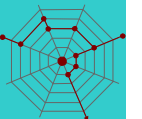


obtaining value from data sources:

remembering the basics

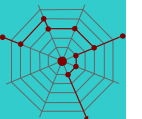
martin bach





has the quality or availability of the data we use in our models improved over this period?

- ‘within the profession’: hard to generate, nationally consistent data sets have evolved, such as:
 - TEMPRO (underpinned by CTripEnd), TRICS
- very much dependent on exogenous data sources, such as:
 - census (and boundary data)
 - employment data (NOMIS – National Statistics)
 - digital road networks (ITN, NAVTEQ)
- and we continue to collect data specific to our modelling, not found elsewhere:
 - traffic counts, roadside interviews, bus/train surveys, household travel diaries etc



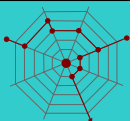
but we are so demanding.....



is there a dichotomy between what we want from the data
and its inherent quality?

- technical demand has increased as we try to be more specific and drill down to deal with complex relationships with increased market segmentation:
 - TIF and income groups
 - interrelationships between:
 - time of day / day of week / mode / car ownership / employment status etc
- the corollary is that we need MORE data, of a high quality, to support these analyses

it has always been difficult to get good data; have we got
any better at getting it?



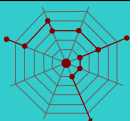
maybe not...



bitter experience tells us that :

- data collection is VERY expensive
- the quality of the collected data is extremely variable
- the data is often trusted far beyond what its quality justifies
- there is substantial inherent bias
- data is becoming harder, not easier, to collect due to:
 - administrative and legal shackles
 - changes in social mores
 - language
 - public sensitivity to private data
- expectation that the process needs 'less time'

there is an increased tension between the need for larger sample sizes and increasing cost and time pressures

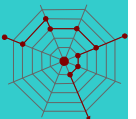




why has so little changed over these fifty years?

why do we still struggle to produce good quality data 'fit for purpose'?

can we do better?



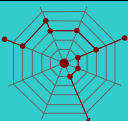
consider...



the data collection process for the 'intercept' survey.

this comprises:

- survey design and methodology
 - data collection
 - data coding
 - processing
 - validation

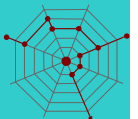




we've done it all before, but.....

- probably re-invented more times than we care to admit
- questionnaire design becomes critical
 - we need to be succinct, but we seek more (complex) information
 - problems of semantics: we know what we mean by a trip. Or is it journey, or perhaps a stage, or a tour....
 - complexity leads to greater scope for misunderstanding by the interviewer or respondent
 - every mistake leads to a significant drop in quality of data
 - how many times have we agonised over dealing with the 'Escort' question?
 - multi-purpose form: direct interview/self-completion

where do we look to benefit from good practice or to learn from others' mistakes?





once upon a time we did this.....

but now it's common for this to be outsourced.

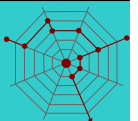
BUT

data collection companies, who are typically NOT users of the data, do not generally understand:

- the relevance or nature of the data being collected
- the importance of every point of detail which is specified in methodology

and getting the job done 'pdq' is often seen as more important than ensuring quality in the data

there is only one natural consequence.....



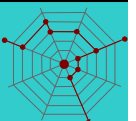


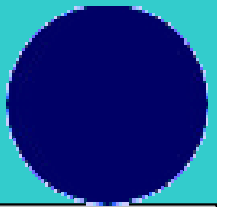
even more ERROR

we now have to deal with errors in

- the 'interview process'
 - inherent data capture issues
 - varied interpretation of the questions by respondent
 - the way the information is captured by the interviewer
- the data entry and coding stage

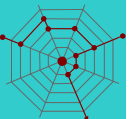
often these errors are detected long after the survey process has finished – when it is too late to take remedial action to prevent further instances



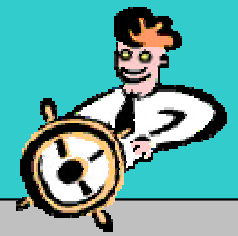


these are non-trivial steps often vastly underestimated

- ‘logical’ consistency
 - internally within a data record and across other records in the set
 - across related survey data sets (e.g. multi-modal trips)
 - spatial sensibility (direction; land use; intercept point)
- geocoding: address/postcode → grid reference → zone
- understanding and treatment of bias
- merge with related data
 - multiple observations (RSI)
 - handling the reverse trip
 - multi-modal trips
- expansion factor calculation



we've done it all before....

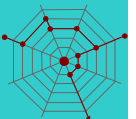


but where do we look to benefit from good practice or to learn from other's mistakes?

can we do better with
design methodology?
process?
toolkit?

basic questions

should we continue to collect such data (at vast expense)?
should we simply develop synthetic processes?
or should we have a mixture of synthetic and new sources?



what we need is to.....



recognise that there is a problem

decide that something should be done about it

evolve a forum for:

- developing standards
- defining processes
- exchanging ‘good’ experience

is webtag the obvious hosting mechanism?

