Measuring the Quality of Public Transport Journey Planning

Dr P Bell*, Nick Knowles*, Paul Everson*

* Trapeze Group, The Mill, Staverton Wiltshire, UK, BA14 6PH
Public Transport Journey Planning systems such as Traveline or Transport Direct have been around for nearly a decade now. All current systems rely on a small number of vendors who have developed the tools necessary to collate timetable information from a large number of disparate sources and to then integrate that data into a single data set that is closely aligned with the specific vendor’s journey planning algorithm.

A number of changes are occurring in the Journey Planning software market that are challenging the status-quo. Google has entered the public journey planning market with its own format for public transport data and its own journey planning software. The UK government’s push to providing open public transport data is also enabling new entrants to the market.

These changes should challenge us to ask whether there is a need to measure the quality of journey planning software. Such a quality measure can be used by bodies procuring journey planning software. Alternatively, in a de-regulated market such quality measures can be used to provide confidence to operators and travellers about the provided journey plans.

In this paper we explore the issues that affect the quality and user’s perception of quality, based on our own experience of developing journey plan software over a decade.

1. Background

Large scale internet public transport journey planning systems have been around over a decade. All current systems rely on the products of a small number of vendors who have developed the tools necessary to collate timetable information from a large number of disparate sources and to then integrate that data into a single data set. This data set is designed to allow the vendor’s journey planning algorithm to quickly and efficiently search for journeys. For this reason the data set and the algorithms are closely linked.

Journey planning engines themselves have evolved rapidly, replacing the simple first generation single mode engines with advanced engines capable of complex evaluation of large route networks. As important, journey planning engines now allow users to optimise across a number of different criteria (shortest journey, least transfers, etc).

A number of changes are occurring in the Journey Planning software market that are challenging the status-quo. Google has entered the public journey planning market with its own format for public transport data and its own journey planning software. OpenTripPlanner offers a basic open source engine using Open Street Map data for the road network. The UK government’s push to providing open public transport data, and the development of the Traveline National Data Set (TNDS) is also enabling new entrants to the market.

Although superficially all journey planners may appear the same to a user, their simplistic journey planners are prone to deliver misleading journeys. Examples include requiring travellers to cross water or motorways where no footpath exists, poorly modelled interchanges leading to impossible connections or choosing the direct route when a more indirect route would get there faster.

There may also be policy motivations, particularly relating to fare pricing – one example is the rail industry which through RSP’s Journey Planning Code of Practice, has been anxious to make planners conform with certain guidelines and enforce certain rules.
We believe that good journey planning is an essential part of encouraging travellers to make the modal shift away from cars. Clearly, a bad experience based on a poor journey plan will delay this.

For this reason, we believe that the development of standardised ways of judging journey plans is essential.

In this paper we explore the issues that affect a user’s perception of journey plan quality, based on our own experience of developing journey plan software for over a decade:

- What do we mean by quality of journey planner results?
- What sort of tools and techniques might be used to evaluate quality?
- How do different techniques for assessing the quality work better for different quality criteria.
- Our own experience of using automated testing to measure the quality of journey plans

There remains a separate issue not covered in this paper around how quality scores might be used (government regulation, industry self-regulation, or some form of independent assessment of the quality).

2. What Quality Parameters Need Measuring?

At the simplest level, a good journey planner allows a user to easily plan a journey for a specific date and time between two arbitrary locations. The returned journey plan must be easy to follow and get the traveller to their destination efficiently. However, this simple statement already shows that quality in journey planning is multifaceted.

A good journey plan depends on the availability and correctness of several distinct types of data, the user’s own priorities, and how the journey planner processes the data and uses it algorithms to search it. We explore these aspects in more detail in the following paragraphs

Ease of data entry: a good gazetteer is absolutely essential to good journey planning. The gazetteer needs a complete set of start and end points (see next paragraph) but also needs to cater for other vagaries of data entry: for instance, references to locations by local convention, common misspellings and locational awareness.

Data coverage and data accuracy: Journey planners ultimately rely on having data that is an accurate representation of the real world transport systems they describe.

Key issues needing addressing include whether the data set is complete (no missing services) and accurate (the service runs at the times the journey planner thinks it runs). Particularly, accuracy can be challenging for complex networks such as the UK. Here services can vary significantly between weekdays and weekends, between term time and school holidays, and between ordinary days and bank holidays. Variations can include services not running at all, running with a lower frequency, running with shortened routes, or numerous combinations of these. One of the key differences between Google’s General Transit Feed Specification (GTFS) and TransXchange is that the latter has a much more flexible data model to describe these variations.

A second important part of data coverage and data accuracy, and one that is often neglected is how well interchanges are modelled. For good journey planning, accurate stops locations and names, interchange transfer paths and transfer times available. For multimodal planners that include access walk legs, it is not just
PT data but also the road and path network data that contributes to journey quality – and minor errors can give glaring mistakes.

**Data Currency:** The scheduling of transport network typically happens as a sequence of progressive refinements, with different versions of data – a long term timetable, a short term timetable, a daily operational timetable and a real-time operations plan that may be successively released to journey planners. Journey planning engines typically transform this data into a highly optimized binary representation suitable for searching and the computational effort to do this may be materially significant (for example Google’s static pre-computation takes several hundred hours for a major urban area).

For the journey planner to provide an accurate journey plan for a specific day, these different updates to the timetable need to be processed efficiently. This challenge is as much about the quality of the organisational processes of the journey planner’s operator, as it is about the strength of the journey planning algorithms.

**Algorithmic effectiveness:** The user’s concern is to the best journey - but “best” may depend on widely different assessment criteria; for example, ‘fastest’, ‘easiest’ (e.g. least changes, most accessible), ‘cheapest’, ‘most comfortable’ (e.g. least crowded). More complex criteria may also be germane such as ‘most reliable’ (e.g. how likely to be late), ‘most resilient’ (i.e. what happens if I miss a connection), ‘most diverse’ (representing a choice rather than a minor variation). Differences in the algorithm and the data models used to support the algorithm are of course fundamental to the results and the main theme of this paper. For dense networks such as commonly found in the UK there may be a very large number of possible journeys and the task of ranking alternatives to show a manageable few is fundamental and can vary significantly between journey planning algorithms.

Different users may want different optimisations so ideally the algorithms will be parameterised to allow the different criteria to be specified for user preferences.

Diversity is an important but subtle aspect of a journey planner. For example, travellers would find a journey planner unhelpful if in response to a query it returned the same train journey and just varied the route to the station slightly.

A classic example of the issues around algorithm effectiveness can be seen with journey plans from central London to Heathrow Airport. These will frequently route passengers via the Heathrow Express at Paddington even when using the Piccadilly Line from central London actually provides an overall simpler journey.

**Performance:** Users very often use journey planners iteratively to explore and refine their understanding of their choices. Speed of response is therefore important for usability and the overall quality of experience.

3. **Quality Measurement techniques**

Given the variety of system aspects that need testing, there are basically three types of test that might be applied to a journey planning system.
A key question in deciding how to measure quality is whether the quality score should be absolute or relative. An absolute score is useful, if one is trying to measure an existing journey planning provider in a ‘monopolistic’ situation and assess how well they are doing.

However, for many of the quality criteria we outlined above, it is very difficult to develop an absolute score. Therefore, a relative scoring of journey planning systems provides a better way forward. This can even be used in the above monopolistic scenario to allow a comparison between different versions of a journey planning system.

**User Testing**

Observing real users interacting with the system can provide useful feedback on key aspects of the system. This has the advantage of providing real feedback, but lacks the ability to apply a strict numerical algorithm to the scoring.

**User Feedback**

A good way of getting a subjective measure of the quality of the journeys is to ask users to provide feedback on the journeys produced by the journey planner. Ideally, users would be selected in a completely randomised way to provide feedback on the journeys they undertook. This approach would clearly identify issues with the journey planner (for example, bad modelling of interconnects) but may not necessarily identify where better journeys may have been available but not proposed by the Journey Planner.

**Process Audit**

Running an efficient journey planning system is as much about the organisation as it is about the journey planning algorithm. A process audit would look at such aspects as the ability of the organisation to cope with updates, ability to deal with complaints arising from the journey planner, and quality of algorithm development process.

**Automated, Randomised Test Suite Based on Sampled Real Journey Plan Requests**

Creating a test suite based on real journeys as planned by existing travellers. The sample set should be designed to be representative both of the journeys planned, but also of the times of year (e.g. it should include term time, school holidays, etc). This would give a very representative set of journeys to measure. A key issue would be how to score the returned journeys.

A challenge remains even here how to measure the quality of the journey. For something such as shortest journey, then this is unambiguous so long as all journey planners under test use the same parameters (e.g. walk speed).

However, for more subtle parameters such as easiest journey, then a weighted scoring is needed.

**Edge Case Journey Plans**

The approach of the previous bullet will miss testing some of the infrequent edge cases that cause journey planners to fail, yet are annoying when they happen.

### 4. Quality Parameter vs. Quality Measurement Technique

Different approaches work better for different quality parameters. Table 1 (right) shows how different quality measurement techniques are better suited to different quality parameters and that no single measurement technique will give the entire solution.

The correctness of this view is borne out by Trapeze’s
approach to testing which has been built up over 15 years of experience.

For our release testing of new versions of our ETNA journey planner we use large test sets of sample journeys, comprising both randomly generated Origin/Destination pairs and hand chosen edge cases, built up over time in part as a result of support calls or user feedback that can be run and compared automatically with known results.

We have automated tools that can compare the quality of the journeys produced against journeys generated in previous releases of the journey planner. In this way, we are able to assess quickly and accurately, that changes to the journey planning algorithm will not degrade the overall quality of the solutions delivered.

Since our tools use the JourneyWeb interface, they could equally well be used to compare the outputs of non-Trapeze journey planners that support the JourneyWeb interface.

Other techniques widely used are visualisation. NaPTAN users will be familiar with use of ITO! World to spot anomalies in the placement of stops and in journey timings. Similar techniques can be used to visualize journey results and spot outlying or anomalous results; for example travel isochrones can give a gestalt representation of the results of very many queries in a way that highlights anomalies. However such techniques they rely on eyeballs and are less suitable for automated testing.

Whether using automated comparison tests or generating visualisations for human comparison, a set of cross product comparison tools could be of great interest for highlighting the real differences between planners. It would be important to include a significant component of randomly chosen journeys to prevent gaming of the benchmark.

Conclusions

In this article, we have shown that measuring the quality of a journey is not down simply to which returns the quickest journey. In fact, there are numerous criteria that need to be assessed to measure the quality of a journey planner.

Likewise, there is no one single approach to objectively measuring quality. Instead, a number of different approaches need to be used.