Australasian Transport Research Forum 2017 Proceedings
27 – 29 November 2017, Auckland, New Zealand
Publication website: http://www.atrf.info

Leveraging e-ticketing data to improve patronage and origin-destination survey outcomes

Stuart Muir[1], Elizabeth Stark[1]

[1]Symbolix. 1A/14 Akuna Drive. Williamstown North. Victoria, Australia, 3016.

Email for correspondence: smuir@symbolix.com.au

Abstract

This paper outlines the final stage in a detailed and complex application of e-ticketing transaction data, supported by focussed surveys to generate an overall image of the patronage of a complex, multiply connected and multi-modal public transport network. The work aimed to leverage big data assets into survey design and analysis. Doing this provides the opportunity for detailed, robust (and eventually on-demand) statistics about complex transport networks without exorbitant resource costs.

The work was done in conjunction with Public Transport Victoria and their myki system, however, it has application to any of the e-ticket systems worldwide. We highlight two individual surveying regimes:

1. One results in a patronage estimate that incorporates a correction term to the digital transaction record to account for the difference between digital transactions and physical access (the touch-on/off rate or TOR).
2. The second is an origin-destination (OD) derivation, which combines digital patronage data and a traditional intercept survey. This integrated approach enabled estimates to be generated for the likelihood of trip combinations that were not directly detected in the surveys, increasing efficiency in field and accommodating the ~47,000 potential OD pairs.

Rather than specifics on the survey implementation, this paper provides an overview of the relationship between the surveys and the transaction data assets and provides the generalised statistical methodology for combining the stand-alone touch-on rate and OD surveys with the transaction stream. It is presented to provide specific methods, but also to stimulate discussion about new opportunities for ‘traditional’ survey data collection within a big-data environment.

1. Introduction

E-ticketing for public transport has many benefits to the consumer (e.g. ease of use and reduction in waste). However, it carries far more return for the implementing agency. The deployment of more complex ticketing products that can account for multiple modes, multiple ticketing zones, and different usage levels is just one example. As data systems mature, agencies are increasingly in a position to leverage and combine with other data sources for a constantly updating view of network usage.

All e-ticketing solutions work fundamentally on the same principal:

1. The patron accesses the network through a (controlled) node, travels on the network unmonitored and unhindered, and then exits the network through another (controlled) node.
2. At each control point, the unique identifier of the e-ticket is recorded, facilitating not only a financial record, but a geo-located and time-stamped record.

If this were the complete situation, this digital record would be enough to understand the human flow across the network, including patronage loads and demand cycles. Unfortunately, the systems are more complex.

To avoid dangerous situations and impeding people, some access/egress nodes use a pass-through validator, rather than a physical, automated barrier. These are necessary on vehicles such as bus and tram, but they may also be found at stationary locations such as rail and light-rail stations.

In addition, those access nodes that are attached to a vehicle (naturally) move. Although GPS can tag and locate the device and the transaction, technical issues such as time-lag delay, canyoning (where the vehicle is in the midst of skyscrapers and can’t find a base-station) all lead to an incomplete instantaneous patronage view. It should be noted that, with respect to the financial ledger, these are not issues. But for policy and planning, and general longer term operational aspects of the network, these gaps can lead to misleading analyses and, in particular, inappropriate load and demand views.

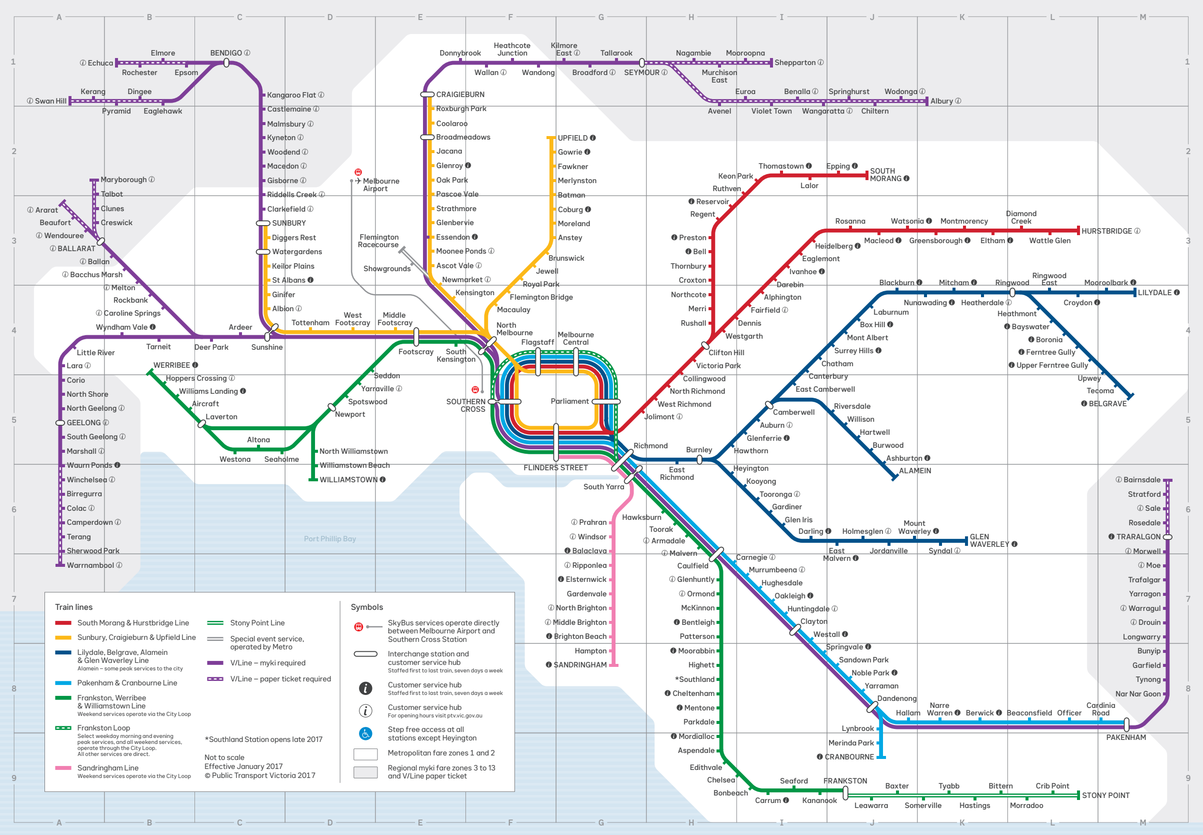
Making use of these data assets requires considerable data infrastructure, but also careful design work – to ensure that each data set collected integrates into the existing data.

This particular piece of work was implemented on the public transport network in metropolitan Melbourne. It was originally commissioned by Public Transport Victoria, and employed Symbolix for sampling design and preparation, analytics and data management. Ipsos Australia managed the origin-destination project team, including the complex field-work components. PTV managed the field deployment of the TOR survey.

Components of this work have been applied to the bus, tram/light rail and rail networks, as well as regional bus networks. We will be focussing on the rail network (Figure 1). It is the simplest but is still multiply connected, with a central loop common to all branch lines. The access nodes are physical gates permanently positioned (as opposed to the added complexity of moving ‘gates’ on trams and buses).

To establish context in what is a complex environment, this paper will first detail the necessary touch-on/off (TOR) correction required for the digital transaction record. From this we can detail the new, origin-destination intercept surveys, and how they can be cast onto this digital record to gain an insight into the entire set of possible origin-destination combinations.

Figure 1 The metropolitan Melbourne train system. Note the hub and branched spoke design and the potential for bi-directional trips in the loop. Image: www.ptv.vic.gov.au


Train lines
South Morang & Hurstbridge Line
Sunbury, Craigieburn & Upfield Line
Lilydale, Belgrave, Alamein & Glen Waverley Line
Alamein – some peak services to the city
Pakenham & Cranbourne Line
Frankston, Werribee & Williamstown Line
Weekend services operate via the City Loop
Frankston Loop
Select weekday morning and evening peak services, and all weekend services, operate through the City Loop. All other services are direct.
Sandringham Line
Weekend services operate via the City Loop
Stony Point Line
Special event service, operated by Metro
V/Line – myki required
V/Line – paper ticket required
*Southland Station opens late 2017
Not to scale
Effective January 2017
© Public Transport Victoria 2017
Symbols
SkyBus services operate directly between Melbourne Airport and Southern Cross Station
Interchange station and customer service hub
Staffed first to last train, seven days a week
Customer service hub
Staffed first to last train, seven days a week
Customer service hub
For opening hours visit ptv.vic.gov.au
Step free access at all stations except Heyington
Metropolitan fare zones 1 and 2
Regional myki fare zones 3 to 13 and V/Line paper ticket
Port Phillip Bay
FLINDERS STREET
SOUTHERN CROSS
Flagstaff
Melbourne Central
Parliament
Richmond
South Yarra
North Melbourne
Melbourne Airport
BENDIGO
SEYMOUR
BALLARAT
GEELONG
SUNBURY
CRAIGIEBURN
UPFIELD
SOUTH MORANG
HURSTBRIDGE
LILYDALE
BELGRAVE
ALAMEIN
GLEN WAVERLEY
PAKENHAM
CRANBOURNE
FRANKSTON
STONY POINT
SANDRINGHAM
WILLIAMSTOWN
WERRIBEE
TRARALGON
Warrnambool
Shepparton
Albury
Echuca
Swan Hill
Ararat
Maryborough

2. Information architecture

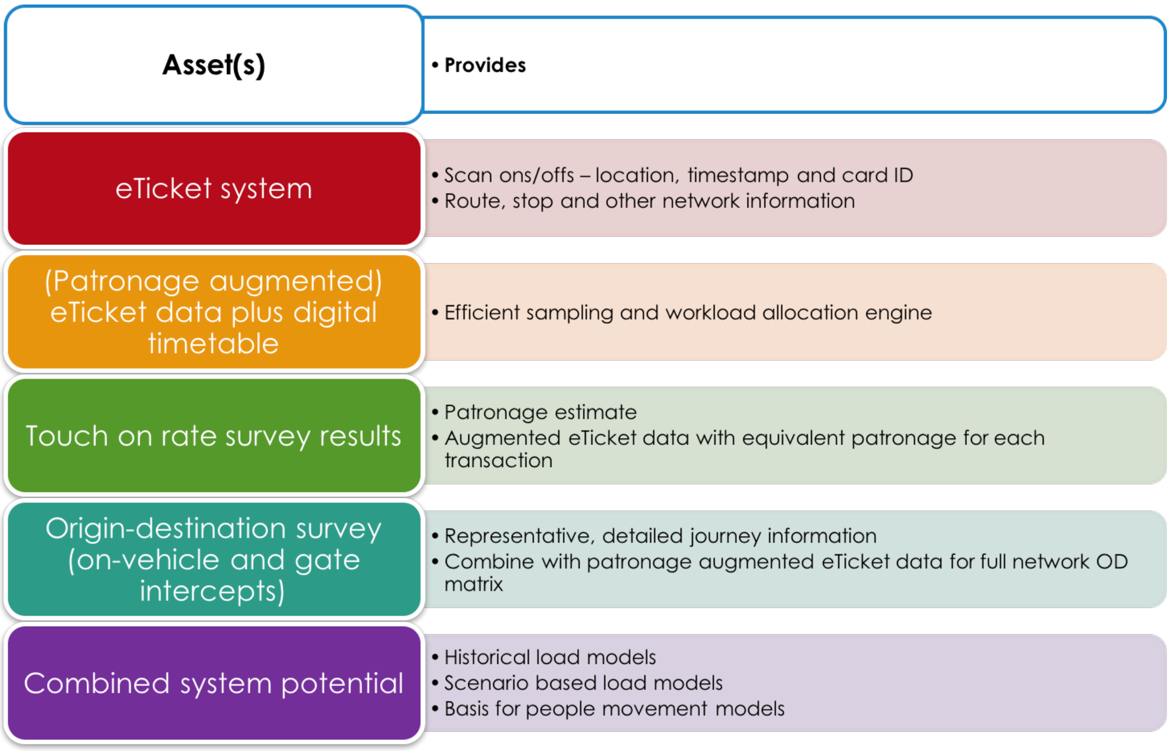
Before we outline the two case studies we demonstrate the overarching data system that relates the e-ticketing system to the two survey datasets (Figure 2).

In order to be able to sample patrons on the network we required an estimate of patronage. Initially this was drawn from earlier incarnations of the patronage surveys (or from timetabled trip counts where patronage estimates were not initially available at the required resolution). As the TOR survey was implemented, the TOR was used to augment every single transaction with a transaction-to-patronage boost factor. This data is fed back into the process to allow patrons to be representatively sampled each survey.

The e-ticketing data was also used to support sampling for the OD survey and to develop the final OD matrix.

Custom staging databases and scripts were developed to manage the sampling and analysis process. Publicly available timetable information (using the General Transit Feed Specification) was also integrated to automate the sampling and workload allocation process.

Figure 2: Relationship between survey and e-ticketing data sets



Asset(s)	Provides
eTicket system	• Scan ons/offs – location, timestamp and card ID • Route, stop and other network information
(Patronage augmented) eTicket data plus digital timetable	• Efficient sampling and workload allocation engine
Touch on rate survey results	• Patronage estimate • Augmented eTicket data with equivalent patronage for each transaction
Origin-destination survey (on-vehicle and gate intercepts)	• Representative, detailed journey information • Combine with patronage augmented eTicket data for full network OD matrix
Combined system potential	• Historical load models • Scenario based load models • Basis for people movement models

3. Touch-on rate correction methods

Within an e-ticketing network, users become effectively invisible to the system whilst within it. After scanning-on at an access node, most ticketing systems allow users to move freely within the network, until such time as they scan-off and an appropriate charge is levelled. Until (or if) a scan-off is registered, there is no information available to the operator about the direction travelled on a line.

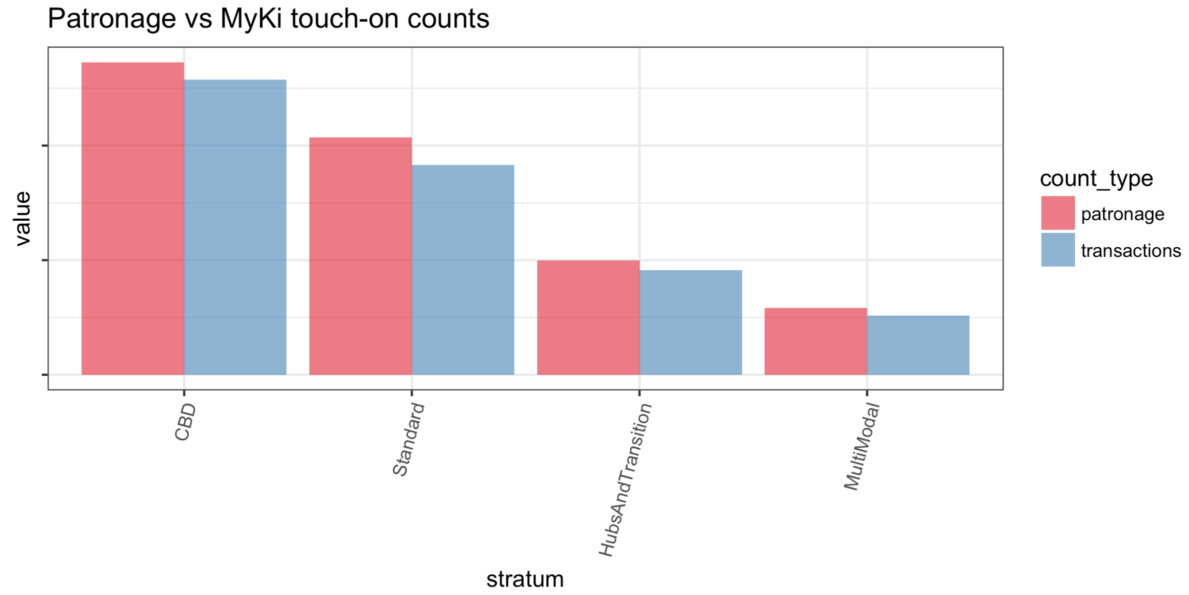
Further to this, the use of e-ticketing facilitates more complex product lines, including weekly or monthly passes. With these tickets, the patron may bypass the scan-on process entirely without consequence, as their e-ticket is valid.

There are therefore many legitimate reasons that the digital transaction record will not directly reflect patronage access. For patronage analysis, this record must be adjusted. We do this through a survey to determine the requisite "boost" factor. Statistically, we will use this as a ratio estimator (as opposed to a more traditional Horvitz-Thompson style), casting onto the digital transaction records to determine patronage count.

Through careful stratification, and a survey design that attempted to minimise the bias impacts due to ratio estimation, we generated a boost factor that can be associated to each and every access node.

In establishing the stratification it is important to note that the strata chosen need to reflect the behaviour pattern of scanning-on, and not the patronage pattern itself. This is to reduce the variance in the estimate. To do this, we stratified on the weekday type, four classes of stations based upon the differences in access mode to that station, and whether the access node was a barrier type or a validator.

Figure 3: Example plot showing the difference in number of transactions and number of patrons. We aim to correct this with a touch-on rate boost correction.



By generating the boost factor in this way, we can generate an access node level correction, which gives us the access patronage, corrected for factors such as ticketing product mix. This approach specifically targets the access (and egress) of the network.

Using capital letters to designate “universal” values i.e. modal patronage (*Y*) and modal touch ons/offs (*X*), and lower case to signify sample values, we have the relationship (see Kish (1965) for statistical background on ratio estimators):

$$Y = \frac{y}{x}X = rX$$

The variance of *Y*, the estimate of patronage, is a product of the variance of *r* only. The sampling error in *X* is zero, as it is a census value taken from the entire financial data set. It contains the entire digital record of things that can only be recorded digitally.

Therefore,

$$Var(Y) = X^2 Var(r)$$

The bias of such an approach at an average cluster can be shown to follow:

$$Bias(\hat{y}) \simeq (1 - \frac{n}{N})\frac{1}{n\overline{x}_U}(rS_x^2 - RS_xS_y)$$

So from this formula, we can make the following statements about bias of the total patronage estimate using this technique:

The bias of the estimate will be small if

- The sample n is large
- The sampling fraction $\frac{n}{N}$ is close to one
- The average number of passengers through a gate is large ($\overline{x_U}$)
- S_X is small (the standard deviation of the digital count)
- The correlation R is close to one

4. Origin-destination sampling methods

Although the transactional patronage record provides origin-destination information, there is still a place for generating a stand-alone OD probability matrix. For example, transiting (the act of changing services) becomes “invisible” within an e-ticketing system. The OD survey data itself can be used to provide more detailed information about the patrons’ end-to-end journey in addition to the estimated OD matrix. Both these assets are important reference data sets. They can be used to validate load models and to provide insight into the drivers for journey decisions.

A traditional approach to generating a matched origin-destination matrix is to approach the access nodes individually (the O’s in origin) and directly intercept patrons. Techniques employed on the London Metro network involved counting patrons as they go through some traffic bottleneck (Department for Transport (UK), 2010). From this minor cohort, individual patrons were intercepted for interview, identifying their individual journey termini. The simultaneous count of people through the funnel serves to weight the relative propensity for that journey OD combination.

The issue for such a survey design is getting enough coverage of the network. The London Area Travel Survey (LATS) employed 300,000 survey responses over 900 stations (Strategic Rail Authority Statistics Team (UK), 2005). The patronage flow is highly time dependent, so one has to generate a large, simultaneous survey count to ensure representative coverage. This issue is exacerbated if the intercepts are done on platforms, as the origin destination combinations are necessarily conditional upon the survey point.

To survey the entire Melbourne rail network using this approach would take 217 person days, surveying each rail station for an entire day. In fact, it is nearly twice this value, as many platforms have a direction associated with them. To truly capture the entire row of the OD matrix will require every platform on the network.

To increase the efficiency we needed to:

1. Increase the number of journey OD combinations that can be intercepted at a single time point, and
2. Remove the need to generate an absolute patronage estimate from the survey collection.

This second point was tackled by generating a **relative** OD weighting, and then using the TORS corrected digital records to generate the universal count of patronage for each OD combination.

This methodology required us to sample patrons proportional to patron/trip density but meant we could reduce the surveying effort significantly.

4.1 Types of intercept

There were two types of intercept used:

1. Traditional gate intercepts, and
2. On-vehicle intercepts.

They perform two very different but related tasks.

Gate intercepts have a simple probability of selection and the benefit that a surveyor can collect multiple surveys from the one origin location without moving. However, standing on a platform (or stop), one can only detect origin-destination pairs that are from that specific origin (Table 1). They are surveying a single row of the traditional OD matrix with no context.

The second type of intercept is an “on-vehicle” intercept. If the interviewer waits until the doors of the vehicle shut, and intercepts a patron between two stops (i.e. before the doors open again) then they are sampling from all OD combinations that **span** that interception point. This is a sub-matrix of the OD matrix, and is more efficient (Table 2). However, it carries the burden that calculating the probability of interception of these journeys is no longer simple.

Table 1: Showing a traditional Access node interception survey made at Node C. Cross hatching indicates the potential for directionality

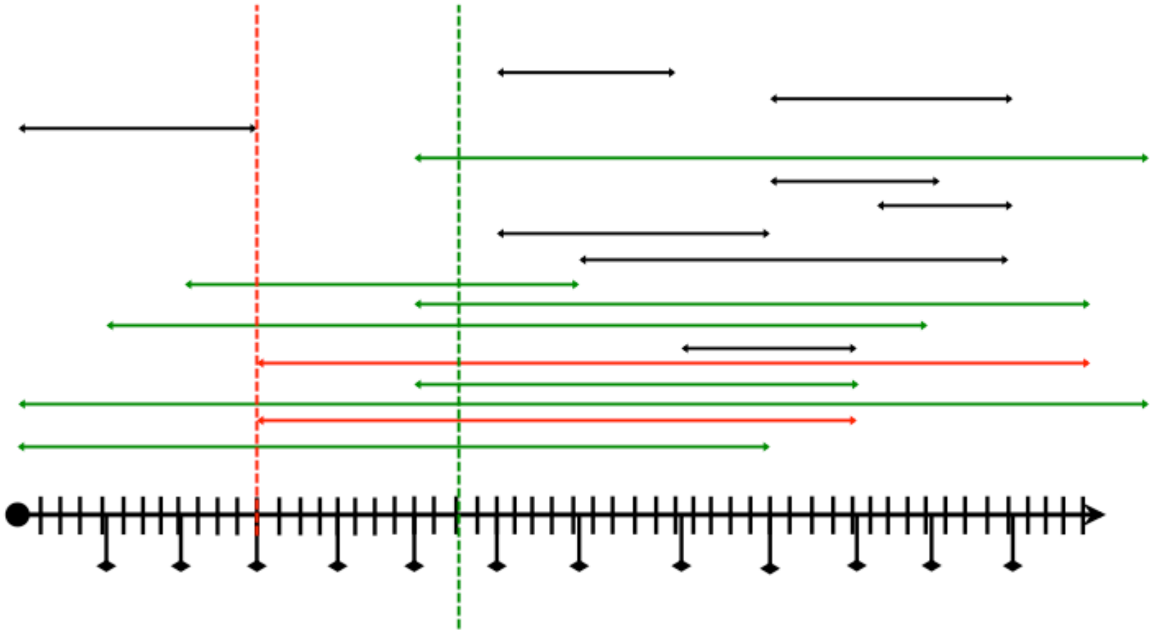
		Destination							
		A	B	C	D	E	F	G	H
Origin	A								
	B								
	C	▨	▨	▨	▦	▦	▦	▦	▦
	D								
	E								
	F								

Table 2 : Showing an on-vehicle interception, between Nodes C & D

		Destination							
		A	B	C	D	E	F	G	H
Origin	A				▨	▨	▨	▨	▨
	B				▨	▨	▨	▨	▨
	C				▨	▨	▨	▨	▨
	D								
	E								
	F								

An alternate view of on-vehicle interception is shown in Figure 4. Note how the gate intercept (red dashed line) can only detect the two red journeys. The green dashed line denotes an on-vehicle intercept, and is capable of intercepting all the green journeys, as well as the two original red trips.

Figure 4: Possible on-vehicle intercepts (green) and at-gate intercepts (red).



4.2 Sampling design

We used two-stage sampling, where the cluster (intercept location) was chosen first, with probability proportional to trip density/patronage, then individuals were randomly chosen in-field.

The probability of intercepting patron *j*'s trip:

$$P(j) = P(I)P(j|I)$$

That is, the probability of intercepting *j* is the product of the probability of selecting the interception point, and the probability of selecting *j* given we are at the interception point.

The interception point is behaving as a cluster. So the conditional probability is given by the number of interviews made divided by the volume of patrons.

$$P(j|I) \propto 1/N_{\text{patrons}}$$

To sample patronage (with equal probability of selection) the probability of selecting an interception point must be proportional to the number of patrons within the interception point, i.e.

$$P(I) \propto N_{\text{Patrons}}$$

To select the two samples we attach a selection weight to each of the interception points. For access nodes, this is simply the TORS corrected patronage count. For the on-vehicle interception points, we need an estimate of point load.

The point load on a vehicle is estimated as the cumulative sum of all serviced access points (TOR adjusted patronage) less the cumulative sum of all estimated egress patrons to that point. This then is a time independent representation of the number of patron trips spanning an interception point. P(I).

$$P(I) \propto \sum_{i=1}^{i<k} TOR * N_{\text{On}} - \sum_{i=2}^{i<k} TOR * N_{\text{Off}}$$

Given a set of interception clusters, we select a time based again on the relative activity rate at that point. In this way, we have a set of clusters that are selected approximately proportional to patronage.

4.2.1 De-weighting long trips during the post-collection analysis

The probability of interception with on-vehicle intercepts is biassed towards trips that span more interception points. So, upon interception, the response needs to be de-weighted by this factor to produce a representative sample of patrons, and hence trips. This is not an issue with the gate interceptions.

From Figure 4 we see how the patron can be captured at a variety of interception points and how shorter trips suffer a reduced detection probability with an on-vehicle intercept design.

The probability then of a vehicular intercept interview j is then

$$\begin{aligned} P(j) &= \sum_{k \in I} P(k)P(j|k) \\ &\approx M_j P(I)P(j|I) \end{aligned}$$

The M represents the number of possible interception points (k) where the patron j **could** have been intercepted. This is unknown until the interception interview itself. It is therefore added post hoc in the analysis stage. Without incorporating it, those taking longer trips will be more likely to be interviewed.

4.3 Generating field workloads

The two interception methods were employed within a single work-shift by a single interviewer/field officer.

Shifts were created using custom scripts that statistically selected either an intercept or a gate to begin the shift. From this point, field officers were directed onto services to generate on-vehicle intercepts at particular points on the network, or back out onto gates based upon the patronage estimates. The achieved sampling fraction for each required strata were dynamically updated and used to adjust the probability matrix for the next shifts.

For safety reasons the shifts were generated such that field officers were not surveying on fully laden vehicles during peak periods. During these times gate interceptions were selected preferentially. A load model (under development) was used to predict high load services and adjust the shift accordingly.

Field officers collected geo-location data that allowed us to verify the actual responses against the expected sample.

Although complex, such a design enabled us to move the field officers across and through the network. They were able to move in such a way as to keep in touch with the average patron at that time, keeping the sampling representative.

4.4 Verification through simulation

To verify that this form of interception could work, a simulation experiment was run. In it we created a known OD patronage distribution. We create a survey agent who, based on a nominated interception point, randomly selects an allowable OD-pairing. From this we attempt to regenerate the original OD matrix. The code simulates a large number of ‘surveyed’ journeys.

As can be seen from Figure 5 we see the approach is non-biassed (the simulation results on the x-axis reproduce the simulated OD patronage on the y-axis).

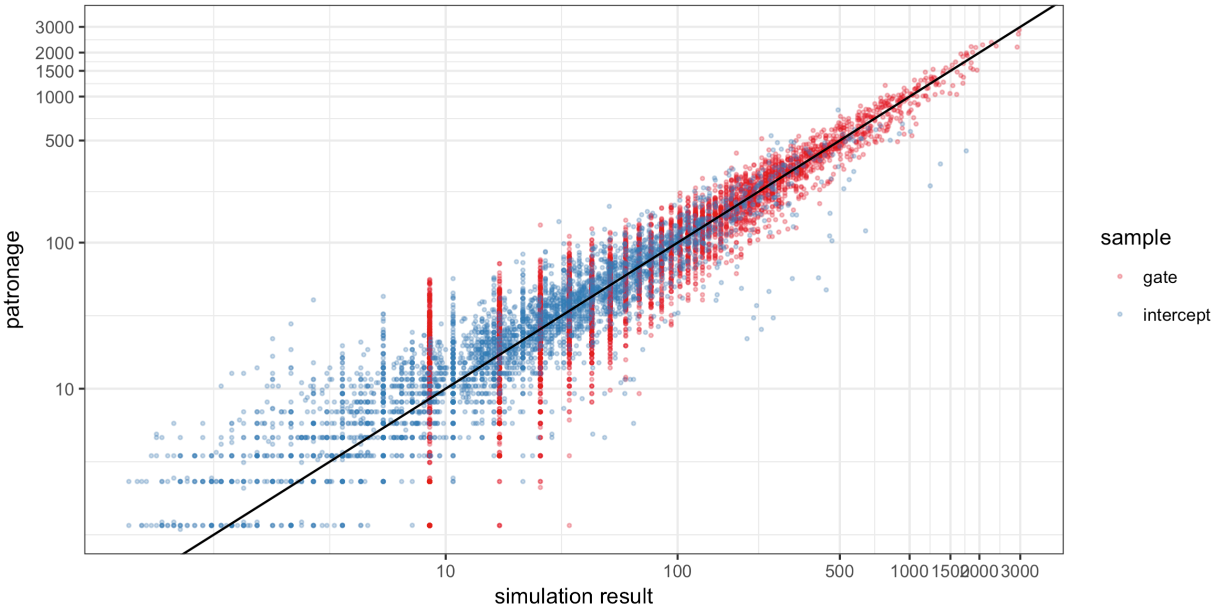
Secondly, it has a much lower variance around high patronage OD pairs. This follows precisely the same pattern as the TOR sampling approach, which is designed to generate tighter relative confidence bounds around commonly used access nodes, than less common ones. This is an important feature of both surveys as a 20% relative error has far greater implications for planning on a million transactions-a-month gate, than it does on a 40 transactions-a-month gate.

The use of log-transformed axes is not to control for variance, as one might expect with count and proportional data, but merely to manage the large spread in response values and produce an interpretable chart.

The horizontal and vertical artefacts correspond to the discretisation effects of generating proportions from integer-based counts.

Attention is also to be drawn to the odd on-vehicle intercepts (blue) that appear to under-predict quite significantly (to the right of the figure). These are left in the chart to highlight just how much information is needed regarding the network topology. The Melbourne train network is multiply connected, with a central loop (Figure 1). This means there are different numbers of interception weights between two given OD pairs, depending upon how the patron travelled from O-to-D. In these cases, the simulation averaged the possible paths to generate the down-weight. In practice, this value is part of the survey response and is known explicitly. It was unavailable to the simulation, which had only an OD matrix to start with. These points are left to highlight the sensitivity of the approach to this knowledge (or lack of it) and as a point of discussion about the volume of external data that this approach rests upon.

Figure 5: Verification simulation results. Blue are simulated on-vehicle intercepts and red are at-gate.



The simulation also gave us insight into how to generate a high coverage of the network using these two techniques in tandem. This is represented by the relative location of the gate (red) and the on-vehicle (blue) points in the above figure.

With a hub-and-spoke style network with the hub at the CBD, performing a gate intercept at the hub generates coverage of the vast majority of possible journeys. Simply, most patrons travel from a spoke and exit in (or near) the hub (or vice-versa). Gate surveys are very efficient at these locations. They complement the on-vehicle interceptions, which are more efficient towards the lesser-used spoke ends.

The OD survey then uses both gate intercepts and on-vehicle intercepts as replicate surveys to combine and generate the complete, relative OD matrix.

4.5 Analysis methods

To generate a sample of every OD pair directly would be prohibitively expensive. One of the driving motivations for attempting such a survey design extends both from the increasing size of modern networks (the metropolitan Melbourne rail network has over 46,000 possible combinations), and we required a technique that returned a (non-zero) estimate for low proportion OD combinations.

Following data collection, we generated two replicated survey tables, holding the proportion of OD trips on the network for each intercept type. These tables contained the proportions as directly measured by the survey responses, down-weighted by the number of interception points.

In its simplest form, each cell of the OD matrix, γ_{ij}, represents a proportion, which we would normally model using a beta distribution. We can, however, use the additional constraint that the row must total to unity to use a generalised distribution, the multinomial beta, or Dirichlet. This extends from the truism that every patron accessing the network at origin O must, necessarily exit the network somewhere.

To provide an idea of the analysis method, let us focus on just one row, which describes the proportion of destinations, ***j***, given an origin, ***i***.

We invoke a vector of parameters, called β, whose components, β_j describe this row of destinations. We also describe a pseudo-population (the reasoning behind this terminology will become clear later) given by $K = \sum_j \beta_j$.

From this we can generate the expectation for the proportion of each destination given the origin as $\gamma_j = E(\pi_j) = \frac{\beta_j}{K}$. This generalises to the entire OD matrix through $\gamma_{ij} = E(\pi_{ij}) = \frac{\beta_{ij}}{K_i}$.

4.6 Incorporating TORS corrected e-ticketing data

To generate an OD matrix that covers all possible trips, we added the patronage corrected e-ticketing data. From this we extracted all existing OD pairs. This data set is not perfect but serves as a first-guess at the O-D matrix, allowing us to infer O-D probability even if a particular pair was not surveyed. If we did not include an informative prior guess, we would require many more survey observations or accept a statistically (much) weaker result.

Bayes theorem states that the expectation of the OD proportion (the posterior) is also a Dirichlet density (Agresti, 2002), implying

$$E(\pi_j | n_1, \dots, n_N) = \frac{n_j + \beta_j}{n + K}$$

Here, we have used the symbol, n_j to denote the observed cell (row x column) population, and n to denote the relevant, row sample size. The key to this OD integration is the manipulation of the above to

$$E(\pi_j | n_1, \dots, n_N) = \frac{n}{n + K} p_j + \frac{K}{n + K} \gamma_j$$

The second term is our prior information (the e-ticketing generated OD matrix), and the first term is the sample effort information, which can be seen to be a weighted average through the relative sizes of the sample, *n*, and the pseudo-population, *K*, of the prior information.

The problem was that K may be very large and dominate the study, more so than our expression of true confidence in it. The myki data (or any e-ticketing data) contains millions of paired transactions that swamp the size of the survey. We descaled this population, so that it gives the relevant insight, and allows each survey cohort to contribute its own information.

4.7 The shrinkage parameter

As alluded to in the introduction, the transactional set, even when corrected by the TORs boost factor, is not a complete set. There are many instances where there is legitimately no scan-off record associated with a scan-on, boosted or otherwise.

Let us presume that we wish to bias our results to the prior assumption by 10%, i.e. 1:9 relative weighting. To do this our pseudo-population for the prior must be

$$K = \frac{10\% \times n}{(100\% - 10\%)}$$

This enables us to control how much influence our prior information holds over our observation. The sample size is fixed, so we manipulate the prior only, and generate a new estimate. A natural estimate of our confidence is the proportion of paired scan-ons within the transactional set.

The variance for any point estimate of proportion under these constraints is then

$$Var(\pi_{ij}) = \frac{\pi_{ij}(1 - \pi_{ij})}{(n + K + 1)}$$

4.7.1 Important Notes on the use of shrinkage

The sampled control vector is relative weighted to accommodate the difference in trip interception probability. However, this relative weighting does not alter the effective sample size, *n*, which remains true to the survey (i.e. 100 sample count, even individually relative down-weighted, still sum to 100 weight).

The shrinkage parameter can be used repeatedly, to “shrink” multiple surveys and data streams together. For the survey presented here we used this approach to combine the two intercept types onto the e-ticketing data. Care must be taken that the overzealous use of this control does not overwhelm the analysis with false confidence. A good rule of thumb is that at no time should the resulting pseudo-population (*n*+*K*) increase much beyond that of the sample sizes.

4.8 Results

To finally present a transport matrix, the relative scaling of the rows of the current proportional matrix must be re-enforced. The use of the Dirichlet structure means that the proportional OD matrix has each row independently combining to unity, or equal weighting across rows.

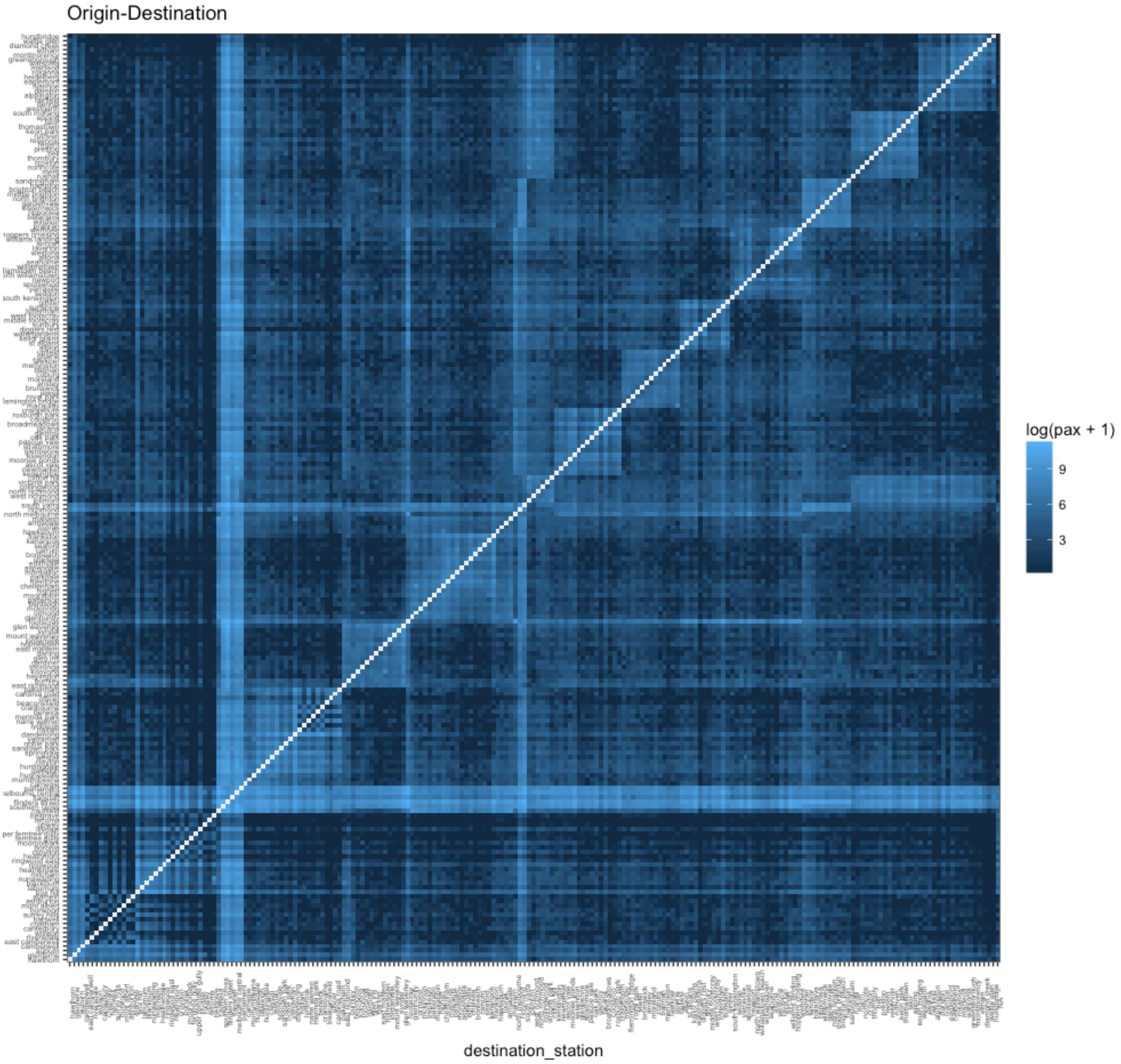
Each row sums to 100%, and indicates the proportion of destinations given the access point. The e-ticketing data, with the relevant TORs correction gives the physical numbers of patrons using that access point.

By multiplying this number onto each row, a traditional transport OD is generated, correctly reflecting the amount of patronage within each cell. Importantly, this also provides an avenue to generate the number of trips between an OD pair for a time-period where there was no survey. Provided the analyst is satisfied that the OD proportions remain reasonably constant, an OD proportion matrix generated by a survey in November could be multiplied onto the patronage for February to estimate the number of patrons attempting each journey in the latter month.

Lastly, this approach also ensures that ‘one source of truth’ for patronage assessments within the organisation is maintained. That is, the estimated patronage from summing the OD matrix will always agree with the official, estimated patronage as generated from the focussed TOR/patronage survey. This is a critical aspect of modern data analytics, ensuring that different tools do not generate a plethora of interpretations for fundamental operational metrics.

Figure 6 shows a complete OD picture of the metropolitan Melbourne train network. The lighter colour indicates greater numbers of patronage. While the axis text is illegible at this resolution, the figure still serves to highlight the connectivity of the network, and distinctive OD patterns are clearly visible. On each axis, stations in close geographic proximity to each other are presented near to each other on the axis, which is why there are distinctive squares, showing the journeys between those locations. The city loop stations are also easy to pick out, as they have a lot of patrons travelling in both the Origin and Destination directions, from all the other stations.

Figure 6: A complete OD picture of the metropolitan Melbourne train network.



5. Acknowledgments

A number of people with expertise in statistical sampling, software development, field data collection and public transport policy have contributed to this work. They include Andrew Molloy and Julie Cain (PTV), Jennifer Brook and Rene Durrant (Ipsos Australia), and David Cooley (Symbolix).

6. References

Agresti, A 2002, '*Categorical data Analysis, 2nd Ed.*' John Wiley & Sons, New Jersey

Department for Transport (UK) 2010, *National Rail Travel Survey Overview Report*, viewed 30th July 2017, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/73094/national-rail-travel-survey-overview-report.pdf

Kish, L 1965, ‘*Survey Sampling*’, John Wiley & Sons, New York.

Strategic Rail Authority Statistics Team (UK) 2005, *London Area Travel Survey National Rail Results*, viewed 30th July, 2017, http://www.asinfo.ru/upload/iblock/44b/London%20Area%20Travel%20Survey.pdf