



# Rail Demand Forecasting Estimation (redacted)

Phase 1 Report – September 2015

Leigh | Fisher



# Index



	Page
Introduction	3
Literature Review Summary	5
RUDD Data Review	22
NTS Data Review	38
Phase 1 Summary	55
Phase 2 Recommendations	59

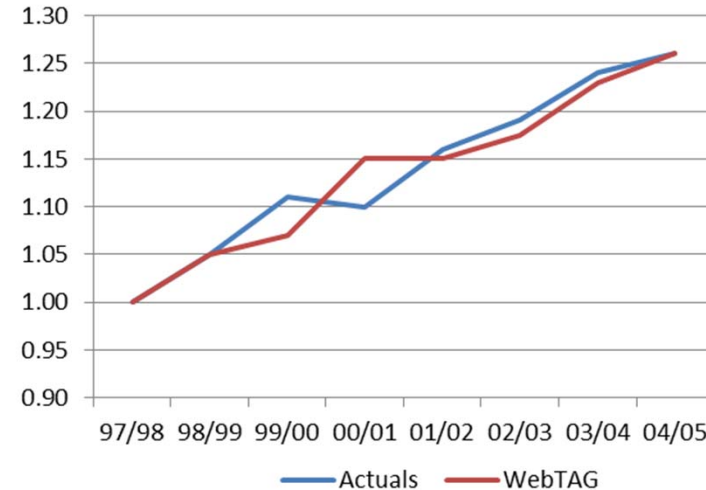
# Introduction



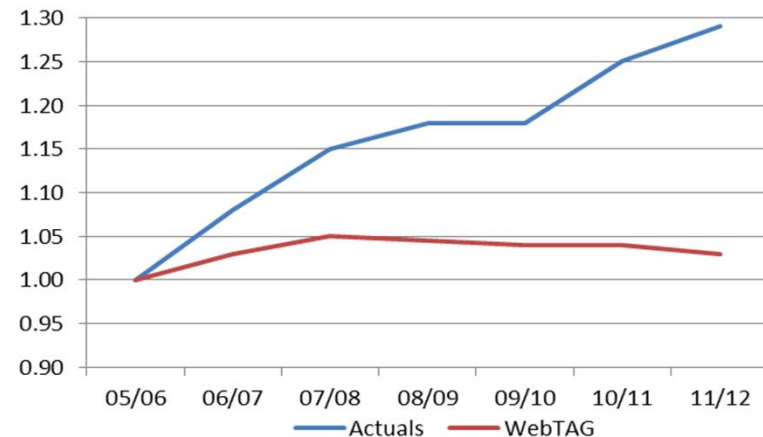
## Purpose of Rail Demand Forecasting Estimation Study

- Since circa 2005 actual demand growth has exceeded aggregate predictions using PDFH approach
- Performance improves with market segmentation but under-forecasting still a major issue
- Purpose of study is to estimate new parameters for external factors
  - ◆ Robust
  - ◆ Suitable for use by DfT
  - ◆ May include updates to existing parameters as well as introducing new parameters

Passenger kms - All Flows



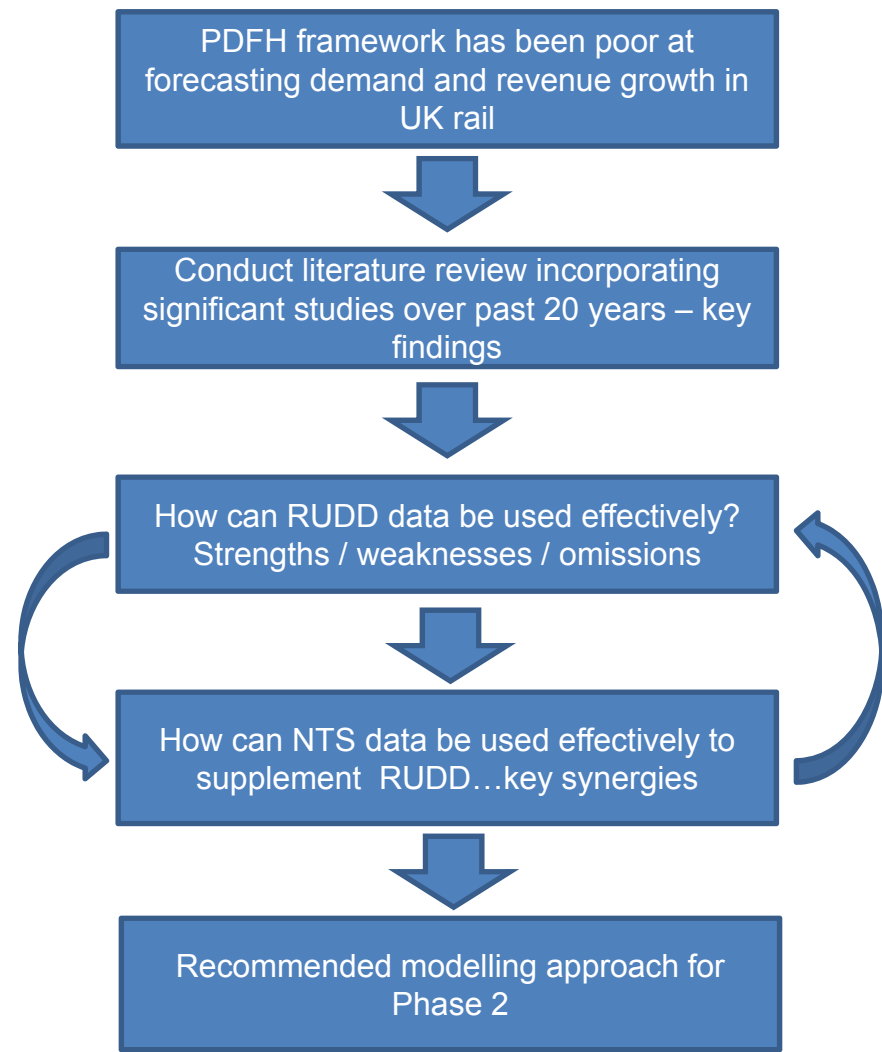
Passenger kms - All Flows



# Introduction

## Approach

- Study split into two phases:
- Phase 1 (two months)
  - ◆ Literature Review
  - ◆ Initial analysis of RUDD data
  - ◆ Initial analysis of NTS data
  - ◆ Workshop to discuss findings
  - ◆ Recommendations for Phase 2
  - ◆ Phase 1 report
- Phase 2 (four months)
  - ◆ Collection of additional data (if required)
  - ◆ Econometric Analysis
  - ◆ Final Report





# Literature Review Summary

Leigh|Fisher





# Scope of Literature Review

## RANGE OF STUDIES

- Econometric Analysis of Ticket Sales (23 studies)
- Analysis of National Travel Survey Data (7 studies)
- Reviews (4 studies)
- Explanatory Studies/Think-Pieces (9 studies)

## COVERAGE OF EVIDENCE

- Existing PDFH parameters

*Income/economic activity (GDP, employment, other income measures)*

*Population Elasticities*

*Car Ownership Effects*

*Cross Elasticities with respect to car and bus*

- New Insights

*Greater Breadth*

*Greater Depth*

# PDFH Forecasting Problems: Not for the First Time!

- In mid and late 1990s, PDFH seriously understated rail demand growth
- PDFH was then based on (positive) GDP/Employment elasticity and (negative) trend
- The negative time trend discerned:
  - Strong growth in car ownership*
  - Road building and new car journey opportunities*
  - Lower fuel and operating costs*
  - Increased coach competition*
- But the trend reflected conditions prior to 1990s. In the 1990s:
  - Fuel price increases (fuel duty escalator)*
  - Congestion increasing and more widespread*
  - Car ownership growth slowing (saturated in some rail markets)*
  - Other positive trends for rail (eg, road unreliability, increased environmental concerns etc.)*
- Industry slow to recognise and respond to the changed environment

# The Correlation Problem

## TCI (1997) Effects of External Factors

The first major study to address the forecasting problem (1997)

Mileage		0-24	25-49	50-149	150+
InterCity London	GDP Trend		■ -0.9%	■ -0.6%	■ 2.3%
InterCity Non London	GDP Trend	■ -1.3%	■ 0.2%	■ 2.6%	■ 3.9%
South East London	GDP Trend	■ -2.1%	■ -1.5%	■ -2.4%	
South East Non London	GDP Trend	■ -2.8%	■ -3.2%	■ -3.3%	
Regional	GDP Trend	■ 3.1%	■ 1.5%	■ 3.9%	■ 4.2%

Correlation of GDP elasticity and time trend -0.83

Persisted into the 2000s!



# Response to Forecasting Problem

- National Passenger Demand Forecasting Framework (1999)

  - Introduced explicit cross-elasticity terms*

  - Revised the GDP/employment elasticities*

  - Fresh econometric analysis*

  - But unaccounted for effects were assigned to time trends*

- PDFH v4 (2002)

  - Adopted a lot of NPdff framework and some parameters*

  - Estimated new models with GDP elasticities consistent with cross-elasticity terms - which were constrained to (NPdff) recommended values*

- Recommended PDFH Approach

  - Changes in rail demand a function of changes in:*

    - GDP/employment*

    - Population*

    - Proportion of households with no car*

    - Car cost and time*

    - Bus cost, time and headway*

    - Air cost and headway*

- Subsequent modelling (as opposed to forecasting) did not really follow this approach

# Evolution of PDFH

- PDFH v3 (1997)
  - GDP/employment elasticity and time trend for 7 flow/ticket types
- PDFH v4 (2002)
  - Adopted a modified NPdff approach. No mention of regional GVA. Airport flows added.
  - Segmentation by distance bands, to and from London and ticket type
- PDFH v4.1 (2005)
  - Encouraged local income measures. Merged ticket type and distance categories
  - Large incremental distance effect on GDP elasticity for long distance London
  - Population elasticity relates to relative population growth.
  - Method for deducing cross-elasticities
- PDFH v5 (2009)
  - Removed distance effect on long distance London GDP elasticity
- PDFH v5.1 (2013)
  - Employment elasticity  $> 1$  for non London short
  - Non London flows split into to and from core cities, between major cities and other
  - Some appreciable increases in cross-elasticities to car time and car cost
- Little change in framework/parameters in past 13 years

# Post PDFH v4 Models (I)

- LOT OF MODELLING BUT NOT ENTIRELY SATISFACTORY RESULTS

*Some studies only specify GDP – interpretation issues of what elasticity represents*

*Some studies specify additional external effects – with very mixed results!*

*The concept of constraining elasticities to best evidence not favoured*

- EXAMPLES OF ‘POOR’ RESULTS IN 2000s

OXERA/ARUP (2010) study 19 Non Seasons models

*Car cost: significant in all but 2 BUT 8 (42%) > 1 and 15 (79%) >0.5*

*Car ownership: significant in four (one wrong sign!) and remaining three average -2.1!*

*Population elasticity: either ■ or ■*

*GDP elasticity: 5 (26%) > 2 and 8 (42%) >1.5*

Wardman and Dargay (2007) 4 Non Seasons models

*Car cost: Average ■ across flows*

*Car time: Wrong sign 2 out of 4 models*

*No Car: Wrong sign 3 out of 4 models*

*Population: ■*

# Post PDFH v4 Models (II)

Mott Macdonald and University of Southampton (Most Recent PDFC Study)

- 8 flow type models by full, reduced, season ticket types
- Car Cost: 54% significant and correct sign
- Car Time: Insignificant  $\frac{3}{4}$  of times
- No Car: Wrong sign significant 67% of times. ■
- Population: ■
- Employment Elasticity: ■
- GVA Full Elasticity: ■
- GVA Reduced Elasticity: ■

# Quality of Econometric Evidence

- No shortage of very precisely estimated GDP and employment elasticities but:
  - Some are not credible*
  - Some must be (are) influenced by correlation with omitted factors or with included factors*
  - But some are wrong sign*
- Fuel price is the most robustly estimated cross-elasticity
- Estimated car journey time elasticities invariably poor
- Population and car ownership terms generally unreliable
- Little success exploring new terms or with more detail

# Problems

The problems with the forecasting approach are:

- Quality of Evidence relating to Existing PDFH variables
- Range of influences and detail absent from PDFH forecasting

The shortcomings of existing models:

- GDP is sometimes the only external factor
- Long run GDP elasticities can be very high
- Correlation of external factors
- Large data sets (covering 'all flows')
- Ticket type models (full and reduced) and non-price competition
- Many factors unaccounted for



# Experiences with NTS Data

Considerable unexploited potential to enhance (or replace!) ticket sales models.

Impact of range of socio-economic, demographic and locational factors on rail demand can be obtained

- Car ownership effect in PDFH obtained from NTS analysis
- Wardman (2006) aggregate analysis of NTS data – income elasticities broadly comparable with ticket sales income elasticities
- TRL ‘White Book’ (2004) Mode choice commuting model, whether make rail trips and how many (by purpose)
- Dargay (2010) long distance (Demand) model uses socio-economic effects on demand from analysis of NTS trip rate data
- RAND Long Distance (Choice) Model based on NTS
- Le Vine and Jones (2012) On the Move Study provided insights
- Network Rail long term forecasting makes much use of NTS trip rate models

Further NTS analysis clearly warranted:

- Need to be careful not ‘double counting’ when enhancing existing models
- Need to be careful with model properties
- Better for socio-economic and demographic than transport system costs and times

# Way Forward: 'More of the Same' Is Not an Option!

- Historical inertia in rail demand modelling approach

  - Consistency with PDFH Approach*

  - Data limitations*

- The Way Forward

  - More explanatory variables (breadth)*

  - Better data (depth)*

  - Improved model specifications*

# Suggested Exploratory Factors

Some are one-off demand impacts, others are ongoing

- Changes in Employment Market

  - Changes in labour market towards jobs where people more likely use train*

  - Changes in labour market to locations better served by rail*

  - Regionalisation with competition in jobs market leading to more rail use*

- Regeneration of Regional Centres

  - Growth in leisure opportunities in locations well served by rail*

  - Competition with local centres leads to more rail use (destination choice identified by Worsley)*

- Digital Revolution and Attractiveness of Rail

- Car Parking Cost and Availability

- Company Car Tax and Business Travel

- Changes in Socio-Economic Characteristics and also Cohort Effects

- Non Price Competition between Tickets

- General rail trends not explicitly in models

  - Rolling stock improvements, Better marketing (easier purchase, better information, special offers etc.), Revenue protection, Environmental attitudes

# Better Data (for Modelling)

More explanatory variables has data implications

For existing PDFH parameters:

- Better car journey time data clearly needed
- Use more locally relevant/detailed data
- Improved data on diversion factors for cross-elasticities
- Exploit NTS data

# Model Specification

- More market segmentation – flow type, ticket type, journey purpose
- Allow elasticities to vary over time (might expect to be increasing)
- Use theoretical considerations more (eg, GVA origin or destination, population and employment relationships in commuting market)
- Outside evidence to enhance forecasting framework as well as modelling
- Careful consideration of directionality
- Careful consideration of non-price competition between products
- Cross elasticities linked to local circumstances (Oxford-London v Ipswich-London)

# Way Forward – This Study

- Re-estimate OXERA/ARUP models with newer data AND constrained parameters
- Compare different input variables (e.g. for income) and different model specifications
- Consider theoretical relationships
- Add new variables (suggested supplementary exogenous / mode choice variables shown below):
  - ◆ **Population by age band**
  - ◆ **Employment by type of occupation (SOC) or possibly industry**
  - ◆ More detail on **car ownership** (e.g. cars per licence)
  - ◆ **Car time/speed**
  - ◆ **Bus cost/speed**
- Add other endogenous variables
- Within project resource, refine cross-elasticity constraints to be route specific (deduced with Dargay (2010) diversion factors and NTS shares)
- Improved segmentation (e.g., extend locational hierarchies, distance, temporal, flow size)
- Improved allowance for non-price competition for full and reduced tickets
- Use NTS to obtain car journey times (Car time model). Also other modes/variables?
- Use evidence on values of time, Hensher parameters to proxy for digital revolution



# Way Forward - Long Run

- Set in motion collection of relevant data for future modelling
- Analysis of Census data and other data sets (FES, labour market surveys etc)
- Set in place purpose collected data on ongoing basis to develop rail trip models (as in long distance travel survey of many years ago)
- Achieve better understanding of cross-modal diversion factors by distance, purpose and relative attractiveness of modes
- Source existing data to enhance models (eg, local income data)
- Continual refinement of ticket sales models as data and resources become available (no more inertia!!)
- Possible exploitation of cross-sectional models – access/egress, catchments
- Compare enhanced PDFH with other approaches (LDM, NR, Dargay) and backcast as part of broader appraisal of rail forecasting
- Development of forecasting models based entirely on disaggregate approaches



# RUDD Data Review

1. Outcome from a backcast
2. Trends in the data
3. Description of the dataset
4. Fitness for purpose

Leigh | Fisher



  
UNIVERSITY OF LEEDS

# RUDD Data Review



## Introduction

In order to:

- provide context for the changes in rail demand in the last two decades, which our phase 2 work will seek to better explain;
- discuss what RUDD (Rail Usage and Drivers Dataset) contains that can be used in such work

This section:

1. Uses RUDD to undertake a 'backcast', seeing how rail demand has outperformed given the exogenous drivers and the elasticities prescribed in PDFH/WebTAG
2. Describes the trends that can be seen in RUDD that will inform our modelling
3. Discusses the data that are contained in RUDD and how useful they would be in backcasting and in econometric modelling.

# RUDD data: Description

- RUDD (Rail Usage and Drivers Dataset) is a balanced panel
  - ◆ 21,000 flows (origin-destination pairs)
  - ◆ for 21 years – rail years 1995 through 2015
  - ◆ with more than 800 descriptive variables
- There is a lot of duplication in the dataset (one 2GB .csv file), because most of the descriptive variables are attributes of the origin or destination in that year, and each origin or destination
- Many of the descriptive variables cover only one year or only part of the dataset
- Some of the flows have no demand in part of the data (as they are to/from stations like West Brompton that were not open throughout the period)
  - ◆ although they still have descriptive variables
- It is based on TOAD (The OXERA-ARUP Dataset), and the 21,000 flows cover £6.7bn of revenue in rail year 2014 – about five-sixths of the UK total
  - ◆ Most of the remainder will be internal to the Travelcard area

# RUDD data: Description

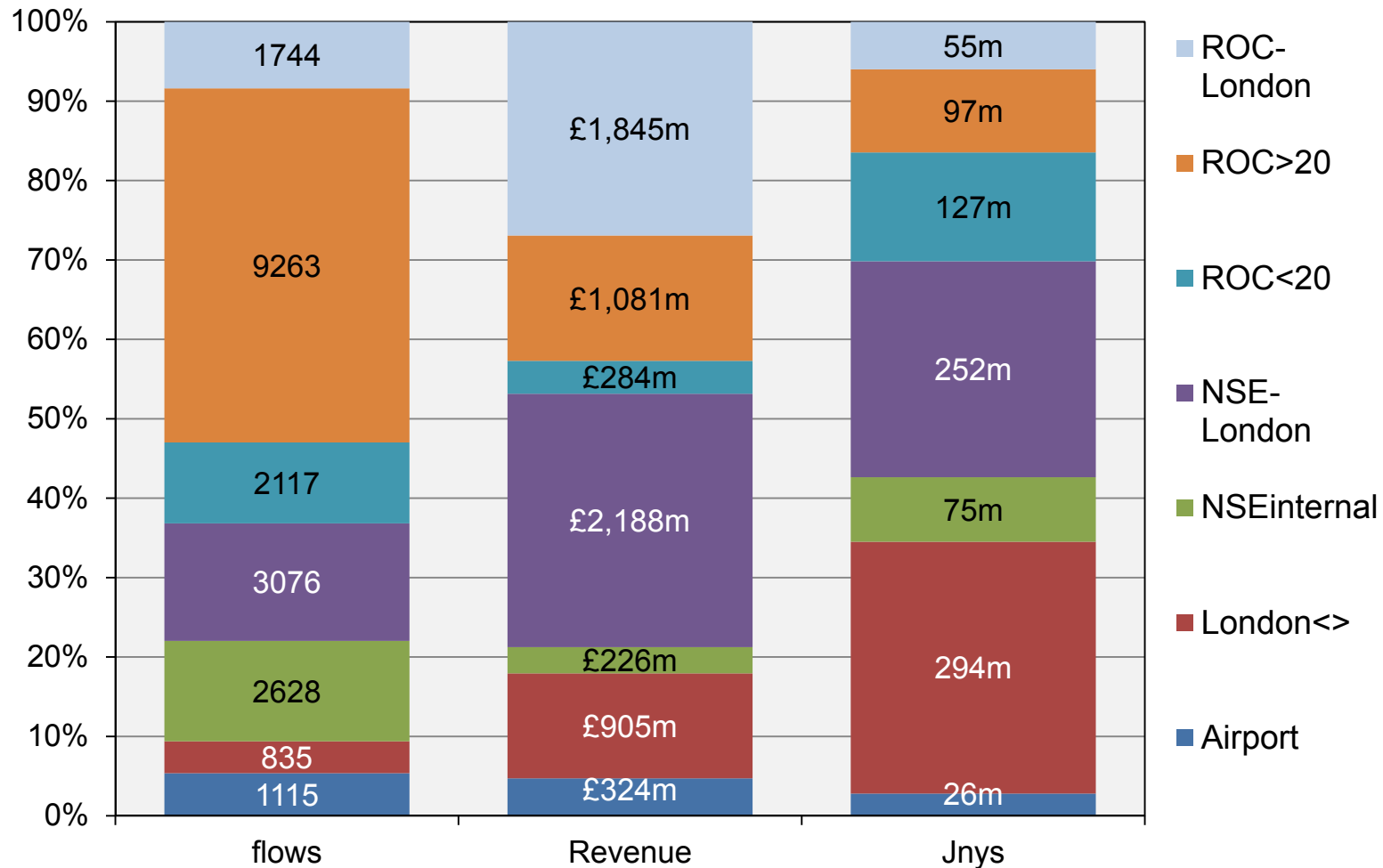
## Coverage

- The data largely exclude the Travelcard area due to data quality issues
  - ◆ The only flows which **are** included are to **or** from group stations – largely London BR and “London R1” (which in RUDD groups Travelcards and Tube stations) but also Croydon BR
  - ◆ We would probably expect the remainder of the Travelcard area flows not to be used because of remaining data quality concerns
  - ◆ Travelcard area represents the bulk of the ‘infills’ **which are not in RUDD until 2005**
- **RUDD does include** PTE areas, but not any O-D infills for PTE products

# RUDD data: Description

## Coverage

- The flows are distributed as follows (in 2014):



**Commercially Sensitive** □ **Not Quality Assured**



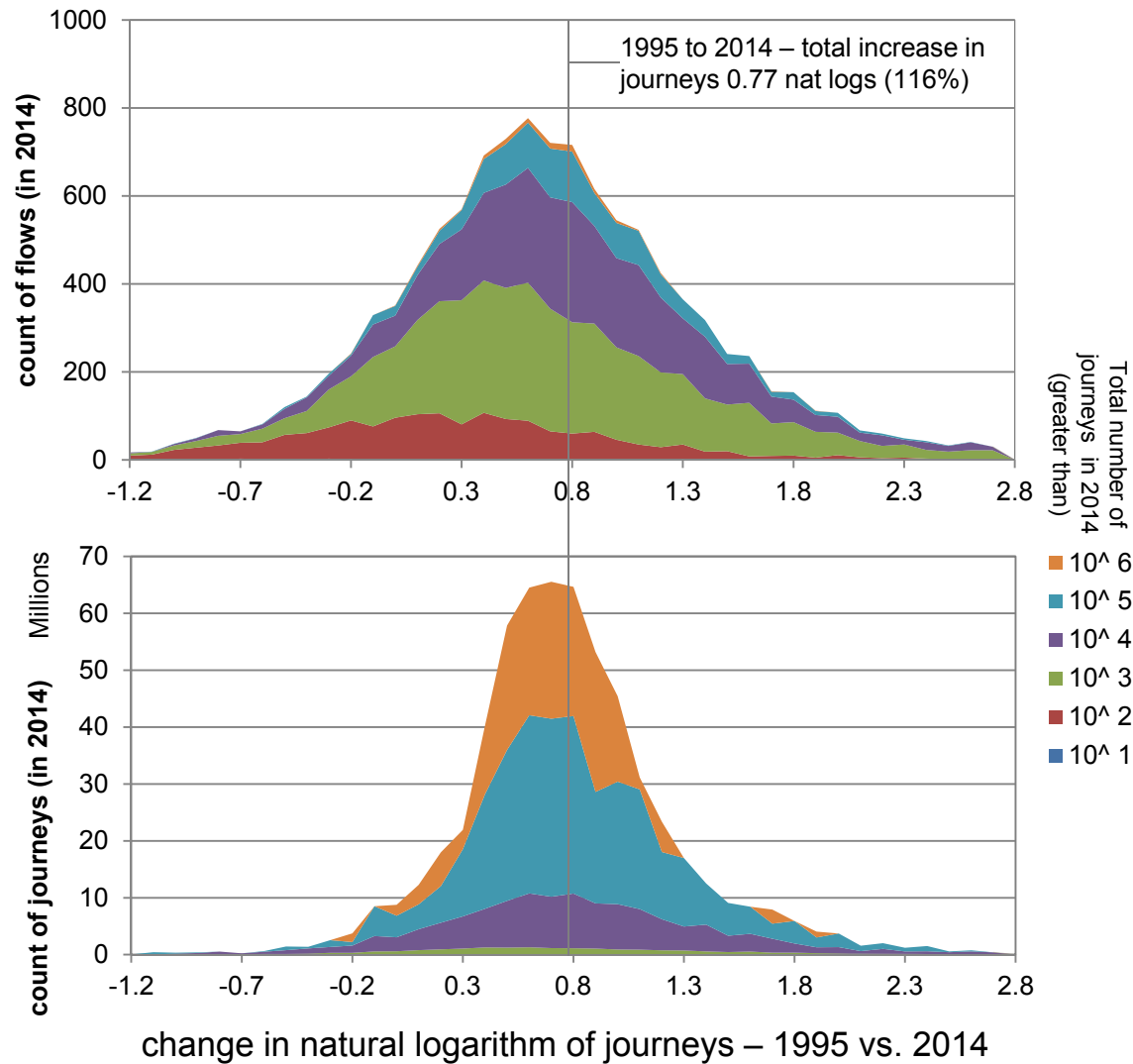
# RUDD data: Description

## Flow size

- Most studies have excluded flows below a certain size (as the RUDD database does) because their behaviour might not be normal
  - ◆ There are likely to be econometric problems associated with ticket type – flow combinations that have no demand in some years (or negative demand because refunds were processed in a subsequent rail year). This is a particular problem for first class and long distance season tickets.
- Behaviour of flows of which larger volumes are sold does seem to be more “typical” than low volume flows

# RUDD data: Description

## Flow size



Commercially Sensitive  Not Quality Assured

# RUDD data: Description

## Flow size

- Most studies have excluded flows below a certain size (as the RUDD database does) because their behaviour might not be ‘normal’
- Behaviour of flows of which larger volumes are sold does seem to be more “typical” than low volume flows – intuitively, flows which incorporate journeys made by more people will be less prone to idiosyncratic changes in individuals’ behaviour
- Weighting the econometrics to minimise the variance of error weighted by (say) passenger miles may be more robust than excluding certain flows – otherwise part of the market is neglected (although we have not looked for evidence that small flows really behave any differently from large flows)
- There is a trade-off between forecasting individual flows well, and forecasting the impacts on total rail travel, implied in the way one undertakes econometrics

# RUDD data: Description

PDFH “External Factors”

## GDP *per Capita*

RUDD contains data on GVA and GDI (Gross Disposable Income) at GOR (Region), NUTS2 (County) and NUTS3 (District) levels, together with population (to derive GVA/GDI per capita). No data before Rail Year 1998.

## Population

RUDD contains mid-year population estimates at GOR, NUTS2, NUTS3, County & District levels through 2013. RUDD also contains population for 1km, 5km and 10km station catchments, but only for (Census) Years 2002 and 2012 – not clear what use this would be. RUDD contains the following divisions of population data:

<b>Age</b>	<b>Ethnic Group</b> (census years only; not all areas)	<b>Other</b> (2005-2013 only)
0-14		Working Age
15-29	1 White	In Higher Education
30-44	2 Indian/British Indian	Employed
45-64	3 Pakistani/British Pakistani	Unemployed
65+	4 Bangladeshi/British Bangladeshi	
	5 Chinese/British Chinese	
	6 Other Asian/British Asian	
	7 Black (British) African	
	8 Black (British) Caribbean	
	9 Black (British) other	
	10 Other	

# RUDD data: Description

PDFH “External Factors”

Leigh|Fisher



UNIVERSITY OF LEEDS

RUDD contains estimates for employment:

- ◆ At GOR, County & District levels for 1995 through 2013 from LFS, **but with a break between 2003 and 2004 due to a methodology change**, making backcasting from earlier very difficult
- ◆ For 1/5/10 km catchments from 2004 only
- ◆ From NTEM for the entire period (though these are modelled figures, every five years with interpolation in the interim)
- ◆ based on residence *and* workplace

RUDD also contains employment disaggregated by sector, occupation type, full and part time, for the same geographies:

Employment

## Sector

1. Agriculture, forestry & fishing
2. Mining, energy and water supply
3. Manufacturing
4. Construction
5. Wholesale; motor vehicles; accommodation and food services
6. Transport, storage, information, communication
7. Financial, insurance, real estate; professional, scientific & technical; administrative, support services
8. Public admin, defence; social security, education, health, social work
9. Other services

## Occupations

1. Managers, Directors, Senior Officials
2. Professional Occupations
3. Associate Professional & Technical
4. Administrative and Secretarial
5. Skilled Trades
6. Caring, Leisure and Other Service
7. Sales and Customer Service
8. Process, Plant, Machine Operatives
9. Elementary

# RUDD data: Description

PDFH “Inter-modal competition”

Leigh|Fisher



UNIVERSITY OF LEEDS

## Non-car ownership

RUDD contains estimates of non-car ownership from NTEM – i.e. modelled numbers every five years with linear interpolation in between. There are also NTS data on licence holding. These should be adequate for back-casting, but not for estimating elasticities.

## Fuel cost

*“the combined effect of fuel prices and fuel efficiency”*

RUDD contains two sets of car cost data:

1. Derived from car time and speed-cost curves. Increases from 7.5p/km to 12.5p/km from 1995 to 2014
2. DECC pump (petrol) price data. Increases from 52p to 135p

The car time data are not complete (see below).

**No information on car parking prices.**

## Car time

RUDD contains data on car journey times, sourced from TrafficMaster from 2007 and NTM before then, with the NTM journey times rebased to make a continuous series.

Movements seem modest and in-line with intuition.

There are no car time data for trips to, from or within Scotland, and for some flows/years not captured in TrafficMaster.

The data do not distinguish between car times by time of the day.



# RUDD data: Description

PDFH “Inter-modal competition”

## Bus cost

No data on bus or coach cost are included in RUDD. DfT publish a bus fares index, but not a *yield* index. For bus and coach, the ‘DOCX’ component of RPI provides a single national index. Both of these datasets would exclude any specific effects on specific flows.

## Bus time

No data on bus or coach journey times are included in RUDD. It might be reasonable to assume that the change has been similar to car time.

## Bus headway

No data on bus or coach headway are included in RUDD. DfT publish data on bus (but not coach) vehicle miles at GOR level; headway can then be derived with car times as a very rough estimate

## Air competition

RUDD matches long distance rail journeys with corresponding airport flows, although this may not cover the choice of London airport completely. CAA data are included on the following:

- ◆ Frequency for 1996 to 2014 (only when it was at least one flight per day)
- ◆ Air fare for 2003 to 2011 (based on observations in 2003, 2007 and 2011)
- ◆ Air route passenger volumes for 2002 through 2014
- ◆ Share of travellers on that air flow accessing the airport by rail

# RUDD data: Description

Rail factors / Network variables

## Fares

RUDD includes revenue and journeys data, which can be used to derive yield, by three (F/R/S) and six (1O/1S/2A/2F/2R/2S) ticket types

There are no data on advance quotas or other ticket restrictions, consistent with the usual experience.

## GJT

RUDD includes Monday to Friday GJT for each of the three ticket types (F, R, S) divided between journey time, frequency and interchange penalties

## Performance

RUDD contains AML (which is used in schedule 8 to calculate the revenue impacts of poor performance), PPM and CaSL data. Some years are missing for some flows, and no data are included before 2003 (2006 for PPM) or after 2009.

## Crowding Rolling Stock Station Facilities Marketing

No data on these factors in RUDD

# RUDD data: Fitness for purpose

RUDD is a rich, useful dataset, with a large amount of aggregate data on demand drivers, however

- There are some drivers of rail demand which are not included in RUDD. Some of these are simple to source from elsewhere, but for others data are limited/non-existent
  - ◆ There is a risk of neglecting certain drivers of demand because they are not covered – such as crowding or marketing – with a risk of Omitted Variable Bias
- There are some drivers of rail demand for which RUDD data are limited in coverage
  - ◆ For example, AML and car journey time
  - ◆ It would be necessary to decide whether these require that flows be dropped as data points or that the relevant variable be ignored
- The aggregate nature of most of the independent variables should lead to few very large movements which might adversely affect estimation
  - ◆ The only exceptions are yield/GJT – large movements in yield on small volumes are unlikely to reflect fare change; “large” movements in GJT may lead to changed rail-heading on nearby flows
- There are some drivers of rail demand where the RUDD data are likely to be sufficient for controlling for certain variables, **but not estimating effects**
  - ◆ For example, data on carless households are modelled (estimates) – they are probably unbiased but any estimate effect would be attenuated. When the variables are correlated with other variables, then estimated effects could vary drastically from the true impact

# RUDD data: Fitness for purpose

RUDD is a rich, useful dataset, with a large amount of aggregate data on demand drivers, however

- There will always be data points for which the relationship between aggregate variables and the disaggregate variables that affect individuals' decisions break down

GDP *per Capita*

Population

Employment

- Even when 'average' aggregate relationships are holding up, there is an omitted variable problem – where the omitted variable is one that described the disaggregate relationships correctly
- Because the aggregate variables are correlated with one another, we may not be able to estimate 'average' aggregate relationships very well, and get completely outlying results

# RUDD data: Fitness for purpose

- Some continued concerns about robustness of the data
  - ◆ **Infills for the airport segment** are only included from 2007 – difficult to provide sensible analysis in periods where some sales are not captured
  - ◆ **RUDD excludes PTE infills** – although growth in season tickets have been robust where they do not exist
    - lingering concerns that point-to-points have grown differently from PTE products;
    - Excluding non-PTE cities will leave only a limited pool of major travel-to-work destinations



# NTS Data Review

Leigh | Fisher

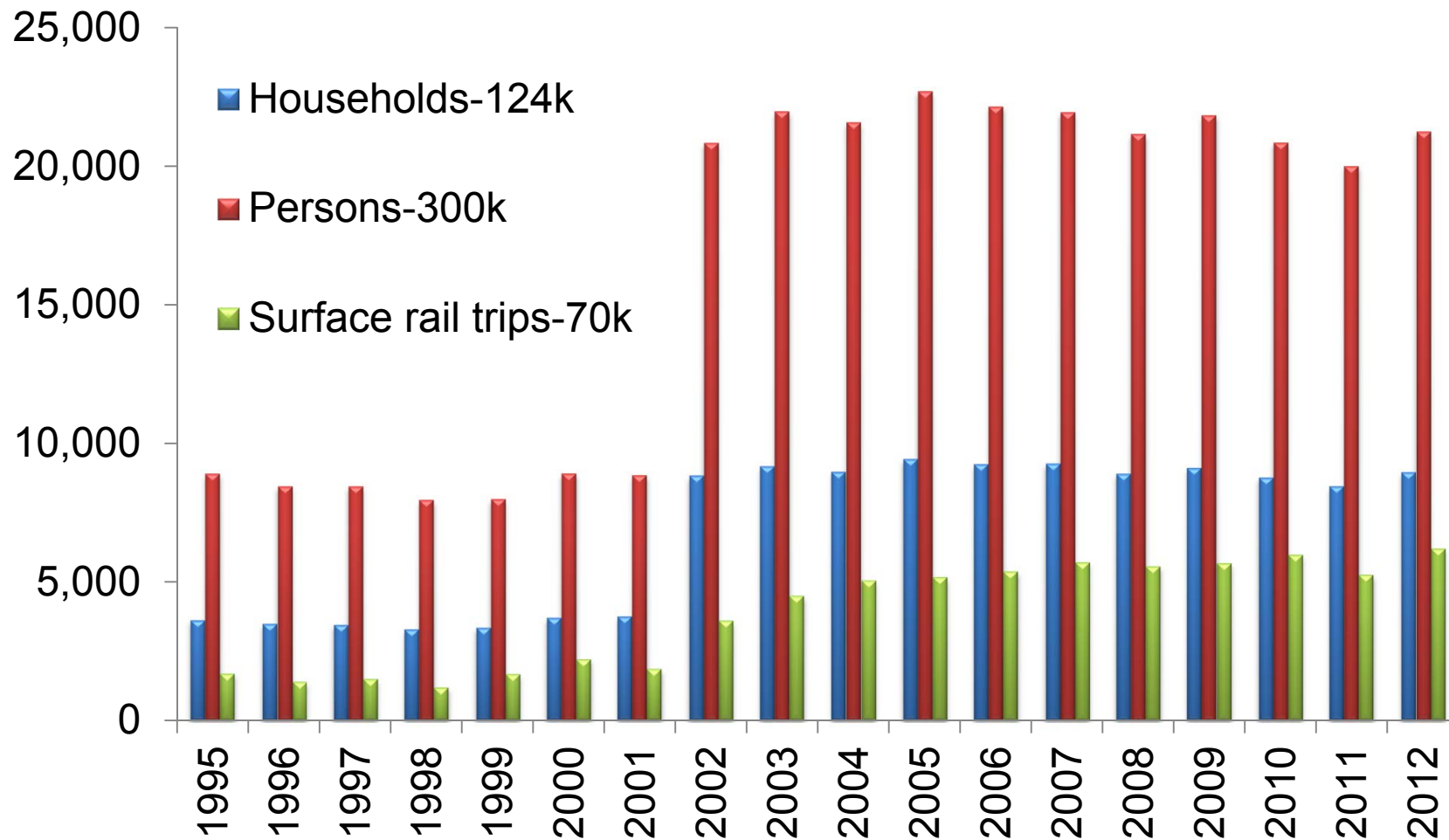


# NTS data

## Introduction

- Rich dataset for analysis with information recorded at a number of levels
  - ◆ One of the main sources of data on personal travel patterns in Great Britain
  - ◆ Commissioned in 1965/66 and carried out on ad-hoc basis till 1986.
  - ◆ Continuous survey since 1988
- Detailed information on the key characteristics of
  - ◆ Households
  - ◆ Vehicles
  - ◆ Information collected for all individuals in the household
    - Proxy information for children under 11
  - ◆ Travel diary: 7 day travel record
    - Long distance journeys have recall period
- Very rich in socio-economic detail
- But difficult to relate to network variables
  - ◆ Resource constraints, very limited location detail
  - ◆ Probably limited to aggregate data for rail, road (and maybe air)
  - ◆ Propose to source information on rail GJT, rail fare, car-time and car-cost from RUDD database

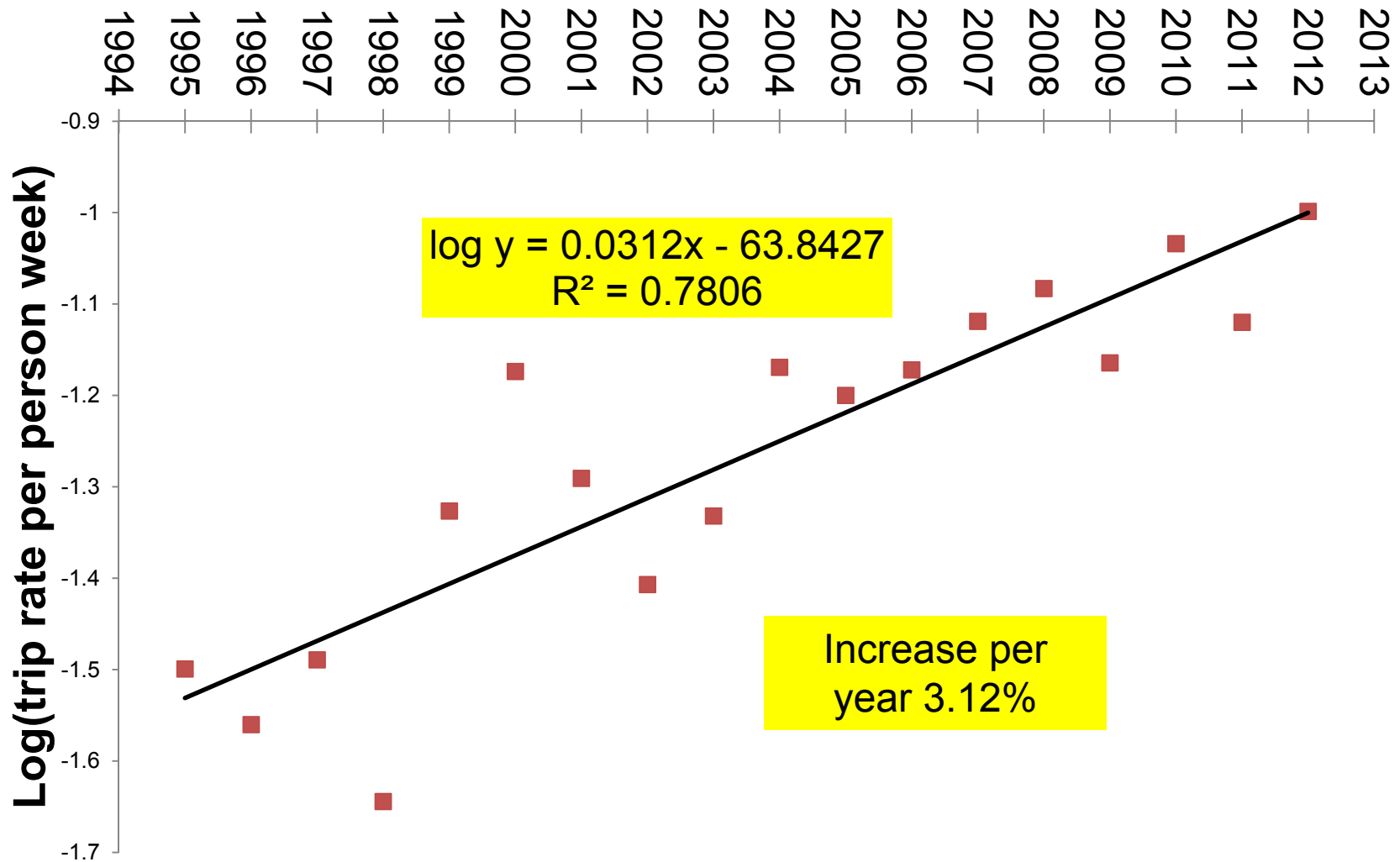
# NTS data: Sample size (1995-2012)





# NTS data: Trends

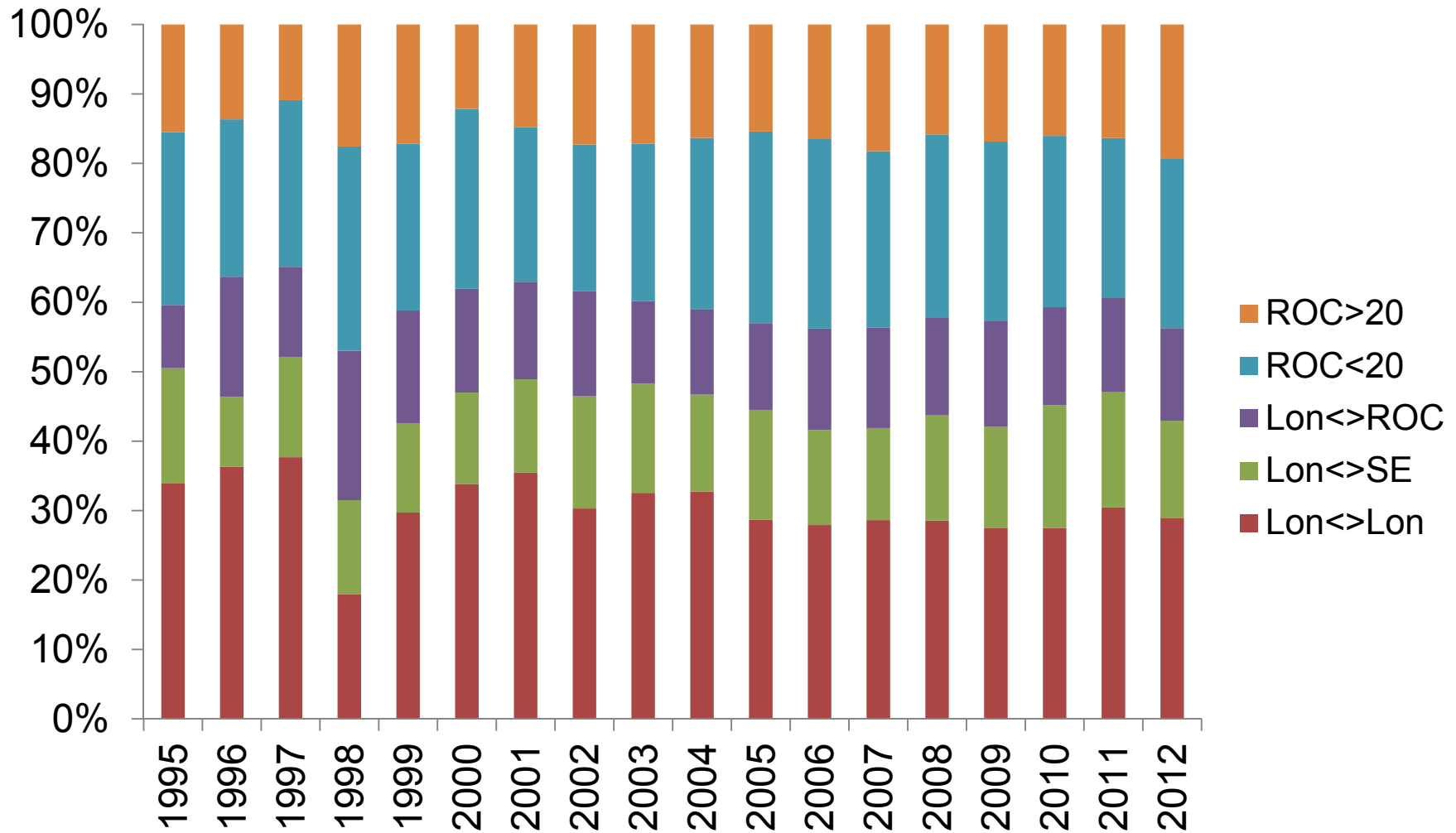
Surface rail trip rates (broadly in line with RUDD when adjusted for population growth)



# NTS data: Trends



Surface rail trips by approximate RUDD flow type



# NTS data: Description

Variables for modelling include

## Socio-economics

Age

Gender

Education level

Ethnicity, but not ideal

Income, personal or household

Employment, number of hours and the type of employment

## Geographic

Urban density / area type

Region of residence, in particular London/non-London

Region of destination, in particular London/non-London

## Car-ownership

Personal or household cars

Company-cars-  
Individual/household

Licence holding

## Trip

Purpose

Length

Approximate flow type

Ticket type- full, reduced or season

## Time

Year, to indicate trends not otherwise captured

# NTS data: Description

## Socio-economics

### ■ Age

- ◆ Continuous variable
- ◆ RUDD bands can be derived easily

### ■ Ethnicity

- ◆ Available from 1995-2012
- ◆ 15 groups - RUDD types fit within these
  - White British
  - Other white background
  - White and Black Caribbean
  - White and Black African
  - White and Asian
  - Any other mixed background
  - Indian
  - Pakistani
  - Bangladeshi
  - Any other Asian background
  - Caribbean
  - African
  - Any other black background
  - Chinese
  - Any other

### ■ Income

- ◆ Unique to NTS data
- ◆ Banded household and personal incomes available from 1995 to 2012
  - extended from 21 to 23 bands from 2002

Less than £1000	£8000- £8999	£30,000 - £34,999
£1000- £1999	£9000- £9999	£35,000 - £39,999
£2000- £2999	£10000- £12499	£40,000 - £49,999
£3000- £3999	£12500- £14999	£50,000 - £59,999
£4000- £4999	£15000- £17499	£60,000 - £69,999
£5000- £5999	£17500- £19999	£70,000 - £74,999
£6000- £6999	£20000- £24999	£75,000 +
£7000- £7999	£25000- £29999	

Bands in green are available from 2002 only

# NTS data: Description

## Socio-economics: Employment

### Economic status

Employees: full-time

Employees: part-time

Self-employed: full-time

Self-employed: part-time

ILO unemployed

Economically inactive: Retired

Economically inactive: Student

Economically inactive: Looking after family/home

Economically inactive: Permanently sick/disabled

Economically inactive: Temporarily sick/injured

Economically inactive: Other

### Occupation (SOC classification)

Managers and senior officials

Professional occupations

Associate professional and technical occupations

Administrative and secretarial occupations

Skilled trades occupations

Personal service occupations

Sales and customer service occupations

Process, plant and machine operatives

Elementary occupations

- Standard Industrial classification (SIC) is available in NTS
  - A - Agriculture, hunting and forestry
  - B – Fishing
  - C - Mining and quarrying
  - D – Manufacturing
  - E - Electricity, gas and water supply
  - F – Construction
  - G - Wholesale and retail trade
  - H - Hotels and restaurants
  - I - Transport, storage and communication
  - J - Financial intermediation
  - K - Real estate, renting and business activities
  - L - Public administration and defence; compulsory social security
  - M – Education
  - N - Health and social work
  - O - Other community, social and personal service activities
  - P - Private households with employed persons
  - Q - Extra-territorial organisations and bodies
- Occupational classification is the same in RUDD and NTS
- RUDD sector classification fits within SIC in NTS

# NTS data: Description

## Trip: Ticket types

- Ticket type is a stage variable in NTS
  - 19 ticket types are available
  - **Can be approximately mapped to RUDD data**
  - Class not given
- Ordinary adult
  - Ordinary child
  - Reduced ordinary adult
  - Reduced ordinary child
  - Special category reduced
  - Other (including free)
  - Season ticket
  - Travel card
  - Combined season/travel card
  - Railcard
  - Concessionary - Employees
  - Other non concessionary
  - OAP pass
  - Scholars pass
  - Disabled persons pass
  - Subsidised travel tokens
  - Other concessionary

# NTS data: Description

Trip: Propose to maintain segmentation by purpose

- Purpose is an important NTS insight
  - ◆ And NTS is limited in ticket type info (class is missing)
  
- At least business, commute, other
  - ◆ Not consistent with ticket sales data which is by ticket type
  
- Important to maintain purpose insights, but can produce ticket-type-segmented models if needed for compatibility
  - ◆ Or ticket-type-segmented elasticities
  
- Possibly segment by flow type as well
  - ◆ If significant
  - ◆ If resources permit

# NTS data: Fitness for purpose

## Modelling possibilities using NTS

- The objective is to model rail demand in NTS
  - ◆ To confirm, challenge or enhance ticket sales data
  - ◆ To investigate effects, primarily socio-economic, that cannot be addressed in ticket sales data
  - ◆ With aim of enhancing operational models
- NTS data can be modelled in aggregate or disaggregate
  - ◆ Differences are not fundamental
  - ◆ Aggregate quicker to run, but may require more resources to set-up
  - ◆ Disaggregate quicker to set up, particularly with many variables
  - ◆ Disaggregate retains more variance (no averaging)
  - ◆ Prefer disaggregate but useful to run limited comparisons
- Basic idea is to model either expected trips or probability of specific number of trips as a function of data



# NTS data: Extracting Information

What modelling possibilities are there?

- Taking account of the nature of the data:
  1. Models of expected numbers of trips
    - ◆ As close as possible to ticket sales data models
    - ◆ Like Wardman (2006), maybe with improvements
      - e.g. Daly, Sanko & Wardman 2015, non-linear functions
  2. Simple 2-stage extension
    - ◆ First train users, then number of trips
  3. Model NTS as count data
    - ◆ Recognising that people make 0, 1, 2.. rail trips in a week
- **NOT** a multi-modal model or a household model
  - ◆ Because of resource constraints (also refer to other models)

# NTS data: Extracting Information

We recommend a choice model

Leigh|Fisher



UNIVERSITY OF LEEDS

- Choice between no rail trips and some rail trips, then conditional choice of how many
  - ◆ Repeated binary logit
- Facilitates separation of two key issues
  - ◆ Who are the rail users?
  - ◆ How many trips do rail users make?
  - ◆ We do not know which of these drives growth
- Separate models for separate purposes
- This model form has been used in several studies
  - ◆ So we know it is feasible
  - ◆ And we have staff expertise and efficient software
  - ◆ Resources can be focussed on segmentation, hypothesis tests and elasticities
- Output comes in two forms
  - ◆ Qualitative identification of key features of the market
    - with appropriate significance tests
  - ◆ Quantitative elasticity measures
    - for any variables found to be significant
    - for comparison with ticket sales models
    - and potential addition to those models

# Enhanced (PDFH) Rail Demand Models With NTS Data and Analysis

Leigh|Fisher



  
UNIVERSITY OF LEEDS

Three possibilities to enrich forecasting performance of PDFH forecasting using NTS:

- Improve historic independent variable evidence (a constant 'frustration')
- Provide demand parameters (e.g. elasticities) for use in modelling (and forecasting)

*Fresh Insights*

*Evidence to overcome problems estimating some effects on rail demand*

- Better understand trends (especially hypothesised effects)

# Improving The Independent Variables

- A common concern is the availability and quality of relevant independent variables.
- Classic example is car journey time
  - Several studies recognise poor quality data used*
  - And lots of very poor car journey time cross-elasticity estimates*
- NTS has respondent reported journey time data over many years (car speed is a more robust measure, given changing trip lengths over time)
- If car times (speed) have been changing, it should be detectable
- Develop a model of reported car journey times as a function of:
  - Time of travel*
  - Distance*
  - Year*
  - Area*
  - Etc ...*
- Use the model to enhance the car journey times in ticket sales models
- Explore whether NTS can provide other such insights

# Provide Demand Parameters

- We have demonstrated the ‘limitations’ of freely estimating the set of PDFH external factor parameters
- Note that the current PDFH forecasting recommendations contain:
  - A parameter denoting the effects of non-car ownership*
  - Efforts to estimate this important effect in rail models have not met with great success*
  - The parameter used in PDFH was estimated on **NTS data** (independent of income)*
  - The parameter is then used with historic data on car ownership to account for this effect*
- This study would explore how rail trip rates vary with:
  - Socio-economic, demographic factors and locational factors*
  - Income (even though PDFH provides this)*
  - Possibly cohort effects – are young people now more likely to use train than 15 years ago*
- Same procedure followed as for car ownership
  - Include the relevant historic series and constrain the parameter (although attempt free estimation)*
- NTS can provide qualitative (e.g. “company cars are relevant for rail business demand only”) as well as quantitative insights

# Mopping Up

- In some cases, we will not have historic data to exploit NTS based parameters
- For example, on local employment trends apparent in NTS
- We should use such NTS insights to help account for differences between actual and expected demand changes
- Modelling might 'remove' such effects



# Phase 1 Summary

Leigh | Fisher



# Phase 1 Summary

A substantial demand “gap” has opened up in recent years

- Phase 1 has highlighted that there is a demand driver “gap” in past 10 years.
- Studies reviewed suggest need to include a wider range of factors in model estimation
- Need to make use of evidence from other studies / sources where possible
- Findings from Phase 1 point towards the need for
  - ◆ More variables, better data, improved model specification



# Phase 1 Summary



## Suggested wider influences on rail demand and revenue (and potential data sources)

Structural And Land Use	<ul style="list-style-type: none"> <li>Structural changes in employment market: due to centralisation effects and other trends in the labour market, white collar employment increasingly dominated by regional CBD, well served by rail with often congested roads.</li> <li>Increasing reliance for (good) jobs on the regional centres where the propensity to travel to work by rail is greater, fuelling recent strong growth in rail commuting to regional centres</li> </ul>	NTS RUDD
	<ul style="list-style-type: none"> <li>A similar trend in the shopping and entertainment sectors. Strengthening regional centres and the corresponding decline of local opportunities again benefits rail - not merely switching destinations but creating new trips</li> </ul>	This study
	<ul style="list-style-type: none"> <li>The digital revolution – rail a more attractive mode as it facilitates productive and worthwhile use of travel time through new technology in a way other modes do not – stimulating rail demand</li> </ul>	Hensher VoT study
Demographic	<ul style="list-style-type: none"> <li>Demographic changes and cohort effects, such as the retired now being far more likely to travel than in the past, students used to train travel but cannot afford a car, and higher participation rates in the labour market by females who tend to be more predisposed to rail</li> </ul>	NTS
Car Costs	<ul style="list-style-type: none"> <li>Increases in car park costs and reductions in spaces</li> <li>Recent trends towards more business travel by train because of less favourable tax treatment of company car</li> </ul>	N/A NTS
Environmental	<ul style="list-style-type: none"> <li>Environmental and sustainability considerations and viewpoints and green travel plans leading to more use of rail</li> </ul>	?
Endogenous Effects	<ul style="list-style-type: none"> <li>Revenue protection / rolling stock and station quality / marketing / information access / smart ticketing</li> </ul>	TOCs

# Phase 1 Summary

## Better Data and Model Specification

### **Better Data**

More variables requires better data. Specific need for better and more systematic data collection:

- Car journey time and car parking data
- More locally relevant data for employment (if it exists) and socio-economic data for non-commuting trips
- Improved data on diversion factors to deduce cross elasticities
- Exploit NTS data

### **Model Specification**

- More market segmentation – flow type, ticket type, journey purpose
- Allow elasticities to vary over time (might expect to be increasing)
- Use theoretical considerations more (eg, GVA origin or destination, population and employment relationships in commuting market)
- Outside evidence to enhance forecasting framework as well as modelling
- Consideration of directionality
- Consideration of non-price competition between products
- Cross elasticities linked to local circumstances (Oxford-London v Ipswich-London)
- In long run, potential development of 'rival' trip rate, mode choice etc model



# Phase 2 Recommendations

Leigh | Fisher



# Phase 2 Recommendations

## Ticket sales models

- Re-estimate OXERA/ARUP models with newer data AND constrained parameters
- Compare different input variables (e.g. for income) and different model specifications
- Consider theoretical relationships
- Add new variables (suggested supplementary exogenous / mode choice variables shown below):
  - ◆ **Population by age band**
  - ◆ **Employment by type of occupation (SOC)** or possibly industry
  - ◆ More detail on **car ownership** (e.g. cars per licence)
  - ◆ **Car time/speed**
  - ◆ **Bus cost/speed**
- Add other endogenous variables
- Within project resource, refine cross-elasticity constraints to be route specific (deduced with Dargay (2010) diversion factors and NTS shares)
- Improved segmentation (e.g., extend locational hierarchies, distance, temporal, flow size)
- Improved allowance for non-price competition for full and reduced tickets
- Use NTS to obtain car speed (Car speed model). Also other modes/variables?
- Use evidence on values of time, Hensher parameters to proxy for digital revolution

# Phase 2 Recommendations

## Underlying data

GDP *per Capita*

Population

We plan to use GVA, GDI (each) and Population data at the District level. We expect to use the age bands included in RUDD. We will need to collect population estimates for 2013/14 (from NOMIS) to complete the data in RUDD.

Employment

We prefer to use Employment data at District level. We would expect to include the occupational divisions in RUDD as this may be useful for explaining changes in rail demand. However, the data in RUDD are not consistent with what we have the APS/LFS data in NOMIS, total employment numbers are not in RUDD prior to 2005, and the RUDD data includes zeros for occupational divisions with imprecise estimates which are omitted from LFS output. **Thus**, we will download employment data ourselves, using district totals and county occupation splits. We will adjust the LFS numbers prior to 2005 to match APS in the same year.

# Phase 2 Recommendations

## Underlying data

### Non-car ownership

RUDD contains only modelled car ownership from RUDD. We can extract licence-holding at a UA level from NTS, and also household levels of car ownership (taking moving averages if appropriate to ensure the series are smooth)

Licences per car or adults per car can be extracted from NTS which would be important in assessing the impact of second/third cars.

### Fuel (car) cost

We have the speed-cost curves from RUDD, although the implied movements in efficiency are not consistent with the pump price data. We can use car journey times (average speeds per mile) from NTS to estimate changes in car cost.

It would be best, though, to ensure that the fuel price component of the cost curves is consistent with the pump price data, which is possible if we had the underlying speed-fuel curves.

### Car time

We can calculate changes in average car speed from NTS for different UA-UA, county-county or region-region trips, as appropriate (more aggregate as the samples decline in size).

Obviously this cannot capture the selection effect (if rail use has increased *when/where* car time has increased) but this is the best available data.

# Phase 2 Recommendations

## Underlying data

Bus cost  
Bus time (speed)

NTS data contains bus stage distance, journey time and fare. We should be able to use this to derive bus cost indices for regions for Short trips (e.g. <10 miles), medium distance (e.g. 10-50 miles) and coach (e.g. 50+ miles) fares, which would incorporate the reduction in average bus fares when free travel was introduced. We can do the same for bus journey time (1/speed) indices.

Bus headway

There are no historical public data on bus headway.

Air competition

We are not planning to study air competition.

Other rail factors

We are not planning to include crowding or marketing as suitable data are not available – the former would have severe endogeneity problems too.

It may be possible to identify instances of major improvements at the top few stations (managed by Network Rail) and instances of major new fleets being introduced (as service codes are included and can be mapped to TOCs) to allow us to test the sensitivity of the analysis to such events. We will aim to obtain revenue protection / gating initiatives for key stations.



# Phase 2 Recommendations

## Other data issues

### Directionality

On long distance flows (where advance tickets are dominant) and for season tickets, Ticket origin and destination may not reveal much about direction. We can test a dataset containing both 'bidirectional' and 'unidirectional' data, although deciding how to use exogenous factors for origin and destination is more difficult in those cases.

### Price deflators

We expect to use the GDP deflator (CPI based) for income/output variables and CPI for other variables. However, fares elasticities would come from another study and we would expect to use consistent price deflators.

### PTE infills

We are awaiting PTE infills which are not in the RUDD data. They are however available for only part of the dataset; we will need to test the sensitivity of our findings to the inclusion of PTE infills.



# Phase 2 Recommendations

## Use of NTS data

Exploit NTS data to improve understanding. This could take the following form in the short run:

- ◆ Compare NTS Trends with ticket sales trends
- ◆ Examine rail commuting mode choice propensity and how this might explain rail demand trends
- ◆ Use NTS to provide evidence on journey time variations (where current evidence is weak) and also other possible 'independent variable' evidence that might be used to enhance rail models
- ◆ Determine whether NTS evidence is long run (or more likely a mix of long and short) and identify whether independent of effects already in rail demand models
- ◆ Use NTS to obtain independent income elasticities by purpose/distance etc and also 'elasticities' for other socio-economic categories
- ◆ Use NTS to identify what changes to socio-economic, demographic and locational factors might have led to strong rail demand growth
- ◆ Enhance rail demand models with evidence on how rail trip rates have varied – so add in extra terms in our estimated models with, say, change in company cars along with a parameter that indicates how company cars impact on rail demand

# Phase 2 Recommendations

## Use of NTS data

This section outlines ideas on how evidence from modelling of NTS data might enhance or indeed challenge models based on the analysis of ticket sales data

The functions here set out are not intended to be comprehensive nor indeed the best ones that could be developed. This is preliminary thinking to set out some ideas.

Starting with income ( $Y$ ) where we actually have evidence, the usual constant elasticity model is:

$$V = Y^\alpha \quad (1)$$

Where  $V$  is the change in volume of rail demand between two stations in some year and  $\alpha$  is the income elasticity.

NTS can provide an estimate of  $\alpha$ , both cross-sectional and / or longitudinal. This might be useful if, amongst other things, we believe the income elasticity from ticket sales models is discerning other effects. **So the model would constrain  $\alpha$  to be the NTS provided value.**

# Phase 2 Recommendations

## Use of NTS data

NTS should be able to provide more detail on the income elasticity. Suppose the income elasticity differs between blue and white collar workers at the origin. We might specify a model of the form:

$$V = Y^{\alpha + \beta W} \quad (2)$$

where  $W$  is the proportion of white collar workers and  $\beta$  is the incremental effect on the income elasticity. We have two possibilities given how reliably we can estimate  $\alpha$ :

- constrain both  $\alpha$  and  $\beta$  to be the NTS derived values
- constrain  $\beta$  to be the NTS derived value and freely estimate  $\alpha$

Where there is sole reliance on NTS data, equation 2 could be reformulated as:

$$V = Y^{\gamma W + \delta B} \quad (3)$$

Where  $B$  is the proportion of blue collar workers and here  $\gamma$  and  $\delta$  are income elasticities for these two groups.

# Phase 2 Recommendations

## Use of NTS data

We might go a step further, and argue that NTS provides accurate estimates of both  $\alpha$  and  $\beta$  up to a scale transformation (not least to convert from cross-sectional to time-series). The model might then take the form:

$$V = Y^{\lambda(\alpha + \beta W)} \quad (4)$$

**Where  $\alpha$  and  $\beta$  are constrained to NTS evidence and  $\lambda$  is a freely estimated scale parameter.** This is analogous to a generalised cost approach where a generalised cost elasticity is estimated to a composite term constructed using evidence from elsewhere.

These are essentially interactions on the income term. But the white collar effect might be additive:

$$V = Y^\alpha e^{\beta W} \quad (5)$$

Here **NTS provides the  $\beta$**  which indicates the proportionate effect on rail demand of a unit change in the proportion of white collar workers.

There is no point here estimating a scale to the  $\beta$ . If we could do that, there would be no point for the NTS estimate. But what if the propensity to make rail trips depends on, say, three age groups, with  $A_1$ ,  $A_2$  and  $A_3$  denoting the proportion in each. If the effect is additive, then the model can simply be specified as:

$$V = Y^\alpha e^{\beta_2 A_2 + \beta_3 A_3} \quad (6)$$

# Phase 2 Recommendations

## Use of NTS data

**It would be the purpose of NTS to provide the  $\beta_2$  and  $\beta_3$ .**

We could now estimate a scale:

$$V = Y^\alpha e^{\lambda(\beta_2 A_2 + \beta_3 A_3)} \quad (7)$$

**NTS would again provide the  $\beta_2$  and  $\beta_3$ .** But now it is essentially providing the relativities.

# Quality Assurance – Phase 2

QA to incorporate the following:

- An audit of the data sets on which the econometrics have been undertaken, comparing the raw data with the final cleaned data to ensure that they match, or that differences are adequately documented.
- QA review of back-cast model.
- Replication of a sample of key estimated parameters using the same source data.
- QA of draft reports.

This will be an independent work-stream undertaken by Jacobs.