change 1 character per plate.
3. Residue matches using control table to allow on 2 characters changes per plate.
4. Residue match using no control table allowing a wild, any character match of one character on a plate. This suggests that if a plate is matched in all bar one of the characters, no matter how different the other character is, then it is assumed to be the same plate.
5. Residue match using no control table allowing a wild, any character match of two characters per plate (As will become apparent, a 2-character wild (open) substitution of any plate character match produces too many highly dubious matches to be considered acceptable hence rendering 3 or similar match types as a waste of endeavour).

What does the fuzzy matching look like

Table 2: Examples of changes that are made by various fuzzy matching techniques

ID	Numb	1fuzzy+table	1fuzzy+table	1fuzzy no table	1fuzzy no tab	2fuzzy+table	2fuzzy+table	2fuzzy no table	2fuzzy no table
13469	F56KR							F56FB	digits-4&5 ---(KF)(RB)-
13612	FJ8833							FJ8931	digits-4&6 ---(89)-(31)
13761	ER8055			EN8055	digit-2 -(RN)----			EN8055	digit-2 -(RN)----
13800	ED3280							EO3287	digits-2&6 -(DO)---(07)
14127	F99MG							F98QG	digits-3&4 --(98)(MQ)--
14148	CM2338			EM2338	digit-1 (CE)-----			EM2338	digit-1 (CE)-----
14431	B94LU	B94LQ	digit-5 ----(UQ)-	B94LQ	digit-5 ----(UQ)-	B94LQ	digit-5 ----(UQ)-	B94LQ	digit-5 ----(UQ)-
14477	FM8669			FH8669	digit-2 -(MH)----			FH8669	digit-2 -(MH)----
14478	C67QS							C61QR	digits-3&5 --(71)-(SR)-
14598	E91JF							E92JR	digits-3&5 --(12)-(FR)-
14645	F72TI	F72TL	digit-5 ----(IL)-			F72TL	digit-5 ----(IL)-	F72HE	digits-4&5 ---(TH)(IE)-
14665	FQ0414	FO0414	digit-2 -(QO)----	FO0414	digit-2 -(QO)----	FO0414	digit-2 -(QO)----	FO0414	digit-2 -(QO)----
14879	F22EN			F22NN	digit-4 ---(EN)--			F22NN	digit-4 ---(EN)--
14984	F87AL							F87NI	digits-4&5 ---(AN)(LI)-
15041	D3381							D33HU	digits-4&5 ---(8H)(1U)-
15078	E20EM			E20PM	digit-4 ---(EP)--			E20PM	digit-4 ---(EP)--
15172	D71PW	D71PM	digit-5 ----(WM)-	D71PM	digit-5 ----(WM)-	D71PM	digit-5 ----(WM)-	D71PM	digit-5 ----(WM)-
15182	F37PI			F32PI	digit-3 --(72)---			F32PI	digit-3 --(72)---
15251	EH3872			BH3872	digit-1 (EB)-----			BH3872	digit-1 (EB)-----
15266	A22LJ			D22LJ	digit-1 (AD)-----			D22LJ	digit-1 (AD)-----
15407	E93FL							E93SW	digits-4&5 ---(FS)(LW)-
15418	E08FM					F08EM	digits-1&4 (EF)--(FE)--		
15439	E8565					E85GS	digits-4&5 ---(6G)(5S)-	E85QM	digits-4&5 ---(6Q)(5M)-
15533	A8ZZB	A82ZB	digit-3 --(Z2)---			A82ZB	digit-3 --(Z2)---		
15773	A60BQ	A60BO	digit-5 ----(QO)-			A60BO	digit-5 ----(QO)-		

Table 2 exemplifies how fuzzy matching algorithms can change a recording to achieve a match.

- Line 14645 shows a credible (if the travel time is acceptable) match where "L" in the number plate F72TL could reasonably be read as an "I" and as a number plate with F72TI was recorded then a match could be called. This was captured in all four fuzzy matching techniques.
- Line 14879 shows number plate F22EN being recorded and later a number plate F22NN. It is not traditional to expect that an "E" would be confused with and "N" and so is not "matched" via the control table. But it is matched if you assume that one mismatched digit is OK no matter how different. Another consideration that is not addressed in the fuzzy matching is that with another "N" next to it, there might have been an entry problem. Should this be included?
- Line 13469 has two number plates that have two different digits; F56KR and F56FB. In the last fuzzy matching process, they are assumed to be the same plate. But there is a means to further distinguish an acceptable/unacceptable fuzzy match using travel time statistical considerations.

3. The impact of fuzzy logic in a specific example

3.1 The survey

Austraffic conducted an O/D survey which collected 40,597 registration plate records at 4 locations. Each location was surveyed in both directions, in two 4-hour periods, to cover the AM and PM peak times. The result was a data capture rate of 95.97% with an open road network.

The data was entered into a contiguous data file to facilitate return matching vehicles. Therefore, there was no opportunity to reconcile all events (as with a closed network or data capture over the entire day) which highlights limitations with data interpretation.

3.2 The first sweep through the data

A quick look at the capture rates can identify where to check the data for erroneous entries. The project data set had been checked for duplicate entries prior to producing table 3. The remaining errors that give rise to an imperfect capture rate are likely to have been a masked plate record (due to sun glare, dirt, bike rack, tow ball, trailer) or a human error of misreading the characters as the vehicle passed by at highway speed (as opposed to being viewed via a video record, that could have been paused and reviewed multiple times).

Table 3: Capture rate at 15-minute intervals – morning period – stations 10 and 13

	Stations										
	10N			10S			13N			13S	
Time	All Plates	Good Plates	Capture Rate	All Plates	Good Plates	Capture Rate	All Plates	Good Plates	Capture Rate	All Plates	Good Plates
06:00	67	64	97.25%	92	91	99.40%	10	10	100.00%	24	24
06:15	96	94	98.70%	90	89	99.40%	10	10	100.00%	18	18
06:30	126	124	99.15%	170	169	99.70%	19	19	100.00%	46	44
06:45	136	135	99.60%	173	168	98.50%	36	36	100.00%	48	48
07:00	113	113	100.00%	175	172	99.10%	31	30	97.90%	40	40
07:15	179	179	100.00%	209	206	96.80%	32	32	100.00%	54	53
07:30	189	187	99.45%	288	286	99.65%	55	55	100.00%	88	87
07:45	270	269	99.80%	300	297	99.45%	77	75	98.60%	123	123
08:00	253	251	99.55%	367	363	99.40%	95	94	99.45%	131	130
08:15	303	301	99.65%	373	369	99.45%	138	137	99.60%	165	163
08:30	356	356	100.00%	416	411	99.35%	146	146	100.00%	209	209
08:45	335	331	92.00%	291	291	100.00%	194	190	98.90%	143	143
09:00	262	259	96.65%	236	233	99.30%	105	102	98.45%	116	115
09:15	223	220	97.10%	220	214	98.50%	107	106	99.50%	101	101
09:30	180	178	95.55%	234	227	94.80%	103	102	99.50%	108	107
09:45	183	183	100.00%	194	186	97.80%	107	106	99.50%	102	101
10:00	188	187	99.70%	205	199	98.45%	107	107	100.00%	97	97
10:15	182	182	100.00%	216	204	93.15%	109	109	100.00%	106	106

A detailed review showed that of the 40,597 plates, 39,145 were clearly identified, which is a very good sample to interpolate into a meaningful trip assignment table.

For this survey the data file was reviewed to be clean of multiple vehicle entries (sometimes a recording is duplicated as can be identified as two records having the same vehicle at the same time).

3.3 The raw data trip table

Having started with 39,145 plates, and the clean raw matches accounting for (2 x 6168) 12,336 plates, we had Residue Unmatched vehicle events totalling 13,093 inbound and 13,726 outbound plates. The inbound and outbound numbers are not the same as some vehicles could have been inside the area before the survey period started or remained within the area after the survey period had ended. Also, the survey area was not a closed cordon, so while the major roads were recorded there were other minor access points.

The trip table of matched trips from the raw data is shown in table 4 below.

Table 4: Matched trips from the raw data

Table of Matched trips-					
Row Labels	**10N**	**13N**	**15E**	**22S**	**Grand Total**
10S	1441	366	101	610	2518
13S	171	647	74	47	939
15W	114	111	606	55	886
22N	739	114	65	907	1825
Grand Total	**2465**	**1238**	**846**	**1619**	**6168**

Now let's start trying to apply some fuzzy logic to see if we can capture extra plates that could be reasonably assumed to be the same but were not recorded as such.

3.4 Results from different fuzzy matching approaches

Each of the following tables gives three sets of numbers:

1. The additional matches of "through" trips in each element of the trip table arising from fuzzy matching; the total for each station and the overall total.
2. The percentage of all through trips between each station pair.
3. The percentage change in the percentage of trips in between each station pair, with respect to the base line raw exact data matches.

Table 5: Added matches using a control table on 1-character substitution per plate

1fuzzy Character Adjustment using Control Table - Additional vehicle records matched

Count of TravelTime	Colum	abels				
Row Labels	10N	13N	15E	22S	Grand Total	
10S	64	16	7	45	132	
13S	7	12	3	5	27	
15W	3	2	12	1	18	
22N	27	7	5	57	96	
Grand Total	101	37	27	108	273	Additional Vehicles

	10N	13N	15E	22S
10S	23%	6%	3%	16%
13S	3%	4%	1%	2%
15W	1%	1%	4%	0%
22N	10%	3%	2%	21%

Variation trip assignment 1 Fuzz CT Substitution

	10N	13N	15E	22S
10S	0%	-1%	57%	67%
13S	-8%	-58%	-8%	140%
15W	-41%	-59%	-55%	-59%
22N	-17%	39%	74%	42%

While the total number of trips increases by only 273 vehicles (4.4%) there are changes in the percentage of trips in each element of the table by significant amounts. For example, trips from 10S to 22S were 10% (1 in 10) of the total, now they are 16% of the total (1 in 6). A 67% increase.

Table 6: Added matches using 1 fuzzy character un restrained substitution – i.e. No Control Table

1fuzzy - Wild Character- No Control Table - Additional vehicle records matched

Count of TravelTime	Colum	abels				
Row Labels	10N	13N	15E	22S	Grand Total	
10S	142	71	30	129	372	
13S	39	37	12	25	113	
15W	17	18	25	16	76	
22N	108	36	22	106	272	
Grand Total	306	162	89	276	833	Additional Vehicles

	10N	13N	15E	22S
10S	17%	9%	4%	15%
13S	5%	4%	1%	3%
15W	2%	2%	3%	2%
22N	13%	4%	3%	13%

Variation trip assignment 1 Fuzz Unrestrained

	10N	13N	15E	22S
10S	-27%	44%	120%	57%
13S	69%	-58%	20%	294%
15W	10%	20%	-69%	115%
22N	8%	134%	151%	-13%

Here in table 6, the additional matched trips are more significant at 833 (being an increase of 13.5% of the original total matches) but again the split in the percentage of total trips into each element has shifted significantly.

The matched trip from 13S to 22S were 1% but are now 3%. While this represents small numbers, it is a change in emphasis. More significantly trips in and out of station 10 have shifted from being 23% of the total to 17% of the total.

Table 7: Added matches using a control table on 2 characters substituted per plate

2fuzzy Character Adjustment using Control Table - Additional vehicle records matched

Count of TravelTime	Column Labels					
Row Labels	10N	13N	15E	22S	Grand Total	
10S	82	20	11	63	176	
13S	11	13	4	8	36	
15W	3	3	13	1	20	
22N	36	12	7	74	129	
Grand Total	132	48	35	146	361	Additional Vehicles

	10N	13N	15E	22S
10S	23%	6%	3%	17%
13S	3%	4%	1%	2%
15W	1%	1%	4%	0%
22N	10%	3%	2%	20%

Variation trip assignment 2 Fuzz CT substitution

	10N	13N	15E	22S
10S	-3%	-7%	86%	76%
13S	10%	-66%	-8%	191%
15W	-55%	-54%	-63%	-69%
22N	-17%	80%	84%	39%

Perhaps the most notable factor here is that the increase in matched trip of 361 (5.8%) is up on raw data but it is not nearly as much as on the using one character with uncontrolled fuzzy matches.

Table 8: Added matches using control table on 2 fuzzy un restrained

2 Fuzzy- Wild Character -No Control Table - Additional vehicle records matched

Count of TravelTime	Column Labels					
Row Labels	10N	13N	15E	22S	Grand Total	
10S	236	128	57	208	629	
13S	68	54	17	54	193	
15W	33	44	30	27	134	
22N	190	67	32	147	436	
Grand Total	527	293	136	436	1392	Additional Vehicles

	10N	13N	15E	22S
10S	17%	9%	4%	15%
13S	5%	4%	1%	4%
15W	2%	3%	2%	2%
22N	14%	5%	2%	11%

Variation trip assignment 2 Fuzz Unrestrained

	10N-OutBou	N-OutBou	E-OutBou	2S-OutBound
10S-Inbou	-27%	55%	150%	51%
13S-InBou	76%	-63%	2%	409%
15W-InBou	28%	76%	-78%	118%
22N_Inbou	14%	160%	118%	-28%

The use of statistics for this single dimension of trip assignment is compelling, but when we add the second dimension of duration of travel time for such forced fuzzy matches the analysis clearly shows the pursuit of an additional 1392 matched events will grossly distort the statistical profile of those additional matches as compared with the baseline exact raw data matches.

3.5 Travel time a potent aide for validation

Travel time analysis can assist with the validation of data event matching, but can also provide an insight as to a likely trip purpose within in an interpretive broad context as distinct to simple through traffic determination predicated on a defined match time of say 30 minutes

The travel times that can be established from the data can be used for more than just averaging the results. If a very long travel time is recorded then the vehicle is likely to have stopped which is a different trip to a continuous one (which for example, might be one likely to use a by-pass).

An open match time will include all vehicle matches throughout the whole survey duration and will be a very different set of numbers to a shorter match time that may be more reflective of the through traffic assignment.

A point of qualification is to highlight the significant role of a match time can have within a matching algorithm. This parameter can influence the sequencing of plate pair matching and produce a varying result dependent upon the logic and value of such parameter- A subject worthy of its own paper. Reinforcing the need to understand and question “what treatment has been applied to my data ?”

Statistical averages alone can mask the data problem. Following the theme of peeling back the layers three statistical measures were deployed. The last thing any data set review with meaning needs is another average that masks dud data, left within a data set or erroneous distortions from an over working of the logic to recover optimum matches. The impact could be the creation of an

even less credible average that will pass unnoticed. Hence this analytical review utilised three simple statistical measures as follows:

In table 9 the travel time results are shown in three measures:

1. The average travel times
2. The 85th percentile travel time
3. Standard deviation

Table 9: Travel times for the raw data and each fuzzy logic matching conditioned option.

Exact

Count of T	Column Labels														
	10N			13N			15E			22S			Grand Total		
Row Labe	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev
10S	4:41:49	9:27:58	4:09:09	3:48:28	9:01:22	3:51:52	1:54:59	5:52:05	3:00:46	0:25:52	0:22:28	1:12:30	3:25:22	9:06:31	3:59:45
13S	4:39:28	9:03:08	3:53:13	3:21:06	8:55:32	3:58:15	2:36:37	7:47:34	3:29:27	2:12:00	6:57:13	3:05:29	3:28:25	8:54:29	3:55:55
15W	1:20:30	1:55:00	2:17:30	1:50:17	4:38:28	2:53:37	2:54:20	8:50:08	3:40:44	0:39:44	0:48:03	1:23:41	2:25:53	8:26:02	3:24:46
22N	0:25:58	0:15:51	1:15:57	1:42:49	2:09:28	2:47:18	1:05:46	1:07:16	2:22:21	5:37:38	9:37:06	4:00:35	3:07:05	9:03:41	3:57:18
Grand Tot	3:15:38	9:03:51	3:58:12	3:12:00	8:49:29	3:49:22	2:37:22	8:41:51	3:32:45	3:24:05	9:08:36	4:02:13	3:11:52	9:00:32	3:54:38

1fuzzy+table-fuzzy only

Count of T	Column Labels														
	10N			13N			15E			22S			Grand Total		
Row Labe	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev
10S	3:58:39	9:20:16	4:00:57	3:38:11	7:20:43	3:26:58	4:20:09	8:10:59	3:28:46	0:43:23	1:23:57	1:24:20	2:50:45	8:40:20	3:35:44
13S	2:23:07	5:30:29	3:16:39	4:13:16	8:36:10	3:56:13	4:41:05	6:48:52	2:40:30	3:33:41	7:55:39	3:36:34	3:40:28	8:12:09	3:40:57
15W	0:16:46	0:23:40	0:08:05	0:38:33	0:50:05	0:16:29	3:33:06	9:24:20	4:11:30	10:05:00	10:05:00	0:00:00	3:02:45	9:29:04	4:04:12
22N	1:49:22	7:17:03	3:20:51	2:15:37	5:41:15	2:58:13	3:41:53	7:40:43	4:02:49	4:59:53	9:23:16	3:58:55	3:50:16	8:57:32	4:01:24
Grand Tot	3:10:53	9:03:47	3:53:32	3:24:14	8:11:19	3:34:12	3:54:29	8:55:19	3:52:02	3:11:51	8:48:41	3:48:29	3:17:23	8:52:43	3:49:13

1fuzzy no table-fuzzy only

Count of T	Column Labels														
	10N			13N			15E			22S			Grand Total		
Row Labe	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev
10S	3:00:11	7:57:18	3:29:49	2:57:00	6:11:52	2:51:26	1:58:37	5:36:06	2:28:15	1:27:21	4:29:40	2:17:38	2:22:25	6:30:49	3:00:37
13S	3:16:54	7:11:30	3:07:45	3:41:20	8:21:00	3:36:50	3:54:55	6:27:31	3:06:42	2:14:00	5:48:52	2:49:50	3:15:01	7:05:46	3:17:09
15W	2:47:13	6:04:28	2:52:45	3:24:21	7:43:21	3:40:15	2:09:40	6:29:08	3:03:21	3:20:59	7:27:05	3:21:18	2:50:46	6:54:32	3:16:54
22N	2:02:30	5:46:04	2:40:06	2:56:08	6:39:13	2:55:58	4:07:18	7:57:14	3:04:34	3:26:43	8:49:32	3:33:34	2:52:31	7:11:28	3:11:36
Grand Tot	2:41:14	7:10:01	3:11:01	3:09:59	7:04:51	3:10:24	2:49:13	6:32:16	3:02:40	2:24:00	6:52:29	3:05:51	2:41:58	6:57:18	3:09:01

2fuzzy+table-fuzzy only

Count of T	Column Labels														
	10N			13N			15E			22S			Grand Total		
Row Labe	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev
10S	4:10:40	9:18:55	3:59:01	3:40:43	7:41:07	3:30:32	5:00:57	8:38:17	3:38:10	1:03:21	1:51:42	1:55:59	3:03:21	8:43:16	3:38:29
13S	4:01:26	9:36:59	4:08:12	4:02:25	8:28:31	3:50:02	5:21:37	7:19:11	2:35:43	4:00:58	8:01:10	3:33:36	4:10:36	8:16:37	3:46:42
15W	0:16:46	0:23:40	0:08:05	1:07:28	1:44:13	0:43:03	3:17:30	9:17:14	4:07:36	10:05:00	10:05:00	0:00:00	2:51:16	9:14:52	3:54:55
22N	1:54:14	8:05:13	3:22:59	2:41:04	7:14:24	3:12:30	4:10:56	9:12:43	4:00:36	5:02:50	9:30:16	3:57:41	3:54:12	9:01:14	4:00:15
Grand Tot	3:27:22	9:03:49	3:57:18	3:22:06	8:10:52	3:30:31	4:14:53	9:01:26	3:53:25	3:18:11	8:48:58	3:47:05	3:27:33	8:53:45	3:49:57

2fuzzy no table-fuzzy only

Count of T	Column Labels														
	10N			13N			15E			22S			Grand Total		
Row Labe	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev	Avg	85%ile	StdDev
10S	2:05:12	6:14:41	3:06:07	2:01:27	5:41:24	2:38:21	1:37:35	4:56:46	2:25:43	1:16:53	3:30:36	2:10:54	1:45:57	5:17:56	2:41:45
13S	2:28:08	6:06:41	3:01:28	2:50:11	6:32:20	3:24:35	3:06:16	6:16:56	3:07:56	1:54:57	4:56:31	2:33:13	2:28:23	6:09:08	3:03:12
15W	1:52:46	5:14:55	2:35:33	2:08:08	5:55:45	2:59:30	1:50:34	6:02:17	2:52:47	2:26:10	6:30:02	3:01:30	2:04:03	6:06:32	2:53:18
22N	1:22:41	4:41:35	2:16:04	1:47:16	5:35:11	2:37:03	3:01:53	6:41:51	3:09:17	2:49:35	8:24:01	3:26:24	2:03:02	5:58:13	2:54:53
Grand Tot	1:52:03	5:22:31	2:48:50	2:08:11	6:00:11	2:52:01	2:11:22	6:04:15	2:52:51	1:57:09	5:36:30	2:51:00	1:58:56	5:45:05	2:50:44

As with the results from matched trips, the variations are very significant and drastically affect the interpretation of the survey results. The more the fuzzy matching constraint is relaxed, the bigger the difference in the data statistics from that of the clean RAW matches. In terms of travel times the additional matches produce variations in both increased and decreased travel times, extremes at either end of the time durations between matched stations. The 4 fuzzy logic conditions tested clearly show that only a control substitution table should be considered for deployment and any attempt to run unrestrained character substitutions should not be applied.

Another form of presentation for the travel time statistics is shown in the following tables. In Figure 2 we compare three travel time statistics (average, 85th percentile and Standard deviation) for the raw data and the most extreme fuzzy matching we have considered – changing any two digits no matter how dissimilar. Within each encircled pair of stacked bars, ideally the colours and heights of each pair should be similar. They are not.

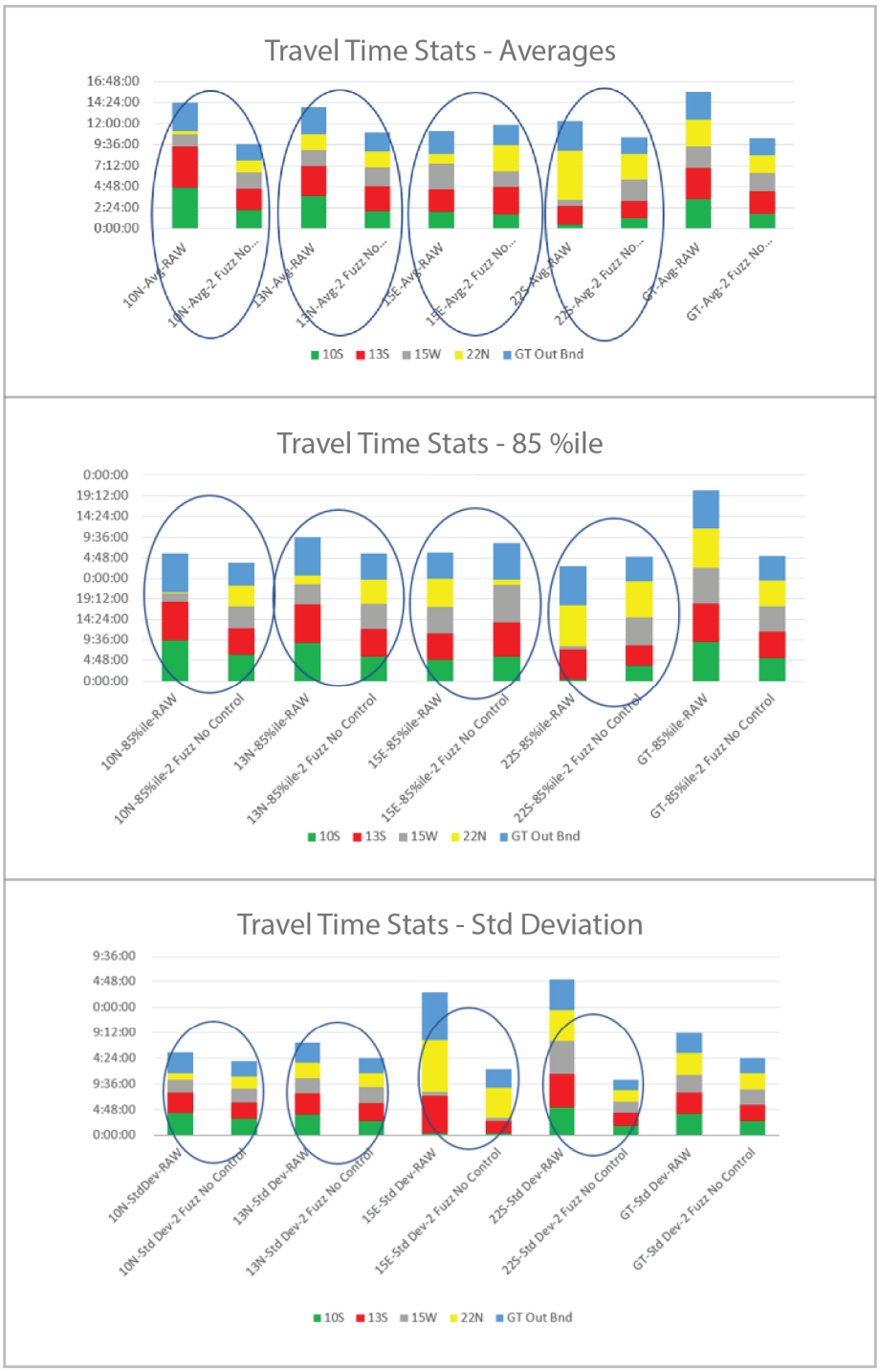

Figure 2: The difference between raw data and uncontrolled fuzzy matching of two digits in travel time statistics

The following charts show how the overall pattern of travel times is affected with fuzzy matching.

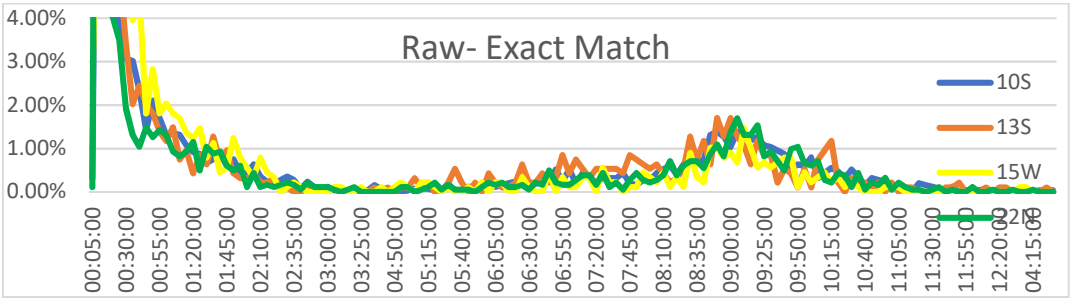

Figure 3: Percentage of trips for the range of travel times from the raw data – Base Line

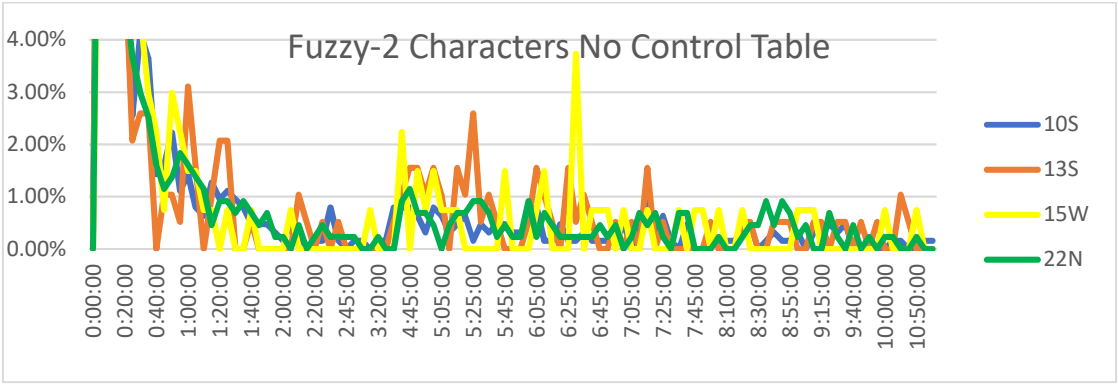

Figure 4: Percentage of trips for the range of travel times from the 2-character fuzzy match with no control table.

The raw data appears to have a good pattern while the fuzzy matched data appears to have many more “outliers” (records that appear to be beyond base line variations).

It is interesting to look at some individual records in table 11 to make this point.

- ED4?56 In@ Stn 22N may look like a nice match for ED4366 Out@Stn10N, but not with a 6 min 29sec travel time when the raw match had an average of 25 min 51 sec With a travel time discrepancy of 19 min 32 sec this fuzzy matching fails the pub test.
- A23WQ (In@13S) should never be substituted as a pair with A21WV (Out22S) with a ridiculous travel time of 1 min 55 sec.

Table 11: Examples of fuzzy matches with unrealistic travel times

Inbound Vehicle		:bound Vehicle		Travel	Substitution	
Plate	Station	Plate	Station	Duration	Position	Characters
EM5377	22N	EM5967	10N	0:00:08	digits-4&5	---(39)(76)-
E39DO	22N	E31PO	10N	0:00:12	digits-3&4	--(91)(DP)--
E63DG	22N	F63QG	10N	0:00:13	digits-1&4	(EF)--(DQ)--
ED4?56	22N	ED4366	10N	0:06:29	digits-4&5	---(?3)(56)-
F75PW	22N	F75HG	10N	0:00:20	digits-4&5	---(PH)(WG)-
D76RD	22N	D83RD	10N	0:00:32	digits-2&3	-(78)(63)---
D43PH	22N	D43ME	10N	0:00:40	digits-4&5	---(PM)(HE)-
FP5068	22N	FL9068	10N	0:00:42	digits-2&3	-(PL)(59)---
A63SK	22N	A68SR	10N	0:00:49	digits-3&5	--(38)-(KR)-
A23WQ	13S	A21WV	22S	0:01:55	digits-3&5	--(31)-(QV)-
F79PL	22N	A79PL	10N	0:00:55	digit-1	(FA)-----
A48AW	22N	A48TO	10N	0:00:55	digits-4&5	---(AT)(WO)-
F10CG	22N	F10QZ	10N	0:00:55	digits-4&5	---(CQ)(GZ)-
C49YU	22N	C49TV	10N	0:01:41	digits-4&5	---(YT)(UV)-
E52TZ	22N	A22TZ	10N	0:01:41	digits-1&2	(EA)(52)----
D40FG	22N	D40KT	10N	0:02:07	digits-4&5	---(FK)(GT)-
C67YK	22N	C67AK	10N	0:02:35	digit-4	---(YA)--
B67GW	22N	B67SL	10N	0:02:48	digits-4&5	---(GS)(WL)-
E91RF	22N	E91EI	10N	0:03:12	digits-4&5	---(RE)(FI)-
F36TQ	22N	F31TV	10N	0:03:29	digits-3&5	--(61)-(QV)-
B72WC	22N	F72AC	10N	0:04:19	digits-1&4	(BF)--(WA)--
F64LX	22N	F64WY	10N	0:04:22	digits-4&5	---(LW)(XY)-
F93TH	22N	F53PH	10N	0:04:22	digits-2&4	-(95)-(TP)--
E69NI	22N	E69SS	10N	0:04:36	digits-4&5	---(NS)(IS)-
C67QS	22N	C61QR	10N	0:04:37	digits-3&5	--(71)-(SR)-
B72HE	22N	B32HL	10N	0:05:00	digits-2&5	-(73)--(EL)-
F72TI	22N	F72HE	10N	0:05:04	digits-4&5	---(TH)(IE)-
A03ND	22N	A00NG	10N	0:05:08	digits-3&5	--(30)-(DG)-
EM5257	22N	EN5257	10N	0:05:27	digit-2	-(MN)----
ED49CX	22N	ED4901	10N	0:05:29	digits-5&6	----(C0)(X1)
C53LQ	22N	C53LO	10N	0:05:45	digit-5	----(QO)-
EQ2536	22N	ER2576	10N	0:05:55	digits-2&5	-(QR)--(37)-

Table 11 shows more records that don't make sense and unworthy to be acceptable matches. These plates should not be matched as the trips have unrealistic travel durations, in some case they are downright ridiculous.

These examples alone highlight the impact of the single dimension of travel duration as a means to accept or reject a plate character substitution.

3.6 Impact on aggregated data

The next step for the data workup was pointed toward a review of the statistical impact of data aggregation.

The segregated data from each of the conditional character substitutions were aggregated individually with the initial good raw data, to assess the impact of the fuzzy matches.

A charted summary of 2 conditions are shown in Fig 5 & 6 below depict the Variation of the trip distributions with respect to the statistics of the exact raw match data.

The use of fuzzy logic to try and identify extra matched trips, would reasonably expect the additions to be similar in the assignment distribution as per the "real" matched trips. Figures show the impact of aggregation of trips with fuzzy matching of two digits via a control table resulted in only small variations but for "wild" matches of any two digits, the variations ranged between -14% to 75%.

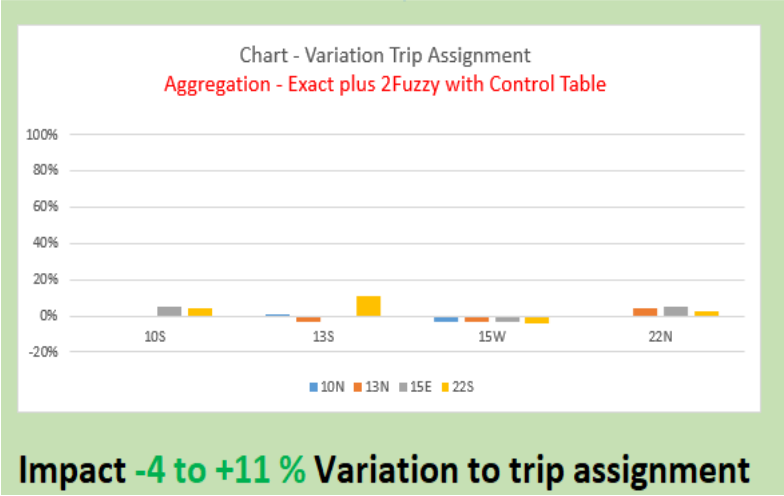

Figure 5: Variation of trip assignments including fuzzy matching of two digits with a control table.

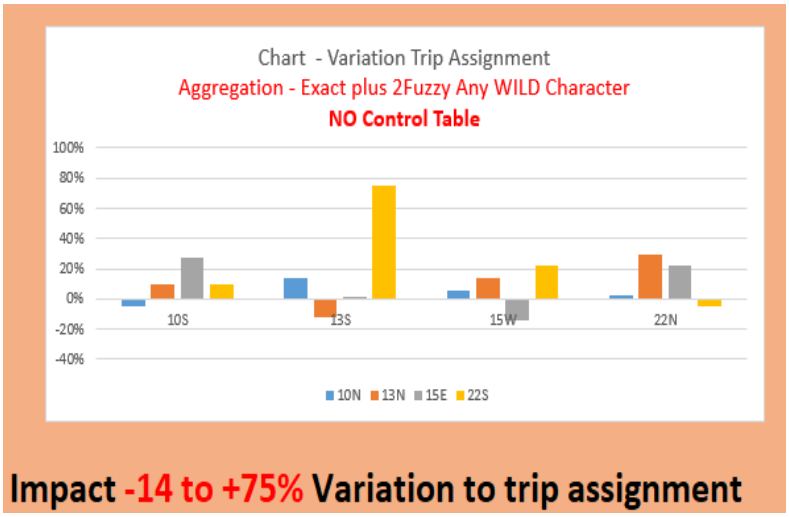

Figure 6: **Variation of trip assignments including fuzzy matching of two digits with any characters.**

4. What should you do to attain an appropriate data result?

Summary:

- View data as an onion:
  - Peel back the layers of any data set.
  - Statistically validate each data layer, where possible.
  - Avoid aggregated data averaging as an input or validation check.
  - Don't aggregate a good data set with a cleansed or filtered date set until both are deemed credibly worthy.
- Smart use of statistical testing the data. Segregate the good data from suspicious or known errors and compare the statistics. Recovered matches should be analysed for statistical reasonable consistency.
- Know the condition of your RAW data, e.g. sight data capture and accuracy rates over say 15 minutes increments throughout that data set duration.
- Sight a RAW data match table and travel time distributions, aggregated per stations of origin and then per station of destination as matched.
- Sight a listing of character adjustment tables.
- Sight a listing of all plate alternates as then fuzzy logic matched, correlated with time of day, station of origin and assigned station of destination, with travel time included.
- Origin-Destination trip tables should have clear transparency as to what treatment has been applied to the final data product.

- To which limited and clearly defined use of character substitution should be applied to force optimising the residue plate matches.

5. The pressures to have a 'complete' answer

There can be professional pressure to maximise the numbers within a trip table matrix. It is a natural expectation that the raw numbers will default to a result that is to understate the match rate, due by direct correlation to the fact that a plate record error will prevent plates being matched.

An O/D trip assignment will be configured, predicated on the results of an O/D survey with significant consequences if the numbers have been over corrected and changed the characteristics of the assigned trips and correlated durations of travel.

6. Conclusions

- Advanced systems of data collection do not guarantee 100% accuracy. In fact, they can create errors that are often buried under a huge number of recordings.
- Data quality is an issue not just for individual survey companies. A respect for the complexity of collection and processing of data must be driven by institutionalised care for credible results.
- Students and professionals should be taught about the history and the processes of data collection to understand how numbers are arrived at and the problems that may arise.
- It is often said that there are three factors in providing a service: A quick turnaround, a cheap price and a quality product; but you can only ever have two at once. This is a reality that our profession seems to have lost.
- Even with the best processes and the latest technology you do not get a 100% sample and 100% accuracy. We need to be testing just how good some results are so that we can deal effectively with the results that are produced.
- Quality data is no accident. Quality data is the product of a quality process, with adequate resources and doses of integrity in the application.
- The transport planning and management profession should look to independent quality organisations to help develop standards for data collection and processing.
- The underlying theme of data, its validation, the filtering and blending, requires diligence and transparency.

7. References

1. AITPM (2017a) – Proceeds of the 2017 AITPM National Conference – Melbourne 15-17 August 2018
2. AITPM (2017b). Video news story "Bridging the gap between data and wisdom" - https://www.youtube.com/watch?v=odARKEOZO9Y&t=1s
3. Ampt, L., (2017), Personal email commenting on the ISCTSC (2017) Program from "11th International Conference on Transport Survey Methods" Esterel Canada, September 2017,
4. ASQ undated: ISO 9000 Quality Management Principles. http://asq.org/learn-about-quality/iso-9000/overview/quality-management-principles.html
5. Benjamin, S., (2013). "Real time and historic data collection in a blender" Proceeding of the AITPM National Conference: Perth 30 - July to 2 August and IPENZ Transportation Group Conference, Shed 6, Wellington – 23 – 26 March, 2014
6. Brown (2018): AITPM April 2018 Newsletter – Editor's Reflections https://us8.campaign-archive.com/?u=c7d03ee5f6283f4eaeba1f983&id=0182009af4#Fees
7. Delbosc, Dr A, Young, Prof W., (editors) "Traffic Engineering and Management (Edn 2017) Chapter 31 "Traffic Surveys" "Traffic Engineering Methods" Text Book, Monash University 2017 ISBN 978-0-6481898-0-0 https://www.monash.edu/engineering/its/publications/tem2017
8. Recode (2017): "These University of Washington professors are teaching a course on bullshit" https://www.recode.net/2017/2/19/14660236/big-data-bullshit-college-course-university-washington
9. Reid, J., (2017a), "Where is the wisdom we have lost in big data" – Proceeds of the AITPM 2017 National Conference, Melbourne 15-17 August 2018
10. Reid, J. (2017b). Speech to the ITEANZ President's Dinner 15 November 2017.
11. Reid, J., (2018a) "Big Data – Size is not everything". Proceeds of the Transportation Group NZ 2018 conference, Queenstown 21 – 23 March 2018
12. Reid, J. (2018b). Personal memory of a question from the floor during a presentation at the Transportation Group NZ 2018 conference, Queenstown 21 – 23 March 2018
13. Shakespeare, W. (c1591). Romeo and Juliet - Act III, Scene 1. Publisher: Penguin Books Ltd ISBN10 0141396474 - ISBN13 9780141396477

Acknowledgements

Yihui Zhang, Principal Analyst Austraffic**;** M.Eng- Distributed Computing - Melbourne University