ARUP

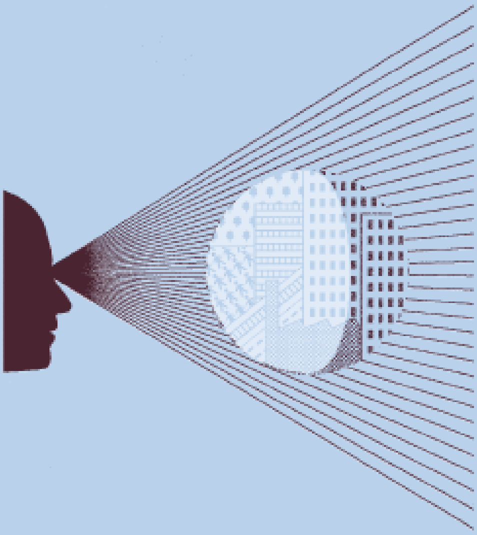
Oxera

Is the data capable of meeting the study objectives?

Data capability report

Prepared for
the Department for Transport,
Transport Scotland, and
the Passenger Demand Forecasting Council

March 2010



Oxera Consulting Ltd is registered in England No. 2589629 and in Belgium No. 0883.432.547. Registered offices at Park Central, 40/41 Park End Street, Oxford, OX1 1JD, UK, and Stephanie Square Centre, Avenue Louise 65, Box 11, 1050 Brussels, Belgium. Although every effort has been made to ensure the accuracy of the material and the integrity of the analysis presented herein, the Company accepts no liability for any actions taken on the basis of its contents.

Oxera Consulting Ltd is not licensed in the conduct of investment business as defined in the Financial Services and Markets Act 2000. Anyone considering a specific investment should consult their own broker or other investment adviser. The Company accepts no liability for any specific investment decision, which must be at the investor's own risk.

Contents

List of tables

List of figures

1 Introduction

Oxera and Arup have undertaken a study, 'Revisiting the Elasticity-Based Framework', by the Department for Transport (DfT), Transport Scotland, and the Passenger Demand Forecasting Council (PDFC). The primary aim of the study is to update and estimate the fares and background growth elasticities contained within the Passenger Demand Forecasting Handbook (PDFH).

The study has a number of secondary objectives, which include:[1]

- exploring the use of innovative or alternative econometric techniques;
- re-specifying and extending the core elasticity-based framework;
- improving the underlying data.

As part of this study, a number of reports have been produced, detailed below, which form key elements in the formulation of the overall final forecasting framework, and are referenced a number of times here.

Reports prepared by Oxera and Arup for the 'Revisiting the Elasticity-Based Framework' study:

- 'What are the findings from the econometric analysis?' (the *Findings* report)
  - 'Is the data capable of meeting the study objectives?' (the *Data capability* report)
  - 'How has the preferred econometric model been derived?' (the *Econometric approach* report)
  - 'What are the key issue for model specification?' (the *Model specification* report)
  - 'How has the market for rail passenger demand been segmented?' (the *Market segmentation* report)
  - 'Does quality of service affect demand?' (the *Service quality* report)
- 'How should the revised elasticity-based forecasting framework be implemented?' (the *Guidance* report)

The key question addressed in this report is whether the data is capable of being used to test the hypotheses of the economic model set out in the *Model specification* report, and the segmentation set out in the *Market segmentation* report. Hence, this report asks whether the data is sufficient to enable the study to meet its objectives? In particular, the report seeks to demonstrate that the study team has achieved the third objective set out above—to improve the underlying data. This includes detailed information on the availability of data, the process of data gathering, and the scope for improving the existing evidence base. This is intended to highlight the strengths and weaknesses of the data available and, in turn, to present the data that has been available for use in the econometric re-estimation exercise. The accompanying *Model specification* report aims to achieve a similar objective with respect to the other study objectives. A service quality measure is discussed in the *Service quality* report. These reports, which form key elements in the formulation of the overall final forecasting framework, are referenced a number of times here, and hence are referred to in the report by their name.

[1] Department for Transport (2008), 'Rail Passenger Demand Forecasting: Revisiting the Elasticity-Based Framework Request for Proposal and Statement of Requirement', July, pp. 12–13.

This report is structured as follows.

- Section 2 discusses the demand data available for use in this study.
- Section 3 provides a statement of the explanatory variables available for use in the estimation. This section is split into two sub-sections: explanatory variables of direct interest, and explanatory variables which are being included in the estimation as control variables to avoid biasing the coefficient estimates of the variables of interest.
- Section 4 discusses whether the data is suitable to meet the study objectives set out above.
- Section 5 sets out how the data might be used to form the basis of rail demand forecasts.
- Section 6 provides practical recommendations for improvements to the evidence base.
- Section 7 discusses the final dataset compiled by the study team.
- Section 8 concludes and provides recommendations for next steps.

Each section begins with a non-technical summary of the section.

2 Demand data

The raw demand data for this study was provided to Oxera by DeltaRail, and consists of data covering a period of 18 years for more than 22,000 origin–destination station pairs of flows, broken down into six ticket types: first class season, first non-season, standard full, standard reduced, standard Apex and standard season. The study team has estimated the econometric models by aggregating these ticket types so that they are consistent with the recent simplification to the fares system introduced by ATOC—ie, anytime, off-peak, and season tickets.

This data was extracted from the LENNON computer system and included zero and negative values, due to the treatment of refunds. However, these negative and zero observations are not useful for this study because forecasting negative demand is not meaningful. In addition, as accurate forecasts of very low levels of demand are likely to be difficult to produce, the data has been processed to create a dataset that can be used as the dependent variable for the econometric analysis. This processing consisted of two stages:

– removing all negative and zero values;
– removing observations which display very large year-on-year percentage changes.

The rationale for the first stage is discussed above, but the rationale for the second stage is that these observations are likely to exert an undue influence on the econometric analysis. (There is a legitimate rationale for removing these observations in that they may be caused by data errors or very small values. This means that analysis based on these outliers/extreme values may not give robust model results.) Hence, removing these observations should improve the accuracy of the parameter estimates. The cleaning rule used here is that observations which increase by more than 100% year on year are removed from the dataset.

To ensure that the estimated parameters are not unduly affected by this processing of the dataset, some sensitivity tests have been carried out comparing the results of a 500% filter—ie, only observations where the percentage change was greater than 500% year on year were removed. The results of this analysis are reported in the *Model specification* report, but the filter does not appear to unduly influence the parameter estimates.

The explanatory variables which could have been included in the econometric modelling are discussed in section 3.

This section sets out the details of the demand data that has been gathered for this study, including the length of the time series, level of disaggregation, and any known issues or problems with the data, together with the processes of demand data collection, cleaning and selection undertaken by the study team.

2.1 Raw data

DeltaRail has provided the study team with data on a large selection of flows (origin–destination pairs).[2] Data on journeys and nominal revenue is available at the flow level, for six ticket types (standard full, standard reduced, standard season, standard Apex, first season, first non-season) and for the total of these. Data is available for 18 financial years, from 1990/91 to 2007/08.

The raw data was extracted from the LENNON database, and the following processing steps were undertaken by DeltaRail prior to the data being provided to the study team:[3]

– flows for which the origin is the same as the destination were removed;

[2] DeltaRail (2009), 'Rail Demand Data Extension', June 25th.

[3] More detail is provided in DeltaRail (2009), op. cit.

- stations which are not valid stations and cannot easily be converted to valid stations were removed;
- individual stations within a group were collated into that group (revenue and journey allocation below this level is likely to be spurious);
- flows for which the origin or destination is within the London Travelcard area, except London BR or London Zone R1, were removed;
- flows for which total revenue is less than £10,000 in 2005/06 were removed for each year.

There are a number of issues with the raw data, in particular the inclusion of negative journeys and revenue due to the treatment of refunds within LENNON. There are also certain flows for which there is zero revenue but positive journeys within a particular year. Therefore, a process of cleaning the data was required. This process is described in section 2.2.

Table 2.1 presents summary statistics on the data of the raw journeys. This shows that the minimum value of journeys for all ticket types is negative, for the reason mentioned above. The high standard deviation relative to the mean indicates that the variables are distributed widely around the mean. The high skewness and kurtosis[4] scores demonstrate that the distribution of each variable is positively skewed and 'peaky'—ie, most of the data lies at the low end of the distribution.[5] This may have implications for some diagnostic tests in the econometric analysis, such as significance testing, because these tests either assume normality or rely on a law of large numbers and a central limit theorem for validity. If the normality assumption is not met, then the tests are only valid asymptotically—ie, if the sample size is very large (strictly, tending to infinity). See the *Econometric approach report* for details of diagnostic testing in the chosen analytical framework.

The statistics presented below include the mean; the number of observations (N x T—ie, flows multiplied by the number of periods in which the flow data is available); standard deviation (sd); skewness; kurtosis of the distribution; the value of selected percentiles (eg, p1 corresponds to the value of the first percentile, p5 to the value of the fifth percentile, and so on); and iqr (ie, the interquartile range—the difference between the value of the third quartile and that of the first quartile). These summary statistics help to understand the distribution of the data and hence the likelihood of outliers, and, therefore, any implications which this may have for the econometric analysis if left untreated, as explained above.

[4] Kurtosis is a measure of the 'peakiness' of the distribution—ie, the higher the kurtosis, the spikier the distribution.

[5] For reference, the skewness of a normal distribution is zero, and the kurtosis (k) of a normal distribution is $3\sigma^4$. σ=1 for a standard normal distribution N(0,1) and therefore, for a standard normal distribution, k=3.

Table 2.1 Summary statistics of journeys: raw data by ticket type (no. of journeys)

	First non-season	First season	Standard Apex	Standard full	Standard reduced	Standard season
Mean	266	179	262	4,976	10,935	10,045
N	413,928	413,928	413,928	413,928	413,928	413,928
Sum	110m	74.1m	108m	2,060m	4,530m	4,160m
Max	563,236	153,118	242,651	2,686,213	21,962,419	20,264,176
Min	–97	–377,958	–12	–3,242	–1,901	–124,795
Standard deviation	3,021	2,266	3,261	22,127	79,185	111,673
Skewness	66	4	33	41	147	102
Kurtosis	8,646	3,245	1,412	3,033	37,166	15,988
p1	0	0	0	5	7	0
p5	0	0	0	21	100	0
p10	0	0	0	36	187	0
p25	0	0	0	99	447	0
p50	6	0	0	519	1,347	41
p75	31	0	6	3,274	5,184	1,524
p90	143	0	195	11,422	19,521	8,945
p95	466	103	521	22,558	42,788	26,484
p99	5,122	4,032	3,768	66,078	159,156	214,776
iqr	31	0	6	3,175	4,737	1,524

Note: The number of observations (N) equals the number of flows multiplied by the number of years for which data on those flows is available.
Source: Oxera analysis.

A dataset with negative or very low values is unsuitable for use as the dependent variable, because forecasting negative journeys is meaningless and it is difficult to predict demand on flows with a small number of journeys. Therefore, this dataset required some processing to be suitable for the econometric analysis. This cleaning process is described in section 2.2 below.

2.2 Clean dataset

The study team has undertaken a process of cleaning the dataset to remove observations that are either outliers or are not useful for this study, such as negative values of revenue and journeys. This process has been discussed and agreed with the study team's academic advisers, Professors Subal Khumbakar and Anindya Banerjee.

There is a trade-off between the analysis being applicable to a higher proportion of journeys made on the rail network and being able to forecast passenger demand accurately. This trade-off arises because of the difficulty in predicting demand on flows with low levels of demand, and hence a cut-off must be applied at the point at which demand can no longer be accurately predicted.

The data cleaning was undertaken in two stages: removal of negative values and zeros; and removal of large percentage changes.

Negative values and zeros

All observations (not flows) containing negative values for journeys and revenues were removed, as well as observations for which there are no journeys or revenues (ie, one or other is coded as negative or zero). Predicting demand where the observed value is small is likely to be misleading because it will increase the likelihood of predicting meaningless negative passenger journeys or revenues.

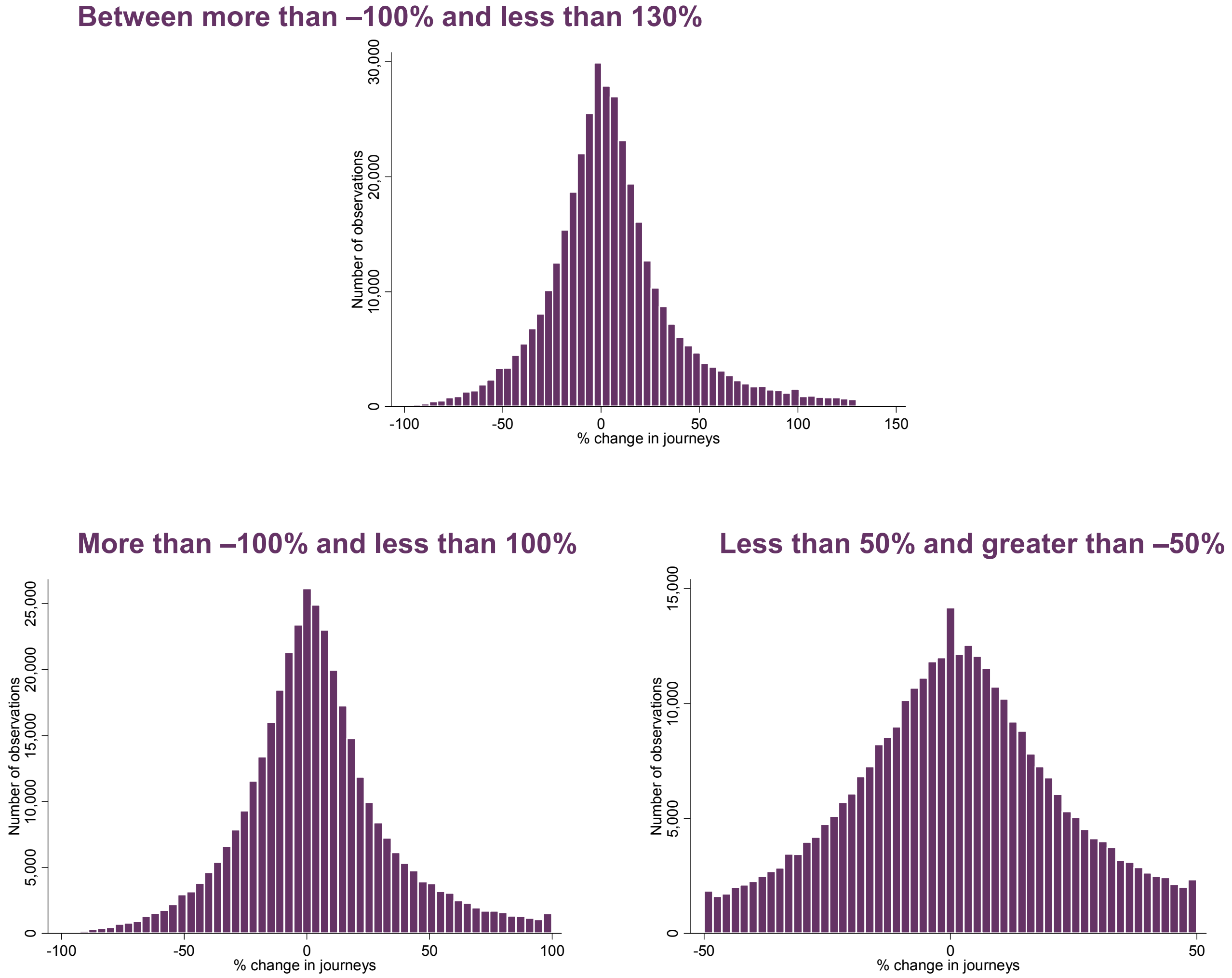
Large percentage changes

Outliers in the data (the precise definition of which is explored below) will exert an undue influence on the parameter estimates obtained in the econometric analysis. For this reason it is important to explore the data thoroughly and remove any outliers before the econometric analysis begins.

The study team has excluded outliers from the dataset by dropping observations with extremely large year-on-year percentage changes. Failure to exclude these extreme changes will artificially skew the distribution of journeys and unduly influence the parameter estimates if left in the dataset. However, the choice of the cut-off point is also sensitive to the loss in the number of observations associated with it and the consistency with which it can be applied to each individual ticket type.

Oxera has applied a cut-off of ±100% year-on-year changes to journeys for all ticket types. This is based on an analysis of graphs showing the results of using different cut-off points. For example, Figure 2.1 below illustrates the difference between using a cut-off of +130%, ±100%, and ±50%.

Figure 2.1 Histogram of percentage change in journeys (standard full)



Between more than –100% and less than 130%

More than –100% and less than 100%

Less than 50% and greater than –50%

Note: Negative and zero levels removed.
Source: Oxera analysis.

It can be seen that using a cut-off of ±50% results in truncated tails of the distribution of flows, while a cut-off of –100% to +130% results in a more symmetric distribution. However, using a cut-off of ±100% has the advantage of being appropriate for all ticket types without a substantial loss in the number of observations. [6] For this reason, it was selected as the recommended cut-off. Many of the dropped flows are likely to be relatively small flows where the addition or removal of a small number of passengers results in a large percentage change.

Table 2.2 provides the same summary statistics as Table 2.1, but in respect of the dataset following completion of the cleaning process.

[6] Applying a cut-off of –100% and +130% would result in a loss of 11.8% of the observations. A cut-off of ±100% would result in a further loss of only 1.4% of the observations. However, using a cut-off of ±50% would result in a further loss of 12.2% of the remaining observations, which is substantially greater than the loss from the ±100% cut-off.

Table 2.2 Summary statistics of cleaned data for journeys by ticket type (no. of journeys)

	First non-season	First season	Standard Apex	Standard full	Standard reduced	Standard season
Mean	457	2,654	999	5,109	10,637	19,658
N	232,909	27,186	100,506	388,868	385,334	197,197
Sum	106m	72m	100m	1,990m	4,100m	3,880m
Max	563,236	153,118	242,651	2,686,213	20,906,447	19,619,643
Min	1	2	1	1	1	1
Standard deviation	3,791	7,847	6,316	21,628	69,663	146,090
Skewness	46	7	17	38	147	70
Kurtosis	4,571	66	364	2,722	40,929	8,113
p1	1	10	1	7	20	2
p5	1	10	2	23	118	10
p10	2	10	4	39	201	41
p25	5	45	31	105	468	246
p50	18	393	117	557	1,399	1,318
p75	72	1,642	352	3,443	5,324	6,013
p90	353	6,592	1,151	11,861	19,726	25,999
p95	1,152	12,500	2,681	23,233	42,735	69,426
p99	9,703	37,764	16,872	67,196	152,936	383,951
iqr	67	1,597	321	3,338	4,856	5,767

Note: The cleaned data excludes zeros, negative values and outliers. The number of observations (N) equals the number of flows multiplied by the number of years for which data on those flows is available.
Source: Oxera analysis.

Table 2.3 shows the distribution of flows available for each ticket type, by the number of years of data available after outliers have been removed. The maximum period covered by the data is from 1990/91–2007/08, therefore a maximum of 18 observations is available for a flow. For example, the standard Apex result of 1,155 flows with two years of data available shows that there are 1,155 flows for which there are two observations—ie, data is available for the flow for only two years. Where there are fewer than 18 years of data, this table does not provide information on where these gaps occur in the data series; it indicates only that they are present.

It is important to note that the removal of observations with zero or negative values results in a substantial reduction in observations for first class, season tickets and standard Apex tickets. This is likely to arise because these tickets are not available on many of the flows in the dataset.

Table 2.3 Number of flows available in each ticket type, by number of years in the revised dataset (excluding zeros, negative values and outliers)

Number of years of data	First non-season	First season	Standard Apex	Standard full	Standard reduced	Standard season
1	1,058	1,970	1,853	8	9	401
2	668	934	1,155	11	19	522
3	512	530	647	37	30	6,612
4	480	346	512	32	33	907
5	465	232	439	70	51	657
6	553	193	417	84	81	545
7	507	179	455	80	115	582
8	593	132	501	31	47	606
9	622	130	580	68	68	752
10	819	104	650	74	69	944
11	1,035	111	890	107	150	1,191
12	1,405	87	1,193	197	210	1,533
13	1,712	113	1,429	438	352	1,687
14	2,011	83	1,104	744	543	2,007
15	1,948	96	707	1,426	995	3,358
16	1,735	86	0	3,180	3,406	454
17	1,658	115	0	6,702	11,638	69
18	1,973	215	0	9,985	5,457	15
Sum	19,754	5,656	12,532	23,274	23,273	22,842

Note: The data in the table corresponds to the number of flows, not the number of observations.
Source: Oxera analysis.

Standard full tickets have 23,274 flows in total, but only 9,985 of these flows have data covering the full 18-year period. Similarly, while the other ticket types also have data on a large number of flows, very few of these cover the full time series. Nevertheless, this distribution of flows should be sufficient to estimate the forecasting framework. Although there are only 15 years of standard Apex data, and only a few flows with the full 18 years of season ticket data, previous analysis has identified lags of up to five years;[7] hence, 15 years of data should be sufficient to estimate these lag structures. See the *Econometric approach* report for details of the chosen econometric methodology, which has been developed to cope with such large-N, small-T datasets. The study team has estimated models by aggregating ticket types in a way that is consistent with the fares simplification introduced by ATOC—ie, anytime (first non-season and standard full), off-peak (standard reduced and standard Apex) and season (first and standard season).

To provide a further sense check, Oxera has analysed the top 20 flows from the revised dataset by ticket type for 2008. These appear sensible, with only those flows that might be expected to be in the top 20 occurring (see Table A2.1).The data for the years before 1994/95 appears to be of dubious quality, with large changes in journeys for no apparent reason (see the *Findings* report for more details). This process has resulted in a dataset that can be used as the dependent variable for the econometric phase of the study. However, to ensure that the results of the econometrics are not unduly influenced by this filtering process,

[7] Oxera, Transport Studies Unit, Centre for Transport and Society and Institute for Transport Studies (2005), 'How do Rail Passengers Respond to Change?'

a sensitivity test has been undertaken to filter out flows where the annual percentage change is more than 500%.

The explanatory variables available for use in the econometric analysis are considered next.

3 Explanatory and control variables

This study focuses on the background growth drivers of rail demand (ie, income, demographics, fares, etc) including competition with other modes, and the extent to which these influence demand for passenger rail travel. However, rail demand is also affected by other variables which will need to be accounted for in order to avoid biasing the parameter estimates of the variables of interest to the study.

The variables which have been considered and for which data has been collected are:

- fares;
- generalised journey time;
- income;
- demographic variables (such as population, working-age population, etc);
- jobs and employment by industry type;
- car ownership, journey time and cost;
- air travel;
- flow distance;
- service quality;
- rail performance.

This dataset, known as The Oxera Arup Dataset (TOAD), provides better coverage of many variables than has previously been available for use in the rail industry in Great Britain. Particular improvements have been made in the coverage of the car mode, with new journey time and cost variables developed for this study, and rail service quality. Service quality indices have been developed for each market segment and the construction of these indices is described in the *Service quality index* report.

Given that this study will produce a forecasting framework, the issue of whether there are forecasts available for the variables of interest is critical (this is considered in section 5). This section considers the data available, and section 4 comments on how this dataset allows the study objectives to be met.

This section sets out the data on the explanatory and control variables that are available to the study team. It includes a detailed commentary on the availability of data for each of the variables of interest, its degree of disaggregation, its sources, and the time period for which data is available. Section 4 considers the extent to which this data can meet the study objectives.

This section is split into two parts: the first considers the explanatory variables that form the focus of the study—ie, economic, socio-demographic, and other modes; the second considers those variables that are not the focus of the study (‘control variables’), but which are included to avoid biasing the estimated elasticities of those variables that are the focus of the study.

Table 3.1 below sets out the historical data, and the level of disaggregation at which it is available.

Table 3.1 Data availability

Category	Variable	Time period	Level of aggregation	Source
Dependent variable				
Rail passenger demand	Passenger journeys	1991–2008	Flow	DeltaRail
Explanatory variable				
Rail characteristics	Fare (revenue divided by journeys)	1991–2008	Flow	DeltaRail
Income variables	Gross disposable income, gross value added (GVA), GVA per employee Disposable income per capita, disposable income	1995–2007 1980–2008	NUTS3 GOR	ONS
Demographic variables	Population, working age population, employment, unemployment, households	1995–2007 1980–2008	NUTS3 GOR	ONS

Gender, age composition, and occupational category	Children (0–15 years);
	Males (16–46), full-time equivalent (FTE);
	Males (16–46), part-time equivalent (PTE);
	Males (16–46), students,; Males (16–46), unemployed; Males (65+);
	Females (16–46), FTE; Females (16–46), PTE; Females (16–46), students; Females (16–46), unemployed; Females (65+)
		1991–2011	TEMPRO zone	TEMPRO

Industry composition	Total jobs (E01), Total households (E02), Primary & secondary education (E03), Higher education (E04), Adult/other education (E05), Accommodation (E06), Retail (E07), Health/Medical (E08), Services (E9), Industry, Construction, & Transport (E10), Restaurants & Bars (E11), Recreation & Sport (E12), Agriculture & Fishing (E13), Other (E14)	1991–2011	TEMPRO zone	TEMPRO
Car ownership	0 cars,1 car, 2 cars, 3+cars, total cars	1991–2011	TEMPRO zone	TEMPRO
Car journey/cost	Car journey time, car journey cost	1991–2008	TEMPRO zone	Oxera
Pump price	Price per litre	1991-2008	National	ONS
Air travel	Fare	1996–2002	Route	CAA
Airport throughput	Passengers per year	1991-2007	Airport	DfT
Control variables				
Rail characteristics	Generalised journey time (GJT), distance	1991–2008	Flow	DeltaRail
Service quality	Station quality index, train quality index	1997–2008	Segment	Oxera
Performance	Passenger Performance Measure (PPM) and PEARS results	1997–2008	British Rail Sector and service group since 2001	Network Rail

Note: NUTS3 is a sub-local authority area, a classification used by Eurostat and other EU bodies; GOR, government office region. Where appropriate, variables are available at both the origin and destination.
Source: Oxera.

The availability of forecasts for these variables will be critical in determining whether they can be usefully included in the forecasting framework. Section 5 discusses whether the variables can be meaningfully forecast at the level of aggregation required. The use of the forecasting framework to produce forecasts is discussed in the *Guidance* report.

3.1 Variables

The econometric analysis is likely to need to include a number of variables, many of which are interrelated (see the *Model specification* report). These variables, which are based on economic theory and industry knowledge, include the following.

– **Income**: the most preferable measure is likely to vary by journey purpose. Increased income is likely to result in greater expenditure on rail travel.

– **Population**: there are a number of considerations around the population variable, such as household size and number. Increased population may imply more passengers using each mode of transport, and hence imply greater rail demand.

– **Employment**: a number of types of employment exist (eg, white or blue collar, employment in different industries, etc). Increased employment may result in greater commuting to work, and hence more rail demand. The link to income is also important.

– **Other modes (cost, time, reliability)**: other modes to consider include air, car and bus/coach; taxi is a possible alternative in some cases. Passengers may choose between different modes on the basis of these characteristics, and so they are important to capture within a forecasting framework.

– **Fares**: it is necessary to capture the interactions between the ticket types. Passengers choose between modes, at least to some extent, on the basis of fare.

– **'Hard' endogenous factors**: these include crowding, reliability and generalised journey time (GJT) and affect both value for money and the generalised cost of travel by rail; hence, they may affect rail demand.

– **Other aspects of service quality**: station quality and train quality are reviewed separately in the *Service quality* report. Service quality is likely to affect the value for money of the service, and hence demand for rail travel.

The first five of these are of particular interest to the study. The others (6 and 7) are included in the econometric analysis to avoid biasing the elasticity estimates of the variables of interest. These variables are further discussed below.

3.2 Variables of interest

3.2.1 Income

The PDFH suggests that measures of economic performance, such as personal income, are likely to be important drivers of rail demand.[8] However, economic theory suggests that the appropriate measure of income may differ, according to the purpose of the journey.

– For leisure, disposable income per capita, deflated by a consumer price index (such as the retail price index or consumer price index), is likely to be the preferred measure of income.

– For business, workplace GVA per employee, deflated by an industry-weighted average of the GDP deflator, is the preferred theoretical definition.

[8] ATOC (2009), op. cit.

– For commuting, employment is more likely to be the driver than income, although income may be a proxy for the changing characteristics of employment in Great Britain—for example, the gradual move towards more 'white collar' employment.

However, whether these theoretical differences can be replicated in the data is an empirical issue.

The Office of National Statistics (ONS) produces GVA figures on a workplace basis at the GOR, NUTS2 (local authority) and NUTS3 (sub-local authority) levels. As GVA is published at the GOR level by 15 industries, it is possible to deflate regional GVA by an industry-weighted average of the GDP deflator. However, for consistency, the study team has deflated all monetary series by the RPI.

The study team's understanding is that the current forecasts supplied to the PDFC by Oxford Economics are based on GVA on an employment basis, divided by a residence-based population to provide income per capita. To the extent that people do not work in the region in which they live, it is not entirely clear what this variable is measuring.

The ONS also produces estimates of disposable income on a residence basis, at the NUTS3 level. Therefore, theoretically satisfactory measures of income are available for use within the study. However, their forecastability is of considerable importance to this study, and this will be considered further in section 5.

The length of the time series of the data is of considerable importance since the shorter the time series, the fewer observations are available for use in the estimation of the model, and hence the fewer degrees of freedom are available. For example, although GVA is available for all years of the dataset, the series of disposable income per capita at origin finishes one year before the demand data, and using this variable therefore reduces the estimation period by one year.

The next issue concerns the most appropriate level of aggregation to use for the income variables. The key questions here are: at what level is data actually collected, and which levels are just a scaling of higher levels of aggregation; and at what level are robust forecasts available?

In answer to the first question, this differs between GVA and disposable income, with GVA being scaled down from the GOR level, but disposable income being available at the NUTS3 level without being scaled down. This implies that the income variables used may be at different levels of aggregation. This does not pose significant problems for the econometric estimation since in a number of cases variables at different levels of aggregation will be used. However, using data at a higher level of aggregation does result in a loss of information relative to more disaggregate data; hence, all else being equal, more disaggregate data is preferred.

The answer to the second question is more problematic and needs to be carefully considered when producing forecasts. See section 4 for more details.

A final issue arises of how, if robust data is available only at a relatively high level of aggregation, it can be assigned to the origin and destination of the rail trips of interest. This can be done either by assuming that income per person is constant across the region, or by using the socio-economic and -demographic characteristics of, for example, the TEMPRO zone in which the origin or destination is categorised, in order to interpolate what the value may be.

3.2.2 Population

The 'Rail Trends Report' considers in detail the changes in population in Great Britain, and finds the following general trends:[9]

– increases in net migration;
– an increasing proportion of elderly people;
– increases in long-distance commuting;
– a faster increase in population, in general, outside rail station catchment areas than within rail station catchment areas.

The PDFH also discusses issues relating to population:

– whether population dispersion (density, sparsity) affects rail demand (considered only in relation to forecasting absolute levels of rail demand);
– whether interactions between origin and destination populations increase with the product of the populations;
– the impact of population type (age, socio-economic classifications) on rail demand (ignored by the PDFH due to the long-run nature of changes).

The PDFH concludes that an elasticity of unity should be used for the number of trips originating in a zone, and that, for commuting trips, the relevant variable is a change in population *relative* to a parallel route.

Much of the population (and employment) data used in this study has been taken from TEMPRO (v 5.4) (see Box 3.2 for a summary of TEMPRO and the data that may be sourced from it).

Box 3.2 TEMPRO data

TEMPRO provides modelled data on many socio-economic/-demographic variables using the National Trip End Model to produce forecasts of the number of trips by zone.

A key assumption underlying TEMPRO is that of the constant generalised cost. While this is not particularly important for many of the socio-economic variables, it is important for forecasts of car ownership.

TEMPRO uses inputs from the ONS, particularly the Census. Data between Census years is not observed, but is modelled, and is therefore dependent on the robustness of the underlying model.

The study has also used data from the ONS for demographic variables such as population, working age population, employment, unemployment and number of households. These are also available at the GOR and NUTS2 levels. This again raises the issue of the appropriate level of aggregation, discussed in Box 3.1, which is important within this forecasting framework.

3.2.3 Employment

Employment is likely to be a particularly important driver of the demand for travel by season ticket because as more people are employed, more people need to travel to work. Since some of them might be expected to use the train, this may result in an increase in commuter demand. However, it is important to recognise the difference between 'workers' (ie, the number of people in employment) and 'jobs' (the number of roles filled, including part-time roles), due to the different implications that these may have for the demand for rail travel. This is because a worker filling one job may be more likely to use the train to travel to their workplace than someone with many jobs in various locations.

[9] Ove Arup & Partners Ltd (2009), 'Rail Trends Report', March 30th.

Of equal importance to the availability of historical data is that of forecasts. When considering forecasts of employment, it is important to remember that macroeconomic forecasts typically assume that employment responds to income changes with a lag. Although there is unlikely to be a lag in actual (as opposed to forecast) responses going into a recession, there is likely to be a lag in employment growth as the economy emerges: therefore, forecasts received from Oxford Economics, for example, will reflect this effect already.

It is unclear whether changes in types of employment can also explain changes in rail demand. This was explored in the 'Rail Trends Report'.[10]

The study team has acquired data on employment from both ONS and TEMPRO. This is available at a detailed level and is broken down by industry from TEMPRO, and at a higher level of spatial aggregation from the ONS. Therefore, the issue of aggregation arises again here.

3.2.4 Other modes

Three major modes compete with rail travel—car, aviation, and bus/coach—and these are considered in turn below.

Car

The conceptual framework developed by the study team is set out in the *Model specification* report. In brief, car ownership reflects the long-term costs of owning and operating a car (available from TEMPRO) and short-run changes to the use of cars are a response to changes in short-run marginal costs and journey time.

The study team has enhanced the coverage of car variables by developing car journey time and car cost variables, and constructing a matrix of car journey times for each year, and origin–destination pair in the rail demand dataset, as follows:

- speed–flow curves and data from TEMPRO were used to create an estimate of journey time within each TEMPRO zone in Great Britain;
- geographic information systems (GIS) were used to generate the route that a car driver is likely to take (using only trunk roads where possible) from the origin to the destination station;
- from the GIS mapping, a list was generated of the TEMPRO zones that the driver is likely to pass through, and the proportion of distance travelled through that zone during the journey was used to create a weighted average of car journey time increases by flow.

This methodology has been used to create forecasts of journey time by flow to 2041 (when the forecasts in TEMPRO end).

The car journey time variable uses actual speed data for one year (2003), and a combination of speed–flow relationships from the National Transport Model (NTM) and TEMPRO trip growth estimates, assuming that the fastest available road is used by those taking the car alternative.

Full details of how the car cost variable has been calculated are included in Appendix 3. A brief description is as follows:

- the NTM contains speed–cost relationships that translate km/h to p/km in order to calculate the monetary cost of road journeys. Adjusting these speed–cost curves to represent changes in the cost of road journeys over time (incorporating movement in both the price of fuel and vehicle efficiency), and combining them with journey times on

[10] Ove Arup & Partners (2009), op. cit.

an origin–destination basis, generates an origin–destination-specific car cost time-series variable;

- a car journey time growth variable has been separately generated for use in the study, as explained above. This takes the station-to-station road journey as equivalent to a given rail trip and models increases in journey time based on changes in traffic levels. These journey time growth rates were applied to NTM 2000 base journey times to give modelled journey times. These were then used to calculate journey speed and, through the time series of speed–cost curves, journey cost.

This is a substantial development on previous data. Current PDFH practice is to focus on the proportion of households without a car as a driver of rail demand, along with fuel price and journey time. It is conceivable—and an empirical issue—that rail use does not decline uniformly with the number of cars per household in an area. However, it might be difficult to disentangle high incomes associated with higher rail use and larger numbers of cars per household. The revised data allows for a more detailed understanding of which car variables affect the demand for rail travel.

The PDFH recommends that, for business and first class passengers, car availability has reached saturation point, and sets the relevant elasticity to zero. It also recommends an exponential function such that, as the proportion of households without a car tends to zero, the effect on rail demand also tends to zero. This is important since it may point to market saturation (discussed more generally in the *Model specification* report).

The study team noted in the proposal that a separate car availability variable may not be conceptually accurate since the car ownership decision is also a function of other variables under consideration, including road congestion and car running costs. The preferred approach implicitly assumes that short-run marginal costs (ie, fuel cost) drives car usage, but that fixed and long-run marginal costs (eg, purchase price and servicing) drive purchase decisions. This is consistent with the data developed for this study.

In summary, the study team has invested considerable time in enhancing the data coverage of cars, given the importance of the car as an alternative to rail travel. This data should allow for a more robust coverage of the cross-elasticities from this mode to be applied to rail demand.

Aviation

A number of recent events are important to consider when analysing the impact of aviation on the demand for passenger rail travel, including the following.

- Increases in flight access times associated with the additional security measures introduced since September 11th.

- The frequency of flights on a route, and the associated punctuality levels, are likely to affect rail demand.

- The level of fares charged over time. This may be particularly important given the move by all domestic carriers to yield management over the period in question.

However, rail is often a complement to air travel (eg, passengers often travel by train to an airport). Current PDFH guidance for travel to and from airports is to include an elasticity of unity to passenger volumes at airports. This is standard practice, although it neglects those instances where a policy requirement exists to increase the market share of rail at an airport, and where a greater elasticity might be justifiable. In addition, the PDFH suggests that, for airports affected by road congestion in the vicinity, the elasticity might also be higher. Whether elasticities differ with respect to airport access and the size of the airport throughput elasticity are essentially empirical questions.

The Civil Aviation Authority (CAA) provides data on air frequency and punctuality for all domestic passenger routes on a monthly basis. To supplement this, the DfT has provided fares data extracted from CAA passenger surveys for 14 routes within the UK for the years 1996, and 1999–2002. Therefore, an exercise in interpolation has been undertaken, whereby if a route has more than seven observations, the remaining observations are interpolated. However, no extrapolation has been carried out, which means that a full time series is not available for each route. This approach represents a trade-off between having more data available for the econometric analysis and the robustness of that data.

The next issue concerns the matching of the data for the flight routes with that of the rail routes. The cost of air travel should only affect rail demand where flying is a viable alternative to travelling by train. Having collected data on air fares from a number of sources, it was not clear for which flows this data should be matched.

CAA passenger surveys record the ultimate local authority (LA) origins and destinations of air passengers, as well as the airports used. This data was therefore used to develop ‘catchment areas’ around airports at both the origin and destination ends of a journey. Rail flows originating in the catchment of one airport and terminating in the catchment of another can be identified and matched with the relevant fares data for that air flow.

The CAA survey is annual, but is conducted on the basis of a different set of airports each year, in addition to a set of core airports. Therefore, data from a number of different surveys was used, with the latest available survey being used for each airport.

The CAA survey was not usable in the format in which it was delivered. Foreign flows, including those to Northern Ireland, the Isle of Man and the Channel Islands, had to be removed, and a look-up generated to convert LA names in the CAA survey to be compatible with those used for stations. In addition, a number of filters had to be applied to remove apparent anomalies, once rail and air flows had been matched.

Rail flows were matched to air flows of matching LA origin and destination. For example, if an individual were flying from Edinburgh to London Heathrow and their ultimate destination was Swindon, the LHR–EDI fare would be matched with the Edinburgh–Swindon rail flow. This approach, however, led to very large catchments around airports, and to many apparently anomalous results. The first step in removing these anomalies was to drop LA origin–destination pairs if they constituted less than 0.1% of total passengers using either the origin or destination airport.

It was not possible, however, through this process, to remove a number of somewhat surprising results among intra-Scottish flows. The Highlands & Islands area of Scotland is classified as one LA, and this led to rail flows being matched incorrectly with air flows. These rail and air flows were all de-linked, but were not removed from the air origin–destination matrix since this would have reduced the number of passengers flying from each airport and thus would have affected the filtering of small flows.

Some anomalies still remained after these adjustments, particularly where survey respondents had given the same origin and destination when they were undertaking a return journey. Inspection of the data suggested that implementing a cut-off of 100 miles would remove these flows, leaving only those rail journeys that might realistically be expected to compete with air travel.

In summary, the underlying data on air fares is not as comprehensive as might be hoped due to the sporadic nature of the surveys on some routes.

The final question that needs to be reviewed is whether robust forecasts for this data are available. This is addressed in section 5.

Bus/coach

As is the case in aviation, long-distance coach operators now use yield management systems to vary fares in line with available capacity. Therefore, a survey-based approach is most likely to pay dividends in terms of obtaining a time series of fares.

Local bus fare structures are typically simpler, although there are clearly issues in areas under the control of a Passenger Transport Executive (PTE).

A number of factors may be important when considering bus and coach competition:

- there may be an observable effect from the introduction of increasing discounts on off-peak concessionary fares;
- headway is also likely to be important, according to the PDFH;
- as with aviation, elasticities are expected to vary with the market shares of the two modes.

Data on bus vehicle-kilometres and bus passenger journeys is available at the level of individual transport authorities, but little fares data is available on the bus and coach mode. This may be a key area for future investigation, particularly given the increasing use of yield management by long-distance coach operators, who may be competing for some sections of the rail market.

In summary, the quality of the data on competing modes is variable, with the study team considerably enhancing the quality of the coverage of the car, but coverage of other modes is of poorer quality due to data constraints.

3.2.5 Fares

While not ideal, the study will continue with the practice of assuming that fares are represented by revenues (deflated by RPI), divided by journeys. The issue with this variable is that movements in within it will, by definition, capture not only fare changes in real terms, but also any switching at different rates between ticket types, and switching to and from rail.

The six ticket types for which the study team has data are first season, first non-season, standard season, standard anytime, standard off-peak, and advance purchase. The ‘Rail Trends Report’ suggests that the types of trips taken using these tickets have changed over time; for example, in the recent recession, there has been a tendency for business users to take advantage of cheaper advance purchase fares. [11]

It is interesting to note that these ticket types are consistent with the eight ticket types being used for MOIRA replacement.

The revenue data from DeltaRail has been cleaned to remove outliers, using the same procedure described in section 2.1 for journeys. Observations that record negative revenues and greater than 100% year-on-year changes have been dropped. Data on revenues has been used to calculate fares by dividing revenues by journeys. Therefore, a separate fare is available for each flow, for each year, for each ticket type. T ensure compatibility across all years in the dataset, the analysis has been conducted on three ticket types (which are compatible with the simplified structure introduced by ATOC): full fare (incorporating standard class full fare and first class non-season tickets); reduced fare (incorporating standard class reduced and Apex tickets); and season ticket (which include first class and standard class season tickets).

[11] Ove Arup & Partners (2009), op. cit.

3.3 Control variables

It is important to account for variables other than those of direct interest, to avoid biasing the parameter estimates of those variables. This section discusses those variables which the study team has included in the dataset, and which it will use in order to minimise omitted variables bias.

3.3.1 ‘Hard’ endogenous factors

‘Hard’ endogenous factors are those which the railway industry can influence and measure relatively easily. This section considers GJT, performance and crowding.

Generalised journey time

GJT is defined in the same way as in the PDFH and the ‘old’ MOIRA—ie, it is the sum of total station-to-station journey time, the service interval penalty, and the sum of interchange penalties. As interchange and service interval penalties in the PDFH change by ticket type, GJT will be different for the same flow for passengers travelling on different tickets. For example, service interval penalties currently differ for anytime and season ticket users in comparison with off-peak ticket users, while interchange penalties differ for season ticket users in comparison with anytime and off-peak ticket users.

Table 3.2 shows how the journey types from the GJT data have been matched to the ticket types from the demand data:

Table 3.2 Journey types

Ticket types	Journey type
Standard full	Full
Standard reduced	Reduced
Standard season	Season
Standard Apex	Reduced
First season	Season
First non-season	Full

Source: Oxera analysis.

Performance

There are a number of definitions of performance:

- PPM will be positively correlated with rail demand, but ignores certain factors (eg, days suffering significant disruption);
- average minutes’ lateness (AML) refers to the difference between the arrival time at destinations, and the publicly available timetable; this will be negatively correlated with rail demand;
- ‘delay minutes’ are measured by monitoring points along a route, and capture the difference between the actual times achieved and the working timetable. They will also be negatively correlated with rail demand.

The rate of cancellations, and the distribution of minutes’ lateness (eg, the proportion of long delays), are also likely to affect rail demand.

The performance data provided by Network Rail has been pooled into three groups: London and the South East, regional, and long-distance/intercity. An aggregate estimate of these three groups is also available. There are two variables of interest: the number of services run, and the number of services arriving on time. These two variables have been used to produce the PPM. A moving annual average of PPM is also available. This data has been matched to the relevant flows to provide a time-series measure of performance, at a relatively high level. Performance data is also available at the TOC level (from the PEARS system), which provides a more disaggregate measure but which is available only after 1999. Therefore, the issue of which is the most appropriate level of aggregation arises again.

Crowding

Crowding will not be modelled in the forecasting framework but will be captured through the inclusion of the service quality index (of which crowding is an element). This implies that forecasts account for crowding. However, changes in crowding and their impact on demand will need to be modelled separately (see the *Service quality index* and *Guidance* reports*)*.

3.3.2 Service quality

The quality of the service may be expected to influence the perceived value for money of rail relative to other modes, and therefore it is important to attempt to capture its effect within the econometric modelling. However, measuring quality of service is not straightforward.

The study team has created three indices of service quality for each market segment: train, station and overall. The data for these indices has been taken from the NPS and combined with willingness-to-pay data from external sources.

There are two key issues with the service quality indices:

– the three indices are highly correlated—with a correlation coefficient of greater than 0.99 in many cases. This makes it difficult to include more than one index in the econometrics and obtain reliable estimates of the effect of each index;

– during the interviews conducted by Oxera with experts from across the industry, performance was identified is one of the key factors in determining the responses to the NPS. To avoid double-counting the effect of performance (which will be dealt with elsewhere in the model), it is important to remove the impact of performance on the service quality indices.

The creation of these indices is detailed in the *Service quality index* report, together with a discussion of the issues and assumptions made during this analysis.

3.3.3 Summary

This section has discussed the data available for use in the econometric estimation phase of the study, and whether it can be used to estimate the models identified in the *Model Specification* report. Section 4 assesses the extent to which the data allows the study objectives to be met, while section 5 discusses how forecasts could be generated using the framework.

In looking at the data which was available to the study team for use in both the market segmentation analysis and the econometric phase of the study, the key question is whether the available data is suitable for estimating the models identified in the *Model specification* report, and hence for satisfying the study objectives. While there are still a number of weaknesses in the data (examined below), the data available has been used to estimate the econometric models and to produce elasticities. These can be used to meet the study requirements by producing a robust elasticity-based framework in which as many elasticities as possible are freely estimated within the econometric analysis.

The next section considers whether the available data can be forecast, while section 5 provides recommendations for improving the evidence base, given the weaknesses which remain.

4 Forecasting

One of the key issues in building a forecasting framework is whether forecasts are available for the data which is used to estimate the framework; if not, the framework is unlikely to be fit for purpose. Sections 2 and 3 examined the historical data available for estimation, while this section looks at whether forecasts are available for the variables used in the econometric analysis.

The conclusion is that forecasts are available for most of the variables that would be required to generate forecasts of rail demand. The sources of these forecasts are varied, ranging from publicly available sources such as TEMPRO to macroeconomic forecasters such as Oxford Economics, or are likely to be produced privately by the train operating companies or the DfT. The application of the forecasting framework is set out in the *Guidance* report.

Potential improvements which might be made to the evidence base are considered in section 5.

Section 3 considered the historical data available to the study team; however, the purpose of the study is to generate a forecasting framework. For this framework to be useful, users will need to be able to use forecasts of the explanatory variables. This section examines which variables can be forecast on a comparable basis, and which cannot (see Table 4.1).

Table 4.1 Data forecastability

Category	Variable	Forecastable?	Where are forecasts available from?
Rail characteristics	Revenue, GJT	✓	TOC/DfT
Income variables	Gross disposable income, GVA Disposable income per capita, disposable income	✓	Macroeconomic forecaster—eg, Oxford Economics
Demographic variables	Population, working age population, employment, unemployment, households, disposable income	✓	TEMPRO

Gender, age composition, and occupational category	Children (0–15 years), Males (16–46), FTE;
	Males (16–46), PTE; Males (16–46), Students;
	Males (16–46), Unemployed; Males (65+); Females (16–46), FTE; Females (16–46), PTE; Females (16–46), Students;
	Females (16–46), Unemployed; Females (65+)
		✓	TEMPRO

Industry composition	Total jobs (E01), Total households (E02), Primary & secondary education (E03), Higher education (E04), Adult/other education (E05), Accommodation (E06), Retail (E07), Health/Medical (E08), Services (E9), Industry, Construction, & Transport (E10), Restaurants & Bars (E11), Recreation & Sport (E12), Agriculture & Fishing (E13), Other (E14)	✓	TEMPRO
Car ownership	0 cars,1 car, 2 cars, 3+ cars, total cars	✓	TEMPRO
Car journey/cost	Car journey time, car journey cost	✓	DfT[1]
Service quality	Station quality index, train quality index	✓	TOC/DfT
Performance	PPM	✓	Network Rail/TOC
Air travel	Fare, airport throughput	x/✓	CAA

Note: Passenger journeys are the output from the forecasting framework. This can then be combined with yield forecasts to produce revenue forecasts.[1] Other proxies are also available, such as pump price forecasts. Source: Oxera.

Table 4.1 indicates that most of the required variables can be forecast, either from publicly available sources (eg, TEMPRO), specialist macroeconomic forecasters (eg, Oxford Economics), or by TOCs and/or the DfT, depending on the circumstances under which the variable is required.

TEMPRO provides forecasts that differ substantially from many of the other forecasts set out above, insofar as they are long-term and lag significant economic events by up to two years. Other forecasts are updated more rapidly in light of changing economic events.

The variables that may not be forecastable using public sources are primarily for other modes. These may be important, depending on the flow in question and the degree of competition from other modes, in which case a view will need to be taken on the likely evolution of the variable over the time period of the forecast.

5 How does the current data meet the study objectives?

The objectives of the study are to:

- update and estimate the fares and background growth elasticities contained within the PDFH;
- explore the use of innovative or alternative econometric techniques;
- re-specify and extend the core elasticity-based framework;
- improve the underlying data.

The data collated for this study offers a substantial improvement on the data which has previously been available, by improving the coverage of alternative modes and service quality, and by increasing the number of years for which data is available; it therefore meets the fourth objective.

The other objectives of the study are not directly data-related; however, it is important that the data offers as solid a foundation as possible for the econometric analysis. The study team believes that the data collected for this study has enabled the other study objectives to be met.

The availability of forecasts for the data is considered in section 6, while section 7 provides thoughts on how the evidence base might be improved.

This section provides a commentary on whether the data is suitable to meet the study objectives, and whether it is able to support the models identified in the *Model specification* report.

The study objectives are:[12]

- to update and estimate the fares and background growth elasticities contained within the PDFH;
- to explore the use of innovative or alternative econometric techniques;
- to re-specify and extend the core elasticity-based framework;
- to improve the underlying data.

There are a number of areas in which the existing evidence base is weak—primarily in the coverage of competing modes. The study team has taken steps to enhance the evidence base where possible, given the timescales of the study. These steps are discussed in detail in section 3 above.

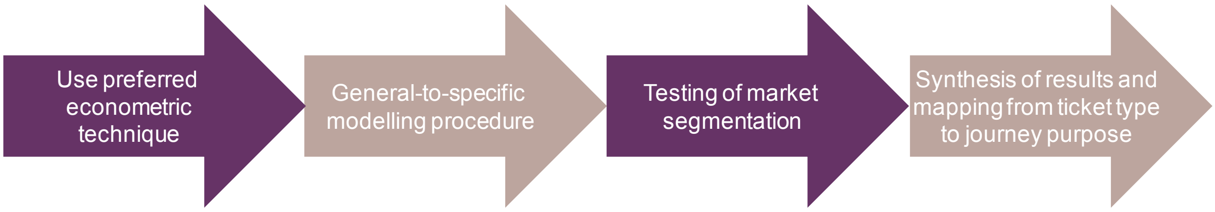
The underlying data has been substantially improved; particular enhancements include coverage of car cost and journey times, together with service quality.

The other objectives will be met through the econometric process, but it is important that the data can provide the base for this analysis. The study team believes that the other objectives have been met, and that while there are still weaknesses in the data (as identified in section 3), the data available to this study is a substantial improvement on that used in previous studies.

The econometric process is illustrated in Figure 5.1. The hypotheses identified in the market segmentation exercise are tested in the econometric analysis. Following this, the preferred model specification and econometric technique are identified, and a general-to-specific modelling procedure followed. In addition to this process, the study team has taken advice from its panel of experts, particularly Professor Banerjee, on the most appropriate way to deal with some of the modelling issues, such as the stationarity of the data and the unbalanced nature of the panel. (See the *Econometric approach* report for more details.)

[12] Department for Transport (2008), ‘Rail Passenger Demand Forecasting: Revisiting the Elasticity-Based Framework Request for Proposal and Statement of Requirement’, July, pp. 12–13.

Figure 5.1 Econometric analysis process


Use preferred econometric technique
General-to-specific modelling procedure
Testing of market segmentation
Synthesis of results and mapping from ticket type to journey purpose

Source: Oxera.

Section 6 provides details on the remaining weaknesses in the data, and what action can be taken to fill in the remaining gaps.

6 Recommendations for improving the evidence base

This section considers those instances where the evidence base would benefit from further expansion. The key areas for which substantial weaknesses in the evidence base remain are:

- disaggregated performance data before 1999;
- service quality measures before 1999;
- coverage of the price and performance of other modes;
- a fares series which is not calculated as revenue divided by journeys.

The contents of the final dataset, and what it can and cannot be used for, are set out in section 7.

Possible enhancements to the evidence base could include measurements of disaggregated rail performance and service quality before 1999, improvements in the coverage of the price and performance of other modes (ie, bus and coach), the total time of aviation journeys, and the development of a fares series.

The use of yield as a measure of fare (ie, revenue/journeys) introduces certain issues (such as endogeneity) into the econometric analysis because 'journeys' appears on both the left- and right-hand sides of the equation.[13] The development of a fares series, which is collected separately from journeys, would provide a useful supplement to the data available for use in studies such as this. However, it is worth noting that the value of the dataset will be improved by keeping it up to date as new data on the variables contained within it become available.

Section 7 sets out the final dataset which the econometric analysis has used in the next phase of the study.

[13] Endogeneity of the explanatory variables means that certain assumptions of the classical regression model are no longer met, and hence instrumental variables procedures need to be used instead. However, the use of instrumental variables is not without its problems (such as that of weak instruments), which can mean that parameters estimated using this method can be unreliable.

7 Requirement for final dataset

The final dataset for this study provides a wide range of variables—covering economic, demographic, and competing modes—and other control variables, such as performance, service quality, etc. This offers a rich dataset which can be used to estimate a wide range of econometric models.

The study team has considered a range of panel data models which could be used in this study, taking into account the suitability of the data for use with these modelling techniques.

However, the dataset cannot be used to estimate conventional multi-modal models due to the different dependent variables and data structure.

Some conclusions on the data capability are offered in section 8, together with presentation of the next steps in the analysis.

The final dataset contains variables on rail and other modes of transport, as well as data on socio-economic variables such as income, employment and population. The dataset is an unbalanced panel—ie, not all of the variables will cover the same length of time.

The dependent variable contains data on passenger journeys (taken from LENNON) between 1990/91 and 2007/08. Appendix 1 lists all the variables contained in the final dataset, their coverage, and any associated problems or issues.

Variables of particular interest that are new to this study include car cost and car journey time, along with train and station quality indices. These were discussed at length in section 3 and so are not considered further here.

As set out in sections 3 and 5, a number of weaknesses remain in the dataset, such as a lack of disaggregate performance data for the period before 1999, on the price, frequency and reliability of bus and coach journeys, and on the overall cost of air travel; and the endogeneity of the fares variable.

This dataset has enabled the study to meet the study objectives—in particular, the requirement to estimate as many elasticities within the econometric framework as possible—by having variables representative of the alternative modes and socio-economic factors of interest.

The other study objectives, including the following, can be met using the available data:

- exploring the technical specification of the forecasting framework;
- re-specifying and extending the core elasticity-based framework;
- improving the underlying data;
- investigating the use of new econometric techniques;
- improving the market segmentation;
- investigating whether there is any support for variable elasticities;
- specifying dynamic models.

The final dataset offers considerable scope for extending the elasticity-based framework through the use of more variables and improved data. Since the publication of the PDFH v4.1 in 2004, there has been considerable development in both theoretical and applied panel data econometrics. This study has considered a wide range of possible models and modelling techniques, a detailed discussion of which is presented in the *Econometric approach* report.

However, this dataset is not suitable for use in estimating conventional four-stage multi-modal models due to the different dependent variables used in this modelling approach.

8 Conclusions

This report has described the construction of the dataset which will be used in the econometric analysis. The dataset has been considerably enhanced, thus meeting one of the core study objectives, and provides a solid base for the econometric analysis and to enable the study to meet its other objectives.

This report has considered the construction of the dataset for use in the econometric analysis of this study, and has examined whether this dataset allows the study to meet its objectives.

The dataset has been considerably enhanced, with a longer time series of data and a number of new variables created for this study. Of particular note are the improvements to the coverage of car journey time and car cost, given the importance of this competitive mode of transport.

This dataset enables the study objectives to be met by providing a solid base for the econometric analysis to generate the elasticity estimates. The economic model estimated by the study team is set out in the *Model specification* report, while the econometric modelling that has been undertaken is detailed in the *Econometric approach* report.

A1 Data sources

Variable	Time period covered	Units	Price base (if appropriate)	Level of aggregation	Modelled or measured	Source	Comments/issues
Passenger journeys	1991–2008	Passenger journeys	n/a	Flow	Measured[1]	DeltaRail	1994 is less robust than other years due to widespread strikes by signallers in that year Potential issues in 2005 due to journey allocation problems in LENNON
Revenue	1991–2008	£m	Real	Flow		DeltaRail	The data is at constant 2007 prices
GJT	1991–2008	Minutes	n/a	Flow	Measured[2]	DeltaRail	There are 442 negative GJT observations. In only 130 of these does the flow have the same origin and destination
Distance	Same for all years	Kilometres	n/a	Flow	Measured	DeltaRail	
Gross disposable income	1995–2007	£m	Real	NUTS3	Measured	ONS	There are 10,040 zero observations. The data is at constant 2007 prices. ONS code QWND
Disposable income per capita	1995–2007	£'000s	Real	NUTS3	Measured	ONS	The data is at constant 2007 prices. ONS code C8G5
GVA	1980–2008	£m	Real	GOR	Measured	ONS	The data is at constant 2007 prices. ONS code ABML.
Population	1980–2008	No. of people	n/a	GOR	Measured	Oxford Economics	Oxford Economics, taken from the ONS
Working age population	1980–2008	No. of people	n/a	GOR	Measured	Oxford Economics	Oxford Economics, taken from the ONS
Employment	1980–2008	No. of people	n/a	GOR	Measured	ONS	ONS code DYDC
Unemployment	1980–2008	No. of people	n/a	GOR	Measured	ONS	ONS code BCJD
Households	1980–2008	No. of households	n/a	GOR	Measured	ONS	http://www.communities.gov.uk/housing/housingresearch/housingstatistics/housingstatisticsby/householdestimates/livetables-households/
Disposable income	1980–2008	£m	Real	GOR	Measured	ONS	The data is at constant 2007 prices
Children (0–15 years)	1991–2011	No. of children	n/a	TEMPRO zone	Modelled	TEMPRO	
Males (16–46)—FTE	1991–2011	No. of males	n/a	TEMPRO zone	Modelled	TEMPRO	
Males (16–46)—PTE	1991–2011	No. of males	n/a	TEMPRO zone	Modelled	TEMPRO	
Males (16–46)—Students	1991–2011	No. of males	n/a	TEMPRO zone	Modelled	TEMPRO	
Males (16–46)—Unemployed	1991–2011	No. of males	n/a	TEMPRO zone	Modelled	TEMPRO	
Males 65+	1991–2011	No. of males	n/a	TEMPRO zone	Modelled	TEMPRO	
Females (16–46)—FTE	1991–2011	No. of females	n/a	TEMPRO zone	Modelled	TEMPRO	

Variable	Time period covered	Units	Price base (if appropriate)	Level of aggregation	Modelled or measured	Source	Comments/issues
Females (16–46)—PTE	1991–2011	No. of females	n/a	TEMPRO zone	Modelled	TEMPRO	
Females (16-46) – Students	1991–2011	No. of females	n/a	TEMPRO zone	Modelled	TEMPRO	
Females (16-46) – Unemployed	1991–2011	No. of females	n/a	TEMPRO zone	Modelled	TEMPRO	
Females 65+	1991–2011	No. of females	n/a	TEMPRO zone	Modelled	TEMPRO	
Total jobs (E01)	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	
Total households (E02)	1991–2011	No. of households	n/a	TEMPRO zone	Modelled	TEMPRO	
Primary & secondary education (E03)	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	
Higher education (E04)	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	
Adult/ other education (E05)	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	There are 180 zero observations
Accommodation (E06)	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	There are 22 negative observations
Retail	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	
Health/Medical	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	
Services	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	
Industry, construction, & transport	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	
Restaurants & bars	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	There are 40 negative observations
Recreation & sport	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	There are 10 negative observations
Agriculture & fishing	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	
Other	1991–2011	No. of jobs	n/a	TEMPRO zone	Modelled	TEMPRO	There are 70 negative observations There are 38 zero observations
0 cars	1991–2011	No. of households	n/a	TEMPRO zone	Modelled	TEMPRO	
1 car	1991–2011	No. of households	n/a	TEMPRO zone	Modelled	TEMPRO	
2 cars	1991–2011	No. of households	n/a	TEMPRO zone	Modelled	TEMPRO	
3+cars	1991–2011	No. of households	n/a	TEMPRO zone	Modelled	TEMPRO	
Total cars	1991–2011	Cars	n/a	TEMPRO zone	Modelled	TEMPRO	
Bus vehicle	1991–2007	Kilometres	n/a	Transport authority	Measured	DfT	
Bus passenger journeys	1991–2007	No. of journeys	n/a	Transport authority	Measured	DfT	

Variable	Time period covered	Units	Price base (if appropriate)	Level of aggregation	Modelled or measured	Source	Comments/issues
GVA—Production	1995–2007	£m	Real	NUTS3	Measured	ONS	The data is at constant 2007 prices
GVA—Construction	1995–2007	£m	Real	NUTS3	Measured	ONS	The data is at constant 2007 prices
GVA—Distribution, transport and communication	1995–2007	£m	Real	NUTS3	Measured	ONS	The data is at constant 2007 prices
GVA—Business services and finance	1995–2007	£m	Real	NUTS3	Measured	ONS	The data is at constant 2007 prices
GVA—Public administration, education, health and other service	1995–2007	£m	Real	NUTS3	Measured	ONS	The data is at constant 2007 prices
Performance	1997–2009	PPM	n/a	BR Sector	Measured	Network Rail	
Car cost	1991–2008	£ per journey	Real	Flow	Modelled	Oxera calculations	Based on TEMPRO, at constant 2007 prices
Car journey time	1991–2008	Minutes	n/a	Flow	Modelled	Oxera calculations	Based on TEMPRO
Service quality	1999–2008	Index	n/a	Ticket-type segment	Modelled	Oxera calculations	Based on NPS data
Air fare	1996–2002	£	Real	Route	Measured	CAA	Incomplete time series—data missing for 1997 and 1998

Note: [1] Ticket sales are measured, not actual journeys. Journeys are inferred from ticket sales using conversion factors which are constant through time. [2] GJT is made up of on-train journey time plus a penalty for changing trains and headway if appropriate. Monetary variables have been deflated to constant 2007 prices using the RPI CHAW series from the ONS.
Source: Oxera.

A2 Top 20 flows in revised dataset

Table A2.1 Largest 20 flows in 2008, by total revenue

	Origin	Destination
1	London BR	Manchester BR
2	Gatwick Airport	London BR
3	Manchester BR	London BR
4	London BR	Stansted Airport
5	Leeds	London BR
6	London BR	Leeds
7	Reading BR	Zone R1 London
8	London BR	Birmingham BR
9	London BR	Gatwick Airport
10	Newcastle	London BR
11	Birmingham BR	London BR
12	London BR	Newcastle
13	York	London BR
14	Milton Keynes Central	Zone R1 London
15	London BR	Liverpool BR
16	Liverpool BR	London BR
17	Croydon BR	Zone R1 London
18	Brighton	London Br
19	St Albans	Zone R1 London
20	Peterborough	Zone R1 London

Source: Oxera analysis.

A3 Car cost variable

The NTM contains speed–cost relationships that translate km/h to p/km in order to calculate the monetary cost of road journeys. Adjusting these speed–cost curves to represent changes to the cost of road journeys over time (incorporating both changes in the price of fuel and vehicle efficiency) and combining them with journey times on an origin–destination basis makes it possible to generate an origin–destination specific car cost time-series variable.

This appendix details the equations for the speed–cost curves in the NTM and how they have been adjusted to generate time series for the different journey purposes used in rail modelling.

A car journey time growth variable has been separately generated for use in the study. This took the station-to-station road journey as equivalent to a given rail trip and modelled increases in journey time on the basis of changes in traffic levels. These journey time growth rates were applied to data from the NTM for the year 2000 on base journey times to give modelled journey times, which were in turn used to calculate journey speed and, through the time series of speed–cost curves, journey cost.

It was not possible to match all the rail flows to a journey time in the NTM 2000 base data. Section A3.2 provides further details on the matching process, and outlines how unmatched flows were treated. Section A3.3 discusses the relative merits of using the average speed of an entire journey as opposed to the speed through each zone in that journey.

A3.1 Data sources

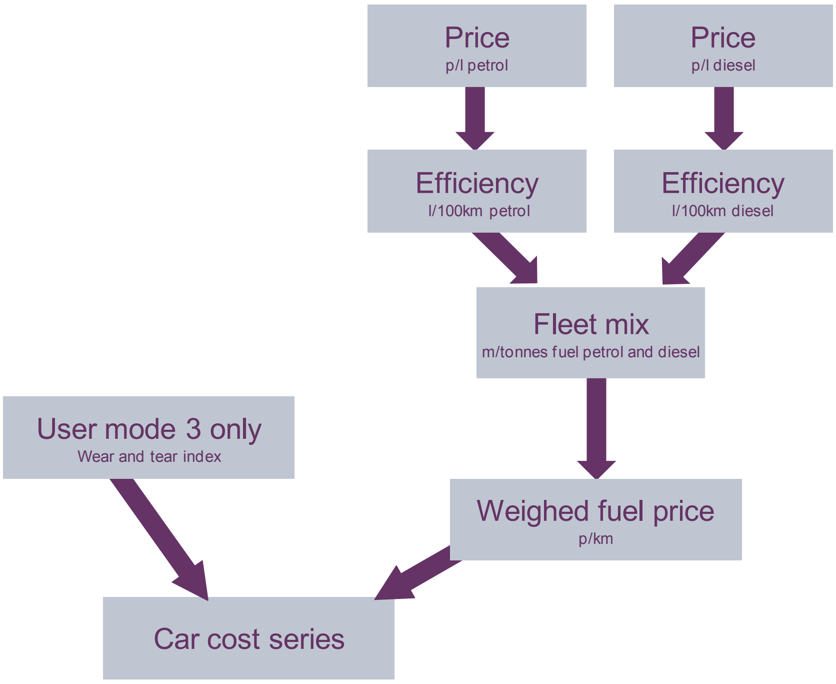
The first stage involved calculating a time series of the fuel cost in p/km. Fuel price data was sourced from DECC, with prices given in pence per litre (p/litre).[14] Efficiency data, in litres per 100 kilometres (l/100km), was sourced from the DVLA and Society of Motor Manufacturers and Traders (SMMT) for 1995–2007, and used to convert to p/km. [15] For the period 1990–95, data was available for new petrol cars only. To obtain a complete dataset, it was assumed that this was representative of all vehicles, and trends in this data were applied to the efficiency of all petrol and diesel cars. Finally, both price and efficiency data were split by fuel type. Therefore, it was necessary to weight for fleet mix, and this was done using data from AEA Energy & Environment.[16]

[14] Taken from DECC's *Quarterly Energy Prices,* Table 4.1.1, and excluding tax, available at http://stats.berr.gov.uk/energystats/qep411.xls

[15] Fleet efficiency: l/100km data was compiled by the EAE branch of the DfT combining data from the DVLA and SMMT. New car fuel efficiency is available at http://www.berr.gov.uk/files/file47218.xls. Fleet efficiency data was not available for 2008; therefore, in order to obtain a complete dataset, it was assumed that there was no change from 2007.

[16] Fleet mix: data is compiled by AEA Energy & Environment on behalf of DECC, using DfT vehicle licensing and road traffic use. Fleet efficiency data was not available for 2008; therefore, in order to obtain a complete dataset, it was assumed that there was no change from 2007.

Figure A3.1 Producing the car cost series



Source: Oxera.

The NTM generates short-run marginal costs for three different user modes: User Modes 1 and 2 are for travel in non-work time, for high- and low-income individuals respectively. The relevant cost here is simply the cost of fuel and the speed–cost relationship, and is given as:

$$p/km(v_t)_{1\&2,t} = c_{1\&2,t} – 0.223v_t + 0.00158v_t^2$$

where v_t is the speed in km/h and $c_{1\&2,t}$ is a constant representing the fuel cost in year t.[17]

User Mode 3 is for in-work travel time. As such, the VAT element of the fuel cost is omitted, but a term for 'wear and tear' is included.

$$p/km\ (v_t)_{3,t} = c_{3,t} – 0.190v_t + 0.00135v_t^{\ 2} + 107.66m_t / v_t$$

where v_t is the speed in km/h, $c_{3,t}$ is a constant representing the fuel cost, and m_t the wear and tear index in year t.[18]

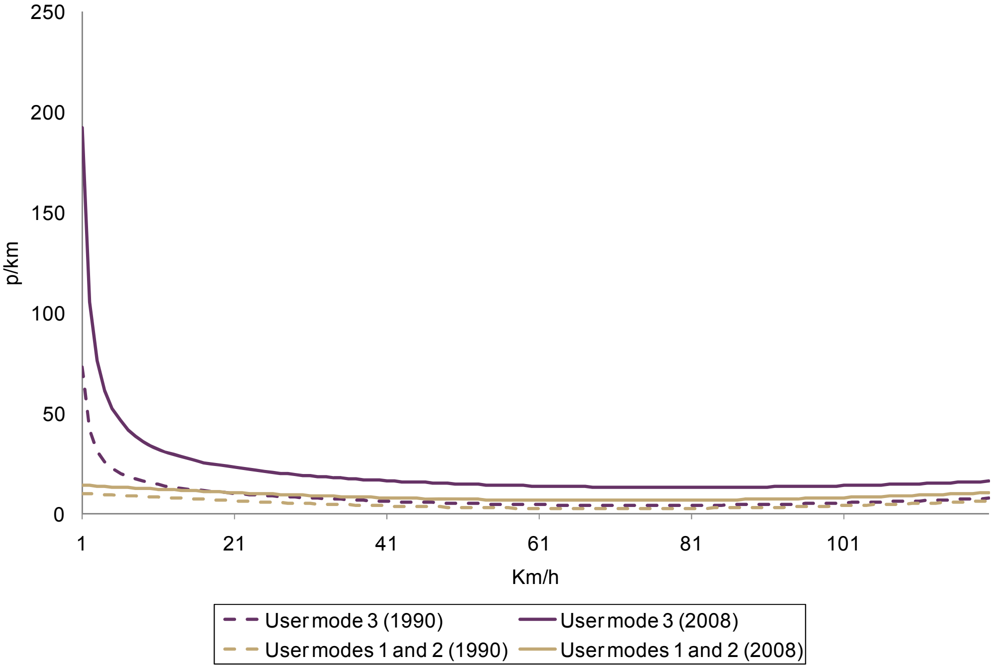
The wear and tear index was taken from table 1.19 in the DfT's 'Transport Statistics Great Britain' (TSGB).[19] This gives data for 1996–2007 and, as there was no data available for the period 1990–1996, the trend for the existing data was extrapolated backwards to 1990. Figures A3.2 and A3.3 below present the results of this process. Figure A3.2 shows the real speed–cost curves for 1990 and 2008 for User Modes 1 and 2, and for User Mode 3. Figure A3.3 shows the time series of real cost in p/km at 70km/h for each user mode that arises from application of the speed–cost curves.

[17] $c_{1\&2,2000}$ = 12.98 in the NTM. At 49km/h this leads to a price/km of 5.85p, equal to the weighted fuel cost in that year. The time-series of speed–*cost* curves has been constructed to preserve that relationship so that $c_{1\&2,t} = 7.13 + p_t^{VAT}$, where p_t^{VAT} is the weighted fuel price in year *t*.

[18] The wear-and-tear element of cost also included a constant term, and this has been incorporated into the term c3,t so that c3,t = 6.06 + ptNO VAT + 3.93mt, where ptNO VAT is the weighted fuel price excluding VAT and mt is the maintenance cost index in year t.

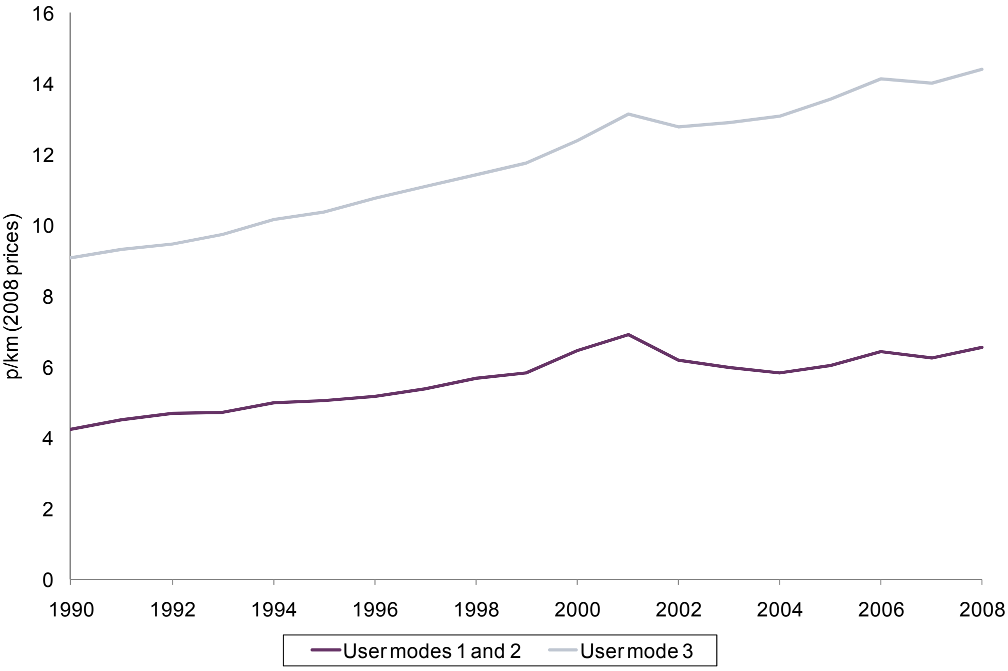
[19] Department for Transport (2008), 'Transport Statistics Great Britain', November, p. 29.

Figure A3.2 Real speed–cost curves in 1990 and 2008 for different user modes



Source: NTM.

Figure A3.3 Time series of real car cost (p/km) at 70km/h for different user modes



Source: NTM.

The different user modes were used to generate car cost variables for the journey purposes typically employed in rail modelling. Table A3.1 shows the combinations of speed–cost curves and base journey times for each journey purpose.[20]

[20] The car journey time growth rates do not vary by peak/inter-peak, so the distinction between commuting and leisure is likely to be small.

Table A3.1 NTM curves and base data used for each journey purpose

	NTM curve	Base journey time data
Business	User Mode 3	AM peak
Commuting	User Modes 1 and 2	AM peak
Leisure	User Modes 1 and 2	Inter-peak

Source: Oxera.

Base journey times and distances in the NTM were provided by user mode. Therefore, for commuting and leisure trips, which use the speed–cost curves for User Modes 1 and 2, there are two base journey times and two base journey distances. In these cases the appropriate base journey time and distance were calculated as the flow-weighted average of the journey times and distances of the two user modes. The exceptions to this are instances where:

- one of the two user modes had a zero flow (ie, no traffic on it), in which case the journey time and distance of the non-zero flow were used;
- both user modes had zero flow, in which case the simple average of the two journey times and distances was used.

Business trips use the User Mode 3 speed–cost curve, and so only had one base journey time and distance.

A3.2 Matching flows

Car journey time growth was modelled for each rail origin–destination pair, but in order to calculate costs these journey time growth rates had to be converted to times, then speeds, and finally to costs. This required the rail origin–destinations, and their journey time growth rates, to be matched with origin–destination pairs in the NTM data, which was provided at TEMPRO zone level.

However, the corresponding TEMPRO zones for some rail origin–destinations were not included in the NTM base data, and this led to around a quarter of the flows not being matched with journey time growth rates.[21]

The origin or destination TEMPRO zone was replaced for these unmatched flows to give the closest equivalent journey that was included in the NTM base. The modelled journey time growth rates were then applied to this alternative base journey time.

Where there were many unmatched flows with a common origin or destination, it was not always possible to match every one with an alternative.[22] Some intra-Scottish flows were simply not included in the NTM base data. These were matched to data from the Transport Model for Scotland (TMfS). The TMfS data was used only for flows not included in the NTM base data. Where data was available from both models, the NTM data was used in order to ensure the greatest consistency across the dataset.

TMfS data from AM peak skims was used for all three journey purposes (as opposed to the NTM data, where inter-peak data was used for the leisure series). Also, TMfS does not employ user modes in the same way as the NTM, but rather gives distance, time and cost for ‘in work’ (iw), ‘not in work commuting’ (nwc) and ‘not in work other’ (nwo). Table A3.2 summarises the TMfS data used for each journey purpose.

[21] Of 20,733 flows, 5,043 were not originally matched with journey time growth rates.

[22] A total of 628 flows still remained unmatched after providing alternative zones for each origin or destination station.

Table A3.2 TMfS base data used for each journey purpose

	TMfS type	Base journey time data
Business	In work	AM peak
Commuting	Not in work commuting	AM peak
Leisure	Not in work other	AM peak

Source: TMfS.

Once rail flows were matched to journey times and then were converted to costs, these nominal costs were then deflated to 2008 prices using the CHAW RPI index.

A3.3 Average speed versus zones

This approach has two potential advantages over simply using the pump price of petrol:

- it includes terms that allow for increases in engine efficiency to reduce the cost of driving and the mix of petrol and diesel vehicles;
- it allows for costs, as well as the journey times, to vary with changes in congestion.

Due to the non-linear nature of the speed–cost relationship in the NTM, there will be more variation arising from the second point if the speed (and therefore the cost) is considered for different parts of the journey separately. This could be done by calculating the average speed through each TEMPRO zone which constitutes a part of the whole journey.

However, initial work suggested that calculating costs on the basis of average journey speed was leading to cross-sectional differences between the flows. It was therefore decided that any additional detail arising from calculating the costs by zone would not add enough value to exceed the considerable costs in terms of computing time and power.

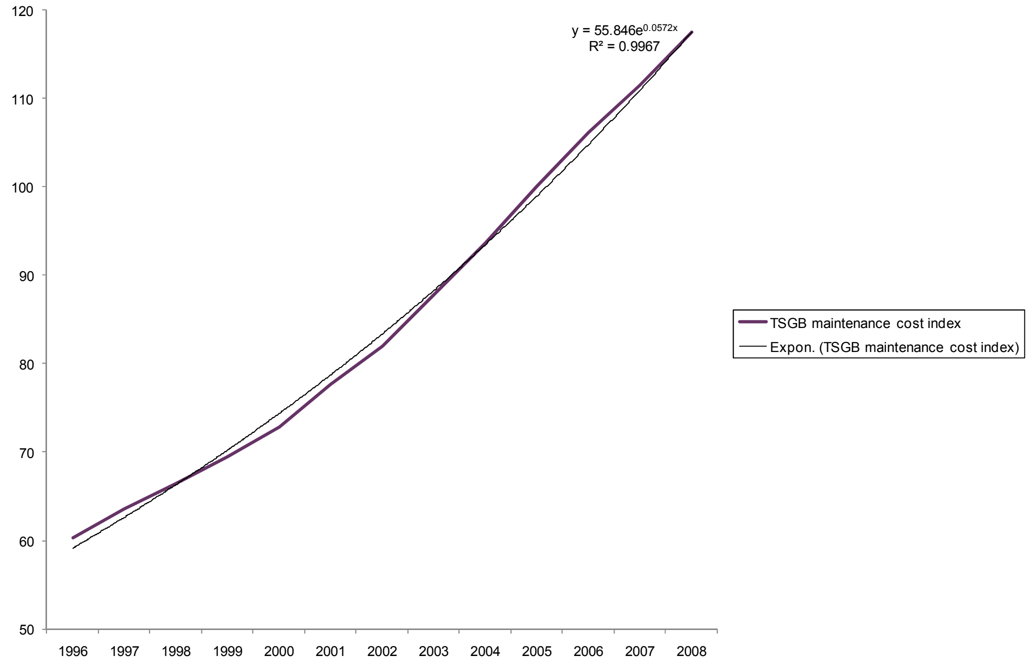
A3.4 Maintenance cost index

Table 1.19 from TSGB gives a maintenance cost index used in RPI calculations, but only from 1996 to 2008. In the absence of an alternative data source, and in order to have a complete dataset, it was therefore necessary to extrapolate the trend in this data back to 1990. Analysis of the available data found that the trend was best characterised by an exponential function of the form:

$$y = 55.85 \exp(0.0572x)$$

Figure A3.4 plots this data and the fitted trend.

Figure A3.4 TSGB maintenance cost index 1996–2008 and fitted exponential trend



Source: Oxera

The constant average rate of growth associated with this function is 5.89% per annum. This growth rate was then projected back to 1990. Table A3.3 shows the complete dataset.

Table A3.3 Maintenance cost index

Year	Maintenance index
1990	**42.0**
1991	**44.4**
1992	**47.0**
1993	**49.8**
1994	**52.7**
1995	**55.8**
1996	60.3
1997	63.5
1998	66.5
1999	69.4
2000	72.8
2001	77.7
2002	82
2003	87.7
2004	93.7
2005	100
2006	106.2
2007	111.5
2008	117.5

Note: Bold figures (in red) indicate extrapolated data.
Source: Oxera and TSGB.

Park Central
40/41 Park End Street
Oxford OX1 1JD
United Kingdom

Tel: +44 (0) 1865 253 000
Fax: +44 (0) 1865 251 172

Stephanie Square Centre
Avenue Louise 65, Box 11
1050 Brussels
Belgium

Tel: +32 (0) 2 535 7878
Fax: +32 (0) 2 535 7770

Thavies Inn House
7th Floor
3/4 Holborn Circus
London EC1N 2HA
United Kingdom

Tel: +44 (0) 20 7822 2650
Fax: +44 (0) 20 7822 2651

www.oxera.com