ARUP

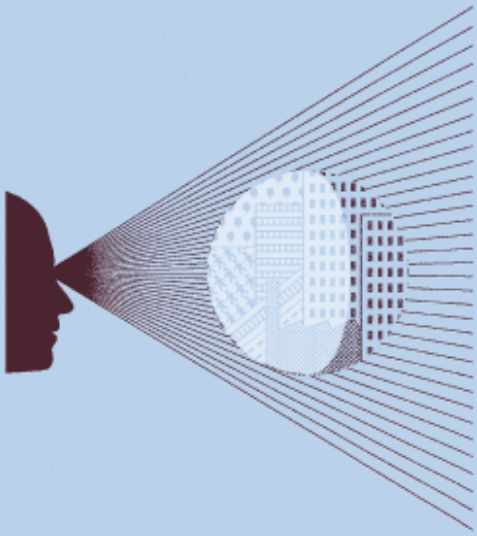
Oxera

How has the market for rail passenger demand been segmented?

Market segmentation report

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

March 2010



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).

When forecasting demand for passenger rail travel, the PDFH[1] segments the market in Great Britain into the London Travelcard area, the rest of the South East, outside the South East, and areas covered by a Passenger Transport Executive (PTE) (now known as an Integrated Transport Authority). This is mostly the result of historical administration and not necessarily based on an economic rationale. Such a rationale is important because it allows for a segmentation based on behaviour, and hence may result in better forecasting performance.

The aim of the market segmentation is to examine how responses to a change in a demand driver vary between different elements of the rail market, allowing the demand for passenger rail travel in Great Britain to be divided into segments—according to an economic rationale and on the basis of the available data. While this study focuses on forecasting demand for passenger rail travel in Great Britain, the principle of market segmentation is applicable to a wide range of other contexts, such as understanding the effect of a policy change or merger.

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 report is structured as follows.

- Section 2 sets out the analytical framework on which the market segmentation exercise is based; and section 3 presents the data that has been collated and matched for this exercise, and outlines the initial development of hypotheses using partial derivatives.
- Section 4 reports on the cluster analysis undertaken to test the initial hypotheses on segmentation from the partial derivatives work.
- Section 5 concludes, with a proposed market segmentation.

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

[1] ATOC (2009), 'Passenger Demand Forecasting Handbook Version 5', August, Section B0, p. 3.

2 Analytical framework

Elasticities of demand reflect the **behavioural response** of consumers to a change in a demand driver (such as income or fares). As consumers are likely to have different preferences for different products, their responses (and hence their elasticities) will be expected to vary. However, allowing all consumers to have different elasticities is not a practical way to forecast rail demand.

Market segmentation can be considered as the process of grouping together individuals with similar responses. Although the data to do this for individual passengers is not available for this study, flows with similar responses can be grouped together, following the same concept.

The PDFH is the industry source providing the analytical framework and parameters to enable forecasts of passenger rail demand in Great Britain to be generated. It suggests that journey purpose (business, leisure or commuting) is the optimal market segmentation. This is supplemented by distance bands and geographical variations for some demand drivers.

While segmentation by journey purpose is desirable (as it reflects industry expectations and some behavioural evidence), comprehensive data with which to estimate these elasticities is not available. Therefore, the existing practice of estimating elasticities based on ticket types, for which comprehensive data is available, and then converting to journey purpose through the use of survey data, will be continued in this study.

This study differs from others by taking a comprehensive, flexible approach to market segmentation, thus allowing the data to test hypotheses about what drives segmentation, while also including the economic characteristics of rail users.

2.1 Why would an elasticity vary?

Consumers are likely to have different behavioural responses to changes in demand drivers because of differences in preference for the product in question. However, as it is not practical to forecast rail passenger demand on an individual basis, there is a trade-off between accuracy and practicality (eg, allowing for different elasticities and having sufficiently few market segments to create a forecasting framework that can be both robustly estimated and easily applied).

Market segmentation is the process of grouping together consumers who express similar preferences for a good or service. While this report adheres to this general concept, the data available does not allow the grouping of individual passengers; hence, this study focuses on grouping flows, where passengers display similar responses to changes in demand drivers.

2.2 Current market segmentation

Recent versions of the PDFH use a market segmentation based on products—ie, ticket types, geographic areas and journey purpose. Some of the inconsistencies previously observed between different sections of the PDFH (for example, the use of a distance factor in the income elasticity but distance bands in other sections, such as non-London flows)—have been resolved in the latest version (version 5). However, potential issues remain, one being the evolution of the market segmentation from what were initially administrative segments with no clear underlying economic rationale. This may result in inaccurate forecasts if the relationships between passengers' behaviour and rail demand changes.

Box 2.1 Market segmentation in the PDFH

The recent update of the PDFH suggests that 'the principal segmentation is by journey purpose'; with geographic and distance segmentations added to journey purpose in order to arrive at a market segmentation.

The considerations of market segmentation extend beyond these journey purposes (as is also set out in the PDFH), with the following important characteristics:

- journey purpose—business, leisure, commuting (where leisure could be split further into shopping, visiting friends and relatives, holidays and other; and commuting into work or education);
- passenger characteristics—age, sex, socio-economic group;
- group size;
- availability of other modes of transport, especially car;
- baggage or other restriction on mobility;
- geography.

However, the key journey purposes contained within the PDFH are:

- business;
- leisure;
- commuting;
- travel to and from airports.

The PDFH provides elasticities for a number of geographically defined market segments:

- within the London Travelcard area;
- rest of the South East to/from the London Travelcard area;
- within the South East (excluding the London Travelcard area);
- outside South East to/from London;
- non-London inter-urban;
- non-London short distance.

While the PDFH suggests that 'the principal segmentation is by journey purpose', the principal source of demand data available for use in the British railway industry is the LENNON ticket sales database, which contains data by ticket type, rather than journey purpose. The PDFH provides conversion tables from ticket type to journey purpose, based on National Rail Travel Survey (NRTS) data. However, this data is insufficient for the purposes of this study, which requires consistent data over time. Therefore, an alternative market segmentation must be developed.

This study provides a new approach to market segmentation: setting out an economic framework and testing it using data and industry knowledge, and providing recommendations for an updated market segmentation where appropriate. Section 2.3 considers current knowledge on what a proposed new market segmentation might look like.

2.3 Possible market segmentation

The expected structure of the market segmentation can be considered using economic theory, insights from previous research, and market analysis.

Economic theory suggests that the demand for a product is a function of the price (and/or generalised cost[2]) of the product, the price (and/or generalised cost) of substitutable and complementary products, and the income of the consumer. Therefore, a segmentation accounting for these factors is likely to be appropriate. However, as later analysis will show,

[2] Generalised cost is a term that relates to a combined concept of price and non-price characteristics of a product. In the case of rail passenger transport, generalised cost can also include access and egress, in-vehicle and waiting times, and service reliability.

these economic characteristics may also be captured within the econometric analysis by using an appropriate functional form. This highlights one of the key trade-offs in this study, between what can be captured within the econometric analysis and what is captured by prior analysis, such as market segmentation. For example, distance bands have traditionally been used to provide a market segmentation based on distance.[3] However, a distance effect could also be modelled within an econometric framework,[4] which is the proposal for this study.

An alternative to segmenting the market on the basis of economic characteristics (such as income or employment) is to segment according to geography, by location or distance. This approach has a notable history within rail demand forecasting, with a number of sources (including the PDFH v4.1 and Regional Flows work)[5] using this method. However, the interactions between variables are complex, and the direction of causality is often difficult to determine. For instance, it may not be straightforward to determine whether the most appropriate market segmentation should be on the basis of price, distance, or journey time since all three are highly correlated. Therefore, it is important that the segmentation is based on a rigorous conceptual framework grounded in economic theory.

Segmentation based on journey purpose is appealing, as this would provide an economically robust way of modelling passengers' responses to a change in a demand driver, and has previously been identified as a preferred way of presenting a market segmentation.[6] However, as noted above, this segmentation is not practical given the data available for this study. Therefore, the existing approach—estimating elasticities on the basis of ticket types and converting them to journey purpose (through the use of surveys such as the NRTS)—will be continued in this study.

Several recent studies suggest possible market segmentations, including work by MVA Consultancy (2009), Steer Davies Gleave (2009), Scott Wilson et al. (2008), and the Institute for Transport Studies, Leeds (2007).[7] These studies have suggested that journey purpose, distance and income may be important, but, additionally, long-distance travel may be a separate segment requiring further disaggregation. Moreover, a hypothesis suggested by MVA Consultancy is that non-London core cities may be an appropriate market segment for further investigation, when compared to the existing PDFH segmentation. Following this review of the literature, the study team conducted an extensive process of data evaluation to arrive at a proposed market segmentation. This process is set out in the sections 3–5, together with the results.

[3] See, for example, ATOC (2009), op. cit., Tables B0.4–B0.8.

[4] See, for example, the distance effect on the income elasticity in version 4.1 of the PDFH. ATOC (2005), 'Passenger Demand Forecasting Handbook Version 4.1', June.

[5] MVA Consultancy (2009), 'Regional Rail Demand Elasticities', September.

[6] Steer Davies Gleave (2008), 'PDFH Update – Phase 1', June.

[7] See MVA Consultancy, op. cit.; Steer Davies Gleave (2009), 'Recession Impacts: Final Report', August; Scott Wilson, RAND Europe and H-G-A (2008), 'Modelling Longer Distance Demand for Travel: Feasibility Study', June 18th, p. 7; Institute for Transport Studies (2007), 'Revealed Preference Study to Assess Impact of Reliability on Passenger Rail Demand', November 30th, p. 2.

3 Data and partial derivatives

As set out in section 2, market segmentation can be considered as grouping together flows which display similar responses to changes in demand drivers.

The first step is the creation of partial derivatives, which show how demand varies in response to a small change in a demand driver (for example, income or fares), while all other factors are held constant.

The bivariate nature of this analysis means that the causality of any identified relationships is difficult to establish. For example, if a relationship between X and Y is identified, it is often unclear whether X causes Y, Y causes X, or whether there may be another variable affecting both X and Y such that it appears as though the observed relationship is causal. This problem is present in most applied statistical work, but is particularly acute in bivariate analysis.

The hypotheses generated by this bivariate analysis are then refined using more sophisticated analysis, which is described in section 4.

This section gives a brief description of the data and the rationale for examining partial derivatives, as well as the process for their creation, to provide a preliminary assessment of segmenting variables. Partial derivatives are a measure of how one variable (the demand for passenger rail travel) varies with a small change in another variable (price, income, etc).

Having calculated average partial derivatives for each flow in the dataset, these were grouped to ascertain whether there is likely to be a different response to a change in the demand driver, depending on the level of the variable or the type of flow (eg, long-distance inter-urban, to airports, etc). There is an important distinction between the drivers of rail demand, such as income and fares, and the characteristics of the flow which may affect the response to changes in these drivers, such as the level of fares or type of flow.
This process has generated several hypotheses, and these are further tested in the next step of the market segmentation process (explained in section 4).

When considering these results, it is important to remember the bivariate nature of the analysis (and the limitations this implies) in terms of whether the identified relationship is causal.

3.1 Data

The dependent variable for this study is the annual number of journeys for six ticket types at the flow level.[8] The number of flows is more than 20,000 for the longest time series of 18 years, although many have available time series over a shorter period.

The dataset created for this study is detailed in the *Data capability* report. The next section discusses the rationale for creating partial derivatives (the analysis of which proposed initial hypotheses for consideration).

3.2 Partial derivatives

A partial derivative shows how the dependent variable (rail journeys) varies with a small change in the explanatory variable (eg, income), assuming that all other variables are held constant.

[8] The ticket types are: first class season; first class non-season; standard class full price; standard class reduced price; standard class Apex; and standard class season.

3.2.1 Purpose

The rationale for estimating partial derivatives is that they allow investigation of how passenger demand (number of journeys) on a particular flow responds to a change in a demand driver. The demand drivers have been identified (according to economic theory and industry knowledge) as:

- income;
- demographics;
- monetary and time cost of rail travel.

The following variables have been identified in the 'Rail Trends Report'[9] as potentially contributing to the difference between flows in the response of passenger demand to changes in each of these drivers:

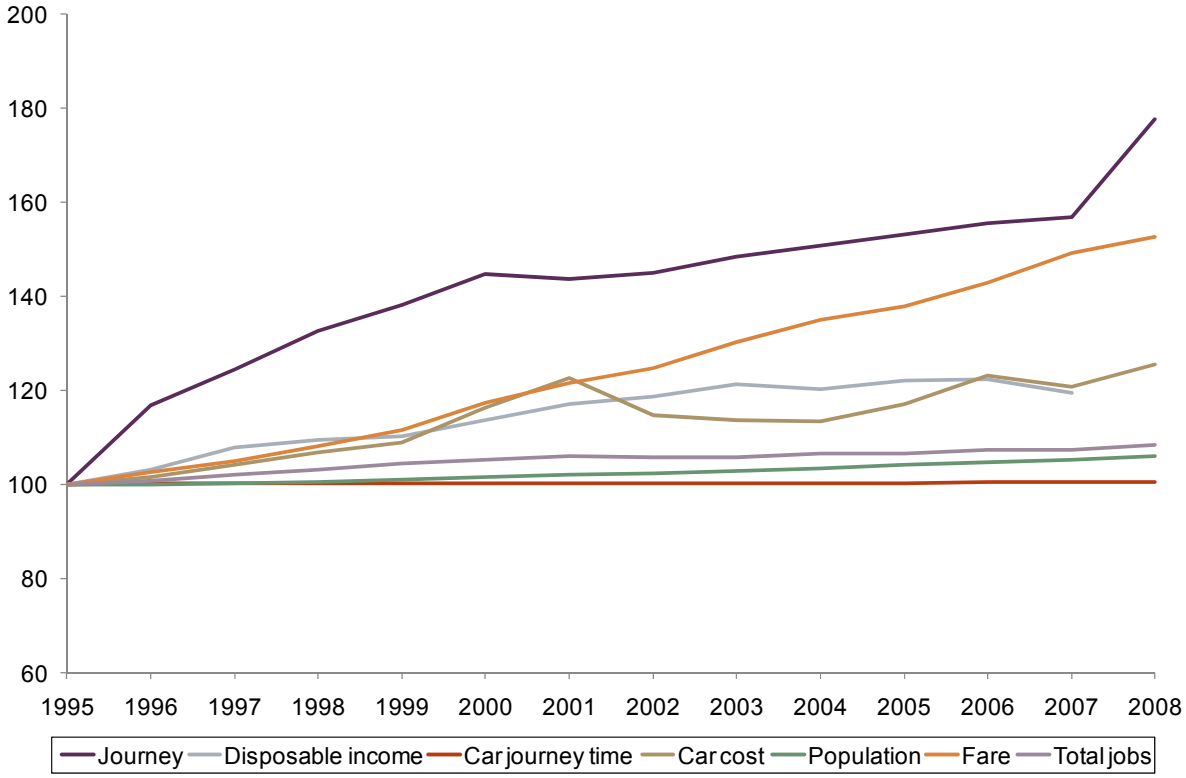
- workplace gross value added (GVA, gross and per capita);
- financial workplace GVA (gross and per capita);
- residential GVA (gross and per capita);
- disposable income (gross and per household);
- population;
- employment;
- number of households;
- car ownership;
- average fare;
- generalised journey time (GJT).

3.2.2 Methodology

To reduce the possibility of spurious correlations between these variables and demand, both the driver and demand have been detrended. This is necessary because of the upward trend present in many of the variables over time, as shown in Figure 3.1 below.

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

Figure 3.1 Trends over time



Source: Oxera analysis.

The variables were detrended by regressing the natural log of the variable on a linear time trend and a constant, and then calculating the residual from this regression. These detrended variables were then used to calculate the partial derivatives, as explained below.

The partial derivatives have been calculated as the ratio between the natural log of the detrended dependent variable (journeys) and the natural log of the detrended driver for each flow:

$$[\text{Log}(\text{journeys}_{i,t}) - \text{Log}(\text{trend in journeys}_i)] / [\text{Log}(\text{driver}_{i,t}) - \text{Log}(\text{trend in driver}_i)]$$

This gives a partial derivative for each flow for each year in the dataset, $p_{i,t}$, from which an average was calculated as follows:

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p_{i,t}$$

This gives an average partial derivative for each flow. The next section describes how these average partial derivatives were used to generate hypotheses.

3.2.3 Process of generating hypotheses

The average partial derivatives were then grouped in various ways (explained below) to establish whether there were differences between groupings. This exercise was repeated for different characteristics, and the following flow characteristics were identified as potential segmenting variables for this analysis:

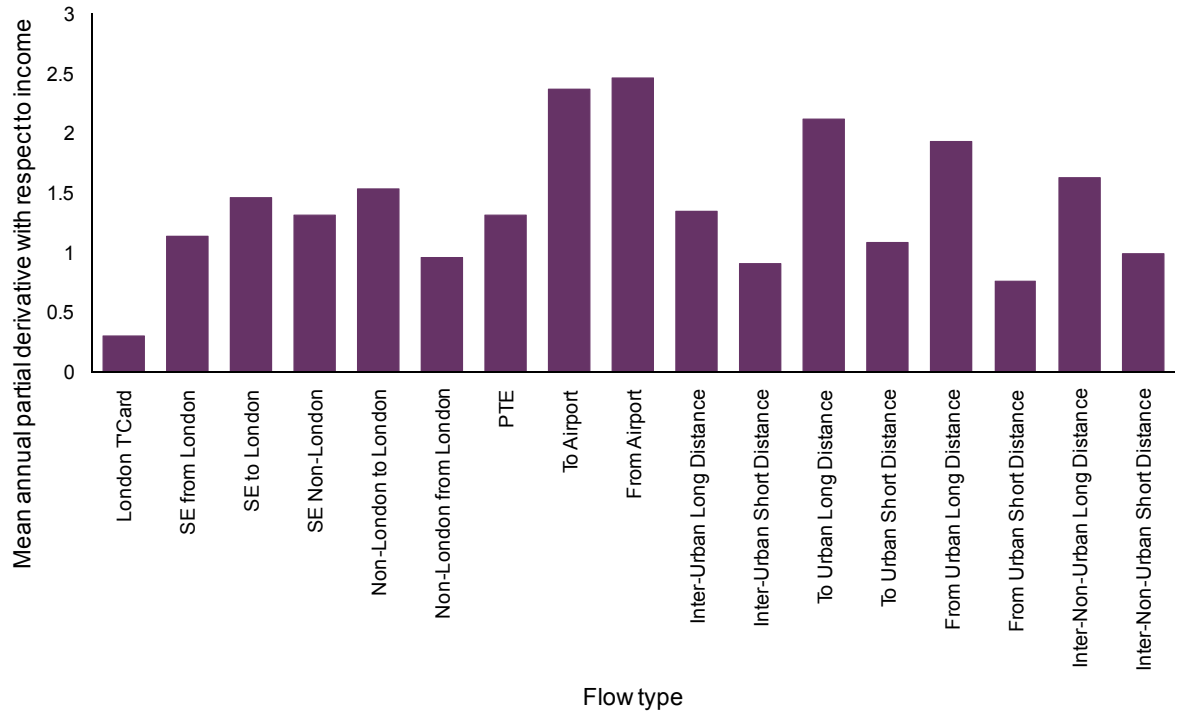
- distance;
- type of flow (eg, to/from London and the South East, urban, or airport);
- government office region (GOR) at origin/destination;
- PTE flow;

– corridor;
– average fare;
– GJT;
– workplace GVA at origin/destination;
– residential GVA at origin/destination;
– job density at origin/destination.

These variables are slightly different from those set out in section 3.2.1 because they are characteristics affecting the response of passenger demand to a change in a demand driver, rather than drivers of rail demand per se, although the two sets of variables have some elements in common. For example, income is both a driver and a characteristic because economic theory suggests that the higher a consumer's income, the more likely they are to purchase goods, and hence a change in income can be expected to change the demand for a good. However, since different consumers have different levels of income, income is an important characteristic of a consumer as well as a demand driver.

The flows were then grouped by segmenting variable (listed above), the average partial derivative for the group was calculated, and the results plotted. This provides a visual means of making an initial judgement on the likely usefulness of the segmentation. For example, Figure 3.2 shows that there are differences in the average income partial derivative, depending on the flow type. This provides some indication that flow type warrants further investigation in the multivariate cluster analysis, described in section 4.

Figure 3.2 Income partial derivative by flow type

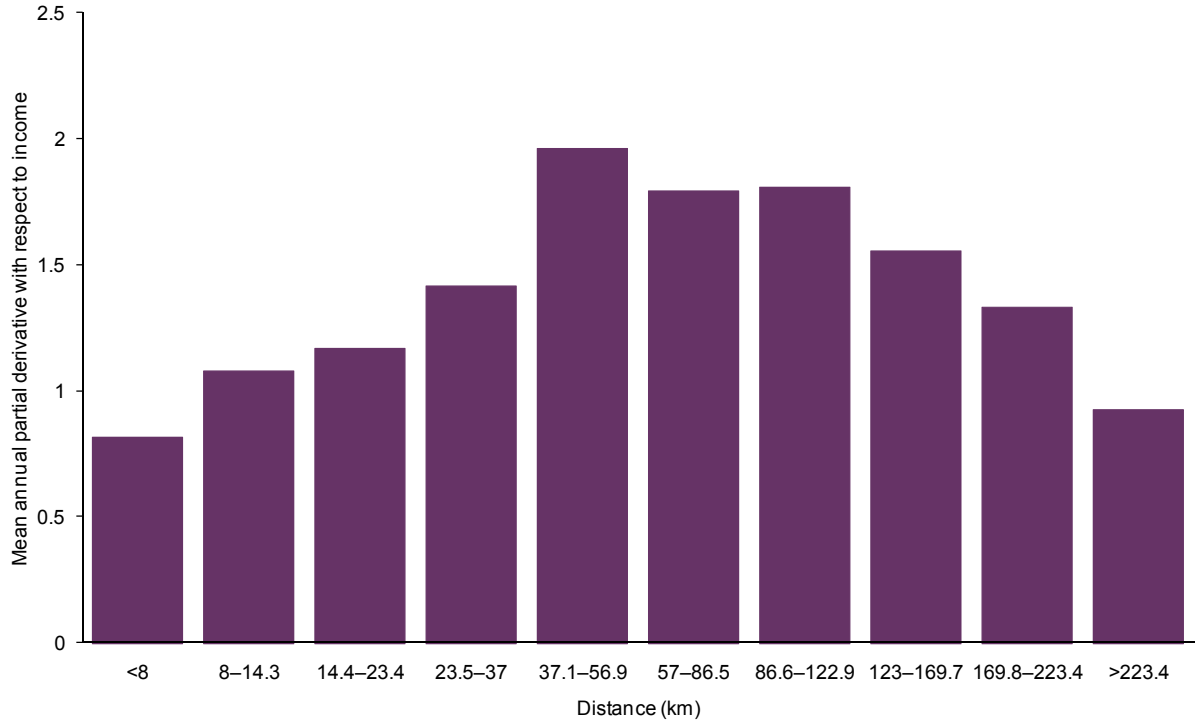


Source: Oxera.

Figure 3.2 plots the average partial derivative with respect to income across 17 flow types over 18 years. The mean annual partial derivative for 'to airport' flows is approximately 2.4. This suggests that if income increases by 1%, trips to airports increase by 2.4%. This figure may include abstraction from other modes and/or an increase in absolute trip numbers. It could also suggest that as income increases so does demand for air travel. Figure 3.2 suggests that there may be different responses in demand to a given change in income depending on the type of flow under consideration. For example, the average response to a 1% change in income on flows to/from airports appears to be 1.5 percentage points greater than that on Urban Short Distance flows, suggesting that income elasticity varies according

to flow type. However, it is important to bear in mind the bivariate nature of this analysis because, while the relationships identified may hold, there may also be a third factor driving the identified relationship. To demonstrate this, the apparent differences across flow types in Figure 3.2 may be driven by differences in the average level of income at the origin or destination of the flow, or by the journey purpose on the different types of flow, rather than the type of flow per se. The use of cluster analysis allows further investigation of issues of this type (see section 4).

Figure 3.3 Income partial derivative by distance



Source: Oxera.

Figure 3.3 shows that the responsiveness of demand to changes in income appears to vary, depending on the distance of the flow, with middle-distance flows being the most responsive. Again, it is important to remember the bivariate nature of the analysis, as this may suggest that modal competition is most effective at short and long distances, or that there may be another factor to be considered.

The hypotheses drawn from this process are as follows:

– distance appears to be important in generating different responses to changes in the drivers;
– there appears to be a broad similarity in responses between ticket types—for example, standard class season and first class non-season appear to exhibit a similar response. However, in some cases, such as standard class season compared with other ticket types, the magnitude of the responses appears to be different;
– average fare per mile does not appear to influence the responses, and hence does not appear to be a useful segmenting variable;
– flow type appears to be an important factor for further consideration;
– the regions of origin and destination appear to be an important factor in explaining the differences in responses;
– wealth at origin and job density at destination do not appear to be useful segmenting variables.

To further examine the hypotheses identified in this exercise, cluster analysis was used to enable them to be tested in a multivariate setting. Section 4 explains this process.

4 Cluster analysis

Following on from the bivariate analysis in section 3, cluster analysis was used to investigate the identified hypotheses further.

Cluster analysis is a statistical technique which allows different flows to be grouped according to how similar they are. Having grouped flows with similar characteristics into clusters, the characteristics of each cluster were analysed to produce further hypotheses of market segmentation. The hypotheses generated in the bivariate analysis described in section 3 were then used to inform the hypothesis testing in the second step of the process.

The number of clusters was assumed to be 15, following analysis of a dendrogram (a graph depicting the similarity between different groups of flows). 15 clusters offers a trade-off between complexity (offering sufficient numbers of clusters to enable useful analysis to be undertaken) and robustness (allowing many of the clusters to have a sufficiently large sample size—ie, greater than 50 flows—such that robust inferences can be drawn from the process).

The process was initially conducted by ticket type, although first class season tickets did not provide a sufficient number of clusters with a large enough sample size to allow robust inference, and so this type was not analysed in detail. For the econometric analysis, the ticket types were aggregated into three products which are consistent with the fare simplification introduced by ATOC: full fare, reduced fare and season tickets.

A consistent market segmentation appears to arise from this process:

– London, the South East, and the East of England;
– the Midlands;
– the rest of the country;
– flows to and from airports.

Further analysis was conducted to determine whether the Midlands was a separate market segment, or whether, because the Midlands is dominated by the Birmingham conurbation, a segment of non-London core cities may be more appropriate (as per recent work undertaken by MVA Consultancy).[10] This analysis involved studying whether the patterns of flows, distances and fares were different between large cities around Great Britain and the surrounding areas; and were similar to other large cities. The conclusion from this analysis is that there is a greater proportion of inter-urban travel from large cities than the surrounding areas, together with, on average, longer distance and higher fare flows. This suggests that non-London core cities present a distinct market segment; hence, the proposed market segmentation is:

– London, the South East, and the East of England;
– non-London core cities;
– the rest of the country;
– airports.

A number of other factors of interest arose during the analysis, which were then investigated further in the econometric analysis:

– distance appears to be an important factor to be taken into account in the analysis;
– income and fares responses appear to vary, depending on the level of the variable.

Following on from the bivariate analysis in section 3, cluster analysis was used to further investigate the identified hypotheses.

[10] MVA Consultancy (2009), op. cit.

Box 4.1 Cluster analysis

Cluster analysis is a non-parametric statistical technique used to group 'similar' objects into clusters. 'Similarity' can be defined in a number of ways, but is essentially the distance between measurements of characteristics of the objects, such as the income at origin for two flows. A formal definition is difficult as there are many ways of formulating the distance metric.

In this analysis, the chosen distance metric is Euclidian distance, which is defined as:

$$\left[\sum_{p=1}^{P}(x_{i,p}-x_{j,p})^2\right]^{1/2}$$

where the analysis is comparing observation I to observation j.

Having grouped flows with similar characteristics into clusters using cluster analysis, the characteristics of each cluster were analysed to produce further hypotheses of market segmentation. This process used the hypotheses generated in the bivariate analysis described in section 3 to inform the hypothesis testing in the second step of the process.

The initial approach of the cluster analysis in this case was to use a number of different statistics[11] to decide on the optimal number of clusters within each ticket type. However, this approach resulted in a relatively small number of clusters for which the 'sensible' (ie, values consistent with economic intuition) partial derivatives were placed in one or two clusters, with the large outliers placed in the other clusters.

Therefore, the approach was changed to assume that there were 15 clusters for each ticket type, and to investigate those clusters containing more than 50 flows. The rationale for this assumption is discussed in section 4.1.3; 15 clusters offers a trade-off between complexity (offering sufficient numbers of clusters to enable useful analysis to be undertaken) and robustness (allowing many of the clusters to have a sample size sufficiently large—ie, greater than 50 flows—such that robust inferences can be drawn from the process).

The process was initially conducted by ticket type, although first class season tickets did not provide a sufficient number of clusters with a large enough sample size to allow robust inference, and so this type was not analysed in detail.

4.1.1 Proposed market segmentation

A consistent market segmentation appears to arise from the process described above:

- London, the South East, and the East of England;
- the Midlands;
- the rest of the country;
- flows to and from airports.

Further analysis was conducted to determine whether the Midlands was a separate market segment, or whether, because the Midlands is dominated by the Birmingham conurbation, a segment of non-London core cities may be more appropriate (as per recent work undertaken by MVA Consultancy).[12] This analysis consisted of studying whether the patterns of flows, distances and fares were different between large cities around Great Britain and the areas surrounding them; and whether these patterns were similar to those present in other large cities. The conclusion from this analysis is that there is a greater proportion of inter-urban travel from large cities than the surrounding areas, together with, on average, longer distance and higher fare flows. This suggests that non-London core cities present a distinct market segment; hence, the market segmentation is:

[11] The Duda–Hart, Caliński–Harabasz and pseudo t-squared statistics were used.

[12] MVA Consultancy (2009), op. cit.

– London, the South East, and the East of England;
– non-London core cities;
– the rest of the country;
– flows to and from airports.

This proposed segmentation has been tested within the econometric analysis.

A number of other factors of interest arose during the market segmentation analysis:

– distance appears to be an important factor to be taken into account in the econometric analysis;
– the response to a change in income or fares appears to vary depending on the level of the variable.

4.1.2 Approach to arriving at hypotheses of market segmentation

Following the allocation of rail flows to clusters, a four-stage approach was used to arrive at hypotheses of market segmentation:

– in the **first stage**, frequency tables were produced showing the percentage of flows in each GOR at the origin and destination of the flow, and flow type for each cluster. This was supplemented with a chart showing the standardised deviation from the mean of each characteristic (eg, income, employment);

– the **second stage** involved a detailed examination of these initial outputs for deviations from the average of the ticket type, to determine the characteristics which define the clusters and which thus define the market segmentation;

– a summary table for each ticket type was then produced in the **third stage**;

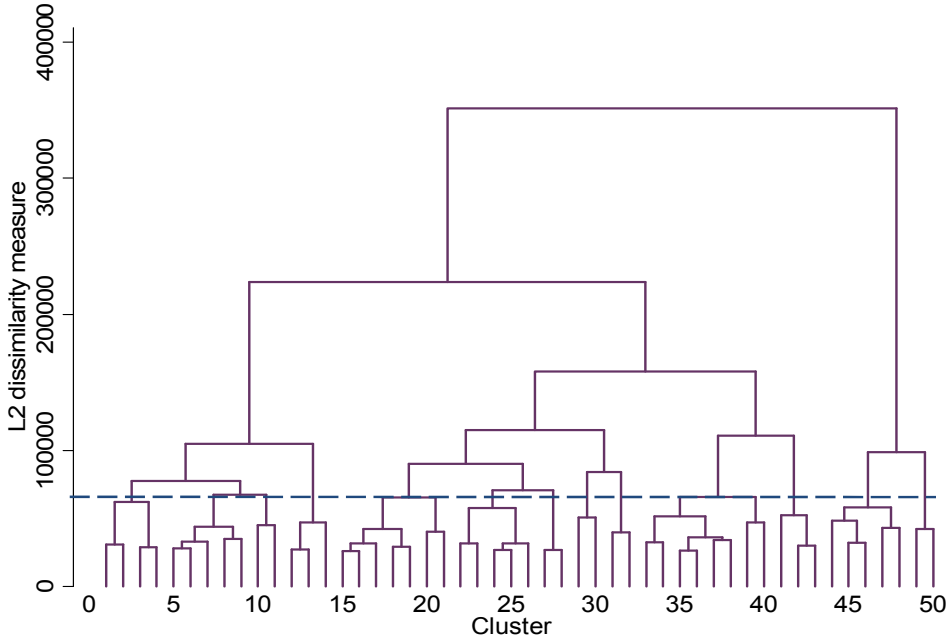
– in the **fourth stage**, these tables were compared across the different ticket types, allowing tentative conclusions to be drawn about possible market segments for testing in the econometric phase of the study.

A detailed worked example of the entire process provided in section 4.2.

4.1.3 Optimal number of clusters

As explained in the introduction to section 4, the initial approach—of using a number of statistics together with expert judgement to decide on the optimal number of clusters—produced results with one large cluster (with ‘sensible’ partial derivatives) and a number of other clusters containing outliers. The approach was therefore changed to selecting the optimal number of clusters based on the examination of dendrograms, an example of which is given in Figure 4.1 below.

Figure 4.1 Standard full dendrogram



Source: Oxera analysis.

Figure 4.1 is a dendrogram which can be read as the degree of similarity between clusters being on the y axis—ie, the smaller the vertical distance between two horizontal lines, the greater the similarity of the characteristics of the two clusters. Clusters are matched together sequentially by merging the clusters that are most similar—ie, where the vertical line is shortest between the clusters. For example, the group of clusters on the far right of the x axis (clusters 44–50) form a separate grouping because the vertical line joining them to the others is long, which implies that these clusters have different characteristics to the others. It is important to note that *the clusters do not have any direct meaning* at this stage of the analysis. The aim of the four-stage approach is to provide an interpretation of the different clusters. For this reason, it is not important which of the clusters in Figure 4.1 are the 15 clusters used in subsequent analysis.

The use of 15 clusters provides a trade-off between having many clusters from which to draw inferences and having a robust sample size within most of the clusters. The second half of the decision process concerns the sample size within each cluster. Although the sample sizes of the clusters are not presented in Figure 4.1 (they have been analysed separately), in many of the clusters they are sufficient for robust inferences to be drawn.

4.2 Worked example

This section works through an example of the first three of the four stages in the process for the standard class full fare ticket type. Tables 4.1–4.4 and Figure 4.2 provide an example of the outputs for an individual cluster. Cluster 11 has been used for this example, but the process is identical for each cluster. The exact cluster is unimportant, and this analysis has been repeated for each cluster with a sample size of more than 50 flows for each ticket type.

Table 4.1 GOR at origin (% of flows)

	Example cluster	Standard class full fare tickets
East Midlands	4.02	4.32
East of England	9.55	10.43
London	3.02	14.14
North East	6.03	3.22
North West	12.06	10.52
Scotland	**24.12**	**6.43**
South East	10.55	23.69
South West	3.02	9
Wales	2.01	3.06
West Midlands	4.02	6.58
Yorkshire & Humberside	**21.61**	**8.6**
Total	**100**	**100**

Note: Totals may not sum due to rounding.
Source: Oxera analysis.

Table 4.1 demonstrates that the percentage of flows which begin in Scotland and Yorkshire & Humberside (highlighted in bold) are much higher than for the ticket type as a whole.

Table 4.2 GOR at destination (% of flows)

	Example cluster	Standard class full fare tickets
East Midlands	0	4.28
East of England	0	9.14
London	0	17.58
North East	0	2.89
North West	0	11.32
Scotland	**41.21**	**6.51**
South East	0	21.95
South West	0	8.32
Wales	0	2.8
West Midlands	0	7.42
Yorkshire & Humberside	**58.79**	**7.78**
Total	**100**	**100**

Source: Oxera analysis.

Table 4.2 indicates that the percentage of flows which end in Scotland or Yorkshire & Humberside are higher than the ticket type as a whole (highlighted in bold in the table). This, combined with the information from Table 4.1, suggests that the cluster includes a substantial proportion of flows which begin and end in Scotland and/or Yorkshire & Humberside, and hence that geography may be an important variable to account for in the market segmentation. The next step in the analysis is to compare the types of flow contained within the cluster with the average of the ticket type (see Table 4.3 below).

Table 4.3 Flow type (% of flows)

	Example cluster	Standard class full fare tickets
From Airport	2.01	2.57
From Urban Long Distance	0	2.54
From Urban Short Distance	0	0.91
Inter-Non-Urban Long Distance	0	0.18
Inter-Non-Urban Short Distance	0	0.29
Inter-Urban Long Distance	**59.3**	**35.66**
Inter-Urban Short Distance	4.52	3.17
London Travelcard	0	4.02
Non-London from London	3.02	3.48
Non-London to London	0	4.83
PTE	8.04	5.37
South East Non-London	0	13.18
South East from London	0	6.34
South East to London	0	8.58
To Airport	0	3.46
To Urban Long Distance	**19.1**	**4**
To Urban Short Distance	**4.02**	**1.42**
Total	**100**	**100**

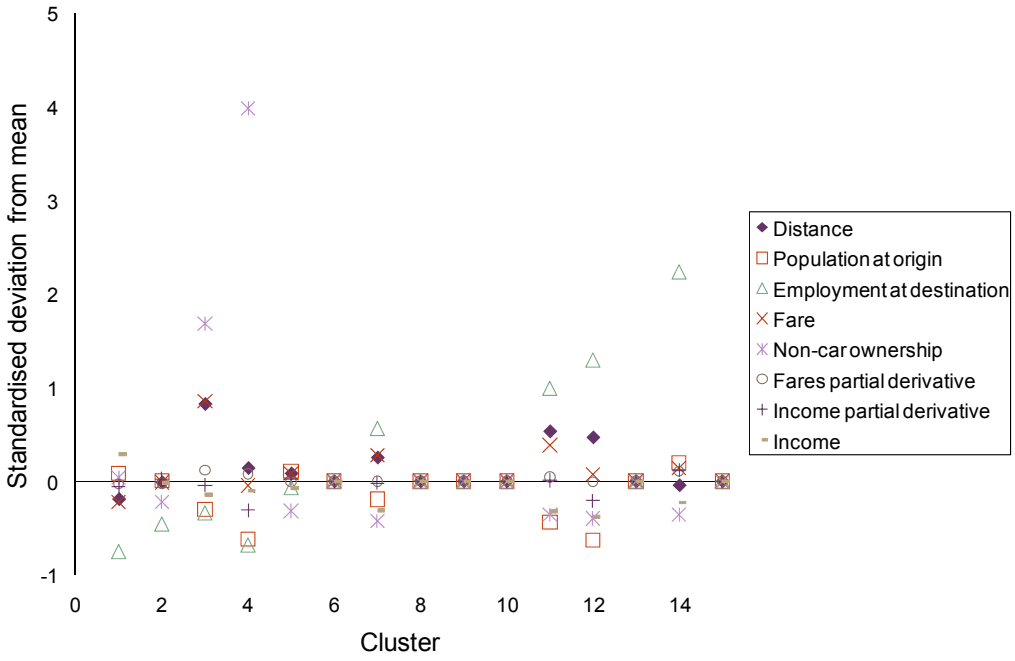
Source: Oxera analysis.

Table 4.3 suggests that the cluster includes a higher proportion of long-distance flows to/between cities (highlighted in bold). This implies that distance may be important in the market segmentation.

In summary, the cluster analysed in this worked example has a greater number of flows originating in Yorkshire & Humberside and Scotland than is the average for the ticket type. Furthermore, the cluster has more flows for which the destination is in Yorkshire & Humberside or Scotland than the average for the ticket type. In addition, there is a difference between the type of flows that make up this cluster and the type that make up the whole ticket type, with Long Distance Inter-Urban and Long Distance to Urban tickets comprising a much larger part of this cluster—78% for the cluster compared with 40% for the ticket type as a whole. This analysis suggests that geography and flow type may be important variables to consider within the market segmentation.

The next step is to examine the deviation of the characteristics from the mean of the ticket type, as shown in Figure 4.2 below.

Figure 4.2 Characteristics of standard class full fare



Source: Oxera.

Figure 4.2 shows the standardised deviation from the mean of the ticket type, for each characteristic by cluster, expressed as:

$$x_{i,standardised} = \frac{(x_i - \bar{x})}{s_x}$$

where x_i is the value of the characteristic in cluster i, $\bar{x}$ is the mean for that characteristic, and s_x is the standard deviation of characteristic x for the ticket type.

Points above the x axis show where the value of the characteristic is greater than the average for the ticket type, while values below it indicate that the value of the characteristic is below the average for the ticket type.

Figure 4.2 shows that cluster 11 differs from the mean of the ticket type in having high total jobs at destination, long-distance and high-fare flows, combined with low population at origin, non-car ownership and income at the origin of the flows. It should be emphasised that these are all *relative* to the average for the ticket type, which is shown by the x axis.

The next step in the analysis is to collate this information into a table for all the clusters within the ticket type. Table 4.4 below provides an example of this summary table for standard class full fare tickets. This table is then examined to determine whether there are patterns across the clusters which may provide hypotheses about market segmentation. For example, one geographic grouping which appears a number of times is the South East grouped with the East of England (see clusters 1, 2, and 14). Other hypotheses which can be drawn are that distance is important—with a long-distance cluster (3), medium-distance clusters (1, 4 and 11), and short-distance clusters (5 and 14), which can be determined by looking at the flow types combined with the GORs at origin and destination—and that income may be an important factor to consider in the market segmentation. However, as discussed further below, income and fares appear to be related to the functional form of the demand equation—ie, this is an issue to be dealt with in the econometrics rather than in the market segmentation.

Table 4.4 Summary table: standard class full fare

Cluster	Sample size too small (Y/N)	GOR(s) at origin	GOR(s) at destination	Flow type(s)	Other characteristics
1		London/North West	East/South East/South West	From Urban Long Distance, From Urban Short Distance, Non-London from London, SE Non-London, SE from London	Low total jobs at destination, high income
2		East/South West	East/South East	Inter-Urban Long Distance, SE Non-London, SE from London, SE to London, to airport	Low total jobs at destination
3		North West/Scotland/ Yorkshire & Humberside	East/North East/South West/ Yorkshire & Humberside	Inter-Urban Long Distance, Non-London to London, PTE	
4		Scotland/West Midlands	East Midlands/Scotland/ West Midlands	Inter-Urban Long Distance, PTE	High non-car ownership, low population at origin, low jobs at destination
5		South East	London/West Midlands	Inter-Urban Long Distance, London Travelcard Area, SE to London	Low non-car ownership
6	Y				
7		North West/Yorkshire & Humberside	East Midlands/North West/South West/Wales/Yorkshire & Humberside	Inter-Urban Long Distance, PTE, To Urban Long Distance	High population at origin, low income
8	Y				
9	Y				
10	Y				
11		North West/Scotland/ Yorkshire & Humberside	Scotland/Yorkshire & Humberside	Inter-Urban Long Distance, PTE, To Urban Long Distance	High jobs at destination, long-distance, high fare per km, low income
12		Scotland	Scotland	PTE, To Urban Long Distance	Low population at origin, high total jobs at destination and long-distance, high fare per km
13	Y				
14		East/London/South East	London/West Midlands	SE to London	High total jobs at destination, low income

Source: Oxera analysis.

The next step in the analysis is to compare the tables across all ticket types and analyse whether there are consistent hypotheses emerging from across the range of ticket types. This forms the basis of the next section.

4.3 Emerging conclusions

The final step in the market segmentation analysis is to compare the summary tables (see Table 4.4 for one example) across all ticket types and analyse whether there are consistent hypotheses emerging from across the range of ticket types. Table 4.5 summarises the conclusions drawn from the ticket-type summary tables.

Table 4.5 Proposed market segmentation

	Geographic split	Distance	Other[1]
Ticket type			
First class non-season	South East/East of England/London; the Midlands; rest of country	Short-/medium- and long-distance	Above- and below-average population at origin/ employment at destination, car ownership
Standard class full fare	South East/East of England/London; the Midlands; rest of country	Short-/medium- and long-distance	Car ownership
Standard class reduced fare	South East/East of England/London; the Midlands; rest of country	Short-/medium- and long-distance	Car ownership
Standard class Apex	South East/East of England/London; rest of country	Short-/medium- and long-distance	Car ownership
Standard class season	South East/East of England/London; rest of country	Short- and long-distance	Above- and below-average population at origin/ employment at destination, car ownership

Note: [1] These are not proposed as segmentation variables, but provide an indication of what other factors appear to be important in determining the differences between the clusters. The first class season ticket type does not provide enough clusters to allow robust analysis to be undertaken.
Source: Oxera.

In Table 4.5 there appears to be a consistent geographic split across ticket types, with the South East, London and the East of England belonging to one segment, the Midlands belonging to another segment, and the rest of the country belonging to yet another.

The Midlands—in particular, the West Midlands—is dominated by the Birmingham conurbation. To identify whether the observed responses to changes in demand are a feature specific to the Midlands, or whether they are more typical of large urban areas, further analysis was conducted by looking at the distribution of distances and fares which originate and end in major cities (eg, Birmingham, Cardiff, etc) to ascertain whether these are different to journeys originating and ending in their surrounding areas and to other large urban areas. The aim was to determine whether the appropriate geographic segmentation is by (as above), or by non-London large cities and other areas.

This analysis (an example of which is presented in Tables 4.6 and 4.7) demonstrates that flows which start or finish in the major urban areas of Great Britain both tend to be longer and have higher fares than is the average for the area in which the city is located. This effect occurs across cities and ticket types, and lends support to the hypothesis that large non-London cities should be classified as a segment within the forecasting framework. For example, Table 4.6 suggests that the average distance of flows originating in Newcastle may be longer than those in the North East as a whole. Table 4.7 shows that 21% of flows from Newcastle are From Urban Long Distance, compared with 5% for the North East as a whole.

This implies that the characteristics of travel from Newcastle are substantially different to those from the North East as a whole, which, in turn, suggests that a non-London core cities market segment may be more appropriate than a geographic-based segment.

Table 4.6 Raw distance and fare for the Newcastle area

Variable	Mean (km)	Standard deviation
Distance (Newcastle)	47	39
Distance (North East)	33	28
Fare (Newcastle)	3.74	3.03
Fare (North East)	2.75	2.39

Source: Oxera analysis.

Table 4.7 Flow type for the Newcastle area

Flow type	% frequency: Newcastle	% frequency: North East
From Urban Long Distance	21	5
From Urban Short Distance	21	9
Inter-Urban Long Distance	42	36
Inter-Urban Short Distance	16	29
PTE	0	1
To Urban Long Distance	0	11
To Urban Short Distance	0	9

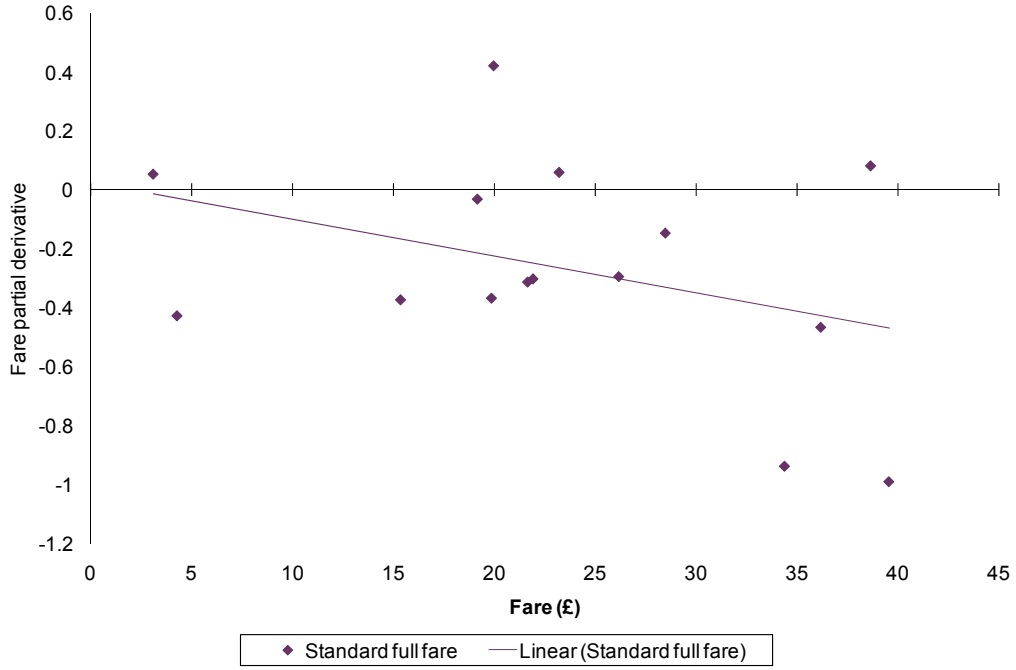
Source: Oxera.

The study team have looked at 16 cities (see Appendix 1). For many of these, there is a consistent pattern of longer distance and more inter-urban travel than is the average for the surrounding region. Distance also appears to be an important determinant of the differences between clusters across a number of ticket types. Long-distance flows are consistently different from short- and medium-distance flows. In some cases (eg, for standard class full fare ticket types), short- and medium-distance flows also appear to be separate. In others (eg, standard class Apex), there does not appear to be a difference between short- and medium-distance flows. There is no *a priori* reason to expect differences in elasticities between flows simply because one flow is slightly longer or shorter than another. For example, two flows (one of which is one mile longer than the other) are unlikely to have different responses to a change in a demand driver simply because they straddle the boundary of the distance bands. Therefore, the impact of distance has been modelled within the econometric analysis.

4.3.1 Functional form

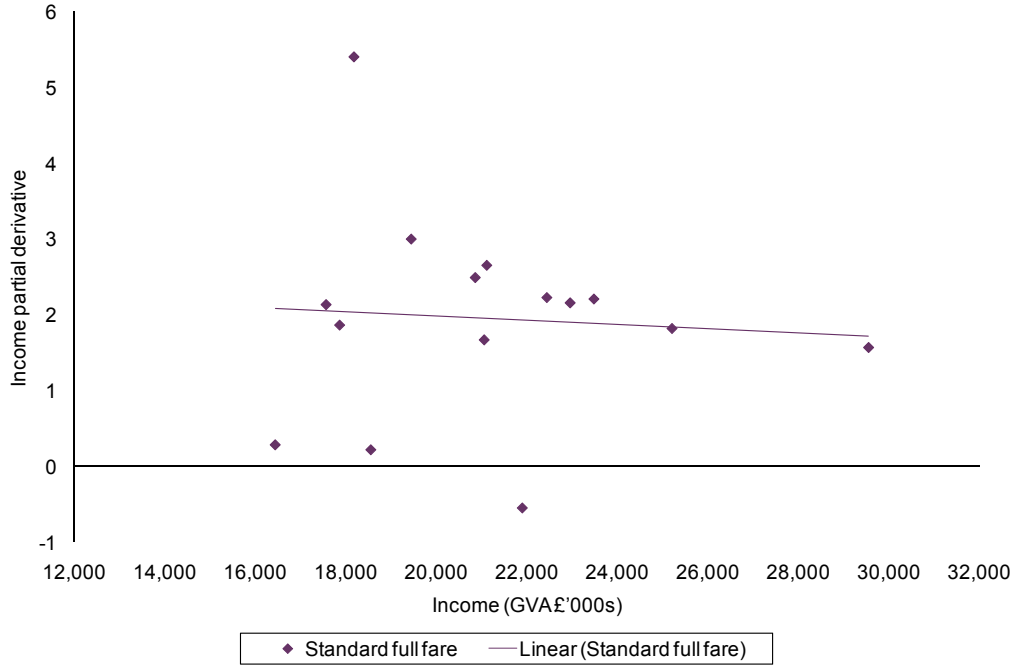
Income appears to be an important segmentation variable, but this may relate to functional form since there appears to be a linear relationship between the level of income and the income partial derivative (although there appears to be some variability in the relationship), and, similarly, between the level of fare and the fare partial derivative (see Figures 4.3 and 4.4 below). Therefore, income and fares have been treated as functional form issues rather than segmentation variables.

Figure 4.3 The relationship between fare partial derivative and fare level



Source: Oxera.

Figure 4.4 The relationship between income partial derivative and income level



Source: Oxera.

In drawing up this market segmentation, Oxera has considered a number of other characteristics, although these appear to be less important than the issues of distance and geography outlined above. Therefore, the proposed market segmentation is:

– London, the South East, the East of England (LSEE);
– core (non-London) cities;
– the rest of the country;
– airport flows.

Many variables have been considered as part of this analysis, and they are often inter-related. The decision on which to use as possible segmentation variables, and which to capture within the econometrics, is driven by a desire for consistency. Consequently, as much as possible has been included within the econometric analysis.

5 Conclusions

This report describes the process followed in this study to arrive at a proposed market segmentation for the demand for passenger rail travel in Great Britain. The proposed segmentation is as follows:

– London, the South East and East of England;
– Non-London large cities;
– airport flows;
– other.

This proposed market segmentation implies some substantive changes to that currently used in the PDFH—in particular, with the inclusion of non-London core cities as a separate segment.

The process has consisted of several stages, with each stage building on the preceding one. The bivariate analysis using the partial derivatives followed an extensive review of the existing industry literature, and suggested hypotheses to be investigated further under the multivariate cluster analysis. The multivariate cluster analysis has suggested a market segmentation that is consistent with previous industry research. The consistency of the market segmentation across ticket types, which has been identified in this study, provides confidence that the aggregation of ticket types into the three combinations (full fare, reduced fare, and season tickets) for the econometric analysis is unlikely to have introduced aggregation bias into the elasticity estimates.

The process followed has resulted in a market segmentation based on data and supported by economic theory. The proposed segmentation builds on previous industry knowledge and research.

A1 Cities analysed for non-London core cities segment

Cities identified with an asterisk (*) were examined, but are not included in the proposed non-London core cities segment as they do not appear to share the same characteristics (of longer distance, inter-urban flows) as the other cities.

Birmingham

Blackburn*

Bristol

Cardiff

Coventry*

Crewe*

Durham*

Edinburgh

Glasgow

Huddersfield*

Hull

Leicester

Leeds

Liverpool

Manchester

Middlesbrough*

Newcastle

Nottingham

Sheffield

York

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