HARD-DATA IN A WORLD OF TRUTHINESS

Richard Young Beca Ltd

ABSTRACT

Truthiness - the degree to which information appears to be true irrespective of its actual accuracy.

As a profession we now have access to unprecedented volumes of information relating to every aspect of journeys and people's movements.

We are being tasked by clients to provide ever more rigorous justifications to demonstrate why any particular intervention is going to produce the most effective solution.

Whilst there is little doubt that many new data sources can provide fresh insights into the world of transportation there is an increasing trend to set aside data sources that were once considered as 'gold standards' and replace them with opaque data surrounded by unanswered questions.

Added to this are the relentless pressure to deliver results fast, more accurately and more cost effectively than our competitors.

This paper will explore several data sources, both old and new and scratch below the glossy hype of “Big Data” to propose a set of criteria that will enable the New Zealand Transportation Industry to apply some commonly accepted tests around data quality.

INTRODUTION AND THE ETHICAL OBLGATION

As a profession within New Zealand, Professional Transportation Engineers are entrusted to provide advice that is impartial and accurate. When we work under the auspices of Engineering New Zealand we are mandated to comply with the following Ethical Obligations (Obligations, Engineering New Zealand, 2016):-

- Possess sound engineering knowledge applied with skill, diligence and care.
- Keeping engineering knowledge up to date through structured learning.
- Understanding our limits of competence.
- Accepting personal responsibility for the work we do.
- Ensuring that we do not misrepresent our areas or levels of experience or competence.

The procurement, capture, processing, analysis, utilisation and interpretation of traffic data to provide impartial and accurate advice rests firmly within these Obligations. Increasingly, clients are reliant on Professional Transportation Engineers to navigate them through evermore complex environments where current traffic data is being relied upon as key inputs into models and forecasts that in their turn influence investment decisions on projects that not only will cost many millions of dollars, but affect the environment for decades to come.

This paper explores this challenge and by way of some examples develops a structure for Professional Transportation Engineers to fulfil their Obligations in this area.

TRUTHINESS

For the avoidance of doubt or confusion ‘truthiness’ as coined by Stephen Colbert in his seminal broadcast presentation (Colbert, 2005) and is being taken to mean…

> “the belief or assertion that a particular statement is true based on the intuition or perceptions of some individual or individuals, without regard to evidence, logic, intellectual examination, or facts.”

As professionals, we need to be able to scrape away ‘truthiness’ and determine what lies beneath the all-pervading glossy packaging of many of the traffic data products that are in the marketplace and identify what is reliable and what we can professionally rely on to inform the decisions and advice we provide to our Clients.

THE IDEAL TRAFFIC DATA SET

In an ideal world we would have traffic intelligence that covers the duration, route, occupancy, mode, purpose and timing of every vehicle on the road, plus a multitude of other factors. This would cover the whole of New Zealand and be continually up to date and 100% accurate.

Whilst being a perfect data set it is guaranteed to fail to gain a social licence to exist and additionally fall foul of the Privacy Act (New Zealand Government, 1993). There are few, if any democratic countries that would aspire to gather this level of insight into their population.

By necessity we are therefore required to use sampling techniques to gather statistically meaningful data sets that are appropriate for the work we perform for our Clients.

Lies, Damn Lies and Statistics

To be a Professional Transportation Engineer and meet one’s Obligations will undoubtedly require a reasonable understanding of statistics; although it is not the purpose of this paper to provide a 101 lesson on statistical analysis there is some merit in setting out some fundamental concepts.

Concept 1 – You don't need to measure every vehicle
Whilst a 100% sample rate may appear to be ideal, the time and cost to gather it is seldom justified. Added to this there is always the likelihood that even with a 100% sample rate a proportion of that data will be miscounted or misclassified.

In this juncture, the metrics of Margin of Error and Confidence Level are of paramount importance when sampling data, it is taken as read that the reader is familiar with these standard definitions. By way of example, **Error! Reference source not found.** provides the sample size required to deliver a Representative Sample from a Population of 10,000 (such as a road with and AADT of 10,000 vpd). A sample rate of **16**% is sufficient to deliver a Representative Sample with a Margin of Error of 3% and a Confidence Level of 99%. This level of sampling and reliability should certainly be the minimum targeted where a client will be reliant on traffic data for investment decisions.

	Margin of Error				
Confidence Level Size	1%	2%	3%	4%	5%
95%	4,900	1,937	965	567	370
99%	6,240	2,932	**1,557** **(16%)**	940	623

Table 1 Sample Size to deliver a representative sample from a Population of 10,000 (based on classical statistics and the Normal (Gaussian) Distribution Theory)

Where technologies commonly used for traffic surveys such as ANPR offer sample rates of 90% (rarely higher) (Reid, n.d. – presentation at Conference) and a reading accuracy of 90% these could combine to give a data sample of >80% as different vehicles are detected at the start and end. But increasing the sample rate from 16% to 80% may not materially improve the accuracy of the gathered data. The author has seen contracts requiring 85% sample rates but lacking any understanding of the additional costs required to provide this even though there is no significant increase in data quality.

When smaller representative samples can be obtained by less obtrusive and more cost effective technologies there needs to be an assessment of value for money and efficiency.

Concept 2 – Avoiding Sample Bias
Where a Representative Sample is being sought by sampling a smaller numbers of vehicles some care needs to be exercised to demonstrate that there is no significant bias in the vehicles being sampled. An effective and pragmatic method of testing for Sample Bias is to repeat the sampling at a range of locations and times of day. For example with if similar sample rates are obtained on freight dominated routes (nights on key State Highways), hire car routes (Central Otago) and busy commuter routes then this is a reasonable indication that the sampling technique is capturing a Representative Sample. It is not a fool-proof approach but it does demonstrate a high degree of consistency across the vehicle fleet.

BIG DATA – BIGGER ISSUES

The ability to utilise data that has been gathered for one purpose but can be commercially packaged for another purpose has significant attractions. Collecting data is often expensive and companies that can exploit the commercial value of this data often will do so.

Limits of Use of Commercial Data
Understandably, commercial organisations selling data will often place stringent limitations around how data can be used and how long it can be retained.

For example, one in-car navigation company (Waterfield, 2011) harvested speed data from their subscribers, data which it then sold onto the national Dutch police, who in turn used that data to site speed cameras to catch subscribers. The company's resultant apology to their subscribers acknowledging that they had breached their subscribers' trust also came with a commitment that their information would never be handed over the police again.

With today's emphasis in New Zealand on Safe Roads and the Safe System Approach the ability to use these data sources to highlight areas of heavy braking is undoubtedly assisting the focusing of resources, but is it acceptable that this data not also be used to target speed compliance?

Big is not always Better, We Have Your Number….

There are undoubtedly benefits in being able to draw on truly anonymised data streams from telecommunications providers to capture large scale movements of people around cities.

To briefly summarise this data, most telecommunication operators (telcos) manage the radio traffic between mobile devices and one (or several) nearby mobile phone base stations (towers / masts). A teleco can track a moving phone as it hands-over between nearby base stations, but often they have no access to the data that is being transmitted. Embedded within that radio traffic (and not generally accessible to the teleco) is data that may include GPS locations, speed etc.

This area-to-area hand-over movement is readily captured as mobile devices move between a succession of cell-phone base stations, even if no call is being made. There is some degree of modal data as some base stations are located at public transport hubs and specifically designed to cover small areas. This is undoubtedly useful data.

Beyond this macro level data, there are techniques to estimate the route taken by that device based on the sequence of base-stations used and the strength of the radio signal received. The Professional Transportation Engineer does needs to understand that this is often 'second order' data reliant on inferred locations rather than actual data on route and location. This data can thereby come with caveats around positional accuracy.

There can also be specific issues when applying this type of technique to dense urban areas, more base stations do not always mean better positional resolution as radio signals are more affected by buildings in urban areas.

The Ubiquitous Smartphone GPS and those Global Players

Most (if not all) smartphones have Global Positioning System (GPS, US Government, 2018) or similar positioning system installed enabling the devices position to be determined to an accuracy of better than 10m. More accurate positioning is possible, as are situations which degrade accuracy – both are outside the scope of this paper.

This location data can be shared by the device with any location based application active on the phone which provides the application provider with a continuous stream of location data for the phone. Whilst there is a commonly held belief that these individual 'snail trails' of data are available for purchase we have found no evidence that this is correct.

What is commonly available are summaries of segmented and aggregated route data. Often any quantitative information on sample size, accuracy, collection period is unavailable. Whilst the Professional Transportation Engineer can utilise this data they also need to exercise their professional Obligations in doing so.

THE NZ PRIVACY ACT

Within New Zealand we have the Privacy Act (New Zealand Government, 1993) which sets out the Principles which determine what data can, and more importantly cannot, be collected without the consent of indviduals. This paper is not intended to be a full legal briefing on the Privacy Act, but to provide some guidance and some examples of what good practice could be.

A simple example would be the collection of vehicle registration numbers or digital IDs[1] (MAC addresses) in mobile devices in vehicles driving along a road. Both of these unique identifiers may be Personally Identifiable Information and require the person to give (or withhold) their consent.

In the case of digital ID's (a prime area of expertise of the author), to comply with the Act and the Principles of Privacy the sensors used that are focused along the road to minimise detection of any devices not on the target road. This is a physical feature of the equipment used combined with the appropriate siting of the sensors to minmise unwanted data collection.

When the sensors do detect a digital ID, they undertake a real-time one-way roadside encryption of these MAC addresses to generate a unique (but anonymous) #Hashed ID. That is further anonymised with only part of the #Hashed ID transmitted across an encrypted link to the servers for processing and matching to other detections of that device.

It is impossible to 'reverse engineer' a partial #Hashed ID back to a MAC ID, so this is an example of removing any potential Personally Identifiable Information whilst still being able to retain highly useful data. This appoach does not materially affect the sample rate but it does demonstrate that the collected data addresses the Act's privacy principles.

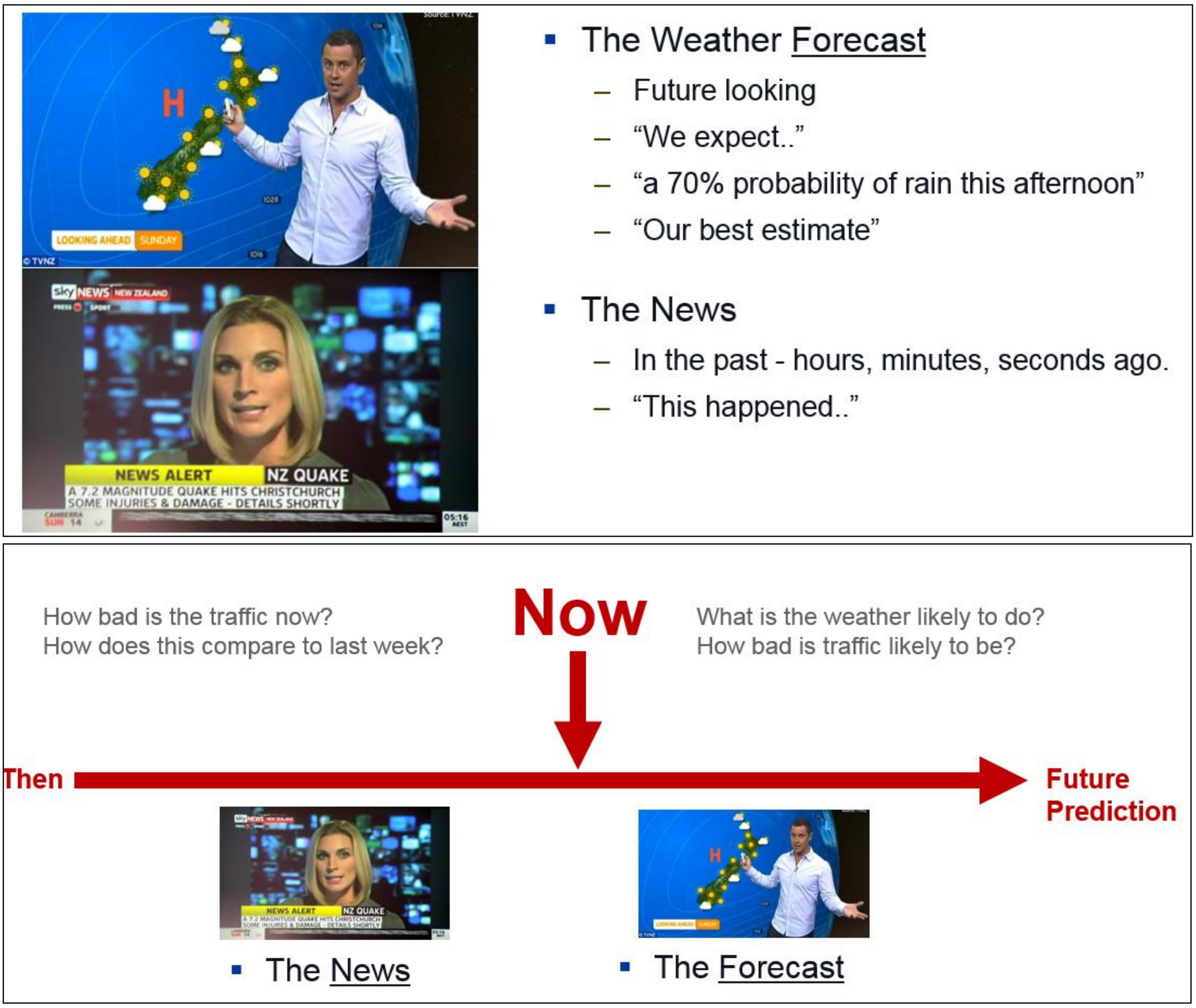
This example is of one approach used, and has been agreed with the New Zealand Transport Agency. It certainly isn't the only approach, but it does illustrate the need for a pro-active approach to ensure that the parties can practically demonstrate that Personally Identifiable Information is not being collected.

Whilst there may be a justified case of collecting other Personal Identifiable Information the gatherer should consider whether there is a genuine reason why that would be necessary. The gatherer should also consider whether an alternative data collection approach or some real-time anonomysation is required to demonstrate adherance to the Principles of the Privacy Act.

WEATHER FORECASTS vs THE NEWS

We are all familiar with the television Evening News and Weather (**Error! Reference source not found.**), they are a staple across many channels and countries. The reason for this analogy is to highlight the fundamental difference between a <u>Forecast</u> and a <u>Fact</u>.

[1] Typical examples would be Hands Free Bluetooth kits built into the vehicles or smartphones with Wi-Fi enabled.



- The Weather Forecast
  - Future looking
  - "We expect.."
  - "a 70% probability of rain this afternoon"
  - "Our best estimate"
- The News
  - In the past - hours, minutes, seconds ago.
  - "This happened.."

How bad is the traffic now?
How does this compare to last week?

Now

What is the weather likely to do?
How bad is traffic likely to be?

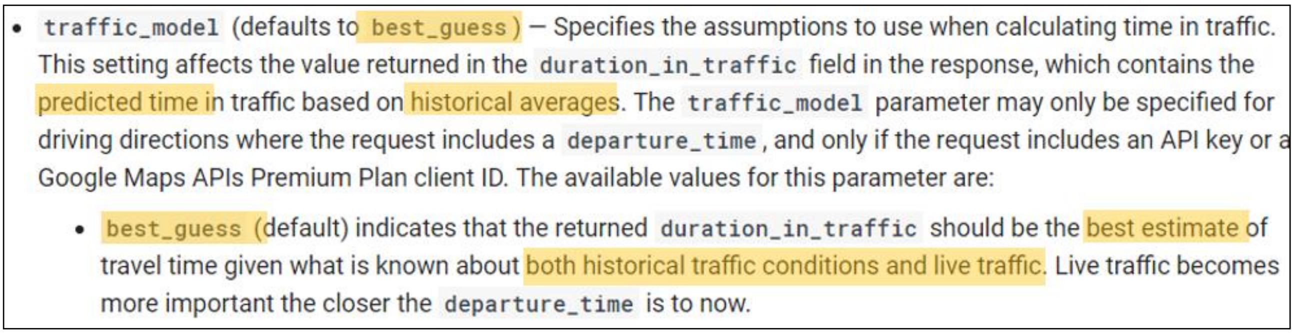
Then

Future Prediction

- The News
- The Forecast

Figure 1 The subtle but important differences the Weather Forecast and the News (One News, 2014, (SKy News, 2011))

Commercial offering's for Journey Time data are often made available in the future looking forecast environment (**Error! Reference source not found.**) rather than a factual record of what the journey was recently completed.



- `traffic_model` (defaults to `best_guess`) — Specifies the assumptions to use when calculating time in traffic. This setting affects the value returned in the `duration_in_traffic` field in the response, which contains the predicted time in traffic based on historical averages. The `traffic_model` parameter may only be specified for driving directions where the request includes a `departure_time`, and only if the request includes an API key or a Google Maps APIs Premium Plan client ID. The available values for this parameter are:
  - `best_guess` (default) indicates that the returned `duration_in_traffic` should be the best estimate of travel time given what is known about both historical traffic conditions and live traffic. Live traffic becomes more important the closer the `departure_time` is to now.

Figure 2 Definition of a commonly quoted journey time Forecasting service (Google , 2018)

The storage (caching) of this type of data is also not without its challenges; commercial vendors generally places stringent restrictions on how long any data can be stored with further restrictions on how it can be used, manipulated and presented (Google, 2018). The data provider referenced in Figure 2 places a standard maximum storage time of 30 days for that particular data source. Government bodies may have agreed longer retention periods, but as Professional Transportation Engineers we need to be careful that we don't fall foul of the vendor's small print and have an

unexpected visit by a lawyer working for the vendor.
More fundamentally, whilst useful in some real-time applications to predict travel times, it is a challenge to accept that a series of stored forecasts (rather than actual records) would be an acceptable data source for a Professional Transportation Engineers to utilise and still satisfy their Obligations.

Comparison of Sensor and non-Sensor based traffic data

In a recent comparison, data of peak period journey times was recorded along a corridor subject to a significant road layout change (Figure 3).

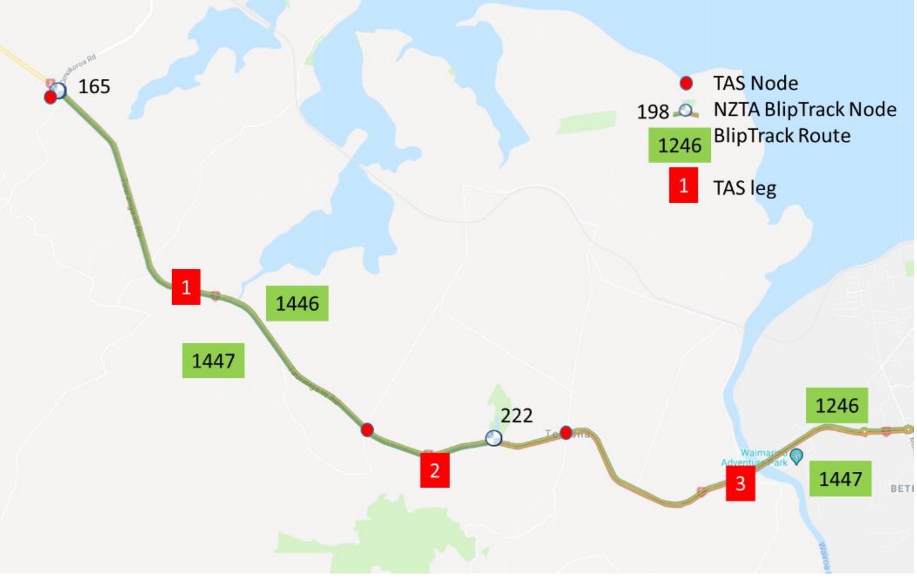


Figure 3 Comparison of Point to Point sensor and Sensorless traffic data

Two digital ID point-to-point recording sensors (165 and 222) were set 6.3km apart (Route 1446) (Beca Ltd, 2018) . Figure 4 shows the individual records of vehilces as dots, with trend highlighted and outliers shown in red. Outliers would typically be vehicles that stopped en-route (slower) or those not constrained by traffic queues (faster).

The consistency and volume of the trend data delivered by the filtering algorithm is considered by the author to be compelling, and this data set has been taken as the ground truthed data.

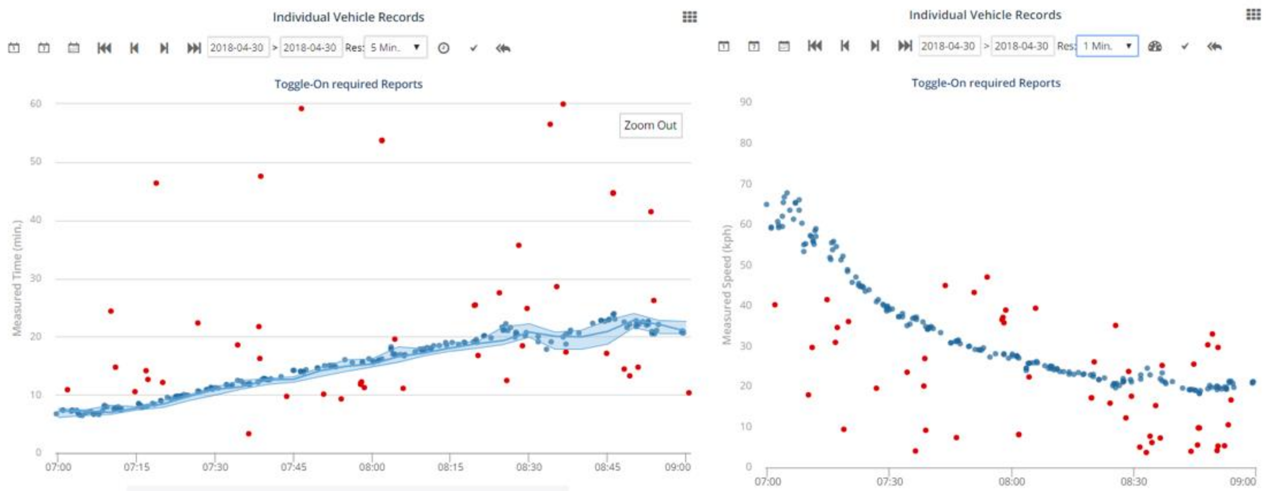


Figure 4 Point-to-point tracking sensor results showing trend and removal of Outliers.

To provide a comparison with sensorless data (understood to be derived from Google) was also provided. A common start point (165) was used but a longer route (Leg 1 + Leg 2) was utilsied. This sensorless route was 800m longer than the point-to-point Route (1446).

Based on the road geometry, posted speeds and the 800m longer route the sensorless journeys data should always have recorded at least 1 minute longer journey times than the point-to-point sensor data. A comparison of point-to-point sensor data (dashed) and the sensorless data (solid) is provided in **Error! Reference source not found.**.

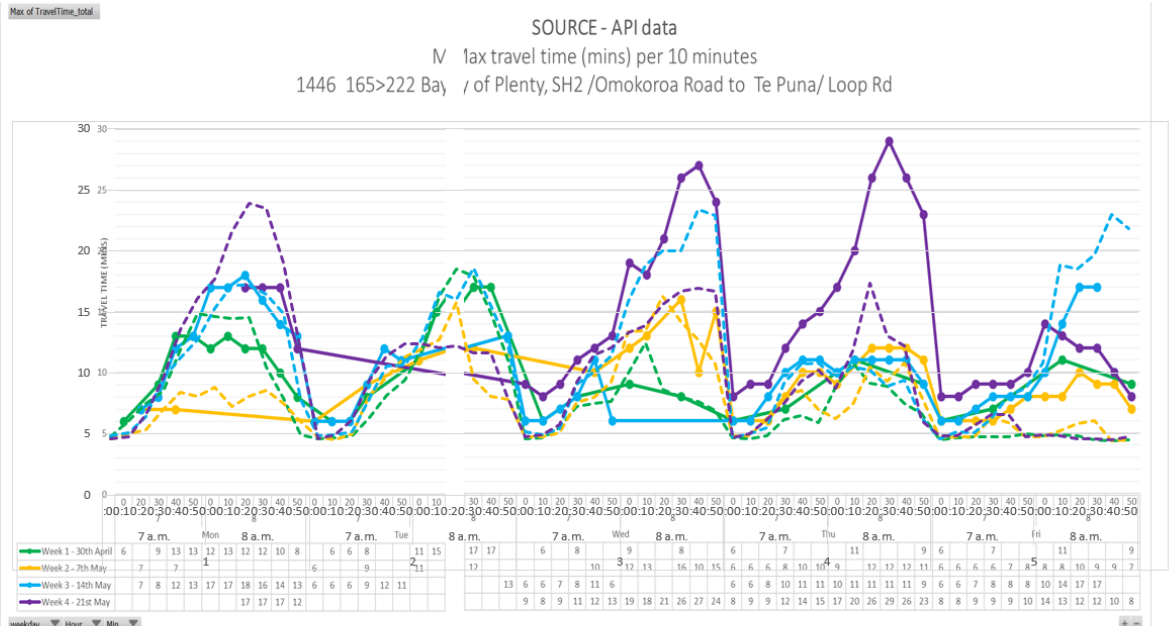


Figure 5 Comparison of point-to-point sensor data (dashed) with sensorless data (solid)

By analysing these graphs an assessment was made of the percentage of the peak period (in 10 minute intervals) where the sensorless data could be considered plausible when compared to the ground truthed point-to-point data.

Table 2 Ten minute periods of Peak Time (07:00-09:00) that sensorless data was plausible (>= point-to-point + 1 min)

Week	Monday	Tuesday	Wednesday	Thursday	Friday	Total of Plausible results	Total TAS results	% Plausible TAS results
1	6	4	3	5	4	24	30	65%
2	0	2	3	7	8	20	37	54%
3	3	4	1	7	4	19	45	42%
4	0		12	11	12	35	40	48%
Total						**98**	**152**	**64%**

The results in one direction indicated that 64% of the sensorless data was plausible, it was significantly less in the other direction. Overall the sensorless data only provided plausible travel times 54% of the time, the remainder of the time these forecasts were providing travel times that were incompatbile with the ground truthed point-to-point sensor data.

IT ALL TO COMPLEX - LET’S JUST PUT SOME MORE TUBES DOWN AGAIN

Having read the foregoing paper the reader could be excused for abandoning anything more sophisticated than deploying a few pneumatic counting tubes across a road and retreating back to the 1970’s!

Unfortunately that approach may also be flawed, prevailing Health and Safety best practice obligates the Professional Transportation Engineer to look at pro-actively mitigating the risk to the 'tube-putter-downer' by considering options that do not necessitate exposure to working on a live road.

Technology may come the aid of the Transportation Professional, the use of radar based counting systems (**Error! Reference source not found.**) is a growing alternate solution; especially the smarter systems that track vehicles along the road and determine size, speed as well as count traffic. The real time capabilities of these technologies is also becoming more attractive to Clients.

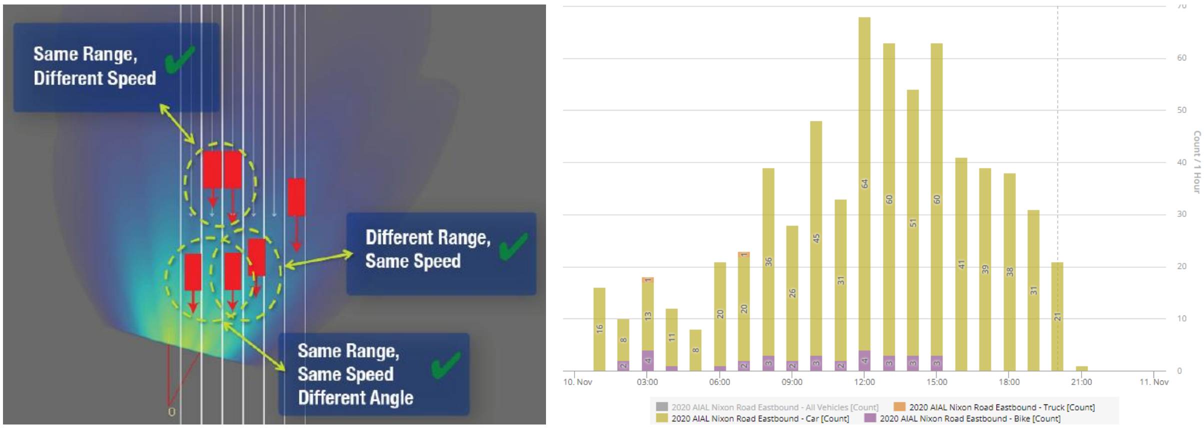


Figure 6 Use of advanced radar counter in place of tube counters

RECOMMENDATIONS FOR PROPOSED PRINCIPLES FOR BEST PRACTICE IN SOURCING AND UTILISING ROAD TRAFFIC DATA

Drawing from the above, these Principles are proposed to provide a clear and consistent framework for Transportation Professionals to source and use Road Traffic Data:-

1. Road Traffic Data is defined as including (but not exclusively)
    a. Traffic Counts.
    b. Traffic Classification by mode & vehicle type.
    c. Journey Times between set points or along defined corridors.
    d. Reliability (the variability of Journey Times across time)
    e. Route Choice (either for a whole Journey or a part of a Journey in the study area)
    f. Mode Spilt (by patronage or vehicle count)
    g. Repeated Journeys made by individual vehicles over time, including their variability
    h. Any other collected data relating to how roads are used.

2. Privacy – If any Traffic Data that could reasonably be construed as being 'personal information' is being collected the Privacy Act recommends that a Privacy Statement to be produced. This should be provided to the Client by both the Transportation Professional and the organisation providing the data.

3. Where Traffic Data needs to be sought for a Client, the Transportation Professional should identify several appropriate data sources and, whenever possible, provide the Client with a range of options setting out the relative reliability, longevity, cost, potential for re-purposing and any technical or commercial limitations to using that data.

4. A Traffic Data source which would require the Transportation Professional to compromise their Ethical Obligations will not be presented to the Client or used.

5. Where the Transport Professional can source Traffic Data in-house this data shall be subject to

the same Obligations as data sourced externally.

6. Client Provided / Instructed Data Sources. To provide an impartial professional service, the Transportation Professional is obliged to undertake their own checks on Client provided / instructed data in the same way as they would with data sourced from third parties on in-house.

7. Fitness for Purpose – To meet their Obligations, the Transportation Professional shall have a sound understanding of any the collection, processing, strengths and limitations of any Traffic Data that is used and as part of their Obligations that the data is suitable and appropriate for the Client's needs.

CONCLUSIONS

This paper highlights the Obligations of Transportation Professionals to provide Client's with impartial advice based on data that is fit for purpose.

The Transportation Professional needs to understand the strengths and weaknesses of various data sources and to cut through the 'truthiness' of data to assess its true value.

A comparison between two data sources has been presented, one taken from sensors one from a sensor-less commercial services. The analysis indicated that the commercially sourced sensor-less data only provided plausible results 54% of the time.

A set of seven Principles are proposed to establish a consistent and best practice approach to gathering certain types of traffic data.

Beca Ltd, 2018. *Review of Two Data sources for Travel Times monitoring on SH2 through Te Puna,* Hamilton: Beca Ltd.
Colbert, S., 2005. *The Colbert Report.* New York: Spartina Productions.
Engineering New Zealand, 2016. *Engineers and Ethical Obligations Practice Note 8, Engineers and Ethical Obligations,* Wellington: IPENZ.
Google , 2018. *Maps Platform, Web Services Directions API.* [Online]
Available at: https://developers.google.com/maps/documentation/directions/intro
[Accessed 10 11 2018].
Google, 2018. *Cloud Google Maps Forum - Term of Service.* [Online]
Available at: https://cloud.google.com/maps-platform/terms/?__utma=102347093.29900861.1541830735.1541831556.1541831556.1&__utmb=102347093.0.10.1541831556&__utmc=102347093&__utmx=-&__utmz=102347093.1541831556.1.1.utmcsr=productforums.google.com|utmccn=(referral)|utmcmd=
[Accessed 10 11 2018].
New Zealand Government, 1993. *Privacy Act.* 2018 Reprint ed. Wellington: New Zealand Government.
One News. 2014. [Film] Directed by -. -: TVNZ.
Reid, J., 2018. *Big Data Size Isn't Everthing,* Queenstown: Engineering NZ.
SKy News. 2011. [Film] New Zealand: Sky NZ.
US Government, 2018. *GPS.gov.* [Online]
Available at: https://www.gps.gov/
[Accessed 10 11 2018].
Waterfield, B., 2011. *The Telegraph.* [Online]
Available at: https://www.telegraph.co.uk/technology/news/8480702/Tom-Tom-sold-drivers-GPS-details-to-be-used-by-police-for-speed-traps.html
[Accessed 10th November 2018].