Big Data – Size Isn't Everything

Author

JOHN REID

Managing Director
Austraffic & Global Counting Systems

John.Reid@austraffic.com.au

John@globalcountingsystems.com

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

ABSTRACT

One of the biggest changes for traffic engineers and transport planners (and many other walks of life) in the last decade has been the enormous increase in the amount, nature and availability of data which includes a wealth of data coming from non-traditional sources.

Just because we are recording many more samples of a transport task, however, this does not automatically guarantee perfect accuracy nor that it is showing a depth of understanding about the reason people are travelling and thus the impact of any changes to the system

This paper takes up on a quote from T S Eliot (Eliot 1917) who wrote:

> *Where is the wisdom we have lost in knowledge?*
>
> *Where is the knowledge we have lost in information?*

With respect to the great writer, the paper adds to the above progression with:

> *Where is the intelligence we have lost in big data?*

This approach is a continuation of a process to compile examples and ideas on why we need to maintain traffic and transport engineering research, skills and resources to ensure that we are using accurate data and it is used with wisdom.

1 THOUGHTFULLY EMBRACING BIG DATA

I am not rejecting big data by any means. It can be very helpful, but we must not take it as automatically showing the truth, the whole truth and nothing but the truth. There is highly credible work being done around the world on the need to ensure that we are not blinded by something, simply because it comes from (and is analysed by) the latest technology. That a book "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy" (O'Neil 2016), has been written, reviewed and highly praised, might awaken us from any stupor we might have in this regard. One reviewer referred to it as "this urgent and necessary book". I refer to other international research in the course of this paper.

When I raised this issue at the AITPM National Conference in 2017, I received strong support that it is a subject that we had to address including one senior transport executive who said that he was going to present the material I had compiled to raise this issue with all his young professionals.

In some discussions I have had with professionals in preparation for this conference, one senior engineer and planner from New Zealand spoke of "plenty more examples" of the misunderstanding of big data and the subsequent inappropriate application of the results.

The international consulting firm WSP has just published a white paper "New Mobility Now – A Practical Guide, launched at the ITS World Congress 2017 in Montreal. (WSP 2017).

Scott Benjamin (who used to work for Austraffic) is WSP's Technical Director for Intelligent Transport in Australia and New Zealand and he is one of the authors of the report.

One of the insightful quotes from the white paper is as follows:

> "Big data is the biggest technological trend right now. We don't have to collect data on everything, but we do need to develop a data requirements specification to define what is needed, how often and its source." (WSP 2017).

You can hear an interview with Scott on this subject (Driven Media 2017)

2 CLEAN DATA

Every time a machine records something, that does not mean it is actually registering what you think it is. Data, no matter how many records it has, can be misleading when it has a significant number of records that are not recording what you think they are .

Here are some brief examples. My paper at the 2017 AITPM National Conference (Reid 2017) goes into more detail on the factors effecting the accuracy and credibility of the data

2.1 Bluetooth information

In the old days, if you wanted to do the job properly, a lot of manual work went into reviewing the results to see if they were credible. Today things are collected much faster, more frequently and in much greater volume. In the quest for efficiency and reduced costs there is inadequate recognition of the time to review work and to commit time to reflect and consider what the numbers really mean.

At the AITPM National Conference in 2013 Scott Benjamin (Benjamin 2013), presented research on the initial use of using Bluetooth devices to measure the origin and destination of vehicles by recording the presence of travelers' "smart" devices as they passed various points.
Where motorway roads and ramps were near each other, there were specific interference issues. Origin/destination trip assignments could be distorted by vehicles travelling on nearby facilities.

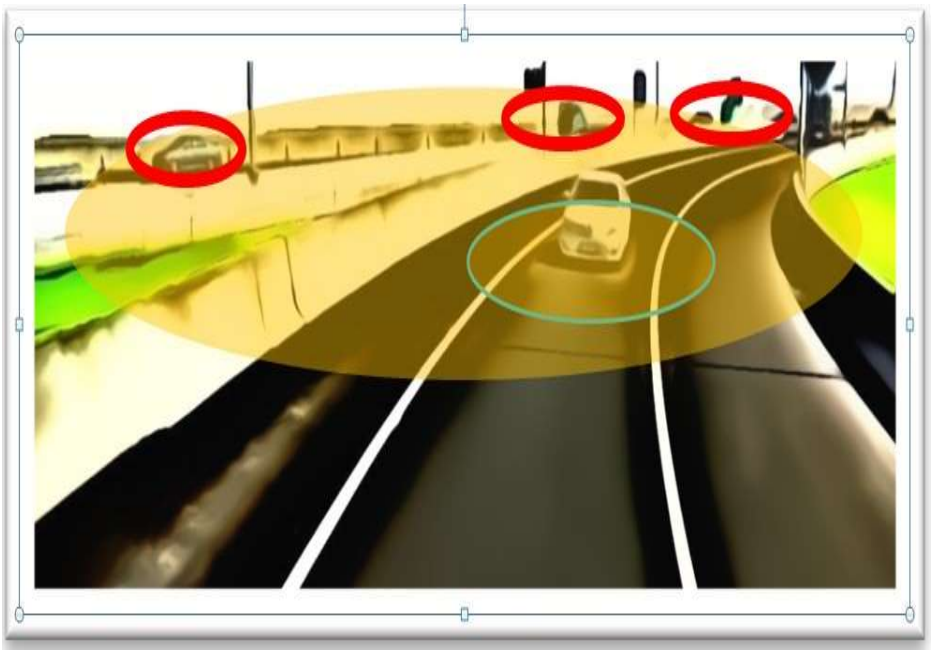

Figure 1 Extraneous sensed probes (Red) not in the corridor of interest (Green), Report by Scott Benjamin (Benjamin 2013), (then Austraffic now WSP).

To identify and categorise probe events is to harness the strength of such results, not to dismiss them. A large group of Bluetooth devices passing in close proximity could be a bus for example. This does not detract from the data as bus speeds are also an important measure, perhaps the most important. So, the results can be more revealing if we spend time to understand what they mean.

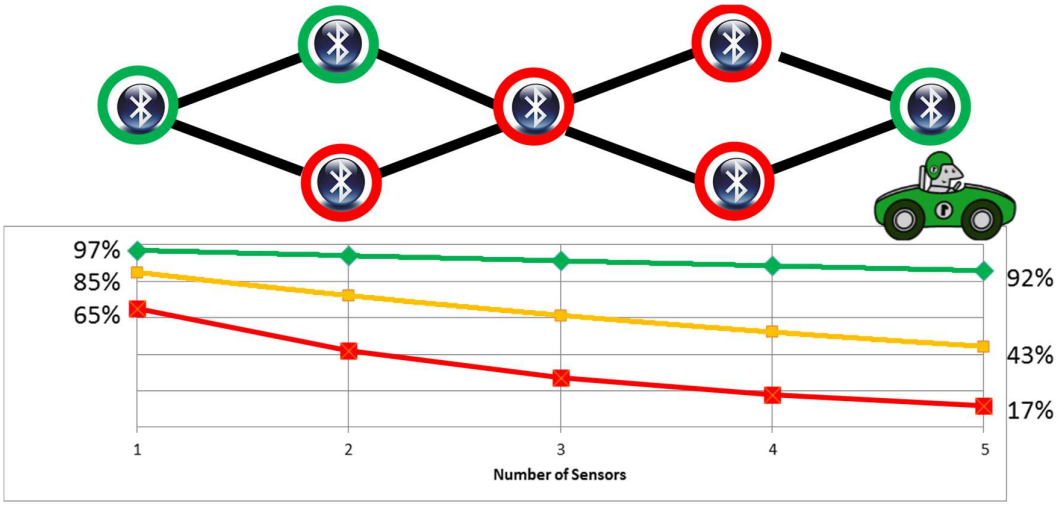

Figure 2 Blue Tooth (BT) accuracy impact on 1-5 station effective match rate

The key theme with probe sensed data is we don't know the sample rate at each location. Probe based survey data lacks definition and context without correlation to another data source. We need some other data source with which to validate the sample rate and any associated trip assignment, calibration factor. Fig 2 depicts the variability of match rates correlating to the degradation of accuracy.

2.2 Bluetooth (BT) and probe sensing technologies identifying and filtering "Outliers"

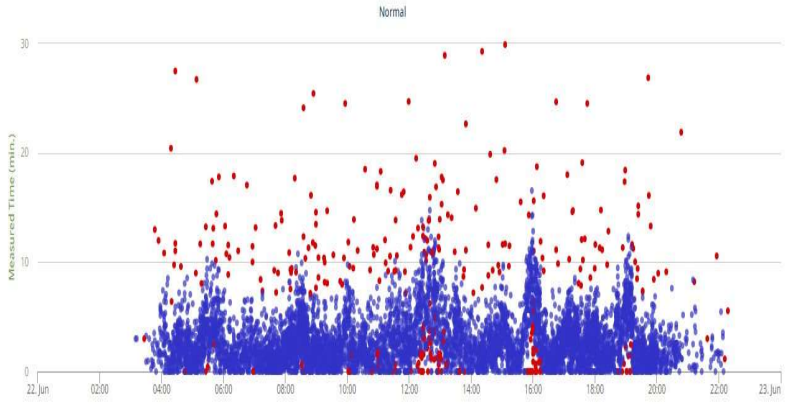

Figure 3 Outliers in the data – Which ones should we filter out?

Figure 3 & Figure 4 are great examples of outliers (data records that are beyond the expected range) with probe sensed data. It could be the same graphic for any probe sourced data set be it mobile, Bluetooth (BT) or WiFi, or navigational.

In this case, the experts at BlipTrack have built a very sophisticated filter engine that recognises results that are unexpectedly atypical from other results recorded soon before and afterwards. An atypical result is a genuine result but one that can be excluded based on set parameters. As a simplest example if you stop for a pie between two sensors your recorded journey is genuine but atypical and it would be valid to exclude it from the filtered data set.

Note outliers that exist high and low on the plot chart are shown in RED.

Typical conventional banding techniques may cull most of the high outliers while culling large travel durations that were real. The skill is making a call on the short duration red dots close to the horizontal axis.

Quality data cleansing is imperative to empowering today's and tomorrow's data scientist with a credible data set that will produce useful statistics and trend analysis.

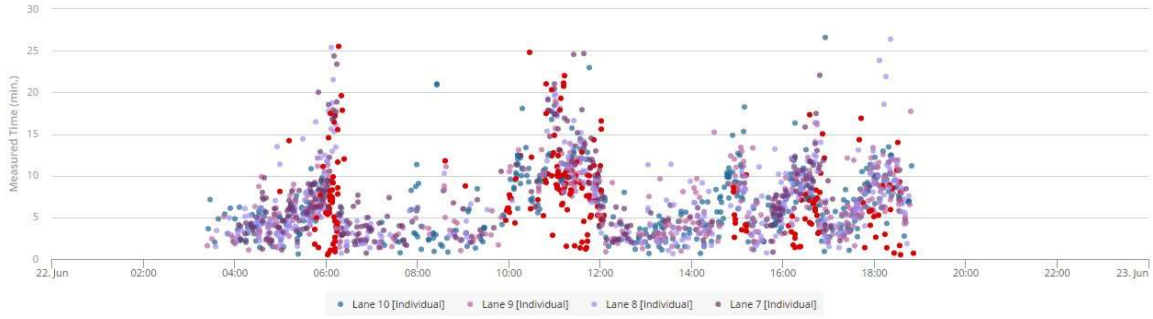

Figure 4 - Lane by lane measured times by time of day

If you look at figure 4 where would you begin to try and recognise the erroneous data? How much time would you give to the exercise?

This data is from static Bluetooth sensors, pairs of sensors that form to provide bidirectional zone based data along the corridor of interest

- The Austraffic website has the following A video on BlipTrack (Blip Track) titled "Bluetooth Devices and Driving Behaviour Help Ease Traffic Congestion". (Austraffic Undated a)
- Bluetooth Tracking System Review (Austraffic Undated b)
- Blip Systems Announces Bluetooth Partnership with Austraffic (undated c)

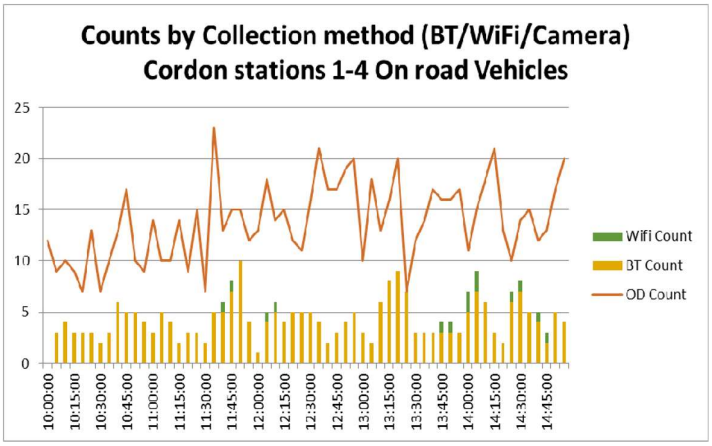

Figure 5 - Probe point sensor sample rates show another perspective for big data outcomes: what is the sample and what is it representative of?

Austraffic's on-road experience highlights the vagaries of sample rates between probe point sensors, these are reflected in Figure 5

This relatively low volume match rate is shown to highlight the variations between a camera (OD) derived count versus that from Bluetooth (BT) and WiFi.

WiFi for on-road data collection is of dubious consistency as compared with either camera (OD) or BT.

2.3 ANPR – Two examples

Raw ANPR, to-date, fails the data scientists. This will change with time. I have two examples to consider.

2.3.1 Car park confusion

Figure 6 – The real number plate and three different records of it through ANPR

Number plate recognition at car parks saves a lot of hassle at barriers and gives better control by restricting cars that are regularly driving in and out to try and avoid high charges.

If it does not work properly, it leads to bad data and very frustrated drivers.

Here we see the real registration plate and the three examples (in red) of how they were read wrongly by an ANPR based system at an international airport car park. These have all been collected in the past year.

The camera image shown in Fig 7 is intriguing. Just what am I looking at?

Figure 7 - A tradesman's utility vehicle and trailer passing a commonly used law enforcement camera. In the top left-hand corner of the picture is the camera's ANPR's perspective of what it believes is a rego plate. In fact, it is the checkered plate of the side rear portion of the vehicle. The ute created several recorded events, all dumped into an electronic file. One record was

correct; others derived from the checked plate, signwriting, fences and grass on the opposite kerb added further records.

2.4 GPS Data and plotting the results

Figure 8 - is of the road network immediately south of the Harbour Bridge in Sydney. There are two trips (highlighted with a yellow boundary) that appear to have deviated from the defined route. The recording was from a survey vehicle that we knew did not take this path off the digitised road network nor through buildings.

Because we know what the vehicle did, it was easily identified and fixed. But what if the data set was generic probe-tracked data? It may not even be wrong, as it could have been a device carried on foot or cycle or other?

What about Google's ubiquitous Traffic Data?

There is an unquestioned belief that traffic data from Google (and others) is live, accurate and perfect. In reality there is little (if any) published quantitative papers that have access to the base data to demonstrate that the information available can be used for anything other than live driving estimates. Google themselves call the data stream a 'best guess' but some in this industry are advocating it is a service that makes high quality surveys for multibillion dollar investments redundant. More at https://developers.google.com/maps/documentation/directions/intro

3 "FIXING" YOUR DATA

As data is usually only a sample and we know where some errors can creep in, we have developed considerable experience in expanding samples into representing the entire situation.

But through a lack of knowledge, limits in funding, and/or time constraints, inappropriate application of correction factors can be as misleading as collecting error-full data.

As an industry we need to be upfront about what it takes to do a good survey and not lapse into survey practices that are dominated by getting a quick result at a price that prohibits adequate review and thoughtful analysis.

As we have automated many data collections process we run the risk of taking numbers, as read from cameras and computer algorithms that are running on auto pilot. We run the risk of losing our understanding of what we are measuring and what "analysis" is being done to produce a result. i.e., losing the Wood for the Trees

4 WE NEED DEPTH AS WELL AS WIDTH

As shown above in the example of Bluetooth recording travel time surveys, big data can produce a lot of numbers but they may not have the depth of information to understand the details rather than just the average. A skilled practitioner will understand the algorithms to tag Outliers, and validate those for the required application, have full access to all the raw data and be able to identify when to use Medians, Means, Percentile results based and how each can be affected by road geometry and light/heavy mix.

With reference to Workshop B7 at the 11th International Conference on Transport Survey Methods in Transport (ISCTSC 2017), part of its intention was:

> This workshop examined data capture mechanisms and methods to fuse data as we enter into an era of active as well as passive survey data collection mechanisms.
>
> Examples of multiple source data collections in the actual field were reviewed. The workshop harnessed the collective experience of leading survey analysts to forge a pathway forward in this fast developing and exciting area of multiple, mixed, and heterogeneous data sources.

5 MANIPULATING THE DATA TO YOUR ADVANTAGE

Efforts to get data to tell us what he wanted to hear in the first place, is a now well-known phenomenon in our culture.

The informed media and academia are starting to see this issue. The Australian Financial Review

ran a story on the 10 April 2017 with the heading "A Masterclass in calling bullshit" (AFR 2017). It made reference to a course at Washington University simply titled "Calling Bullshit"

The intent of this course is to call out "exaggerated or foolish talk; nonsense or deceitful or pretentious talk". The need for such a course arose when Professor Carl Bergstrom and Assistant Professor Jevin West started to notice more 'bullshit' in the articles they were reviewing. "We think science is, sort of, it's ... at risk a little bit," West said. *(The Chronicle 2017).* The report went on further and added:

> "One big problem: Big data (one of the buzzwords of the century, which at its simplest refers to big sets of data, but has likely also been overhyped in its potential for revolution). He said he noticed methods of statistics meant for smaller data sets being applied to "big" data sets with millions or billions of examples, where it's easy to force a correlation that isn't necessarily accurate.
>
> He also observed situations where machine learning algorithms were "overfitting" data. Basically, you can have an algorithm that so specifically matches a particular data set, meaning it reflects even errors or noise, it fails when applied to another data set where you would otherwise expect it to work".

We must move to 'wisdom' not only because it is the way to the best answers but also because wisdom lets the world see that we are clear in our thinking, we understand where we are at, our processes are not dogmatic and the end result we wish to achieve is to serve their needs.

It is wisdom that will show the world that we are not sitting in our ivory towers developing solutions that we have determined are good for other people.

6 THE WISDOM OF THE ELDERS

It is easy to belittle what has been done in the past. The huge freeway building projects of the fifties and sixties were naive in their understanding of how we can get a total solution to our transport needs. Yet when we look back and condemn decisions, we can place ourselves in a situation where we think we now know it all.

We have seen stereotypical categorisations such as "the older members of our profession are locked in the past and the younger people having a greater grasp of the new technology". Or conversely, "we have lost the depth of understanding that is inherent in the wisdom of the elders while young people are caught up with a narrow approach to favouring technology for its own sake".

Of course, we say that we should embrace both, but this has to be more than a few warm and fuzzy comments made occasionally at conferences and dinners. This needs to have very clear

processes where we don't just express different opinions and hope we can convince everyone else of our own side of the argument.

With specific reference to road safety, David Brown's paper at this conference (Brown, 2018) calls for a move away from the lecture style of communication, the "adult-to-child" statements, towards a more collaborative approach.

7 CONCLUSIONS - WHAT DO WE HAVE TO DO NOW?

This paper has tried to show that data quality is a major issue and that we need to ensure we are giving enough resources to ensure that we know what we are collecting and we have removed the erroneous data.

It goes on to look at how this is part of the balance of respecting and using the experiences we have built up in the past while being agile enough to make full use of new technological developments.

But rather than just talk about broad strategies, we need to develop specific principles and directions.

The list could include the following:

- New technology is not only about how we can do the old things in a quicker or more efficient way.
- Numbers are not good data if they are not measuring what you think they are. It needs to reflect what we think it is measuring.
- Data compiled and presented fairly in graphs and tables and representing trends is information. If it is distorted or represented in a way that tries to obscure its true meaning then it is misinformation.
- It is not just counting traffic but understanding people and their needs and approaches to using transport. At the AITPM National Conference in 2017 the key note speaker, Brent Toderian from Canada, said that you cannot be a good transport planner unless you understand people.
- Price is not the only measure. We have seen that outsourcing activities may reduce short term costs but it can significantly reduce the value to the customer, can have unintended consequences and can end up costing more. Organisations that are only aiming at the short-term cost reduction are not supporting opportunities for deeper thinking and more comprehensive control of research and development projects. We need to commit to developing our knowledge and provide opportunities for learning.
- We have seen that some of our old ways of doing things need to be enhanced but also that

some bright new ideas lack the wisdom of the elders that is essential to maintaining a focus on the real outcomes not just fancy technology.

- Knowledge is the theoretical or practical understanding of a subject. It is something that is gained through experience or association. Knowledge is not just anything that supports your opinion.
- Wisdom needs knowledge and good judgement and an ability to know how to facilitate the value of wisdom toward the right solutions being embraced and adapted.
- Wisdom is not just an improved form of information or knowledge.
- Wisdom is rarely just a one-off judgement.
- Wisdom includes the ability to know when to say nothing.
- Wisdom can provide guidance although may not give the ultimate answer.
- Wisdom is getting to the best answer by the best processes in a way that people who are affected, own the solution.

8 REFERENCES

1. Austraffic "Bluetooth Helping To Manage Traffic Congestion" Austraffic Video https://austraffic.com.au/news/bluetooth-helping-manage-traffic-congestion

2. Austraffic "Blip Systems Announces Bluetooth Partnership with Austraffic" Press release https://austraffic.com.au/news/blip-systems-announces-bluetooth-partnership-austraffic

3. Austraffic (undated c) "Blip Systems Announces Bluetooth Partnership with Austraffic" Austraffic video https://austraffic.com.au/news/blip-systems-announces-bluetooth-partnership-austraffic

4. Australian Financial Review (AFR) (2017) "A Masterclass in calling bullshit". 10 April 2017 http://www.afr.com/leadership/a-masterclass-in-calling-bullshit-20170410-gvhjso

5. BlipTrack "Bluetooth Devices and Driving Behaviour Help Ease Traffic Congestion". Video Presentation Benjamin, Scott (2013) "Real time and historic data collection in a blender" Proceedings of the AITPM National Conference 2013 http://austraffic.com.au/system/files/REAL%20TIME%20AND%20HISTORIC%20DATA%20COLLECTION%20IN%20A%20BLENDER%20ver1-3_0.pdf

6. Brown, David (2018). Proceedings of IPENZ Transportation Group 2018 conference, Queenstown 21 – 23 March 2018

7. Driven Media (2017). 'New Mobility Now – A practical Guide – talking to one of the authors' Audio recording play across Australia on the community radio network http://drivenmedia.com.au/wp/new-mobility-now-a-practical-guide-talking-to-one-of-the-authors/
8. Eliot, T.S. (1017) – Choruses from The Rock – From "The Complete Poems and Plays of T.S. Eliot" Publisher FABER & FABER ISBN10 0571225160 - ISBN13 9780571225163 Text is available at http://courseweb.ischool.illinois.edu/~katewill/spring2011-502/502%20and%20other%20readings/eliot%20choruses_from_the_rock.pdf
9. Gilbert (2017) Syndicated cartoon that appears in over 2,000 newspapers, in 57 countries, and in 19 languages http://dilbert.com/
10. ISCTSC (2017) Program from "11th International Conference on Transport Survey Methods" Esterel Canada, September 2017, Workshop B7 "Data Fusion: Needs and Challenges in a New Transportation Data Landscape" http://www.isctsc2017.ca/
11. O'Neil, Cathy (2016) "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy", Publisher: Crown; 1 edition (September 6, 2016). ISBN-10: 0553418815, ISBN-13: 978-0553418811
12. Reid, John (2017) Proceedings of AITPM 2017 National conference, Melbourne 15-18 August 2017
13. The Chronicle (2017) "The Chronicle article- fine art of sniffing out crappy science. The Chronicle of Higher Education https://www.chronicle.com/article/The-Fine-Art-of-Sniffing-Out/238907
14. WSP (2017) "New Mobility Now – A practical Guide". ITS World Congress 2017 in Montreal. https://www.wsp.com/en-GL/news/2017/new-mobility-now-is-the-time-to-take-action

Acknowledgements

1. Richard Hanslip: Life member and past president of AITPM
2. Richard Young: Senior Associate, Hamilton - Beca
3. David Brown: Director Driven Media, Life member and past president of AITPM