The Cost of Bus Travel Time Variability

Yaron Hollander

Submitted in accordance with the requirements
for the degree of Doctor of Philosophy (PhD)

The University of Leeds
Institute for Transport Studies

December 2006

Acknowledgments

I would like to pay special thanks to my supervisors, Prof. Peter Mackie and Dr. Ronghui Liu, for their help throughout my work on this thesis. Peter has given me succinct but very accurate guidance on thesis-writing; Ronghui has provided me with many intellectual challenges, and at the same time, proved to be a true friend when her help was needed.

While working on this thesis, I had the chance to consult with some of the leading professionals in transport analysis. The long list of people to whom I am grateful for their valuable comments includes Andrew Daly, Gerard Whelan, Richard Batley, Mark Wardman, Tony Fowkes, Peter Bonsall, James Tate, Tomer Toledo, Joseph Prashker, Mogens Fogerau, Michel Bierlaire, John Polak, Stephane Hess, Henk Van Zuylen and Dirck Van Vliet.

Mr. Andy Pike, business development manager for First Group, is greatly thanked for providing input data for the traffic experiments.

Funding my doctoral studies would not have been possible without the generous support of the Overseas Research Students award scheme (ORS); the Tetley and Lupton scholarship; a maintenance award from ITS; the Michael Golan award from the Technion – Israel Institute of Technology; and the Neil Mansfield award from the Association of European Transport.

The difficulties of spending 3 years away from home were made easier with the continuous help from my family and from many dear friends. As such, they don't need to see their names here to know what they mean to me.

Abstract

The reliability of bus systems is a vital issue on the transport agenda, since urban areas are yearning for high quality alternatives for the private car. A key indicator of reliability is a low level of day-to-day travel time variability (TTV). To obtain funds for reducing TTV, it is necessary to give evidence for the benefits from such improvement, but current tools for estimating the cost of TTV are insufficient. This thesis covers issues that arise when analysts need to show that improved bus infrastructure brings benefits from reduced TTV.

The first part of the thesis aims at understanding how the attitudes of travellers to TTV can be converted into monetary terms. The design of a survey is described, where respondents trade-off between TTV and other attributes. A modelling experiment, based on the survey responses, finds that the effect of TTV is best explained using variables that represent trip scheduling considerations. Following is a series of experiments that seek to estimate the willingness-to-pay for reduced TTV in a way that is sensitive to taste variation between travellers. Several Mixed Logit models are estimated, but some doubts about their credibility are raised, and hence the same willingness-to-pay estimates are also computed using nonparametric techniques. Some conclusions are drawn regarding the process of estimating heterogeneous willingness-to-pay and the ability to recognise the willingness-to-pay from survey data.

The starting point for the second part of the thesis is the lack of tools for estimating the level of TTV in hypothetical scenarios. We consider the case for using traffic microsimulation to estimate TTV by running a microsimulation model multiple times, and looking at the variation between runs as an estimate of the variation between different days. Such concept of estimation requires a special calibration methodology, which sets the level of simulated inter-run variability at a similar level to inter-day variability in the real network. A full calibration methodology is developed, tackling methodological, computational and statistical issues.

Finally, the demand and supply methodologies are combined, and it is illustrated how the savings from improved bus infrastructure can be examined. The contribution of the entire study includes methodological and technical insights into modelling the attitudes to TTV, estimating the distribution of the willingness-to-pay and calibrating traffic microsimulation models; but it also brings up policy issues concerning the role of TTV in transport appraisal.

Table of contents

List of tables

List of figures

Chapter 1
Introduction

1.1. Background

In recent years, the reliability of transport systems has been widely recognised as an important issue in transport planning and evaluation. Many studies have shown that the users of a transport system rate reliability as a key feature, which affects their views of the system and their frequency of using it. Reliability issues often come up in the discussion between transport researchers or practitioners, and seem to be on the agenda of transport decision makers. For this reason it is surprising that the tools available to analysts today are still in an early stage of development when it comes to the dealing with the sources and the impacts of unreliability.

Unreliability has many faces, and in the transport literature it sometimes appears under a guise. There are some studies of the *unpredictability* (e.g. Small et al, 1999) of travel conditions, namely the tendency of the state of a transport network to vary with no evident reason. Other works focus on the lack of *punctuality* (e.g. Bates et al, 1995): journeys do not take exactly the same time as expected, or when referring to public transport services, the same time as in the published timetable. There are also papers about *irregularity* (e.g. Laidler, 1999) in the traffic conditions or in public transport performance. Although there are differences between the various terms used to describe unreliability, a fundamental trait of an unreliable transport system that resides in all of them is a high level of day-to-day *travel time variability* (TTV).

If an authority wishes to obtain funds for improving the reliability of a transport system by reducing the level of TTV, it often needs to give evidence for the expected benefits from such improvement. This is equivalent to providing evidence of the costs of TTV before and after the improvement, or evidence of the difference between these costs, at the least. But the current practice in transport analysis does not include sufficient tools for estimating the cost of TTV or of a change in the extent of TTV. Not much is known about how to estimate the level of TTV (or differences in this level) in hypothetical settings, and even if such estimates exist, it is not always clear how to place a monetary value on them, as often required in scheme appraisal.

It is commonly agreed by planners and policy makers that in order to maintain the vitality and sustainability of urban areas, high quality alternatives to the private car must be provided. The ability to achieve a high level of reliability is therefore particularly important in the design of public transport systems. Of the various public transport modes available in the UK and other countries, railway services dominate the interurban travel market, but a large share of journeys in urban and metropolitan areas use bus services. Hence, the efforts to guarantee good levels of reliability for bus users deserve a thorough analytical discussion. As there has not been sufficient focus on this issue in previous studies, this thesis looks into the evaluation and the prediction of bus TTV.

Throughout this thesis, a major role is played by the general concept of stochasticity; namely, by the idea that there is an element of randomness in the behaviour of individuals, the performance of vehicles and the occurrence of phenomena. The focus on TTV is in itself a manifestation of this concept. Another expression of it is an analysis of the distribution of the willingness-to-pay, which is based on the understanding that we cannot truthfully capture the pattern of response to changes in the transport system if we assume that travellers have uniform preferences. The stochastic approach is also apparent in our decision to model the incidence of TTV using traffic microsimulation, which has the ideas of randomness and heterogeneity at its core.

At first glance, the different topics discussed in this thesis might not seem strongly related to each other. One part of the thesis involves demand modelling and analysis of consumer behaviour from an econometric perspective, while another part examines the way we use a supply model of a transport network, from the perspective of a traffic modeller. Some of the chapters concentrate on the principles of surveying or modelling, whereas other chapters investigate very technical issues. The common thread in all the ideas discussed here is that they are all needed in order to present evidence for the benefits from reduced variation in bus journey times. A modest attempt to bring these separate discussions together in an attempt to address a practical problem is presented towards the end of the thesis.

1.2. Definition of travel time variability

There are various ways to define TTV, and it is important to explain at this stage which of them is followed throughout this study. Three main types of TTV can be identified in

the transport literature, as described in the following paragraphs. The definition of the three types is partially based on terminology presented by Bates et al (1987).

Inter-vehicle variability is the variability between journey times experienced by different vehicles making similar journeys at the same time. It is attributed to waiting times at signals, conflicts with pedestrians, differences in driving style and so on. In the current context it should be noted that inter-vehicle TTV is not a powerful indicator of the type of reliability that travellers are likely to be concerned about, which has to with the unpredictable nature of their own travel experience, and not with the difference between them and other travellers.

Inter-period variability (or *within-day variability*) is the variability between the travel times of vehicles making similar journeys at different times on the same day. It is mainly caused by differences in the level of demand, occurrence of accidents and incidents, weather conditions, the level of daylight and so on. We find that this type of TTV is most relevant for examining various policy-related measures, such as flexible working hours or congestion pricing, but not particularly for reliability analysis.

In this thesis we focus on TTV between similar journeys made on different days (*inter-day variability or day-to-day variability*). It is caused by fluctuations in travel demand, variation in driving behaviour, changes in the amount of roadside activity, weather conditions, accidents and incidents and other reasons. In the case discussed here we limit the definition even further: we are not interested in the elements of the day-to-day variability that rational travellers can anticipate, such as the differences between summer and winter, weekend and weekdays, or irregular travel times experienced due to special events. The main interest here is in TTV that remains unexplained after variations due to these predictable elements have been subtracted. This variability is random in nature; its various causes are too subtle or too complex to be expected by travellers.

It is important to note that the motivation for focusing on the abovementioned definition of TTV is not related to whether or not this type of TTV is easy to analyse or to model. Various ways to compromise with technical and computational difficulties are discussed later in the thesis, but the definition itself is uncompromised. The reason why we exclude the effects of predictable events is the belief that this form of TTV is the most accurate in the way it represents unreliability. A very similar definition of TTV is used, for similar reasons, by Fowkes and Watson (1989) and Bates et al. (2001).

Both Bates et al (1987) and Noland and Polak (2002) note that most research about TTV focuses on day-to-day variability. They do not clarify whether or not existing research investigates only the random element of day-to-day TTV; this might cause some confusion, because not all studies clearly mention how they define TTV. It is common to base some of the analysis of TTV on surveys where travellers are asked to make various choices, responding to different levels of TTV; the actual definition of the investigated TTV is then determined not by what the authors state but by what the respondents understand. A detailed definition of TTV is not normally presented in these questionnaires, and it is therefore likely that respondents refer to TTV in its most intuitive meaning. Judging by the way TTV is commonly presented (see also the review in chapter 2), we find that most studies, similar to this thesis, tend to focus on random TTV, whether or not they explicitly mention it.

TTV is most commonly measured as the standard deviation of travel times (namely, of the set of travel time measurements that exhibit TTV as defined in the respective study). This applies to all TTV variables mentioned in the economic review and experiments, presented in the following chapters (unless mentioned otherwise). For the traffic experiments presented later, we develop another measure that meets the particular needs of the proposed methodology.

The definition of TTV has implications also for the way we define the daily periods to be used in the forthcoming analysis. As mentioned above, we wish to distinguish between predictable and random components of TTV. Transport analysts commonly specify periods not shorter than two hours each; for instance, the morning peak period is typically defined as two or three hours long. However, it is clear that the differences between travel conditions in different parts of the entire morning peak are not completely unpredictable; this means that we cannot attribute a certain level of TTV to a whole period because it would contradict our definition of TTV. The analysis in this thesis therefore refers to much shorter periods; we define a period as a time interval which is short enough to be seen by travellers as having uniform travel conditions. Fifteen-minute long periods were used in a similar context by May and Montgomery (1984), and we suggest that any duration up to thirty minutes would be plausible.

1.3. Evidence of travel time variability

Before going further into the analysis of TTV, it is useful to show evidence of what bus TTV is like in an actual setting. Figures 1.1 and 1.2 present real travel time data from bus route 4 in the city of York, on various days in October and November 2004. The horizontal axes represent stops along the route and the vertical axes stand for time from the scheduled departure. Each of the figures contains three graphs, each of whom presents trajectories of buses departing at a single time on multiple days. If buses adhered to their planned schedule, all the trajectories in the same graph would be identical. The extent of variation *between trajectories* in each graph manifests our definition of TTV. Differences *between graphs*, i.e. between different levels of TTV, are what we later try to predict and evaluate.

Passenger demand on the buses going eastbound (figure 1.1) is relatively high (about 450 passenger boardings per hour, with 8-minute headway), and the general traffic along parts of the route is relatively congested. The opposite direction (in figure 1.2) has lower passenger demand (about 150 passengers per hour) and lighter surrounding traffic. Table 1.1 presents simple analysis of the variation in arrival times at a bus stop, based on the same data as in the figures. The standard deviation of arrival times is averaged between all stops, and there is also indication of the stops with the highest and lowest standard deviation. The location of a stop is indicated as its serial number in the sequence of stops along the route. From the figures and the table we can come to some basic insights into the nature of bus TTV.

Departure	Mean S.D.	Most *reliable* stop		Most *unreliable* stop	
		S.D.	Location	S.D.	Location
07:32 eastbound	04:33	02:31	1st of 23	06:19	2nd of 23
08:20 eastbound	02:58	01:04	1st of 23	04:05	19th of 23
09:08 eastbound	02:54	01:48	3rd of 23	03:21	11th of 23
07:38 westbound	01:56	01:03	35th of 35	02:59	9th of 35
08:26 westbound	02:40	01:51	14th of 35	03:07	4th of 35
09:14 westbound	03:23	02:15	13th of 35	04:37	34th of 35

Table 1.1: Standard deviation of the arrival time at a stop

(as observed on route 4 in York, 2004)

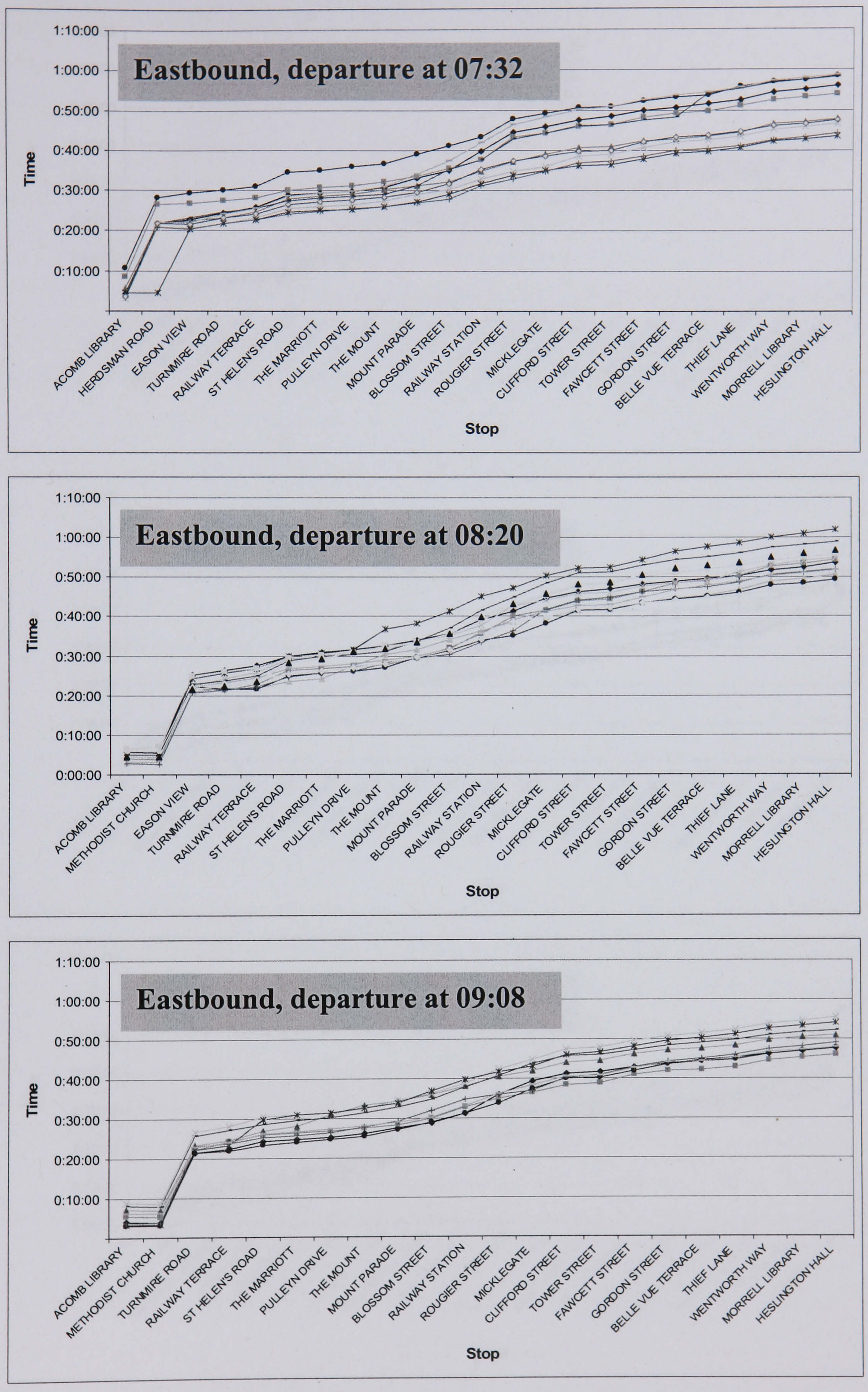
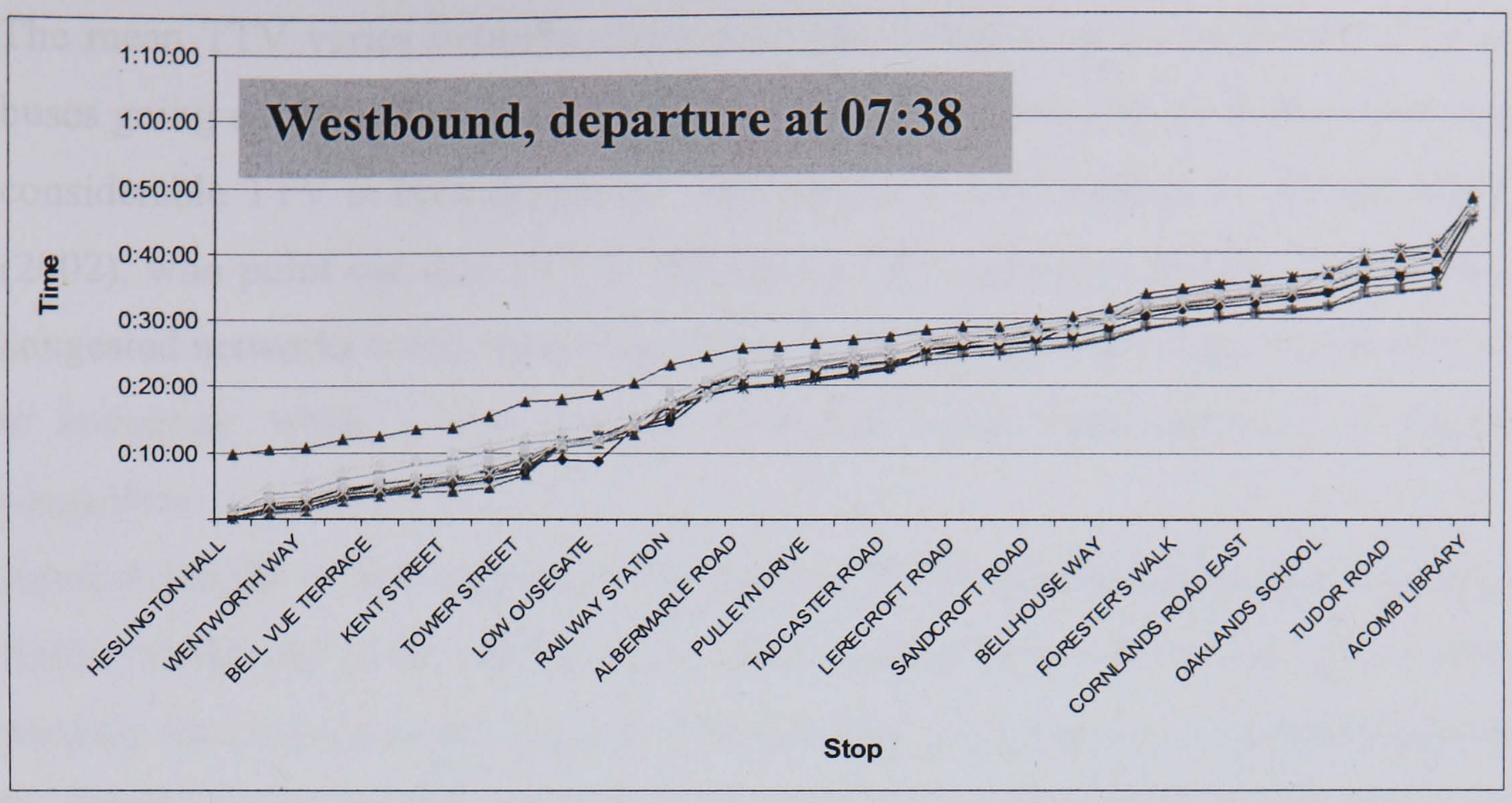
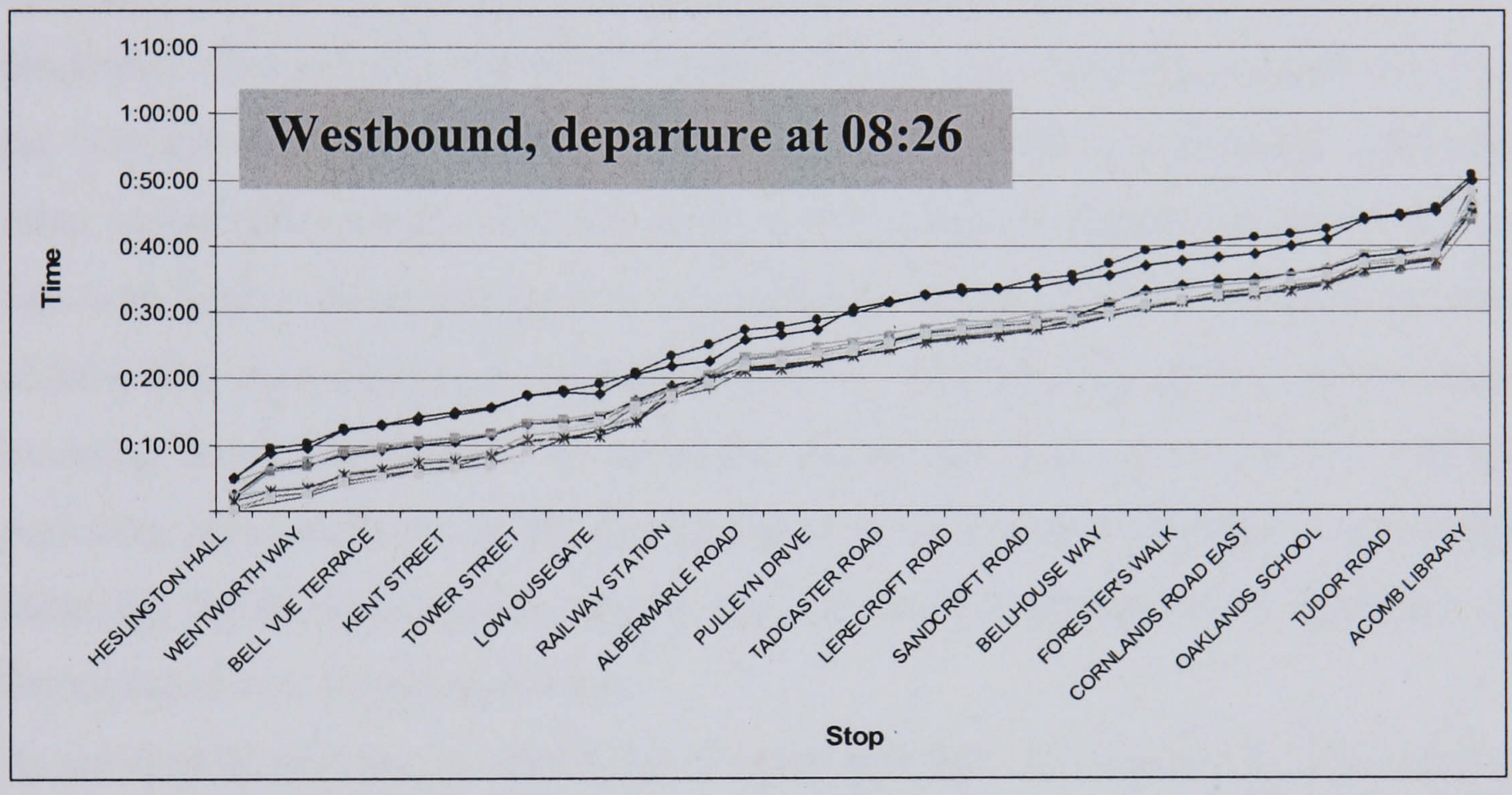


Figure 1.1: TTV on route 4 in York - eastbound





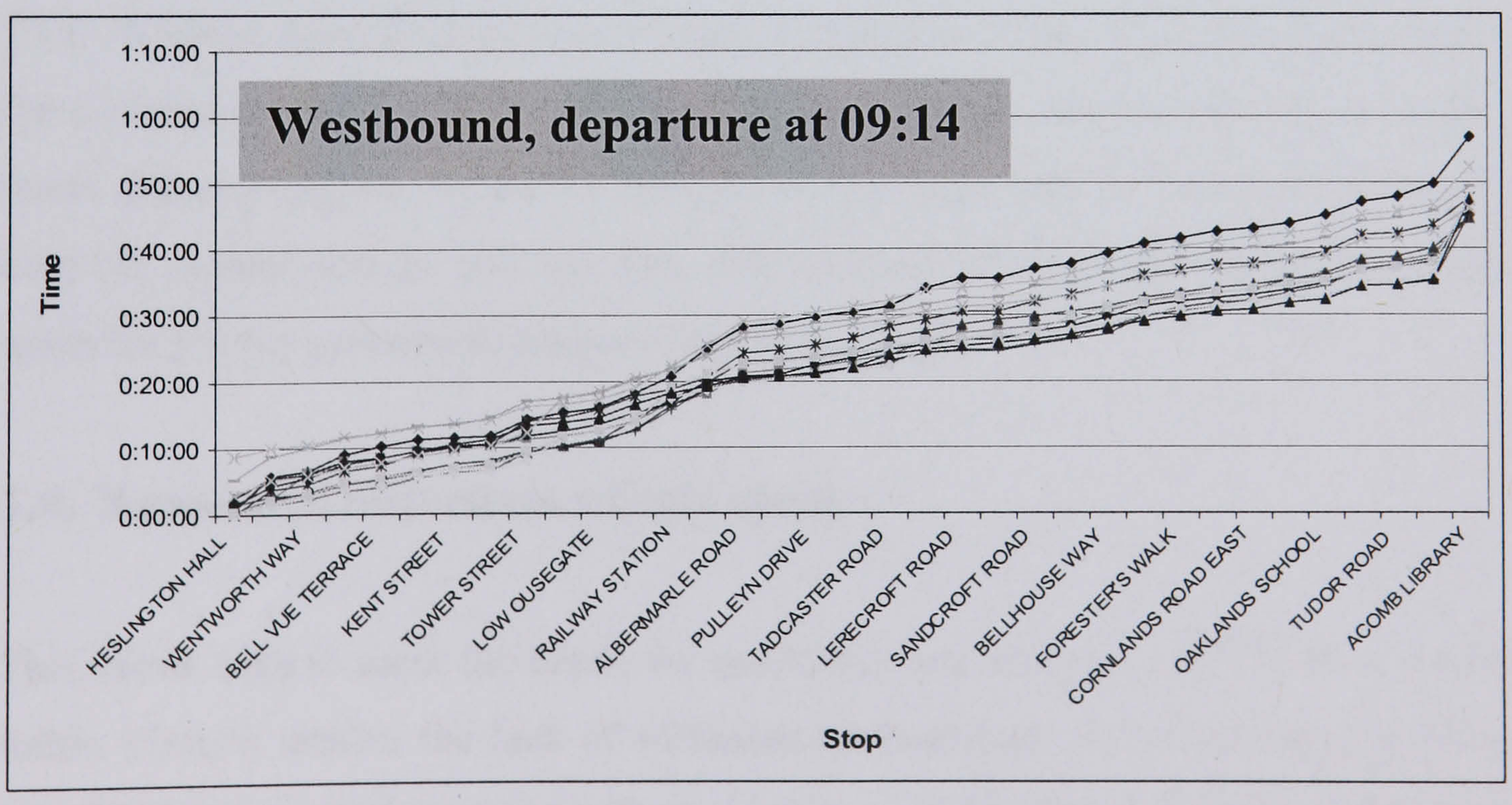


Figure 1.2: TTV on route 4 in York - westbound

The mean TTV varies between departures and between traffic directions. TTV of the buses going eastbound is generally higher than in the other direction. However, there is considerable TTV in both directions. This relates to a discussion by Noland and Polak (2002), who point out that TTV is independent of congestion effects. That is, in some congested networks travel times can still be consistent, and hence not necessarily harder to anticipate, while a high level of TTV can occur when there in no significant congestion.

Another insight is that sometimes the level of TTV tends to increase as the bus goes further along the route, but in some other cases a high TTV at an early section is partially recovered towards the end of the journey. Such recovery is mainly experienced by buses going westbound; but a gradually increasing TTV is apparent in both directions. This possibly has to do not only with the level of traffic congestion, but with the cumulative effect of a delay at one point on the number of passengers boarding at other points downstream. Since in these data the more congested direction is also the one with higher passenger load, it is not certain which of these features has stronger effects. The irregularities in the extent of TTV, and their secondary consequences on boarding times downstream, are probably among the main causes of the well-known *bunching* phenomenon, i.e. the uneven pattern of bus arrival despite an even schedule (although the presented graphs do not show successive departures and therefore do not demonstrate how bunching occurs).

As mentioned, the data presented above is brought here as evidence for the existence of TTV. While it does illustrate some essential features of bus TTV, it mainly proves that the occurrence of TTV does not follow simple patterns, and is not easy to track. This thesis does not include extensive analysis of the causes for TTV, but the network used here for demonstration, and the data collected on the same bus route, are employed again for a more systematic analysis later in the thesis.

1.4. Scope and objectives of this study

This thesis tries to meet the needs for modelling and analysis of TTV in two different fields. First, it tackles the lack of sufficient methodology for converting the effects of bus TTV into monetary terms. Then, it tries to deal with the shortage of tools for estimating the level of TTV in hypothetical scenarios. These form two separate

discussions; the rationale behind this duality is that the two problems are in fact elements of the broader challenge of revealing the benefits associated with improved bus reliability. An additional discussion, that follows the two mentioned parts, tries to combine their findings in a joint experiment.

The objectives of the overall study are as follows:

1. To develop methodology for expressing the impacts of bus TTV in equivalent monetary units, such that they can be used in the appraisal of bus schemes in the study area; and while doing so, to account for the idea that the population in the study area is heterogeneous, with varying tastes and preferences.
2. To develop ability to estimate the expected level of TTV in the study area in hypothetical scenarios, in a way that is sensitive to local factors such as the detailed configuration of the transport network; and while doing so, to account for the idea that traffic phenomena, as well as the vehicles and their users, exhibit a high level of heterogeneity and randomness in their performance.
3. To illustrate the application of our solutions for the two aforementioned problems in a practical case study.

The study area in all parts of this thesis is York, a medium-sized city of 181,000 inhabitants (according to the 2001 census), located in the Northern England county of North Yorkshire. The entire analysis focuses on the commuter population in York and on the morning commuting journey; this includes any journey for work or education purposes, as long as it is made on most working days during the morning peak. The traffic analysis is mainly based on data from bus route 4 in York, provided by its operator. More details of this route and of the network used are given later in the thesis.

Setting the scope of this study also requires elucidating which related issues are not covered. During the preparation of this thesis, questions about the reasons why travellers dislike TTV, and possible differences in this matter between the users of different transport modes, repeatedly came up. For instance, it was theorised that one element of the discomfort attributed to TTV has to do with stress from the unsettled travel experience itself, while a separate element has to do with the late or early arrival at the destination, which is inevitable if travel times are unpredictable. There were also doubts about whether to expect greater sensitivity to TTV among bus users, compared to what is known from other studies about car users, because they are less capable of

changing their route once the journey has begun, or smaller sensitivity, because much of the stress caused by TTV would have to do with the task of driving. Such questions arose in such contexts as choosing survey wording, deciding on model variables, or judging whether modelling results are sound. But although the informal discussion of these issues was an integral part of the modelling experience presented later, it should be stressed that seeking proper answers to these questions in the current scope would be too ambitious. We do not formally attempt to understand the broad psychological background for the aversion that travellers feel towards TTV, only to *model* this aversion. A model is unavoidably (and to some extent also desirably) simpler than the real-world phenomenon it describes, and as other researchers do, we find that a model of the impact of TTV can be adequate even if it leaves many questions unanswered.

The same disclaimer applies to the analysis of the reasons why travel times fluctuate at all. The role of various factors in the evolution of TTV is occasionally brought up throughout the thesis, but in our attempts to create a tool for estimating the level of TTV, no major attention is given to uncovering the full set of causes of TTV.

Another broad area that the thesis is unable to encompass is the variety of indicators of unreliability of public transport systems. Analysis of TTV is very central to the discussion of unreliability, but other important indicators are used too in many studies. There are also works about TTV that measure it using other statistics but the standard deviation or the variance of travel times. Alternative measures for TTV are mentioned in the following chapters, but the whole range was not possible to cover in detail.

1.5. Outline of this thesis

As mentioned earlier, the main body of the thesis comprises two parts. The econometric part (chapters 2 to 4) describes experiments in modelling the effect of TTV on bus users, and the traffic part (chapters 5 to 7) explains issues that relate to the estimation of TTV using traffic microsimulation. Figure 1.3 depicts the structure of the thesis.

Chapter 1 describes the background and the scope of the study.

Chapter 2 reviews some relevant topics from the literature in transport economics. These relate to the appraisal of transport schemes, the role of TTV in it, and the design of surveys that focus on TTV.

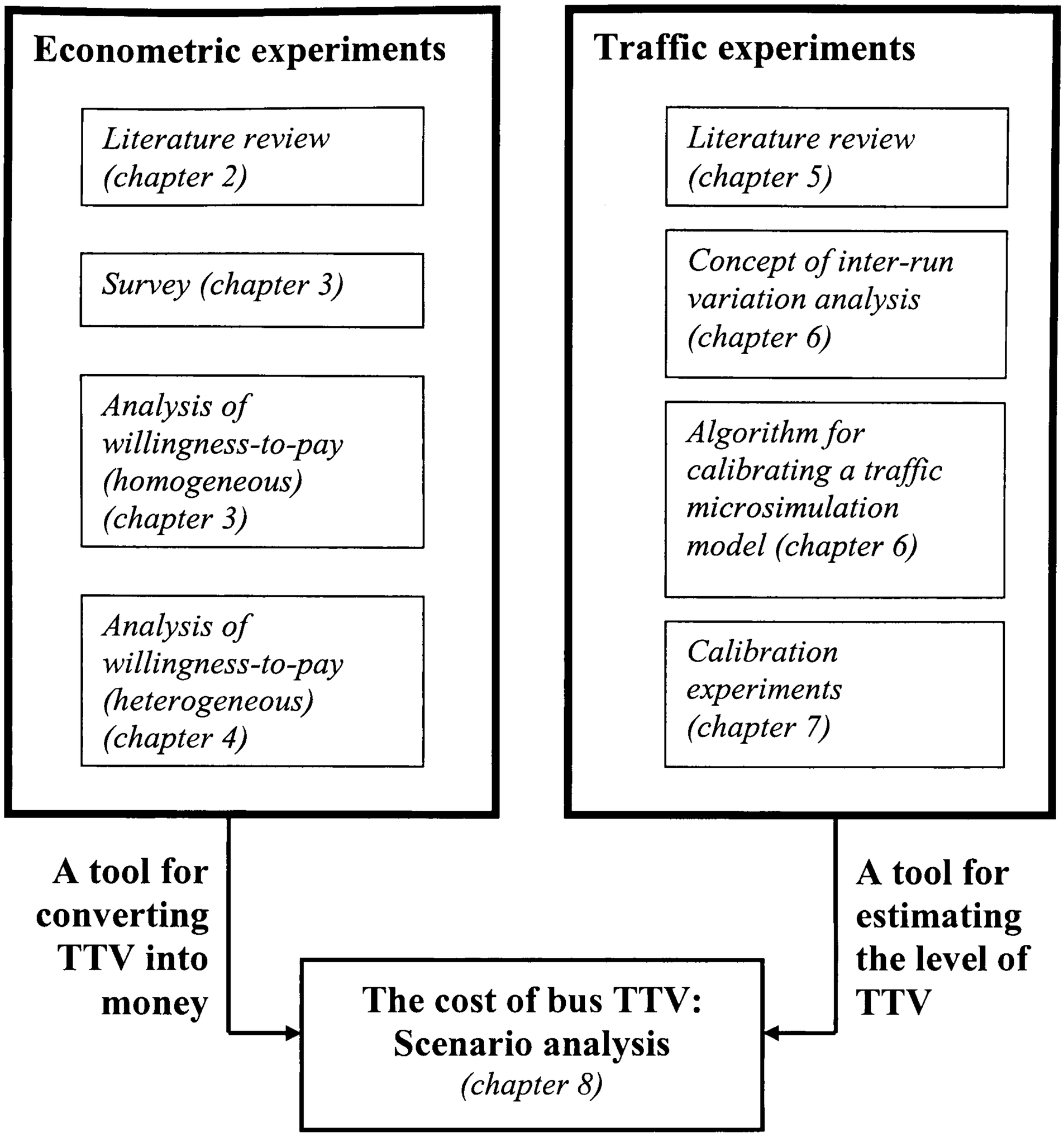


Figure 1.3: Outline of the thesis

Chapter 3 describes the design of a survey, where the respondents are asked to trade-off between TTV and other attributes. Based on the data collected in the survey, this chapter also presents models that account for the attitudes towards TTV either directly or indirectly, and suggests uniformly-distributed estimates of the willingness-to-pay for reduced TTV.

Chapter 4 extends the estimates of the willingness-to-pay obtained in chapter 3 by allowing for taste variation between travellers. Several Mixed Logit models are estimated, but due to some doubts about the credibility of the distribution of the willingness-to-pay implied by these models, attempts are made to derive alternative estimates using two different nonparametric techniques.

Chapter 5 begins a new section of the thesis with a review of existing approaches for the estimation of TTV, and as a separate topic, a thorough review of methodologies for calibration of traffic microsimulation models.

Chapter 6 presents a new approach for the estimation of TTV, based on analysis of the variation between the outputs of different runs of a traffic microsimulation model, such that each run represents a single day. The chapter shows that in order to be able to perform such analysis, the model needs to go through a special calibration process; full calibration methodology, including a solution algorithm, is proposed.

Chapter 7 describes two experiments that implement the approach developed in chapter 6. The first experiment uses artificial data, and is aimed at testing the validity of this approach. The second experiment applies the calibration algorithm with real data and a full-size network, and results in a model adjusted for the study area.

Chapter 8 combines the demand and supply methodologies developed earlier in the thesis. Different scenarios of bus infrastructure schemes are specified, and the costs associated with TTV are compared.

Chapter 9 includes conclusions and suggestions for further research.

1.6. Notation

The following abbreviations are used throughout the thesis:

TTV	travel time variability	MTT	mean travel time
TMM	traffic microsimulation model	CBA	cost-benefit analysis
WTP	willingness-to-pay	MXL	Mixed Logit
DWP	distribution of the willingness-to-pay	SUS	sub-sampling
DTC	departure time choice	ppm	pence per minute
VOT	value of the mean travel time		
ME	mean earliness		
VOE	value of the mean earliness		
ML	mean lateness		
VOL	value of the mean lateness		
MTE	mean travel time and earliness		
VOTE	values of the mean travel time and earliness		

Chapter 2

Travel time variability in the economic literature

2.1. Introduction

This chapter commences a part of this thesis that seeks to determine how to convert a given level of bus TTV into monetary terms, which can be used in an appraisal context. The search for such method includes the design of a survey, followed by a series of modelling experiments. Other works that share their objectives, or the techniques they use, with this study have been undertaken in the last decade, and therefore this chapter precedes the surveying and modelling effort with a review of these previous works.

The review has two main parts. The first part discusses the general idea of appraisal and the willingness-to-pay (WTP), the role of TTV in the appraisal practice, and different approaches for estimating of the cost of TTV. The second part presents some basic features of the type of survey we carry out later, and examines in particular other studies that tackled the challenge of designing a survey in which TTV is a key attribute. The conclusions from this review are to be applied in chapter 3.

2.2. Appraisal of transport schemes

The construction of new transport infrastructure, or the improvement of existing infrastructure, has effects in various fields of modern life. It can cause immediate changes in the living standards of any individual and in the level of prosperity of any business. In the longer run, changes in the transport system may have economic, social, environmental and other consequences, either positive or negative, on parts of the public. The investment in transport projects may require substantial amounts of money, and the sources of finance are perpetually limited. Since this nature of transport investments was recognized, it has been common to carry out a process of appraisal prior to making investment decisions.

Transport investment appraisal always includes the identification of expected costs and benefits from different alternatives of a transport scheme. There must be more than one alternative, since there are always at least the options of "do nothing" and "do

something". Frequently there is more than one "do something" option, such as "build a road" versus "construct a railway". Since the first attempts to formulate consistent appraisal methodologies, in the 1960's and 1970's, till today, a major trend in the evolution of appraisal principles has been the expansion of the range of benefits taken into account. Whereas the main discussion in the 1970's concentrated on benefits from travel time savings and improved road safety, there is today an increasing awareness of the environmental effects and socio-economic impacts, such as equity or economic development (Mackie and Nellthorp, 2001; Grant-Muller et al, 2001). Some of the effects that appraisal techniques are expected to take into account today were traditionally seen as un-measurable (Hotchkiss, 1977).

The range of appraisal techniques has also expanded in other ways. Transport project assessment is now expected to consider systemwide effects rather than focus on the vicinity of the investment area; it is expected to deal with multimodal projects and identify benefits from different transport modes; and it also expected to analyse cost-benefit considerations in projects that are financed by complex mixtures of public and private sources (Mackie and Nellthorp, 2001; Grant-Muller et al, 2001; Nash, 1993; Banister and Berechman, 2000).

Probably the most common appraisal method is the cost-benefit analysis (CBA). The main feature of the CBA is that all expected costs and benefits are converted into monetary units and summarized; by doing so, it is possible to ascribe an overall net benefit to each project alternative. The concept of attaching monetary values to amounts of time savings or to estimates of safety improvement, and the different ways to compile all impacts into a single number (such as a discounted net present value of the proposed project), have been extensively discussed in literature. Conventionally, CBA is based on the concept of WTP, i.e. the assumption that a more desirable situation is one that people would pay more for. Therefore, the monetary equivalent of a one-minute reduction in travel time in a CBA is simply the expected amount of money that individuals would be willing to pay for this reduction. However, transport schemes also have effects that might be undervalued by the concept of WTP, because their importance is not necessarily appreciated by individual travellers; therefore, this concept is often extended by monetising environmental impacts, as well as other externalities, in other ways. The determination of prices per unit can be alternatively founded on social considerations; this can be done, for instance, by taking into account

the different levels of WTP that characterize different income sectors (Mackie and Nellthorp, 2001; Grant-Muller et al, 2001; Nash, 1993).

It is possible to appraise transport projects through a multi-criteria analysis. The multi-criteria analysis framework is more flexible than the CBA in terms of defining the appraisal objectives; multiple simultaneous objectives are allowed, which can reflect a broad range of ideologies or policy perspectives. To form a multi-criteria analysis, a set of objectives should be defined, and a set of evaluation criteria should be attached to each objective. Each objective is also given a weight, which represents its relative importance with respect to the other objectives. Based on the evaluation criteria, each project alternative is given a mark (typically on a 0-100 scale) with respect to each of the objectives. Subsequently, the marks are multiplied by the weights, and all the weighted marks of each alternative are summed to form an overall score for that alternative. Various assessment techniques may be used for marking the alternatives, either based on purely quantitative calculations or on a partially-qualitative rating. Using assessment techniques that are based on the expected market reaction, and converting values into monetary units, as done in CBA, are also possible in a multi-criteria analysis but they are not common. The freedom to choose a variety of objectives, assessment criteria and any desired weights is both a great advantage of multi-criteria analysis and a major disadvantage; it enables a wide range of policies to guide the project judgement, but also gives rise to using improperly calibrated scores and to inconsistency between decisions made by different people or at different times (Grant-Muller et al, 2001).

It is repeatedly found in appraisal studies that travel time savings are by far the most significant benefit from capital investment in transport projects (Georgi, 1973; Harrison, 1974; Hotchkiss, 1977; Glaister, 1981; Pells, 1987b; Polak, 1987a; Nash, 1993; and others). Senna (1994a) and Polak (1987a) estimate that the share of travel time savings among other benefits in a CBA framework is approximately 80%. Pells (1987b) quotes references that estimate that benefits from the reduction of travel time in road improvement projects in Britain are on average 89% of the total benefit. As mentioned above, other common sources of benefit are accident prevention and environmental effects, but in recent years there has also been an increasing interest in socio-economic impacts, influence on land use and development, and equity considerations (Nash, 1993; Banister and Berechman, 2000; Grant-Muller et al, 2001; Mackie and Nellthorp, 2001). The concept of WTP, in particular, has been implemented to estimate numerous sources

of benefit, such as the benefit from improved safety (e.g. Iraguen and Ortuzar, 2004), noise reduction (e.g. Galilea and Ortuzar, 2005), fuel efficiency (e.g. Walton et al, 2004), and more. Although the analysis of WTP is clearly not the only way of evaluating suggested investment alternatives, the current study accepts the CBA framework and the idea of WTP as a basis for project appraisal.

From the key role that time reduction benefits often have in appraisal results we can learn not only about the importance of time savings, but also about the relatively minor role that other considerations but the travel time have in the appraisal methodology itself. It should be emphasized that the outputs of any appraisal study strongly depend on a predetermined list of sought benefits; such list is part of the appraisal guidelines, which vary from place to place and from time to time. The introduction of new potential sources of benefit to this list is a slow process, even if the need for this change is widely accepted. This is mainly due to the need to develop techniques for measurement and assessment, but also because changing appraisal guidelines is a political matter that requires the consent of decision makers and authorities. For this reason, environmental and socio-economic impacts of transport schemes took years to become an integral part in CBA practice, and some of them are still being discussed theoretically without being practically used. This is also the case for the benefits associated with improved reliability, which are at the heart of this study.

2.3. Travel time variability considerations in cost-benefit analysis

A factor that still struggles for a place in the list of potential benefits from transport investment is the reduction of TTV. Statements that these benefits should have a role in project appraisal have appeared occasionally in the transport literature for a few decades. Knight (1974) notes that TTV is "a significant component in the generalised cost of trip-making". A similar opinion is expressed by Harrison (1974), Polak (1987a, 1987b), Noland and Small (1995), Small et al (1999), Bell and Cassir (2000), Grant-Muller et al (2001) and others.

The idea that the effects of TTV should be included in project appraisal has evolved along with the recognition of TTV as a factor that influences choices made by travellers. Starting from the 1970's, variables that represent TTV were introduced in several behavioural models. These models gave evidence that travellers take account of TTV

considerations when choosing a mode of transport (e.g. Prashker, 1979) or a departure time (see detailed review later in this chapter). As stated by Bates et al (1987), it has been shown that the variability in journey duration, and the uncertainty in arrival time that directly results from TTV, are major sources of irritation to travellers, and can therefore be recognized as potential sources of additional travel cost. Several studies have even concluded that a low level of TTV is more important to travellers than travel time itself or the journey cost (Golob et al, 1970; Paine et al, 1976; Bates et al, 2001). It was mentioned above that a basic feature of CBA is the conversion of insights on the preferences of travellers into monetary values; it therefore seems clear that the recognition of the impact of TTV on travel behaviour should in principle result in the inclusion of the cost of TTV in the CBA practice.

Nevertheless, TTV considerations are absent from almost any discussion of what is actually included in the appraisal framework. Historical reviews of appraisal components are brought by Georgi (1973) and Adler (1971); TTV is hardly mentioned. Pells (1987a) explicitly notes that in existing appraisal framework, "benefit is assumed to derive purely from reductions in mean travel duration, and any effect an investment might have on the distribution of travel time is ignored". Nash (1993) reviews common CBA methodologies in Britain and elsewhere, and although he elaborates on various issues that have been raised and developed since Georgi's review, TTV is not addressed. A review of the current state-of-the-art in EU countries is carried out by Grant-Muller et al (2001); it shows that benefits related to TTV are not taken into account in any of these countries.

The absence of TTV from the traditional appraisal agenda means, first of all, that investment benefits are underestimated in all projects that improve reliability. One may assume that this will always be in favour of "do less" investment alternatives. The consequences of ignoring TTV might be more complex when reduction of TTV is expected to be a major outcome of the appraised project, or even more so when several alternative schemes are compared but only one of them includes measures that reduce TTV. Appraisal results in projects that directly aim at improving reliability may be seriously distorted, since a key source of surplus is disregarded, and their likelihood of being promoted is clearly harmed. Noland and Polak (2002) mention that the projects whose benefits will be revealed, if TTV considerations are taken into account, are not the traditional transport projects but investments in improved incident removal, better management of public transport networks, or advanced information systems. There are

numerous documented CBAs of projects in which TTV reduction seems likely to be a source for considerable benefit, but this source is ignored (e.g. Schweiger and Marks, 1999; and Thijs and Lindveld, 1999).

There is some recent evidence showing an increasing awareness of the benefits from improved reliability in CBA practice. A set of guidelines for the appraisal of bus priority measures, presented by Laidler (1999), does state the significance of benefits from the reduction of TTV, but these benefits are only calculated using rules of thumb: they are assumed to equal 30% of the benefits from reducing the mean travel time (MTT) and waiting time in the peak period, and 15% in the off-peak. No attention is paid to the effects on TTV of potential differences between project alternatives or to the effect of any particular traffic conditions. SDG (2001) carried out appraisal of a quality bus corridor scheme and although benefits from reduced TTV were taken into account, it is not entirely clear what methodology was used for attaching a monetary value to TTV savings.

It therefore seems that the CBA of transport schemes today either ignores TTV considerations or handles them in a simplistic manner. Obviously, the reasons for this are not as simple as pure negligence; the exclusion of TTV benefits from the appraisal practice has surely much to do with difficulties in defining measures for TTV, difficulties in generating forecasts of TTV, and difficulties in attaching monetary values to TTV. The methodology proposed in this thesis aims to examine the feasibility of filling these gaps.

2.4. The benefits from reducing travel time variability

2.4.1. Theoretical foundations

We saw that the appraisal framework often does not account for the costs induced by TTV or the potential benefits from its reduction. In this section we look at previous attempts to quantify these costs or benefits. This sub-section reviews theoretical discussions of this issue, and the subsequent sub-sections examine models of a more empirical nature.

It was mentioned earlier that the major source of benefits in the CBA of most transport investment schemes today is travel time savings (that is, reduction of the *average* travel time). The required input for estimating the monetary benefits from such time savings in

a hypothetical scenario is a forecast of the savings in time units. Converting them to monetary units is conventionally done by multiplying by the value of time (VOT), i.e. the WTP for a reduction of one unit of travel time. The use of VOT is based on the idea that there is a trade-off between time spent travelling and money; establishing an equivalent trade-off between money and other features but the average travel time is the main challenge to be faced if we want to include these additional features in the CBA. Knight (1974) states that the difficulty in defining a suitable trade-off between TTV and money is the main reason why TTV is neglected in project appraisal. To a great extent, all attempts to evaluate the benefits from reducing TTV try to define situations where travellers can exchange money, directly or indirectly, with a lower level of TTV.

The economic significance of TTV was initially discussed theoretically by Knight (1974). Knight (1974) refers to an unpublished paper by Rees, which provides theoretical evidence for the existence of benefit from reducing TTV. It is assumed that travellers gain a surplus from travelling (the source of this surplus is further discussed later in this chapter); this surplus is assumed to be a function of trip duration. If surplus were a linear function of time, then the traveller would be indifferent to variability in trip duration, given that the same mean time is guaranteed. However, Rees assumes a convex function, i.e. an increasing slope of surplus loss as trip duration increases. Under this assumption, a traveller "would be indifferent between a trip of guaranteed duration and one of stochastic duration with a *lower* mean". In other words, for a given MTT, the surplus when there is no TTV is higher than in a case where some variability exists.

This is illustrated in figure 2.1. The vertical axis represents surplus and the horizontal axis represents trip duration. The non-linear curve is the surplus function. Let us examine an example in which trip duration is a variable with a 0.5 probability of duration t_0 and a 0.5 probability of duration t_1. The MTT is $\bar{t}$ with surplus $\bar{S} = 0.5\ (S_0+S_1)$. Now let us examine an alternative example, where a constant travel time is guaranteed. In this case, traveller's surplus equals $\bar{S}$ if travel time is t_2, where $t_2 > \bar{t}$. The introduction of variability is therefore a cause for reduced surplus. To eliminate the variation in travel time, the traveller would be willing to pay A units of surplus or B units of time.

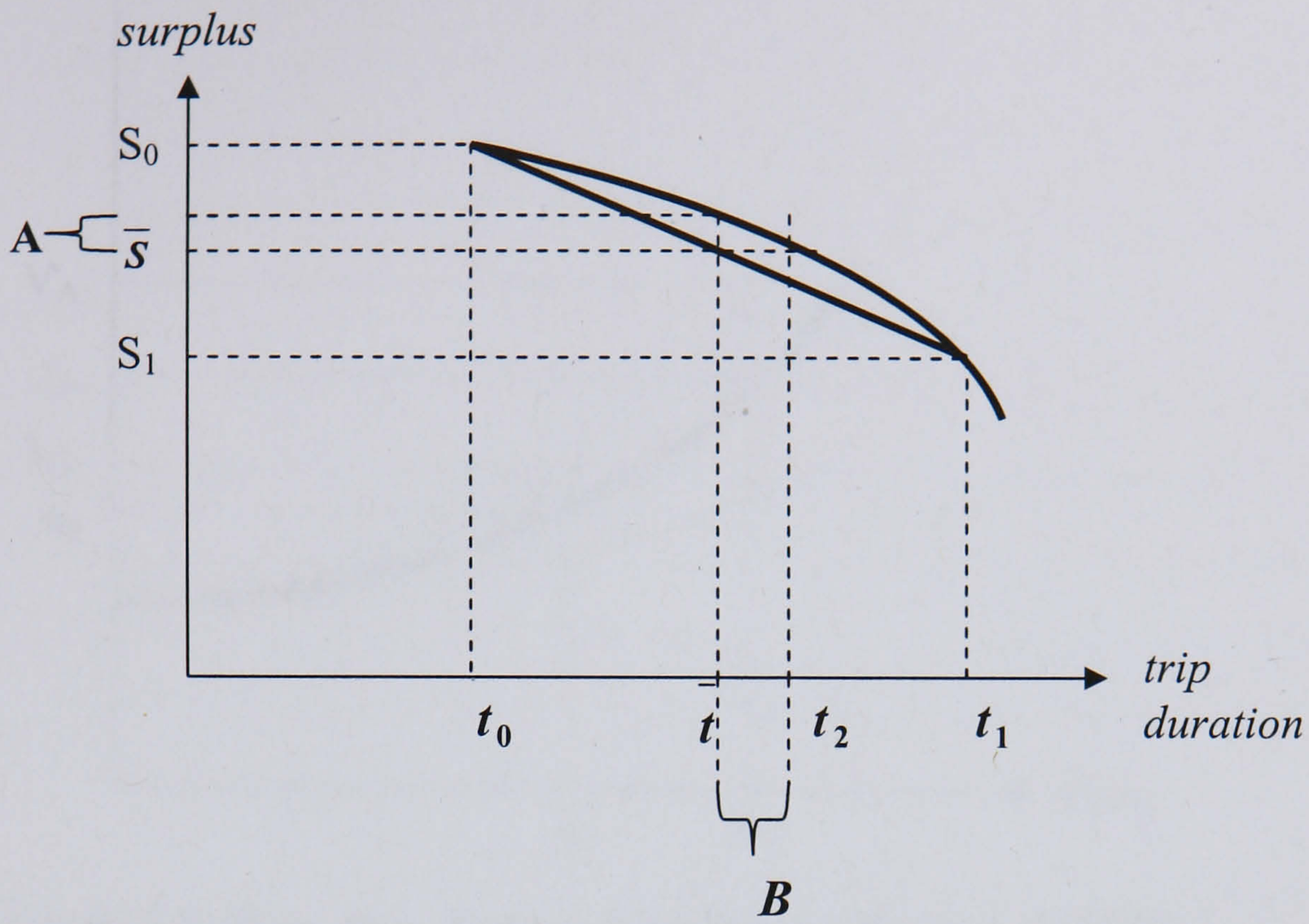


Figure 2.1: Traveller's surplus according to Knight / Rees (1974)

Knight mentions that this discussion is only valid if trip duration is the single relevant variable and if the surplus function is convex. Flexible arrival time to the destination, a high level of substitution between origin and destination activities, and various constraints on the use of time may create circumstances in which the assumptions made by Rees are questionable. However, Knight and Rees do raise an important source of disutility that should be seriously considered, possibly under more realistic assumptions, when accounting for the benefits from investment alternatives.

Another discussion of the outcomes of TTV, that is somewhat similar to the approach presented by Knight and Rees, is brought by Bates et al (1987). It is illustrated in figure 2.2.

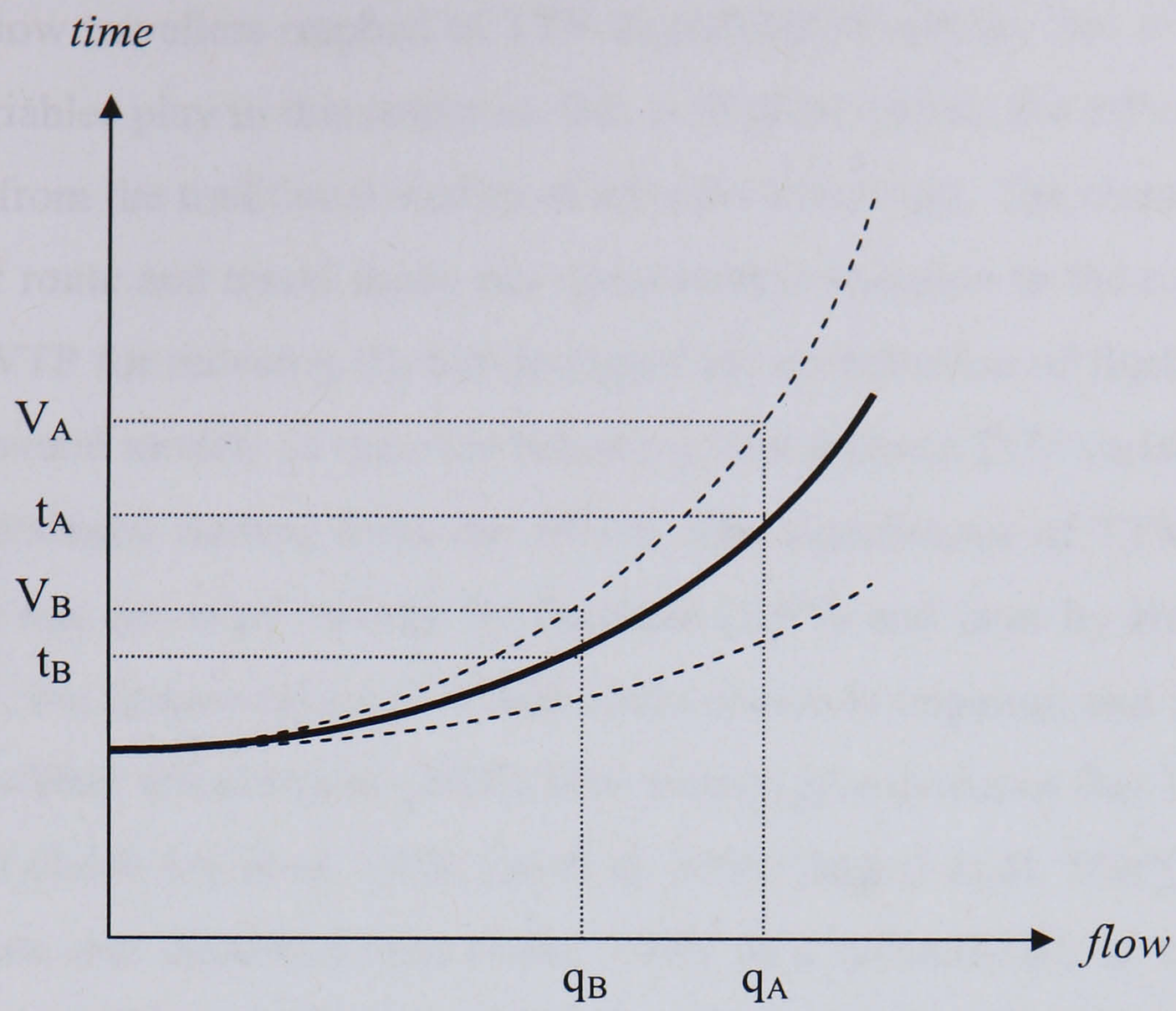


Figure 2.2: Time-flow diagram according to Bates et al (1987)

The horizontal axis represents traffic flows and the vertical axis represents time. The solid curve shows the relationship between MTT and traffic flow on a specific road link. The area that lies between the two dashed curves contains 95% of different travel times observed on the road link. It is assumed that TTV increases with flow. It is also assumed that due to some improvement of the road infrastructure, vehicle density on the discussed road link decreases and the initial flow q_A goes down to q_B. If there were no TTV, time saving would be t_A-t_B. But Bates et al theorise that since travellers are aware of possible variability in trip duration, they are not concerned about the mean time but about the 95% percentile; they use it, for example, to decide on their departure time. As a result, the saving is V_A-V_B, which is more than t_A-t_B. This might have a major impact in a CBA context.

The concept presented by Bates et al, similar to the one discussed earlier by Knight and Rees, describes the behaviour of travellers as a very simple mechanism that responds to a limited number of variables; it ignores more complex considerations that travellers might have when scheduling their journey. Still, these ideas do contribute theoretical reasoning to the additional generalised cost or disutility attributed to TTV.

The abovementioned works do not aim at making practical recommendations concerning the introduction of TTV as an integral factor in a CBA framework. To make such recommendations based on the concept of the WTP, it is necessary to examine

empirically how travellers respond to TTV in real-life situations, and to find what role monetary variables play in this response. But as implied earlier, the influence of TTV is often absent from the traditional studies of traveller behaviour. The classical models for the choice of route and travel mode pay considerable attention to the role of the MTT (and to the WTP for reducing it), but disregard the contribution of fluctuations around this mean. Several models of traveller behaviour that include TTV variables have been gradually introduced starting from the 1970's. The significance of TTV in explaining mode choice was revealed initially by Prashker (1979) and later by Hendrickson and Plank (1984); the discussion of TTV and mode choice is ongoing, and includes recent works such as Bhat and Sardesai (2005). Few studies give evidence that TTV influences route choice (Abdel-Aty et al, 1995; Liu et al, 2004; Bogers et al, 2005), the combined choice of route and departure time (Lam, 2000) or a combination of route, time and mode (Lam and Small, 2001). However, the big majority of route or mode choice models with TTV variables (all but Lam and Small, 2001) do not include monetary variables, and cannot lead to analysis of the WTP for reduced TTV. It seems that even if the extent of TTV does affect these choices, it is not yet clear what this effect implies on the WTP for reduced TTV.

An effect of TTV that has been studied more intensely than the impact on route and mode choice is the effect on departure time choice (DTC). Many researchers (Hendrickson and Plank, 1984; Mahmassani and Stephan, 1988; Noland and Small, 1995; Bates et al, 2001) stated that all other consequences of TTV are secondary to this. In the current study we accept the convention that the sought WTP for reduced TTV can be derived from a model of DTC. But although it is widely accepted that TTV is a key factor in DTC, researchers have not yet agreed on the preferable formulation of DTC models. Two approaches for capturing the effects of TTV in a DTC modelling framework appear in the literature; they are described in the following sections.

2.4.2. The trip scheduling approach

The effect of TTV on DTC was originally discussed by Gaver (1968). Gaver introduced the idea that travellers depart earlier than they would if there were no TTV, allowing some amount of slack time to reduce the chance of arriving late. Gaver proposes a model for determining this slack time (which he calls *headstart*) using a cost minimization formulation. The cost expression is linear, and includes a penalty on arriving at the destination too early or too late; an optimal trade-off is sought under various assumptions concerning probability distribution of congestion delays.

Knight (1974) refers to the early departure slack time as a *safety margin*; this term has been commonly used ever since. Similar to Gaver, Knight hypothesizes that travellers wish to be on the safe side when estimating their arrival time, but they also do not wish to arrive too early. Knight defines the traveller surplus from making the trip as the difference between the surplus from the time spent in both trip ends and the surplus from the time spent at the origin alone, since the latter would be gained if the trip were not undertaken. This definition is the basis for the convex surplus function that was discussed earlier in this chapter.

Knight illustrates the safety margin approach by focusing on the morning commuting trip, under the assumption that work start time is fixed. The marginal disutility of being late for work is determined as the difference between the marginal utilities of time spent at home and time spent at work prior to the fixed start time. The commuter wishes to minimise this marginal disutility by leaving a big enough safety margin, and it is shown that a reduction of TTV results in a smaller safety margin. Knight is the first to suggest that DTC considerations should be accounted for when evaluating the benefit from a reliability improvement. Although this is not expressed explicitly, it is indirectly implied by Knight's discussion that the benefit from reduced TTV should be calculated by multiplying the reduction in slack time by some monetary value associated with a unit of slack time. However, Knight does not carry this insight much further.

Hall (1983) explores the effect of TTV on accessibility, and dedicates part of the discussion to modelling the safety margin that traveller place on their departure time. The model determines the magnitude of the safety margin based on a maximum risk approach: travellers are assumed to delay their departure from the origin as much as they can, as long as their risk of being late does not exceed a certain limit. The WTP for reduced TTV is not discussed.

Pells (1987a) expands the discussion about the differences between values of time spent at home, at work when arriving early and at work when arriving late. Pells states that the value of time at work when arriving late is higher than the value of time at work when arriving early; this is the reason why travellers allow a safety margin of slack time on their departure time. Time spent at home has a higher value than time spent at work when arriving too early; this is the reason why travellers want to keep this slack time a short as possible. Based on this reasoning, Pells (1987b) develops two choice models for evaluation of two WTP elements: the value of slack time and the value of lateness. Pells' work is the first attempt to explicitly calculate the monetary benefit from a reduction in TTV. For doing this it uses the simple formula implied by Knight (1974): the benefit from a reliability improvement equals the reduction in slack time multiplied by the value of slack time. The value of slack time is the difference between the value of time at home and the value of early time at work.

Polak (1987a, 1987b) formulates a DTC model that takes into consideration the attitudes of travellers towards risk. A main claim in this study is that such attitudes can not be incorporated into cost minimisation models, where the measuring units of all components are fixed and the only freedom allowed is to change the cost coefficients. A non-linear expected utility maximisation formulation is therefore proposed, which is in fact a concave transformation of Gaver's linear function. Different forms of utility expressions are examined, in which both quadric and exponential functions of MTT and TTV are studied. Using different parameters enables representing varying levels of risk aversion or proneness. The discussion of these formulations is theoretic and is not tested emirically; there is also no reference to CBA practice.

Noland and Small (1995) develop a cost minimisation DTC model. The minimised cost function is:

$$C = \alpha \cdot MTT + \beta \cdot ME + \gamma \cdot ML + \theta \cdot P_L \tag{2.1}$$

Where

C	cost
MTT	mean travel time
ME	mean earliness to destination
ML	mean lateness to destination
P_L	the probability of late arrival
$\alpha, \beta, \gamma, \theta$	parameters

The authors do not calibrate the model using observed data, but estimate parameter values based on evidence from other works. Still, this expression for the cost associated with trip scheduling considerations has been accepted by many other researchers in later works as a succinct description of the factors that affect DTC (see following paragraphs). Note that the distribution of travel times influences more than one component in this function. The discussion by Noland and Small takes into account the fact that different choices of a departure time lead to different levels of congestion; two alternative assumptions regarding the distribution of delay times are examined - uniform and exponential – and optimal departure times that minimise the cost function are derived analytically under both assumptions. The authors calculate and compare the expected MTT costs and scheduling costs (i.e. lateness and earliness) under various assumptions. Scheduling cost is found to be relatively minor under the assumption of uniform delay distribution but quite large under the exponential distribution assumption.

Noland et al (1998) calibrate the cost function developed by Noland and Small (1995) using data from a stated-preference study. They apply the calibrated model in a simple imaginary network, where various congestion conditions are generated using a simulation model, and calculate the contribution of each element in the cost function to the total cost. It is shown that under various conditions, scheduling considerations account for a major part of travel cost. However, it should be noted that only the cost model, not the simulated TTV, is calibrated; hence the computed costs do not represent a realistic situation.

Small et al (1999) develop another DTC model based on similar principles to those proposed by Noland and Small (1995). The model is calibrated and validated for highway users. In addition to the linear elements in the original cost function, the potential contribution of quadric values of the same variables is examined. It is found that the predictive power of the model improves if the cost function includes a quadric element of the amount of earliness to the destination. In addition to the variable expressing the probability of arriving late, previously used by Noland and Small (1995), the authors also examine here a variable that represent the probability of an extra-late arrival. An upper limit to the reasonable amount of lateness is determined individually for each respondent, and an extra-late arrival is one that exceeds this upper limit. Both lateness probability elements appear significant. The proposed cost function therefore includes MTT, early arrival (in minutes), squared early arrival, late arrival (in minutes), probability of late arrival and probability of extra-late arrival. Different model

parameters are estimated for travellers with different employment statuses, trip purposes and demographic background.

A common feature of all works described in the paragraphs above is that they try to model the cost or disutility of TTV without explicitly using a TTV variable. This indirect approach, that is sometimes called *the scheduling approach*, is based on the concept that if travel times vary from day to day, arriving always at the destination exactly at the desired time is infeasible. Travellers can therefore react to TTV by choosing to shift their departure from home backward or forward, and by doing this, to change their chance of arriving too early or too late. The scheduling approach claims that this choice is the main manifestation of the impact of TTV. The cost attributed to TTV, according to the most recent scheduling models, is associated with the WTP to reduce the amount of earliness and lateness. This is derived from the utility function, equivalently to the more traditional computation of the VOT:

(2.2)

$$VOT = \left(\partial U / \partial MTT\right) / \left(\partial U / \partial C\right)$$

$$VOE = \left(\partial U / \partial ME\right) / \left(\partial U / \partial C\right)$$

$$VOL = \left(\partial U / \partial ML\right) / \left(\partial U / \partial C\right)$$

Where:

U	utility function
MTT	mean travel time
ME	mean earliness to the destination
ML	mean lateness to the destination
VOT	value of the mean travel time
VOE	value of the mean earliness
VOL	value of the mean lateness
C	cost

It is important to note that not all DTC models take TTV consideration into account. Bates et al (1987) review several modelling attempts in which variables that represent TTV appear not to be insignificant. Several DTC models (Vickrey, 1969; Small, 1982; Hendrickson and Plank, 1984) do not assume that travellers place a safety margin on their departure time, but they still make a contribution to the development of the

scheduling approach by introducing the idea that the cost or disutility gained from the journey is influenced by early or late arrival to their destination.

Although they have not yet been fully implemented for appraisal purposes, scheduling models form today the primary tool for estimating the cost of TTV based on behavioural reasoning. Yet, it has also been claimed that modelling only scheduling decisions does not capture the whole range of traveller responses to TTV. A modelling approach that tries to determine a monetary value for TTV per se is described in the following section.

2.4.3. The mean-variance approach

The alternative modelling approach claims that travellers see TTV per se as a source of inconvenience, independently of its consequences at the origin or the destination, similar to the way they look at the MTT; this concept is commonly referred to as the mean-variance approach. Bates et al (2001) and others note that this approach ascribes the cost of TTV to the anxiety or stress caused to travellers by the uncertainty of travel conditions. Mean-variance models use utility or cost functions that deal with TTV directly, often using a variable that stands for the standard variation of the journey time. Although most of the models reviewed here are DTC models, a mean-variance formulation can also be used in principle for modelling other choices. The main point in the distinction between the mean-variance and the scheduling approaches is the issue of whether or not the cost attributed to TTV can be adequately accounted for by the attitudes towards early or late arrival.

A mean-variance model is developed by Jackson and Jucker (1982). They calibrate utility functions for a route choice model, based on a survey in which travellers are asked to choose between different combinations of MTT and TTV. The distribution of the TTV coefficient is determined, but there is no discussion of any monetary values. Black and Towriss (1993) improve the surveying techniques (see later in this chapter for a review of surveying issues) but do not introduce changes to the modelling methodology.

Senna (1994a, 1994b) creates a model that combines the mean-variance approach and the modelling concepts proposed by Polak (1987a, 1987b). That is, he develops a utility maximisation model that incorporates MTT and TTV with other components that stand for travellers' risk aversion or proneness. Alternative formulations of the utility function are examined and compared. In addition, attention is paid to the possibility of flexible arrival times to the destination. Senna formulates the mean-variance equivalents of

formulas (2.2), expressing the WTP for improved travel conditions as a direct function of MTT and TTV:

(2.3)

$$VOT = \left(\partial U / \partial MTT\right) / \left(\partial U / \partial C\right)$$

$$VOV = \left(\partial U / \partial TTV\right) / \left(\partial U / \partial C\right)$$

Where:

U	utility function
MTT	mean travel time
TTV	travel time variability (e.g. the standard deviation of travel times)
VOT	value of *MTT*
VOV	value of *TTV*
C	cost

However, in the part of Senna's work that illustrates the estimation of the costs and benefits, the value of TTV is calculated not according to the abovementioned formula but as the difference between two VOT estimates, one determined under the assumption that there is no TTV, and the other one determined assuming that TTV does occur. Such approach seems sensible but the rationale behind it is not explained.

Noland et al (1998), whose scheduling model was presented previously, investigate the assumption that in addition to scheduling costs there is also a cost element that they call *planning cost*. They describe this supplementary cost, which is a function of the mean and the standard deviation of travel times, as related to the pure nuisance of not being able to plan one's activities precisely, which is an outcome of TTV. The authors use their stated-preference survey data to estimate a model that includes the typical variables of both scheduling and mean-variance approaches. The results show that the planning cost is not as powerful as the other variables in explaining commuters' decision-making. Much of the reaction to the uncertainty in commuting trip duration seems to be explained better by the scheduling delay and lateness probability variables.

Small et al (1999), in the study described in the previous section, use their survey data to compare two model versions, one having a simple mean-variance formulation and the other including both mean-variance and scheduling variables. When the utility function does not include scheduling considerations, the TTV variable is found significant; the value that travellers ascribe to improvement in TTV is twice as high as the value that

they place on MTT. But when scheduling costs are explicitly accounted for, the TTV measure no longer has explanatory power. The authors strictly conclude that scheduling costs account for all the aversion to TTV, and that "in models with a fully specified set of scheduling costs, it is unnecessary to add an additional cost for unreliability".

Noland and Polak (2002) show that in a special case, where there is no lateness penalty and where changing the choice of departure time does not lead to a change in recurrent congestion, the scheduling and mean-variance approaches are equivalent. Bates et al (2001) note that empirically, the sum of the earliness and lateness components in the utility function of a scheduling model can often be approximated by a single component expressing the standard deviation of travel times. Although these insights are only valid under certain conditions, it is still interesting that a utility function with scheduling variables can be reduced to an expression that includes only the MTT and TTV. This might suggest that the mean-variance approach is a simplified form of a scheduling model, or that the costs captured in mean-variance formulations are in fact implicit estimates of scheduling considerations.

2.4.4. The cost of travel time variability for public transport users

Most of the models cited above focus on the behaviour of car users. An intermodal comparison of values of TTV is derived by Black and Towriss (1993) from their mean-variance models of car, bus and train users; scheduling considerations or departure time choice are not explicitly addressed. There are very few studies of attitudes to TTV among public transport users. For bus users, Pells (1987a, b) introduces innovative ideas regarding scheduling behaviour, but does not use, within a single cost function, the entire set of scheduling variables that were later found necessary by other authors. For rail travellers, the only work known to us is described by Cook et al (1999) and analysed further by Bates et al (2001).

The model presented by Bates et al (2001) attempts to take into account a major source of complexity in modelling the DTC of public transport users: the need to consider a given timetable of the service being used. Unlike car users, that can choose departure times along a continuous time scale, public transport users make a discrete choice of the departure time from their initial boarding point. The authors note that the discrete nature of public transport DTC is both a modelling difficulty and an additional source of disutility for travellers. They propose a utility maximisation model for the simultaneous choice of a railway service and a departure time. The utility function is the following:

$$U = \beta \cdot ME + \gamma \cdot ML + \phi \cdot f + \eta \cdot h + \upsilon \cdot \mu \tag{2.4}$$

Where

U	utility function
ME	mean earliness to the destination
ML	mean lateness to the destination
f	fare
h	headway
μ	mean delay, i.e. the difference between actual and scheduled arrival
$\beta, \gamma, \phi, \eta, \upsilon$	parameters

This model is calibrated based on a series of stated-preference studies. A variable expressing the MTT is not included. A direct representation of TTV is excluded as well, due to a high level of correlation between the standard deviation of travel times and the mean delay variable. An important finding from the modelling experiment is that a nested structure is strongly implied, i.e. change in DTC is much more likely than a change in the choice of a rail service. This resembles the conclusions reached in the aforementioned DTC models for car users.

Another important finding is that the variables related to scheduling are not the only factors that appear significant. Earliness and lateness do have a major role in the model, but the variable that represents the mean delay in arrival time contributes to the predictive power of the model as well. This contradicts the findings reached by Noland et al (1998) and Small et al (1999) concerning DTC of car users. Bates et al (2001) conclude that railway users associate some further disutility with TTV per se, over and above the contribution of scheduling variables.

2.4.5. Summary

The main studies reviewed above are summarised in table 2.1. It is apparent from this review that there has recently been a significant improvement in the ability to model the way TTV affects decisions made by travellers. There has also been considerable progress in estimating the WTP for a reliable journey. Despite this progress, it is still unanswered whether a scheduling model can satisfactorily explain the reaction of most travellers to TTV. While for modelling the effect on car users there seems to be enough evidence for the sufficiency of scheduling variables, the little documented evidence for rail users is to the contrary; for bus users there is hardly any evidence at all. There is great interest in finding whether bus users will appear to be motivated by scheduling consideration only, similar to car users (according to Noland et al, 1998, and Small et al, 1999), or also by the nuisance caused by TTV per se, similar to rail users (According to Bates et al, 2001). We probe into this issue in chapter 3 of this thesis.

The majority of studies reviewed here suggest that scheduling models give a better understanding of the cost associated with TTV. But most practical works that involve evaluation of this cost use mean-variance models; see, for instance, TRL (2004), Atkins (1997) and others. The main reason for this is probably that the implementation of a mean-variance model is fairly straightforward and only requires aggregate estimates of MTT and TTV, while applying a scheduling model requires information on the distribution of preferred arrival times and simulation of the mean lateness and earliness at a disaggregate level. A serious concern, which has not been discussed in any of the studies reviewed here, is whether the common use of the formulation that is normally found inferior leads to any bias in the evaluation of the cost of TTV. This issue is further investigated in chapter 3.

Source	Approach	Formulation	Calibrated?	Attitudes to risk?	Market segmentation?	Calculation of costs or benefits?
Gaver (1968)	Scheduling	Cost minimisation	No	No	No	No
Knight (1974)	Scheduling	Utility maximisation	No	No	No	No, but the idea is introduced
Jackson and Jucker (1982), Black and Towriss (1993)	Mean-variance	Utility maximisation	Yes	No	No	No
Hall (1983)	Scheduling	Joint minimisation of time and risk	No	No	No	No
Pells (1987a, b)	Scheduling	Utility maximisation	Yes	No	Yes	Yes
Polak (1987a, b)	Scheduling	Utility maximisation	No	Yes	No	No
Senna (1994a, b)	Mean-variance	Utility maximisation	Yes	Yes	Yes	Yes
Noland and Small (1995)	Scheduling	Cost minimisation	Partially	No	No	Yes
Noland et al (1998)	Scheduling + mean-variance	Cost minimisation	Yes	No	No	Yes
Small et al (1999)	Scheduling + mean-variance	Utility maximisation	Yes	No	Yes	No
Cook et al (1999), Bates et al (2001)	Scheduling + mean-variance	Utility maximisation	Yes	No	No	Yes

Table 2.1: Models for the effects of TTV

Note that in the abovementioned models for the attitudes to TTV there is little account for the heterogeneity of tastes and preferences among travellers. A theoretical discussion of possible ways to account for varying levels of risk aversion across the population is presented by Polak (1987a, b) but not calibrated with real data. Senna (1994a, b) continues this discussion and calibrates models for travellers with different conceptions of risk, but only uses mean-variance formulations, and does not derive any estimates of the distribution of the WTP. Few other studies, such as De Jong et al (2004), Hess et al (2004) and Rohr et al (2005), present models for the choice of mode and time of day, which use a Mixed Logit formulation that fully considers heterogeneity of preferences across the studied population. These models do make an important contribution by allowing for a distribution of individual attitudes to the extent of earliness and lateness. However, the discussed lateness and earliness depend only on the MTT; the surveys on which the models are based do not present distributions of journey times, and therefore, the effects of TTV cannot be captured even if scheduling variables are included. The modelling attempts described in chapter 4 of this thesis try to meet the need for a study of variations between travellers in their attitudes to TTV.

2.5. Travel time variability in stated preference surveys

2.5.1. Basic challenges in survey design

We now shift the focus of this review from the evaluation of the cost of TTV to the stage of data collection that normally precedes the evaluation. In chapter 3 we describe a survey conducted in order to investigate DTC considerations of bus users and their attitudes to TTV; in this section we describe several earlier works that inspired the design of our survey.

In order to estimate a choice model that includes TTV variables it is necessary to find situations where travellers have a choice between well-defined alternatives that differ in the extent of TTV. Bates et al (2001) point out that using revealed-preference (RP) techniques, which examine the choices travellers have actually made in observed situations, is not practical for TTV valuation because it is extremely hard to find such situations. They state that for modelling the effects of TTV, stated-preference (SP) techniques are "de facto almost always the only realistic possibility for data collection". This is indeed the case in the current study, as we did not have detailed information on

any situations in which travellers in the study area could trade-off between TTV and money. Still, it should be mentioned that the difficulty in using RP data for modelling the effects of TTV is a technical problem and not a fundamental problem. Although most models in this area are based on SP experiments, several RP-based models do exist, such as those presented by Brownstone and Small (2005) and Small et al (2005). Brownstone and Small (2005) compare values of time and reliability from several SP and RP models; their review confirms that econometric analysis of the effects of unreliability based on RP data has been undertaken, although in a very small number of studies (which were reviewed earlier in this chapter).

SP techniques offer a higher level of flexibility than RP since they analyse the responses to hypothetical scenarios. A typical questionnaire in an SP experiment presents a series of choice situations; in each situation the respondent faces two or more choice alternatives. Many works (e.g. Fowkes and Wardman, 1988; Fowkes and Preston, 1991; Hensher, 1994; Hensher, 2004; Caussade et al, 2005; Cirillo, 2005; and others) have discussed issues and challenges in the design of SP surveys. The major decisions that need to be made when designing such a survey include the following:

- Setting the number of attributes that define each of the presented alternatives, and deciding what would each attribute stand for. Normally, the attributes correspond to potential variables in the choice model that the survey designer wishes to base on the survey responses.
- Determining the number of alternatives in each choice situation.
- Defining the task the respondent is asked to perform. The common options are *choosing* one alternative, *rating* the alternatives or *ranking* them.
- Setting the number of different levels for each of the attributes presented in each choice situation, and choosing the particular values for these different levels.
- Deciding to what extent the series of choice situations would cover the entire set of combinations of feasible levels of the attributes. *Factorial* designs cover the whole set of combinations, but are rarely used since the full number of combinations might require a very lengthy questionnaire. *Fractional factorial* designs only use a subset; if a fractional factorial design is used, the total number of choice situations in the questionnaire should also be determined.

There seems to be an underlying contradiction between the different criteria that the survey designer should follow when making the abovementioned decisions. On one

hand, the primary objective of the survey is always to get as much data as possible. On the other hand, designs that try to extract an excessive amount of information from the respondents might actually achieve the opposite. If the questionnaire is either too long or too hard to understand, the responses might not be credible, because of the "limited ability of consumers to process the information presented in the experiment" (Caussade et al, 2005). The need to balance between the desire for a rich dataset and the risk of an incomprehensible questionnaire is the incentive for much of the recent discussion of SP concerns.

In order to make the survey easy to understand and to assure that the respondent does not get tired or bored while filling in the questionnaire, it is inevitable to limit the number of choice situations and the number of alternatives per situation. Hensher (1994) discusses some of the consequences of the attempts to make the SP experiment manageable; these involve some loss of statistical efficiency, since the designer has to assume that certain interactions among the attributes are not significant. Hensher states that the designer has to be creative in selecting a limited number of combinations of the attributes.

Since we want to construct a rich dataset despite these restrictions, the task of determining the levels of the presented attributes often becomes the most critical element in the design process. The attribute levels are normally not constrained by complexity considerations, and they have a direct impact of the ability to identify, at a later stage, the behaviour patterns of the respondents. Most of the studies cited above (especially Hensher, 1994; and Caussade et al, 2005) have therefore highlighted this issue from various perspectives. However, Cirillo (2005) illustrates a serious obstacle to an optimal design of the attribute levels: since it is common to use the responses to estimate non-linear choice models, the best design depends on the values of the model parameters, but these are obviously unknown at the stage of design.

2.5.2. Presenting the idea of travel time variability

SP surveys with TTV attributes have already been designed in some of the studies mentioned earlier in this chapter. These surveys conventionally ask the respondents to make hypothetical choices that reveal the way they trade-off between TTV and other elements of the generalized cost of travel, such as the MTT or the journey cost. SP methodologies that concentrate on the trade-off between sources of discomfort to travellers are very common, but in most of them the respondents are asked to exchange

MTT and money; the description and presentation of such attributes in a questionnaire is relatively straightforward. The introduction of TTV an attribute raises the issue of how to present the level of TTV in a clear and simple way. The difficulty lies in the fact that analysts measure TTV using terms such as standard deviation, that are not intuitively understood to many of the respondents (Cook et al, 1999). Hence, many researchers of this area focused their discussion on how to illustrate the concept of a probabilistic travel time distribution to respondents.

Early SP experiments described a given distribution of travel times by noting the *usual travel time* and the *extent and frequency of delay*. The usual travel time in these experiments is sometimes defined in an informal manner, e.g. as the "approximate time it takes most of the time" (Jackson and Jucker, 1982). The frequency of delay is given in terms such as "once a week". A typical question in surveys of this type (that we denote "type 1") would ask respondents to choose one of the following alternatives (Jackson and Jucker, 1982):

Alternative	Usual travel time (in minutes)	Delay
A	50	None
B	40	20 minutes once a week

Table 2.2: Typical choice situation - type 1

It is common in such questionnaires that in one of the presented alternatives there is no delay at all. A slightly different version of such questionnaire (that we denote "type 1a") fixes the frequency of delay at a constant level in the introduction to the main questionnaire, and then presents choice situations that differ only in the usual time and the amount of delay.

In recent years, most experiments used a more explicit formulation, which was originally proposed by Benwell and Black (1994). This methodology (that we denote "type 2") demonstrates the TTV pattern as a sequence of several journey times. The number of times presented to define a single distribution varies in different works from 5 to 10. A typical choice scenario would be the following (Noland et al, 1998):

Alternative	Departure time (in minutes before the desired arrival time)	Typical travel times on 5 different days
A	15	12, 13, 14, 16, 20
B	10	5, 7, 9, 12, 18

Table 2.3: Typical choice situation - type 2

Studies that use questionnaire type 2 state that it is better understood, since it gives a plain illustration of the unpredictability of travel times. Still, there are also discussions of its disadvantages. Noland et al (1998) mention that using a list of travel times to describe a distribution creates an artificial certainty about the maximum possible delay. Many authors emphasize that the order of the presented travel times is important: some respondents mistakenly assume that the times are displayed in descending or ascending order when they are actually not, and thus do not read carefully the entire sequence. The most recent design (Bates et al, 2001) tries to solve this problem by listing the travel times not as a sequence, but using a circular "clock-face" presentation (see figure 2.5).

Both questionnaire types 1 and 2 do not explicitly present the levels of all the variables that are subsequently used for modelling. The modeller normally does not employ the set of presented attributes, but a different set of variables that captures the same information in a way that is statistically more concise. The modelling set explicitly includes variables such as the standard deviation of travel times or the probability of late arrival, whose direct presentation is avoided.

Note that although in the general SP literature there are many examples for surveys where more than two alternatives are presented in each choice situation, most the surveys reviewed here tend to have only two alternatives. This is clearly because the implicit presentation of a TTV variable requires displaying more information per alternative than is normal in other SP experiments; since each alternative is relatively complex, most authors prefer to present the smallest possible number of alternatives. With only two alternatives per question, the only relevant task for the respondent to perform is to choose the preferable alternative; ranking and rating are irrelevant in the case if two alternatives. Pells (1987a, b) and Senna (1994a, b) ask the respondents to choose one of five options, but these are based again on only two alternative sets of travel conditions, and the five options merely specify different levels of preference

("definitely choose A", "probably choose A" etc.). Cook et al (1999) and Bates et al (2001) include three optional departure times for each level of TTV, and thus even though only two distributions of travel times are displayed, the number of alternative combinations amounts to six.

2.5.3. Full survey design

The most common attributes that form each of the presented alternatives, in surveys that investigate the attitudes to TTV, are the MTT, TTV, departure time and the cost of the journey. As mentioned earlier, for a powerful choice model it is important that the different levels of the survey attributes are carefully designed. The levels should cover the entire hypothetical range of circumstances in which the model we wish to calibrate should work. With the increasing use of computerised surveys, it has recently become common to ask the respondents about their normal MTT, desired arrival time to the destination and departure time, and then base the presented levels on the reported values, so that the choice situations are adjusted to travel conditions that each respondent is familiar with.

Most authors base the design of the attribute levels on three preset levels of the feasible values of each attribute – high, medium and low (Senna, 1994; Noland et al, 1998; Small et al, 1999). TTV levels (measured as the standard deviation of travel times) typically vary from 0% to 50% of the usual travel time (Senna, 1994) or from 10% to 30% (Small et al, 1999). The range of departure times can be fixed such that the range of expected arrival times will start two standard-deviation-long time units before the usual arrival time (i.e. very early departure) and end at the usual arrival time (i.e. very late departure) (Noland et al, 1998); or alternatively, such that the shift of departure time due to TTV varies between 0% to 15% of MTT (Small et al, 1999). The levels of the MTT are determined by Small et al such that the middle level is equal to the MTT reported by the respondent, and the lower level does not fall under the free flow travel time. Within the range of levels for each variable, Small et al find that the design is more powerful when the low, medium and high levels are not evenly spaced, i.e. the medium level is not exactly the average of the low and the high.

If three levels of n attributes are identified, the number of choice alternatives would be 3^n. The full set of alternatives can be reduced to a much smaller set, in which no alternative dominates another in all n variables. Once this subset is identified, several pairs of alternatives can be randomly chosen to formulate each choice situation in the

fractional factorial design. Most surveys documented in literature include between 6 and 10 situations.

Not all the abovementioned SP experiments include a cost variable. The exclusion of the cost from the set of survey attributes helped several authors (e.g. Noland et al, 1998) reduce the amount of information displayed, and thus alleviate the cognitive burden on the respondent. However, surveys where the alternatives that the respondent chooses from do not differ in their cost cannot be used for determining the WTP.

Most of the early designs presented the distribution of travel times without explicitly noting the departure time. Noland et al (1998) were the first to state the departure time, described as "15 minutes before your usual arrival time". This approach was later employed by other authors, some of which worded it simply as "depart at 08:10". In principle, it is possible to estimate a DTC model even if departure time itself is not specified; but since most recent models also account for the occurrence of early or late arrivals, the explicit departure time appears an important element in the SP design (Bates et al, 2001). In contrast, Small et al (1999) choose to leave the exact departure time explicit, fearing that too much information is already presented.

Due to the complexity of the information displayed in surveys with TTV attributes, it is common to precede the main body of the questionnaire with an introductory section. This section is meant to orient the respondents to the idea of TTV and familiarise them with the way it is presented in the main questionnaire. Jackson and Jucker (1982) mention that it is also important to clarify the basic assumptions of the survey, e.g. that it is not possible to avoid the delay by switching route or to predict when the delay will occur.

Most of the latest ideas in the design of surveys that look at the attitudes to TTV are implemented in the study of rail users' choices, described by Bates et al (2001). Parts of this computer-based experiment are demonstrated in the following figures. Figures 2.3 and 2.4 show two of the introductory pages presented before the main questionnaire to present the idea of unpredictable travel times. These pages illustrate the same level of TTV in different ways, but only the last format is then repeatedly used in the main questionnaire. Figure 2.5 shows a typical choice situation in the main questionnaire, where a "clock-face" diagram is used to depict the distribution of the amounts of earliness or lateness to the destination.

2.5.4. Summary

Characteristics of the reviewed SP methodologies are summarised in table 2.4. Many elements in the existing methodologies, such as most of their statistical properties, seem generally suitable for the current study. However, despite recent developments that aimed at making the idea of TTV easy to grasp, it still seems that this objective is not fully met. Even in advanced designs such as the one proposed by Bates et al (2001), the respondent is expected to read twenty daily arrival times (of the format "15L", "6E" etc.), six scheduled departure times, six scheduled arrival times and two fare levels prior to making a single choice. This is repeated in every choice situation, and one might therefore doubt whether all respondents are able to process this amount of information and provide credible responses.

Note that none of the surveying methodologies mentioned here uses graphical presentation to explain the difference between different distributions of travel times. The "clock-face" display in figure 2.5 makes a step in this direction, but the circular form of the presented times is only meant to imply that attention should be paid to their order. The shape of the circle does not change between alternatives or between choice situations, and it therefore does not illustrate the level of TTV itself. In the design of the survey described in chapter 3, an attempt is made to deliver the same information using a lighter presentation, and to use a graphical display to make the information understood even if the respondent does not rigorously read all the presented numbers.

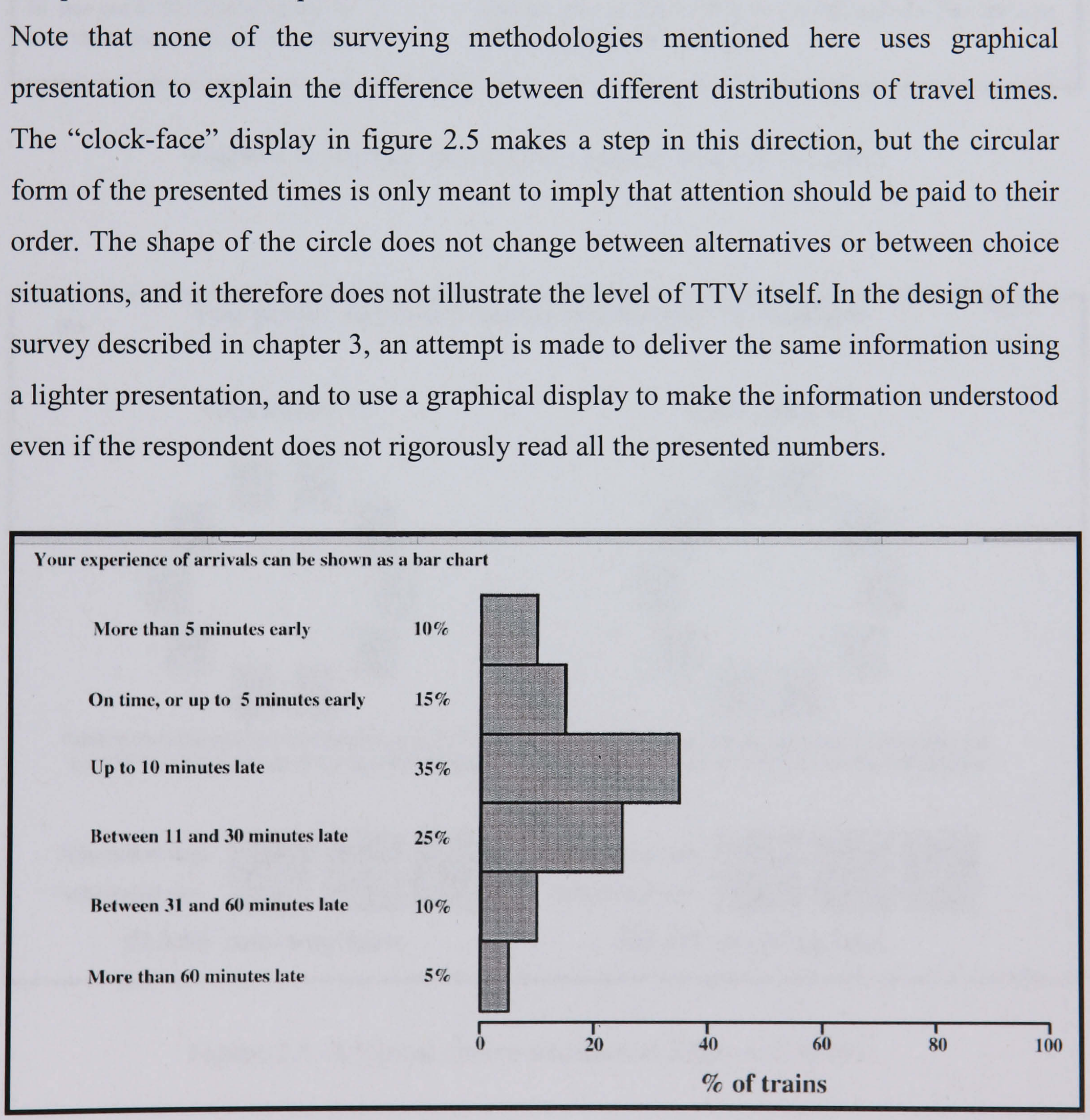
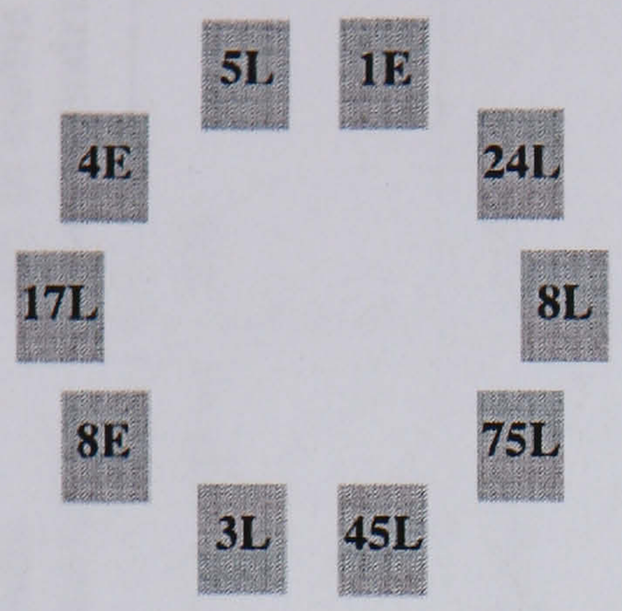


Figure 2.3: One of the introductory pages at Bates et al (2001)



More than 5 minutes early	10	%
On time, or up to 5 minutes early	15	%
Up to 10 minutes late	35	%
Between 11 and 30 minutes late	25	%
Between 31 and 60 minutes late	10	%
More than 60 minutes late	5	%

Later in this interview, we are going to use cards like those above to represent late arrivals of trains...

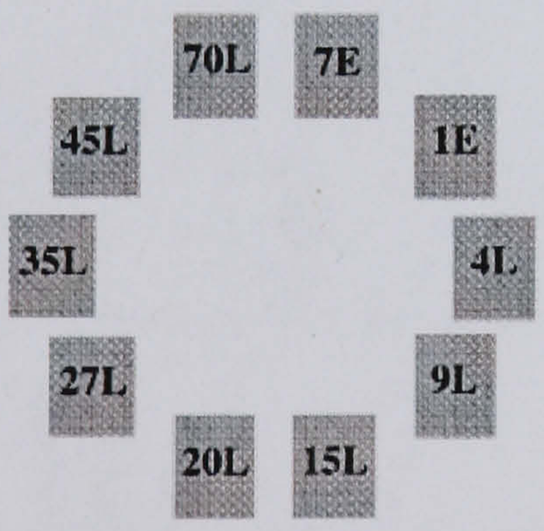
... and we'd like you to think of the cards as being just another way of describing an arrival pattern like the one you originally gave us above

Figure 2.4: Another introductory page at Bates et al (2001)

10a

You prefer to be at London Paddington at 11.00am

Operator A

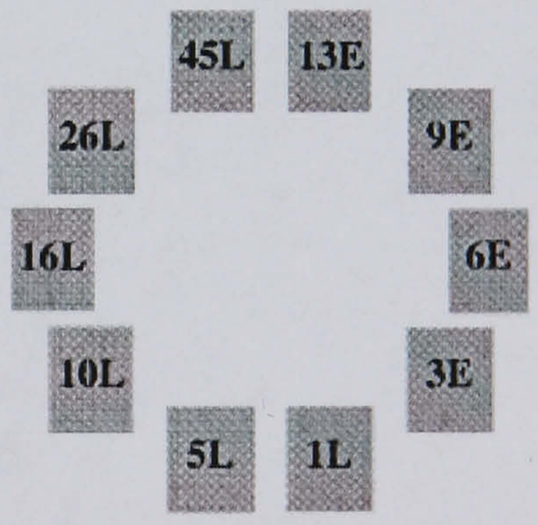


Pattern showing number of minutes early/late for typical ten train arrivals of London Paddington

Scheduled dep.	0704	0804 1 st	0904
Scheduled arr.	0940	1040	1140

£13.00 one-way fare

Operator B



Pattern showing number of minutes early/late for typical ten train arrivals of London Paddington

Scheduled dep.	0634	0804 2nd	0934
Scheduled arr.	0910	1040	1210

£15.50 one-way fare

Figure 2.5: A typical choice situation at Bates et al (2001)

Source	Includes cost?	Modes	Form of questionnaire	Departure time noted?	Respondent's choice	Model description	No. of usable questionnaires
Jackson and Jucker (1982)	No	No distinction between modes	Type 1	No	Choose 1 of 2 alternatives	Mean-variance	200
Pells (1987a, b)	Yes	Bus	Type 1	No	Choose 1 of 5 options: "definitely choose A", "probably choose A" etc.	Scheduling	207
Bates et al (1987) and Johnson et al (1989)	No	Car	Type 1a	No	Choose 1 of 2 alternatives	Scheduling	Unknown
Black and Towriss (1993)	Yes	Car, bus and rail	Type 2 with 5 items	No	Choose 1 of 2 alternatives	Mean-variance	606
Benwell and Black (1994)	No	Rail	Type 2 with 10 items	No	Choose 1 of 3 alternatives	Unknown	Unknown

Table 2.4: Surveys with TTV attributes

Source	Includes cost?	Modes	Form of questionnaire	Departure time noted?	Respondent's choice	Model description	No. of usable questionnaires
Senna (1994)	Yes	No distinction between modes	Type 2 with 5 items	No	Choose 1 of 5 options: "definitely choose A", "probably choose A" etc.	Mean-variance	301
Atkins (1997)	Yes	Bus	Type 1	No	Choose 1 of 2 alternatives	Mean-variance	156
Noland et al (1998)	No	Car	Type 2 with 5 items	Yes	Choose 1 of 2 alternatives	Scheduling / mean-variance	543
Small et at (1999)	Yes	Car	Type 2 with 5 items	No	Choose 1 of 2 alternatives	Scheduling / mean-variance	959
Cook et al (1999), Bates et al (2001)	Yes	Rail	Type 2 with 10 items. Two alternative TTV patterns, each of them having 3 possible departure times – 6 combinations in total	Yes	Rank the 4 best of the 6 combinations	Scheduling / mean-variance	672 pseudo-observations based on 28 questionnaires

Table 2.4 (continued): Surveys with TTV attributes

2.6. Conclusions

The variability of travel times is a source of inconvenience to travellers and there is theoretical and empirical evidence that it affects their travel behaviour. There is great interest in studying the willingness of travellers to pay for reducing this variability, in order to incorporate such effects in the appraisal of transport schemes. The WTP for reduced TTV is conventionally investigated by examining the impact of TTV on departure time choice considerations. Many researchers have studied this issue, but some important aspects of the problem have not been tackled. It is still unclear whether the cost associated with late or early arrival to the destination fully captures the attitudes towards TTV. This is particularly true for public transport users: there is only one study of the behaviour of rail users, and hardly any study of bus users. Since practitioners normally prefer mean-variance models, even if there is some evidence that they are less powerful than scheduling models, it is also important to examine what effect this preference has on the results of appraisal studies that use these models. In addition, new ideas are needed in order to alleviate the cognitive burden on the participants of surveys with TTV attributes, as current questionnaire designs present extensive amount of numerical information with very limited use of graphics. In chapter 3 we hope to use all the information assembled here on these issues for collecting data on bus user preferences, modelling their DTC behaviour, and deriving their WTP for reduced TTV.

The review presented here also reveals that all existing attempts to determine the WTP for improved reliability ignore the variations in tastes and preferences among travellers. The scope for calculating the WTP not as a single average number but as a distribution is explored in chapter 4.

Chapter 3

A monetary value for travel time variability

3.1. Introduction

We are interested in evaluating the willingness of bus commuters in York to pay for reduction in the extent of TTV. Many of the works reviewed in chapter 2 evaluate the WTP through modelling of DTC considerations, and we choose to do the same here. The first part of this chapter describes the design of a survey conducted as the basis for the analysis of DTC and attitudes towards TTV. The modelling work that followed the survey is described in the second part of this chapter. Further analysis of the survey is also presented in chapter 4.

3.2. Survey design

3.2.1. Internet-based surveying

The survey described here was conducted and distributed through the internet. The questionnaire is a series of programs, accessed through the survey website (http://www.its.leeds.ac.uk/survey/). The survey programs are written in PHP (Hypertext Preprocessor), which is a programming language mainly used for server-side internet applications.

Internet-based transport surveys have been used in many recent studies (e.g. Killi and Nossum, 2003; Hojman et al, 2004; Bhat and Sardesai, 2005), but not all the advantages and disadvantages of this medium are recognised. The main motivation for holding a survey on the internet in most cases is the low cost, as there is often no need in such survey to hire additional staff, produce paper-based prints or send large amounts of mail. In the current study, the costs associated with traditional surveying methods made any other option but using the internet impractical, as the entire design, distribution and processing had to be at a zero cost. Additional incentives for an internet survey are the automatic creation of the database, which is ready to analyse in no time; the ability to use colourful, graphical and dynamic presentation without additional cost; the possibility of modifying the survey wording after its distribution has already begun; and the ability to create an automatically customised, respondent-adaptive questionnaire.

A major difficulty with internet surveys is that they cannot reach the potential respondents using traditional methods. Two common methods of distributing internet surveys are either by sending letters that ask the recipients to enter the survey website, or by purchasing lists of email addresses of people or businesses that have given their consent to be contacted for commercial purposes; both methods are pricey and therefore could not be used here. Distribution of the survey was enabled thanks to the cooperation of several contact people in various organisations in York, who were willing to circulate within their organisations an email with a request to fill in the questionnaire, and a link to the survey website. A related disadvantage of an internet survey is a relatively low response rate; many recipients are regularly bombarded with email messages that invite them to follow a link to an unknown website, and cannot be blamed for discarding the message that informed them about the current survey. Among the organisations that took part in the survey, it was generally found that the number of actual responses strongly depended on the identity, rank or position of the person that circulated the link to the survey website; when it was some senior figure, or the person in charge of transport arrangements at the work place, more recipients of the email were willing to take part in the survey. It should be noted that when a questionnaire is distributed using such method, it is impossible to calculate the exact response rate, since the number of people that received the email message in the first place is unknown.

Another key problem with internet surveys is the difficulty in reaching a good representation of the population of interest, given that the sample is entirely consisted of people that are accessible through the internet. In the context of the current study there is a concern that responses from those bus users that have email access do not represent the entire range of bus user preferences. We discuss this concern, and the actions taken to deal with it, later in this section.

3.2.2. Format of the survey

The review in chapter 2 illustrated that previous studies have already discussed many surveying issues that the needs of the current study bring in. For instance, it has been made clear that it is hard to find records of choices made by travellers in situations where there are several travel options with different levels of TTV. This is often tackled by concentrating on hypothetical situations, using SP experiments. Although many aspects of the design have been discussed by others, all features of the survey were re-specified and adjusted to the current project.

A key consideration in determining that survey format is the complexity of the concept of TTV. Independently of the method chosen for presenting different levels of TTV, it is inevitable that the amount of information that needs to be displayed to illustrate these levels, and the amount of time the respondent is expected to spend processing this information, are higher compared to the other presented attributes. The main concern is that too many details per choice situation might make the SP experiment incomprehensible. The main decision this led to was to stick to the relatively simple format of two alternatives per choice situations. Some surveying methods that enable extracting more information from each respondent, such as using rating or ranking of the presented alternatives, are only available when there are three or more alternatives. But in the case when TTV is one of the attributes, the additional information might come at the expense of loss of credibility. Therefore, each situation presented in our survey will include two alternatives only, and the choice between them will be a straightforward "A or B", without rating or ranking.

A related decision has to do with the total number of choice situations in the questionnaire. The review in chapter 2 showed that in other surveys with TTV attributes, nine choice situations for each respondent are very common. This number was therefore accepted as is (although later in this chapter it is verified that the total time a respondent needs to complete the entire questionnaire is reasonable). Another feature that was borrowed from other recent works is the choice of the main attributes that define each choice alternative: the MTT, TTV, cost and departure time. The first three attributes are used in all SP experiments that examine attitudes to TTV. Although most of these experiments are used for estimating DTC models, the departure time itself is not always stated, especially in paper-based questionnaires, where it is not possible to adjust the times presented in each question to the normal travel conditions the respondent is used to experience. Since the current survey is computer-based, it was felt that explicitly presenting the departure time will make the respondent feel more familiar with the choice situation. Besides, including the departure time as one of the survey attributes is essential for the potential inclusion of variables relating to earliness and lateness to the choice model we wish to estimate.

The main element in the survey format where existing methodologies seem unsatisfactory is the presentation of the extent of TTV. As explained in chapter 2, in all recent SP experiments that investigate the attitudes to TTV, the only means for delivering the level of TTV is through display of a list of numbers. These numbers,

which represent travel times on different days, are organised on the paper or the screen in various ways, but in all cases the respondents must read and grasp all of them, one by one. If even one of the presented numbers has not been read properly, the distribution of travel times is likely to be misinterpreted. The problem with experiments that only display a list of number is that they assume that the respondents are concentrated and patient enough to read the whole list throughout the entire questionnaire. In a survey with unknown respondents, that were contacted in a random manner and have varied skills and an unproven willingness to cooperate, there are no grounds for this assumption. The survey conducted here includes a numerical display similar to previous experiment, but this is accompanied by a graphical display which has not been used previously.

The graphical display used here is meant to provide all the information related to the journey time (i.e. the departure time, MTT and TTV) in a way that can be intuitively understood even by a respondent that hardly reads any of the presented numbers. It has also been decided not to assume that the respondent knows how to interpret a diagram, i.e. to avoid presenting values on horizontal and vertical axes. The chosen formant, which tries to meet these requirements, is shown in figure 3.1. Any particular pattern of travel conditions is described as a set of five possible daily journeys; indeed some previous studies used up to ten items, but the gain from a more accurate distribution might be marred by reduced comprehensibility. Each daily journey is represented by a vertical bar; a longer bar implies longer journey, and when two alternative sets of travel times are presented in one choice situation, a late departure is shifted downwards along an invisible vertical axis. The journey time itself is not explicitly stated, only hinted by the length of the bar, but the exact departure and arrival times are displayed. To deliver the idea that journey times are unpredictable even if the departure time is unchanged, the departure time is written separately above each bar; but since this information is redundant, and to ensure that attention is mainly paid to the differences between the arrival times, the departure time is printed in a much smaller font. None of the statistics that might be included later in the choice model (MTT, TTV, mean earliness, mean lateness, probability of late arrival, and more) is displayed.

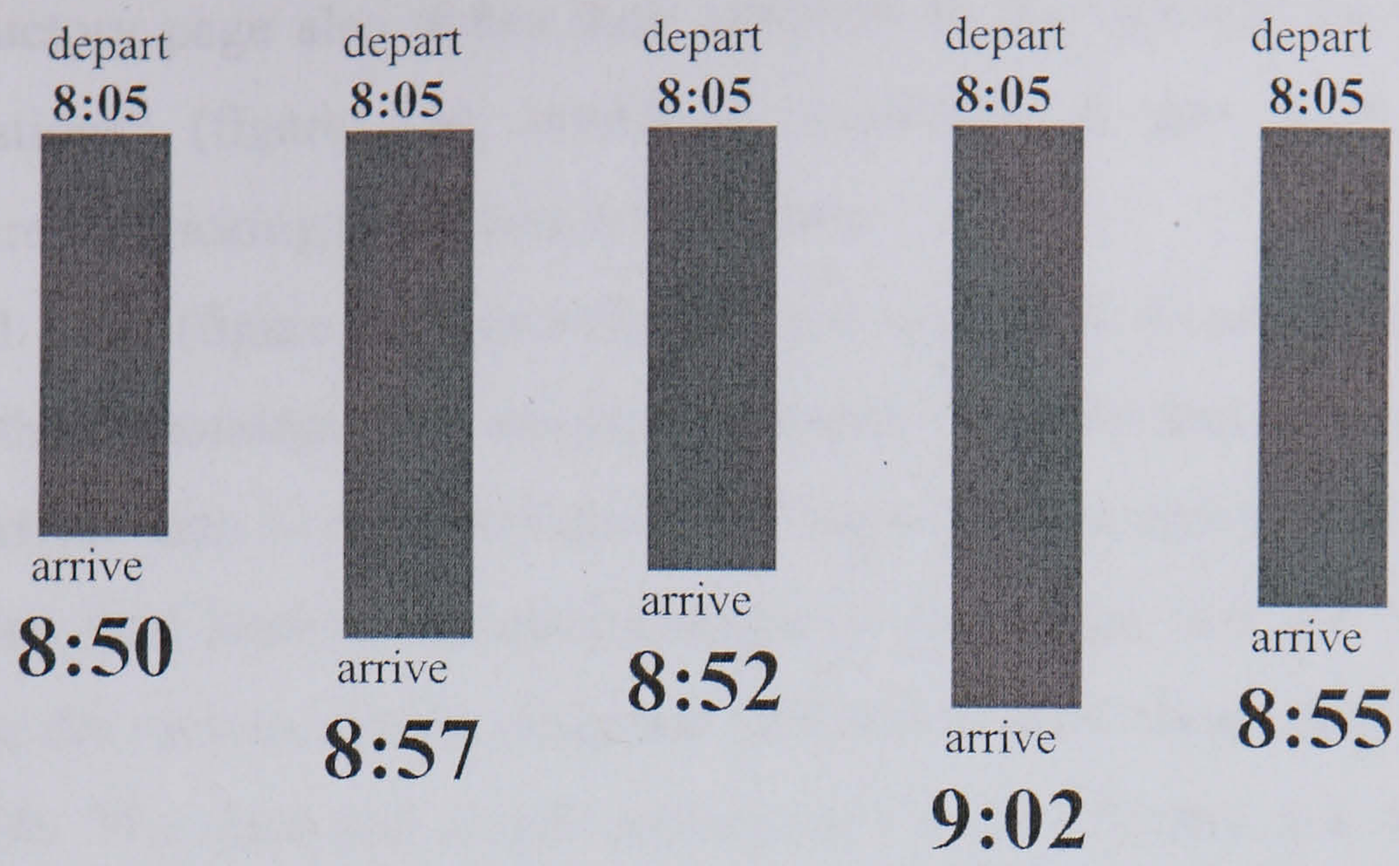


Figure 3.1: Presentation of a distribution of travel times

3.2.3. Full survey design

The entire internet-based questionnaire was prepared in two different versions. One version is addressed to bus passengers only, whereas the other version is for users of bus, car or rail. The purpose of collecting data from car and rail travellers was to enable intermodal comparison of the WTP; this analysis is briefly described in appendix A. The main analysis described in this chapter is based on responses from bus users that filled in either of the questionnaires. Figures 3.2 to 3.8 are taken from the intermodal version of the survey, and hence the wording used is slightly more general; but note that there are only very subtle differences between this version and the version for bus users only.

The questionnaire starts with a series of four introductory pages. The objectives these pages are aimed at are the following:

1. Give a very general introduction to the current research.
2. Verify that respondents that are not included in the population of interest do not proceed to the main part of the questionnaire.
3. Ask each respondent about his/her daily commuting journey. The details of this journey are used later in the main questionnaire.
4. Orient the respondent to the idea of TTV and to the way it is presented in the main questionnaire.

The first introductory page (figure 3.2) is mainly used as a filter: it clarifies that the questionnaire is to be filled in by certain travellers only. Readers are asked to halt if they do not use the travel modes of interest (in the version for bus users only this is clearly stated at this point), or if they are not regular commuters in the study area. The

first introductory page also draws their attention to the optional page of "Frequently asked questions" (figure 3.8), which is accessible at any point throughout the questionnaire by clicking the "More info" button.

The second page (figure 3.3) includes several questions about the daily commuting journey of the respondent. The sought information includes that normal travel time, the preferred arrival time to the destination, and cost of the journey. The answers to these questions are used later in the questionnaire to make sure that the random attributes presented in the various choice situations fluctuate around values that the respondent is familiar with. The third and fourth introductory pages (figures 3.4, 3.5) demonstrate, both verbally and graphically, the idea of TTV and the way it is displayed throughout the questionnaire.

The main body of the questionnaire consists of nine choice situations, all of the same structure. In each of them, the respondent is asked to consider two alternative sets of travel conditions and costs, displayed in different colours, and choose one of them. In the inter-model version, the alternatives are entitled "Option A" and "Option B"; in the bus-only version they are entitled "Red bus" and "Green bus". Typical choice situations are presented in figure 3.6. The levels of the survey attributes (MTT, TTV, departure time and cost) are adjusted such that their averages are similar to the levels stated by the respondent in the introductory page. The MTTs of the two alternatives presented in each choice situation are chosen randomly within the range from 70% to 130% of the respondent's normal time. The TTV, i.e. the standard deviation of the journey times in each alternative, is generated through a random draw between 1 minute and 40% of the MTT of the same alternative. The fare is chosen randomly between 50% and 160% of the actual fare paid by the respondent. The five displayed times that define each distribution are determined as follows. The first is chosen randomly subject to lying no more than two standard deviations away from the mean (either over or under). The second needs to be within 1.5 standard deviations from the mean, and the third within one standard deviation. The forth and fifth are determined such that the predetermined MTT and TTV are kept. Departure times are drawn such that earliest possible departure allows the maximum trip length (out of the five displayed) before the desired arrival time, i.e. constitutes a conservative estimate of the worst possible travel conditions, and the latest possible departure allows the mean, i.e. constitute a realistic estimate. A set of constraints guarantees that in each choice situation, none of the two presented alternatives has dominance over the others in all variables. The constraints also verify

that no choice situation resembles one that has already been presented to the same respondent.

After the series of nine choice situations, a concluding question (figure 3.7) asks the respondent about his/her level of income. It is intended to use the responses to this question for checking whether there is risk of considerable bias, caused by potential insufficient representation of bus users with low income.

Responses to all questions and choice situations are automatically fed into a database. The survey program sends the responses to the databases immediately after every time the "continue" button has been pressed, and not only after the entire questionnaire has been completed. This is meant to enable using the answers of respondents that did not press the "continue" button at one of the last pages. It was decided to accept choices made by respondents that did not state their level of income, and also to accept responses from users who completed successfully any eight of the nine main choice situations.

In figures 3.2 – 3.8, the symbols at the upper corners of each page include links to the Leeds University website, the Institute for Transport Studies website, the "Frequently asked questions" page, and the email address of the author, respectively.

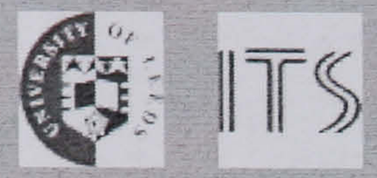


Figure 3.2: First introductory page

ITS

More info | Write to us

Before we go on, we'd like to ask you a few questions about your morning journey.

Even if some of the details change from day to day, please refer to a typical day.

The destination of this journey is the place where you ◉ **Work** ○ **Study**

At what time do you normally have to be there? 9 : 00

How much time (in average) does this journey normally take? 30 **minutes**

Which of the following modes of transport do you use:

☐ **car** ☐ **rail** ☑ **bus**

(if, for example, you use Park & Ride, you can tick more than one mode)

How much do you estimate that this single trip costs you? £ 2.00

(Please fill in the cost even if your employer pays for it.)

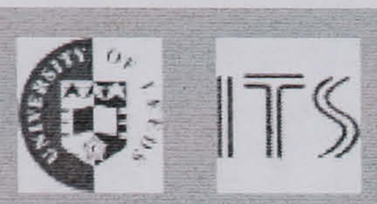
If you use bus or rail, the cost is the price of your ticket. If you buy a ticket for the whole day/week/etc., please try to estimate how much you pay for a single journey.

If you use a car, the cost might include petrol, parking, car maintenance and so on.

You must enter some cost in order to proceed.

CONTINUE

Figure 3.3: second introductory page



More info

Write to us

Normally, the time your morning journey takes on one day is not exactly the same as on another day.

There are many reasons why it is hard to know precisely when you will get to your destination.

In the following pages, we will ask you to choose between 2 alternative travel options.

We will tell you at what time you have to leave home for each of these options, and also how much they will cost.

But since we can't be sure about the exact journey time, we will show you 5 possible times.

CONTINUE

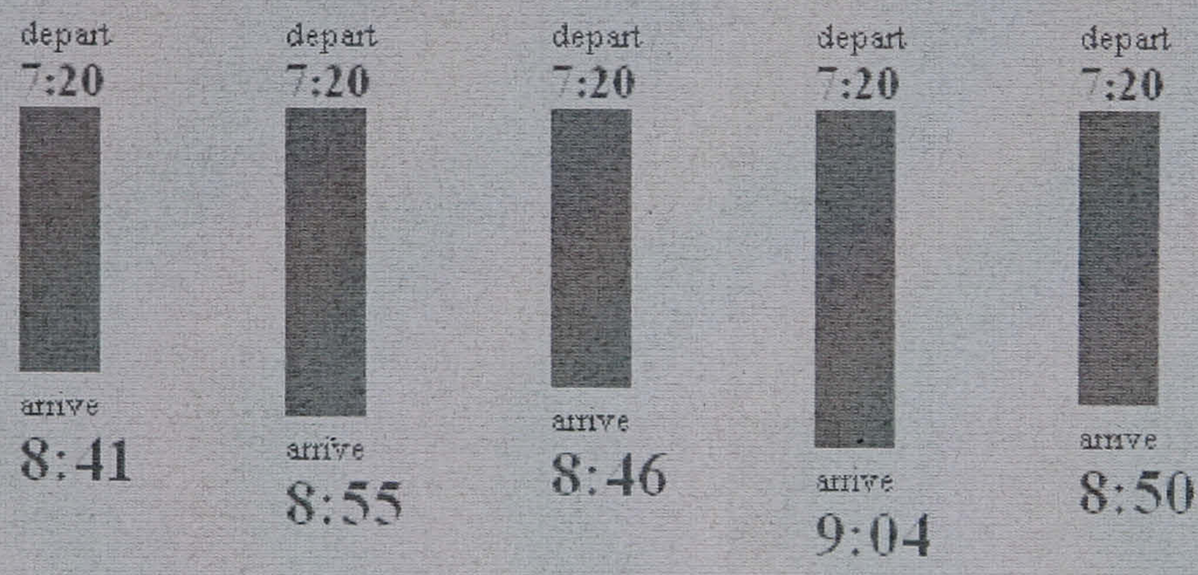
Figure 3.4: Third introductory page

ITS

More info | Write to us

Here's as example of how we will present the 5 possible times.

Your departure time from home is written above each bar. Your arrival time to the destination is written under each bar.



A longer bar means that your journey time is longer. You don't need to think about the time it takes to walk to or from the station, bus stop, or car park - it's already included in the presented time.

If you travel on 5 different days, you are likely to experience all these 5 journey times, but we don't know in what order.

In each of the following questions you will simply have to look at 2 such diagrams, and decide which one of them you prefer.

Are you ready to start?

CONTINUE

Figure 3.5: Forth introductory page

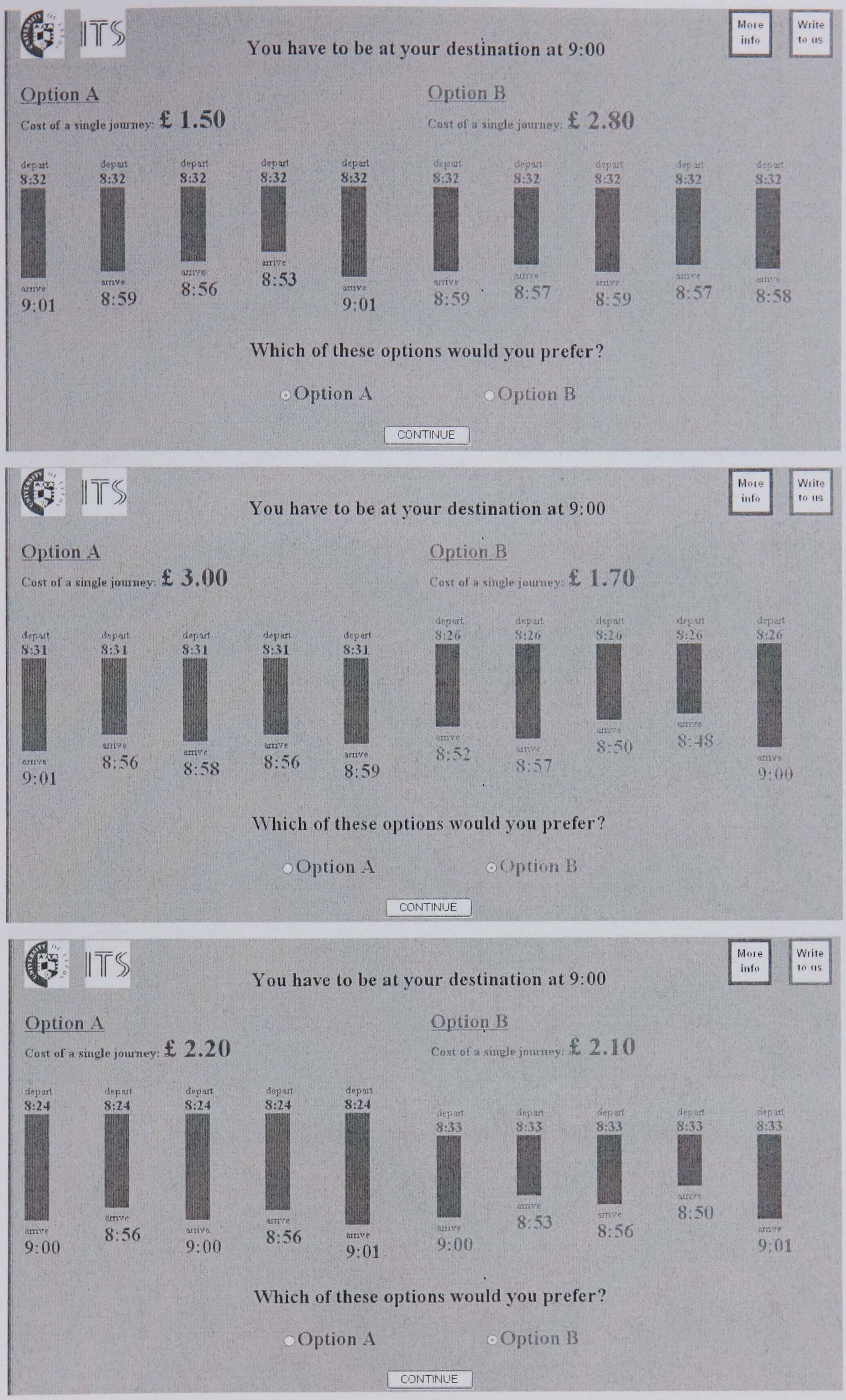


Figure 3.6: Typical choice situations

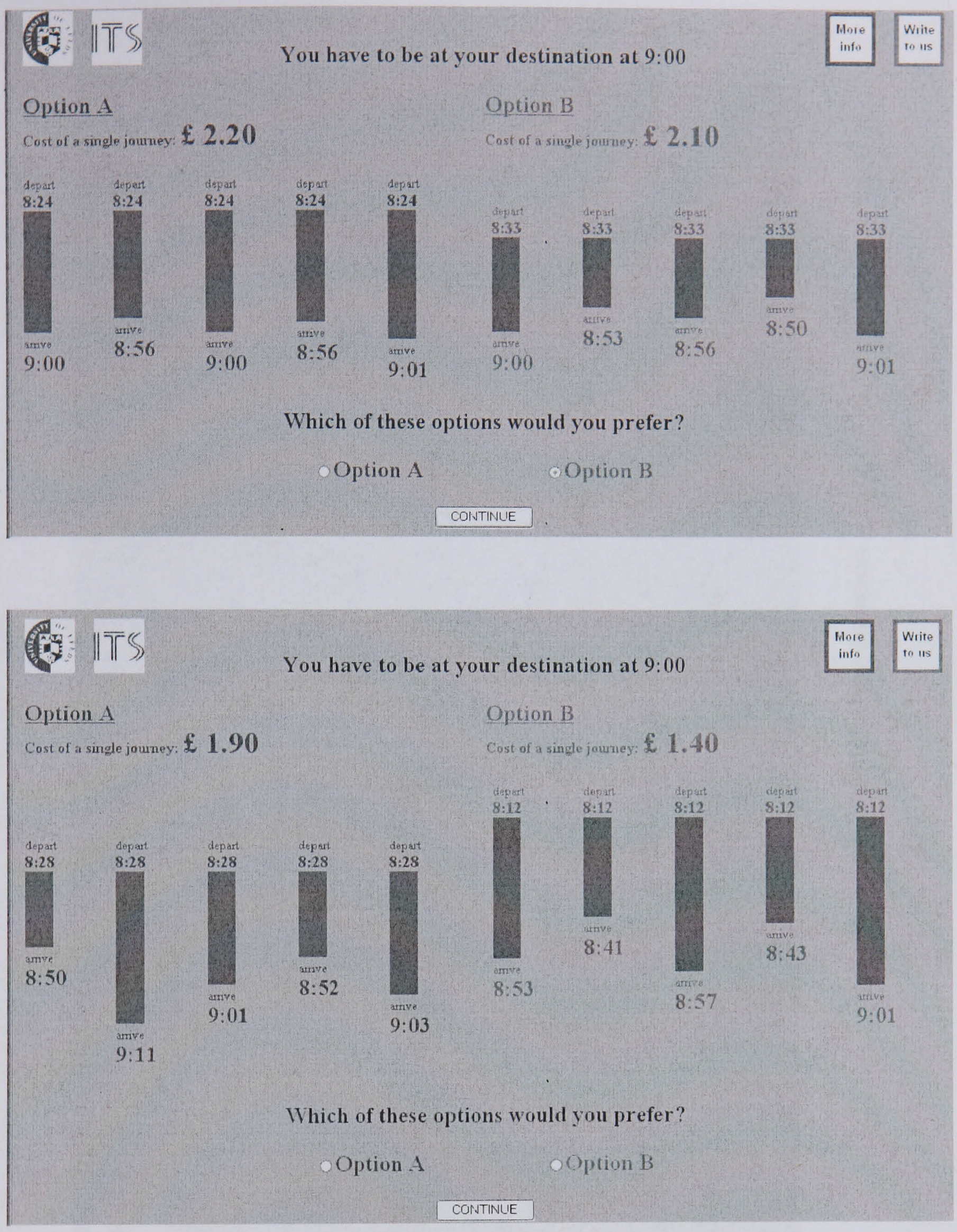


Figure 3.6 (continued): Typical choice situations

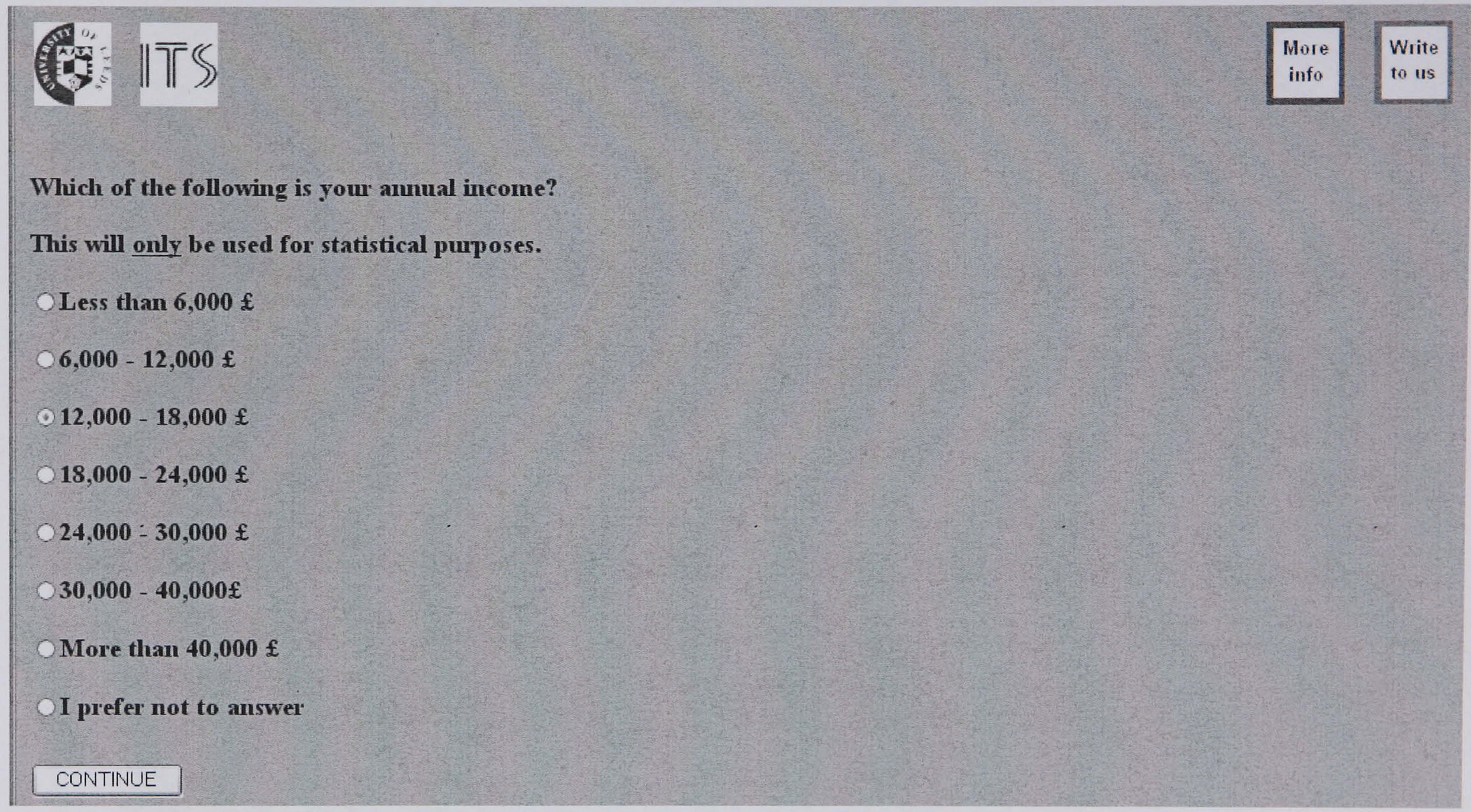


Figure 3.7: concluding question

ITS

More info | Write to us

Here are the answers to some frequently asked questions.

About this Survey

What is the purpose of this survey?

The survey will help us understand the influence of various factors on the way travellers choose how and when to travel to work (or to university). The main factor that we focus on is the day-to-day variations of the travel time. Understanding people's preferences is important in order to make sure that the money invested in improving the transport system is used in the most appropriate way.

Who conducts this survey?

The survey is part of an academic research, carried out by the Institute for Transport Studies at the University of Leeds. The analysis is performed by Mr. Yaron Hollander and supervised by Prof. Peter Mackie and Dr. Ronghui Liu.

Will my personal details be stored somewhere, or used again?

We are not asking for any identifying details. The only details sent to us when you fill in the questionnaire are your answers. We do not have your name or email address. We will not contact you again, unless you send us an email with questions or comments about the survey and ask us to reply.

I'm not satisfied with the level of public transport services in York. Will this survey lead to improvement?

We hope that at the end of our research, we will be able to suggest ways of improving your travel conditions. We will present our recommendations to York City Council, bus operators and so on. However, it is not expected that you will see the effects of this research in the immediate run.

Filling in the questionnaire

Sometimes I walk to work in the morning, and travel by bus on the way back. Can I answer the questions referring to my afternoon journey?

No. The factors that we analyse differ greatly between the morning journey and the afternoon journey. We would ask you to refer only to journeys made in the morning, before 10:30 AM.

I travel to work just two mornings a week. Would you still like me to fill in the questionnaire?

No. Since we focus on the issue of day-to-day variations in your journey time, we are mainly interested in the views of people who travel every day.

I use Park and Ride every day. Should I complete the questionnaire?

Yes. When you are asked about the modes of transport that you use, please tick all the modes involved in your daily journey (for example, car and bus).

I'm using two different buses on my way to work. How do I answer the questions about my journey time?

When we mention your departure time, we refer to the time you leave home, and when we mention your arrival time, we mean the time you get to your final destination (working place, university etc.). Everything between the departure and arrival is part of the journey, even if it includes changing buses on the way.

My travel expenses are paid by my employer, so it doesn't cost me anything. But the survey doesn't let me go on if the cost I enter is 0.

Figure 3.8: The "Frequently asked questions" page

We would ask you to enter your journey cost even if you don't pay for it yourself.

Will the train/bus company that I travel with know about what I'm writing here? I'd like to take this opportunity to make a complaint.

Bus companies do not take an active part in this part of the research, so this might not be the best way to contact them. Some respondents have emailed us various comments about bus and rail services in York. We will inform the companies that operate in York about these comments, but without mentioning the name of the person who wrote to us.

The way I travel changes from one day to another, so I have different travel costs on different days. But in the survey I can only enter one travel cost.

If you have a typical way of travelling on most mornings, please refer to it. But if your journey details vary greatly between days, please do not complete the questionnaire.

The times presented are not realistic – it actually takes much more (or less) to get to my destination. Why is that?

The scenarios that we present are entirely imaginary. You should think what would you do if they were real.

I do not have a fixed arrival time to work. How can I answer this question?

Simply enter the time when you prefer to arrive. If the exact time is not crucial to you, we will probably understand that when analysing the results.

In some of the questions the cost is much higher than in real life. I'm worried that this might lead to raising prices, because fares are very high anyway.

The results of this survey will not be used for considering changes in fares or other costs.

CONTINUE

Figure 3.8 (continued): The "Frequently asked questions" page

3.2.4. The pilot survey

The limited budget of the survey did not enable conducting a full pilot study. Since it was feared that there were not going to be enough responses from travellers in the study area, it was preferred not to "waste" potential respondents on the pilot survey. Instead, a limited pilot experiment was carried out outside the study area, featuring 35 students and members of staff at the University of Leeds. The participants of the experiment were asked to fill in the survey without being given any preliminary information; in addition, they were asked several additional questions informally, after completing the questionnaire.

The pilot experiment was used to examine the credibility of the survey design in the following ways:

- Unlike the questionnaire used for the main survey, as described above, the pilot experiment included several choice situations where one of the alternatives dominated the other in all attributes (at any level of WTP that seemed feasible). The responses from all 35 participants show that the dominating alternative was always chosen. This gives evidence that the presentation and the wording of the questionnaire are clear enough.

- The survey program keeps a record of the time it takes each respondent to fill in the entire questionnaire (including the time it takes to read the introductory pages). Among the 35 participants of the experiment, the shortest time was around 3 minutes, the longest was close to 15 minutes, and the median was around 5 minutes. This seems short enough to avoid any significant effects of tiredness or boredom.
- Some of the respondents were given a longer questionnaire with 20 choice situations and were asked, after completing it, whether they felt concentrated enough to give serious answers throughout the whole questionnaire. While some of them stated that they did feel they gave credible answers to all questions, some said that the repetition of the same task was palling after about 10 choice situations. This confirms the choice of nine choice situations per respondent as a satisfactory number to avoid loss of credibility.
- After completing the questionnaire, the participants were asked, in an informal manner, whether they were mainly following the numerical display of departure and arrival times, or the graphical display, or both. About a third of the participants said they started with reading all the information, but then they realised that the graphical display included most of the necessary information, and therefore towards the last choice situations they started referring mainly to the graphical display. These participants were still having an occasional peek at the numerical display, mainly if there were either very small differences between the two alternatives, or very big differences among the five typical days that form each alternative. Another third of the participants said they always got a first impression of the travel conditions from the graphical display, and then had a look at the presented numbers to confirm this impression. The other participants said they equally used the graphical and the numerical display. All these responses seem to support the assumption that the graphical display alleviates the need to process the relatively complicated presentation of a distribution of travel times.
- It is worth verifying that the different colours used for the display of the two alternatives in each choice situation did not have an effect on the actual choices. This is particularly important in the version of the survey that is used by bus users only, because the two alternatives are entitled “Red bus” and “Green bus”,

and these might be wrongly identified with specific bus operators, that the participants might like or dislike independently of the presented choice situation. To check this, a simple Logit model was estimated based on the experiment data, featuring all variables that are discussed in great detail later in this chapter, and an additional dummy variable that gets a value of 1 if the "Green Bus" alternative is chosen and 0 otherwise. If this dummy variable was found significant, it would mean that the participants have a consistent tendency towards (or against) one of the colours. However, the dummy variable was found most insignificant, indicating that there was no such bias.

All in all, it appears that the pilot experiment, despite its limited scope, confirms that the survey design is sound.

3.3. Survey results

The full-size survey, featuring respondents which are bus commuters to or in the city of York, was conducted from November 2004 to February 2005, excluding holidays. After sifting improperly or partially filled questionnaires, the database included 250 questionnaires. Most of the questionnaires include choices made in nine situations each, but in several cases, one or two choices per respondent were omitted from the database due to suspected errors. The final data file that was used for modelling included responses from 2165 choice situations.

In order to verify that the WTP derived from this sample can represent the true WTP, the income distribution of the respondents was compared to the general income distribution in the study area, as published on the Office for National Statistics website. It was found impossible to prove that a significant difference between the two distributions exists. The share of low-income respondents, that were feared to be insufficiently represented in the sample, seemed to be reasonably accurate. Many efforts were made to find additional information on the income distribution of bus users only, as they are obviously very likely to have lower income levels than the general population in the area. Unfortunately, no available data were found, and the survey-based income distribution could therefore not be compared to another source. Hence, the analysis presented here is based on the assumption that the participants of the survey do form a credible sample of all bus users in York. In the absence of robust evidence for

it, our judgment of how realistic this assumption is will be mainly based on paying special attention to whether the WTP estimates derived later seem too high.

Before using the survey results for modelling, an attempt was made to identify behavioural patterns that stick out without using advanced mathematical tools. First, it was checked whether there were travellers that consistently followed a simple decision rule throughout the entire questionnaire, i.e. always chose the alternative that had advantage over the other option in a specific attribute, ignoring all other attributes. It is found that 13% of the respondents choose the alternative with the lower fare in each of the choice situations, no matter what other apparent differences existed between the two alternatives. Minimising the cost was found the only simple decision rule: all other respondents (87%) respond to the value of more than one attribute when choosing their preferred bus service.

It was therefore examined which of the time-related attributes, namely all attributes apart from the cost, were most frequently used when choosing a bus service. This was done by noting in which attributes the chosen alternative in each choice situation was better than the alternative that was not chosen, and then checking which attribute appeared as a choice criterion in more choice situations than other attributes, among the nine choices of each respondent. If two or three attributes were used by a respondent more than the other attributes, and at the same frequency, a weight of 0.5 or 0.33 was ascribed to each one of them. The different choice criteria used by the survey respondents, and the share of respondents using them, are presented in table 3.1.

Choice criterion	**Share of respondents that used this as most frequent criterion**
Minimise MTT	46.7%
Minimise TTV	9.9%
Depart as late as possible	3.7%
Minimise mean lateness	25.4%
Minimise mean earliness	14.3%

Table 3.1: Criteria for choosing bus service

The last three rows of table 3.1 describe scheduling-related decision rules. They add up to 43.4% of the respondents, who tend to prefer travel alternatives that optimise common scheduling considerations, while only 9.9% follow the minimum TTV more frequently than other criteria. This is our first evidence that an indirect representation of TTV, through the amounts of lateness and earliness, enables better understanding of the effect it has on the behaviour of bus users.

3.4. The scheduling model

In chapter 2 we saw that it has not yet been investigated which set of variables can satisfactorily explain the reaction of bus passengers to TTV. Following the data collection described above, a series of modelling attempts was performed, aimed at founding a basic Multinomial Logit model on the survey responses.

A decision that had to be made concerning the model formulation, prior to the actual estimation, was whether travellers with fixed arrival times should be treated separately from travellers who have some flexibility in choosing when to arrive at their destination. The flexibility in choosing arrival time is crucial to the choice of departure time; in previous works, different authors interpreted this in different ways. Some works (Black and Towriss, 1993; Bhat and Sardesai, 2005; and others) distinguish between travellers that ascribe different levels of importance to on-time arrival, while most other researchers do not make this distinction. Models that discuss the cases of fixed or flexible arrival time separately have a better behavioural reasoning, since this is clearly a consideration that travellers take into account when choosing a departure time. However, such models are harder to use for prediction, because they require information about arrival time flexibility as input, and this is hard to find. Data about the level of flexibility are hard to obtain also because there are actually not only two strictly contradicting cases; in different working or studying places, a whole range of attitudes to late arrival exists. If we wanted to account for this accurately, we would have to either carry out extensive data collection or create a disaggregate model that predicts the level of flexibility, but both these options are not possible in the current scope. It was therefore decided to ignore the different levels of flexible arrival, and cover all of them in one model. Although this might at first glance seem simplistic, ignoring this issue

might actually help the real distribution of flexibility reveal itself more truthfully, as it is indirectly incorporated in every choice the respondents made in the survey.

All modelling attempts discussed in this chapter were carried out using the Alogit 4.1 and Alogit 4.2 packages. A sample of test models was also run with other software tools (Biogeme and Gauss) to verify their consistency. For more details on the software used, see chapter 4.

Many model specifications were attempted; we cannot present all of them here, but some interesting specifications and the estimation results are shown in table 3.2. In each cell of the table, the value in brackets is the t-statistic; and the value in italic print, in rows that stand for variables measured in minutes, is the WTP in pence per minute (ppm). An expected outcome of the modelling experiments is that the fare parameter and the MTT parameter are found the most significant; logically, these parameters have negative values. Several attempts were made to formulate a utility function that explicitly includes a TTV variable. Some models with a TTV variable were found very successful in terms of their final likelihood, but the significance of the TTV variable itself (as reflected by its t-statistic) was low, and we could not accept these as models with a satisfactory explanatory power. Insufficient significance of the TTV variable was found both in models where scheduling variables were included and in models where these variables were omitted. A single specification (model 3 in table 3.2) did lead to a model with a higher significance of the TTV variable, but the likelihood of the model as a whole was very low, and it was decided not to accept this specification, too. In contrast, the mean lateness (ML) consistently proved to make a significant contribution to the model; this led us to the understanding that bus users are concerned about the effect of TTV on the way they schedule their trip, and not about TTV per se. This resembles the conclusions of Noland et al (1998), Small et al (1999) and others regarding car users' preferences; it contradicts what Bates et al (2001) conclude as to rail users.

Many different forms of variables that denote the amount of lateness or earliness were probed; table 3.2 shows only some of these attempts. In addition to the ML variable, modelling attempts included the probabilities of arriving too early or too late by various amounts of time, expressed either in minutes or as a percentage of the MTT. Squared values of the mean lateness and earliness, and various combination where lateness and earliness are represented by more than one variable, were tried too as they were found a significant contribution to the model in other works (such as Small et al, 1999). The

conclusion from these experiments was that there is no statistical justification for introducing most of these variables, including some that do seem rational, such as the various variables that represent the size of the right-hand tail of the distribution of arrival time (e.g. probability of extra-late arrival).

Variables that represent earliness were not found as significant as the ML in any of our model specifications. It is clearly logical the travellers are less concerned about earliness than about lateness; but in a departure time choice model, the penalty on late arrival must be balanced by some penalty on early arrival, even if a small one. Travellers are unlikely to be indifferent to very early arrival, since this also implies very early departure. Due to the difficulty in specifying a model that directly penalises earliness, we specified a model in which the sum of MTT and the mean earliness constitutes one variable, which we denote MTE (model 7 in table 3.2); this model performed well statistically. It can be shown that MTE is always equal to the time from the departure till the moment with the mean lateness; this means that some correlation exists between the MTE and ML variables, since changes in the mean lateness affect both of them. However, since changes in the extent of lateness can only occur together with changes in travel time, the way MTE is affected by lateness differs between short and long journeys, while ML represents pure lateness independently.

The fact that the MTT and the mean earliness are covered in this successful model by a single variable is not a very strong indication that the WTP that corresponds to these two separate elements is equal. It certainly implies that the difference between the attitudes to earliness and to the MTT is likely to be much smaller than the difference between earliness and lateness (or between MTT and lateness). But what is more strongly understood from this experience with the MTE variable is that the amount of available data was not sufficient for making clear distinction between the MTT and earliness elements, whose contributions were apparently not very different from each other. By incorporating the MTE variable we probably merely helped the estimation tool identify, without introducing a massive bias, a model more easily. Indeed, a rational guess of the relationship between the penalties on MTT and on earliness would suggest that the penalty on MTT should be the higher of the two (this is also discussed in detail by Pells, 1987b). We test the logic in the MTE variable again in chapter 4, where the decision finally made is to leave its two components separate; but since the Multinomial Logit model with the MTE variable performs here better than other specifications, we leave it as is for now.

	Model 1	Model 2 (final mean-variance model)	Model 3	Model 4	Model 5	Model 6	Model 7 (final scheduling model)
Fare	-1.174 (-12.0)	-1.179 (-12.1)	-1.009 (-11.7)	-1.196 (-12.2)	-1.144 (-13.6)	-1.245 (-13.0)	-1.375 (-14.2)
Mean travel time	-0.0626 (-3.8) ***5.3***	-0.0821 (-12.3) ***7.0***	-	-0.0636 (-12.7) ***5.3***	-0.0586 (-13.1) ***5.1***	-0.0687 (-12.5) ***5.5***	-0.0717 (-11.5) ***5.2***
Mean earliness	0.0447 (1.6) ***-3.8***	-	-	0.0332 (1.2) ***-2.8***	-	-0.0601 (-0.3) ***4.8***	
Mean lateness	-0.1384 (-1.3) ***11.8***	-	-	-0.1935 (-3.2) ***16.2***	-0.1237 (-2.7) ***10.8***	-0.1594 (-3.6) ***12.8***	-0.1974 (-4.1) ***14.4***
Median travel time	-0.0010 (-0.1) ***0.001***	-	-0.0488 (-12.0) ***4.8***	-	-	-	-
TTV	-0.0103 (0.4) ***0.88***	-0.0077 (-0.5) ***0.65***	-0.0192 (-2.1) ***1.9***	-0.0118 (-1.3) ***1.0***	-	-	-
Probability of late arrival	-0.2030 (-1.0)	-	-	-0.2231 (-1.4)	-0.2240 (-1.5)	-	-
Probability of arriving more than 5 minutes late	-0.6042 (-0.9)	-	-	-	-	-	-
Probability of arriving more than 10 minutes later	-0.7771 (-1.2)	-	-	-0.5429 (-0.7)	-0.5665 (-0.8)	-	-
Squared mean earliness	-0.0014 (-2.5)	-	-	-	-0.0008 (-1.2)	-	-
Squared mean lateness	-0.0082 (-0.5)	-	-	-	-	-	-
Initial likelihood	-1534	-1534	-1534	-1534	-1534	-1534	-1534
Final likelihood	-1342	-1359	-1381	-1349	-1346	-1369	-1369

Table 3.2: Multinomial Logit models

(in brackets – t-statistic; in italic print – WTP in pence per minute)

Table 3.2 also shows an attempt to examine the potential effect of replacing the MTT variable with a variable that stands for the median of travel times. This was based on a hypothesis that bus users might be more sensitive to the median than to the mean. The median was generally found capable of replacing the MTT in the utility function, but the median-based model was less powerful than the mean-based model, and the median variable was therefore left out.

The scheduling model with fare, MTE and ML variables (model 7 in table 3.2) is therefore the best description we could obtain of the attitudes of bus users to TTV. To assess whether the final scheduling model and the WTP derived from it make good sense, it would be useful to compare them with the findings of other studies. But this is found rather difficult, as no previous work examined the same elements of WTP in a similar area at a similar time. Given that no other study is truly equivalent to the current one, the following is a general comparison of our findings to those reached in the only available studies that seem somewhat comparable. A recent study of the value of the MTT (namely the VOT) of bus users was carried out by Mackie et al (2003). They suggest several values that apply in various cases, all are around 3 ppm, in 1994 prices. The ratio of the Nominal Gross Domestic Product (GDP) per capita in the UK between 2004 and 1994 is 1.653 (£19461 to £11773 per person), and we would therefore expect the respective VOT of a bus passenger, at the time of our own survey, to be around 5.0 ppm. This is very close to the value of 5.2 ppm that our scheduling model implies.

The only study to which we can directly compare the estimates of the values that bus users place on earliness and lateness is Pells' work from 1987. Pells estimates the value of earliness at 1.5 ppm and the value of lateness at 7 ppm. The ratio of GDP between 2004 and 1987 is 2.634, and the respective values in 2004 prices would therefore be 4.0 ppm for earliness and 18.4 ppm for lateness. Although these estimates do not perfectly match ours, they clearly are in the same order of magnitude. Pells' estimate of the value of earliness supports the hypothesis made earlier, that our MTE estimate overvalues the penalty on earliness, although not to a great extent. The ratio of values of lateness and earliness according to Pells is 4.67, and according to our model it is 2.77. Despite the considerable difference, the comparison does strengthen the finding that lateness is much more heavily penalised than earliness. Unfortunately, Pells does not present VOT estimate and therefore it is not possible to compare all three WTP elements to his results.

The attitudes of public transport users to TTV were also studied by Bates et al (2001). This study focused on train passengers, and similar to Pells, it presents values for earliness and lateness but not for VOT. The authors founds that train users value earliness at 56 ppm and lateness at 113 ppm; it is not surprising that the order of magnitude of these values is significantly higher than that of our estimates for bus users, as it is well known that railway travel in the UK has a very different market from bus travel. Still, the fact that the ratio of the values for lateness and earliness is 2.018, gives some additional evidence that the relatively high penalty placed in our own model on lateness is plausible.

3.5. The consequences of using a mean-variance model

In chapter 2 it was mentioned that although most of the research about the effects of TTV on car travellers has found that these effects are best modelled using scheduling variables, applications of mean-variance models are more common in practice. The reason for this is presumably the difficulty in the implementation of scheduling models, which normally requires using simulation-based methods, and depends on detailed information about individual preferred arrival times. The modelling experience described above confirms that for bus users too, a model based on scheduling variables is more credible than a model where TTV is represented directly. It is thus important to understand the consequences of the common use of the less powerful model. To do this, we look at the differences between model 2 in table 3.2, which is a mean-variance model, and model 7, which is our scheduling model. The likelihood of the mean-variance model is higher, but as explained above, we could not accept this specification because of the low t-statistic of the TTV variable. Examining the WTP implied by each of the models shows that the mean-variance model ascribes a higher value to MTT and a very low value to TTV. It seems that in the absence of the appropriate variables, the mean-variance model associates most of the monetary effect with the only included variable that is found significant enough (i.e. the MTT variable), and by doing this, it undervalues the importance of reliability.

The following experiment aims at demonstrating the influence of using a mean-variance model on the assessment of TTV-related costs in a realistic scenario. The survey data file, which was used for model estimation, is used here again, but this time we only use

the answers to the introductory questions about the respondents' daily travel experience: their MTT, preferred arrival time and fare. The choices in the hypothetical situations are ignored, i.e. the survey data are merely used since they contain information about the true distribution of journey times and costs. The analysis included the following stages:

1. A spreadsheet was prepared, with a record for each of the 250 individuals that took part in the survey. The record for each person includes the MTT, preferred arrival time and fare of his/her commuting journey.
2. A random level of TTV was added to each record. TTV was not mentioned in the introductory page of the survey and therefore the data file did not include information about the real TTV experienced by each individual. The random values were drawn in a similar way to the levels of TTV generated in the survey, i.e. as a proportion of the MTT, with the same upper and lower boundaries. This range of values seemed a reasonable representation of real conditions, as described in chapter 2.
3. For each individual (j), 10 random travel times were drawn, assuming a lognormal distribution with MTT_j and TTV_j, i.e. the true mean and variance of the daily journey of this respondent. For reasons why a lognormal distribution was chosen, see chapter 5.
4. For each individual, 150 evenly-spaced (non-random) feasible departure times were generated. The earliest departure for individual j is $2 \cdot MTT_j$ before his/her preferred arrival time, and the latest is $0.5 \cdot MTT_j$ before the preferred arrival time.
5. For each feasible departure time of each individual, ML and MTE were computed, based on the 10 random travel times. The ML and MTE were added to the record of each individual.
6. For each individual, the optimal departure time (of the 150 alternatives) was chosen deterministically, as the one that maximises the utility function of the scheduling model. Note that it is assumed here that as our model implies, actual choices are best explained by the scheduling model, even if the cost might be calculated (in stage 7 right away) according to the mean-variance model.

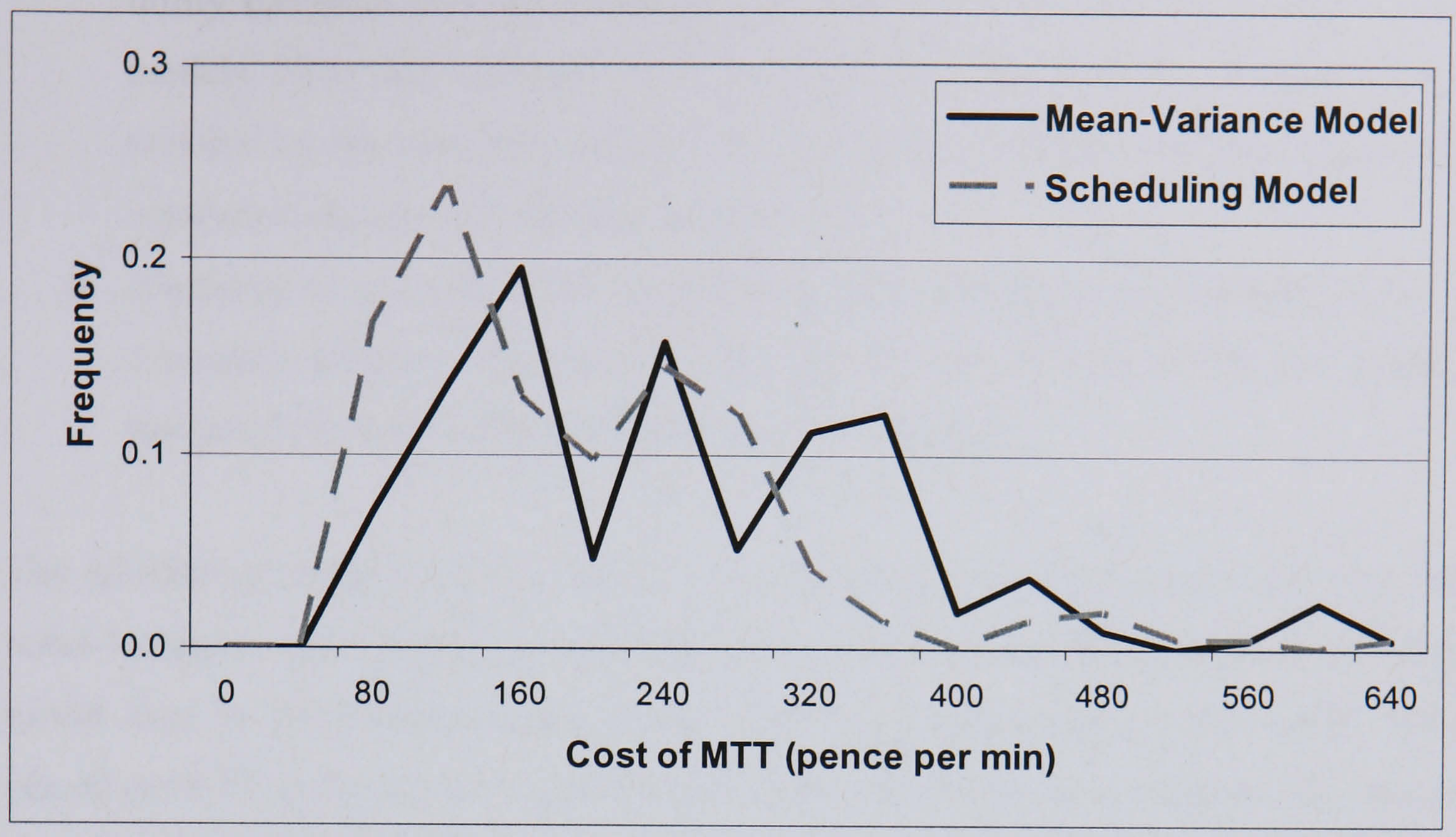


Figure 3.9: The cost of MTT according to the scheduling and the mean-variance models

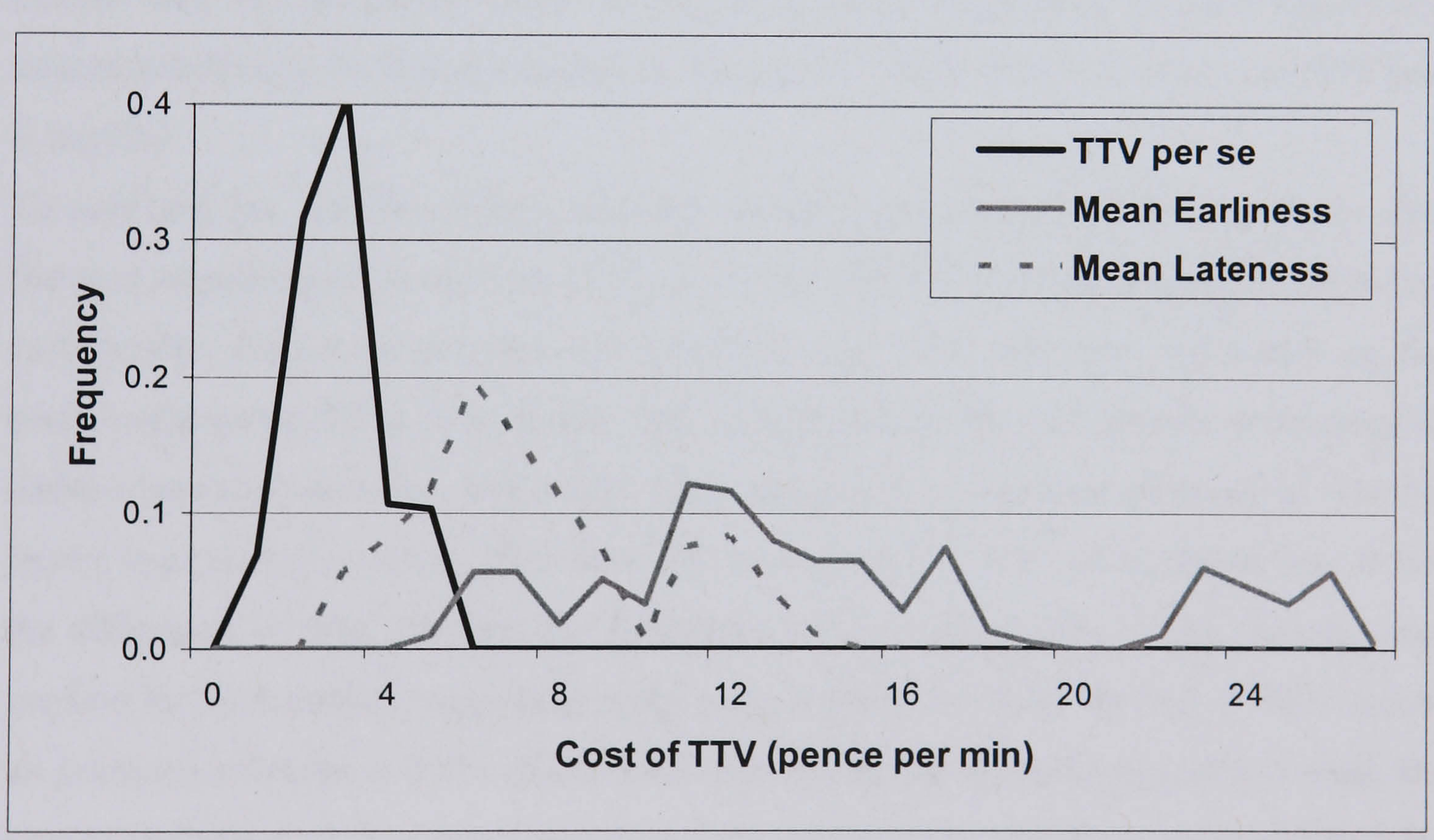


Figure 3.10: The cost of TTV according to the scheduling and mean-variance models

7. Based on the chosen departure time of each individual, the cost of each of the utility elements was calculated by both the scheduling and the mean-variance models. Note that although MTT and the mean earliness are included in one variable in the scheduling model, in the current analysis they are considered separate elements with the same cost per unit.
8. Elements of the total cost, according to both models, were analysed using a frequency diagram. Figures 3.9 and 3.10 show how each of the two models interprets the cost of the analysed sample of journeys.

The scheduling model curve in figure 3.9 is slightly more skewed to the left than the mean-variance curve; high cost of MTT is less frequently implied by the scheduling model than by the mean-variance model. This is a direct result of the higher value placed on MTT in the mean-variance model, and is therefore not surprising. The picture painted in figure 3.10 is much stronger: if we accept the aforementioned argument that the scheduling model is more powerful, figure 3.10 shows that the mean-variance model immensely undervalues the effects of TTV. If the relations between mean-variance models that are practically used for transport scheme appraisal to their equivalent, unused scheduling models are similar to the relation disclosed here, then a massive bias is implied.

To conclude the current analysis, the total journey cost across all 250 trips in the data file was summarised, as well as the cost of only the TTV-related elements. There is a considerable difference between the models in the total cost: the cost based on the mean-variance model is 21% higher. But even if one of the two models overvalues or undervalues the total cost, this might apply equally to all scheme alternatives that the model is expected to asses. Therefore, the more critical source of potential bias is not the difference in total cost but the difference in the composition of the various costs implied by each model. According to the mean-variance model, the cost of TTV across all journeys amounts to 1.0% of the total cost. When the scheduling model is used, the total contribution of the mean lateness is 3.8% of the total, and the mean earliness adds up to 7.2%, hence the indirect cost of TTV is 11.0% of the total cost. This means that when we compare alternative investment schemes, the mean-variance model is likely to be very insensitive to differences between the alternatives in terms of their effects on TTV, while the scheduling model does exhibit such sensitivity. Since we find the statistical significance of the TTV variable in the mean-variance model inadequate, we

conclude that the scheduling model should be preferred owing to its ability to distinguish between "more reliable" and "less reliable" alternatives.

3.6. Conclusions

The main contribution of the presented experiments is the estimates of values of MTE and ML, as the attitudes of bus users to TTV have not been discussed elsewhere in recent literature in an economic context. We find that the values placed on the mean time and the mean earliness are 5.2 ppm, and the value placed on ML is 14.4 ppm. We have raised the suspicion that the value of earliness might be slightly overestimated, but overall, judging by common sense and by comparison to other studies, the derived values seem plausible. The fact that the late arrival is heavily penalised sheds some light on a major element in the attitude of bus users to TTV.

On a broader view, our results show that the effects of TTV should be converted into monetary terms *indirectly*, through analysis of the consequent pattern of lateness and earliness. Models with scheduling variables are not the easiest to implement, because they rely on disaggregate input data; but it is found here that monetising the effects of TTV using the alternative, mean-variance approach, is inappropriate since it leads to a serious underestimation of the importance of TTV. The presented experience with a mean-variance model should be therefore treated as a warning to practitioners who use models where TTV is assumed to have a direct cost.

Along the way towards the estimates of the WTP, several other issues were brought up. These include some ideas about the formulation of a departure time choice model and about survey design, particularly with respect to the use of a computerised survey and to the graphical presentation of the idea of TTV.

The values derived here can be used to analyse whether investment in improved bus infrastructure brings considerable reliability benefits. But we choose not to do so before we extend the analysis of the same data, and try to reveal variations in tastes and preferences between individuals, which have not been considered here. This extended analysis is presented in the next chapter.

Chapter 4

The distribution of the willingness to pay

4.1. Introduction

This chapter describes the attempts to extend the estimates of the WTP, obtained earlier in the thesis, to estimates of the entire *distribution of the WTP* (i.e. the DWP). The collection of data on the attitudes of bus users to TTV, as well as some Multinomial Logit modelling experience, were presented in chapter 3. The analysis in the current chapter starts from the point reached at the end of the previous chapter: it uses the same data for the same purpose, but it tries to account for variation in preferences between the travellers, and thus to represent the true WTP more accurately, using other techniques.

The first section of this chapter reviews previous studies that inspired our DWP analysis. Since obtaining credible estimates is found here a challenging task, the other sections try to tackle the estimation of the DWP from a number of perspectives. First, several specifications of a Mixed Logit model are presented. Then, the use of a sub-sampling technique as an additional tool for testing model fit is considered and illustrated. Subsequently, the DWP is investigated in an experiment that is not based on the parameters of a choice model, and thus attempts to avoid unnecessary assumptions. A concluding discussion tries to decide which of the different estimates of the DWP should be used in forthcoming chapters.

It is worth mentioning that unlike chapter 3, which presented econometric analysis using well-established methodologies, this chapter deals with a field of transport economics whose exploration has only just begun. The publications mentioned here in the context of the estimation of the DWP are very recent; many of them were made public when the analysis presented here was almost complete. As before, our main interest is in the evaluation of benefits from improved bus TTV, but the undiscovered nature of the techniques required to reliably estimate the DWP shifts the main focus in this chapter from the phenomenon of TTV to the econometric technique itself. Nevertheless, as the forthcoming sections disclose, the technical difficulties described here are closely related to the essence of TTV, because the attitudes towards TTV are not captured in a single variable, and the key source of complexity is therefore the fact that these attitudes compose a multi-dimensional WTP space.

4.2. Literature on the distribution of the willingness to pay

4.2.1. Mixed Logit models: strengths and weaknesses

The current section introduces the techniques used in the various experiments presented later in the chapter. This sub-section introduces the Mixed Logit (MXL) model, and the subsequent sub-sections present some alternative nonparametric approaches and some relevant software tools.

The idea that the WTP is heterogeneous, namely that different travellers have different levels of WTP, has gone through several phases in the last decade. Traditionally, it has been very common to derive the WTP from the utility function of a Multinomial Logit model; in such function there is a single value for each parameter, uniformly distributed across all travellers, and therefore the WTP is homogeneous too. To account for differences in the WTP between the users of different travel modes, or between journeys made for different purposes, it has been common to calibrate separate models (or separate utility functions) for each. It also became common practice to segment the studied population into smaller groups according to their income level or other socio-economic characteristics. By estimating a different model for each group, it is possible to find several different levels of WTP without abolishing the traditional methods. This traditional approach does not allow for variation in the WTP that results from taste variation among travellers or from any other traveller characteristics that there is no explicit information about. This makes them a rather simplified representation of the true WTP, since it is impossible to include in the model all variables that have effect on individual WTP.

In recent years, significant attention has been given to models that allow random variation in individual preferences, independently of whether the causes for this variation are explicitly modelled. This increased popularity was gained thanks to the development of simulated-based techniques for the estimation of MXL models and other similar models with random parameters. The foundations of MXL are extensively discussed in the literature (Train, 2001; Hensher and Greene, 2003; Bhat, 2001; Batley et al, 2001; Walker, 2002; and many others); the following is only a brief description of its general form. MXL models generally use the utility function $U_{njt} = \beta_n \cdot X_{njt} + \epsilon_{njt}$, whose value varies between travellers (n), choice alternatives (j) and choice situations (t). X_{njt} is a vector of explanatory variables, and ϵ_{njt} is an error

term with an iid-extreme-value distribution. β is a vector of preference parameters that varies over travellers with $\beta_n \sim D(\theta)$, where D is the set of distributions of the elements of β (each parameter of the utility function can have a different distribution) and θ is a vector of the parameters of these distributions. Conventionally, θ includes the mean and variance of each distribution, although θ might also include boundary values or other constraints, as we discuss later. Note that the terminology that distinguishes between D and θ does not appear in this form in existing literature; it is made here to clarify ideas expressed in the forthcoming paragraphs.

Theoretically, MXL models seem to meet the need for tools to estimate the range of preferences, and hence the distribution of the WTP, in a heterogeneous population. However, the estimation is sometimes more challenging than it seems at first glance. Note that which parameters are included in θ, and whether or not it includes any bounds or constraints, are conventionally considered part of the specification of D, *not* the specification of θ. Determining θ only involves finding the best values for a predetermined set of parameters. A serious difficulty lies in the fact that the available tools for estimating MXL models perform evaluation and statistical analysis of θ, but a specification of D is required as input; the common practice is thus to *choose* D (not systematically) and then *estimate* θ (systematically). Another difficulty is that the estimation tools mainly use the same measures of statistical fit that have been traditionally used to assess the performance of models with fixed parameters. This is only natural, as there has so far been no apparent need for new statistical measures. But some recent experience (Hess et al, 2005; Train and Weeks, 2005; Sørensen and Nielsen, 2003) suggests that MXL models that are deemed successful by those measures do not necessarily exhibit good performance.

Several researchers have tackled problems relating to the choice of D or to the assessment of MXL model fit (although this is not always the main focus of their discussions). Hess et al (2005) illustrate why caution should be taken when specifying the parameter distributions, especially when the parameters are later used to derive the WTP. The authors analyse the case where the model ascribes positive parameter values to some of the population; this often occurs when the estimated mean is negative but relatively close to zero. Positive parameters lead to negative WTP, and the authors stress that this contradicts the economic theory of time valuation; they state that the positive values should not be seen as evidence that some travellers truly wish to experience longer journeys. The paper also discusses the use of bounded distributions to

avoid the excessive share of positive parameters. Using distributions with fixed bounds is generally found inappropriate because it might force the estimation procedure to yield flawed results. The recommended solution is to use bounded distributions where the bounds are estimated from the data; some successful experience with Johnson's S_B distribution is reported.

The discussion by Hess et al (2005) concerning the choice of distribution introduces valuable insights, but some expansion of its scope is required. A first reason for this is that the analysis of the difference between the true and the modelled distributions is mainly based on comparing the 95 percentiles, i.e. examining the tails of the distribution; it should be important to examine the differences using other two-sample tests. Second, the *true* behaviour that Hess et al try to replicate is made up artificially; it is not clear whether the examined distributions perform similarly with data representing real behaviour. Third, further discussion is needed about how we can know more about the true range of traveller preferences and about whether a certain specification of a MXL model fits the true behaviour.

Not all works on the distribution of the WTP are equally rigorous about positive values of time-related parameters. As mentioned earlier, Hess el al (2005) recommend treating such values with suspicion. Similarly, Batley et al (2001) find a positive tail in the distribution of a mean lateness parameter, and decide as a result not to allow any distribution of this parameter. In contrast, Cirillo and Axhausen (2004) state that it is acceptable that a small share of a population does not value time savings or would even rather extend the journey. They bring evidence of the existence of such preference, and also cite other studies with a similar finding. In a model presented by Bhat and Sardesai (2005), positive travel time parameters are attached to 27% of the population, and the authors do not see this as a reason to reject the model.

Train and Weeks (2005) and Fosgerau (2005) examine modelling in *WTP space* as an alternative approach for estimating the distribution of the WTP; this is based on a concept originally introduced by Ben-Akiva et al (1993), although it has so far been uncommon in transport analysis. Instead of estimating parameters of a utility function and then deriving the WTP as a ratio, this concept directly estimates individual WTP values and the cost parameter, and can then derive the other parameters by multiplication. One of the appealing features of modelling in WTP space is that it avoids the need to calculate the WTP as a quotient of two parameters; as illustrated later in this chapter, the seemingly-straightforward division of two numbers can cause

serious difficulties if the distribution of the cost parameter, which is used as the denominator, includes values close to zero. Formally, modelling in WTP space and the more traditional modelling in parameter space should be equivalent, but in practice this is compromised due to the distributional assumptions made prior to estimation. If, for instance, we estimate a model in the parameter space and assume that the parameters distribute lognormally, the resulting distribution of the WTP will have the quite complicated shape of a ratio of two lognormal curves; if the model is estimated in the WTP space and lognormal assumptions are applied to the WTP, it is then the parameters and not the WTP that have a more complicated distribution. Train and Weeks (2005) find that estimation in WTP space leads to better estimates of WTP, although some common statistical tests fail to detect this better fit. However, although this concept opens an interesting avenue as an alternative modelling approach, note that some theoretical and practical issues still need to be tackled before modelling in WTP space can be treated as an available approach for model estimation. These issues include investigation of the performance of different statistical distributions in WTP space and assessing the fit of utility parameters when they are derived from estimates of WTP. Note also that the software tools used in the current study (as described in section 4.2.4) do not enable estimation of models in the WTP space; this made it impossible for us to focus on WTP estimation based on this concept within the scope of this work.

4.2.2. Sub-sampling techniques

Later in this chapter we present experiments that use a method which is part of a bigger group of methods, called sub-sampling (SUS) or re-sampling techniques. SUS is employed here in an attempt to obtain an external estimate of the distribution of the WTP, independently of the MXL model. SUS methods are typically applied to a dataset when additional statistical analysis is required but there is no further data; the analysis is thus performed using subsamples of the available dataset. Cirillo et al (2000) explain the essence of SUS, denoting the observed data records in the full original dataset by $(x_1, x_2, \ldots, x_n)$. When the full dataset is used to compute any statistic (e.g. mean, variance, parameter in a MXL model, etc.) these records are normally given trivial weights, i.e. $(1/n, 1/n, \ldots, 1/n)$, and the statistic is computed once. SUS replaces the trivial weighting with other weighting rules, and the single computation of the statistic is replaced with an aggregation of its value over multiple evaluations.

The procedure used by Sørensen and Nielsen (2003) is similar to the SUS technique known as Bootstrap. Bootstrap involves generation of multiple small subsamples based on the original full sample; subsample i with k_i data records can be generated by random weighting rules such as $(1/k_i, 0, 0, 1/k_i, \ldots)$, $(0, 1/k_i, 0, 0, \ldots)$, etc. The distribution of values of the statistic of interest across all subsamples is the Bootstrap estimator of the distribution of this statistic in the full dataset (Cirillo et al, 2000). Sørensen and Nielsen generate a large number of small subsamples by repeatedly dividing the full sample into random parts of similar sizes. Then they estimate a Multinomial Logit model for each subsample, and the distribution of model parameters across these sub-models is their estimate of the real parameter distribution.

The procedure described by Hensher and Greene (2003) resembles the SUS technique called Jackknife. Each Jackknife subsample is created by removing a single record (or all r records that come from the same respondent) from the full dataset, using a weighting rule of the form $(1/(n-r), 1/(n-r), \ldots, 0, \ldots, 1/(n-r), 1/(n-r))$. If a Multinomial Logit model is estimated for each subsample, it is possible to learn about the parameter distribution in the original dataset through sensitivity analysis.

Cirillo et al (2000) compare the performance of Bootstrap and Jackknife and present advantages and disadvantages of the two approaches; the general impression is that the two perform largely the same. In an experiment with up to 40 subsamples, the Bootstrap estimate of the sample variance shows better convergence, but appears slightly biased downwards. The authors confirm, though, that Bootstrap is a more direct method for trying to replicate the distribution of the original dataset, whereas Jackknife mainly tests the sensitivity of the parameter estimates.

The SUS experiments later in the chapter are inspired by the studies of Sørensen and Nielsen (2003) and Hensher and Greene (2003); these studies use SUS techniques for the same purposes as here, but there are several reasons why it is necessary to extend their scope. First, these previous works do not discuss the fact that the techniques they use are private cases in the broader group of SUS techniques. Second, Sørensen and Nielsen do not clarify whether (or how) they avoid the bias caused by the fact that multiple responses in the database come from the same survey respondent (unlike Cirillo et al, 2000, that do elaborate on this bias but in a different context). And third, SUS has so far been used *prior* to model estimation, to examine which specification of the parameter distributions (namely, using the terminology defined earlier, which specification of D) seems the most appropriate for the model. The aforementioned

works ignore the fact that the when the most suitable D (say D_0) is then used for the actual estimation of the model, the parameters θ are re-estimated. The vector of parameters θ_0 that was used in the SUS experiment to choose D_0 is replaced with another vector, θ_1, determined by the estimation tool. There is no reason to assume that $\theta_0 \cong \theta_1$, or that a model specified with $\beta_n \sim D_0\ (\theta_1)$ will perform similarly to the model that had good fit with $\beta_n \sim D_0\ (\theta_0)$. It might even occur that once θ has been re-estimated, the model with $\beta_n \sim D_0\ (\theta_1)$ is inferior to another model, where the specification of the parameter distributions is some unknown D_1. The experiments in presented later try to deal with this difficulty by applying the SUS technique as an additional statistical test, *after* the estimation of the MXL model.

Later in this chapter we discuss to what extent SUS can be reliably used to estimate the DWP. Although we do use it as a key method in the search for the best estimates of the DWP, it should be stressed at this early stage that when the level of variation of some parameter is determined using SUS, the estimate is inevitably biased downwards by the very definition of this technique. Since the subsamples always include several individuals, the preferences of individuals with extreme characteristics will always be balanced by those of individuals in the same subsample whose behaviour is more common. Hence, the far edges of any distribution cannot by captured by its SUS-based replica. However, as we discuss towards the end of this chapter, all methods for estimating the DWP have deficiencies, and a distribution might be considered a relatively good estimate despite an inability to reproduce these far edges.

4.2.3. Nonparametric estimation of the distribution of the willingness to pay

The common way of estimating MXL models is using the method of maximum simulated likelihood. Denote y^{nt} the choice that traveller n makes in choice situation t. y^n is the sequence of choices that traveller n makes throughout the entire set of choice situations. Conditional on the set of parameters $\beta_n \sim D\ (\theta)$, the probability that a particular y^n is chosen is (Based on Train, 1999; Train, 2001; Hensher and Greene, 2003):

(4.1)

$$P\ (y^n \mid D, \theta) = \int L\ (y^n \mid \beta_n) \cdot g\ (\beta_n \mid D, \theta)\ d\beta$$

Where

(4.2)

$$L\left(y^n \middle| \beta_n\right) = \prod_t \frac{e^{\beta_n \cdot X_{ny^n t_t}}}{\sum_j e^{\beta_n \cdot X_{njt}}}$$

and $g\ (\beta_n \mid D,\ \theta)$ is the probability density function that describes the distribution of β_n. The maximum log-likelihood is estimated by simulation because the integral in (4.1) does not have a closed form. The simulated likelihood is calculated using random draws of β from the density function $g\ (\beta_n \mid D,\ \theta)$; the value of $L\ (y^n \mid \beta_n)$ is calculated separately and the results are averaged over all draws. This simulated probability is denoted P^*, and the simulated log-likelihood is:

(4.3)

$$SLL\ (\theta) = \sum_n ln\ P^*_n\ (y^n \mid \theta)$$

The parameters θ are optimised by maximising this expression. The maximum likelihood method required preliminary knowledge of D, or in other words, of the general form of the function $g\ (\beta_n \mid D,\ \theta)$; it therefore belongs to the broad family of *parametric* techniques. The information that parametric techniques require as input is often unknown, and therefore various assumptions need to be made. *Nonparametric* techniques are those that try to determine the estimates of interest without making such assumptions. The SUS techniques mentioned earlier are all nonparametric, but there are many other nonparametric methods (described for instance by Yatchew, 1998).

It has been known for several decades that nonparametric methods can be used as an alternative to Maximum Likelihood estimation (see for instance comments by Ben-Akiva and Lerman, 1985). However, hardly any attempts to take this idea forward can be found in the transport literature. The few abovementioned studies that use SUS do illustrate the general concept of nonparametric estimation, even if SUS clearly has drawbacks, as discussed further later. Thorough study of the case for nonparametric estimation of the distribution of the WTP is beyond the scope of the current study; however, a very simple nonparametric experiment is conducted later, in an attempt to reveal what information truly comes from the input data, as opposed to the information that originates in our own assumptions.

The nonparametric experiment performed later is inspired by two previous works from very different points in the history of WTP evaluation. The first inspiring work is Beesley's discussion of the calculation of WTP (Beesley, 1973), that was published

before it became common practice to derive the WTP from choice models. Beesley is not interested in the distribution of the WTP, only in the average VOT. He shows that in a survey where each respondent is asked to make a choice between hypothetical services, we can identify the *accepted maximum* and the *rejected minimum* of the WTP. Stating that "the idea of acceptance or rejection suggests a method for deriving the measure of central tendency", Beesley searches for the single value that fits the entire sample best. The optimal WTP is determined by "minimising misclassifications of observations", i.e. by checking how many choices in the sample can be explained by every feasible WTP, and then choosing the value that can successfully explain the biggest share. The experiment described later in the chapter demonstrates that despite the remarkable improvements in choice modelling since Beesley's work, what can be deduced from survey responses about the range of individual WTP, without relying on additional assumptions, is not significantly different from Beesley's accepted/rejected boundaries.

The second work that inspired the simple nonparametric experiment is Fosgerau's recent investigation of the distribution of the VOT (Fosgerau, 2006). This is the only work (known to the author) that explicitly describes nonparametric methodology for computing the WTP; this is done without estimating a MXL model. The analysis examines, directly from a dataset collected in a stated choice experiment, for which responses it is possible to identify the WTP that led to the choice. The mathematical and statistical tools used for the nonparametric regression include the Klein-Spady estimator, Zheng test and other advanced features; these require high mathematical proficiency and are not used in the current study. An important finding of Fosgerau's analysis is that the WTP can identified for 87% of the examined population without requiring distributional assumptions. For the remaining 13%, additional input information is necessary.

It should be noted that the work by Fosgerau makes a first important step in a direction that deserves significant further development. Fosgerau uses a dataset from a very simple survey, where respondents are asked to trade off between time and cost only; there is therefore just one element of WTP (namely the VOT). As discussed later in this chapter, the number of WTP elements influences the ability to identify a considerable share of the distribution of the WTP. It can be expected that if 13% remain unidentified in the analysis of VOT only brought by Fosgerau, then in models where several variables have a monetary value, a bigger part of the distribution will not be revealed

directly but only with the help of various assumptions. The hypothesis that in many practical circumstances more than 13% of the distribution of WTP will remain unidentified is also enhanced by the fact that the dataset that Fosgerau uses is almost ideal in terms of its size (17,000 responses from 2000 respondents) and in the sense that the panel nature of the data is ignored.

4.2.4. Software issues

Several software packages are available for estimating MXL models. The differences between the packages relate not only to the user interface, file syntax and so on, but also to their modelling concepts and capabilities.

Alogit is a friendly and powerful tool, which supports most model structures from the Logit family, and is very commonly used worldwide. Alogit has two disadvantages in the current context. The first is that version 4.1, which is currently the most common, does not enable lognormal (and other asymmetrical) parameter distributions. This is partially solved in version 4.2, which is still in limited distribution, but was kindly made available to use in this work by its developer. The second disadvantage is that when a MXL model is estimated with panel data, i.e. with several responses from each respondent, Alogit might ascribe different parameter values to the different responses of the same respondent; this leads to biased estimates, as explained by Cirillo et al (2000). Alogit does feature a module that reduces this bias using the Jack-Knife technique. This is satisfactory in most circumstances but not for the needs of this thesis, as we compare the MXL estimation results with an external estimate based on SUS, and it seems inappropriate to use a MXL model which in itself is based on SUS.

Another software package is Biogeme (see Bierlaire et al, 2004). Biogeme lacks some of the friendly traits of Alogit but has some other merits. It is distributed freely on the internet, and enables greater flexibility in the model specification, including non-linearity and a direct option of modelling with panel data. Unfortunately, some of the model specifications presented in the following chapters were tried with Biogeme and the estimation process did not terminate successfully.

A third optional tool is a code developed by Kenneth Train (see Train et al, 1999). The code is distributed freely on the internet, but must be used with Gauss, which is a commercial package. Many of the capabilities of Alogit and Biogeme are not possible using the code, but it is purposely designed for estimating MXL models with panel data.

The code was therefore found very convenient for the current needs, and it was used for all MXL estimates presented here.

For the Multinormial Logit models that form the sub-models in the SUS experiments, mainly Alogit was used. Various test models were also estimated using other tools; they generally verified that if convergence is reached, then the same specification leads the different tools to very similar results. For the sake of clarity, we refer to all software used in this chapter as *the estimation tool*.

4.2.5. Summary

There is no doubt that different travellers have different levels of WTP, and that it is very hard to include all the factors that influence individual WTP in any choice model. But unlike the well-established methodologies for calculating the average WTP, methods for estimating the distribution of the WTP, independently of its sources, are still in an early state of development. MXL models, which allow random variation in individual preferences, have attracted significant attention in recent years, as they are powerful, flexible and relatively simple to estimate. Nevertheless, some challenges in specifying MXL models, and especially the need to predetermine the family of statistical distributions for the model parameters, have raised doubts in several recent works about the credibility of the derived distribution of the WTP.

It is important to observe that the growing popularity of MXL model has not been so far accompanied by development of new tools for testing their statistical fit. As illustrated later, it seems that there is a real need for such tools. In their absence, it is not always clear how to interpret outputs such as positive parameters (and hence negative WTP) for a certain share of the population. It appears that SUS and other nonparametric methods can be used to obtain alternative estimates of the distribution of the WTP; but this has hardly been done so far in practice.

4.3. The Mixed Logit models

One of the main properties of MXL is that it allows for random variation between individuals in the values of the parameters that represent attitudes and preferences, and it is therefore a natural basis for estimating the DWP. This section describes the attempts to specify and estimate several MXL models based on the same dataset that

was used for the Multinomial Logit models. The scheduling model created in chapter 3 is the starting point for the current modelling experiments; the same variables as in the final model are used here, namely the fare paid for the bus journey, the MTE (sum of the mean travel time and mean earliness to the destination) and the ML (mean lateness). In all the models presented in this section, variation among travellers was found in the fare and ML parameters, but there was not sufficient evidence of variation in the MTE parameter. Leaving one parameter fixed is a common practice in MXL modelling, and it has therefore been decided to leave the MTE parameter fixed.

It was explained earlier that the specification of MXL models, prior to the estimation of the parameter values, involves not only choosing the relevant variables but also determination of the general shape of the distributions of the parameters. The model specifications that were probed here used some of the most common distributions for the fare and the ML parameters, namely the normal, triangular and lognormal distributions. The model with normally-distributed parameters is denoted NU (i.e. normal unconstrained). The t-test and the change in the maximum likelihood, compared to the original Multinomial Logit model, testify that this is a successful model (see table 4.1). However, examining the range of parameter values in the NU model reveals that positive fare parameters are attached to 5.4% of the population (as the mean parameter is -4.0981 with standard deviation of 2.5504). An individual whose fare parameter is positive prefers an expensive journey to a cheap one, when all other attributes are equal; although 5.4% is not a substantial share, it is yet unlikely that the number of people with such unusual behaviour is as high as this. Looking at the distribution of the ML parameter, it is found that 29.1% of the population have positive parameters (the mean is -0.7168 with standard deviation of 1.3028). A person with a positive ML parameter is one that prefers to arrive late to work. Since many working places operate flexible working times, it is likely that some travellers might wish to arrive to work later than the formal starting time; but a portion of 29.1% seems excessive (although we have no solid evidence to support this). In general, the NU model appears statistically successful, but some qualitative judgement implies that the parameter estimates are questionable.

The TU model (triangular unconstrained) replaces the normal distribution used in NU with the symmetrical triangular distribution. The TU model estimation results are very similar to those obtained for the NU model, both in terms of the statistical fit of the model, which seems good, and in terms of the distributions of the parameters, which

seem unlikely. The fare parameter has 5% positive values and the ML parameter has 29%. Both NU and TU models used unconstrained distributions; the occurrence of a large share of positive values is simply a result of the shape of the curves of these distributions. The main conclusion this leads to is that more reasonable estimates might be obtained by constraining the distributions of the parameters. The following model specifications apply different approaches for the constraining of these distributions.

One way of narrowing the range of parameter values is by using a distribution such as the lognormal, which is *constrained by definition*. The lognormal distribution only encompasses positive values, and if a negative sign is attached to it, it only includes negative values. By specifying a model with lognormally-distributed parameters for the fare and ML variables we can therefore avoid the positive tail experienced in the NU and TU models. Unfortunately, the attempts to estimate a model with lognormal distribution were not successful, as the estimation tool was not able to identify a satisfactory model. Difficulties in estimating MXL models with lognormal parameters are very common; similar experience is reported by Small et al (2005) and other authors.

A different approach to avoiding outspread parameters is by using distributions with lower and upper bounds, which are estimated simultaneously with the other parameters. We refer to this approach as *constraining by estimation*. A distribution that can be used efficiently for such estimation is Johnson's S_B distribution (Johnson, 1949), which has been successfully implemented in MXL models described by Hess et al (2005) and others. In many of the common distributions, such as the normal or lognormal, any specific curve can be identified by two parameters, namely the mean and variance; the S_B distribution includes the lowest and highest values as additional parameters and it therefore has four parameters. The appealing property of a distribution such as S_B is that it enables fixing the highest and the lowest values for each parameter at levels that are directly derived from the data. It therefore avoids the somewhat-arbitrary shape of the tails obtained at the far ends of the distribution when an unbounded distribution is used. Nevertheless, the main strength of distributions constrained by estimation is also their main weakness: the fact that each curve of the S_B distribution is defined by four parameters means that during the estimation of the MXL model, there are more parameters to identify. In the current work, all the attempts to specify and estimate MXL models with the S_B distribution were not successful, since the estimation tool was not able to identify a satisfactory set of parameters. A very similar case, where it is

found impossible to identify the parameters in a model that uses the S_B distribution, is reported by Train and Sonnier (2003). It might be assumed that constraining by estimation is more likely to work well when a very big dataset is at hand, such that the need to identify many parameters does not constitute a major problem.

The remaining way to avoid an irrationally large share of positive parameters is through *constraining by imposition*. This approach involves imposing various constraining rules on the distributions of the parameters prior to the estimation of the model. Since the constraints are introduced before the estimation, without an attempt to determine constraints (such as upper and lower bounds) from the data, there are fewer parameters to identify, compared to the case of constraining by estimation. It is clear that constraining by imposition is not the preferable technique. As illustrated in the following paragraphs, the imposed constraints are inevitably arbitrary; specifying models with imposed constraints is considered not because there is any evidence that these constraints make sense, but simply because such model is easier to estimate. Still, as other specifications either led to irrational estimates or did not lead to any model at all, the advantages of constraining by imposition should not be undervalued.

Two models were estimated using a constrained normal distribution. In both models, the method of constraining was forcing the standard deviation of the parameters to be equal to the mean; with the normal distribution this always fixes the percentage of positive values at 16%. In one of these models (denoted NC, i.e. normal constrained), such restriction was imposed on both the fare and ML parameters, and in the other (NP, i.e. normal partially-constrained) it was only imposed on the ML parameter. In the NP model, the proportion of travellers with a positive fare parameter rose to 8.4%. An additional model, where only the fare parameter was constrained, appeared much inferior by both statistical and rational judgement, and is not presented here.

Two additional constrained models were estimated with the symmetrical triangular distribution: a partially-constrained model (TP) and a fully-constrained model (TC). The rule used for constraining was that the spread of the distribution must be equal to the mean; in a symmetrical triangular distribution with a negative mean, this always sets the upper bound at zero and thus there are no positive values. This constraining rule was imposed on the ML parameter in the TP models and on both ML and the fare parameters in the TC model. In the TP model, the share of positive values for the fare parameter, that was left unconstrained, increased to 9.3%. As with the normal

distribution, a triangular specification where only the fare parameter is constrained performed much poorer than the other specifications and is therefore not presented.

The output of the entire modelling sequence is the six MXL models presented in table 4.1. Note that in models with normally-distributed parameters, the error component is the standard deviation, whereas in models with triangularly-distributed parameters the error component is the spread. All the presented models have significantly higher likelihood than the Multinomial Logit model, and reasonably high t-statistic values for all parameters.

	Multinomial	NU	NC	NP	TU	TC	TP
Fare	-1.343 (-4.5)	-4.0981 (-11.0)	-3.4812 (-11.2)	-4.0071 (-11.0)	-4.1859 (-10.9)	-3.6528 (-10.6)	-3.6791 (-11.2)
MTE	-0.0724 (-7.3)	-0.1413 (-11.2)	-0.1377 (-11.0)	-0.1405 (-11.1)	-0.1410 (-11.3)	-0.1244 (-11.4)	-0.1234 (-11.0)
ML	-0.1961 (-2.3)	-0.7168 (-6.0)	-0.8953 (-6.7)	-0.9248 (-7.2)	-0.7285 (-6.2)	-0.6943 (-4.9)	-0.6401 (-4.7)
Fare – error component	-	2.5504 (9.5)	3.4812 (11.2)	2.9074 (9.3)	6.1272 (9.1)	3.6528 (10.6)	6.4779 (10.0)
ML – error component	-	1.3028 (7.9)	0.8953 (6.7)	0.9248 (7.2)	3.0533 (8.2)	0.6943 (4.9)	0.6401 (4.7)
Final likelihood	-1187.9	-999.4	-1010.6	-1004.7	-998.4	-1056.2	-1028.5

Table 4.1: Mixed Logit models (t-test results in brackets)

In order to derive the DWP that we are interested in, it is necessary to choose the single model that seems to perform better than the others. But this choice is found quite challenging, as all six models have serious flaws. Judging by the traditional maximum likelihood test, the unconstrained models (NU and TU) are the best. This is supported by other tests, such as rho-squared, which are not presented as they lead to the same conclusions. But as discussed earlier, in the unconstrained models the positive tail of the

ML parameter seems too outstretched. In contrast, the constrained models (NC and TC) have lower maximum likelihood, and there is also some concern that the arbitrary imposition of the constraining rule, which presets the size of the positive tail, is an artefact that makes these models unreliable. The partially-constrained models (NP and TP) could be seen as a tolerable compromise, but they imply an outsized positive tail for the fare parameter. All in all, the question of which specification should be employed to estimate the DWP is still unanswered.

The various approaches attempted here for constraining the distributions of the parameters are summarised in table 4.2. These approaches have been used in previous studies, but note that none of these previous studies illustrated the fundamental differences between them. For instance, Hess el al (2005) present the advantages of constraining by estimation without mentioning that it normally requires a rich dataset, while Hensher and Greene (2003) use constraining by imposition without stressing that using preset constraints constitutes a serious compromise. Using the terminology developed earlier, observe that constraining by estimation sees the constraints as part of the set of model parameters, θ, whereas constraining by imposition deems them part of the model structure, D.

Approach	Distributions	Advantages	Disadvantages
Unconstrained parameters	Normal, triangular	No interference in the estimation process	Risk of excessive share of positive values
Constraining by definition	Lognormal	No risk of a tail with the wrong sign	Difficulties in estimation
Constraining by estimation	S_B	Upper and lower bounds derived directly from the input data	High number of parameters leads to identification problems
Constraining by imposition	Normal, triangular	Easy to estimate because of the small number of parameters	The preset constraining is an artefact

Table 4.2: Different approaches to constraining distributions

The NC, NP, TC and TP models described above were specified using a relatively simple imposed constraint, which set the standard deviation or the spread at the same level as the mean parameter. Hensher and Greene *(5)* discuss other constraining rules, such as forcing the standard deviation to equal the mean multiplied by a factor. Although this opens an avenue for many new possible model specifications, it was decided not to use this modified approach; the reason is as follows. If the mentioned factor is preset, a distribution is thus imposed, but Hensher and Greene do not devise a procedure of determining the value of the factor prior to estimation. Apparently this requires either a lengthy trial-and-error process whose success is not guaranteed, or an arbitrary decision that will form an intrusive interference in the estimation procedure. If the factor is estimated, then similar identification problems as those with distributions constrained by estimation are likely to occur.

As discussed earlier, the main purpose of the entire series of modelling experiments is to obtain estimates of the DWP for reduced TTV. The distributions of the model parameters are of interest because they are meant to be used later to derive the DWP. Since it was found difficult to choose the best model by looking at the values of the parameters, it might be useful to also explore the different DWP curves implied by each of the six alternative models. All six model specifications include two variables to which we want to attach a monetary value: MTE and ML. We therefore now derive the VOTE (*value of mean travel time and earliness*) as the ratio of the MTE parameter to the fare parameter, and the VOL (*value of mean lateness*) as the ratio of the ML parameter to the fare parameter. Note that although the MTE parameter remained fixed (not random), we do get an entire distribution of VOTE because of the division by the fare parameter. The distributions of VOTE and VOL are derived by repeatedly taking random draws from the distributions of the relevant parameters (that vary according to the specification of the model), calculating the ratios of these random numbers (fixed MTE parameter to random fare parameter or random ML parameter to random fare parameter), and finally analysing the distribution of the resulting ratios. The distributions of VOTE and VOL are presented in figures 4.1 and 4.2 as cumulative frequency curves. A cumulative presentation was chosen because some of the curves include values that lie outside the presented range; the amount of such values can only be emphasized in a cumulative diagram.

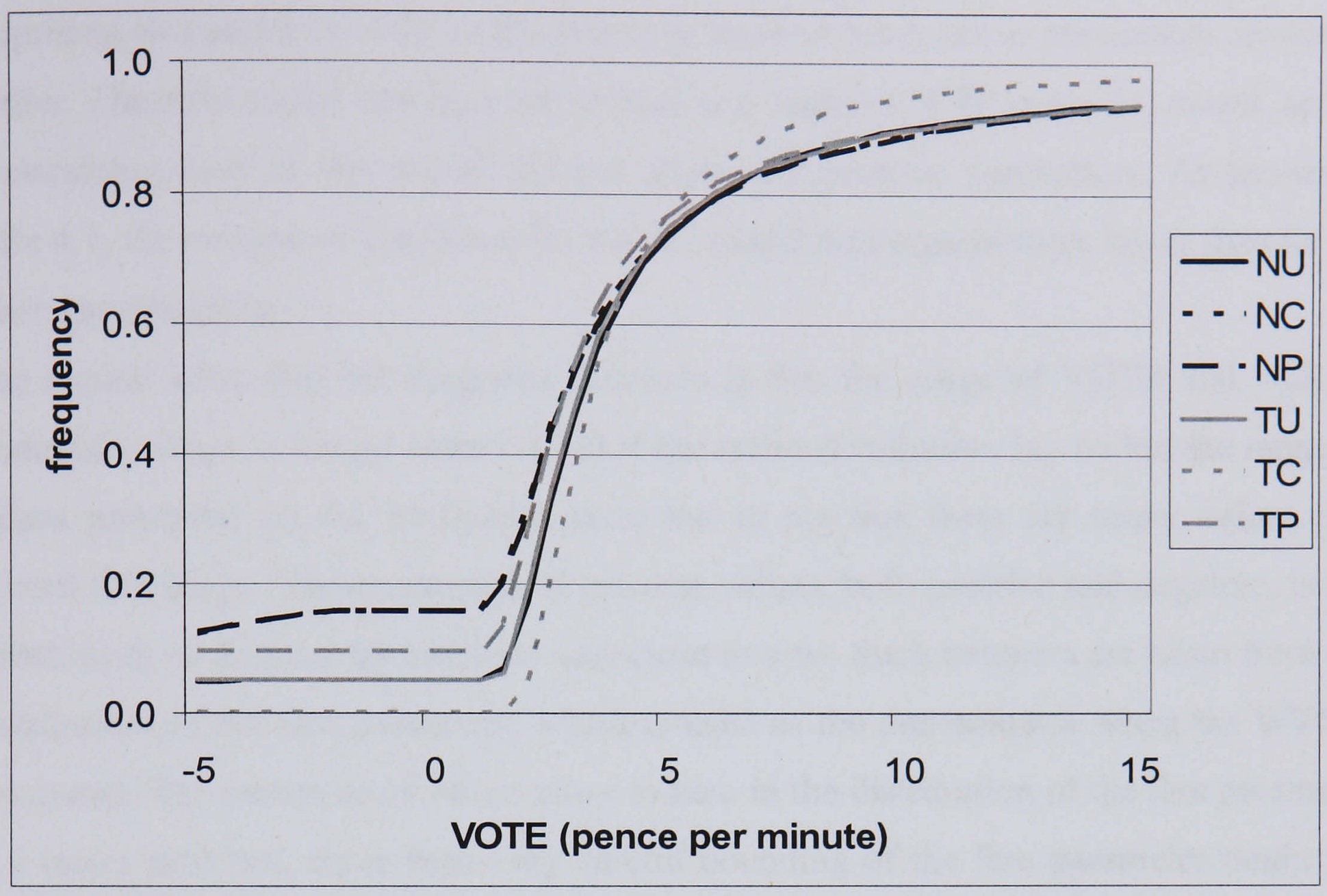


Figure 4.1: The distribution of VOTE

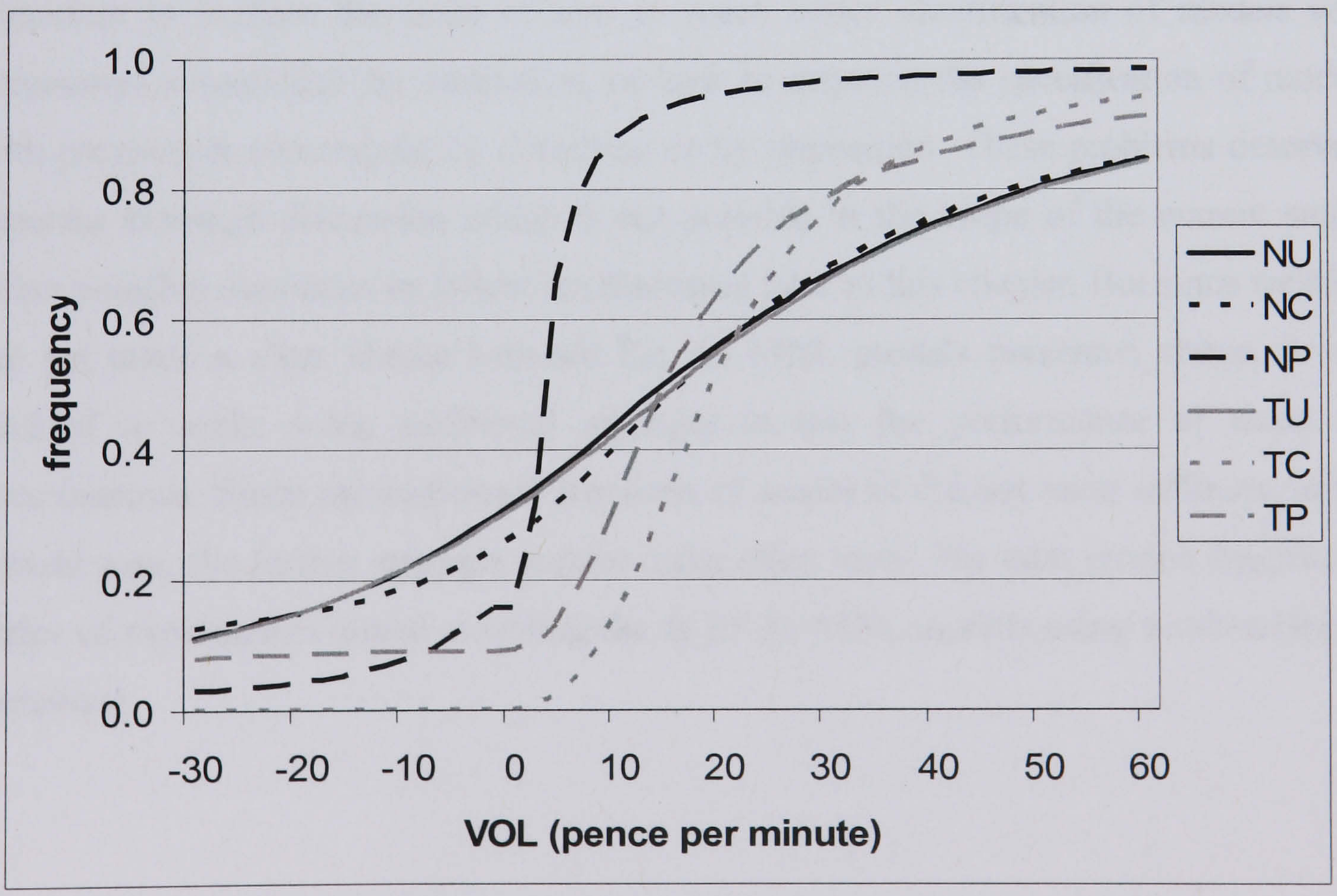


Figure 4.2: The distribution of VOL

The DWP diagrams demonstrate two important issues. First, it can be seen that the distributions of VOTE and VOL include a negative share that seems too big; this is not surprising as a negative WTP is the direct outcome of the positive parameters discussed earlier. The only model that does not include any negative WTP is the TC model, as the constraining used in this model did not allow any positive parameters. As shown in table 4.1, the maximum likelihood for the TC model was significantly lower than for all other specifications.

The second issue that the diagrams illustrate is that the range of VOTE and VOL is irrationally large. It would seem logical if the entire distribution lay within the range of values presented on the horizontal axes; but in practice there are many values that exceed this range. The occurrence of extreme values, both positive and negative, is the direct result of division by numbers very close to zero. Such numbers are taken from the distribution of the fare parameter, which is used as the denominator when the WTP is computed. The existence of values close to zero in the distribution of the fare parameter is a major problem, since even very careful bounding of the fare parameter might not completely prevent it.

The problems encountered while trying to model the attitudes to TTV with MXL can lead to several different directions of further investigation. For instance, it should be important to explore the issue of how to reach better identification of models with parameters constrained by estimation, or how to improve the specification of models with parameters constrained by definition or by imposition. These problems deserve a separate thorough discussion which is not possible in the scope of the current study. Other possible directions to follow are discussed later in this chapter. But since we have not yet made a clear choice between the six MXL models presented above, it was decided to make some additional attempts to test the performance of these six specifications. Since the traditional measures of model fit did not seem sufficient in the current case, the further attempts require using other tools. The next section describes a series of experiments aimed at testing the fit of the MXL models using a sub-sampling technique.

4.4. Sub-sampling experiments

4.4.1. Tests for comparing distributions

The main analysis presented in this section involves comparing distributions based on the MXL models to the respective distributions estimated independently using sub-sampling (SUS). Before describing the SUS experiments it is important to verify that trustworthy methods are used for this comparison. We find that two distributions can differ in two distinctive dimensions, and hence both these dimensions should be compared.

If we describe the two distributions as cumulative frequency curves, the first dimension of potential difference is represented by the vertical axis of the diagram. Namely, a small vertical distance between two cumulative frequency curves signifies good fit between the distributions; this is conventionally measured using the Kolmogorov-Smirnov (K-S) test. In the current context it was decided to consider any K-S test statistic smaller than 15% as indicating satisfactory fit. This is merely a rule of thumb, but since the input data is based on responses from travellers that are often inconsistent, and since the SUS technique used to obtain an external estimate of the DWP is inaccurate, we find it unnecessary to use a rigorous statistical terminology that includes accurate levels of confidence. For a more thorough description of the K-S statistic, see later chapters of this thesis, where this test is used more systematically (although in a different context).

Verifying tolerably small vertical difference is essential when comparing two distributions, but it is not sufficient. Major differences can lie between two compared distributions in another dimension, that we call the *horizontal* dimension, even if the vertical fit is good. This can result from the existence of some very big numbers (positive or negative) in only one of the distributions; the amount of these big numbers might be too small to make a vertical difference, but it can still cause vast differences when various statistics are computed. The DWP is particularly sensitive to poor horizontal fit, since WTP is a quotient of two numbers; as illustrated above, the value used as denominator is sometimes very close to zero, hence it can result in very high values. A simple way to verify good horizontal fit is by comparing the standard deviation of the analysed distributions.

The K-S test and the comparison of the standard deviation, which are used here, are not the only ways to verify good fit between two distributions, and they clearly have some drawbacks. In future research, it should be important to examine other measures. However, it should be stressed that it is essential to examine the discrepancies between the distributions both vertically and horizontally.

4.4.2. Experiment A: the validity of sub-sampling experiments

The essence of SUS was described earlier in this chapter. The series of experiments presented in this section uses SUS to obtain independent estimates of the distributions of the model parameters or of the WTP; then the discrepancies between the SUS-based and the MXL-based distributions are used as an additional test of goodness-of-fit. The SUS technique used here is similar to the Bootstrap technique, and also to the SODA technique proposed by Sørensen and Nielsen (2003), as explained further in the following paragraphs. The first experiment in the series is meant to illustrate the use of this technique and to test whether it is sound to use it to estimate the DWP. Note that there is no intention to undermine other techniques, such as Jackknife. Bootstrap is a more direct method for trying to replicate the distribution of preferences in the original dataset, whereas Jackknife mainly performs sensitivity analysis. The direct output of Bootstrap is an approximated distribution which can then be easily compared with the MXL distribution without further analysis; but it should be emphasized that the use of Jackknife too for similar purposes should be considered in future research.

To perform the first experiment, the full dataset from the SP survey is used, but instead of the actual choices made by the respondents, artificial choices were added, based on an imaginary MXL model. The imaginary model has the same variables as the aforementioned models (fixed MTE parameter and random fare and ML parameters) but the parameter values are made up. These values were determined such that when the MTE and ML parameters are divided by the fare parameter, the resulting VOTE is distributed triangularly across individuals, with mean 3 and spread 3 pence per minute; and VOL is distributed triangularly across individuals, with mean 12 and spread 12 pence per minute. Our interest is in whether examining this synthetic dataset using SUS can reproduce the artificial DWP, which is in this case the *real* distribution.

The sample was randomly divided into 20 subsamples, such that each subsample contained the responses of about 5% of the individuals; responses from the same individual were always kept in the same subsample (to comply with the conventions of

panel data analysis). The random creation of subsamples was repeated 10 times; all together, 200 different subsamples were created, and a Multinomial Logit model with fare, MTE and ML parameters was estimated for each subsample. From the parameters of each model we can calculate VOTE and VOL, derive the distributions of these values across all SUS models, and examine whether these distributions are similar to those that were used to simulate the artificial choices. Note that the number of subsamples used here is much bigger than in the experiments described by Cirillo et al (2000). As mentioned in earlier in this chapter, Cirillo et al found that with a few dozens of subsamples, the Bootstrap estimates tend to underrate the extent of variation in the dataset; the high number of subsamples is used here in an attempt to reduce the risk of such bias.

The mean values in the SUS-based DWP are found very similar to the real values: 2.9 pence per minute for VOTE and 11.5 for VOL. The SUS standard deviation of VOTE is 0.9 (the real is 1.2); the SUS standard deviation of VOL is 8.7 (the real is 4.9). When the SUS distributions are compared to the real ones, K-S test statistic for VOTE is 11% and for VOL it is 12%. We conclude that although SUS distributions do not replicate the real preferences accurately, they do constitute reasonable estimates of the true behaviour. In the absence of better estimates we therefore wish to use SUS to examine how likely different MXL model specifications are, in addition to more traditional tests such as the maximum likelihood.

4.4.3. Experiment B: fitting distributions to sub-sampled parameters

Unlike the illustrative experiment described above, in the second experiment we use the real choices recorded in our database. However, we still do not compare the MXL-based and SUS-based estimates directly (we do this later); the different specifications of the MXL model are ignored at this stage. The experiment follows the methodology proposed by Sørensen and Nielsen (2003), which is meant to be used prior to the MXL model estimation (it is, however, different from Sørensen and Nielsen as it takes into account the panel nature of the data). This basic test examines whether the SUS-based parameter distributions resemble several common distributions; we use normal, triangular and lognormal (with the opposite sign). The logic is that if a specific distribution fits well, it can then be used as a starting point for MXL estimation. Using the notation defined earlier, the current experiment assists in making an intelligent guess of D under the assumption that there is no risk of bias in θ. Note that it is not possible to

compare the SUS distribution to the general shape of a normal, triangular or lognormal distribution; it is only possible to compare with specific curves from these families. In the analysis performed here we give these specific curves the SUS-based, not the MXL-based, mean and standard deviation; namely, as in Sørensen and Nielsen (2003), we assume that the SUS estimate of θ is always accurate.

The test was carried out similarly to experiment A. Based on the full sample (with travellers' real choices), 200 subsamples were created and a Multinomial Logit was estimated for each. The distribution of parameters across these sub-models was compared to normal, triangular and lognormal distributions with the same mean and standard deviation. In the current experiment we do not examine the distribution of the MTE parameter since in the MXL models that we are about to test, the MTE parameter is fixed; its fit is tested later.

The SUS distribution of the fare parameter includes positive values for 1% of the sample; these values are very close to zero and therefore seem negligible. This makes good sense because an increase in the level of utility as the fare increases seems most unlikely for any traveller. The fact that the SUS estimates of the fare parameter do not include a considerable positive share supports the doubts raised earlier concerning the credibility of the some of the MXL specifications. Unfortunately, conclusions from the SUS distribution of the ML parameter are not equally clear. Similar to the NC and NP models, SUS shows that about 18% of the travellers have positive ML parameters; this finding creates a dilemma. On one hand, both the MXL model and SUS have now independently shown that some travellers have a positive attitude towards late arrival. Note that such attitude does not seem as irrational as positive attitude to the fare or to the mean travel time, because in working places that allow flexible arrival times, it is not unlikely that some workers see late arrival as a better use of their time (for reasons related either to leisure activities, carried out before arrival, or to work activities, carried out after arrival). On the other hand, several authors have recently stated that positive parameters for time-related variables (and hence negative WTP) are irrational, as reviewed above. It therefore appears that before deciding which specification of a MXL model is the best, there is a need to decide whether or not we accept the existence of a small share of travellers that are lateness-prone. We return to this issue later in this section.

Table 4.3 and figures 4.3-4.4 present the attempts to fit distribution curves to the SUS-based parameter distributions. The SUS distribution of the fare parameter is

successfully replicated by the normal and triangular curves. The lognormal curve, however, has the infamous long tail which stretches much farther than the lower end of the SUS curve. Similar to the fare parameter, the SUS-based distribution of the ML parameter is reconstructed well by the normal and triangular curves. The lognormal curve is unable to replicate the positive tail of ML, and is also found inappropriate at the negative end (25.1%).

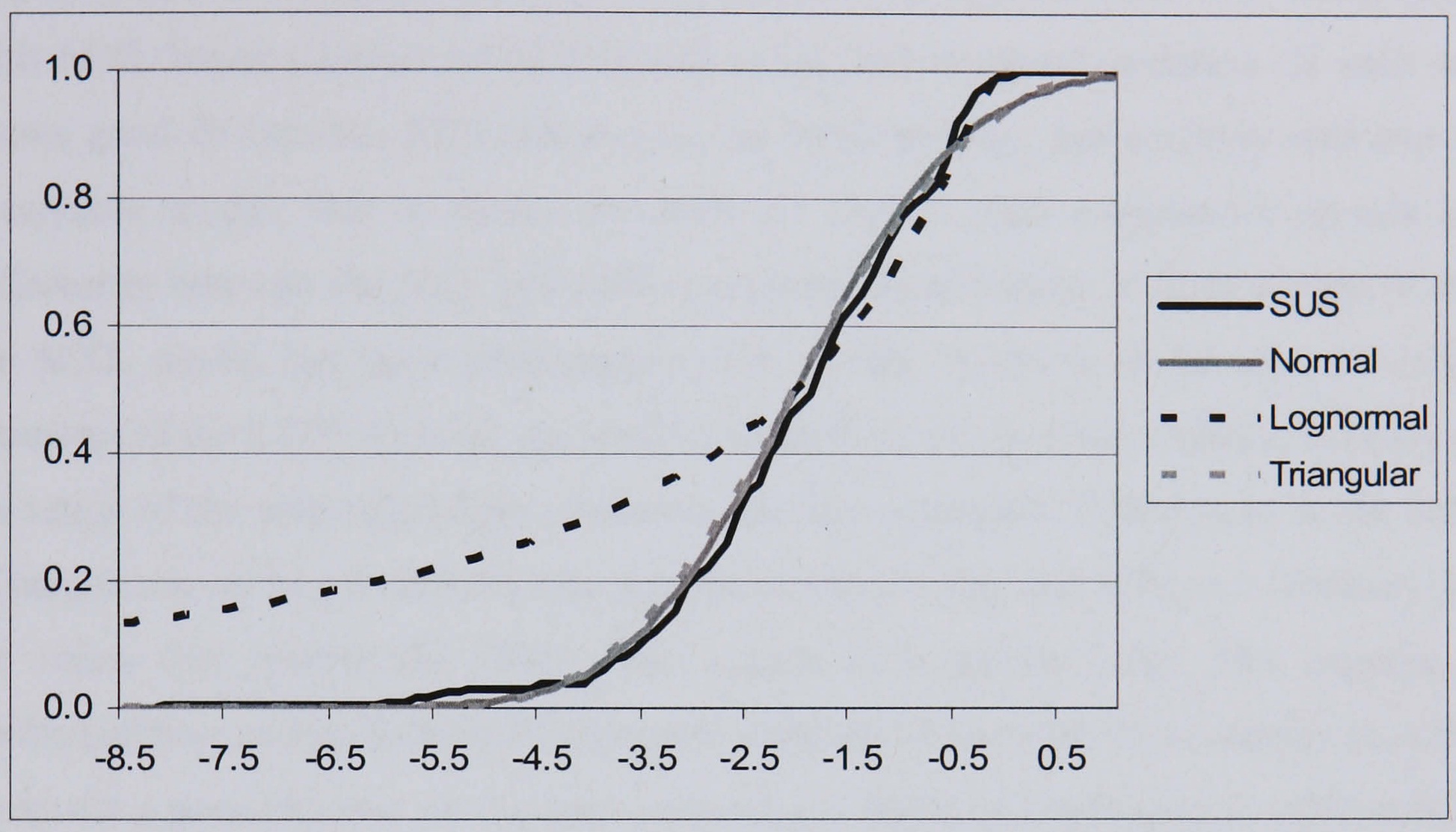


Figure 4.3: SUS distribution of the fare parameter

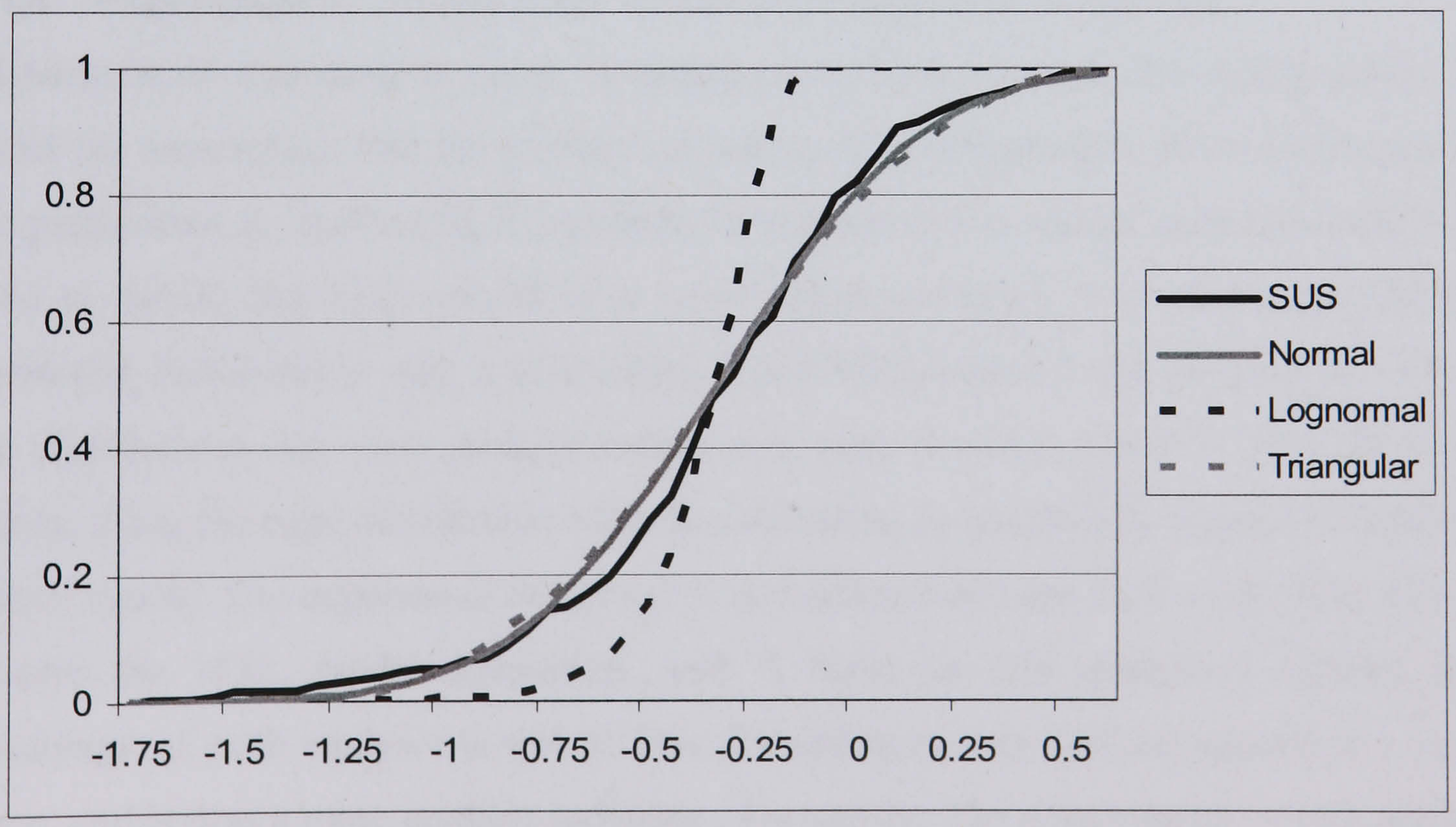


Figure 4.4: SUS distribution of the ML parameter

	Normal	Lognormal	Triangular
Fare	5.4%	25.1%	6.4%
ML	6.1%	29.6%	6.8%

Table 4.3: Experiment B - K-S test statistic

In principle, it is also possible to compare the SUS-based parameters with distributions with MXL-based (instead of SUS-based) mean and standard deviation. If such test shows good fit between SUS and one of the MXL models, this credibly indicates an acceptable model. But it should be observed that if such comparison reveals big differences between the SUS and MXL parameter distributions, it does not prove that the MXL model has poor performance. The sought model is to be used to derive estimates of the DWP; these do not directly depend on the parameter values, but only on the ratios of the time-related parameters to the fare parameter. Differences in the order of magnitude of the parameters between Multinomial Logit and MXL are common, but the ratios that specify the DWP often remain at a similar level. The experiment performed here, which follows experiments proposed in some previous studies, is in this sense not a powerful one. In the next subsection a different experiment is performed to compare not the parameter distributions but the DWP directly.

4.4.4. Experiment C: Mixed Logit versus sub-sampled distributions

Experiment B was used to assist in specifying the set of parameter distributions, D, under the assumption that the problem of setting D is independent of the estimation of the parameters, θ. In practice, if a guess of D is made *before* model estimation and then used to specify the MXL model, θ is re-estimated and there is no guarantee that the parameter distributions with a SUS-based D and MXL-based θ will perform as well as the distributions that were initially tested, with both D and θ based on SUS. In other words, using the right distribution while re-estimating its parameters might still result in a poor model. The experiment described in this subsection uses SUS to test both D and θ *after* the MXL model estimation, with θ based on the estimation results. The advantage of such analysis is that it does not disregard any bias introduced at a later stage, and is thus a more credible indicator of model fit. The disadvantage is that similar to other indicators, such as the maximum likelihood, it only applies to models that have already been specified and estimated; it does not tell us how to specify a better model.

The current experiment tests directly the DWP and not the model parameters, as the need for good estimates of the DWP is the primary motivation for the entire modelling effort. The SUS-based DWP is determined as in experiment A, except that it is based on the true behaviour rather than on artificial choices. The DWP from the different MXL models was derived by simulation. Each of the MXL models presented earlier defines a fixed MTE parameter and a distribution of ML and fare parameters; 10,000 (ten thousands) sets of parameters were drawn from each of the models, simulating the individual choice models of 10,000 travellers. The VOTE and VOL were calculated for each one of them, and the resulting distributions were compared to the SUS distributions.

Figures 4.5 and 4.6 present the cumulative frequencies of VOTE and VOL according to all six MXL models, similar to figures 4.1 and 4.2, but with an added curve for the SUS estimates. The figures demonstrate that the models with the normal distribution overestimate the share of travellers with WTP that lies at the far ends. The NP curve of VOL performs better than the other normal models at the negative end, but it fails to predict the entire part of the curve that lies above the mean. The TU curve implies (not surprisingly) that using the triangular distribution without constraints performs similarly to the unconstrained normal distribution. As mentioned earlier, the reason why many oversized values (positive or negative) are repeatedly found is that most feasible distributions of the fare parameter include values very close to zero, that are used as the denominator when calculating the WTP. The TP and TC models are the best in the way they replicate the right end of the distributions of VOTE and VOL; but the TP model, like most other models, gives too many negative values, and in contrast the TC model is unable to replicate any negative values. As we discussed in the previous subsection, the decision of whether or not the TC model is acceptable depends on another decision, namely whether or not we allow a certain share of lateness-prone travellers.

As explained earlier in this section, the diagrams illustrate the vertical differences between the estimates of the DWP, but they do not examine the horizontal differences. Table 4.4 presents a more systematic comparison between the SUS-based and MXL-based DWP. The vertical fit is measured using the K-S statistic and the horizontal fit is measured by comparing the standard deviation. The results are gloomy: none of the MXL models gives acceptable levels of vertical fit to the SUS estimates; most models also have very poor horizontal fit. The NC estimate of the standard deviation of VOL is 250 times higher than the SUS estimate, and the TP estimate is 1100 times higher than

SUS! NU, NP and TU also overstate the standard deviation, and only the TC model gives an estimate at the same order of magnitude as SUS.

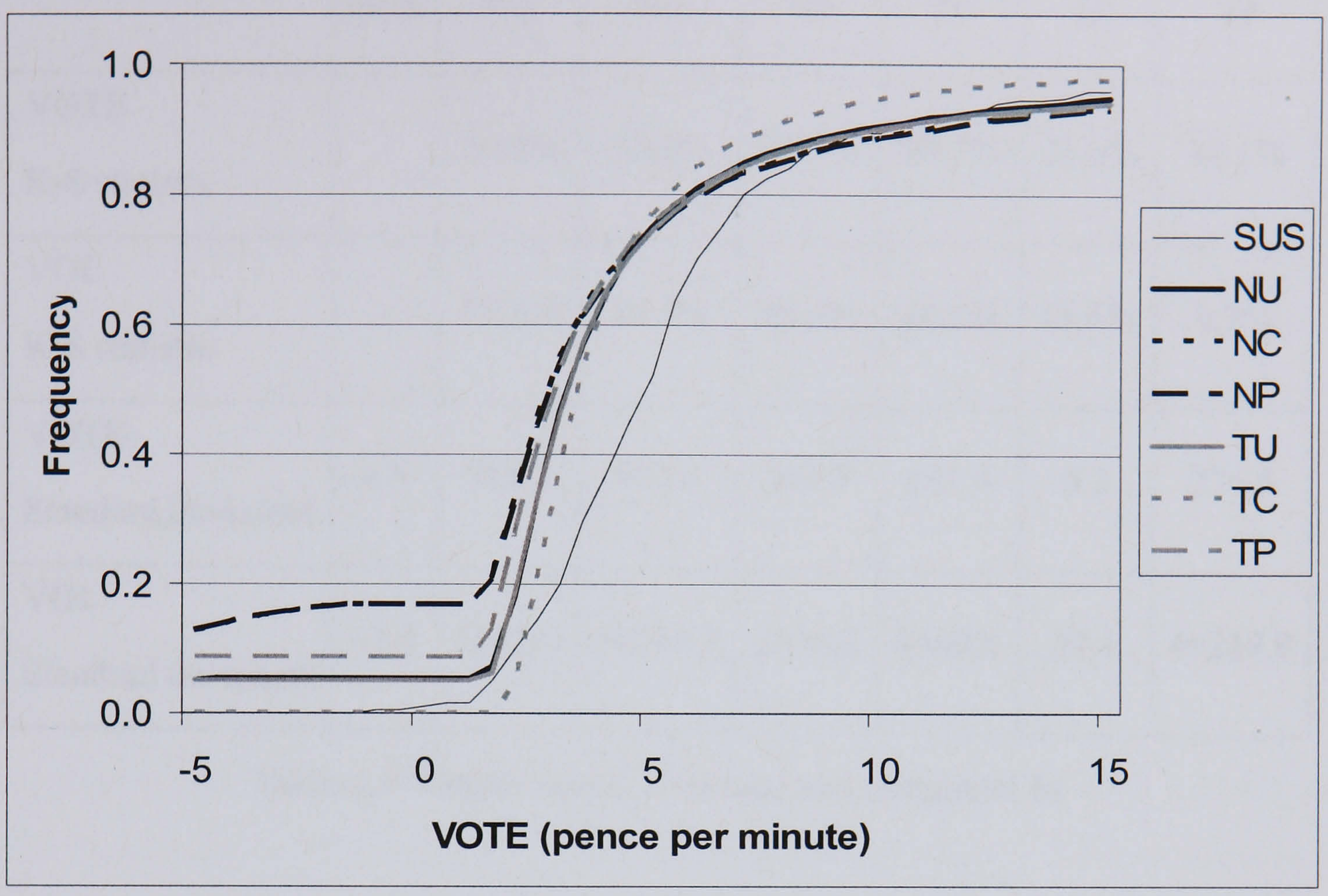


Figure 4.5: MXL-based VOTE versus SUS

Figure 4.6: MXL-based VOL versus SUS

	SUS	NU	NC	NP	TU	TC	TP
VOTE K-S statistic	-	26.8%	34.2%	33.2%	28.7%	26.8%	32.1%
VOL K-S statistic	-	19.4%	15.7%	58.6%	20.1%	16.6%	9.7%
VOTE Standard deviation	6.0	325.5	492.0	502.3	181.4	5.3	276.5
VOL Standard deviation	43.8	1512.0	11286.2	1896.2	6408.0	33.1	49219.9

Table 4.4: Experiment C – vertical and horizontal fit

4.4.5. Discussion

A major issue that the presented analysis raises is the existence of negative WTP. Experiment B shows that the SUS distribution of the fare parameter has only negative values; this enhances our suspicions about the positive tail of the fare parameter, implied by most of the MXL models. SUS results for the ML parameter in experiment B show that similar to the results of models NC and NP, 18% of travellers have positive values. The fact that a considerable amount of positive ML parameters is found by both SUS and some of the MXL models obliges us to rethink whether the argument made in some previous works (reviewed earlier), regarding the flaw in positive time-related parameters, is not too strict. The fair amount of positive parameters might insinuate that there should be a distinction between travel time in general, which is truly unlikely to have positive parameters, and specific elements of it, such as lateness. Since flexible start times are accepted in many working places, it might make sense that some travellers would rather arrive later than earlier. If we had managed to estimate a credible MXL model in which some share of positive VOL was the only feature that contradicted the conventions of traditional demand models, we would have therefore seriously considered accepting that model. But obtaining such a model was found here a real challenge, as each of the six MXL specifications presented here had other flaws.

Since the estimation of the DWP is the main objective of the modelling experiments, the poor fit of the MXL models is a key problem. Nevertheless, the experience described here raises an additional, more general problem: despite the poor fit, the conventional tools imply that the fit is good. Table 4.5 summarizes how different tests rate the MXL models. Judging by the maximum likelihood, TC is the worst model, but according to most other tests, it is the best! The TP model, which is the only model that replicates the distribution of VOL reasonably well, has the second-worst likelihood. The maximum likelihood shows a consistent tendency to penalise constrained models; the more constraints we place on the distribution of parameters, the lower the maximum likelihood. Still, comparison of the DWP between the MXL and SUS distributions testifies that constraining did generally lead to better estimates. The general problem implied by this experience is that MXL specifications that are chosen only by checking the maximum likelihood, and the respective estimates of the DWP, might lead to biased analysis and hence even to wrong conclusions from scheme assessment.

Our analysis used constraining of the parameter distributions as a key method for defining different model specifications from the same family of statistical distributions. It should be pointed out that the entire concept of constraining the cost parameter, in an attempt to reach a more rational DWP, has some deficiencies. One of them is that even if a logical way of constraining is found, the distribution of the cost parameter might still include values close to zero that result in an unbounded DWP. Another deficiency is that the distribution of the cost parameter often affects more than one WTP ratio (e.g., in our case, VOTE and VOL), and the constraints required to adjust different ratios might contradict each other. Such problem is most apparent in models where some parameters are fixed (i.e. in most MXL models), since altering the cost parameter becomes the only way to control the DWP associated with the fixed parameters. For both these reasons, one might claim that in some cases, a cost parameter distribution that leads to a realistic DWP simply does not exist.

Rate	MXL	Experiment C			
		K-S test (vertical test)		Ratio of sta. dev. (horizontal test)	
	Maximum likelihood	VOTE	VOL	VOTE	VOL
1 (best model)	TU (-998.4)	TC (26.8%)	TP (9.7%)	TC (88%)	TC (76%)
2	NU (-999.4)	NU (26.8%)	NC (15.7%)	TU (3023%)	NU (3452%)
3	NP (-1004.7)	TU (28.7%)	TC (16.6%)	TP (4608%)	NP (4329%)
4	NC (-1010.6)	TP (32.1%)	NU (19.4%)	NU (5424%)	TU (14630%)
5	TP (-1028.5)	NP (33.2%)	TU (20.1%)	NC (8200%)	NC (25768%)
6 (worst model)	TC (-1056.2)	NC (34.2%)	NP (58.6%)	NP (8372%)	TP (112374%)

Table 4.5: Rating of the MXL models by different tests

As mentioned earlier, the main reason for the poor performance of the MXL models is insufficient horizontal fit, caused by an excessive share of very big values that result from fare parameters close to zero. The presented analysis detects this problem only in our particular database, but we believe that a similar problem is likely to occur in many models with a random cost parameter. The search for solutions to the problem can lead to different directions, as described in the following paragraphs.

The unbounded nature of the WTP can be solved by not allowing the fare parameter to vary across the population. In the current context this means that VOTE will have a fixed value for all travellers, because an error component for the MTE variable did not appear significant. Since we believe that some variation of VOTE across travellers does exist, this would be undesirable. Besides, Train and Weeks (2005) discuss why it is

generally important to allow for heterogeneity in the cost parameter despite the difficulties it brings in.

Efforts to avoid the problems arising from the sensitive nature of the cost parameter distribution can take another direction, by estimating the model in the WTP space rather than the utility space. The concept of modelling in WTP space, and its main advantages and disadvantages, were mentioned in section 4.2.1. The advantages include the fact that the WTP is computed directly, avoiding the division by small numbers, as well as some evidence of good statistical fit, reported in a small number of studies. The disadvantages have to do with the fact that there is still very limited general experience in application of this approach, and in particular, experience with the performance of different statistical distributions in WTP space. An additional major disadvantage is that techniques for such estimation are currently not offered by the common software. Therefore, although this opens an interesting avenue as an alternative modelling approach, extending our modelling experiments in this direction is beyond the scope of the current study.

At this stage, we leave the problem of choosing the best DWP estimate unsolved. We return to it later in this chapter, after examining the DWP from a different perspective.

4.4.6. Summary

The analysis presented here demonstrates some of the difficulties and risks involved in the process of estimating MXL models. We wanted to allow variation in the cost parameter, but this gave rise to problems in specifying a model that leads to a credible estimate of the DWP. When the MXL model parameters were not constrained, the resulting range of WTP seemed too wide. It seemed that limiting the range of parameter values, by using distributions that are constrained by definition or by estimation, would solve this problem (at least partially); but models with such distributions were found hard to estimate. We therefore specified some models where constraints were imposed on the parameter distributions. We then had to decide whether the constrained models performed better than the unconstrained ones.

Our experience suggests that the maximum likelihood test penalises any constraining of the distributions of the parameters, even if it appears that such constraining improves model fit. Since the true range of traveller preferences is unknown, we found it useful to compare the MXL-based estimates to the distribution of preferences derived from a SUS experiment. The SUS distribution is merely an inaccurate estimate of the true

distribution, and it is always biased downwards, at least slightly, as explained earlier in this chapter; but it is still a good test of the MXL model fit, in the light of the failure of other statistical tools to detect poor model performance. We fear that similar deficiencies to those we found in the models presented here may possibly exist in models presented by others; they might have not been identified because insufficient statistical tests were used.

Previous studies suggested using SUS *before* model estimation, to assist in guessing the true parameter distribution; we remind that this does not guarantee good fit, since the parameters of the distributions are then re-estimated. We recommend using SUS also *after* model estimation. We also remind that when MXL and SUS distributions are compared, both the vertical and horizontal dimensions should be examined; the DWP is prone to poor horizontal fit, a problem that stems from values close to zero of the cost parameter.

Since the traditional measures of MXL model fit are insufficient, it is suggested that SUS should be used, among other tests, when evaluating MXL models. To enable easy use of SUS techniques it should be considered to incorporate them as an integral part in the conventional software tools used for estimation. Some of these tools (i.e. Alogit) already include SUS features but there is scope to extend them such that they enable analysis such as the one performed here. There is also scope for investigating, in future studies, the use of other SUS techniques, such as Jackknife. Finally, since many difficulties are encountered due to the parametric nature of the conventional estimation concept, there seems to be a growing need for exploring the merits of alternative, nonparametric methods. A simple nonparametric investigation of our dataset is undertaken in the following section.

4.5. Deriving the distribution of the willingness to pay under weak assumptions

4.5.1. Deriving individual willingness to pay from survey responses

One of the main findings from the MXL modelling experience described above is that different model specifications lead to very different estimates of the DWP; it seems that the original data leave substantial freedom of interpretation for the estimation procedure. If the derived DWP depends not only on the data, but also on the model

specification and on the estimation procedure, there is great interest in trying to isolate only the information that truly exists in the raw data. The current section presents a very simple experiment that aims at exploring these questions. The experiment is nonparametric, as it tries to avoid any preliminary assumption concerning the shape of the DWP curve.

The experiment performed here accepts the conclusion reached in chapter 3 concerning the most significant variables (fare, mean travel time, mean earliness and mean lateness); other variables are not re-examined here. However, the grouping of the mean travel time and earliness into one variable is no longer required. The grouped variable performed better than two separate variables in the previous modelling experiments, but there is no need to assume that the same will occur when a different approach is used. The aim of this section is therefore to determine the DWP that corresponds to three variables: MTT, ME and ML.

Models which are used to derive the WTP are typically estimated based on data from SP surveys. A very common type of situation presented in such surveys includes two alternative sets of travel conditions, corresponding to the variables that the modeller considers to include later in a cost or utility function; the respondents are asked to choose their preferred set. One of the attributes that defines the travel conditions is conventionally the cost. If there is only one additional attribute, such as journey time, then the survey can lead to one-dimensional analysis of the WTP for time savings (i.e. VOT). Each particular choice situation in such survey draws a border between two ranges of WTP. If, for instance, the respondent is asked to choose between a 30-minute journey that costs 200 pence and a 20-minute journey that costs 300 pence, the generalised costs of the two options are equal when

(4.4)

$$200 + 30 \cdot C = 300 + 20 \cdot C$$

i.e. when C=10 pence per minute (ppm). Fowkes and Wardman (1988) and Fowkes and Preston (1991) refer to this as the *boundary value*. A respondent that chooses the more expensive option can be assumed to have a higher WTP than 10 ppm, and a respondent that chooses the cheaper option can be assumed to have a lower WTP than 10 ppm. It is expected that a series of such choice situations with different attribute levels will create a set of constraints on the feasible range of the respondent's WTP, such that it will be possible to identify a relatively small range of monetary values that the sought WTP lies

in. This is illustrated in figure 4.7, where the responses of a single individual to five different SP choice situations are described. The short solid vertical lines represent the boundary values defined by each choice situation; the respondent's choice in each situation suggests whether the range of this respondent's WTP lies to the right or to the left from the boundary value. Based on all the responses is it possible to seek the most likely individual WTP (the dashed vertical line), and derive the entire DWP.

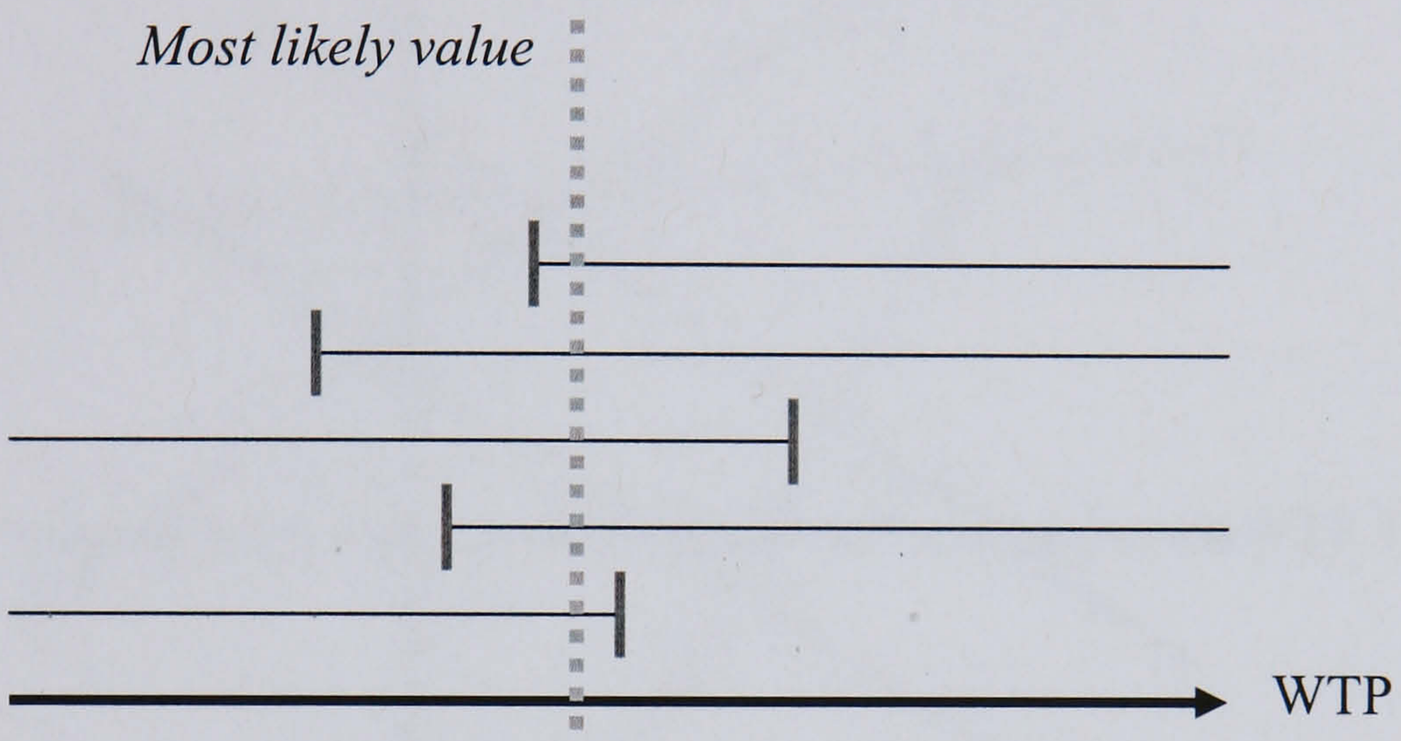


Figure 4.7: One-dimensional derivation of WTP from SP responses

If the modeller attempts to find the WTP for two attributes, such as travel time and the mean lateness to the destination, a new dimension is added to the problem. The boundary determined in each choice situation is no longer a point along a line that represents all feasible WTP levels, but a line in a two-dimensional plane. The response to a single SP choice situation divides the plane into two areas, in one of which the WTP of the respondent is more likely to lie. The analysis of a series of responses in the two-dimensional case is illustrated in figure 4.8. Such analysis is expected to imply in which of the areas bounded between the different lines the most probable combination of WTP levels is.

The same concept applies when the number of WTP elements is higher: if the cost or utility function is linear, then each SP choice situation divides the multidimensional WTP space into two parts, and the respondent's individual WTP is more likely to be in one of them. However, it is common knowledge in SP design that the number of choice situations should be restricted. Respondents tend to become tired or bored, and their answers gradually turn less credible. A reasonable number of choice situations is around nine (Fosgerau, 2006; Noland et al, 1998; Jackson and Jucker, 1982). The direct effect

of this limitation is that when there are several WTP elements, the area in the multidimensional space where we can identify high likelihood of a certain combination of WTP is quite big. In other words, if several elements of WTP need to be identified with a relatively small number of choice situations, it is hard to guarantee sufficient constraining that leads to a properly-defined WTP.

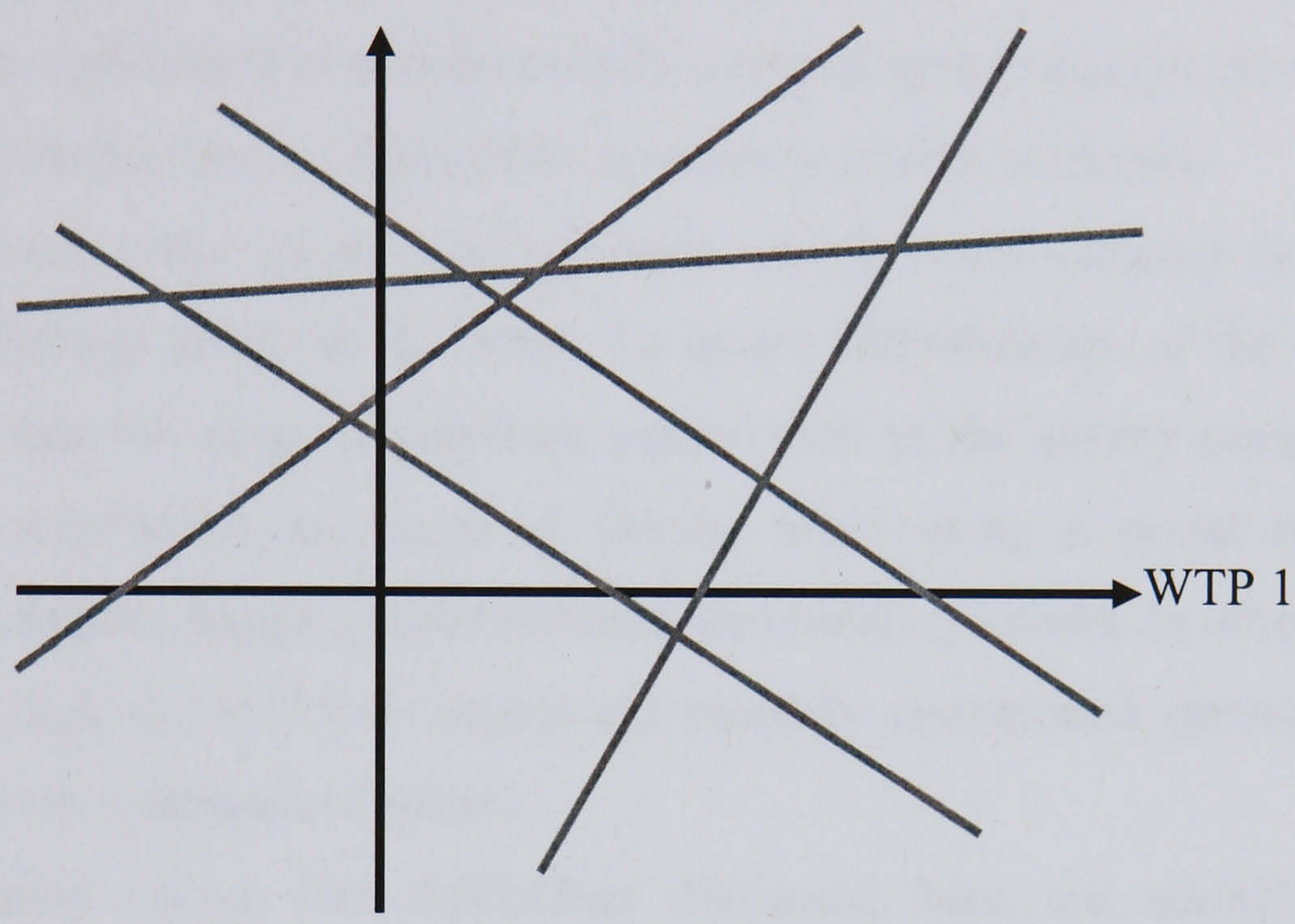


Figure 4.8: Two-dimensional derivation of WTP from SP responses

Good design of the SP survey, that takes into account the composition of areas between the boundary values, can improve the ability to identify the WTP. But given the restriction on the number of choice situations a respondent is willing to face, good design is not expected to completely solve the problem of identification when there are several WTP attributes. Since the responses are unknown at the stage of design, we do not know which boundary values will be effective in identifying the WTP. Even if the boundary values are designed such that they divide the WTP space into many areas, the particular combinations of choices made by each respondent turns only part of these boundaries into strong constraints on the WTP. Almost any combination of the attributes in an SP choice situation can lead to valuable insights about one respondent but not provide any important information about another respondent. The fact that the optimal SP design depends on the values of the parameters, which are still unknown at the stage of design, has been also discussed recently by Cirillo (2005).

An additional difficulty has to do with possible mismatch between the attributes used in survey design and the variables chosen later, at the stage of modelling. In the case

discussed here, the attributes used to generate the SP choice situations were the fare, MTT, departure time and level of TTV; but during the estimation of a choice model it was found better to transform this information into a set of variables that includes ME and ML instead of the departure time and TTV. Ideally, all six attributes had to be accounted for in the survey design. In practice, ME and ML were not explicitly considered when a reasonable composition of attribute levels was verified. Considering this was not practical at the stage of survey design, because at that stage there were many other variables that seemed equally likely to appear significant at later stages, and it was not feasible to use all possible variables as design attributes.

Even in cases where all relevant variables are taken into account as design attributes, another dilemma needs to be faced. To assure identification of the entire DWP, it is important that the range of attribute values used in the survey design is big enough. However, expanding the range of values, while using a preset number of choice situations, implies bigger spaces between the boundary values. In other words, verifying that very high or very low values are properly constrained inevitably weakens the constraints on intermediate values.

As illustrated above, the difficulties discussed here are strongly related to the dimensionality of the problem, i.e. the number of SP attributes or potential model variables. Dimensionality issues in SP design have been recently discussed in several publications, but from a different perspective. Hensher (2004) explores the influence of SP dimensionality on the WTP estimated from a MXL model. The main question being examined is whether the derived monetary estimates vary systematically with the dimensionality of the design, i.e. when there is an increase in the complexity of the tasks the respondent is faced with. In our case the dimensionality was predetermined by other considerations, namely the need to account for the attributes that capture the effects of TTV. Hensher does not investigate the identification of the WTP with the varying number of dimensions. Caussade et al (2005) study how different aspects of dimensionality and complexity in survey design affect the modelling outputs. Again, the WTP is derived in the traditional way and the ability to identify it is not challenged.

4.5.2. Deriving boundaries at different levels of consistency

The simple experiment performed here is similar to the analysis mentioned earlier by Beesley (1973), which seeks the lowest and highest values accepted or rejected in each response in the dataset. But unlike Beesley, we allow each respondent to have different

WTP, so that we can then look at the resulting DWP. We assume that the cost associated with a journey is:

$$Total\ cost = Fare + VOT \cdot MTT + VOE \cdot ME + VOL \cdot ML \tag{4.5}$$

Where:

MTT = mean travel time, VOT = value of MTT

ME = mean earliness, VOE = value of ME

ML = mean lateness, VOL = value of ML.

This is based on the modelling experience described in chapter 3. The experiment included the following stages:

- A set of all feasible combinations of VOT, VOE and VOL was generated. The values are in the range from -10 to +30 ppm, with intervals of 0.1 ppm. Any WTP below -10 or above +30 was considered unlikely when the SP survey was designed and there was therefore no point in accounting for such values in the current experiment, as the survey is unable to identify them. Note that, as discussed in detail earlier, there is a debate in literature about whether or not negative WTP should be strictly rejected, even for a small share of the population. In the current experiment we allowed negative values in order to demonstrate that a "strictly positive WTP" might be the policy of the modeller, but it is not necessarily supported by the data.
- For each single response in the database, it was examined whether the recorded choice made by the respondent can be explained by each feasible combination of VOT, VOE and VOL. In other words, for each response and each combination of WTP it was checked whether the total cost of the chosen alternative is lower than the total cost of the alternative that was not chosen.
- If travellers were entirely consistent in making their choices, we could now check what WTP can successfully explain *all* nine choices made by each respondent. But since we do not expect the respondents to be perfectly consistent, we define a *consistency ratio* (CR) as the ratio of the number of responses in which a certain WTP is accepted to the total number of responses of a specific individual. For example, if a certain combination of VOT, VOE and VOL is accepted by a traveller at CR=0.75, it means that at least 75% of the

responses of this traveller can be explained by these values. Now, for each respondent in the database, we mark all WTP levels that can be explained at CR=1, CR=0.75 and CR=0.5.

- Generally, the choices of each respondent can be explained by many different combinations of VOT, VOE and VOL. Because of the linear nature of the cost function, the accepted values of each one of the three WTP elements form a single segment with global minimum and maximum. For each respondent and for each CR we now calculate the minimum and the maximum accepted levels of VOT, VOE and VOL.
- Finally, at each CR, we examine how common every feasible level of VOT, VOE and VOL turned out, and draw the respective cumulative frequency curves.

The number of respondents, out of the 250 travellers in the dataset, for which a range of WTP could be identified, is as follows -

- At CR=1: 133 respondents (53.2%),
- At CR=0.75: 225 respondents (90.0%),
- At CR=0.5: 249 respondents (99.6%).

The fact that at CR=1 it is not possible to identify the range of WTP for many respondents is not surprising; we did not expected respondents to be consistent. The share of respondents for which a range can be derived at lower CRs seems reasonable, but the main question that arises is how wide the range of WTP is when the acceptance criterion is not strict. The question is answered in figures 4.9, 4.10 and 4.11, which present cumulative frequency curves of the minimum and maximum VOT, VOE and VOL. Note that the experiment examined three-dimensional combinations of the WTP elements and not each element alone, but the figures present the minimum and maximum in a one-dimensional WTP space, for the simplicity of the presentation. Each presented curve is in fact the projection of the extremum contour of a diagram where all three dimensions are presented.

We saw that at lower CRs it is easier to identify the range of WTP; but figures 4.9-4.11 demonstrate that the lower the CR, the wider the recognised range. The range of VOT at CR=1 (the solid curves in figure 4.9) is defined properly for most of the respondents; for around 18% of them the lower or upper bounds of the VOT are outside the examined range. This is only slightly different from the respective finding in the simpler

case discussed by Fosgerau (2006). However, the ability to identify a reasonably-narrow range of VOT at lower CRs is much poorer: at CR=0.75, the upper or lowers bounds of hardly half of the respondents are within the examined range.

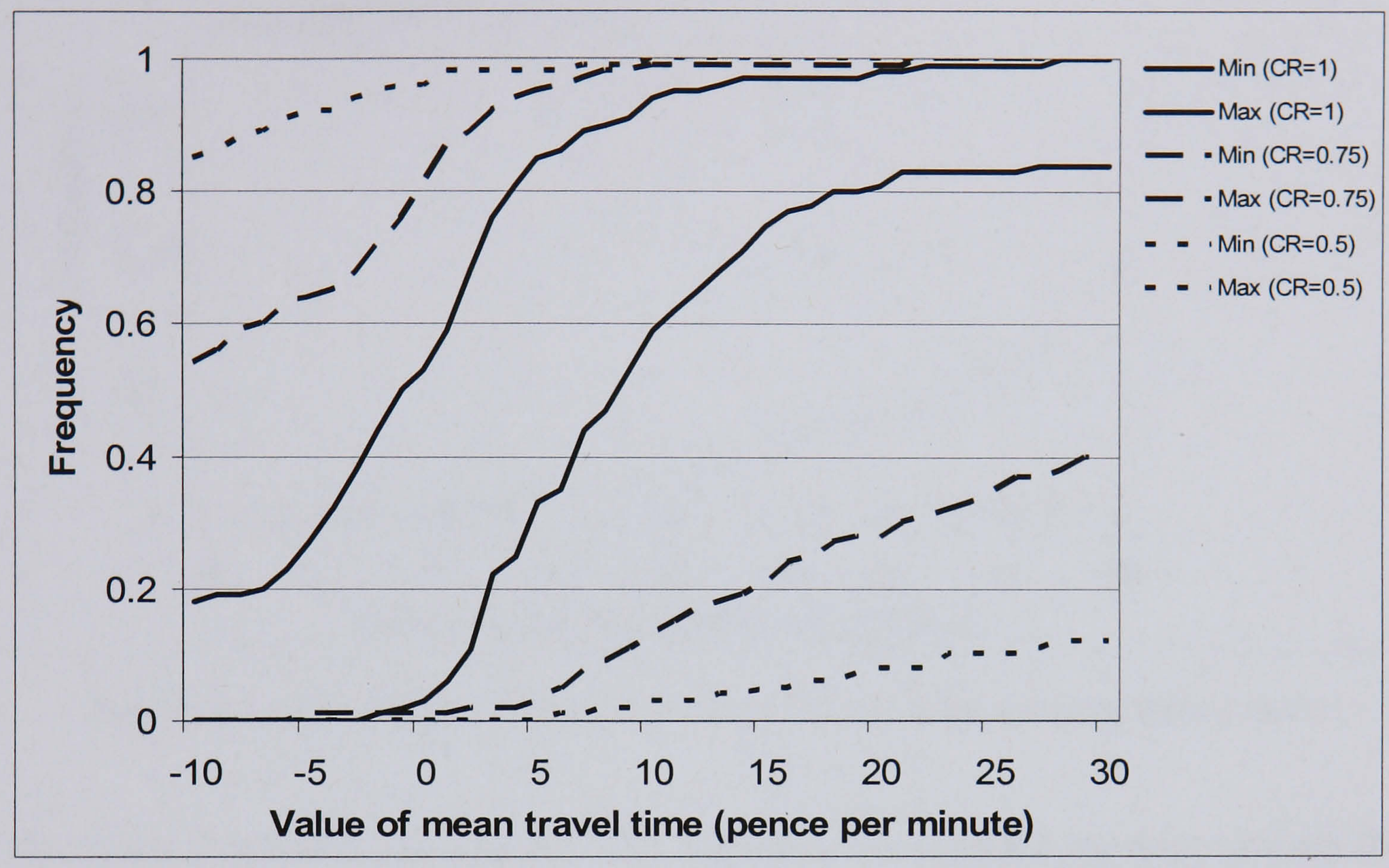


Figure 4.9: Cumulative frequencies of the VOT at different consistency ratios

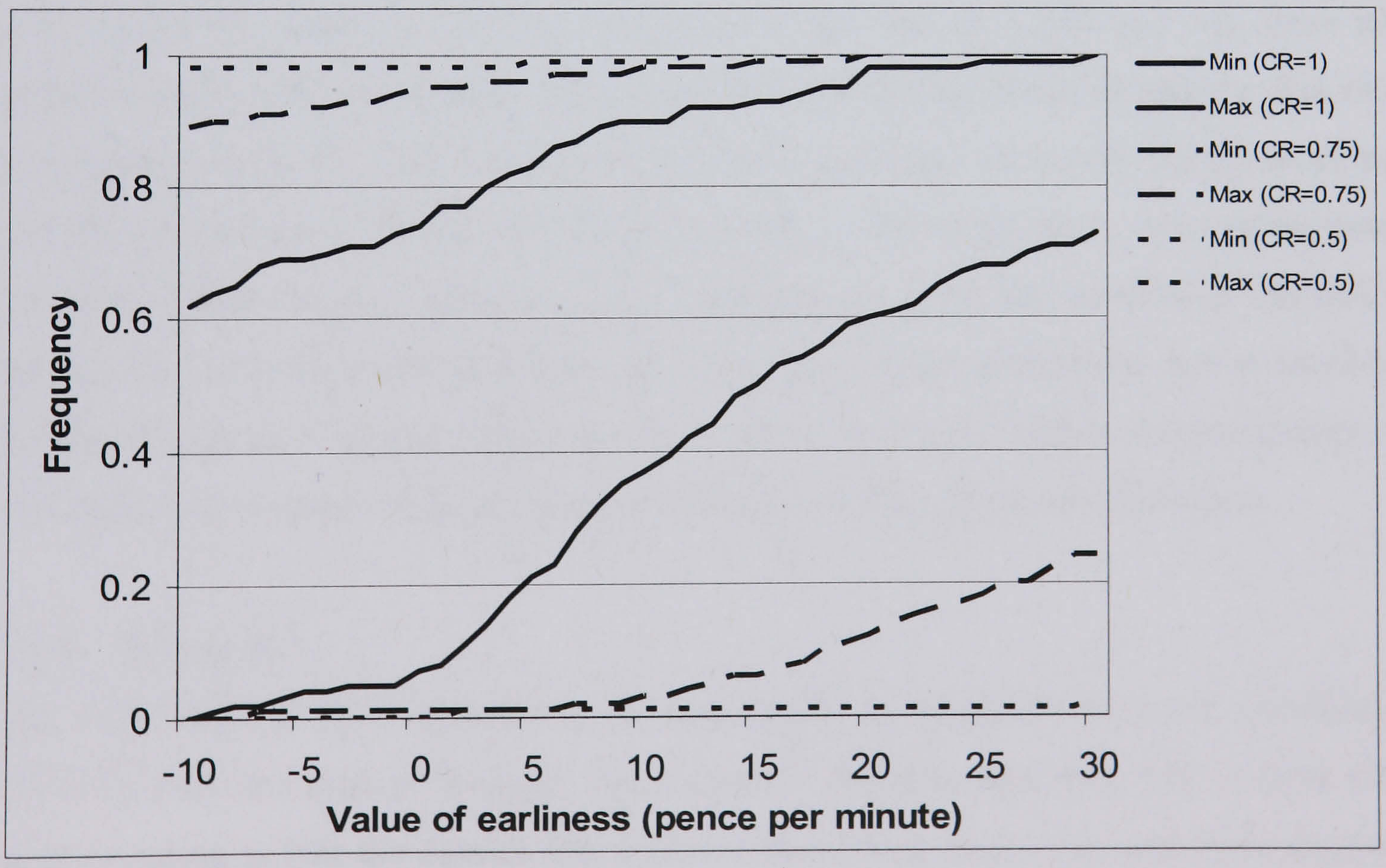


Figure 4.10: Cumulative frequencies of the VOE at different consistency ratios

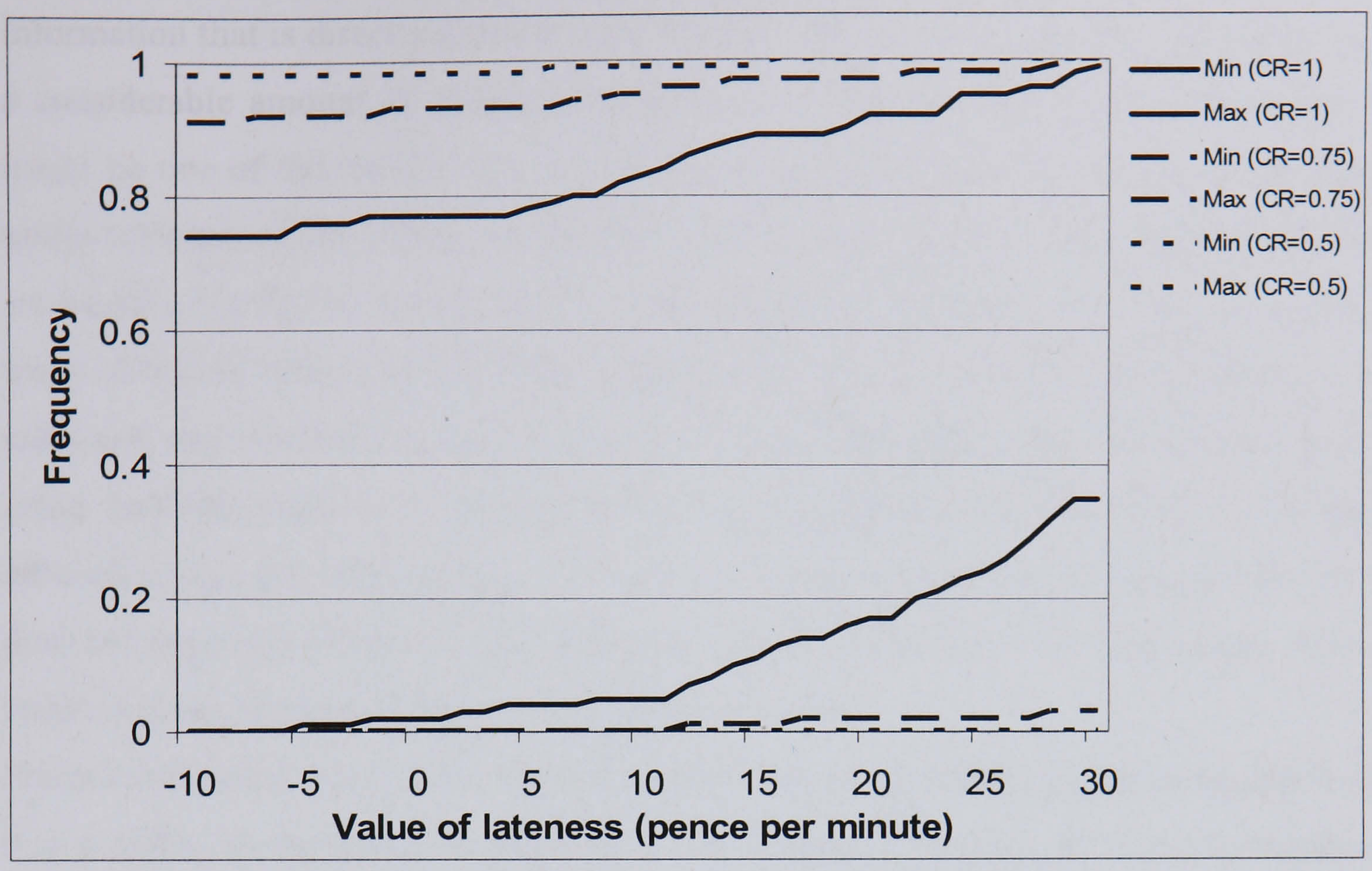


Figure 4.11: Cumulative frequencies of the VOL at different consistency ratios

The width of the range of VOE and VOL appears an even bigger problem: not only the curves for the lower CRs, but also those for CR=1, leave a very wide space between the minimum and the maximum WTP. The difficulty in estimating stricter boundaries for VOE and VOL might be partially explained by the fact that ME and ML were not explicitly used to determine the attribute levels in the design of the SP survey. But still, re-examination of the choice situations that were presented in the SP survey confirms that all the relevant variables – MTT, ME and ML – vary to an extent that encompasses the entire range of likely values. An equal conclusion, about the inability to determine the boundaries of the DWP despite an SP design that seems appropriate, was reached by Fosgerau; and as discussed earlier in this section, due to the higher dimensionality of our analysis it is natural that this problem is bigger in the current circumstances.

4.5.3. Discussion

The wide range of WTP revealed in the experiment described above is not a particular problem with this type of analysis, but a general problem with this type of data. Our understanding is that the dataset that includes the survey responses, and other datasets based on similar surveys, truly do not contain more information that this. The special

feature of the current analysis (similar to the analysis by Fosgerau) is that it shows only information that is directly derived from the data. It reveals that the data leaves us with a considerable amount of freedom to determine the exact DWP curve. This freedom might be one of the reasons for the difficulties that are repeatedly found in obtaining stable estimates of the DWP, and for the big differences between estimates of DWP that are based on different specifications of the parameter distributions. There are clearly many different forms of the DWP curves that can fit into the space between the minimum and maximum curves in figures 4.9-4.11. Parametric estimation of the DWP, using methods such as the maximum likelihood, suggests a likely DWP curve that efficiently uses the information in the dataset; such estimation is very powerful, but it does not explicitly illustrate that a suggested DWP is also based on information whose origin is the model specification and not the data itself.

Practitioners often assume that it is not essential to have estimates of the entire DWP, as it is possible to base scheme appraisal on the estimated preferences of most travellers rather than all of them. It is important to note that if we cannot identify the entire distribution, estimates of the mean or the median are not reliable; therefore the entire shape of the DWP, including extreme values at the far ends, is crucial even if there is only interest in the mean.

This conclusion about the difficulty in identifying the DWP points towards two different directions for further discussion. First, there is a clear need to make any effort to extract all the possible information from the survey data. In the presented experiment some information that exists in the survey data was not used: in cases where a certain level of WTP can explain some of the choices made by a certain respondent, but not as many choices as required by the preset CR, the implications of these choices in determining the likely range of WTP have so far been ignored. Several ideas were brought up as possible ways to employ the unused information to narrow the range of identified WTP:

1. Instead of deriving different WTP curves at different CR levels, it is possible to derive a single WTP curve where the CR is used as a weight. Namely, any level of WTP is accepted, but its frequency is reduced by multiplying it by the CR. We performed this experiment but chose not to present it here, since the resulting range of feasible WTP is even wider than in the presented experiment. This is a consequence of a very high number of WTP levels that were accepted at low CRs.

2. Assume that between the values for the minimum and maximum WTP of each person, there is a single, unknown value which is the most likely WTP, and assume that each person has an individual DWP around this most likely value. The individual-specific distributions can be derived from the curves we already have, based on the assumption that the minimum and maximum distributions we calculated earlier are actually estimates of a collection of the lower and upper boundaries of the individual-specific distributions. However, although this analysis enables more intensive use of the survey data, it requires making assumptions regarding the shape of the individual-specific distribution, while the main incentive of the current analysis is our wish to avoid unnecessary assumptions. This concept was therefore not taken any further.
3. Assume, alternatively, that between the values for the minimum and maximum WTP of each person, all levels of WTP are equally likely. The individual-specific WTP is uniformly distributed, but each person has a different range with a different bandwidth. It is possible to derive from our data file estimates of the individual bandwidth; this is only an upper bound for the real bandwidth, since the real bandwidth (which is narrower) is not fully identified. The bandwidth information has not been used so far and therefore, we could in principle use it to try to narrow the DWP. However it was eventually decided not to use this option, too, since the bandwidth itself is identified for only part of the sample, which means that some undesired distributional assumptions should be made.

Overall, no satisfactory way was devised to reduce the width of the estimated range of WTP, even if some of the information in the data file has not been used. Note that the fact that we look at each individual separately, accounting for the panel data nature of the survey results, is in itself a source for loss of data: if we treated each response separately, there would be no need to ignore responses where some levels of WTP are accepted at a low CR. Observe also that in our experiment, the choice of CR in fact involves a trade-off between an efficient use of the data and a reasonably narrow DWP, since high CR implies that fewer WTP levels are accepted but it also implies that much information is not being exhausted. Nonparametric methods that can make a more systematic use of the data do exist, as illustrated by Fosgerau (2006). However, it seems that existing methods require very high mathematical proficiency and still only perform very basic WTP analysis. All in all, much remains to be investigated in this field.

The second direction for further discussion is towards a better specification of future surveys. Designing the survey held in the current study involved a challenging combination of several difficulties, but at the stage of survey design we were not fully aware of their severe consequences, as these difficulties had hardly been discussed in preceding studies. The nature of the modelled phenomenon required displaying a sequence of travel times for each choice alternative, as well as some other attributes; during the design it was still unknown what variables these attributes will turn into in the final model. The design problem is multidimensional, not only because there are several WTP elements, but also because the determination of appropriate attribute levels should account for any potential variable, even if it is not presented. The risk in designing the questionnaire for a multidimensional problem is that in some of the dimensions not enough information is collected. On the other hand, because of the relatively detailed presentation, the design is also sensitive to the risks of task complexity. Hence, simple ways of coping with the high dimensionality might reduce the credibility of the results due to the effects of boredom or fatigue, as discussed in detail by Hensher (2004) and Caussade et al (2005). These simple ways to extract more data might include any of the following:

1. Presenting more than two alternative options in each question, and asking the respondents to rank the alternatives instead of just choosing one. There are available methods for modelling based on ranking preferences, as used for instance by Bates et al (2001), which are indeed efficient in terms of the amount of information obtained from a given number of respondents. In our survey, where the amount of detail required for defining each alternative is relatively large, presenting more than two alternatives per choice situation might have constituted a considerable burden on the respondent. In addition, the conversion of the ranking data into econometric insights must be based on a set of assumptions that ascribe different probabilities or levels of likelihood to each rank, and this contradicts the nonparametric nature of the current analysis.
2. Presenting more questions. However, as we discussed, the number of choice situations used here is repeatedly mentioned in similar studies as appropriate to avoid reduced credibility of the responses towards the end of the questionnaire.
3. Increasing sample size. The number of respondents in our survey is not significantly different from the sample size used in similar studies (see review earlier in this chapter), and could not be increased due to budget constraints; still, approaching

more respondents seems the only straightforward way of reaching satisfactory estimates of the DWP.

The difficulty in obtaining trustworthy estimates of the DWP stresses the need for more sophisticated surveying techniques. Some basic needs of survey design for MXL models have been statistically specified by Cirillo (2005), who states that "the studies on optimal design for MXL models are in their primordial phases". The main need that future works will have to meet is for surveying methods that enable getting more information from each respondent but without causing the negative effects of complicated surveys.

4.5.4. Summary

The simple experiment presented in this section derived the DWP nonparametrically, directly from the raw dataset. When we relax the distributional assumptions normally made in parametric analyses, it is found that the amount of information that the survey data truly contains is limited. The common tool for estimating MXL models, namely the maximum simulated likelihood, is indeed flawless and powerful in pointing to the best model even with a small amount of data; but this power can mislead us to believe that the derived estimates of the DWP are entirely based on the input data. The special features of the current case study, and the multidimensional space of survey attributes in particular, make this case especially sensitive to insufficient amount of input information. The needs of MXL modelling procedures, and their implications in the design of SP surveys, have hardly been discussed in previous literature. While we realise that this made the design of our own survey imperfect, it also led us into some important insights that might prove useful in the design of future experiments.

4.6. Best-practice estimates

Despite the extensive analysis performed in the previous sections of this chapter, the main objective is still not met, as we still do not have satisfactory estimates of the DWP for reduced TTV. The various MXL models did not perform rationally, either in the vertical dimension, the horizontal dimension or both. The SUS estimates made good sense but given the inexact nature of this technique, there is no evidence that they

accurately represent the true behaviour. The additional nonparametric experiment showed that our dataset indeed does not contain a sufficient amount of information to allow full identification of the DWP. The scope of this study does not enable collecting more data or examining the existing data in new directions. Following all this, is it now necessary to decide whether and how to obtain the desired estimates of the DWP.

As it was found harder than expected to estimate the DWP, one option is to return to the Multinomial Logit model estimated in chapter 3 and to the uniform WTP derived from it. Using WTP estimates that do not vary between travellers would be the safest option in terms of statistical confidence, as the data requirements for estimating such WTP are modest, and the estimates reached in chapter 3 appear reliable enough. However, as discussed earlier, uniform WTP is not a realistic representation of the heterogeneous composition of the population of travellers. In this respect, even an estimate of the DWP that encompasses *most* travellers, rather than all of them, would be a significant improvement compared to a uniform WTP, as long as it can be identified properly; this is the rational behind the final estimates of the DWP, brought in this section. The DWP sought here is a conservative estimate, as it excludes any range of monetary values in which there is no sufficient statistical confidence; but it is more realistic than the MXL models, and it covers a large share of the population in greater detail than the Multinomial Logit model. It can hence be seen as a compromise.

Of the three different concepts we used to analyse the raw data (in sections 4.3, 4.4 and 4.5), the SUS approach was the one that led to the most sensible estimates; we therefore base the best-practice DWP estimates on a new SUS experiment. SUS-based DWP has a major disadvantage: we do not have any external estimate to compare it to, as a test of fit. This is the main reason why the SUS experiment carried out here is more selective than the one in section 4.4. Sub-models that do not perform well enough are not included this time in the group of sub-models from which the DWP is computed.

In the SUS experiments in section 4.4, each DWP estimate was based on 200 subsamples; each subsample contained about 110 responses. Since the multiple sub-models were based on such small samples, their statistical fit was not expected to appear as good as it would be in an equivalent model based on a full sample. In about 75% of the sub-models, all t-test values were above 1.5, and in about 99% of them all t-test values were above 1. In the context of a SUS experiment, where each sub-model is an element in a big sample, this was considered sufficient. Due to the abovementioned reasons, the current experiment is more prudent, and hence only sub-models where all t-

test values are above 2 are used to derive the DWP. In addition, several sub-models that were included in the early SUS experiments are omitted in the current experiment, because of various suspected faults. The new SUS experiment uses the same 200 sub-models as in section 4.4, but as a result of this rigorous selection, the number of effective sub-models is reduced to 100. As in the earlier SUS experiment, the WTP is computed from each sub-model separately and then the DWP is derived. Separate distributions are calculated for VOT, VOE and VOL.

The selective SUS experiment discards any sub-model whose plausibility seems questionable; naturally, the omitted sub-models are mainly those that point to extreme WTP estimates. As mentioned above, it is very likely that the range of WTP revealed by the selective SUS experiment is narrower than the true range. Still, owing to the difficulties encountered at earlier stages, it was decided that it would be better to estimate a DWP that ignores travellers whose behaviour is radically different from the average, than to derive a very wide DWP that comprises some values that do not really exist.

To make the resulting DWP curves easy to describe in the forthcoming application, we now try to find simple curves from common families of distributions (lognormal, normal etc.) that can reasonably replicate them. Describing the SUS curves this way turns this nonparametric experiment into a parametric one. One justification for this is that since a simple curve is fitted to the SUS-based curve *after* the estimation, we are not imposing any preset shape on the DWP, only trying to smooth the SUS-based curve (as opposed to maximum simulated likelihood estimation, where the general shape of the distribution is determined *before* the estimation). Another justification is that as we saw, our dataset does not contain enough information for full nonparametric identification of the DWP, and therefore accepting the shape of an existing curve is unavoidable. The search for the smoothed curve that fits each SUS-based curve best is carried out using a computer program that tries many different curves, calculates their fit using K-S test, and settles on the curve that minimises the test statistic. The selective SUS curves for VOT, VOE and VOL and the chosen smoothed curves are presented in figures 4.12, 4.13 and 4.14. Table 4.6 shows the parameters of the chosen curves and the results of the comparison between the original and the smoothed curves. The K-S statistic values for all three curves are satisfactorily low.

Note that despite the discussion made earlier about the feasibility of negative WTP for a certain part of the population, two of the three chosen DWP curves are lognormal, and

hence do not allow any negative values. The lognormal curves were found best in replicating the selective SUS curves primarily because most sub-models that indicated negative WTP did not meet the rigorous requirements for being included in the selective SUS experiment. It might also seem surprising that a curve that straddles zero was chosen only for VOE, while the earlier discussion of negative WTP mainly referred to the VOL curve. This is probably because in the early SUS experiment, the mean earliness was included in the same variable as the mean travel time, and therefore any unique features of the VOE curve were harder to reveal.

The noticeable negative tail of the new SUS curve for VOE seems rational in the sense that it is indeed likely that a small group of travellers strongly prefers to arrive early. But it should be observed that travellers that have positive VOT, positive VOL and negative VOE will tend to always choose the earliest possible departure, even if it departs irrationally early. The reason why the estimation procedure did not prevent this from occurring is presumably that in the dataset that these estimates are based on, the number of choice situations with exceptionally early departures was not very high. We do accept the distribution of VOE, including its negative tail, as our best estimate of VOE; but we find that it should be implemented in conjunction with a constraint that verifies that among the various departure times that are early enough to guarantee arrival to the destination before the desired arrival time, travellers with negative VOE choose the latest.

	VOT	VOE	VOL
Distribution	Lognormal	Normal	Lognormal
Mean (ppm)	5.0	2.5	20.8
Standard deviation (ppm)	1.6	4.4	2.1
K-S statistic	4.4%	8.7%	5.6%

Table 4.6: Smoothed curves fitted to the SUS-based curves

All in all, despite the weaknesses of the analysis used here to derive the curves in table 4.6, we see these as the best estimates of the DWP for reduced TTV that can be reached in this study.

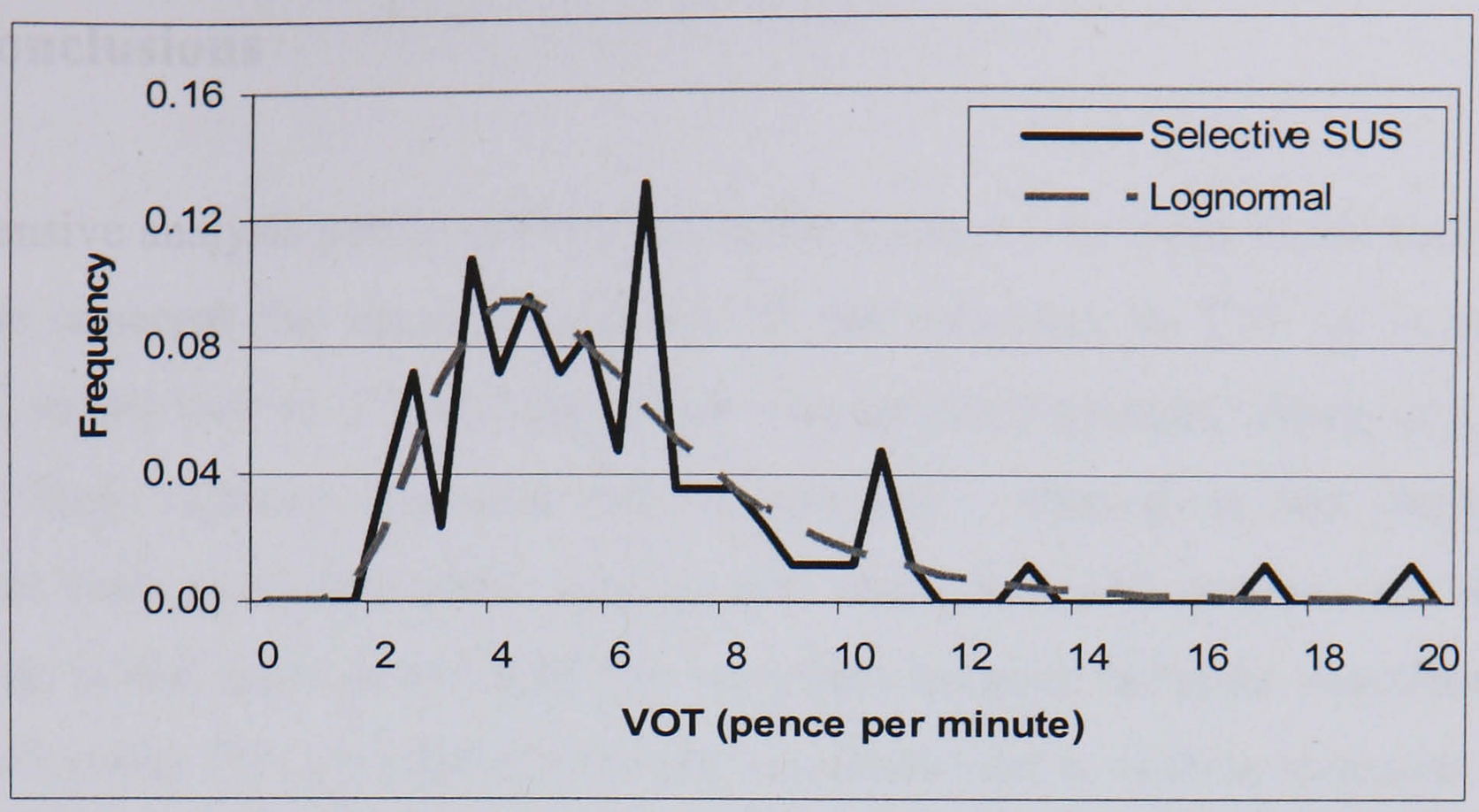


Figure 4.12: Best-practice distribution of VOT

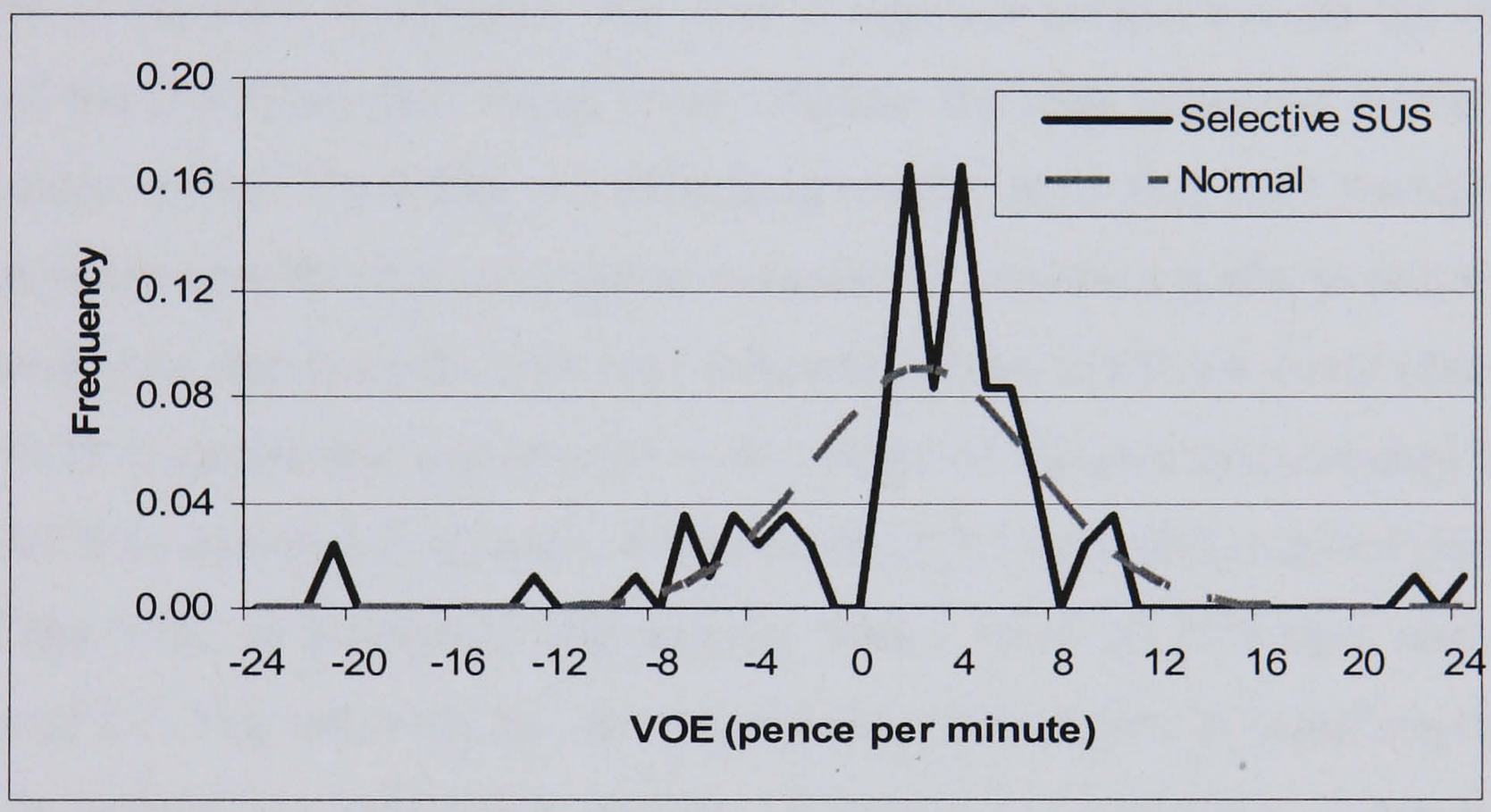


Figure 4.13: Best-practice distribution of VOE

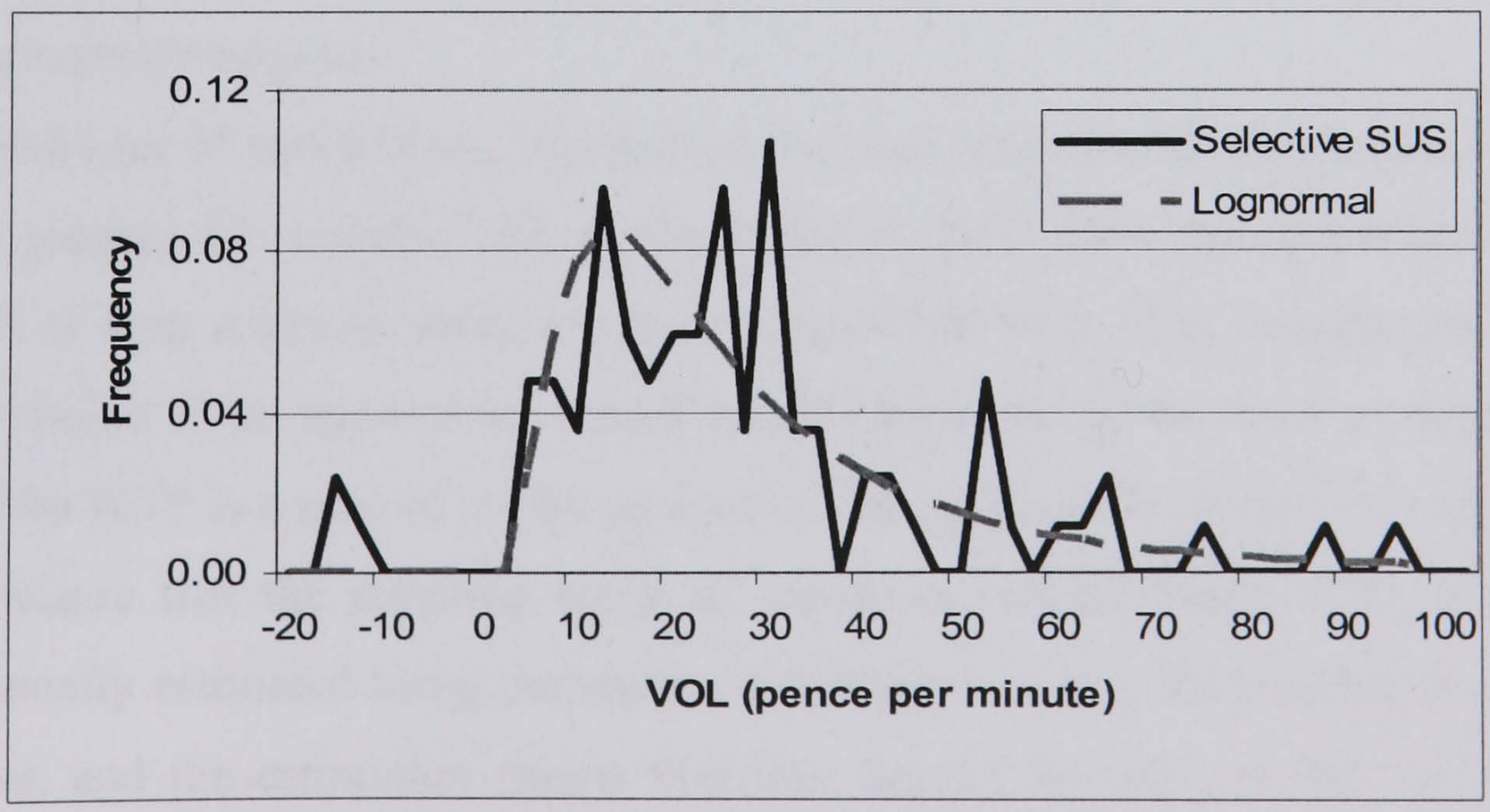


Figure 4.14: Best-practice distribution of VOL

4.7. Conclusions

The extensive analysis performed in this chapter leads to two types of conclusions. The first type concerns the range of attitudes of bus travellers to TTV as an economic problem, rather than as a mathematical (or econometric) problem. Disclosing the true range of these attitudes is a main issue in this thesis, although in this chapter it has somewhat been put aside while dealing with more technical matters. An important conclusion is that there is no doubt that variations between different travellers in their attitudes towards TTV do exist. Obviously, travellers tend to wish to minimise the time spent travelling or any deviation from their desired arrival time to the destination. We found some evidence that not all travellers value savings in the mean travel time or in the extent of earliness or lateness. But from a rigorous perspective on the estimation process of the DWP, we only found solid evidence for such behaviour in the attitudes towards early arrival. Therefore, the modelling results imply that *most* travellers prefer not to arrive too early to their destination, whereas *all* travellers prefer to minimise their mean travel time and lateness. The best estimates of the DWP we could obtain imply that the VOT is distributed lognormally with a mean of 5.0 ppm and standard deviation 1.6; the VOE is distributed normally with a mean of 2.5 ppm and standard deviation of 4.4; and the VOL is distributed lognormally with a mean of 20.8 ppm and standard deviation of 2.1. This confirms the conclusion reached in chapter 3, regarding the severe manner in which any additional minute of lateness is regarded, compared to an additional minute of earliness or of the mean travel time. These estimates of the DWP will be used in chapter 8 to illustrate the role that TTV should have in the assessment of bus infrastructure schemes.

The second type of conclusions, that gained the main attention in this chapter, has to do with the general econometric task of deriving the DWP from the responses of an SP survey. It is very common today to estimate the DWP from MXL models, but we find that the choice of an appropriate model specification should be done cautiously. The fact that the WTP is a ratio of model parameters brings in much sensitivity, and makes it hard to assure that the resulting range of values is well-bounded. MXL models are conventionally estimated using parametric techniques, such as the maximum simulated likelihood, and the estimation results therefore heavily depend not only on the input data but also on the model specification. Attempts to orient the modelling outputs

towards directions that we see as more sensible, for example by using constrained distributions, might be unsuccessful - either because constrained models are mathematically harder to estimate or because the existing estimation tools are unable to identify the improvement in model fit. Future research might make the modeller's life easier by recognising better ways to specify the MXL model in the first place, or by developing improved estimation procedures, or by introducing new tools for testing the goodness of fit. The newly-introduced concept of estimation in the WTP space is another promising avenue for further investigation.

Due to the failure to identify satisfactory MXL models, a series of nonparametric experiments was carried out. First, we used a SUS technique to obtain a crude, but yet useful, independent estimate of the DWP. Unlike previous studies, we do not recommend using SUS simply to start the MXL estimation process with an intelligent guess of the distributions of the parameters, because re-estimation of the parameters makes such use inappropriate. We found that to a great extent, model fit as implied by the SUS estimates contradicts what is implied by the maximum likelihood. We do not suggest here that any of these is necessarily more correct than the other, only that it is advisable to seek confirmation of model fit in other ways but the conventional ones. We also remind that when a MXL-based DWP is compared with an external estimate, it is important to verify good fit in both the vertical and the horizontal dimensions. The nonparametric experiments performed here also looked at the amount of information that truly exists in our raw dataset. Although the sample size and the collection method complied with the common practice in this field, it was found that the amount of available information is far from enabling identification of the entire DWP. We learn from this again that great care should be taken when making conclusions from datasets of this nature, especially in problems that involve a multidimensional WTP space. While it is important to conclude that awareness to the needs of MXL modelling is crucial from the very early stages of SP design, our findings also stress the need for developing more sophisticated methods of data collection.

Chapter 5

Selected topics from the literature in traffic modelling

5.1. Introduction

This chapter opens the second main part of the thesis, where our attention is shifted from the attempt to understand the behaviour of the users of the transport system to an attempt to understand the performance of the system itself. In this chapter and the two chapters that follow, the perspective of an economic analyst is replaced with the point of view of a traffic modeller, who wishes to be able to anticipate what the condition of the transport network will be like under various circumstances.

Following the analysis in chapters 3 and 4, we now have a way to convert a given level of TTV to monetary values. But to carry out full assessment of any suggested change in the transport infrastructure, we also need an estimate of the level of TTV itself. In other words, we have estimated the *costs per unit* of some indicators of unreliability, but we still need to estimate the *number of units*. Tools for estimating the level of TTV are not as common as tools that calculate the MTT, and therefore, in the forthcoming chapters we try to develop such a tool. Note that we seek a method for predicting the level of TTV, even though our economic analysis found that TTV only has an indirect effect, through its influence on how early or late travellers get to their destinations. The amounts of earliness and lateness depend on the extent of TTV, and therefore estimation of TTV is required even if it is later only used indirectly. The conversion of the estimate of TTV into the extent of earliness and lateness is demonstrated in chapter 8.

The next section of this chapter examines previous studies that aimed at developing tools for estimating the level of TTV. The following section includes a thorough review in an area that at this point might seem irrelevant to the problem of predicting TTV: methodologies for the calibration of traffic microsimulation models (TMMs). The motivation for our interest in such methodologies is that the following chapters introduce the concept of estimating the level of TTV using a TMM, and show that in order to make a TMM suitable for this task, it must go through a purposely-designed calibration procedure. The concept that combines the objective (predicting TTV) with the tool (TMMs and their calibration procedure) is introduced later in the thesis. In this

chapter, we explore the role that this objective and this tool, as two separate topics, had in previous studies.

5.2. Prediction of travel time variability

5.2.1. Travel time variability as a function of other attributes

Tools that can generate forecasts of the level of TTV in various hypothetical scenarios are a requisite for evaluation of the benefits from improved reliability. In principle, estimates of TTV do not necessarily have to be the direct output of a model especially built for this purpose: it is theoretically possible to separately estimate travel times in multiple days, or in a big range of settings, and then analyse the variability between them. Mohammadi (1997b) expresses such approach by including models for estimating MTT in a review of methods for making TTV forecasts. However, existing models for estimating MTT are not sensitive or detailed enough to make the repeated estimates of MTT different from each other in a way that truly replicates the real distribution of travel times. Tools purposely designed for the prediction of TTV, and other related works, can be grouped into three categories. The first category is described here and the others in the following sub-sections.

Several works calibrate models where TTV, expressed as the standard deviation of travel times, is a function of the MTT or other attributes. Such models are presented in table 5.1. All the models of this type are empirical in nature, and do not attempt to establish any theoretical foundations of the mathematical relationship they present. Dale et al (1996) identify two different types within these models:

1. Models that estimate TTV of the complete journey. A level of TTV is ascribed to each feasible route through the network.
2. Models that estimate TTV on individual links or junctions.

Although the authors identify only the first type of models as aggregate, in fact both types are based on an aggregate, macroscopic perspective, since they do not look at the behaviour of individual drivers or at the performance of individual vehicles.

All the models in table 5.1 find that raising the MTT to different powers between 0.49 and 1.17, and incorporating the result in a relatively simple linear function, can approximate the level of TTV; but the models differ from each other in any other aspect. This is hardly surprising, since they are based on different assumptions and validated in

diverse test beds. None of the models has been recognised as contributing to the understanding of TTV as a general phenomenon, or gained wide acceptance.

Source	**Mode**	**Expression**	**Comments**
Herman and Lam (1974)	Car	$TTV = 0.36\ MTT^{0.49}$ (to work) $TTV = 0.31\ MTT^{0.70}$ (from work)	Individual road links
Polus (1979)	Bus	$TTV = MTT^{0.5} / 2.478$	Entire journey
May et al (1989)	Car	$TTV = 0.38 + 0.2\ MTT$ (spring) $TTV = 0.41 + 0.14\ MTT$ (summer)	Individual road links
Linaritakis (1995)	Bus	$TTV = 1.049\ MTT^{0.67}\ SP^{0.37}\ (V/C)^{0.36}$ *SP* is the mean bus travel speed. (*V/C*) is the ratio of flow to capacity of the general traffic along the examined corridor	Entire journey
Dale et al (1996)	Mixed	$TTV = 0.0778\ (MTT - TT_0)^{1.166}$ for motorways $TTV = 0.4020\ (MTT - TT_0)^{0.893}$ for other roads TT_0 is the free-flow travel time. Separate models developed for 17 types of junctions	Individual road links, not including junctions. Correlation coefficients are proposed for converting into TTV of entire journey. Model not calibrated with real data
Mohammadi (1997b)	Mixed	Different models for varying vehicle types, journey purposes, times of day, road types, weather conditions and so on. Form of most models: $TTV = a \cdot MTT$ (a constant), or: $TTV = a \cdot MTT + b$ (a, b constants)	Entire journey

Table 5.1: Models that present TTV as a function of other attributes

5.2.2. Fitting a probability density function to travel times

The need to model the unpredictable nature of travel times was tackled in numerous research works by trying to fit a probability density function to observed travel time data. Studies that attempted to determine which common families of statistical distribution curves match observed TTV are reviewed in table 5.2. There are major differences between the datasets used in the various studies: each study focuses on different combinations of road types, times of day, urban or interurban surroundings, levels of congestion and so on. Different authors also report varying levels of statistical significance of their findings. The way TTV is defined varies too: some of the works focus on the distribution of travel times over different vehicles travelling concurrently; some others look at the day-to-day distribution; and many of the studies either combine more than one dimension of variation or simply do not give any clear explanation in this issue. All in all, the diversity of findings is not surprising. It can be generally observed, though, that most researchers choose asymmetric distributions, such as the lognormal or exponential. Apparently this is because it is impossible for the journey to take less time than in an ideal, free-flowing journey, but there is no equivalent limitation on the maximum trip length.

Note that even if sufficient fit is found between a particular probability density function and the observed distribution of travel times, this is not necessarily sufficient as a tool for predicting TTV. Prediction based on a known distribution is performed by drawing random numbers from this distribution; but in such procedure there is no account for local factors, such as the detailed configuration of the road, that make different parts of the distribution suitable for different settings. Unfortunately, when estimating TTV in the context of transport appraisal, sensitivity to these factors is vital.

Source	**Mode**	**Proposed probability density function**	**Based on empirical analysis?**
Gaver (1968)	Car	Exponential	No
Hermand and Lam (1974)	Car	Normal	Yes
Polus (1975)	Bus	Beta	Yes
Richardson and Taylor (1978)	Car	Lognormal	Yes
Turnquist (1978)	Bus	Lognormal	Yes
Anderson et al (1979)	Bus	Lognormal	Yes
Polus (1979)	Car	Gamma	Yes
Taylor (1982)	Bus	Normal	Yes
Mogridge and Fry (1984)	Car	Lognormal	Yes
Guehthner and Hamat (1985)	Bus	Gamma	Yes
Hall (1985)	Bus	Exponential	No
Mora Camino et al (1986)	Car	Shifted Gamma	Yes
Golob et al (1987)	Car	Lognormal (only incident delay)	Yes
Talley and Becker (1987)	Bus	Exponential	Yes
Giuliano (1989)	Car	Lognormal (only incident delay)	Yes
May et al (1989)	Car	Normal	Yes
Bookbinder and Desilets (1992)	Bus	Shifted truncated exponential	No
Strathman and Hopper (1993)	Bus	Lognormal	Yes

Table 5.2: Studies that fit a distribution to travel time data

Source	Mode	Proposed probability density function	Based on empirical analysis?
Mei and Bullen (1993)	Car	Shifted lognormal	Yes
Noland and Small (1995)	Car	Two alternatives: uniform and exponential	No
Mohammadi (1997a)	Car	Lognormal	Yes
Noland (1997)	Car	Lognormal found best but exponential is used	No
Cohen and Southworth (1999)	Car	Gamma (only incident delay)	No
Emam and Al-Deek (2005)	Car	Lognormal	Yes

Table 5.2 (continued): Studies that fit a distribution to travel time data

5.2.3. Other methods

Very few other methods for estimating TTV were found in literature. An interesting approach has been recently proposed by Van Lint and Van Zuylen (2005), who developed an artificial neural network model based on historical day-to-day distributions of travel times. The parametric functions of the neural network model can be used to predict the 10^{th}, 50^{th} and 90^{th} percentiles of the distribution of travel time. Compared to the simpler approaches described earlier, the mathematical complexity of this model makes it potentially more sensitive to the particular characteristics of the different settings in which it can be used. However, it should be noted that similar to the models described in the previous sub-sections, the TTV estimates in the neural network model have no foundation in a behavioural (or other) theory. The model parameters are merely elements in an empirical expression, which do not stand for clear features of the real network. Since there is no theoretical basis, it is not clear how transferable the model is to situations that are different from those that were recorded in the historical

data used for calibration. The potential use of such model as a general tool for prediction is therefore limited.

5.2.4. Summary

This section has included a review of existing approaches to the prediction of TTV in hypothetical scenarios. The common feature of many of the reviewed works is that they try to incorporate TTV in a simple mathematical expression. The fact that hardly any consistency is found between these works might signify that this concept is too simplistic: TTV is not a straightforward function of MTT and it also does not systematically behave according to a well-known probability distribution curve. Even if in a specific time or location such relationship is found, there is no reason to assume that it will hold in other settings.

It is quite surprising that none of the reviewed approaches attempts to estimate TTV in a procedure that considers the causes of TTV. As discussed in more detail later in this thesis, it is clear that not *all* the causes that contribute to the creation of TTV can be modelled; but it seems worthwhile to examine to what extent it is possible to capture a share of the variation by taking direct account of even just *some* of these causes. The failure of the simple, aggregate methods to find an empirical formula for TTV that works in multiple surroundings suggests that TTV is a phenomenon whose magnitude heavily depends on local conditions and on the disaggregate elements of the network. The main incentive for the experiments described in the forthcoming chapters is an interest in how well TTV can be reproduced by microscopic traffic modelling.

5.3. Calibration of traffic microsimulation models

5.3.1. Background

The methodology introduced in the next chapter is based on intensive use of a TMM for the purpose of estimating the level of TTV. We discuss later that such use is only valid if it is preceded by a special calibration procedure of the TMM. The formulation of a new calibration method has a major role in the remaining part of the thesis, and it is therefore important to carefully examine other TMM calibration approaches; this is the main objective of this section.

TMMs are commonly used by researchers and practitioners for a detailed analysis of the performance of transport systems. Microscopic traffic analysis attempts to replicate the dynamics of a real transport system: it generates actions made by users of the system, lets the different actions interact with each other and examines the cumulative effects of these interactions. A key feature of microscopic traffic modelling is that whether the network of interest is as small as a single junction or as big as an entire metropolitan area, the estimates it generates are based on explicit representation of individual behaviour. The behavioural patterns that TMMs try to capture range from the driver's choice of route and departure time (e.g. Hu and Mahmassani, 1997; Liu et al, 2006) to various aspects of the actual driving behaviour, such as vehicle following and lane changing (e.g. Nagel, 1996; Yang and Koutsopoulos, 1996).

Most TMMs include a large number of parameters that stand for various characteristics of the travellers, the vehicles, the transport system and so on. These parameters must be calibrated before the TMM is used as an estimation tool, but until recently, methods for calibration of TMMs were almost nonexistent. In the last few years there has been a wave of valuable research work that developed and discussed procedures for TMM calibration, and it now seems clear, at least to traffic theorists, that no TMM should be used without a preceding stage of proper calibration.

Since many of the newly-proposed TMM calibration methodologies were developed independently, with no well-established principles as a common background, it is hard to view them in a wide uniform context. As demonstrated later in this section, some procedures were devised for the needs of a particular study, whereas some others were created in an attempt to suggest generic calibration guidelines. In addition, some issues that one calibration methodology sees as vital are ignored by others. The review presented here concentrates on several issues that are repeatedly brought up when TMM calibration is discussed. We try to take into consideration the diversity of TMM applications, on one hand, and the particular needs of the current study, on the other hand.

5.3.2. Calibration conventions and underlying assumptions

As mentioned above, a TMM typically consists of several sub-models, each of which tries to reproduce the mechanism of a single decision made by an individual traveller, such as the decision to change lane or to accept a gap in the opposing traffic in order to enter an intersection. Each one of the sub-models that form a TMM normally includes

several parameters, and a complete TMM sometimes includes many dozens of parameters. A fundamental issue is the great difficulty to measure the behavioural parameters directly through observations or field surveys. Direct measurement is very complicated as many of the parameters stand for very subtle features that are hard to isolate from other irrelevant features. Even when a parameter is measurable or observable, this normally requires extensive collection of disaggregate data, which is in many cases impractical because of its cost. Works that directly study the value of a TMM parameter do exist, but we are not aware of any work where it was possible to do so for *all* parameters of a TMM. There is always a need, therefore, to estimate the values of most TMM parameters in a compromised, indirect way.

Essentially, calibration of a TMM is the process of adjusting the values of the model parameters such that the model is able to produce outputs that are reasonably similar to observed data. Due to the abovementioned difficulties, all the studies reviewed here do this with aggregate data, i.e. without using information that explicitly relates to the specific behavioural feature that each parameter represents. Some works that concentrate on calibration with disaggregate data do exist (e.g. Hoogendoorn and Ossen, 2005), but they only deal with the calibration of specific elements of a TMM (such as a car following sub-model) and with a limited number of parameters; we do not include such methodologies in this review as they do not directly aim at calibrating a TMM as a whole. The aggregate data used for calibration typically include such measures as travel times, flows, speeds or queue lengths. In all the methodologies we discuss here, observed values of these measures are compared to the equivalent values in the TMM outputs, and the TMM parameters are modified till there is sufficient match between the field observations and the simulation.

The aggregate nature of TMM calibration has been discussed by Toledo et al (2003), Ben-Akiva et al (2004) and Toledo and Koutsopoulos (2004). These studies describe the general framework of TMM calibration as a process of two stages: first, the sub-models are estimated with disaggregate data, and then the whole model system is calibrated with aggregate data (with possible re-iteration if necessary). Many of the calibration studies imply, even if not explicitly, that although TMMs have disaggregate foundations, aggregate calibration does make good sense because it is only meant to adjust a model that is already theoretically well-established to a specific situation. What has not been sufficiently stressed in any of the reviewed studies is that when a behavioural model is calibrated without using disaggregate data, there is a risk that the

result is plausible as a mathematical tool but not as a powerful behavioural model. This might happen if a set of flawed parameter values happens to have good fit to observed data and is therefore chosen as the best solution of the calibration problem. Since the values are flawed, they might result in unreasonable estimates when the TMM is used as a prediction tool in hypothetical scenarios. As previously mentioned, the lack of data and the inability to measure the parameters directly often oblige us to assume that the empirical calibration does not cause significant bias; but it is vital to always treat the results of such calibration with suspicion. No irrational parameter value should be accepted only because this was the calibration result, and attempts should be made to compare the parameter values rendered by the calibration procedure to estimates of these parameters from other sources.

TMMs, similar to other models, are not free from simplification. In addition to error or bias that might result from the lack of high-quality data, TMM outputs are also inevitably compromised because it is not possible to incorporate in the TMM all the factors that influence the performance of the real transport system. No TMM can always replicate the entire range of factors such as roadside activities or road incidents. When observed data are compared to the simulated outputs during the calibration process, we unavoidably make the assumption that only factors that the model includes exist in the actual network. Since this assumption is erroneous, the result is that the real-world phenomena that are not incorporated in the TMM affect the values of the calibration parameters, although ideally they should not. This is a source of error that we have no means to tackle; it should remind us that there is a need to constantly seek ways to improve the behavioural explanatory power of the TMM itself, independently of the calibration methodology.

TMM calibration is a multidimensional problem, as there are usually more than just one or two parameters to calibrate. Since there is often a limited amount of data, as well as limited amount of time for the calibration process, it is hardly ever possible to calibrate all the parameters. All calibration methodologies reviewed here concentrate on the calibration of a relatively small subset of parameters. Many methodologies stress the importance of a systematic calibration procedure, but no study was found where the subset of parameters to be calibrated is in itself chosen systematically. It is most demanding to investigate the sensitivity of the model outputs to the values of all parameters as a basis for the choice of the parameter subset, because the stochastic nature of the TMM will require a very high number of runs even for a basic sensitivity

test (runtime issues are further discussed later in the chapter). Still, analysts should remember that putting much effort in the development of a powerful calibration methodology can bring little gain if some parameters that strongly influence the traffic measure of interest have not been included in the calibration subset. Therefore, while we focus here on the methodology of calibration, it should be reminded that there is also need to study how the input for the calibration process can be specified more efficiently. The conventions and assumptions discussed in the previous paragraphs are common to all the calibration methodologies reviewed here. In the subsequent sub-sections we discuss issues where major differences exist between the different procedures. A systematic comparison between the reviewed methods and case studies is presented in table 5.3.

5.3.3. Scope of the calibration problem

All the studies summarised in table 5.3 deal with TMM calibration (or validation), but in fact there are considerable differences between these problems. A first major difference lies in the definition of the problem itself: while some studies concentrate on the calibration of driving behaviour parameters only (e.g. Jayakrishnan et al, 2001; Ma and Abdulhai, 2002; Hourdakis et al, 2003; Kim and Rilett, 2003, 2004), some others (e.g. Toledo et al, 2003; Ben-Akiva et al, 2004; Chu et al, 2004; Dowling et al, 2004; Oketch and Carrick, 2005) incorporate this in a broader problem, where a route choice model and/or a demand (origin-destination) matrix are calibrated too. The authors who propose the broader problems present evidence that procedures which simultaneously tackle multiple problems result in stronger models, and that solving the sub-problems separately might lead to biased estimates. However, it should be stressed that the various sub-problems that can constitute a broad calibration problem might differ from each other in their data requirements. For example, to calibrate driving behaviour parameters it is important that the data are collected in a range of traffic settings, while for estimating the demand matrix it is mainly essential that they are collected in a large number of locations throughout the network. Therefore, when there is limited amount of data, the attempt to solve a joint calibration problem might be very ambitious. In addition, the abovementioned risk, of weakened behavioural power of models whose calibration is based on aggregate data, is higher in a problem of a bigger scale. In other words, increased scope of the calibration problem might compromise the ability to

retain the original role of each individual parameter, which is important when the model is used for prediction in hypothetical scenarios.

Among the case studies that accompany the calibration methodologies there is substantial variation in the number of parameters being calibrated. As mentioned previously, no methodology suggests calibrating all parameters; the size of the calibrated subset varies from 4 parameters or even fewer (e.g. Merritt, 2004; and Shaaban and Radwan, 2005) to 19 (Kim and Rilett, 2003 and 2004). Focusing on a smaller number of parameters is less demanding in computational terms, and has also the advantage of enabling to pay more attention to each parameter when its value is modified; in some cases this is done through a manual procedure (see more on this issue later in this section). Bigger parameter subsets are normally calibrated using automated algorithms, which potentially make them more likely to get efficiently closer to an empirically-optimised solution, but also make it harder to follow changes in the value of each parameter very closely. A disadvantage of calibrating only few parameters is the risk that their resulting values are influenced by phenomena that are actually related to some of the non-calibrated parameters. This leads to erroneous estimates, but obviously there is no well-defined minimum for the number of parameters that should be calibrated for the model to be valid. Overall, it seems that when an analyst chooses a set of calibration parameters, the very ambitious task is to choose a number of parameters that is big enough to cover the various behavioural elements in the model, but small enough to enable paying individual attention to the value of each parameter, and also small enough to make the procedure computationally feasible.

Source	The problem	Formu-lation	Compared measures	No. of measure-ment locations	No. of calibration parameters	No. of model runs for a single evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Jayakri-shnan et al (2001)	Calibration	Verbal	Not specified	Not specified	Not specified	Not specified	None	PARAMICS	No
Ma and Abdulhai (2002)	Calibration	Genetic algorithm	Turning flow	20	5	1	Urban street network (470 nodes)	PARAMICS	Yes
Hourdakis et al (2003)	Calibration	Verbal	Flow and speed	21	12	Not specified	One freeway (20-km section)	AIMSUM	No

Table 5.3: Summary of the reviewed studies

Source	The problem	Formu-lation	Compared measures	No. of measure-ment locations	No. of calibration parameters	No. of model runs for a single evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Park and Schnee-berger (2003)	Calibration and validation	Verbal	For calibration: average travel time. For validation: queue length	12	7	5 (50 for a group of best solutions)	Urban arterial (with 12 intersections)	VISSIM	No
Toledo et al (2003)	Calibration including route choice and OD estimation, and validation	Mathema-tical program	For calibration: speed and flow. For validation: flow, average travel time, queue length	Not specified	Not specified	1	Mixed freeway and urban network (size not specified)	MITSIMlab	Yes (not full)

Table 5.3 (continued): Summary of the reviewed studies

Source	The problem	Formu-lation	Compared measures	No. of measure-ment locations	No. of calibration parameters	No. of model runs for a single evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Kim and Rilett (2003)	Calibration	Mathema-tical program, solved with the simplex algorithm	Flow	5	3 in TRANSIMS, 19 in CORSIM	1	One freeway (23-km sections)	TRANSIMS and CORSIM	Yes
Barcelo and Casas (2004)	Calibration and validation	Verbal	Not specified	One example with 700. Not specified in other examples	Not specified	Not specified	One example with interurban network (totally 1800 km of major roads). Other examples with 100-node and 260-node urban networks	AIMSUM	No

Table 5.3 (continued): Summary of the reviewed studies

Source	The problem	Formu-lation	Compared measures	No. of measure-ment locations	No. of calibration parameters	No. of model runs for a single evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Ben-Akiva et al (2004)	Calibration including route choice and OD estimation	Mathema-tical program, solved with Box's Complex algorithm	Small network: speed and density. Big network: flow	Small network: not specified. Big network: 68	Small network: 3. Big network: 6.	Not specified	Small network: freeway and two intersecting arterials. Big network: urban street network (298 nodes)	MITSIMlab	Yes
Chu et al (2004)	Calibration including route choice and OD estimation	Verbal (with quantitative objective functions)	Flow and average travel time	52	Not specified	More than 1 (exact number not specified)	Mixed freeway (5 to 10 km sections) and urban network	PARAMICS	No

Table 5.3 (continued): Summary of the reviewed studies

Source	The problem	Formu-lation	Compared measures	No. of measure-ment locations	No. of calibration parameters	No. of model runs for a single evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Dowling et al (2004)	Calibration including route choice	Verbal	For calibration: Capacity, flow, average travel time, queue length. For validation: average travel time, speed, density.	2	2 in the illustrated example, more in the theoretical discussion	More than 1 (exact number not specified)	Mixed freeway (with two interchanges) and urban (including arterial with 6 intersections) network	Not specified	No
Kim and Rilett (2004)	Calibration	Genetic algorithm	Flow	11	3 in TRANSIMS, 19 in CORSIM	1	Two freeways (23-km sections)	TRANSIMS and CORSIM	Yes

Table 5.3 (continued): Summary of the reviewed studies

Source	The problem	Formu-lation	Compared measures	No. of measure-ment locations	No. of calibration parameters	No. of model runs for a single evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Merritt (2004)	Calibration and validation	Verbal	Queue time, percentage stopped, delay time and queue length	2 inter-sections, 4 approa-ches in each	4	Not specified	Inter-urban arterial (5km section)	CORSIM	No
Toledo and Koutsopoulos (2004)	Validation only	None	Speed	4	None	10	One freeway (4-km section)	Not specified	No
Kim et al (2005)	Calibration	Genetic algorithm	Distribution of travel time	3	6	1	Urban arterial (1-km section)	VISSIM	Yes

Table 5.3 (continued): Summary of the reviewed studies

Source	The problem	Formu-lation	Compared measures	No. of measure-ment locations	No. of calibration parameters	No. of model runs for a single evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Shaaban and Radwan (2005)	Calibration and validation	Verbal	Queue length, total travelled distance	2 sites, 8 periods in each	4	20	Road segment located between two signalised intersections	SimTraffic	No
Park and Qi (2005)	Calibration	Genetic algorithm. Initial set of solutions is generated using a Latin Hypercube Design Algorithm	Average travel time	1	8	5	One signalised intersection	VISSIM	Yes

Table 5.3 (continued): Summary of the reviewed studies

Source	The problem	Formu-lation	Compared measures	No. of measure-ment locations	No. of calibration parameters	No. of model runs for a single evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Oketch and Carrick (2005)	Calibration including OD estimation, and validation	Verbal	link flow, turning flow, average time, queue length	55	Not specified	Not specified	Urban network (8 km^2, including 16 intersections)	PARAMICS	No

Table 5.3 (continued): Summary of the reviewed studies

There are also significant differences between the various calibration studies in terms of their geographical scale. Such differences exist both in the size of the simulation network and in the spread and density of data sources over this network. In terms of network size, the studies vary from a single intersection (e.g. at Ma and Abdulhai, 2002) to an extensive metropolitan area (e.g. at Park and Qi, 2005). The dispersion of sources of input data is sometimes as limited as two observation points in a medium-sized network (Dowling et al, 2004) or, in contrast, dozens of points in a network that is not much bigger (Chu et al, 2004; Oketch and Carrick, 2005). In principle, many calibration methodologies can be implemented in various networks, independently of the size of the network that was used to illustrate their foundations. However, most methodologies are at least partially adjusted for the scale in which they are later implemented: automated calibration is preferred if data is available from many measurement points (e.g. Ma and Abdulhai, 2002 or Ben-Akiva et al, 2004); comparison of multiple traffic measures is used in cases where there is much data but only from a small number of locations (e.g. Dowling et al, 2004; Merrit, 2004); and so on. We discuss these issues further later.

The scope of the calibration problem also has to do with the choice of traffic measures used to compare observed data to the simulation outputs. Some of the proposed procedures use a single measure; for instance, Ma and Abdulahi (2002) and Kim and Rilett (2003, 2004) compare only flows. Some others use more than one measure, normally by performing a sequence of calibration sub-processes, each one of which uses a different traffic measure to calibrate a separate group of parameters. In the procedure proposed by Dowling et al (2004) simulated and observed capacities are compared in the first stage to calibrate driving behaviour parameters, then flows are compared to calibrate route choice parameters, and finally all parameters are fine-tuned by comparing travel times and queue lengths. Hourdakis et al (2003) start with calibrating global parameters (such as maximum acceleration and other vehicle characteristics) by comparing flows; then they calibrate local parameters (such as speed limits) by comparing speeds; an optional third calibration stage is suggested, where any measure chosen by the user can be compared. A similar multi-stage concept is also proposed by Chu et al (2004).

The fact that the calibration methodologies cover a variety of problem sizes could be an advantage if it was clear in which circumstances we should use any of the methods. In practice this is difficult to determine, since the choice of geographical scale and traffic measures often depends primarily on which data is available. When the choice of

calibration methodology is dictated by data availability considerations, the credibility of the calibrated model must not be taken for granted. For instance, it is hard to judge whether a model calibrated using a single traffic measure, such as flows, is proper to use for estimating other measures, such as travel times; or if a model calibrated based on data from one intersection can be used in an entire urban area. Since limited data availability is often inevitable, there would be no use in forming very strict principles for the acceptability of a calibration method. Instead, it should be born in mind that the scope of the calibration procedure determines, to a great extent, the range of circumstances in which the calibrated model can later be reliably implemented.

5.3.4. Formulation and automation of the calibration process

Not all the discussions of TMM calibration include an explicit formulation of a calibration problem. Three groups of studies can be identified in this respect: studies where a full calibration procedure is presented; studies where a full procedure is used but not all of it is presented; and studies where calibration does not rigorously follow a systematic procedure. Many papers (Jayakrishnan et al, 2001; Hourdakis et al, 2003; Park and Schneeberger, 2003; Barcelo and Casas, 2004; Dowling et al, 2004; Merrit, 2004; Chu et al, 2004; Shaaban and Radwan, 2005; Oketch and Carrick, 2005) belong to the latter group. These works form an intermediate stage in the evolution of more cohesive concepts of calibration, as they do stress the need for some consistent judgement throughout the calibration process, unlike earlier studies, where parameters were normally adjusted using manual trial-and-error techniques (see also discussion at Toledo et al, 2003; Ben-Akiva et al, 2004; and Chu et al, 2004).

When a systematic calibration procedure exists (whether or not it is presented), it often has the form of an optimisation problem. It does not necessarily have the explicit form of a mathematical program, although in most cases at least the objective of the calibration is presented in a quantitative way. Systematic calibration procedures must use a *solution algorithm*, i.e. a methodical series of actions taken in order to find the best parameter set. The solution algorithm is normally an automated iterative process, in which the best solution is gradually improved till some stopping criterion is met. It is often described verbally or as a flow chart, even if the approach is methodical and quantitative. As mentioned earlier, the concept of a fully-automated calibration procedure is somewhat inconsistent with the idea that each TMM parameter stands for a particular element in a behavioural sub-model, because such parameters need individual

attention when their values are adjusted. However, the point made here is that this paradox is in the nature of TMM calibration, and that modellers should be aware of the drawbacks of whatever calibration concept they choose.

Although the calibration problem is commonly formulated as an optimisation problem, it is unlikely to lead to a global optimum. Whether or not such optimum exists is an ambitious question, which we leave for other studies. But even if it does exist, the multidimensionality of the solution search space, and the tendency of the observed input data to exhibit various inconsistencies, make many calibration problems more likely to result in a local optimum. In fact, the higher the number of calibration parameters, the more probable it is that all the calibration process achieves is an *improved* set of parameter values, not necessarily even a local optimum. This does not mean that calibration is unnecessary; but as explained earlier in this chapter, it is another reminder that the calibration outputs should be subject to constant logical judgement and comparison to other available sources.

We do not include here a detailed description of all optimisation approaches used in the reviewed calibration studies. However, some specific concepts of optimisation seem particularly relevant for TMM calibration, as they are repeatedly mentioned by different authors. Several studies conduct the search for the best parameter set using a genetic algorithm (Ma and Abdulhai, 2002; Kim and Rilett, 2004; Kim et al, 2005; Park and Qi, 2005). Genetic algorithms borrow ideas from the theory of natural evolution to tackle multidimensional optimisation problems. A population of candidate solutions (namely, in our context, candidate sets of parameter values) is generated and then goes through an iterative evolutionary process. In this process, the chance of any candidate to have offspring in the next generation of solutions depends on some measure of its fitness. This measure is normally called *the fitness function*; each genetic algorithm develops a different function. In the context of TMM calibration, this function should express how similar simulation outputs are to observed field measurements. Candidate solutions with a high fitness value go through mutation and crossover operations, attempting to retain their beneficial traits, while solutions with a low fitness value are likely to become gradually extinct. From a mathematical perspective, genetic algorithms are a powerful search technique; among their advantages is the fact only the fitness function, and not its derivatives, needs to be calculated throughout the process. However, the fact that at each generation there is a need to calculate the fitness value for many candidate

solutions, and thus to run the TMM many times, is a disadvantage as it might cause serious increase of the process runtime.

Another optimisation concept that is used by several authors includes the algorithm known as the Downhill Simplex Method and another technique named Box's Complex Algorithm. These techniques belong to the same family of methodologies, although the calibration studies who use them do not discuss this. The simplex method (used for instance by Kim and Rilett, 2003) presents a set of candidate solutions as a multidimensional geometrical shape, called *simplex*. Each dimension of this space represents the range of feasible values of one variable, i.e. one TMM parameter. If in the calibration problem there are N parameters to calibrate, the simplex is N-dimensional, and it has N+1 vertices, each one of whom stands for one feasible solution. The simplex goes through a series of manipulations, such as reflection and contraction, in an attempt to find vertices that perform best as solutions of the calibration problem. Box's algorithm (used by Ben-Akiva et al, 2004) is a more general version of the simplex method: the simplex is replaced with a *complex*, and at any stage of the process there are *at least* N+1 candidate solutions (as opposed to *exactly* N+1).

In the context of TMM calibration, in which the TMM needs to be run every time a candidate parameter set is evaluated, both the simplex and the complex algorithms are relatively efficient in terms of the number of evaluations per single iteration. This is because they are based on the idea that only few candidates are replaced at a time, as opposed to the substitution of multiple candidates in each iteration of a genetic algorithm. Similar to genetic algorithms, the simplex and complex algorithms do not require calculation of derivatives. However, they are not considered efficient in terms of the number of iterations required to reach convergence. The complex algorithm is more efficient than the simplex algorithm, but requires more evaluations per iteration. There is a clear trade-off between a low runtime per iteration and a small number of required iterations. Since we do not know in advance how many iterations will be required for a satisfactory solution, it is hard to tell which of the methods should be favoured. It generally seems that in applications where each single evaluation takes a considerable time (for instance because it requires multiple runs of the TMM, as we discuss later), any method that requires a small number of evaluations per iteration is preferable despite the tendency of such methods for only a slight improvement of the parameter set between successive iterations. For this reason, later in this thesis we use the simplex algorithm. But note that this is not meant to undermine the complex algorithm, which

might be found more efficient if its improved efficiency compensates for the higher duration of each iteration, and there seems at this point equally suitable.

Naturally most calibration methodologies search optimal parameter values in a continuous space, such that any value within some preset range is considered feasible. However, several methodologies only choose the optimal values from a given set: Merritt (2004) chooses from a set of 10 predetermined values, and Shaaban and Radwan (2005) choose from 3 values. This is done when no optimisation problem is formulated; the selection of the best parameter set is done by checking the difference between simulated and observed measurements for all possible combinations of the discrete values of all the parameters. Note that when the value of a calibration parameter is chosen from a preset discrete list, the level of detail of this list (e.g. 10 or 3 values, in the examples above) is not related to the dimensionality of the entire calibration problem. Whatever the number of preset values, they all constitute a single choice dimension.

A different type of manual search is used by Hourdakis et al (2003) and Oketch and Carrick (2005): the space of feasible solutions is indeed continuous, but the calibration procedure is performed at one location at a time, and the chosen parameter set is obtained after going one-by-one through all sites where input data is available. Although manual calibration does not seem generally efficient, it should be remembered that automated procedures often require a considerable programming effort. If a TMM is needed in a limited small-scale application, the option of manual calibration should not be discarded. As discussed earlier, the risk that an automatic procedure might not be sensitive enough to the behavioural foundations of each parameter is another justification for undertaking manual calibration.

5.3.5. Measuring goodness-of-fit

At the heart of any calibration technique is a comparison between simulation outputs and observed measurements of various traffic measures. The goodness of fit of the simulated measurements, based on any candidate parameter set, to the observed measurements, is the indicator for the fit of the parameter set itself. Various calibration methodologies use different ways to measure the discrepancies between observed and simulated values; the measures they use are summarised in table 5.4. The following notation is used in the table:

x, y	simulated or observed measurement, respectively
N	number of measurements
$\bar{x}$, $\bar{y}$	sample average
σ_x, σ_y	sample standard deviation

The different measures of fit have different sensitivities to various aspects of dissimilarity between the simulated and observed measurements; performing a calibration process with different measures is most likely to lead to different solutions. The following are examples of differences between the measures:

1. Most of the measures will mainly identify poor fit between the central tendencies of the compared samples, while only few measures (especially Theil's indicators) exhibit explicit sensitivity to the variance and covariance. Clearly, whether this has major importance varies between different applications.
2. Some of the measures (PE, ME, MNE) let errors with a similar size but a different sign balance each other. Such measures are useful for detecting systematic bias, but they are not powerful as indicators of the magnitude of an error.
3. Some measures (MAE, MANE) use the absolute value of the difference between the observed and simulated measurements; thus they give equal weights to all errors, whatever their size. In contrast, other measures (SE, RMSE, RMSNE) depend on the squared difference, and hence place a higher penalty on large errors. In the context of traffic modelling, penalising small errors is wrong, since the stochastic nature of traffic phenomena makes small errors inevitable. Using the squared difference between the simulated and the observed measurement is more appropriate, and it is actually surprising that none of the reviewed measures raises this difference to a power higher than 2. Alternatively, avoiding the unnecessary effect of small errors is also possible by examining the probability density function (as in the K-S test) rather than directly examining each individual observation.

Name	Measure	Used by	Comments
Percent error (*PE*)	$\frac{x_i - y_i}{y_i} \cdot 100$	Shaaban and Radwan (2005), Park and Qi (2005), Merritt (2004)	Applied either to a single pair of observed-simulated measurements or to aggregate networkwide measures
Squared error (*SE*)	$\sum_{i=1}^{N}(x_i - y_i)^2$	Ben-Akiva et al (2004), Chu et al (2004)	
Mean error (*ME*)	$\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)$	Toledo and Koutsopoulos (2004)	Indicates the existence of systematic bias. Useful when applied separately to measurements at each location
Mean normalized error (*MNE*)	$\frac{1}{N}\sum_{i=1}^{N}\frac{x_i - y_i}{y_i}$	Toledo et al (2003), Toledo and Koutsopoulos (2004), Chu et al (2004)	Indicates the existence of systematic bias. Useful when applied separately to measurements at each location
Mean absolute error (*MAE*)	$\frac{1}{N}\sum_{i=1}^{N}\lvert x_i - y_i\rvert$	Ma and Abdulhai (2002)	Not particularly sensitive to large errors

Table 5.4: Measures of goodness-of-fit

Name	**Measure**	**Used by**	**Comments**
Mean absolute normalized error (*MANE*)	$\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert x_i - y_i\rvert}{y_i}$	Ma and Abdulhai (2002), Kim and Rilett (2003), Merritt (2004), Kim et al (2005)	Not particularly sensitive to large errors
Exponential mean absolute normalized error (*EMANE*)	$A\,e^{-B\cdot MANE}$ (A, B are parameters)	Kim and Rilett (2004)	Used as a fitness function in a genetic algorithm
Root mean squared error (*RMSE*)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}$	Toledo and Koutsopoulos (2004), Dowling et al (2004)	Large errors are heavily penalised. Sometimes appears as mean squared error, without the root sign
Root mean squared normalized error (*RMSNE*)	$\sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}\left(\frac{x_i - y_i}{y_i}\right)^2}$	Hourdakis et al (2003), Toledo et al (2003), Toledo and Koutsopoulos (2004), Ma and Abdulhai (2002)	Large errors are heavily penalised
GEH statistic	$\sqrt{\frac{2(x_i - y_i)^2}{x_i + y_i}}$	Barcelo and Cases (2004), Chu et al (2004), Oketch and Carrick (2005)	Applied to a single pair of observed-simulated measurements. *GEH*<5 indicates a good fit

Table 5.4 (continued): Measures of goodness-of-fit

Name	Measure	Used by	Comments
Correlation coefficient (r)	$\frac{1}{N-1}\cdot\sum_{i=1}^{N}\frac{(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}$	Hourdakis et al (2003)	
Theil's bias proportion (Um)	$\frac{N\,(\bar{y}-\bar{x})^2}{\sum_{i=1}^{N}(y_i-x_i)^2}$	Hourdakis et al (2003), Barcelo and Cases (2004)	A high value implies the existence of systematic bias. Um =0 indicates a perfect fit, Um =1 indicates the worst fit
Theil's variance proportion (Us)	$\frac{N\,(\sigma_y-\sigma_x)^2}{\sum_{i=1}^{N}(y_i-x_i)^2}$	Hourdakis et al (2003), Barcelo and Cases (2004)	A high value implies that the distribution of simulated measurements is significantly different from that of the observed data. Us =0 indicates a perfect fit, Us =1 indicates the worst fit
Theil's covariance proportion (Uc)	$\frac{2(1-r)\cdot N\cdot\sigma_x\sigma_y}{\sum_{i=1}^{N}(y_i-x_i)^2}$	Hourdakis et al (2003), Barcelo and Cases (2004)	A low value implies the existence of unsystematic error. Uc =1 indicates a perfect fit, Uc =0 indicates the worst fit. r is the correlation coefficient

Table 5.4 (continued): Measures of goodness-of-fit

Name	Measure	Used by	Comments
Theil's inequality coefficient (*U*)	$\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^{\,2}} + \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^{\,2}}}$	Ma and Abdulhai (2002), Hourdakis et al (2003), Toledo and Koutsopoulos (2004), Barcelo and Cases (2004)	Combines effects of all 3 Theil's error proportions (*Um, Us, Uc*). *U*=0 indicates a perfect fit, *U*=1 indicates the worst fit
Kolmogorov-Smirnov (K-S) test	$\max\left(\lvert F_x - F_y \rvert\right)$	Kim et al (2005)	F is the cumulative probability density function
Moses' test and Wilcoxon test	The detailed procedure is described by Kim et al (2005)		

Table 5.4 (continued): Measures of goodness-of-fit

Most measures involve summation of errors over a series of pairs of simulated and observed values. It is not always obvious how to create these pairs. When each measurement is taken at a particular time and place, no such difficulty arises; but if, for instance, each measurement stands for the travel time of a specific vehicle, there are many different vehicles in the simulation outputs that can be paired with any observed vehicle, and each pattern of pairing might lead to a different level of fit. Unfortunately, none of the reviewed methods elaborates on this issue. If a test such as K-S is used, the measurement error that can be potentially caused by the pairing method is avoided, since individual observations are not examined explicitly.

The reviewed methodologies tend to consider the space of simulation outputs as one-dimensional, as only one index (denoted *i*) is used for the series of measurements in all the measures of fit in table 5.4. But in fact the outputs form a multi-dimensional space; in different studies, the index *i* is used in different dimensions. The most common dimension is time (namely, each measurement is taken at a different time interval), as

used by Toledo et al (2003), Chu et al (2004), Hourdakis et al (2003), Kim et al (2005) and others. But sometimes the series of measurements consists of values from different locations in the study network, and in other cases, each element in the series corresponds to a different vehicle. Many of the mentioned measures of fit can be used in any of these dimensions, but obviously in each dimension they have a different meaning. This is most apparent when the measure of fit is sensitive not only to estimates of the mean but also to the variations between the measurements. For instance, calibrating a TMM by focusing on estimates of variation of the travel speed over different time periods will probably lead to different results from calibration that focuses on speed variation between vehicles. It is therefore important to choose not only a measure of fit that is suitable for the particular needs of every application, but also to use it in the appropriate dimension.

The methodology described by Park and Schneeberger (2003) does not use any of the measures in table 5.4 but proposes an alternative concept, which estimates the model parameters without explicit calculation of the goodness of fit. This is done by creating a regression model where the calibration parameters are used as the explanatory variables and a traffic measure is the dependent variable. Calibration of the TMM is performed through seeking the parameter values with which the regression-based value of the traffic measure is the closest possible to the observed value; the fit of a candidate parameter set is therefore evaluated indirectly, via the fit of the regression model.

The procedure presented by Kim et al (2005) is the only one where the evaluation of fit uses the family of statistical techniques known as *two-sample tests*. These tests are more commonly used for validation of the calibration results. It should be stressed that in principle, two-sample tests are as suitable as the other measures mentioned above for measuring the fit between the simulation outputs and the field data. We return to this issue later in this section.

5.3.6. Repeated runs

Due to their stochastic nature, TMMs generate different outputs in every single run (unless the user chooses not to allow this randomness, by repeatedly using the same random seed numbers). As mentioned by Vovsha et al (2002), the heterogeneity of TMM outputs creates an opportunity for realistic representation of the range of likely outcomes in the real transport system. However, most of the reviewed calibration approaches consider this heterogeneity as a burden: Park and Schneeberger (2003), Park

and Qi (2005) and others state that the statistical analysis of the TMM outputs is aimed at reducing variability in the provided estimates, rather than making use of this variability.

Since TMM outputs vary from one run to another, it is necessary to look at the results of more than one run in order to ensure they are credible; the need to run the model several times requires significant time and some additional statistical analysis. The different calibration methodologies are not equally rigorous in this respect: some of them use a single run per one evaluation of the fit of a specific candidate solution, while others run the model up to 20 times for one evaluation. In most works where the model is run multiple times, the subsequent analysis is based on the average values across the series of runs, but Chu et al (2004) use the median rather than the average. The methodologies where the TMM is only run once for each evaluation do not provide reasoning for this. Our understanding is that even if due to the computational burden there is a need for some compromise, a single run is insufficient.

Many of the methodologies use the following formula to determine the required number of runs (Merritt, 2004; Toledo and Koutsopoulos, 2004; Chu et al, 2004; Shaaban and Radwan, 2005):

(5.1)

$$R = \left(\frac{s \cdot t_{\alpha/2}}{\bar{x} \cdot \varepsilon} \right)^2$$

Where

R	required number of model runs
s	standard deviation of the examined traffic measure
$\bar{x}$	mean of the traffic measure
ϵ	the required accuracy, specified as a fraction of $\bar{x}$
$t_{\alpha/2}$	critical value of Student's t-test at confidence level α

When R is calculated with this formula, an estimate of s is necessary as an input; but s is unknown prior to running the model. The papers that use the formula mention three slightly different methods to tackle this difficulty. According to the first approach, s and R are recalculated after every run, and the model is run till the resulting value of R is higher than the number of runs that have already been performed. The second approach

suggests that an estimate of s is calculated first based on a predetermined number of runs, and then it is used to determine how many more runs are still required. The third approach is similar to the first one, but instead of a single run between two successive recalculations of s and R, it uses a randomly-generated number of intermediate runs. All three methods require that if more than one traffic measure is used when comparing simulated and observed measurements, R should be computed for each measure separately, and the highest of all resulting values should be used.

Note that the abovementioned formula only determines the number of runs that is required to achieve a certain level of confidence about the *mean* value of the estimated traffic measure. If there is interest in any other statistic but the mean (such as the variance), it is wrong to use this formula. However, we are not aware of studies that seek the required number of runs for estimating other statistics but the mean. We pay special attention to this issue in chapter 6.

5.3.7. Validation of the chosen parameter set

Whatever parameter values are chosen in the calibration procedure, it is still necessary to confirm, using an independent set of data, that the model with the selected parameter values has a reasonable predictive power; this is conventionally referred to as the *validation* stage. The idea that validation must follow the calibration process is agreed by all, but a variety of techniques are used to implement it:

1. Visual validation (mentioned by Park and Qi, 2005; Oketch and Carrick, 2005; Toledo and Koutsopoulos, 2004; and many others). This is done by eyeballing the graphical presentation of the modelled network as the model runs, trying to spot any unusual behaviour. Most authors agree that visualisation is an efficient way to detect significant errors but cannot replace a more quantitative validation.
2. Validation using measures of fit, like those presented in table 5.4. Toledo and Koutsopoulos (2004) remind that these measures are sometimes used for validation, but in practice we found very few works that do this.
3. Statistical validation by arranging the simulated and observed measurements as two time series and then comparing the series (Barcelo and Cases, 2004). A slightly different version of this approach arranges each of the two compared datasets as a group of series rather than one series, and then compares the groups through bandwidth analysis (Barcelo and Cases, 2004).

4. Statistical validation using two-sample tests (Toledo and Koutsopoulos, 2004; Barcelo and Cases, 2004; Park and Qi, 2005; Park and Schneeberger, 2003). These are tools that examine the level of confidence about the hypothesis that two given samples (i.e. the simulated data and the observed data) have the same statistical properties; they are by far the most common techniques for TMM validation. The most popular is the two-sample t-test, but there are many other available two-sample tests, some of which have not been described in detail in the transport literature (see Scheffe, 1970; Maisel and Gnugnoli, 1972; and Kleijnen, 1995). Some tests are parametric, i.e. designed for cases where we have some preliminary information about the distribution of the measurements in the compared datasets. In contrast, nonparameteric tests do not require such information, but they are less powerful, i.e. require more data for a certain level of confidence. We normally do not know in advance what distribution describes the TMM outputs best, but as Kleijnen (1995) and others point out, it is not uncommon to make some distributional assumption in order to be able to use a parametric test. This is particularly true for the t-test, which can formally be used only for normally-distributed samples, but in practice is not very sensitive to violation of this requirement.
5. Indirect statistical validation. Instead of examining how similar TMM outputs are to field measurements, it is possible to test whether some product of the simulation outputs resembles the respective product of the field data. Toledo and Koutsopoulos (2004) build meta-models that capture relations between various traffic measures, such as the speed-flow relationship or the time evolution of flows; meta-models are estimated independently based on the simulated and the observed measurements, and it is then tested statistically whether the two models might be identical. Earlier versions of this approach were proposed by Kleijnen (1995) and Rao et al (1998).

The review of measures of fit, earlier in this section, shows that the different measures used in the calibration process do not use any uniform scale or consistent criterion to indicate good fit. In contrast, in the validation stage most authors prefer to use statistical tests that state well-defined levels of confidence. We find that this is unnecessary, because in traffic modelling the uncertainty about the *input* data (such as travel demand) is very high and it is therefore impossible to estimate the exact level of accuracy of the *outputs*. Validation is neither more nor less rigorous than calibration, and the requirements from the measure of fit used for calibration and the test used for validation

are in fact the same. Every test used within the calibration process can be also used for validation and vice versa; the various tests obviously have different advantages and disadvantages, but these apply similarly to validation and calibration.

Nevertheless, it is important to ensure that the validation test does not simply repeat what has already been tested in the calibration process. Toledo and Koutsopoulos (2004) mention that validation against the same measure that was used for calibration may lead to overestimating the realism of the model. This is an essential point, but note that it is very ambitious to always require validation using a different traffic measure from the one that has been used for calibration. The basic requirement, which every calibrated TMM must meet, is that it can be successfully validated with a new set of data from the same type. For example, a model that has been calibrated with queue length data from one set of intersections must pass the validation test using queue length data from another set of intersections. A higher standard of validation is reached if it can be confirmed that the model calibrated with queue length data can also give good estimates of speeds, times or flows. But in practice the relations between the different dimensions in the TMM itself are not always reliable enough to achieve such standard. It should therefore be mainly stressed that if validation is undertaken in the same dimension that has been examined during calibration, the TMM can be later used reliably as an estimation tool only in this dimension.

5.3.8. Summary

This section included a review of methodologies for calibration of TMMs and a discussion of similarities and differences between the methodologies. The need for systematic calibration has only gained attention quite recently, and the various newly-proposed calibration concepts have not been so far examined in a uniform context. The reviewed methodologies differ from each other both in principle issues, such as their objective and their scope, and in technical issues, such as their formulation, solution approach and statistical properties. Practitioners who calibrate traffic models often face insufficient amount of data, and the ways to deal with this have a key role in many of the discussed works.

Following the review it can be deduced that whatever the needs and the circumstances of the calibration of a TMM, the issues listed below must always be addressed:

- The set of calibration parameters should be specified with great care. There is little gain from the entire calibration process if the parameters that influence the traffic measure of interest are not calibrated, but in contrast, calibrating too many parameters is computationally inefficient. A parameter should be calibrated only if it cannot be directly measured or observed.
- The scope of the calibration problem is defined by the geographical area it encompasses and the amount of information sought (e.g. whether it incorporates additional features such as origin-destination travel demand estimation). There should be some fit between the scope of the problem, on one hand, and the amount and dispersion of the input data, on the other hand.
- The modeller should be aware of the disadvantages of the method, either automated or manual, used for searching for feasible solutions or for solving the specified optimisation problem. Possible disadvantages are high runtime and inefficient coverage of the solution space.
- If a small number of model runs is used to evaluate the fit of every particular candidate solution, the modeller must appreciate that the statistical credibility is compromised.
- The traffic measures used for calibration should be chosen while accounting for the expected future use of the model. Using more than one measure is advised if the TMM is to be used for multiple purposes. The multidimensionality of the TMM outputs and the input data should also be taken into consideration, as there are variations between vehicles, between points in space and between points in time.
- Given that the calibration process inevitably involves some compromises, it is essential to always treat its results with suspicion and make any effort to crosscheck them with other sources. It is important to remember that appropriate calibration does not solve the weaknesses of the TMM itself.

5.4. Conclusions

Previous attempts to develop tools for the prediction of TTV do not meet the needs that would come up if TTV was to be included in the appraisal of transport schemes. The required tools need to be sensitive to local factors and to detailed network configuration, so that it would be possible to use them to look at the effect of various changes of infrastructure on the level of TTV. We intend to examine whether there is scope for using microscopic traffic modelling for this purpose.

The introduction of TMMs into the discussion brings in the question of how these models should be calibrated if we want to use them for estimating TTV. We reviewed different methods of TMM calibration, and found that the opportunity of analysing the distribution of TMM outputs rather than the mean has hardly been discussed. It appears that particular elements of the TMM calibration process must be purposely adjusted for the task of estimating TTV. For example, special attention has to be paid to the choice of suitable measures of fit and to verifying a sufficient number of model runs. These issues are discussed further in the next chapter.

Chapter 6

Inter-run variation analysis of traffic microsimulation

6.1. Introduction

In chapter 5 we saw that there are currently no available tools that can assist in estimating the level of TTV under various assumptions of network configuration. In the current chapter we therefore propose a new approach for the estimation of TTV. No tool is developed here from scratch; rather, the proposed approach is based on using existing tools of microscopic traffic modelling. The innovative features discussed here are not in the tool itself, but in the concept of how to use it for the purpose of TTV prediction, and in the process it needs to go through before it can be used for this purpose.

Among the main elementary causes for variation in travel times in any transport system are *heterogeneity* and *randomness*. These are apparent in driving behaviour, in the level of demand, in the intensity of roadside activity, in weather conditions and so on. Heterogeneity and randomness are fundamental concepts in microscopic traffic modelling: random values of driver characteristics, vehicle features and so on are drawn in every run of a TMM from preset distributions or samples; running a TMM several times with the same inputs (but with different random seed numbers) will give different outputs. At first glance it might therefore seem only natural to estimate TTV by running a TMM multiple times and then analysing the variation of travel times between the runs. In fact, among practitioners it is not uncommon to run a TMM several times in order to get several different sets of outputs, each of whom is expected to represent feasible traffic conditions on a different day. However, such estimation brings in many issues that need careful examination. In this chapter we discuss whether day-to-day TTV, as defined in chapter 1, can be plausibly estimated through inter-run analysis of a TMM, i.e. the analysis of TTV between different runs. The calibration of the parameters of the TMM has a key role in the discussed approach, but this is not a discussion of a standard calibration problem, as most existing TMM calibration procedures focus on the mean values of the outputs and not on their variability.

This chapter starts with an introduction of the basics of the proposed concept, including the scope and the limitations of the analysis performed here. Then it discusses various technical, statistical and computational issues that must precede the full methodological development. Two experiments and their results are presented: one that has to do with required number of TMM runs, and one that looks at the ability to capture a particular level of TTV through a particular set of parameters. This leads to a full description of a calibration algorithm. The methodology suggested in this chapter is tested and applied in chapter 7.

6.2. Modelling variability using traffic microsimulation

The key objective of this part of the thesis is to examine whether inter-run TTV of a TMM can replicate the level of inter-day TTV in a real transport network. To show that there is a case for such analogy, we need to establish an analogy between a single TMM run and a single day in the real network. This idea brings in several fundamental issues that must be discussed before moving on to more technical matters. Four such issues are raised in the following paragraphs, which are also used to set the scope and the limits of the methodology and experiments presented later.

1. Can TTV be explained at all?

   Independently of the problem of *modelling* TTV, we could ask whether a consistent pattern of TTV exists at all. One might suggest that fluctuations in journey time are chaotic or that they are intractable in nature. If this is the case, then even an seemingly-perfect model of TTV will not be credible, because whatever the explanatory rules it finds in field measurements of travel time, these rules will not hold in other scenarios, and they therefore cannot be used for prediction. The fact that so far no model of TTV has been widely accepted can support this approach. However, the question of whether or not TTV can be modelled at all is too ambitious to be dealt with in the scope of our current work. The analysis in this thesis is based on the assumption that TTV is possible, even if hard, to model to a tolerable level of accuracy. This is taken as an axiom and no evidence for it is given; readers are welcome to make their own judgement based on the presented findings.

2. Can aggregate calibration be valid in a behavioural model?

   There is no reason to assume that inter-run TTV can represent inter-day TTV without appropriate calibration. Such calibration procedure should set the values of the TMM parameters at a level that establishes the analogy between a single run and a single day. This brings in the question of what the calibration procedure should be like. A major strength of TMMs is that every parameter in the model has a distinctive meaning and is not merely an element in a mathematical expression; but most existing methodologies for TMM calibration modify an entire set of parameters simultaneously, ignoring what each individual parameter stands for. Calibrating each parameter separately is extremely demanding in terms of the amount of data and effort required; calibration of a whole parameter set at once seems the only practical option. But as discussed in chapter 5, it should be realised that this is a compromise, since there is a risk that in order to improve the fit of the entire model, individual parameters are assigned wrong values. The calibration experiments described in this thesis do not offer an absolute solution for this risk. An attempt to reduce it is made by limiting the space of feasible solutions such that it only includes values that seem rational in the first place; values that deviate from the predetermined range of feasible values of each parameter are not accepted even if they seem better from the empirical point of view.

3. Can all the causes for TTV be modelled?

   Running a TMM either once or multiple times gives a range of travel time measurements, even if the location and time period of interest are unchanged. There are several sources for diversity in the model outputs:

   a) Randomness in input values, generated by the TMM itself. Some parameters are specified not as fixed values but as a distribution, and the TMM draws different values from this distribution every time it is required.

   b) Randomness in input values, generated by the user. When the TMM does not allow a parameter (or any other element of the TMM) to vary randomly, the user can alter the value of this parameter manually before each run.

c) Heterogeneity within the inputs for a single run. The user sometimes specifies a detailed list of travellers or a list of vehicles as part of the input data for the TMM run. It is often possible to attach different characteristics to each member of the list. The model outputs then enable separate analysis of the individual travel experience of each traveller or vehicle.
d) Special events, such as incidents or accidents, created by the model with a preset probability. Since the occurrence of these events is random, they add to the variability in the outputs.

The first and second sources on the list are essential for the current needs, and are also relatively easy to implement, as illustrated later in the chapter. The third source is important when studying inter-vehicle variability but is not a key issue here. The forth source requires features that are currently not possible in most TMMs (this is discussed further below).

It is important to understand to what extent the sources of variability that are actually generated by the TMM cover the causes of TTV in the real transport system. By efficiently using the first and second sources on the above list it seems possible to encompass a major share of these causes. But the range of factors and phenomena that influence the level of real-world TTV is very wide, and is not directly accounted for in full by any TMM. Daily fluctuations in roadside activity, weather changes, the chance of a missing bus driver and many other factors add additional uncertainty to daily journey times but most of them are currently considered too complicated to be incorporated in a TMM. In addition, most TMM applications do not allow the travellers complete freedom to change their entire set of travel choices (e.g. mode of transport, driving route or departure time) on a daily basis, even if variations in these choices must have some effect on TTV. If some sources of TTV are not modelled, then in the calibration procedure their effect will influence the values of the wrong parameters, and this will result in a compromised model.

The effect of an un-modelled source of TTV is not always a major problem. For instance, there is some sense in letting the calibration procedure find empirically which parameters can bear the influence of variation in weather conditions and roadside activity, as these phenomena truly affect TTV indirectly, via their effect on driving behaviour and vehicle performance. Another example is the direct effect of

accidents and other incidents such as road closure, i.e. the increase in travel times (and therefore in their variability) at the time and place where such events occur. This increase is not part of the scope of TTV discussed here (as defined in chapter 1), since we assume (as other authors do, such as Bates et al, 2001) that travellers are aware that these are independent events that do not necessarily signify poorer reliability. Therefore, if we omit from the observed measurements that we use for calibration those extreme travel times from days when accidents and incidents occurred, then this source of potential bias is avoided.

There are, however, un-modelled causes for TTV that constitute a bigger problem in the current context, such as the *secondary* effects of accidents and incidents, namely the indirect consequences that such events have at other locations or other times than where and when they occurred. These effects are probably quite small in magnitude, but they bring in serious complication because we are neither able to exclude them from the data we use for calibration, nor willing to exclude them from the definition of TTV, as travellers are not aware that they are related to any particular event. In this respect the experiments brought here are compromised, since we use a TMM that does not incorporate these secondary effects. We assume that their influence on TTV is not considerable, and suggest that this issue should be probed separately in the future. Note that to account for secondary effects in the calibration of the model, what needs to be improved is the ability of the TMM to take these effects into account, not the calibration procedure itself. Therefore, to cover the side effects of accidents and incidents in future experiments, the same methodology as presented here can be re-used with a TMM that does model accidents and incidents.

4. <u>Which parameters influence TTV?</u>

In principle all parameters might influence the level of TTV, but the number of parameters in most TMMs is so high that it would be computationally impractical to calibrate them all. Any additional parameter that we wish to calibrate adds a new dimension to the effort of searching a solution for the calibration problem. As the review in chapter 5 illustrates, all existing calibration methodologies concentrate on a subset of parameters. It is natural to exclude from the set of calibration parameters those whose values can be directly observed or measured using available data. But

in most practical circumstances, the remaining set will still be too large, and hence the decision of which parameters should be calibrated is a dilemma.

It would be important indeed to study the particular effect of each parameter on TTV, but this is not possible in the current scope. The calibration set used here is assumed to include the main parameters that significantly influence the level of TTV, but no solid evidence for this is presented. It is likely that the specification of this set can be refined and improved. It is therefore also assumed that with a better specification of the set of influential parameters, the results of the experiments described later could have been stronger.

In summary, this section has defined the scope of an attempt to determine the values of a set of TMM parameters such that the TTV between different runs can replicate inter-day TTV. The starting point for this experiment is the assumption that TTV is a phenomenon that can be plausibly explained by TMM parameters and that calibration of this behavioural model can be performed through a procedure that focuses on a set of parameters concurrently. We have also presented drawbacks of our own assumptions; it is hoped that the results of the experiment will help us determine to what extent were these assumptions logical.

6.3. Definition of objective

Methodologies for calibrating TMMs were proposed in many recent publications, but mainly discussed the estimation of mean values rather than measures of variability. Some of the existing methodologies do consider the distribution of TMM outputs, when testing the fit between simulated and observed measurements (e.g. Kim et al, 2005). But those distributions stand for spatial variability (between different locations in the network) or for temporal variability (between different periods), not for TTV as defined here.

Similar to some existing TMM calibration methodologies, the calibration experiment proposed here is performed through an optimisation process. Examining the objective functions used in other optimisation procedures who aim at calibrating TMMs confirms that most of them do not seek to reproduce a credible distribution of model outputs but mainly to reproduce the mean. A few methodologies use Theil's variance proportion, which does consider differences in the standard deviation of the two compared data series. As mentioned above, those methods look at variability in other dimensions than the one that the current analysis focuses on. In addition, Theil's measure is also unsuitable for our needs because it is insufficiently sensitive to various potential differences between the observed and simulated travel time distributions. If, for instance, the observed and simulated travel times have different distributions with the same standard deviation, Theil's measure will indicate perfect fit. For our optimisation process we therefore specify a new objective.

We wish to minimise the overall difference between the simulated inter-run TTV and the observed inter-day TTV. The required input for the calculation of our objective function includes:

1. Observed travel time measurements from L different locations, P time sub-periods and N different days. Each measurement is taken at a specific location, sub-period and day. It is not required to have data from all sub-periods at each location, but it is strictly required that for each available combination of location and sub-period there are measurements from N days (otherwise, it is not possible to calculate the level of TTV).
2. Simulated travel time measurements from the same L locations and P sub-periods as the observed measurements, each from N different runs. It is strictly

required that the locations and sub-periods are defined equally to the way they are defined in the observed data, and also that for each combination of location and sub-period there should be measurements from N simulation runs.

A later section in this chapter elaborates on how to determine N. The number of locations L and sub-periods P is too difficult to determine analytically; L and P mainly depend on the amount of available data, and the obvious general rule is that the more data is used, the better the results. Since this methodology is developed to be used (in chapter 8) for estimating bus TTV, it is assumed that the input data comes from bus travel time records; hence, each location is defined as a route segment, i.e. the part of a bus route between two consecutive stops.

In this study we only carry out simulation during the morning peak period, and the reason why this period is divided into several sub-periods is that our definition of TTV obliges it. TTV is defined here as random in nature, namely as variation due to factors that a rational traveller cannot predict. If measurements from an early time point during the morning peak period (say 07:10) were treated as part of the same travel time distribution as measurements from a late time point in this period (say 08:50), it would contradict the definition of TTV, since many travellers are clearly aware of reasons for differences in the level of TTV between these two time points. We therefore look at shorter periods and ascribe a different level of TTV to each one of them. The duration of the shortest period that travellers identify with a single set of traffic conditions is unknown; it is suggested that sub-periods of 30 minutes or less are plausible.

For each one of the route segments and sub-periods where data is available, we determine the fit between simulated and observed measurements based on the Kolmogorov-Smirnov test (K-S), which is a very common nonparametric two-sample test. Note that the same test has been used in chapter 4 for different purposes. The range of the K-S test statistic is between 0 and 1, where 0 indicates that the two samples are identical, and 1 indicates the worst fit between the samples. If the simulated and the observed distributions of travel times on a particular route segment at a particular sub-period are presented as cumulative probability density curves, the K-S test statistic equals the maximal difference (i.e. the maximal vertical distance) between the curves; this is illustrated in figure 6.1.

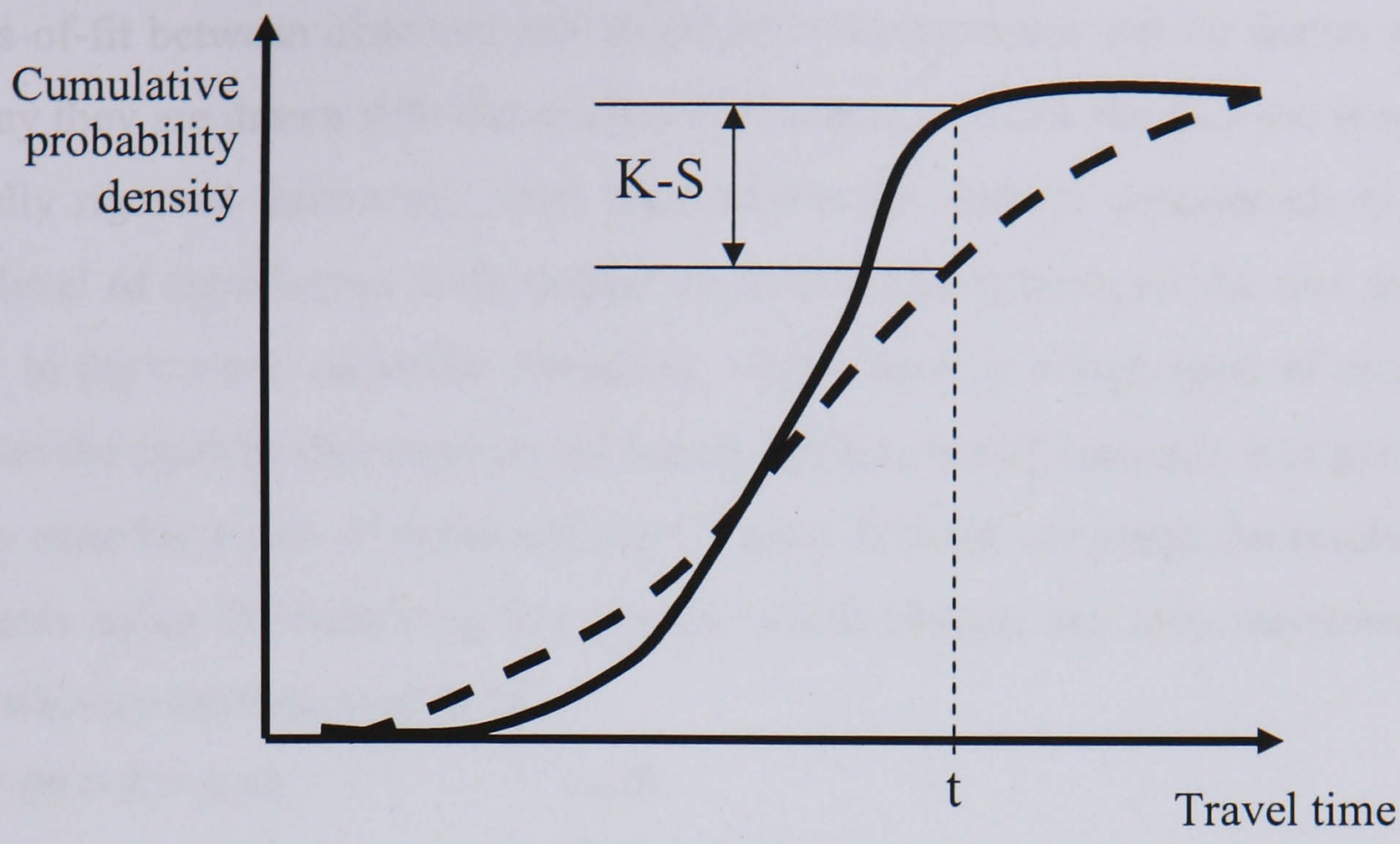


Figure 6.1: Kolmogorov-Smirnov test

In the calibration procedure described here, simulated and observed cumulative probability density curves of this nature are constructed for every combination of route segment and sub-period, and the K-S test statistic is then computed from each such pair of curves. All the curves are based on the same series of N runs of the TMM and N respective days of field measurement. The value of the objective function is an average of the K-S statistic values across all available locations and sub-periods. The objective function is formally defined as follows:

(6.1)

$$\boldsymbol{min\ Z} = \frac{1}{N_{LP}} \cdot \sum_{l} \sum_{p} \left(\max_{t} \left(\left| F_{t,obs}^{l,p} - F_{t,sim}^{l,p} \right| \right) \right)$$

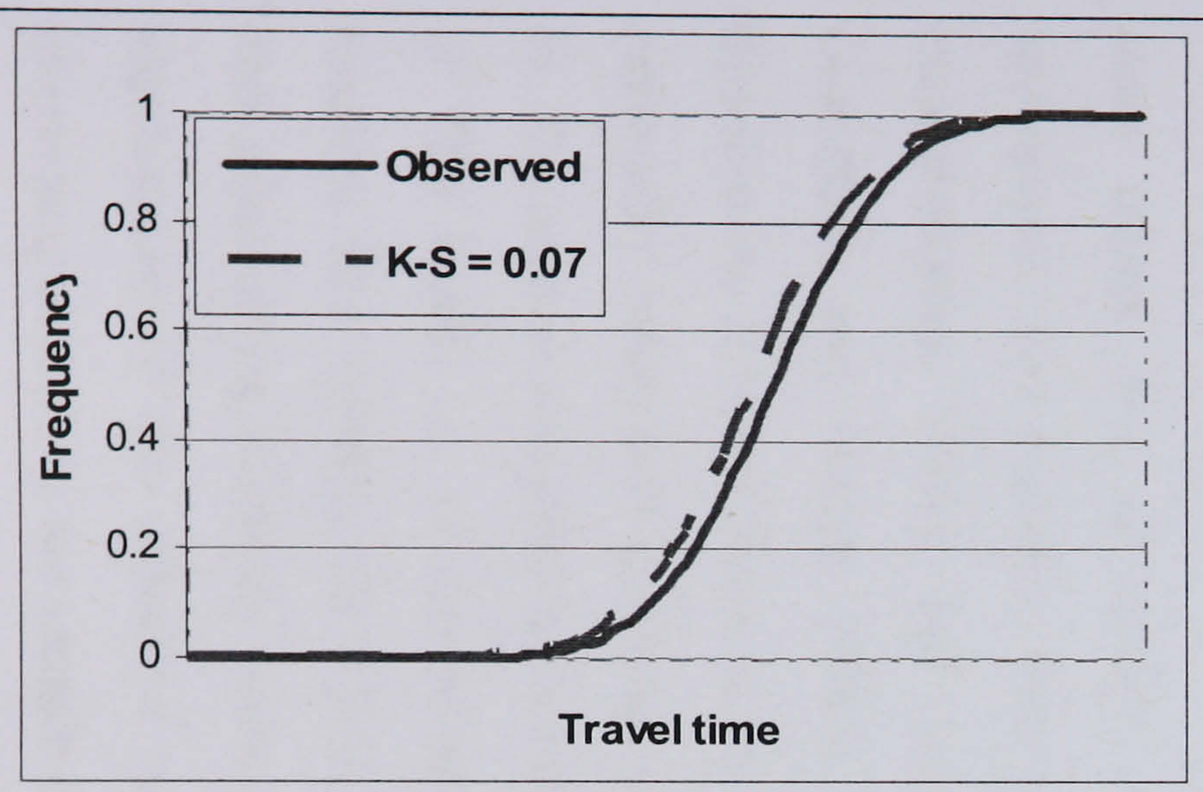
where

Z	value of the objective function
N_{LP}	the number of combinations of location and sub-period with available data
$F_{t,obs}^{l,p}$	cumulative probability density of observed travel time t at location l in sub-period p (based on N days of field measurement)
$F_{t,sim}^{l,p}$	cumulative probability density of simulated travel time t at location l in sub-period p (based on N model runs).

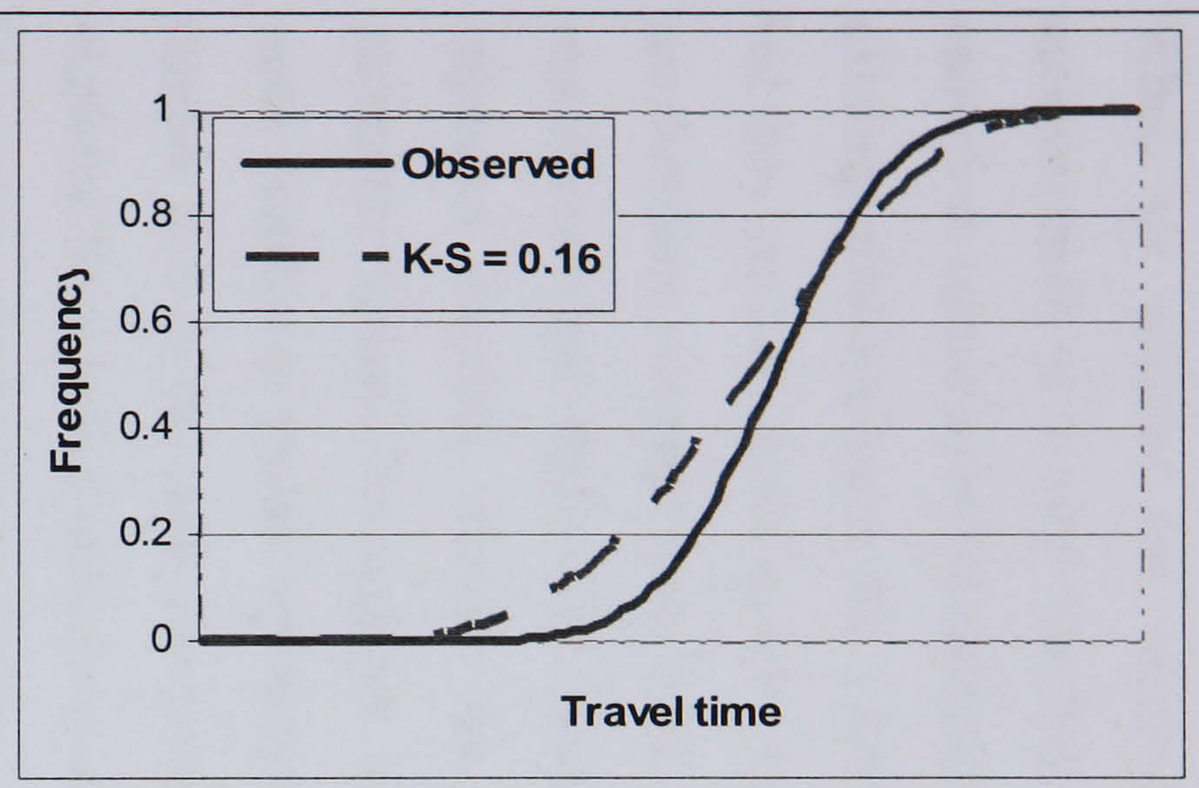
Since the objective value is an average K-S statistic, conclusions regarding the goodness-of-fit between observed and simulated measurements can be drawn similarly to the way they are drawn with the original K-S statistic. When the K-S test is used in a statistically rigorous framework, each level of the test statistic corresponds to a well-defined level of significance with respect to the difference between the two compared samples. In the context of traffic modelling, where there is a high level of uncertainty about even the most fundamental model inputs (such as travel demand), it is pointless to explicitly state the levels of statistical significance. Instead, we judge the results of our experiments using the following key legend, which reflects our own interpretation of how satisfactory the indicated fit is:

1.00 > Z > 0.48	no fit
0.48 > Z > 0.36	very poor fit
0.36 > Z > 0.24	poor fit
0.24 > Z > 0.12	plausible fit
0.12 > Z > 0.00	excellent fit

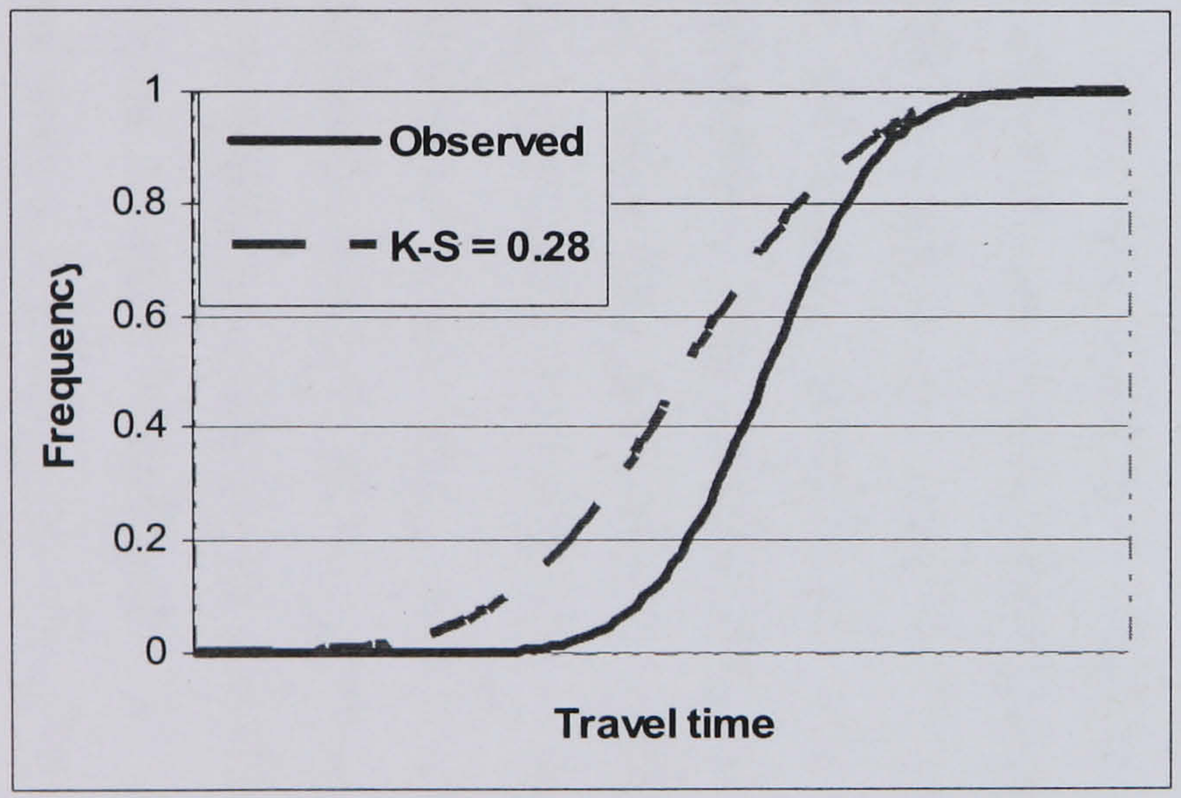
Figure 6.2 illustrates the type of match indicated by each of these levels. The solid curve in all the diagrams represents the cumulative probability density of the observed travel time, and the other curves represent the respective density based on several different sets of simulation outputs. In the upper right diagram, for instance, the K-S statistic is 0.16, which is classified here as "plausible fit". The value of the statistic means that the maximum error found in the cumulative curve of the simulated results, compared to the observed data, is 16%, i.e. for most travel times the cumulative difference between the simulated and observed measurements is smaller than 16%.



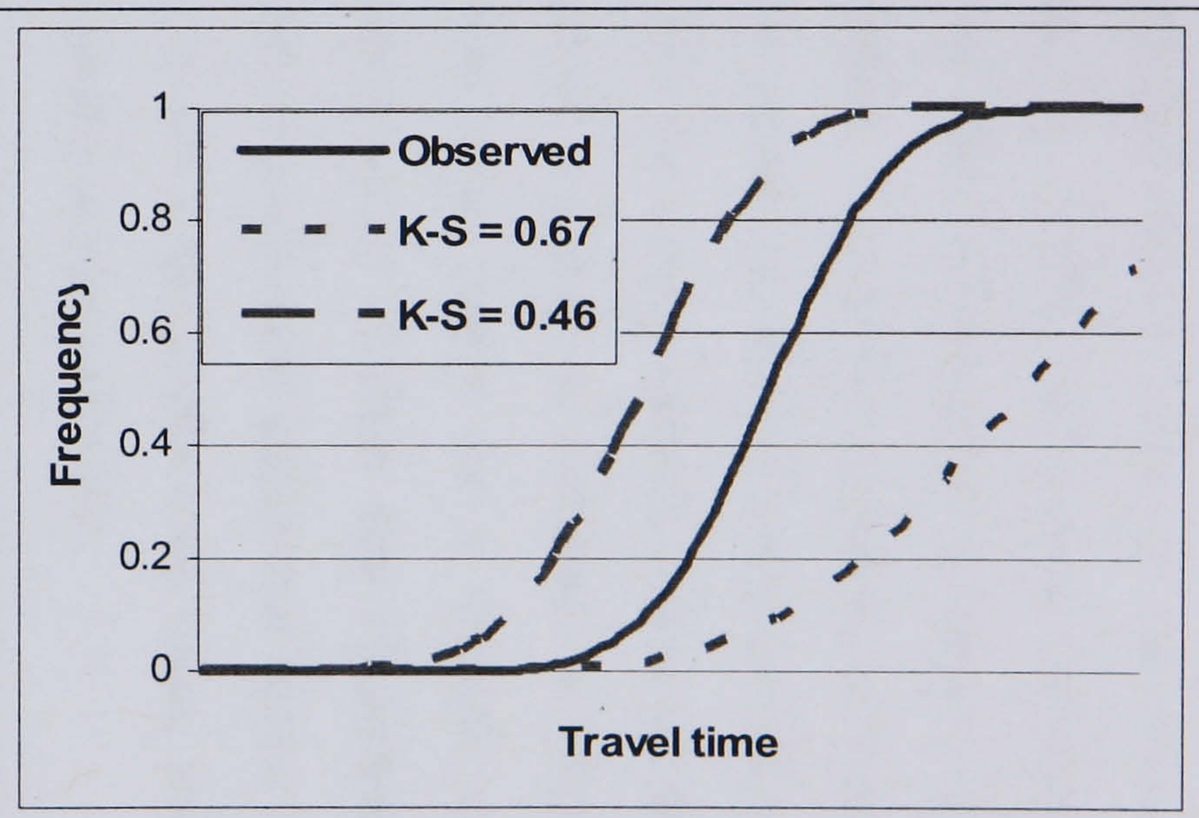
Excellent fit



Plausible fit



Poor fit



Very poor fit / No fit

Figure 6.2: Illustration of different levels of goodness-of-fit

6.4. Determining sample size

When we analyse the variability between model runs or between days of field measurement, it is vital that the sample (i.e. number of days or number of runs) is big enough to ensure credible estimation of the objective function. As shown in chapter 5, some existing methodologies for TMM calibration discuss the required number of model runs, but they are only motivated by the will to guarantee reliable estimates of mean values, and are therefore not applicable here. Note that the sought sample size is the number of model repetitions (and daily field measurements) required to calculate a *single* value of the objective function, namely the objective value based on a specific candidate set of parameter values. To evaluate the goodness-of-fit of other sets of parameter values, the same number of model repetitions should be performed again and again. Determining the sample size in the case discussed here is not a straightforward issue, and it is therefore explored from both a theoretical and an empirical point of view.

For a theoretical estimate of the required sample size, we make two assumptions. First, we assume that the distribution of travel times is lognormal. This assumption is supported by the studies of Turnquist (1978), Strathman and Hopper (1993), Mohammadi (1997) and many others. Not all studies of TTV agree that lognormal distribution is the most appropriate. For instance, May et al (1989) find that normal distribution fits travel time measurements; Talley and Becker (1987) suggest an exponential distribution; and Guehthner and Hamat (1985) suggest the gamma distribution. Still, the lognormal distribution of travel times seems better established than the alternative assumptions as it is repeatedly mentioned in a large number of studies (the full review was presented in chapter 5). The second assumption we make is that it is sufficient to look at the statistical properties of TTV itself, i.e. of the standard deviation of travel times, even if what we actually measure after running the TMM is the objective function and not this standard deviation. This assumption is simply made in order to avoid the mathematical analysis of the level of significance of the objective value, which is more complicated. It would indeed be interesting to look at the sample size issue from a more statistically rigorous perspective, in future work.

If travel times are lognormally distributed then their natural logarithms are normally distributed, and the variance of the logarithms is distributed chi-squared. Based on our assumptions we can therefore determine the sample size by imposing restrictions on the width of the confidence interval of the chi-squared-distributed variance. A confidence interval on the standard deviation in this case is given by (Stephenson 2004; Siegrist 2004; Kendall 2004):

(6.2)

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}$$

where σ is the real standard deviation, s is the sample standard deviation, n is the sample size, α is the desired level of confidence, and χ^2 is taken from the chi-squared distribution with n-1 degrees of freedom. To assure a certain level of significance of the ratio of the real standard deviation to the estimated standard deviation, we determine:

(6.3)

$$\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2}}} < \frac{\sigma}{s} < \sqrt{\frac{n-1}{\chi^2_{\alpha/2}}}$$

The range of ratios of the real and the estimated standard deviation, based on different numbers of model runs and different levels of confidence, is presented in figure 6.3. The horizontal axis represents the number of degrees of freedom, which equals n-1. The two groups of curves show the upper and lower bounds. For example, if for a certain combination of sample size and level of confidence we get an upper bound at 130% and lower bound at 65%, we can deduce that the true standard deviation is not more than 30% above or 35% under the TMM-based estimate.

Figure 6.3 illustrates that up to around 40 TMM runs, any additional run makes quite a substantial improvement in the estimation accuracy. Beyond 40 runs, the slope of all curves becomes milder, hence the increased accuracy might not compensate for the additional runtime.

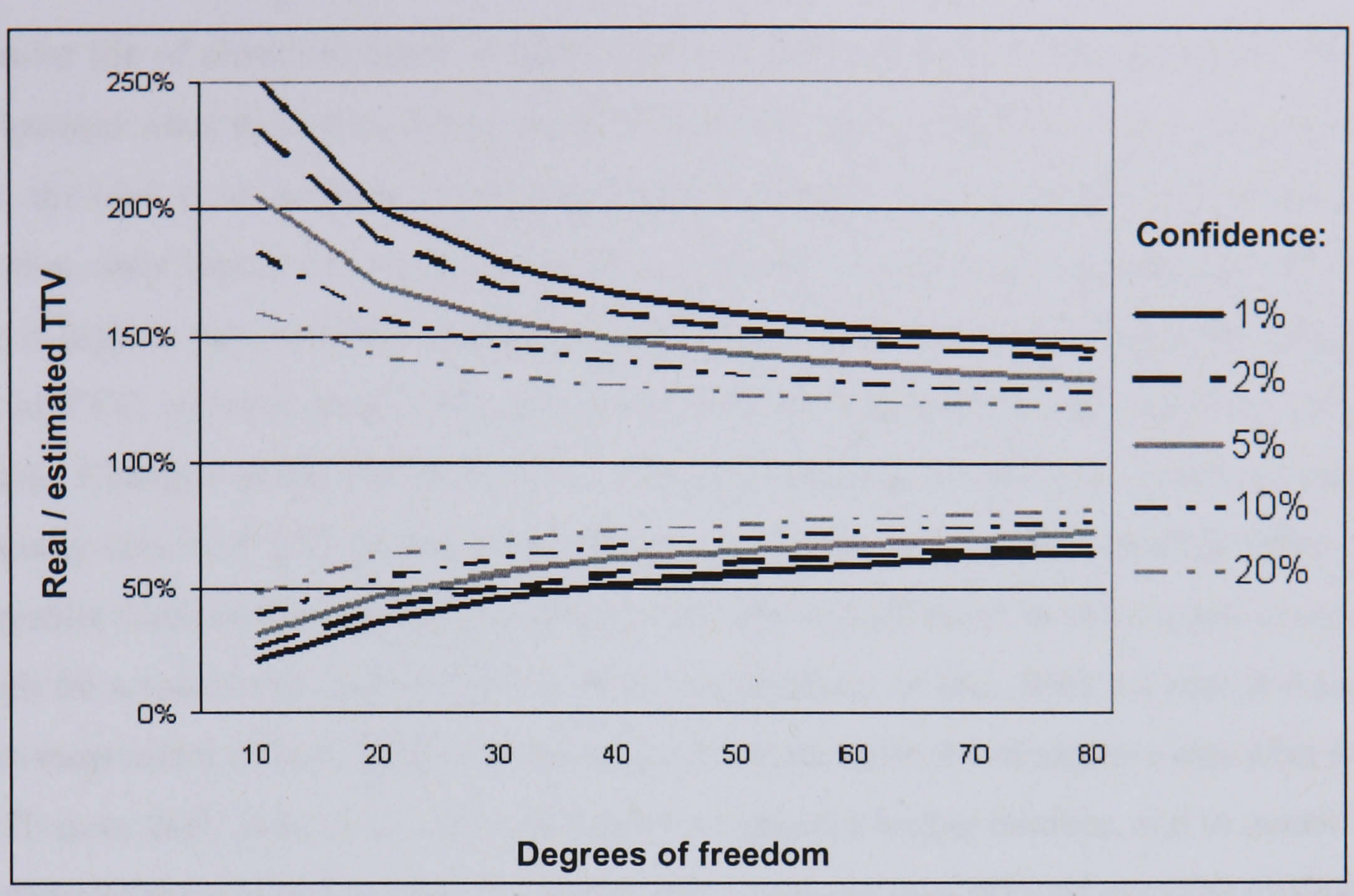


Figure 6.3: Theoretical range (upper and lower bounds) of the ratio of the real TTV to estimated TTV

To examine whether this approximated theoretical estimate of the required sample size is reasonable, we explore the same issue empirically. This is done through an experiment where the objective function is repeatedly evaluated with an increasing number of model runs, and the convergence of the objective value is checked. This experiment involves using the TMM and the procedure that calculates the objective value in the same way as in the full calibration algorithm described later. The TMM, the calibration algorithm, the test network and other relevant features have not been presented yet; therefore, suffice it to say at this point that the TMM used is DRACULA, the calibration set includes 21 parameters, and the test network includes a small section of York city centre. Full details of all these are presented in chapter 7.

Several artificial sets of observed travel time measurements were generated for the experiment, each one corresponding to a different level of TTV. In addition, several sets of values for the 21 calibration parameters were generated; the values are random but within the feasible range for each parameter (as described in chapter 7). Subsequently, several series of model runs are carried out, each series with a particular parameter set and a

particular file of observed times. In each series of runs, the value of the objective function is calculated after each run, based on all runs in the series that have been made till that point; the idea is to see how the objective value changes with an increasing sample size. Note that since this is still not a calibration experiment, it does not matter if the objective value is high or low, or if the artificial files of observed measurements stand for a realistic level of TTV; we only want to see how many runs are necessary for the objective value to stabilise. Changes in the objective value with an increasing sample size, based on various artificially-observed TTV patterns and various parameter sets, are presented in figure 6.4. The results confirm that generally, a series of 40 runs and 40 daily measurements should be enough for a reasonable estimate of the objective function. In fact, from the results it seems that in most series of runs there were no major fluctuations in the objective value after more than 20 runs. Still, since the theoretical analysis implied a higher number, and to assure that the estimates are prudent enough, the forthcoming analysis uses 40 runs per each evaluation of the objective function.

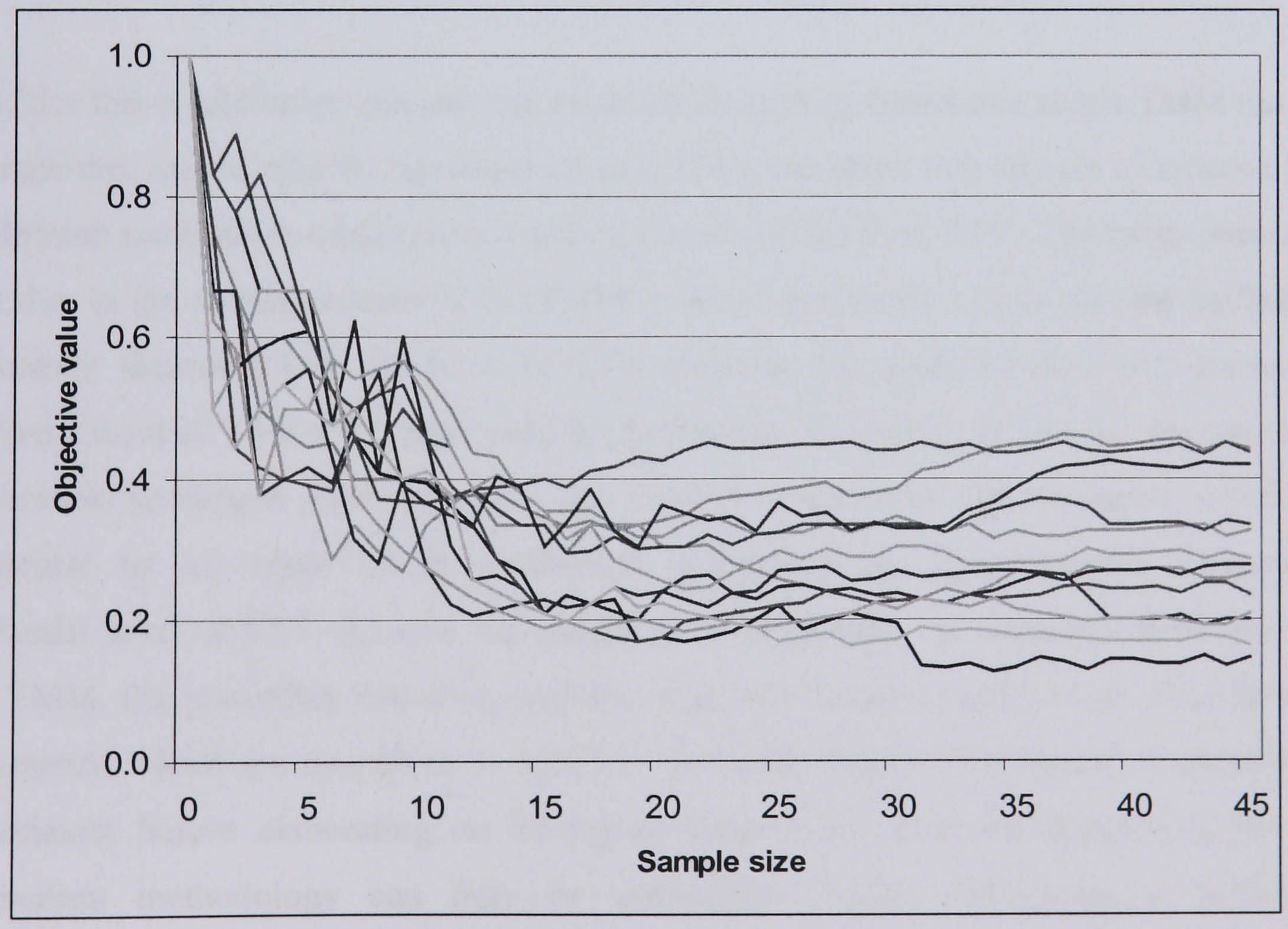


Figure 6.4: The objective value with increasing sample sizes

A brief note should be made on another statistical concern. Some recent studies of TMM calibration and validation (e.g. Toledo and Koutsopoulos, 2004) discuss issues relating to the assumption known as "the i.i.d. assumption", which is a requisite for some common forms of statistical analysis. The essence of this assumption is that the observed and simulated samples are *independent* draws from *identical distributions*. The mentioned recent studies observe that in some analyses of variability in TMM outputs, the i.i.d. assumption is inappropriate, since the entire set of outputs of a particular run is derived from the same random draws, and consequently the estimate of variability is based on dependent measurements. It is important to stress that this problem does not apply to the case discussed here, but only to discussions of temporal or spatial variability. We examine the distribution of travel time measurements taken independently from different runs and based on different random draws, and thus the i.i.d. condition is met.

6.5. Capturing travel time variability in the model parameters

The idea that a calibration process can establish an analogy between a single TMM run and a single day can be reliably implemented only if we can show that any set of values of the calibration parameters consistently stands for a certain level of TTV. The main concern is that due to the random nature of the TMM, a set of parameter values that the calibration procedure identifies with the level of TTV found in the observed data will generate a different level of TTV when later used for prediction. Therefore, before moving on to the calibration procedure itself we carry out another experiment that examines whether a particular set of values of the calibration parameters can systematically represent a particular level of TTV. As with the sample size experiment, the current experiment uses the TMM, the procedure that evaluated the objective function, and the set of calibration parameters which are described in detail in the next chapter. We choose to present the experiment before elaborating on its inputs because the case for developing the full calibration methodology can only be established if this experiment is performed successfully.

For the experiment we use again artificial travel time measurements as well as randomly-generated (but feasible) parameter values. What we check this time is how consistent with each other are 20 different estimates of the same objective value. Each of the 20 estimates is calculated with the same inputs (parameter set and observed times) using 40 model runs (i.e. for each combination of a specific level of TTV and a specific parameter set the TMM is run 800 times).

One set of results of the experiment (out of many sets) is illustrated in figure 6.5. Each one of the three curves represents a different series of 20 estimates of the objective function; for all three curves the same set of parameter values was used, but for each curve, the calculation of the objective value was based on a different set of observed times. These three observed time sets were generated from distributions with the same mean but different levels of variability: the ratio of the standard deviation to the mean was 10% in the first set, 20% in the second set and 30% in the third set. The diagram presents the frequencies of different objective values within the different sets of 20 estimates of the same objective. The general idea is that the easier it is to identify differences between the curves, the more established is the concept that a specific parameter set can be linked to a specific level of TTV.

We come to the following conclusions from the diagram:

1. The fact that the three curves are clearly distinguishable from each other signifies that the examined parameter set stands for a specific level of TTV. When the TTV in the observed times that were used to calculate the objective is similar to the TTV that the parameter set stands for, we get a curve with relatively low values (the left curve in the diagram). When the TTV in the observed measurements and the TTV represented by the examined parameters are very different from each other, we get high values (the right curve). We do not know, at this stage, the level of TTV that the examined parameter set captures, or which parameter values capture the level of TTV in our real observed data; but a calibration process should help us identify it.
2. The fact that each curve encompasses a range of values indicates that a low value of the objective, when a particular set of parameter values is examined, implies a high *chance* that this parameter set stands for the desired level of TTV. Since we are

using a probabilistic model, we cannot expect it to determine with absolute confidence whether a candidate parameter set is good or bad. But since the calibration procedure we develop later is an iterative process, where the objective value is repeatedly re-estimated till there is consistent improvement, even an indicator of high *chance* of success (rather than an indicator of success) can be used satisfactorily.

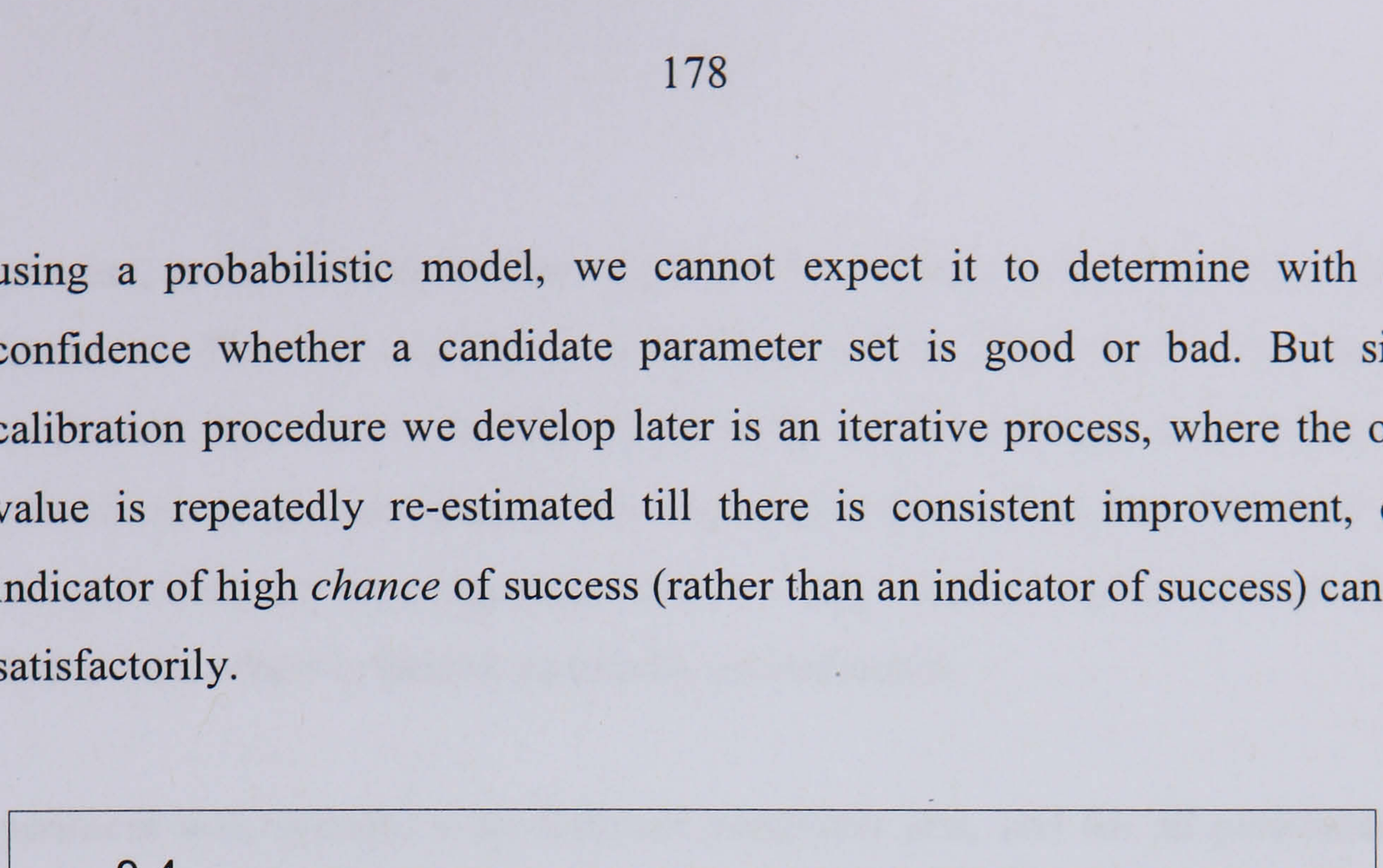


Figure 6.5: Repeated evaluation of the same parameter set with different input data

3. There are overlapping areas between the curves in the diagrams. If for a specific parameter set and specific observed data we calculate the objective value only once, as we intend to do in the calibration procedure, the resulting value might belong to several adjacent curves, some of which imply better fit than others. We could have more certainty about the objective value by calculating it more than once or by using more TMM runs for each value, but this will require higher runtime and is therefore undesirable. As mentioned in the previous paragraph, the limited fluctuations of the objective value are tolerable if they occur along an iterative

process, as the small error they might involve is likely to be corrected in subsequent iterations. The only real risk is in the last iteration, when the final parameter set is evaluated; this stresses the importance of the validation stage. In principle, the same procedure as the one used in this experiment (i.e. calculating the same objective several time) can be adequately used to verify whether a parameter set that seems better than others is indeed repeatedly proved better.

The experiment was repeated with different parameter sets, and for all parameter sets the results were similar to those presented in figure 6.5. When the objective function was repeatedly evaluated with the same parameter set and observed data, the resulting values lay within a relatively narrow range. Changing either the parameter set or the observed data resulted in objective values outside this range. The conclusion is that with the exceptions discussed above, and subject to validation of the chosen solution, a specific parameter set can capture a specific level of TTV.

6.6. Computational difficulties

To calibrate the TMM we wish to minimise the value of the objective function developed earlier. As in many other calibration problems, this is done through an iterative process, where in each stage a particular candidate solution (i.e. parameter set) is examined, and new candidates are repeatedly generated in an attempt to reach improvement in the objective value.

There are numerous methods for solving minimisation problems (for a thorough review see Press et al, 1992); but from a computational perspective, our problem is relatively a complex one, and most methods cannot be applied to solve it efficiently. One reason for this complexity is the multidimensionality of the calibration problem: there is a big number of parameters to calibrate. In the experiments presented in chapter 7, for instance, the calibration set includes 21 parameters even though efforts were made to include only those parameters that seemed essential for the estimation of TTV. The majority of existing calibration procedures (as reviewed in chapter 5) focus on smaller calibration sets. The

space where the optimal parameter set is to be sought has therefore multiple degrees of freedom; many optimisation techniques are not suitable for such problems and thus they cannot be used to solve the problem formulated here.

Another reason why our calibration problem is difficult to solve has to do with the way the objective function is computed. Calculating a single value of this function requires 40 TMM runs; even in a very small network, each run takes at least a few minutes, and the entire series of runs might take a few hours. For illustration: a single run of the DRACULA model with a 17-zone, 60-link test network, on several common types of computers, with some graphical features disabled to reduce runtime, takes about 3.5 minutes (the model and the network are presented in chapter 7). If 40 runs are required to calculate the objective function, it takes 140 minutes to obtain one value of this objective. The minimisation process is likely to require quite a few iterations till a good solution is found, and it is therefore crucial to choose an optimisation procedure that needs the smallest possible number of evaluations of the objective value per iteration. Unfortunately, this requirement implies that the most efficient solution methods cannot be used for our problem. Any minimisation procedure that uses derivatives of the objective function is unsuitable, because estimating the gradient of the objective function numerically requires calculating its value at several points, and the computational effort this involves is excessive. Analytical calculation of the derivatives of the objective function is obviously impossible, since it is not a direct function of the decision variables. As Press et al (1992) point out, "algorithms using the derivatives are somewhat more powerful than those using only the function, but not always enough so as to compensate for the additional calculations of derivatives".

Some efficient optimisation approaches, such as genetic algorithms, are also not suitable for the discussed problem although they do not use derivatives. This is because these methods involve processing a group of feasible solutions at every iteration, rather than focusing on an individual candidate solution at each stage. Again, such approach requires multiple evaluations of the objective function at each iteration, which is impractical when many TMM runs are needed for each evaluation.

The "downhill simplex method", described by Press et al (1992), is one of the few optimisation procedures that seem relatively suitable for the current problem. It is not known as an efficient method: its progress towards an improved solution is slow, and it also does not guarantee convergence to a global optimum. However, it does guarantee improvement in the parameter set, and given the complexity of the problem, this should not be belittled. As briefly explained in chapter 5, the downhill simplex method presents a set of possible solutions of the optimisation problem as a multidimensional geometrical shape, called simplex; each vertex of the simplex represents a single solution. In each stage of the process, the vertex with the worst objective value is replaced with a new vertex through one of several possible geometric manipulations, such as reflection of the simplex through its base, expansion or contraction. In most iterations, the value of the objective should be calculated only once; this is the most appealing feature of this method for our needs.

There are two particular stages in the downhill simplex method where many objective evaluations are necessary. First, there is the starting stage of the process, in which the entire simplex must be initialised, i.e. a value of the objective function must be attributed to each of the vertices of the initial simplex; this is inevitable despite the long runtime it requires. Second, one of the possible simplex manipulations involves changing all vertices but the one with the worst value. This can be seen as a re-initialisation of the simplex, and is performed only if the other types of manipulation do not prove useful. In the current context, the high time consumption of the re-initialisation stage is a serious problem. To avoid the need to re-initialise the simplex in a case where the other available modifications fail, we adjust the simplex method to our needs by introducing an alternative search technique. Our experience with this amended simplex method suggests that in a big majority of the cases where none of the simple modifications can improve the objective value, the alternative search resumes this improvement. The traditional simplex method and our modification are explained in the next section.

6.7. The simplex method and its modification

The downhill simplex method (Press et al, 1992) is commonly used to search optimal solutions for multidimensional minimisation problems. In a problem with N decision variables, this method uses an N-dimensional simplex, which is a geometrical shape with N+1 vertices. This means that throughout the whole process, there are always N+1 candidate solutions. A required input is the objective function we wish to minimise; this function is used to ascribe a value to each vertex. Prior to the search process, the simplex should be initialised, i.e. a set of N+1 vertices should be generated (either randomly or based on previous knowledge) and the objective value for each vertex should be calculated. Each step in the iterative minimisation process tries to replace the vertex with the worst objective value with a new vertex, at a different location in the multidimensional space, where the objective value is lower.

The problems illustrated in chapter 7 have 21 decision variables and therefore they involve a 22-dimensional simplex. But to facilitate the graphical demonstration of the search process, this section considers a problem with two decision variables only. The three-vertex simplex that such problem uses is actually a triangle. The principles of the solution technique demonstrated in this section work equally with any number of variables.

The first attempts to find an improved vertex, in each iteration of our proposed procedure, are based on the traditional version of the simplex method, where new vertices are sought at different points along a single line. This line, that we call *the main search axis*, is presented in figure 6.6. The following symbols are used:

W is the vertex with the worst (i.e. highest) objective value.

S is the vertex with the second worst objective value.

B is the vertex with the best (i.e. lowest) objective value.

C is the centre of gravity of the base on the simplex.

The line along which an improved solution is sought in the current stage is the line that connects points *C* and *W*. We define the *factor* as the ratio $\frac{CV}{CW}$, where *CV* is the distance

from point *C* to any other point *V*, and *CW* is the distance from *C* to *W*. At the original vertex *W* (that we wish to move to a new location) *factor* equals 1.

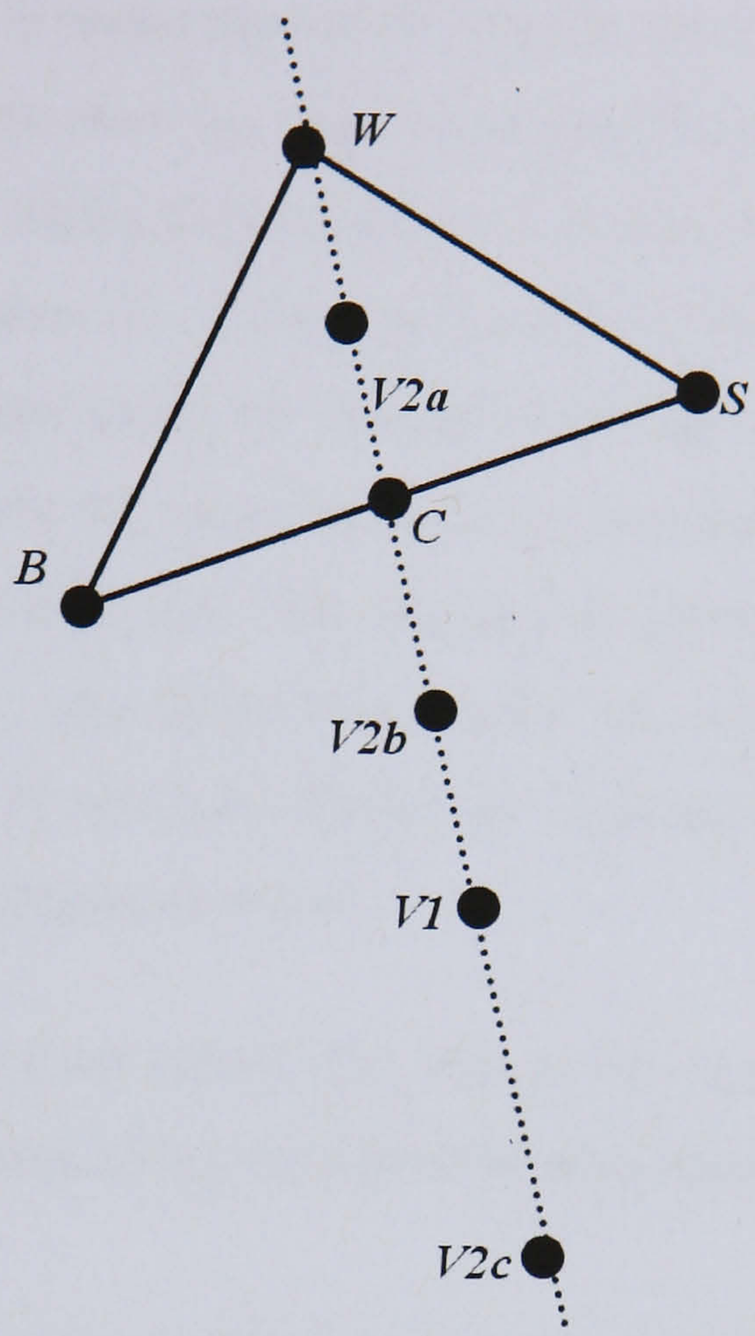


Figure 6.6: The main search axis

The search for solution along the main search axis is based on the following rules:

- The first point we examine as a candidate for replacing *W* is *V1*, where *factor* equals -1. If the objective at *V1* is better than at *S*, we have managed to find a successful replacement for vertex *W*, and we can move on to the next iteration, where a similar process is repeated.
- If the objective at *V1* is better than at *W* but not better than at S, we have managed to improve the worst objective value but not significantly. We therefore examine an additional point *V2b*, where *factor* equals -0.5. Of the two candidates, *V1* and *V2b*, we then choose the best, and move on.
- If the objective at *V1* is better than at *B*, it might signify that the farther we go from *W*, the better the objective value. We therefore examine an additional point *V2c*,

where *factor* equals -2. Of the two candidates, *V1* and *V2c*, we then choose the best, and move on.

- If the objective at *V1* is worse than at *W*, it might signify that searching for solutions with a negative factor does not lead to improvement. We therefore examine an additional point *V2a*, where *factor* equals 0.5. If the value at *V2a* is better than at *W*, we choose *V2a* and move on. If the value at *V2a* is worse than at *W*, we have failed to improve the simplex using the available strategies. In such case, the original simplex method restarts the whole process by re-initialisation. But as we discussed previously, given the high time consumption of calculating the objective value in the current problem, re-initialisation would be most inefficient. If all search strategies along the *CW* axis have failed, we continue the solution search along *the "last resort" axis*, as described below.

Note that only one (or none) of the points *V2a*, *V2b* or *V2c* is evaluated at a single iteration of the process. Hence, the search along the *CW* axis involves no more than two evaluations of the objective function.

If none of the points examined along the main search axis brings improvement, we start a new search along the line that we call *the "last resort" search axis*. This is an alternative to the re-initialisation stage in the traditional simplex method, which would be too lengthy in the current context. The new search axis is the line that connects vertex *B* with vertex *W*, as illustrated in figure 6.7. We re-define *factor* as the ratio $\frac{BV}{BW}$, where *BV* is the distance from point B to any other point *V*, and *BW* is the distance from *B* to *W*. As before, at the original vertex *W factor* equals 1. The idea of seeking alternative solutions along the axis that connects the best and worst vertices was originally proposed by Kaczmarczyk (1999). However, Kaczmarczyk does not use a systematic search along this axis, but only examines specific candidate vertices. In addition, the improvements proposed by Kaczmarczyk to the original simplex method are introduced very briefly, and hence were mainly used as a general inspiration to the technique described here. The following search process along the *BW* axis is different from Kaczmarczyk's suggestions.

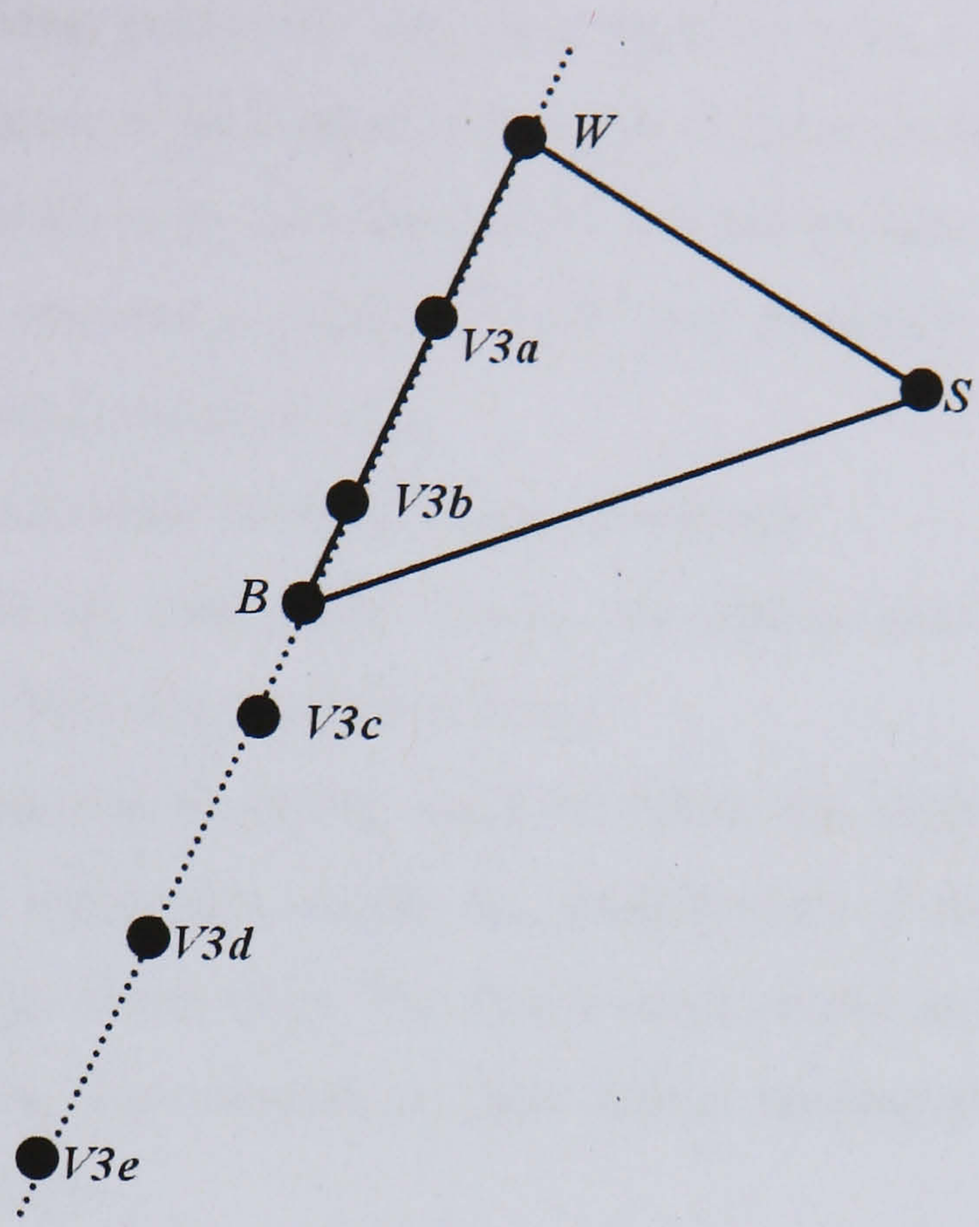


Figure 6.7: The "last resort" search axis

Starting from the original *factor* of 1, the process repeatedly reduces *factor* by 0.4 and re-evaluates the resulting new candidate vertex, till a better vertex is reached. The first examined vertex (*V3a*) has a *factor* of 0.6; the second (*V3b*) has 0.2; then -0.2 (*V3c*), -0.6 (*V3d*), -1 (*V3e*) and so on. In principle, the search process can repeat many times if there is no improvement. However, in our experiments improvement was always experienced within no more than 7 evaluations of candidate vertices along the *BW* axis (i.e. with *factor*>=-1.8), or much less than this in most cases.

6.8. Calibration algorithm

The full calibration procedure, based on the modified simplex method, is described in detail as a flow chart in figure 6.8. The section of the flow chart with dotted frames describes the initialisation stage. The section with solid frames describes the parts of the algorithm that follow the traditional simplex method. The section with dashed frames describes the additional stages in the modified method. The sub-process of calculating the objective

value for a given candidate parameter set, i.e. a single vertex, is described separately in figure 6.9. This sub-process is performed every time in the main process when the value of the objective function needs to be calculated. It is this sub-process that repeatedly calls the external program that runs the simulation itself, and therefore more than 99% of the runtime of the entire process are spent in it.

The inputs to the process include the following information:

1. Initial values for all parameters. These are copied into the first vertex of the simplex, as they form one possible solution.
2. The feasible range and the likely range of values for each parameter. The feasible range is used to verify that during the modification of the simplex no parameter exceeds its realistic boundaries. The likely range, which is narrower, is used during the creation of the first simplex to draw initial random parameter values (for all vertices but the first).
3. A list of observed travel time measurements, used throughout the process to examine the difference between real-world TTV and simulated TTV.

The calibration procedure was programmed in C. The full program is available from the author. The program uses various file structures for inputting data and outputting results; the main structures are described in appendix B.

It was not found necessary to incorporate within the calibration procedure a clearly-defined criterion that causes, once it is met, the process to stop and deem the last vertex as the optimal solution; the reason for this is the following. A key feature of the procedure is that the improvement of the objective value is slow but steady. It is *slow* since the calculation of each value of the objective function takes a relatively long time, and also because between consecutive iterations there is normally quite a modest decrease in the objective value. It is *steady* because there is improvement in every iteration. Our experience is that even after many iterations, there is still gradual improvement. Due to the complex, multidimensional nature of the problem, it is not expected that the process will reach a global optimum. We therefore find that the best practice is to halt the process at the discretion of the user: better solutions can be reached at the price of higher runtime. Note that runtime issues are also discussed in chapter 7.

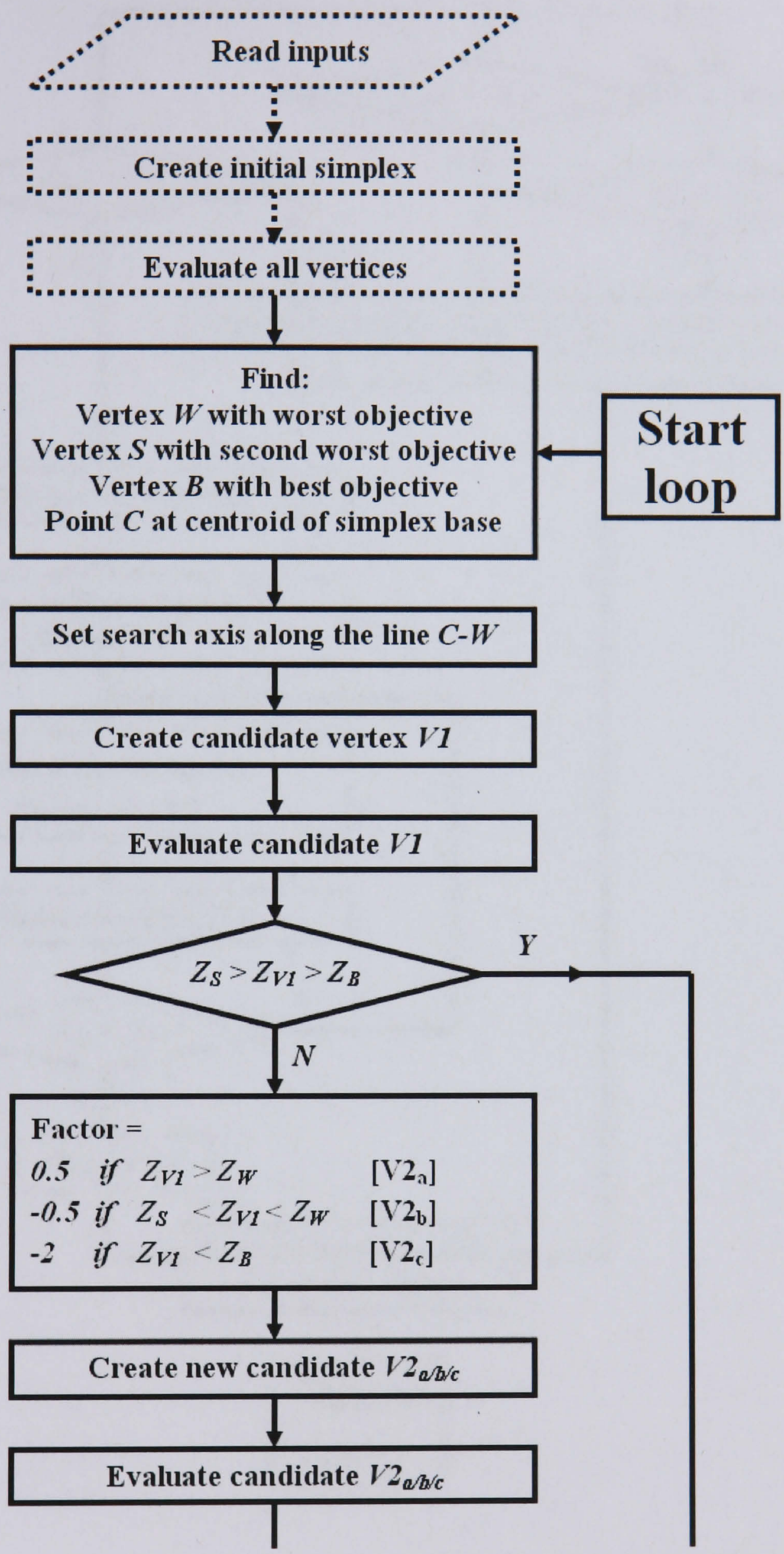


Figure 6.8: The calibration algorithm (continues on next page)

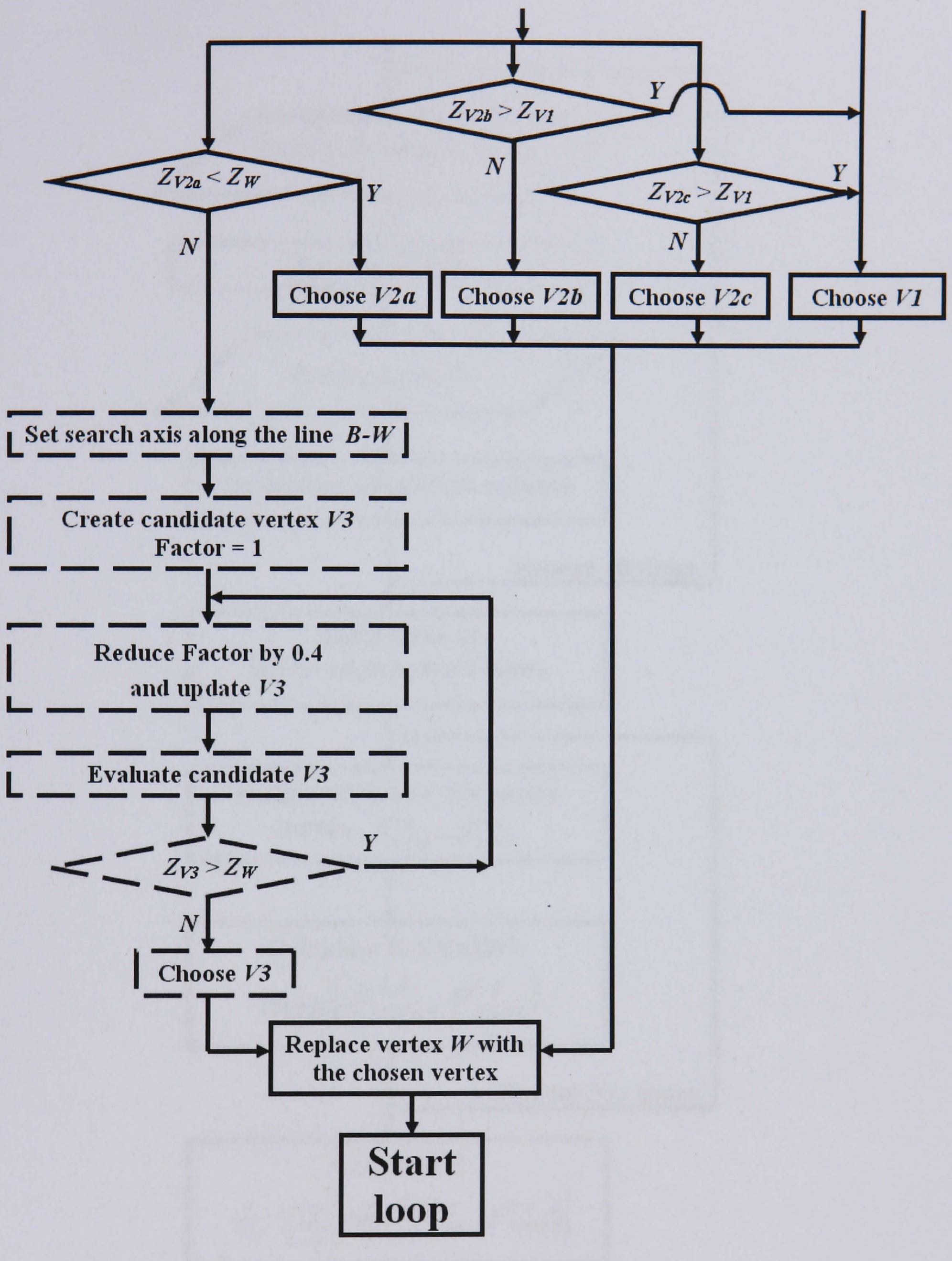


Figure 6.8: The calibration algorithm (continued)

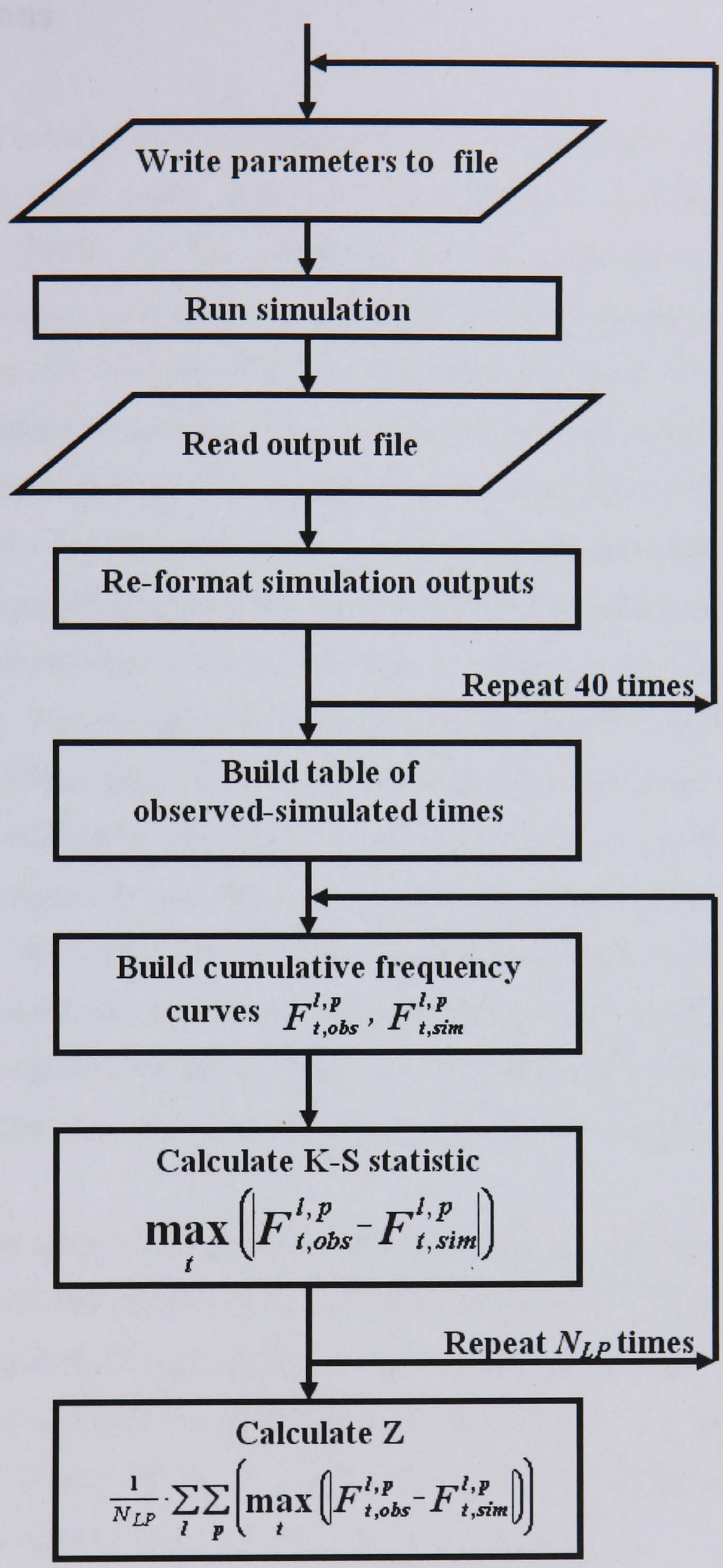


Figure 6.9: Calculating the objective for a single vertex

6.9. Conclusions

Randomness and heterogeneity in the elements of a transport network have a key role in the concept of microscopic traffic modelling. This led us to hypothesize that it might be possible to use TMMs for the ambitious task of estimating the extent of TTV in hypothetical scenarios, by introducing an analogy between a single run of the TMM and a single day in the real network. The fact that some practitioners use TMMs as if such analogy can be taken for granted was an additional motivation for the analysis presented here. To enable estimation of TTV through inter-run analysis of a TMM, the model needs to be calibrated in a way that pays attention to variability in the outputs, as opposed to most existing calibration methods, that focus on analysis of mean values only.

This chapter provided some evidence that there is a solid case for the proposed concept of inter-run analysis. This was done, for example, by showing that running a TMM with one set of parameter values consistently results in a relatively small range of levels of TTV. We presented a full calibration algorithm, where the objective is to minimise the difference between the distribution of simulated times and a given distribution of observed times. It was shown that this calibration problem is computationally challenging, because the solution space is multidimensional and the objective function is non-differentiable and slow to estimate. No common solution method seemed suitable as is for this problem, and the presented algorithm was therefore based on our own modification of the well-known simplex method.

We also discussed some of the limitations of the proposed methodology. The fact that all parameters are calibrated simultaneously, without paying much attention to the features and phenomena that each parameter stands for, is a compromise. There is also some risk of bias since no TMM can replicate the entire range of causes for TTV in the real world.

The methodology developed here has not yet been tested and illustrated. Experiments that use the calibration algorithm are presented in the next chapter.

Chapter 7

Calibration experiments

7.1. Introduction

This chapter demonstrates the application of the methodology developed in chapter 6 for calibration of a TMM. The main goal of the calibration process is to make the variability between time measurements from different runs of the TMM similar to the variability between travel times on different days in the real transport system. If such calibration is performed successfully, then it is possible to use the TMM as a tool that generates forecasts of the level of TTV. Basing TTV estimates on a microscopic traffic modelling might prove very useful in transport analysis, because TMMs are sensitive to various local factors, to the level of demand and to differences between network configurations. As we saw earlier in the thesis, the tools currently available for estimating TTV do not exhibit these features.

Two experiments are described in this chapter. The first experiment is meant to test the calibration mathodology, by separately performing the calibration procedure in three imaginary scenarios and then verifying whether the calibrated parameters in each scenario can pass a validation test. The network used in this test represents a small (but real) part of York city centre. The second calibration experiment is carried out not as a test but as our main attempt to adjust the TMM parameters to the characteristics and behavioural patterns of the York network and its users. The input data for this experiment are real travel time measurements, and the network used covers a major part of the city.

The TMM used for both testing and implementing the calibration procedure is DRACULA (Liu, 2006), which is a package commonly used in the UK. DRACULA has been developed at the University of Leeds since 1993, and its current version includes, apart from sub-models for car following and lane changing, such additional features as demand fluctuation, randomness in vehicle characteristics, complex traffic controls and bus priority measures, and so on.

7.2. Choice of parameters

In the experiments presented here we calibrate a relatively big set of 21 parameters. The choice of these 21 parameters was mainly based on common sense and on some informal past experience. In the current scope we only calibrate parameters that relate to driving behaviour and variable demand; parameters relating to route choice, departure time choice and so on are not considered. As mentioned in previous chapters, it is likely that with a better specification of the parameter set, the results could be improved. The following parameters are included in the calibration set:

1. Four parameters of the gap acceptance model: the normal acceptable gap; the minimum acceptable gap; the time waited before accepting a lower gap; and the time waited before accepting the minimum gap. All these are standard DRACULA parameters, which have to do with the way drivers decide under what conditions they would be willing to use a temporary gap, created between vehicles in the opposing traffic, in order to enter a junction or change lane.
2. Four groups of vehicle characteristic: normal acceleration parameters; maximum acceleration parameters; normal deceleration parameters; and maximum deceleration parameters. Each of the four groups contains four parameters: mean value for cars; coefficient of variation for cars; mean value for buses; and coefficient of variation for buses. All these are standard DRACULA parameters, which define the distribution of features relating to acceleration and deceleration across the various types of vehicles that constitute the general traffic.
3. An additional parameter, which is external to the DRACULA model, stands for the standard deviation of the overall level of car traffic in the network. Although such parameter is not part of the DRACULA package, it seemed important to incorporate this element of travel demand fluctuation into the model as this is presumably a major cause for TTV. DRACULA does include a factor by which all car flows are multiplied in every run, but this parameter is not suitable for calibration, as a fixed value would not result in fluctuation. Therefore, the calibration parameter represents the standard deviation of the DRACULA standard demand factor, and a different

value for the demand factor is randomly drawn in every run from the distribution defined by the calibration parameter. Note that the factor affects car flows only and not the number of bus passengers or bus journeys; and in addition, it applies equally in all parts of the network. In this sense, it is indeed a rather simplistic way of accounting for variations in the general level of congestion, whose real nature is clearly not that simple or uniform. However, allowing for a more realistic pattern of variability in the surrounding traffic will inevitably require calibrating more parameters, and we therefore leave it for other studies.

As explained in detail earlier in the thesis, calibrating the values of all these parameters jointly is not ideal. If it were possible, it would be preferable to determine their values by observation and measurement; but this requires very intensive data collection and analysis, which have so far never been possible in the study area.

The calibration procedure requires five sets of input values for the calibration parameters. The first is a set of initial values, which the algorithm stores as the first vertex of the simplex, since it forms one possible solution. Two more values for each parameter are needed to define the lower and upper bounds of its likely range; all the vertices of the initial simplex, apart from the first one, are generated randomly but within this range. Two additional values define the feasible range (minimum and maximum) of each parameter; during the calibration of each parameter the algorithm makes sure it does not go beyond this range. The input values used here are presented in table 7.1.

Parameter	Initial value	Likely range	Feasible range
Normal acceptable gap (seconds)	3	1 – 5	0 – 60
Minimum acceptable gap (seconds)	0.5	0.2 – 2	0.1 – 60
Time waited before accepting reduced gap (seconds)	30	20 – 40	0.1 – 600
Time waited before accepting min. gap (seconds)	60	40 – 80	0.1 – 600
Mean of car normal acceleration (m/s^2)	1.5	1 – 5	0.1 – 60
Coefficient of variation of car normal acceleration	0.1	0 – 0.3	0 – 2
Mean of car maximum acceleration (m/s^2)	2.5	2 – 5	0.1 – 60
Coefficient of variation of car maximum acceleration	0.1	0 – 0.3	0 – 2
Mean of car normal deceleration (m/s^2)	2	1.5 – 5	0.1 – 60
Coefficient of variation of car normal deceleration	0.1	0 – 0.3	0 – 2
Mean of car maximum deceleration (m/s^2)	5	3.5 – 6.5	0.1 – 60
Coefficient of variation of car maximum deceleration	0.1	0 – 0.3	0 – 2
Mean of bus normal acceleration (m/s^2)	1.5	0.8 – 2	0.1 – 60
Coefficient of variation of bus normal acceleration	0.1	0 – 0.3	0 – 2
Mean of bus maximum acceleration (m/s^2)	1.6	0.8 – 2	0.1 – 60
Coefficient of variation of bus maximum acceleration	0.1	0 – 0.3	0 – 2
Mean of bus normal deceleration (m/s^2)	1.5	1 – 4	0.1 – 60
Coefficient of variation of bus normal deceleration	0.1	0 – 0.3	0 – 2
Mean of bus maximum deceleration (m/s^2)	2.5	1 – 4	0.1 – 60
Coefficient of variation of bus max. deceleration	0.1	0 – 0.3	0 – 2
Demand fluctuation (coefficient of variation of overall demand)	0	0.01 – 0.2	0 – 0.25

Table 7.1: Inputs to the calibration procedure

Figure 7.1: The test network

7.3. Test of the calibration algorithm

The network used for testing the calibration algorithm represents a section of the street network in the centre of York. Figure 7.1 contains a map of the area included in the test network and a plot of its computerised representation. The central intersection, Bootham Bar, is a key location in the centre of York, around which are many public buildings and the Old City, which generate intensive traffic. The roads that intersect at this point are all major urban streets: the A19 arterial (Bootham) to the north-west, Gillygate to the north-east, and St. Leonard's Place to the south.

The route sections, on which observed and simulated measurements are compared in the experiment, are the sections 1-2, 2-3 and 3-4 along one bus route, where 1 ,2, 3 and 4 represent bus stops. A DRACULA model for this test network has been prepared in previous works; the starting point for the current calibration experiment is the parameter values and other network features as determined in the previous works.

Three imaginary scenarios were defined, and an artificial set of observed travel time measurements was generated for each scenario. The mean travel time on each route section (1-2, 2-3 and 3-4 in figure 7.1) is the same in all scenarios, but the standard deviation differs between the scenarios. In scenario 1 the standard deviation of travel times is around 10% of the mean on each section, in the scenario 2 it is around 17% and in scenario 3 it is 25%. We regard these three scenarios as reflecting low, medium and high levels of TTV, respectively. With these sets of input data, three separate runs of the calibration algorithm were launched.

The progress of the calibration process in all three scenarios appeared similar. For illustration, figure 7.2 shows the gradual changes in the objective value throughout the calibration process for scenario 2. The coloured area above the number of a particular iteration describes the range of objective values in the simplex during that iteration. The upper contour of the entire coloured area connects the worst vertices at difference stages of the calibration process, which we try to replace. Along the lower contour, each value is the objective that corresponds to the best vertex found till that point. A new vertex is accepted only if it is better than the worst one, but naturally in most cases the new vertex is not as good as the best. Therefore, the slope of the upper contour is bigger that the slope of the lower contour. If in certain parts of the diagram it looks like there is no progress between two consecutive iterations, it is because there are sometimes several

vertices with very similar objective values, and when one of them is replaced with a better vertex, the other vertex with an equally-bad objective value still remains to be replaced.

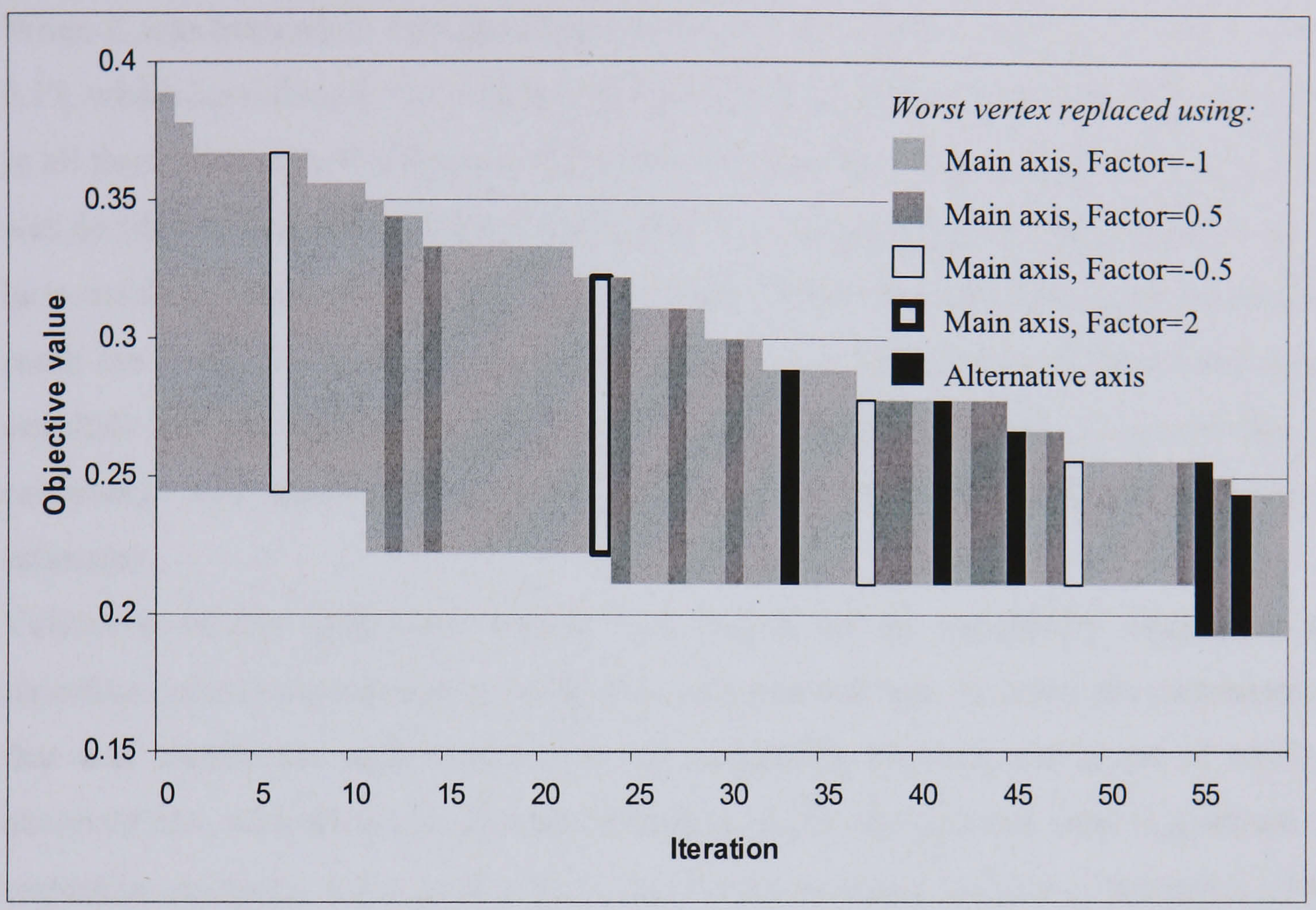


Figure 7.2: Progress of the calibration experiment

The different shapes and colours used in figure 7.2 show what type of simplex manipulation was used in each iteration; this also implies how many times the objective value had to be calculated. In most iterations, one calculation of the objective value was enough, as the candidate vertex created with *factor=-1* along the main search axis was found a successful substitute. But it seems that as the process moves on, it gradually becomes harder to improve the worst vertex, and a second candidate vertex (with other factor than -1) is evaluated more frequently. When a factor of 0.5, -0.5 or 2 is used, there are two evaluations of the objective per iteration: one unsuccessful attempt with *factor=-1* and then one successful attempt with the new factor. When the alternative search axis is used, the number of evaluations per iteration is three or higher. However, it should be reminded that the alternative axis was introduced here as a substitute for the re-initialisation stage in the original simplex method, that in the current case would require 21 evaluations of the objective. Our experience is that the number of evaluations

along the alternative axis is normally between 1 and 3, and hence we find this modification of the simplex method very efficient.

The calibration experiment in all three scenarios included around 60 iterations. The number of evaluations of the objective function during the 60 iterations is around 90. When it was decided to halt the process, the objective function reached values around 0.17, which according to the terminology described in chapter 6 signifies "plausible fit". In all three scenarios, there was still gradual improvement of the objective value when it was decided to stop the process; it seems that if we allowed additional runtime, it would be possible to reach some further improvement. However, it does not seem practical to reach the range of objective values defined earlier as "excellent fit". While we cannot conclude that the calibrated model perfectly fits the observed data, it appears that the calibration was successful in preventing the model from rendering seriously biased estimates.

Validation of the calibration results was carried out by repeatedly calculating the objective values that correspond to the chosen parameter sets. Namely, the parameter set that was chosen for each scenario in the calibration process, and a set of artificial observed data with the same standard deviation as the one that was used to generate the respective scenario, were used a few more times as input for new evaluations of the objective value. Results of this new series of runs, for all three scenarios, are almost identical to those depicted by the left curve in figure 6.5: different evaluations give different objective values, but the distribution of these values mainly encompasses values that indicate plausible fit. This gives evidence that the three calibrated sets of parameters consistently fix the TTV in the model outputs at a level that is adequately similar to the desired level.

Table 7.2 presents the values of all relevant parameters at the end of the calibration process. Due to the extreme complexity of the 21-dimensional solution space, we cannot expect the results to follow simple rules; the difference between scenarios in the value of a certain parameter cannot always be easily explained. We pay special attention to the parameters that stand for direct sources of variability in the model (these appear in the table in cells with a dark background). For some of these parameters, such as the coefficients of variation of bus normal acceleration and deceleration, it is apparent that the calibration process chose a higher level of variation in the scenarios where TTV is higher. This makes good sense, given that the optimisation process was based on comparing bus travel times. For some other parameters, such as the coefficient of

variation of car normal acceleration, differences between the scenarios seem less logical, and our understanding is that these parameters do not have a significant contribution to the extent of simulated bus TTV. The parameters that indicate high level of variation in scenarios with low TTV, as well as parameters whose values in the different scenarios are almost the same (such as the variation of car maximum deceleration) are probably of low importance for TTV estimation.

7.4. Calibration with real data

The basic experiments described in chapter 6, and the test described in the previous section, have shown that there is a serious case for the concept of inter-run variation analysis and the calibration algorithm that applies this concept. Although some features of this concept can benefit from further investigation and development, we leave this for future work. In the rest of the thesis the calibration algorithm is used as a legitimate tool for estimating the level of TTV. The current section is aimed at preparing a DRACULA model for our area of interest in the city of York, which can be used for generating TTV forecasts. This is done by running the calibration algorithm with a network of a considerable size, which covers a large part of York, and with real travel time measurements recently taken in the study area.

The travel time data used here are based on travel time records generated by a system installed on several buses that work on route number 4 in York. Route 4 connects Acomb in west York with the railway station, the city centre and the University of York in the east. It is a frequent service, with around 8 departures per hour during the morning peak. The scheduled runtime of a journey along the entire route is approximately 45 minutes. The system that generates the time records takes a note of the time when the bus arrives at each stop, and can therefore be used to generate a profile of the progress of the bus on each journey. Joining the records from multiple days forms the input required for TTV analysis; note that some basic analysis, based on the same data used here, was presented in chapter 1. The itinerary of route 4 (like the itineraries of all routes in the area) is coded in detail in the DRACULA network, and the calibration is carried out by comparing simulated and observed TTV along various sections of this route. The network used here and the itinerary of route 4 are presented in figures 7.3 and 7.4.

Parameter \ Scenario	Low TTV	Medium TTV	High TTV
Normal acceptable gap (seconds)	3.15	2.39	2.99
Minimum acceptable gap (seconds)	0.84	0.61	1.75
Time waited before accepting reduced gap (seconds)	30.43	31.96	37.52
Time waited before accepting minimum gap (seconds)	47.56	68.52	50.92
Mean of car normal acceleration (m/s^2)	4.41	2.48	4.28
Coefficient of variation of car normal acceleration	0.22	0.20	0.07
Mean of car maximum acceleration (m/s^2)	3.43	3.99	3.88
Coefficient of variation of car maximum acceleration	0.11	0.20	0.17
Mean of car normal deceleration (m/s^2)	2.12	3.63	3.77
Coefficient of variation of car normal deceleration	0.03	0.18	0.15
Mean of car maximum deceleration (m/s^2)	4.75	4.53	4.51
Coefficient of variation of car maximum deceleration	0.22	0.18	0.17
Mean of bus normal acceleration (m/s^2)	1.86	1.57	1.64
Coefficient of variation of bus normal acceleration	0.08	0.18	0.24
Mean of bus maximum acceleration (m/s^2)	0.90	1.14	1.82
Coefficient of variation of bus maximum acceleration	0.24	0.16	0.11
Mean of bus normal deceleration (m/s^2)	1.61	2.63	1.35
Coefficient of variation of bus normal deceleration	0.06	0.07	0.21
Mean of bus maximum deceleration (m/s^2)	2.25	2.87	0.48
Coefficient of variation of bus maximum deceleration	0.27	0.12	0.20
Demand fluctuation (coefficient of variation of overall demand)	0.04	0.06	0.08

Table 7.2: Values of the calibration parameters at the end of the process

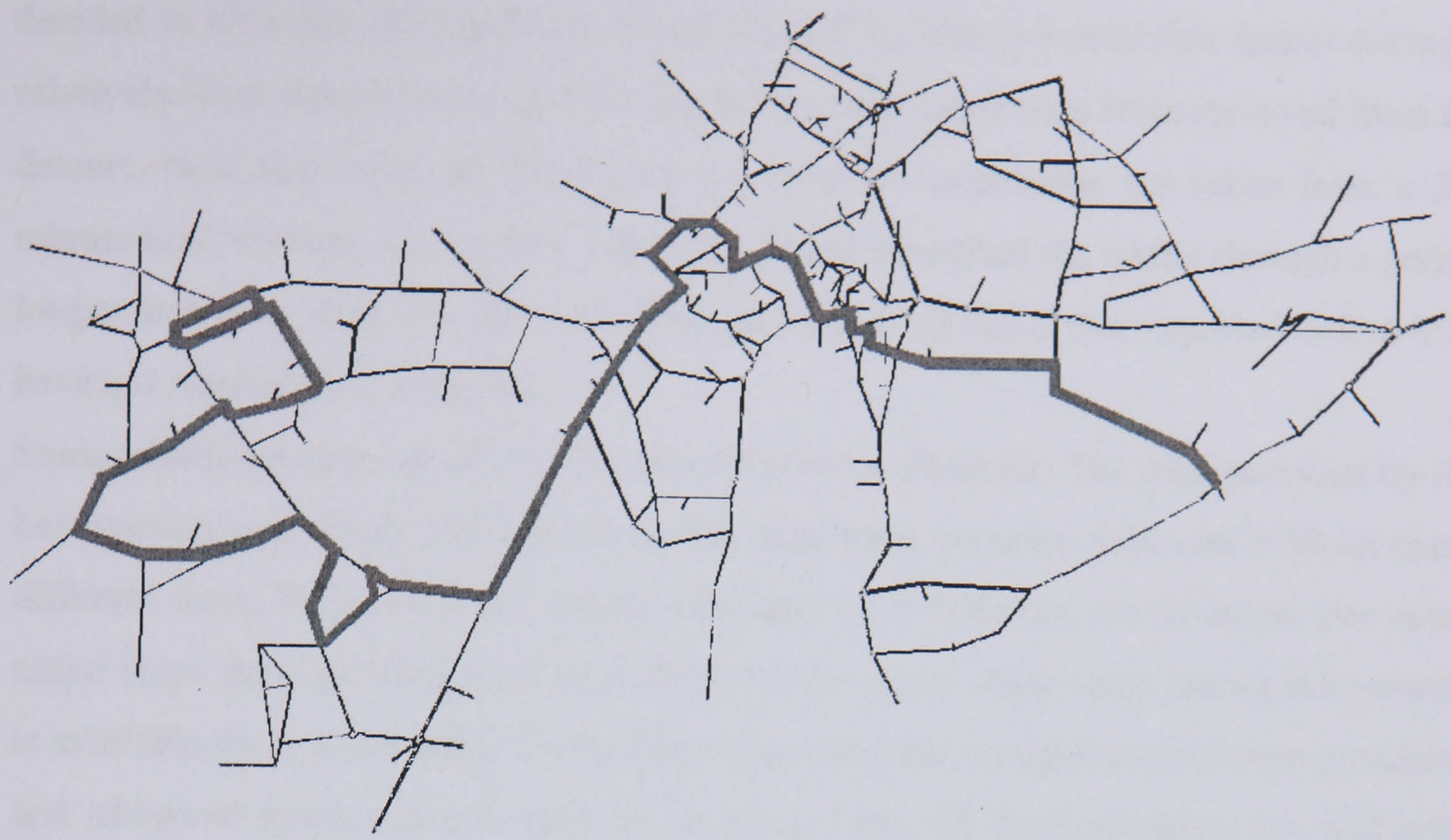


Figure 7.3: The network used for calibration and the itinerary of route 4

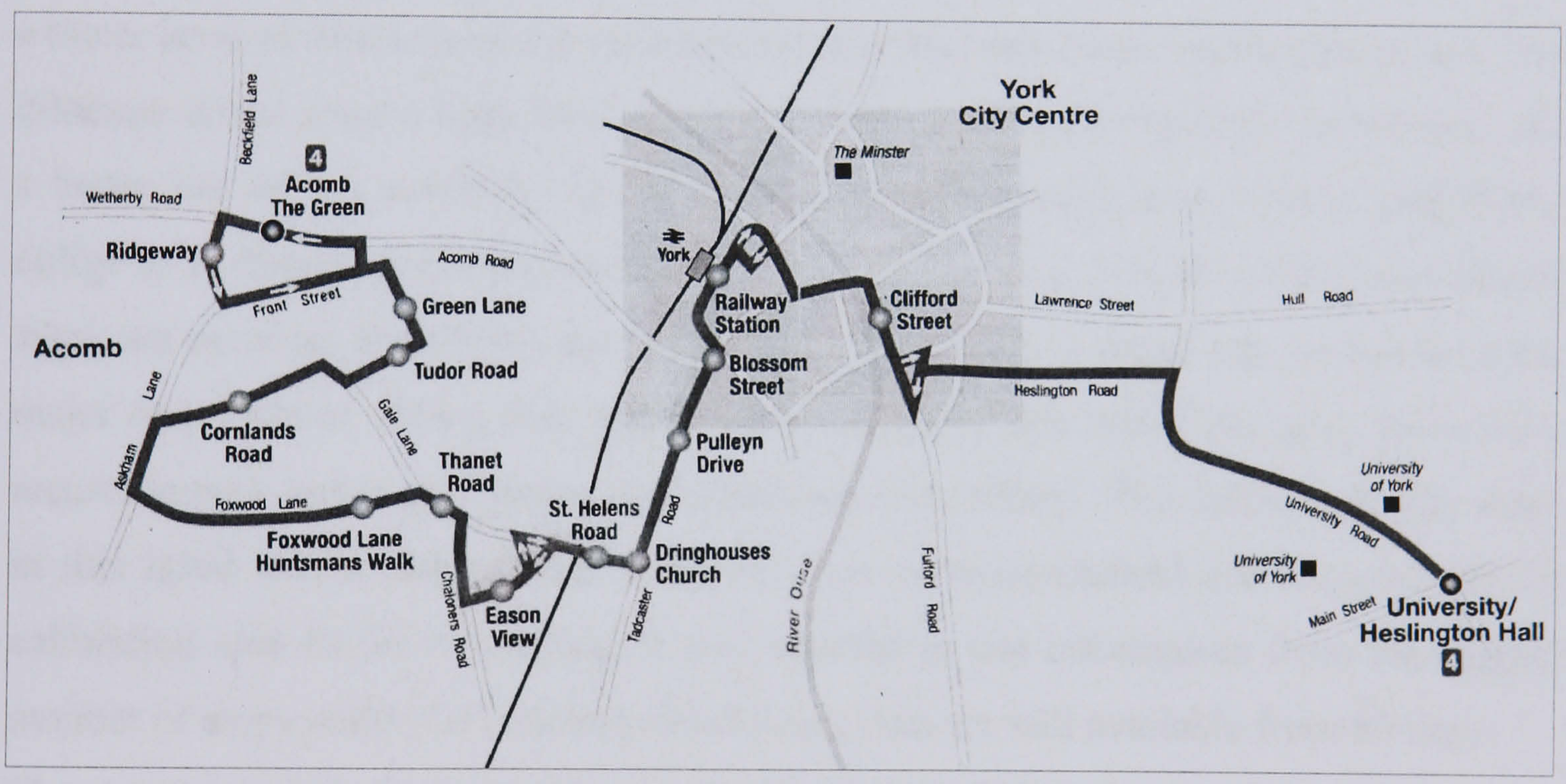


Figure 7.4: The itinerary of route 4, as published by the operator

Prior to the calibration itself, extensive re-editing and processing of the raw data was necessary. First, some simple pruning of the dataset verified that only measurements from normal working days are used, omitting any weekends or holidays. In addition, the DRACULA model has information of traveller demand during the morning peak period only, and it therefore had to be confirmed that the observed data used for calibration only cover bus journeys that depart during the morning peak. After examining the amount of daily measurements available from different departures of route 4, it was

decided to base the calibration on travel times from bus journeys that depart during a relatively short period, between 8:00 and 8:30. Other departures were removed from the dataset. Note that although the departures used for calibration are taken from a 30-minute time window, each actual run of the TMM simulated the traffic through a period longer than this, since the simulation runs as long as all buses that departed before 8:30 have not reached their terminals.

Some additional pruning of the raw data involved a dilemma. The data provided by the bus operator contained information on the departures between 8:00 and 8:30 on many different days, but each daily record consisted of a different list of stops. For some major stops there are data from most days, but for many other stops timing information is available from fewer days. To facilitate the automatic comparison between simulated and observed measurements, the list of stops from all days needs to be uniformly formatted as a set of sections. The sections can be specified either at a high level of detail, such that each section is the part of the route between two consecutive stops, or at a lower level of detail, such that each section also includes some intermediate stops. The dilemma arises since a high level of detail would enable more accurate calibration, and a better use of the available information about intermediate stops, but it would also oblige us to discard measurements from days when the data of some of the intermediate stops are missing. Specifying the sections at a low level of detail can be based on the major stops whose timing data appear on most days, and hence can give more daily measurements, but it also makes the calibration less refined. The decision finally made in this issue was a compromise: since 80 days of measurement are required (40 for calibration and 40 for validation), it was decided to use information from the biggest number of stops under the condition that timing data are still available from 80 days.

The input needed for running the model also includes the number of passenger boardings at each stop along the analysed route. The bus operator provided us with boarding data for route 4, but the stops were grouped into clusters according to fare zones, and the required number of passengers at each stop was not stated explicitly. To tackle this problem we used a slightly older study (Sinha, 2004) that contained a more detailed demand profile of the same route. We followed the internal distribution of boarding within each cluster of stops in the older study, to convert the information of boardings per cluster into information of boardings per stop. The overall number of passengers in each cluster was left as in the newer data obtained from the operator. The boarding profile, clustered into route sections as given by the operator, is presented in

figure 7.5. The modified profile, as fed into the calibration procedure, is presented in figure 7.6. The columns painted with a different pattern in figure 7.6 stand for bus stops that are shared by route 4 and some other low-frequency routes. When several routes share the same stop, DRACULA models passenger boarding without direct account of the desired destination of each passenger. To comply with this method of modelling, the number of passengers boarding the other routes at these stops (based on data from the operator) was added to the profile.

All the information prepared as input for calibration, including journey times and demand data, reflects the state of the York network in October and November 2004, during the morning peak of a normal working day. The data processing described above completed the necessary preparations for running the calibration procedure.

The progress of the calibration process is described in figure 7.7. The coloured area shows the upper, lower and all intermediate objective values in the simplex at every stage of the calibration. The different colours and patterns represent the different ways of modification of the simplex along the process (as in figure 7.2). The process was stopped after 75 iterations, when the best objective value was around 0.16. As in the test experiment, it appears again that at the beginning of the process, simple modification of the simplex with *factor=-1* performs well most of the time, and only one evaluation of a new candidate vertex is needed per iteration. Modification with *factor=0.5* (and also, less frequently, with *factor=-0.5*) is increasingly used as the process moves on, requiring two evaluations per iteration. The alternative search axis was not used frequently throughout the experiment, but its use in iteration 46 saved considerable runtime, since an improved vertex was found immediately (with *factor=0.6*), and the need to re-initialise the simplex (as the traditional simplex method would require) was avoided. All in all, the objective function was evaluated 113 times during the calibration process. The best set of parameter values is presented in table 7.3.

Figures 7.8 to 7.16 illustrate the improved explanatory power of the calibrated model. Three sections of route 4 were chosen for this illustration. Section 1 is from Acomb Library to Chaloner's Road; this part of the route is in a residential area, without significant congestion but with a high number of boarding passengers. Section 2 runs from Eason View to the Railway Station; along a major share of this section, the route goes through a bus lane where it has its own right of way, but towards the end of the section it join the general traffic entering the city centre. Section 3 lies between the Railway station and Clifford Street, and is entirely within York city centre.

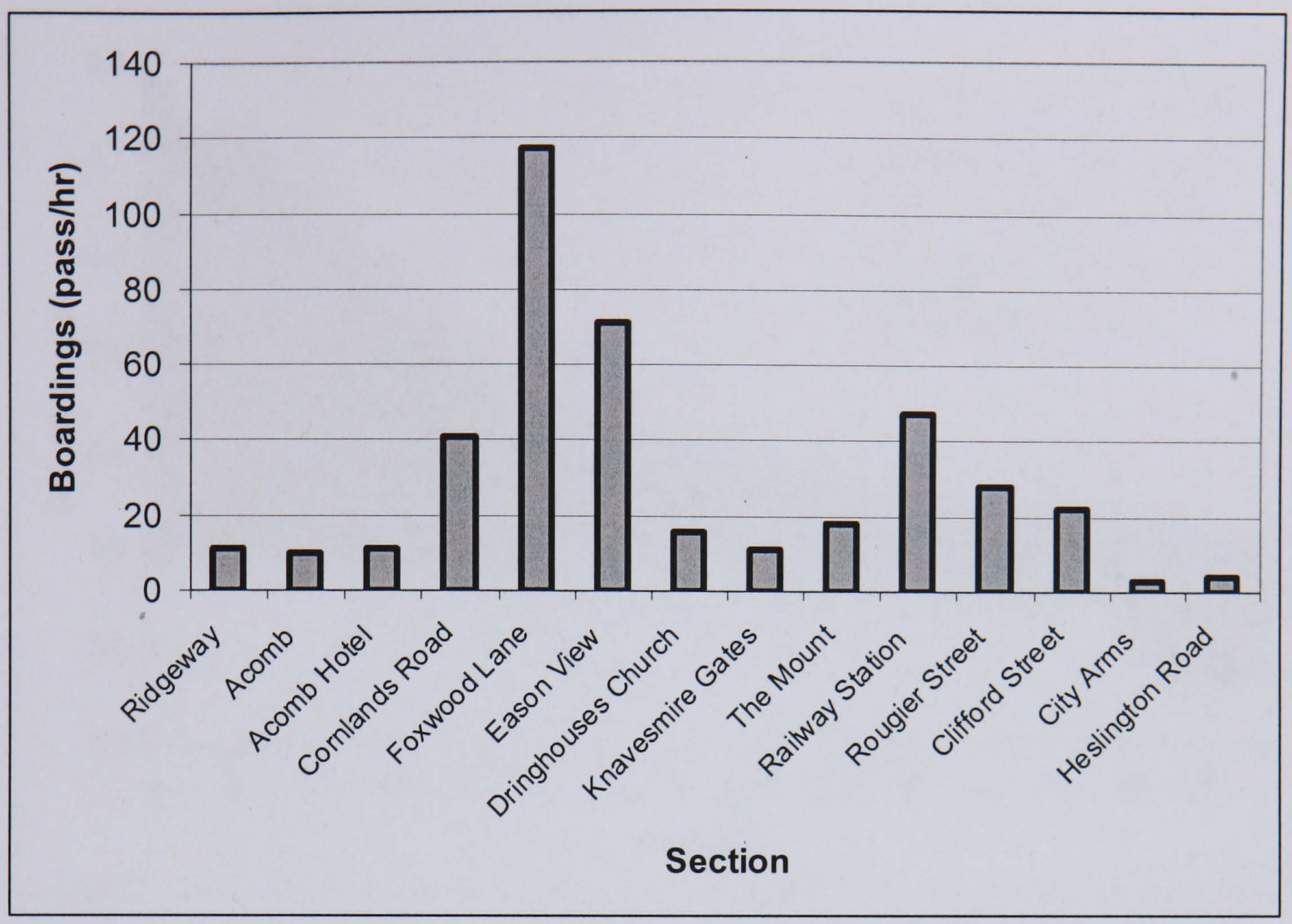


Figure 7.5: Original boarding profile

Figure 7.6: Modified boarding profile

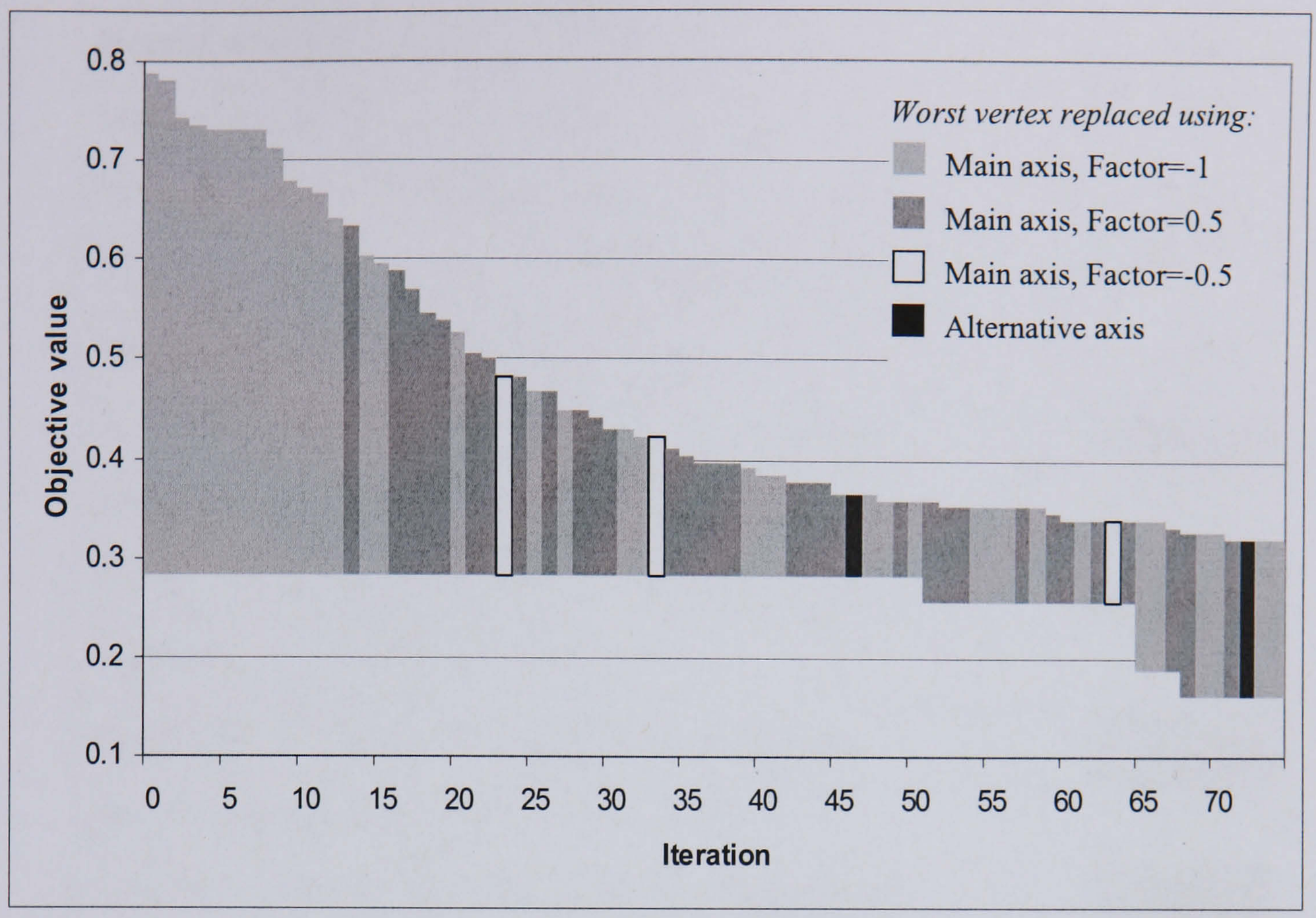


Figure 7.7: Progress of the calibration with real data

Figures 7.8 to 7.10, which compare the observed TTV with TTV simulated using the original parameters, show that before the calibration the model had a very limited predictive ability. In contrast, figures 7.11 to 7.13 demonstrate fairly good fit between the observed curves and those estimated using the calibrated model. This is confirmed by the validation results, in figures 7.14 to 7.16. The nine pairs of curves were compared using the K-S test; the test statistics are presented in table 7.4. The initial values indicate no fit at all between observed and simulated levels of TTV, whereas all values for the calibrated or validated model indicate plausible fit.

Note that although the discrepancies between the observed and simulated TTV curves are relatively small, the simulated distributions of travel times in figures 7.11 to 7.16 are always narrower than the observed curves. This means that despite the good fit, simulated estimates of TTV tend of slightly underestimate the real level of TTV. It seems that the TMM does not do well in reproducing the few days with extreme travel conditions. This is apparent especially on section 3, presumably because this section is in the congested city centre and is prone to such conditions.

Parameter	**Value**
Normal acceptable gap (seconds)	3.75
Minimum acceptable gap (seconds)	1.14
Time waited before accepting reduced gap (seconds)	32.02
Time waited before accepting minimum gap (seconds)	56.66
Mean of car normal acceleration (m/s^2)	3.00
Coefficient of variation of car normal acceleration	0.13
Mean of car maximum acceleration (m/s^2)	3.66
Coefficient of variation of car maximum acceleration	0.12
Mean of car normal deceleration (m/s^2)	3.35
Coefficient of variation of car normal deceleration	0.10
Mean of car maximum deceleration (m/s^2)	4.51
Coefficient of variation of car maximum deceleration	0.12
Mean of bus normal acceleration (m/s^2)	1.35
Coefficient of variation of bus normal acceleration	0.12
Mean of bus maximum acceleration (m/s^2)	1.43
Coefficient of variation of bus maximum acceleration	0.20
Mean of bus normal deceleration (m/s^2)	2.29
Coefficient of variation of bus normal deceleration	0.16
Mean of bus maximum deceleration (m/s^2)	2.15
Coefficient of variation of bus maximum deceleration	0.16
Demand fluctuation (coefficient of variation of overall demand)	0.06

Table 7.3: Parameter values after calibration

	Section 1	Section 2	Section 3
Before calibration	0.821	0.590	0.564
After calibration	0.128	0.162	0.179
Validation	0.231	0.179	0.197

Table 7.4: Testing the calibrated model using K-S statistic

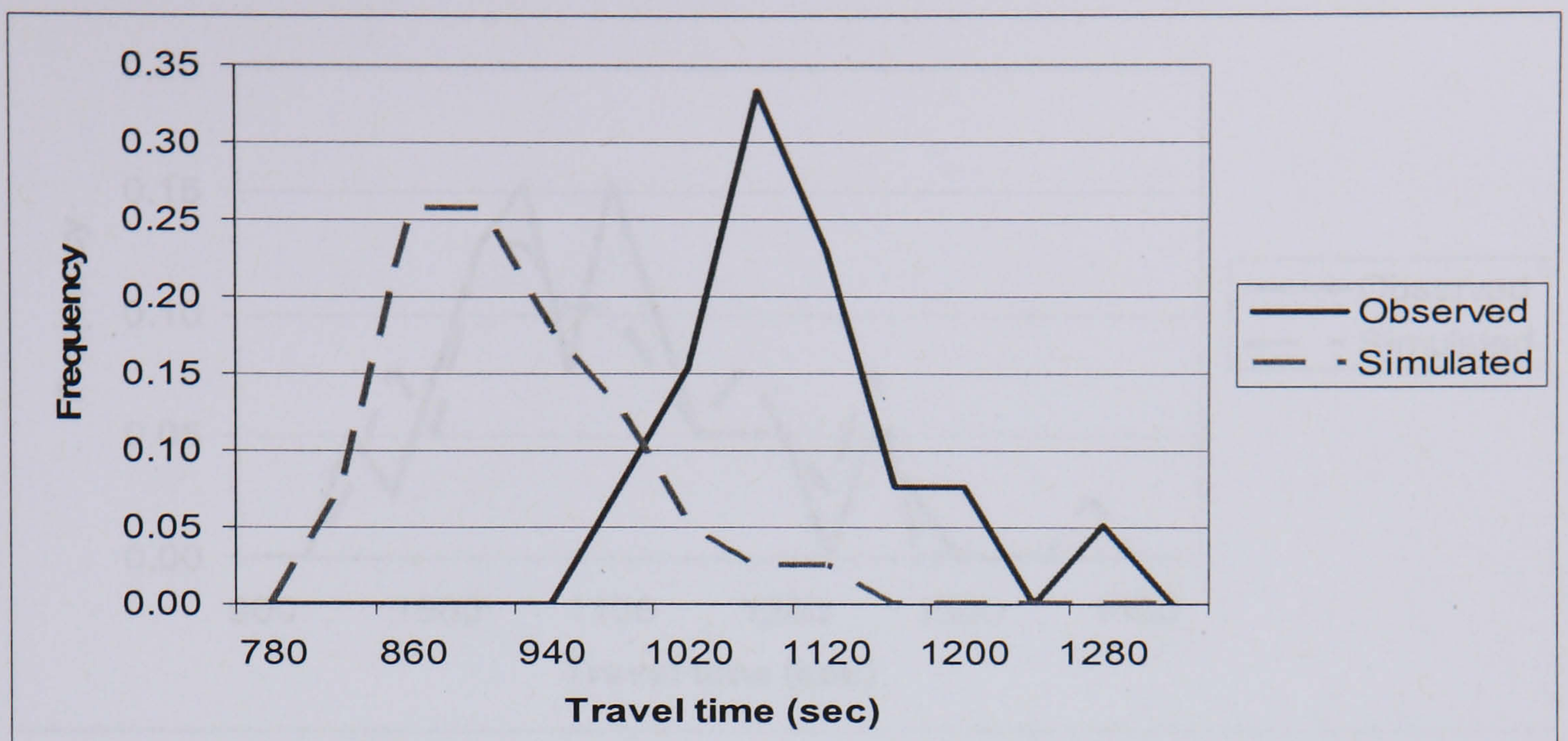


Figure 7.8: TTV on section 1 before calibration

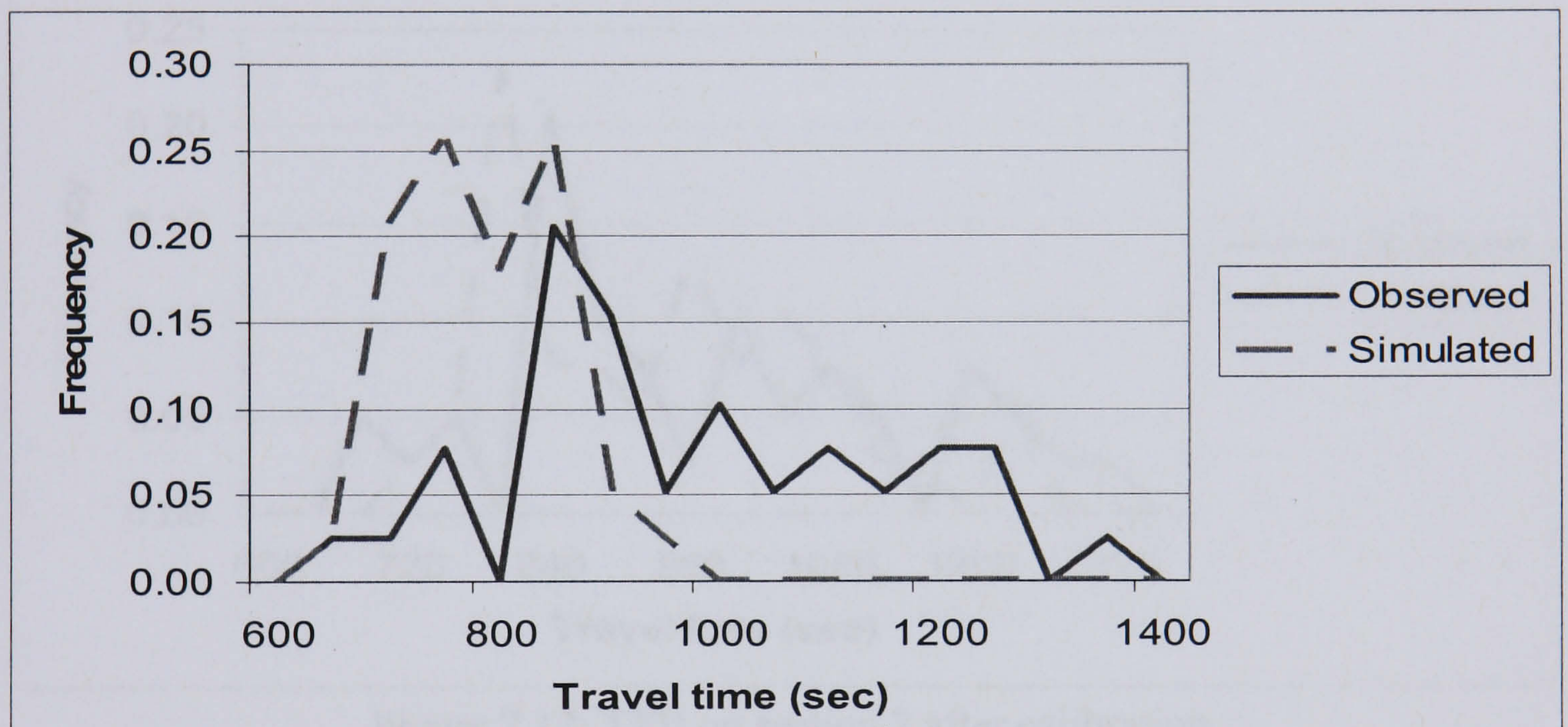


Figure 7.9: TTV on section 2 before calibration

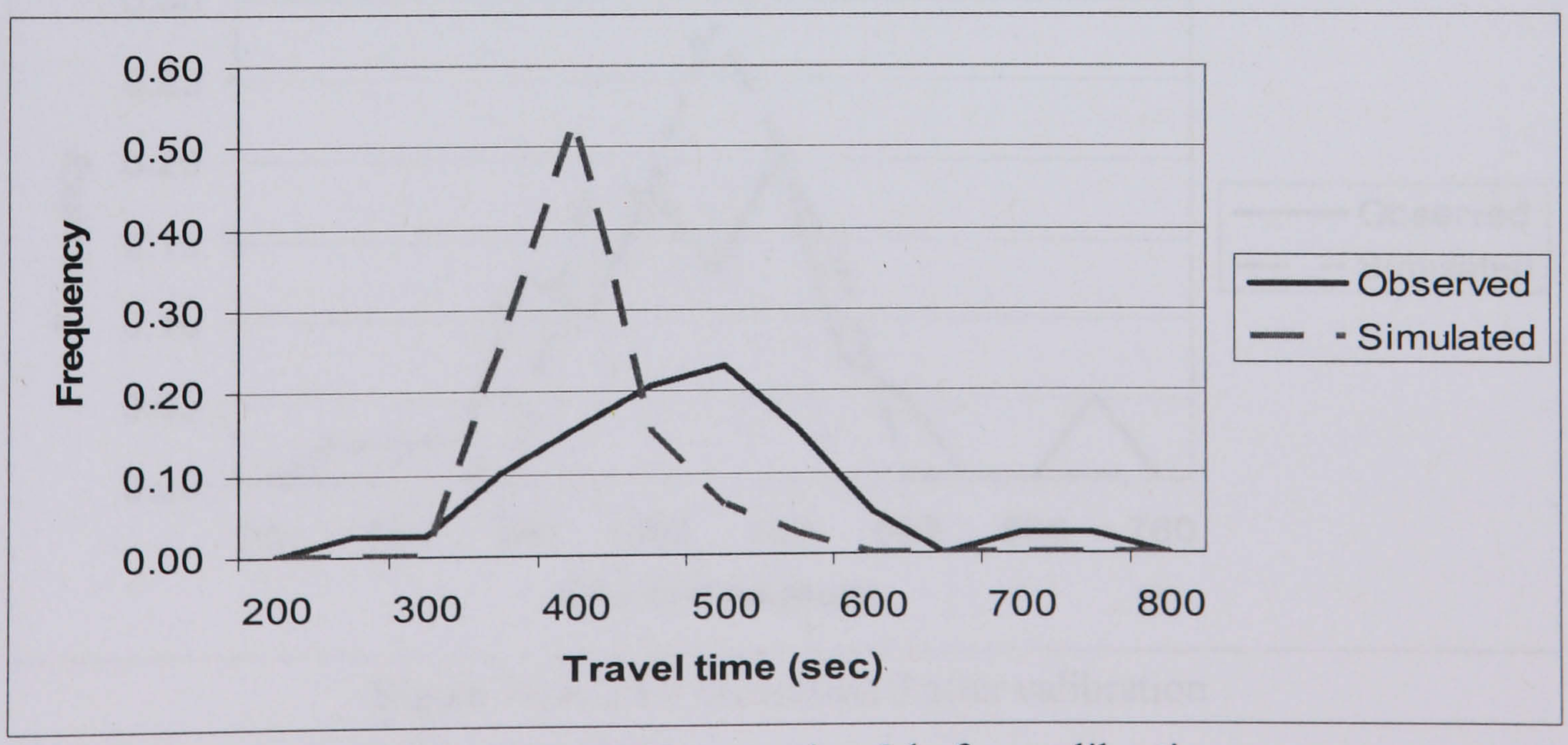


Figure 7.10: TTV on section 3 before calibration

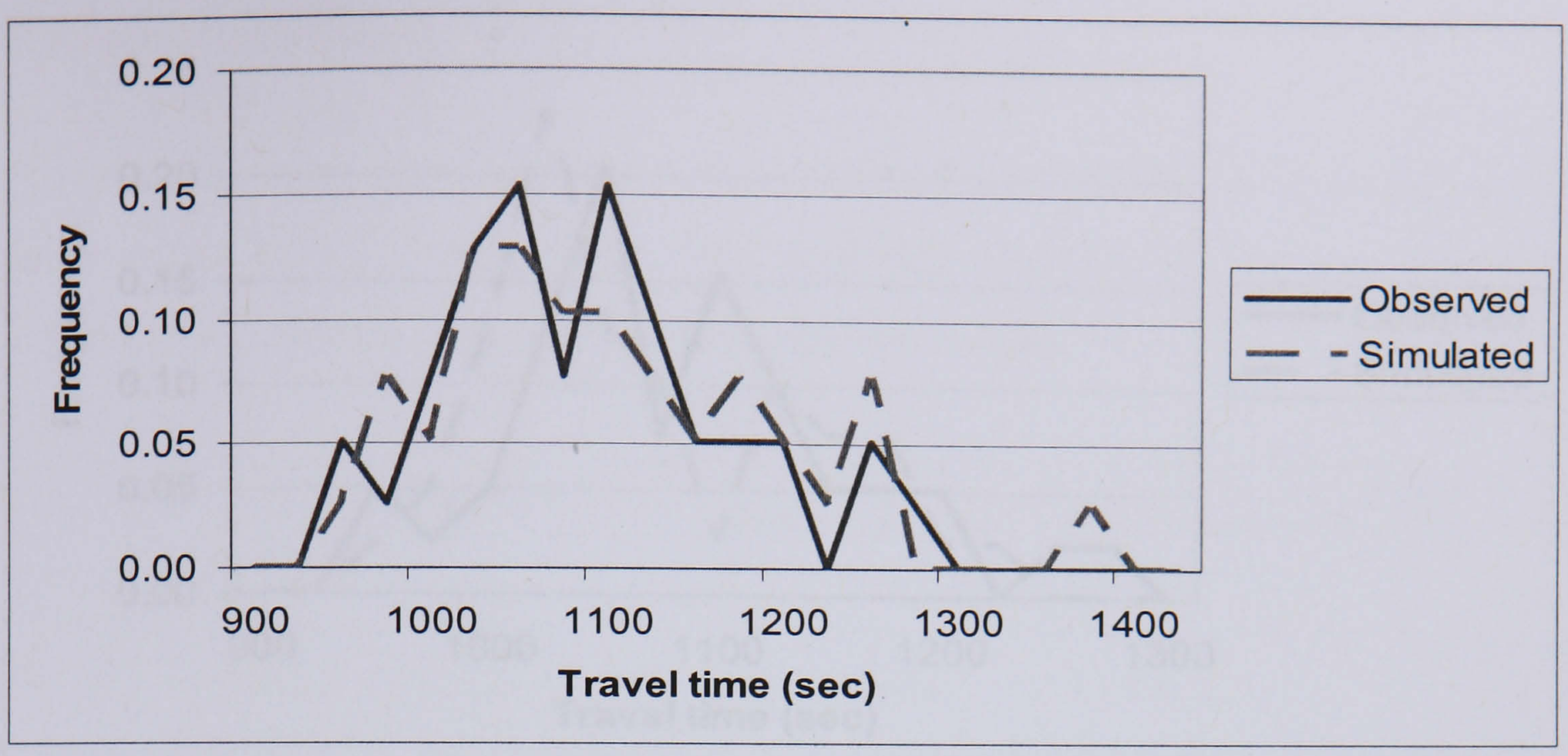


Figure 7.11: TTV on section 1 after calibration

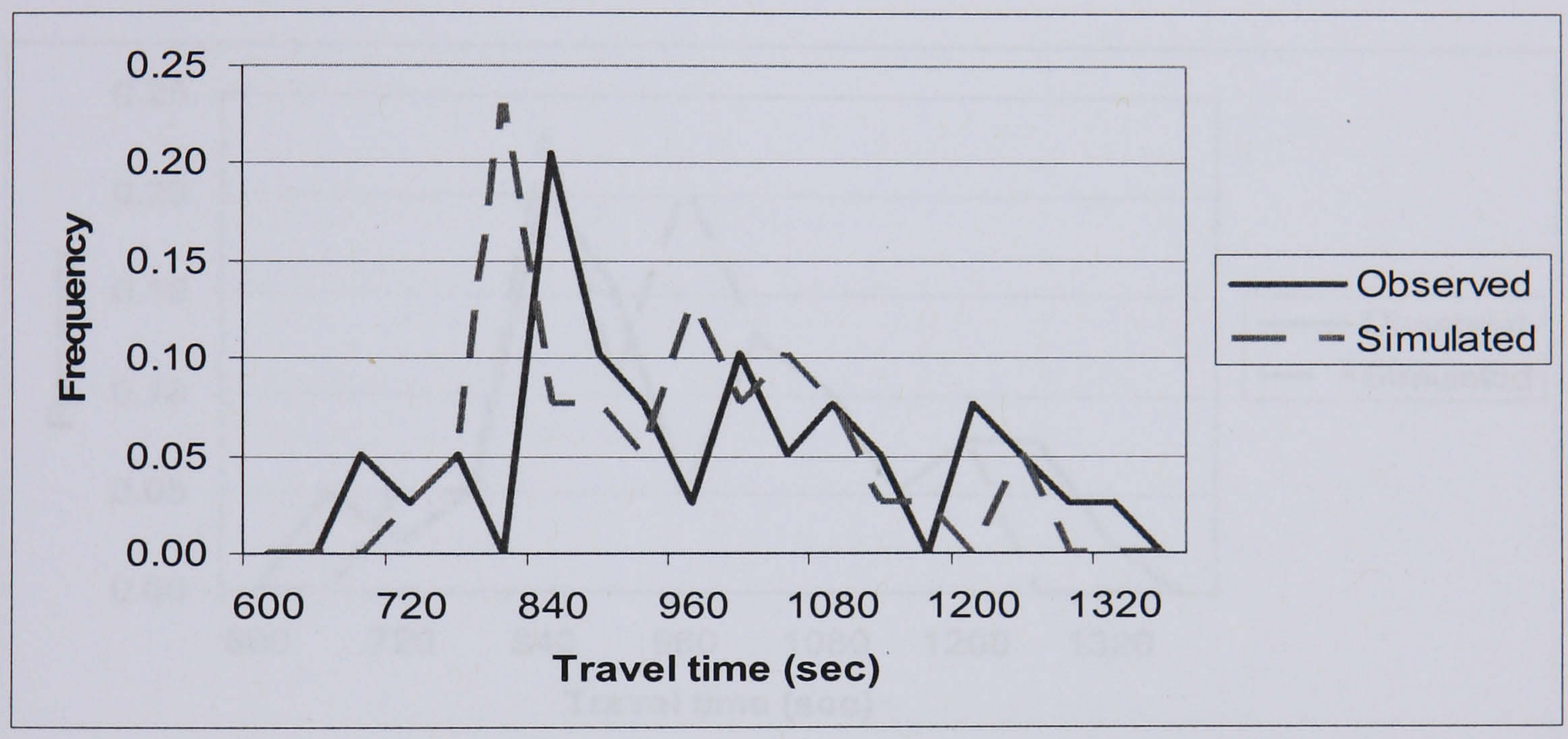


Figure 7.12: TTV on section 2 after calibration

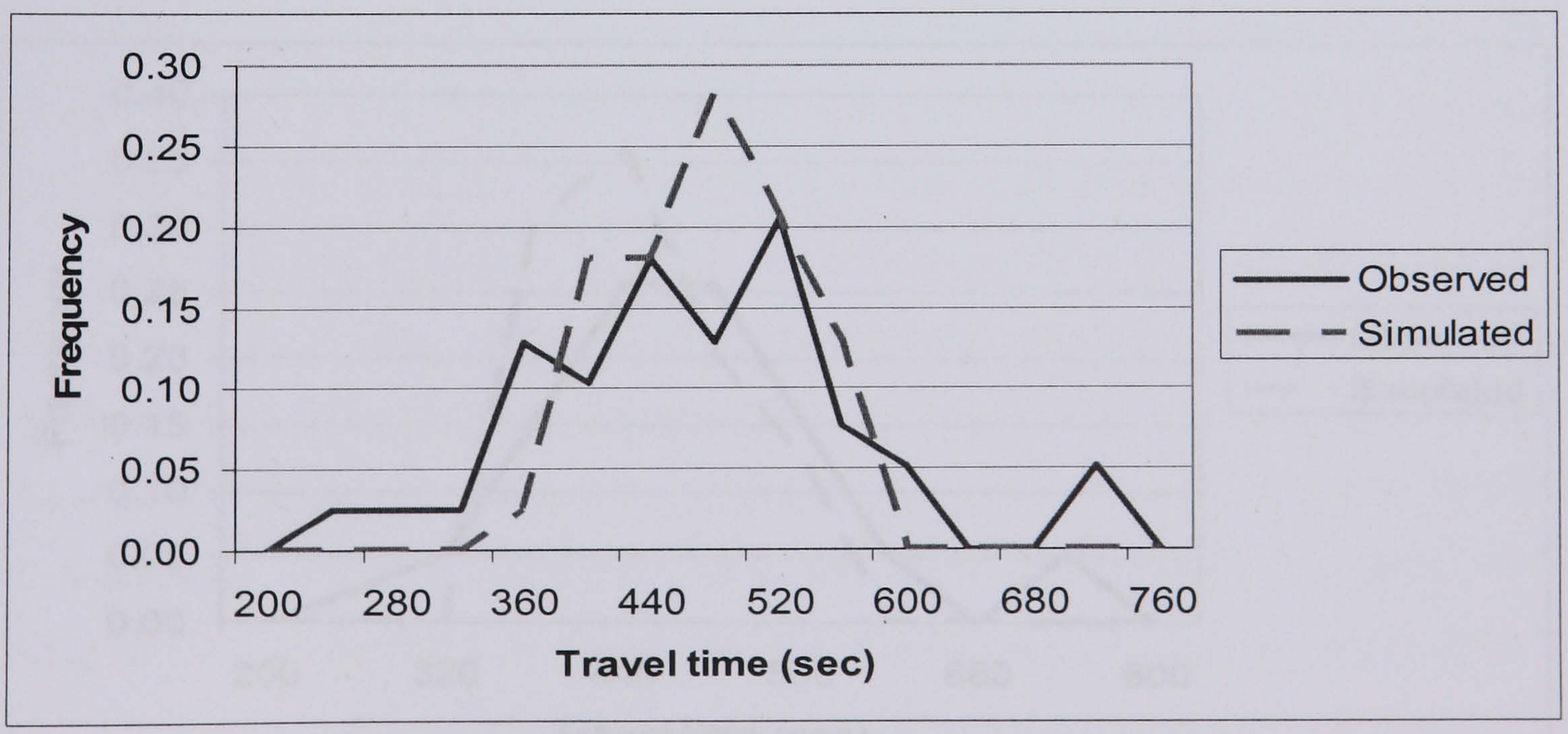


Figure 7.13: TTV on section 3 after calibration

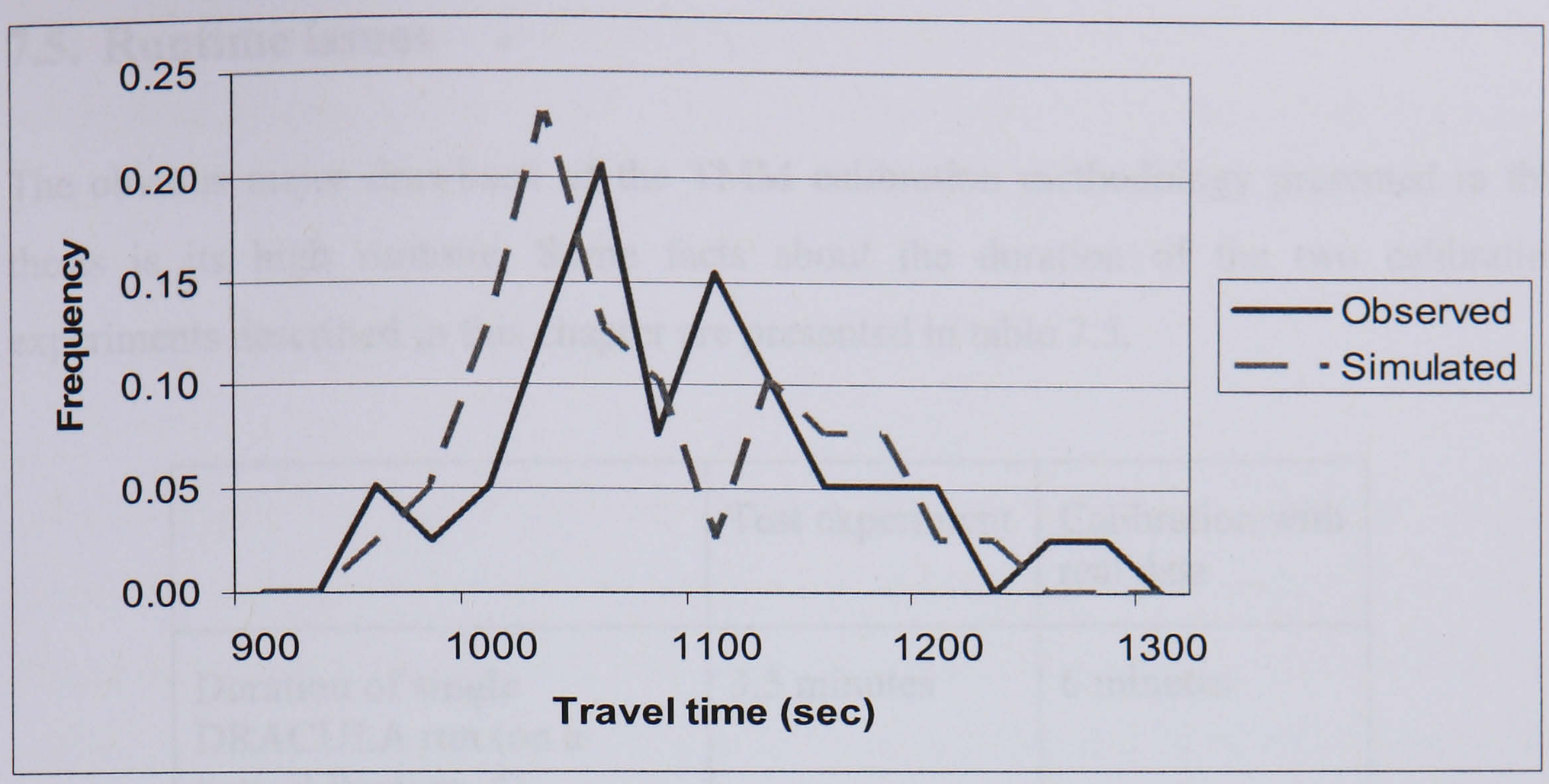


Figure 7.14: TTV on section 1 - validation

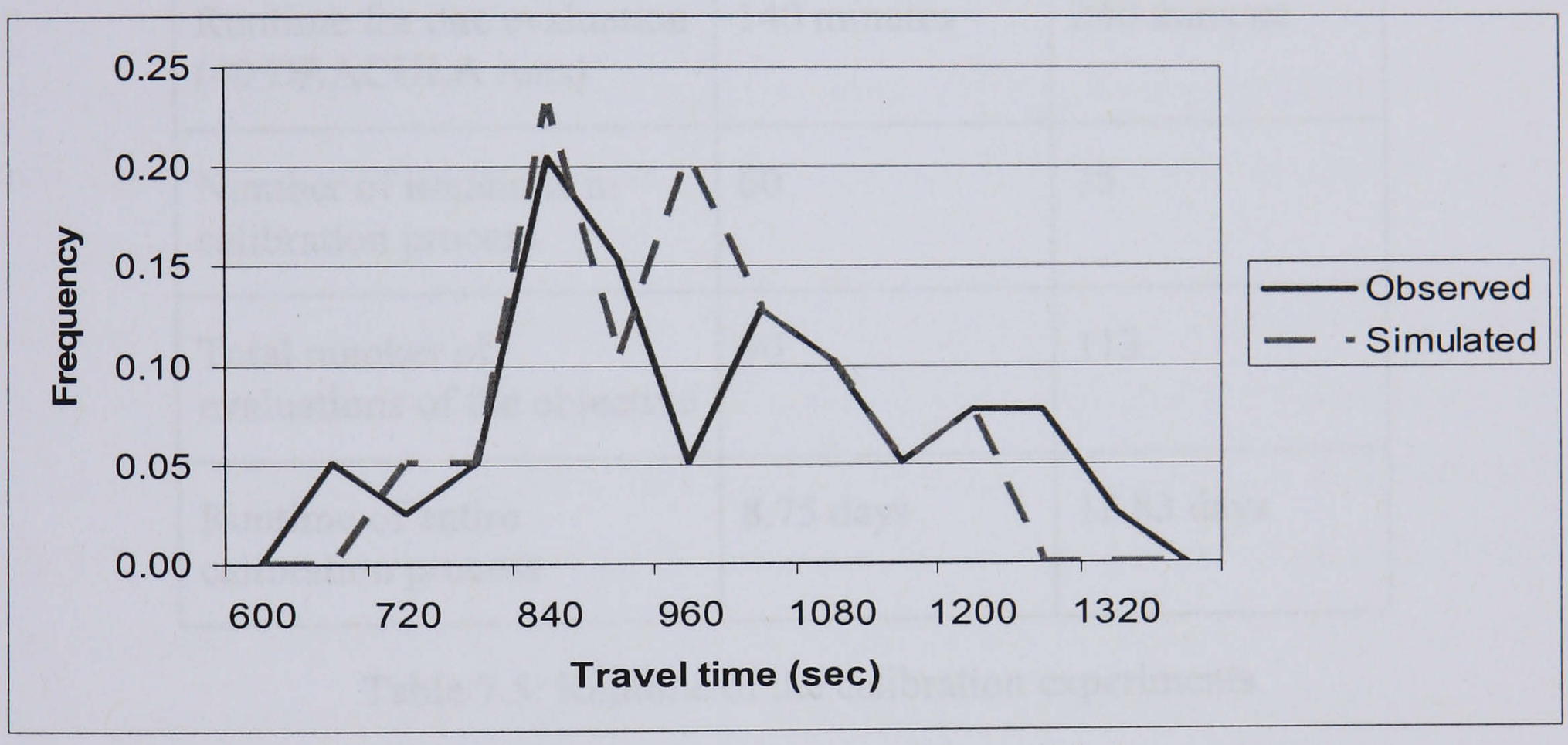


Figure 7.15: TTV on section 2 - validation

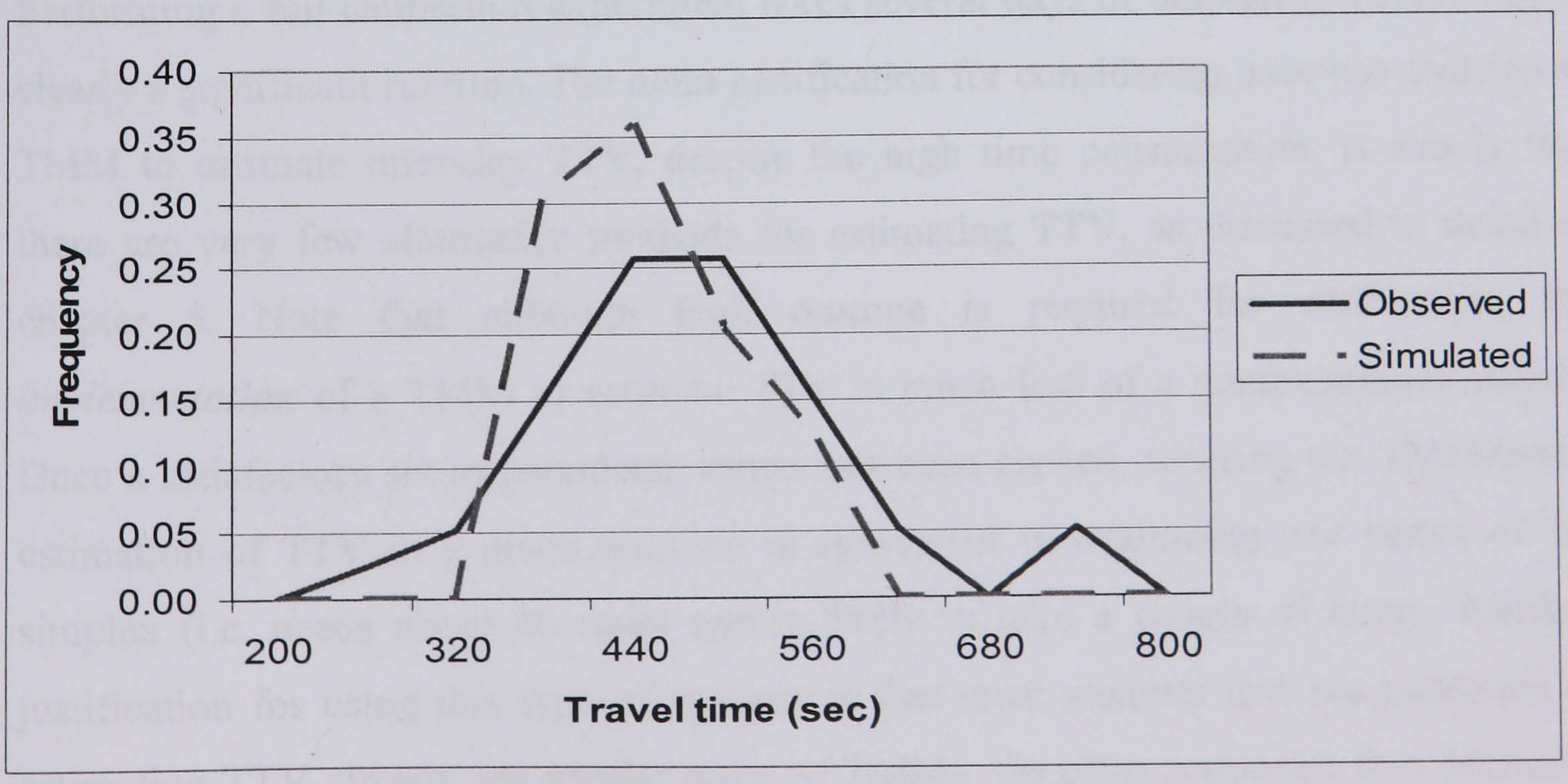


Figure 7.16: TTV on section 3 - validation

7.5. Runtime issues

The obvious major drawback of the TMM calibration methodology presented in this thesis is its high runtime. Some facts about the duration of the two calibration experiments described in this chapter are presented in table 7.5.

	Test experiment	Calibration with real data
Duration of single DRACULA run (on a typical Pentium 4)	3.5 minutes	6 minutes
Runtime for one evaluation (40 DRACULA runs)	140 minutes	240 minutes
Number of iterations in calibration process	60	75
Total number of evaluations of the objective	90	113
Runtime of entire calibration process	8.75 days	18.83 days

Table 7.5: Runtime of the calibration experiments

Performing a full calibration experiment takes several days of uninterrupted run; this is clearly a significant runtime. The main justification for considering inter-run analysis of TMM to estimate inter-day TTV, despite the high time consumption, is simply that there are very few alternative methods for estimating TTV, as discussed in detail in chapter 5. Note that although high runtime is required for *calibration*, the *implementation* of a TMM to estimate TTV is much less of a computational burden. Once a satisfactory set of parameter values has been chosen, carrying out TMM-based estimation of TTV in a given scenario is equivalent to evaluating one vertex of the simplex (i.e. needs about 40 runs) and is likely to take a couple of hours. Another justification for using this type of analysis is that most analysts that are interested in estimating TTV already are regular users of TMMs, for other purposes; they normally keep and maintain TMM networks and other required input data. The relatively lengthy

calibration procedure described here might be actually considered a much quicker way for them to develop a tool for estimating TTV, compared to the alternative of creating new tools from scratch.

7.6. Conclusions

In this chapter we tested and then implemented the calibration algorithm and the modelling concept developed in chapter 6. Calibration was performed in three imaginary scenarios and in a real network with data from the city of York. Although the statistical fit between the outputs of the calibrated models and observed measurements was not perfect, the calibrated models did provide very reasonable TTV estimates. We found that the simulation-based estimates of TTV, after the calibration, were slightly underestimated, especially on a congested section of the analysed bus route. Still, in the absence of other tools for generating forecasts of TTV, we see the results as most satisfactory.

During the test of the calibration procedure, some of the calibration parameters performed more rationally than others. It seems, for instance, that the extent of bus TTV is positively influenced by the coefficients of variation of bus normal acceleration and deceleration, and from the level of fluctuation in the overall demand. However, it should be reminded that the scope of the current study did not enable us to choose the set of calibration parameters systematically. A study of the sensitivity of TTV estimates to the level of different parameters would constitute an important extension of the current work. Implementing the calibration methodology described in this thesis with a set of parameters that has been chosen more carefully would enable a more focused study and would presumably also result in a model with a better explanatory power.

It is also worth mentioning that there is always need to improve the TMM itself, since any model is only a simplified version of the real transport system. Some features that apparently affect the level of TTV, such as accidents, various incidents and their consequences, are currently not modelled in DRACULA; incorporating them in the model will inevitably result in a more powerful tool for estimating TTV.

Undoubtedly, the main weakness of the concept demonstrated here is the high runtime required for full calibration of a TMM. It is hoped that improvement in computing power will significantly reduce the time consumption of the proposed algorithm in the

future. Nonetheless, we remind that only the calibration takes significant time; application of the calibrated model does not require many hours of running, and it has the advantage of relying on an existing software tool rather than having to develop a new one.

Through the calibration experiments we also examined a modification of the well-known simplex method. The modification is suitable for problems such as the one tackled here, where the evaluation of the objective function is time consuming due to high dimensionality and nondifferentiability. It was found that in stages of the minimisation process where the original simplex method requires very lengthy re-initialisation, the modified approach avoids this and moves on much faster by searching for a solution along an alternative geometrical axis.

We deem the model calibrated here for York ready for application. In the next chapter this model is combined with the economic findings of chapter 4, for evaluating the cost associated with TTV in a hypothetical scenario of bus infrastructure investment.

Chapter 8

Combining the demand and supply tools

8.1. Introduction

Although there clearly is much scope for extending and refining our econometric analysis (chapters 2 to 4) and traffic analysis (chapters 5 to 7), we have addressed two of the objectives specified in the introduction to this thesis: we have estimates of the willingness of bus users to pay for reduction in the level of TTV, and we have a tool that can be used for estimating the level of TTV in hypothetical scenarios. The third objective specified in chapter 1 is to illustrate how the outputs from these two separate sets of experiments can be used together in an attempt to determine the benefit from reducing TTV in a practical setting; we attempt to do this in this chapter. The next section formulates methodology for the joint application of our demand and supply tools. The following section describes a simple case study in which the methodology is applied, and presents a full set of results of this application. Some conclusions are presented in the last section of the chapter.

8.2. Methodology

We have seen that TTV affects the way travellers choose their departure time on their commuting trip. We introduce here an iterative procedure, in which travellers repeatedly choose departure times and travel conditions are repeatedly estimated based on their choices, till a stable choice pattern is reached. The travel conditions are estimated using the calibrated traffic microsimulation model (TMM) described in the previous chapter, with the same concept that sees a single run of the TMM as representing a single day in reality (it should be borne in mind that we still follow the definition of TTV described in chapter 1, i.e. we consider unpredictable variation between days but not systematic variation). The choices of the travellers follow the behavioural model developed earlier in the thesis, which ascribes penalties to the mean travel time, earliness and lateness. The gradual improvement of the choice estimates from one iteration to another is based on cost minimisation: the individual cost of travel, based on our econometric model, is calculated for each traveller, and his/her current choice is replaced with a different one

if it offers a lower cost. The total travel cost, summarised across all travellers, can be used in an appraisal framework to compare between different investment scenarios.

A key principle in the methodology is the redistribution of departure time choices (DTCs) at each iteration. Bus journey times cause some travellers to change their departure times, but these changes affect bus journey times, and the process of determining DTC therefore needs to consider this loop. In the suggested estimation procedure, travellers repeatedly shift to earlier or later departures, in order to reduce their travel costs. The overall number of passengers wishing to board each departure of the analysed route is re-calculated at each iteration; the updated passenger demand per departure is fed into the next iteration and hence affects journey times and their variability. The process is repeated till the change in the distribution of DTCs between successive iterations is relatively small.

The methodology outline is presented as a flow chart in figure 8.1. The principles of the methodology and the calculations involved are described in the following paragraphs. It is assumed that the methodology is to be applied for a single bus route, as we do in this chapter, although only minor changes would be required to amend it for the needs of a more general case.

Notation. The following notation is used:

I	size of a given sample of users of the analysed bus route
i	index for an individual user
N	number of runs of the TMM, each run represents a day
n	index for a single run or day
D_i	desired arrived time of user i to his/her destination
B_i	boarding stop of user i
A_i	alighting stop of user i
VOT_i	value of the mean travel time of user i
VOE_i	value of mean earliness of user i
VOL_i	value of mean lateness of user i
j	index for a particular bus journey (with a particular departure time)
j_i^s	index for the s^{th} bus journey to depart after (or before, if $s<0$) the departure chosen by user i
j_i^0	index for the bus journey chosen by user i
j^F	index for the first bus journey during the analysis period

j^L index for the last bus journey during the analysis period

$T_{j,p,n}$ clock time when bus j arrives at stop p on day n

MTT_i^j mean travel time for user i when bus j is used

ME_i^j mean earliness for user i when bus j is used

ML_i^j mean lateness for user i when bus j is used

COT_i^j cost of mean travel time for user i when bus j is used

COE_i^j cost of mean earliness for user i when bus j is used

COL_i^j cost of mean lateness for user i when bus j is used

C_i^j total cost for user i when bus j is used

R_j the proportion of all users that chooses bus departure j

δ_i^j a dummy variable that equals 1 if user i chooses bus departure j, and 0 otherwise

Inputs. The input for the process includes the following:

1. Specification of the analysed network and time period. The network inputs include the configuration of the street network, the bus timetables and the general trip demand during the analysed period.
2. Characteristics of the user demand for the discussed route, given as a sample of I users. The inputs include features of the journey made by each user, namely D_i, B_i and A_i.
3. The products of an economic model that describes the individual willingness of each of the I users to pay for an improved journey: VOT_i, VOE_i and VOL_i.
4. The products of a traffic model that can be used to obtain estimates of travel times and their variance.

Simulation outputs and initial guess of departure time choice. At each stage of the algorithm described in figure 8.1, the TMM is run *N* times and a set of travel times is derived. The travel times obtained from the series of runs have the form of clock times. The main loop performed in the algorithm is used to iteratively determine the departure time chosen, such that each stage tries to improve the estimates obtained in the preceding stage. In order to start this loop, some initial guess of the pattern of DTCs is required. We make this first guess of the journey chosen by each bus user by assuming that he/she chooses the departure that minimises his/her mean travel time. That is, the initial individual j (namely the initial j_i^0) is chosen for each i as follows:

(8.1)

$$j_i^{\,0} = \arg\min MTT_i^{\,j} = \arg\min \frac{1}{N}\sum_{n=1}^{N}(T_{j,Ai,n} - T_{j,Bi,n}) \qquad j^F \le j \le j^L$$

Calculating individual mean travel time, earliness and lateness. At each stage of the algorithm, individual journey attributes are calculated using the following formulas:

(8.2)

$$MTT_i^{\,j} = \frac{1}{N}\sum_{n=1}^{N}(T_{j,Ai,n} - T_{j,Bi,n})$$

$$ME_i^{\,j} = \frac{1}{N}\sum_{n=1}^{N}\max\left(0, \quad D_i - T_{j,Ai,n}\right)$$

$$ML_i^{\,j} = \frac{1}{N}\sum_{n=1}^{N}\max\left(0, \quad T_{j,Ai,n} - D_i\right)$$

The same formulas are used both with $j = j_i^{\,0}$, namely for calculating the attributes that correspond to the currently chosen journey, and with $j = j_i^{\,s}$ $(s \neq 0)$, namely for calculating the attributes that correspond to other alternative departures.

Calculating individual costs. The individual costs that correspond to each possible departure are calculated using the following formulas:

(8.3)

$$COT_i^{\,j} = MTT_i^{\,j} \cdot VOT_i$$

$$COE_i^{\,j} = ME_i^{\,j} \cdot VOE_i$$

$$COL_i^{\,j} = ML_i^{\,j} \cdot VOL_i$$

Checking alternatives and updating departure time choices. At each iteration of the algorithm, the total cost of the bus journey currently chosen by each passenger is compared to the total costs of alternative journeys that depart earlier or later. The procedure does not allow sudden radical changes in the DTC, as these do not seem to

represent real behaviour. Thus, the procedure restricts the range of alternative departures examined at each stage to those that are no more than two departures earlier or later than the currently chosen one. The process does enable a greater change of the DTC, but only as a gradual change, a small step at a time. The individual choice of a new j at each iteration can be formulated as follows:

$$j_i^0 = \arg\min C_i^j = \arg\min \left\{ COT_i^j + COE_i^j + COL_i^j \right\} \tag{8.4}$$

$$j \in \{ j_i^{-2}, j_i^{-1}, j_i^0, j_i^{+1}, j_i^{+2} \}, \qquad j^F \le j \le j^L$$

Updating user demand per departure. Once new DTCs have been made, the new demand per departure is calculated as follows:

$$R_j = \frac{1}{I} \cdot \sum_{i=1}^{I} \delta_i^j, \quad (\delta_i^j = 1 \ \ if \ \ j = j_i^0, \ \ 0 \ otherwise) \tag{8.5}$$

It should be noted that the total number of passenger, *I*, is not fed into the TMM runs at any point. The number of passengers that use the analysed route in the TMM is specified separately, based on passenger counts at bus stops, and is independent of the sample of *I* passengers discussed here. In our algorithm, the choices of the *I* bus users change only the proportion of the overall demand that uses each particular departure, not the overall level of demand. Namely, *I* is not the number of passengers but merely the size of a sample that should be big enough to account for the diversity of levels of WTP among travellers. Changing the value of *I* will affect the level of accuracy of the analysis but not affect the level of demand.

Convergence test. The natural indication that the process should be terminated would be when all (or most) travellers cannot reduce their cost of travel by changing their DTC. However, in some test runs of the algorithm it was found that such convergence does not occur, since the choices of many travellers enter an endless loop. For some travellers it seems that the costs can fall if they shift to another departure, but in the next iteration, once several passengers have switched to the same departure, there is apparently no improvement and they return to the departure chosen previously. In the test runs some travellers repeated the same unsuccessful attempt to reduce the cost again

and again, and hence, convergence was not reached. This clearly does not represent realistic behaviour, and therefore a very simple constraint was introduced to the algorithm, to account for a more logical learning mechanism. The constraint determines that if a shift to a specific departure has been tried and did not lead to reduced cost, it cannot be tried again. Once this constraint was added, convergence of the cost estimates for about 90% of travellers has been finally achieved (i.e. for 90% of travellers it was not possible to reduce travel cost by changing DTC). It was decided that for the needs of the example presented here, trying to pursue convergence for the remaining 10% is not of major importance. Note that the fact that we seek convergence of individual costs rather than total costs implies that we deal with a user-optimal (as opposed to system-optimal) problem.

The methodology has been devised for the purpose of comparing the costs of MTT and TTV between different scenarios of investment in bus infrastructure. If the methodology is applied to different network configurations, and convergence is reached, then the difference between the final cost estimates can constitute an important appraisal feature. The fact that the costs associated with TTV are explicitly accounted for, together with the DTC mechanism that determines these costs, is the main innovative element in this methodology. However, this is done here in an illustrative manner; the methodology is obviously limited in the sense that other effects of TTV, apart from the effect on DTC, are not considered. The main justification for this source of simplification is that, as we discussed in chapter 2, DTC seems to be the most direct effect of TTV. Changes in mode choice, route choice and other secondary responses to the level of TTV would need to be captured in a broad, multimodal and fully-elastic model which is beyond the scope of this thesis.

The presented methodology clearly involves some other compromises. These are discussed in more detail towards the end of this chapter.

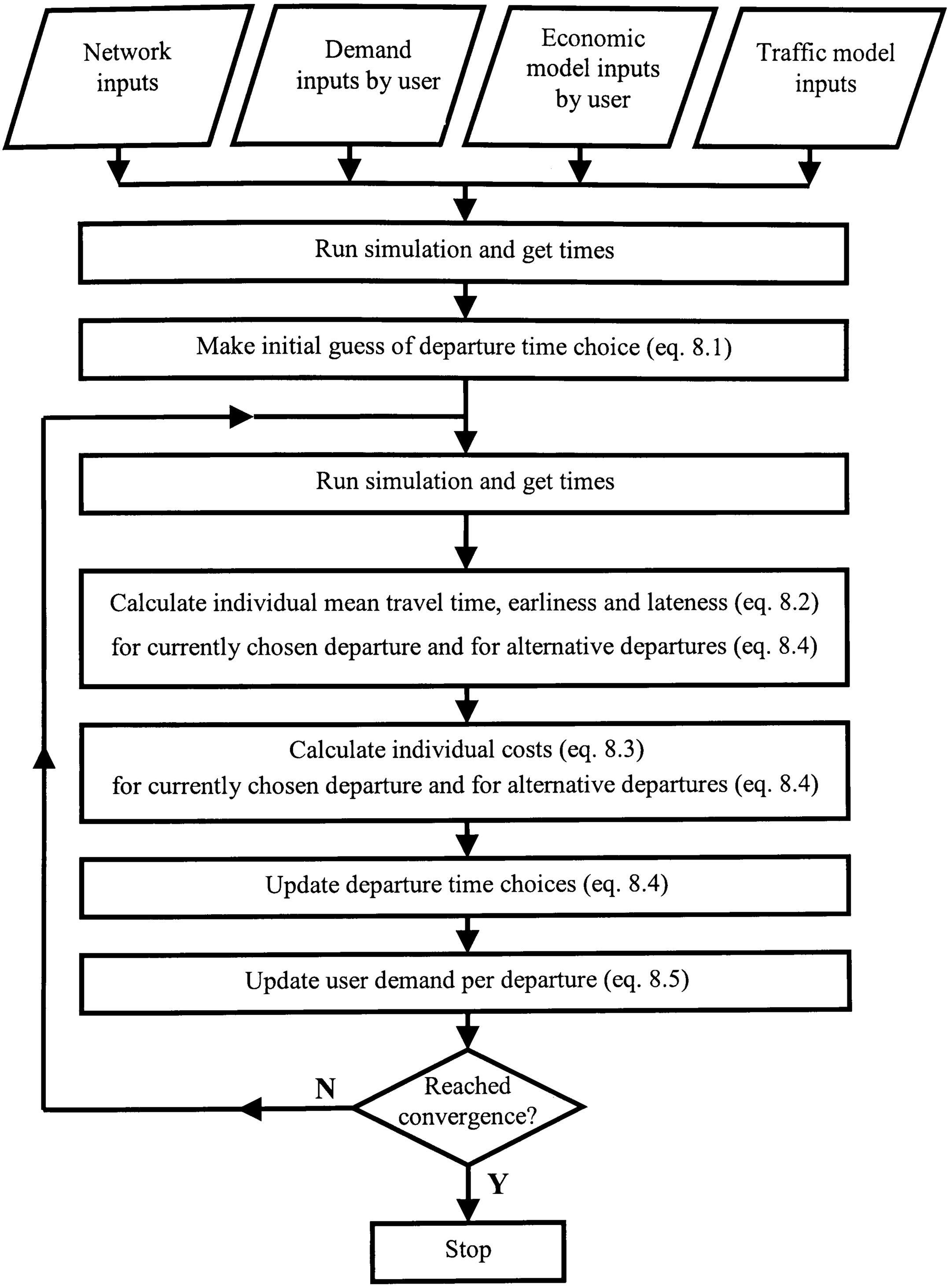


Figure 8.1: Procedure for estimating DTC and travel cost

8.3. Demonstration of a scenario analysis

8.3.1. Inputs

This section demonstrates an application of the procedure presented above. We examine a simple case, where an extension of an existing bus lane is considered; we wish to assess the potential benefit from such extension, in general, and the benefit from reduced TTV, in particular. This is therefore a comparison between a "basic" scenario and an "extended" scenario. The possible extension of the bus lane does not include introduction of any other measures, such as priority in intersections and so on. We carry out this analysis in the same geographical area that all previous parts of the thesis have focused on, namely the city of York. This makes our econometric and traffic models apt to be used here. The bus lane extension is considered along the itinerary of the same bus route that has been previously discussed, namely route 4. This is a major commuting route, that operates at a high frequency (the headway during the morning peak is 8 minutes), and crosses the city from West to East via the city centre. We only examine one direction of the route here, and only examine its performance during the morning peak period.

Figure 8.2 shows the location of the basic and the extended bus lanes on a map of York and of route 4. A bus lane already exists between points A and B; this bus lane serves buses on their way to the city centre, but it does not stretch into the centre itself. We now consider extending the bus lane to point C, situated next to the York Railway Station, closer to the heart of the city centre. The length of the proposed new bus lane section is about 500 meters; it is suggested to construct it adjacent to, and not instead, the existing all-traffic lanes. It should be emphasized that this is only a conceptual scheme; we have not examined whether the physical right of way that would be needed for a bus lane along this section is actually available.

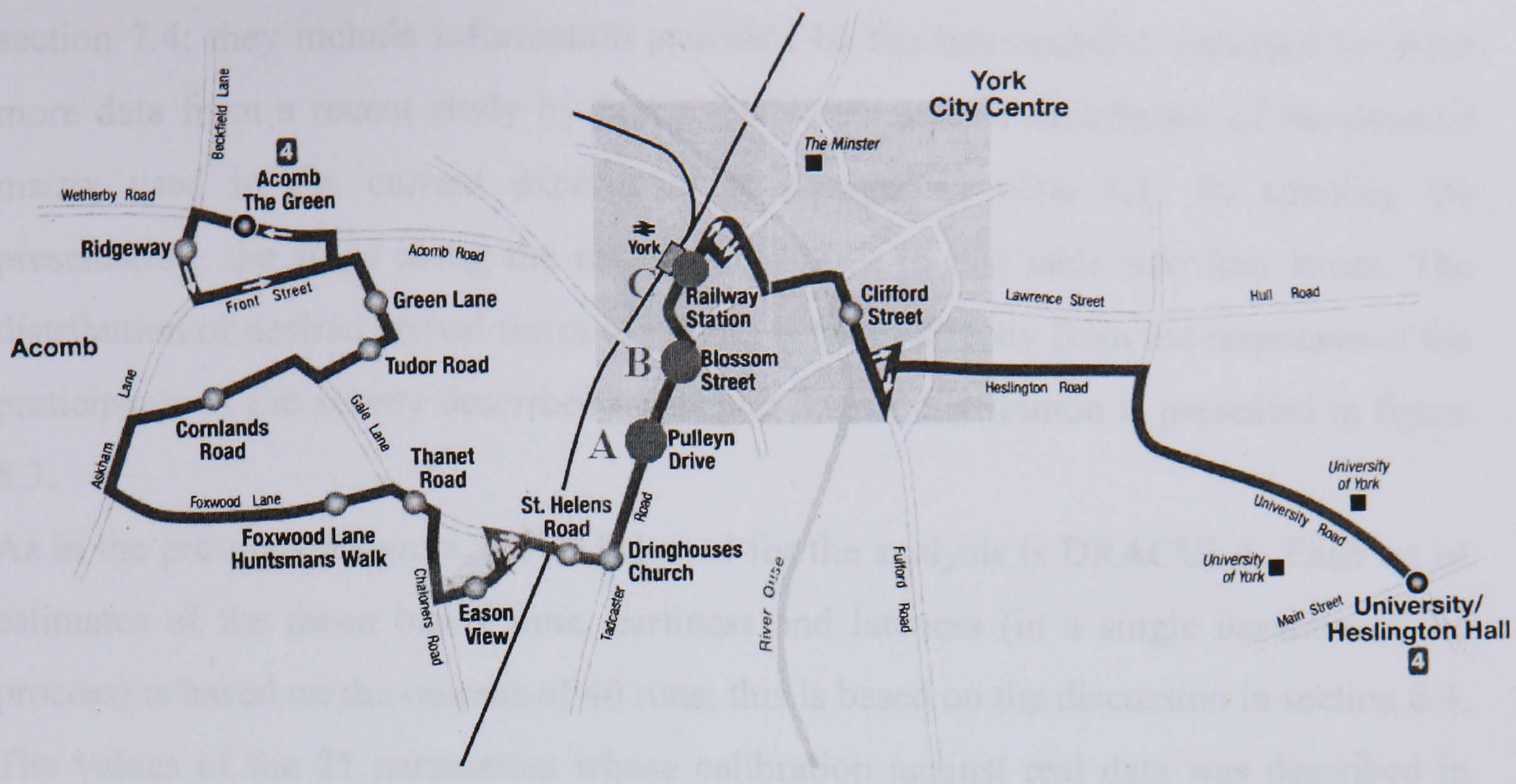


Figure 8.2: Location of the basic and extended bus lanes

As input for the scenario analysis we generate a random sample of 1000 users that make their commuting trip on route 4 in the morning peak period. The actual number of passengers on this route in the morning peak (around 440 per hour in total) was presented in section 7.4, and is not equal to the sample size used here. The current analysis examines the choices made by passengers as a proportion of the entire sample and then applies these proportions to the actual demand; therefore the sample size was mainly determined so that it can reflect a large enough variety of preferences and origin-destination patterns. Each of the 1000 users in the sample has an individual monetary value for the mean travel time, the earliness and lateness. We assume here that the distributions of these WTP elements are the same as the distributions determined earlier in the thesis, in table 4.6 and in figures 4.12, 4.13 and 4.14. These distributions were obtained by fitting *smoothed curves* to the results of a *selective sub-sampling* experiment, as described in detail in section 4.6. We remind the notion made in section 4.6, that these distributions are not rigorously consistent with a full model of departure time choice. Due to the difficulties we faced when trying to reach a model from which the entire distribution of the WTP can be identified, we based our best-practice estimates on a compromised technique; these estimates are not likely to include the extreme values of people with a very high or very low WTP, but still, they are most likely to truthfully capture the preferences of the majority of the population.

Each user also has individual boarding stop, alighting stop and desired arrival time at the alighting stop. The sources for data on boarding and alighting were described in

section 7.4; they include information provided by the bus operator, enriched by some more data from a recent study by Sinha (2004). A general description of the demand matrix used in the current experiment is brought in table 8.1. To simplify the presentation, the stops along the route are grouped in this table into four zones. The distribution of desired arrival times used here is taken directly from the responses of the participants of the survey described in chapter 3; this distribution is presented in figure 8.3.

As in the previous chapters, the TMM used for the analysis is DRACULA. Each set of estimates of the mean travel time, earliness and lateness (in a single iteration of the process) is based on the outputs of 40 runs; this is based on the discussion in section 6.4.

The values of the 21 parameters whose calibration against real data was described in section 7.4 are fixed at the levels presented in table 7.3.

The values of the main attributes we use as input for this demonstration of the methodology are summarised in table 8.2.

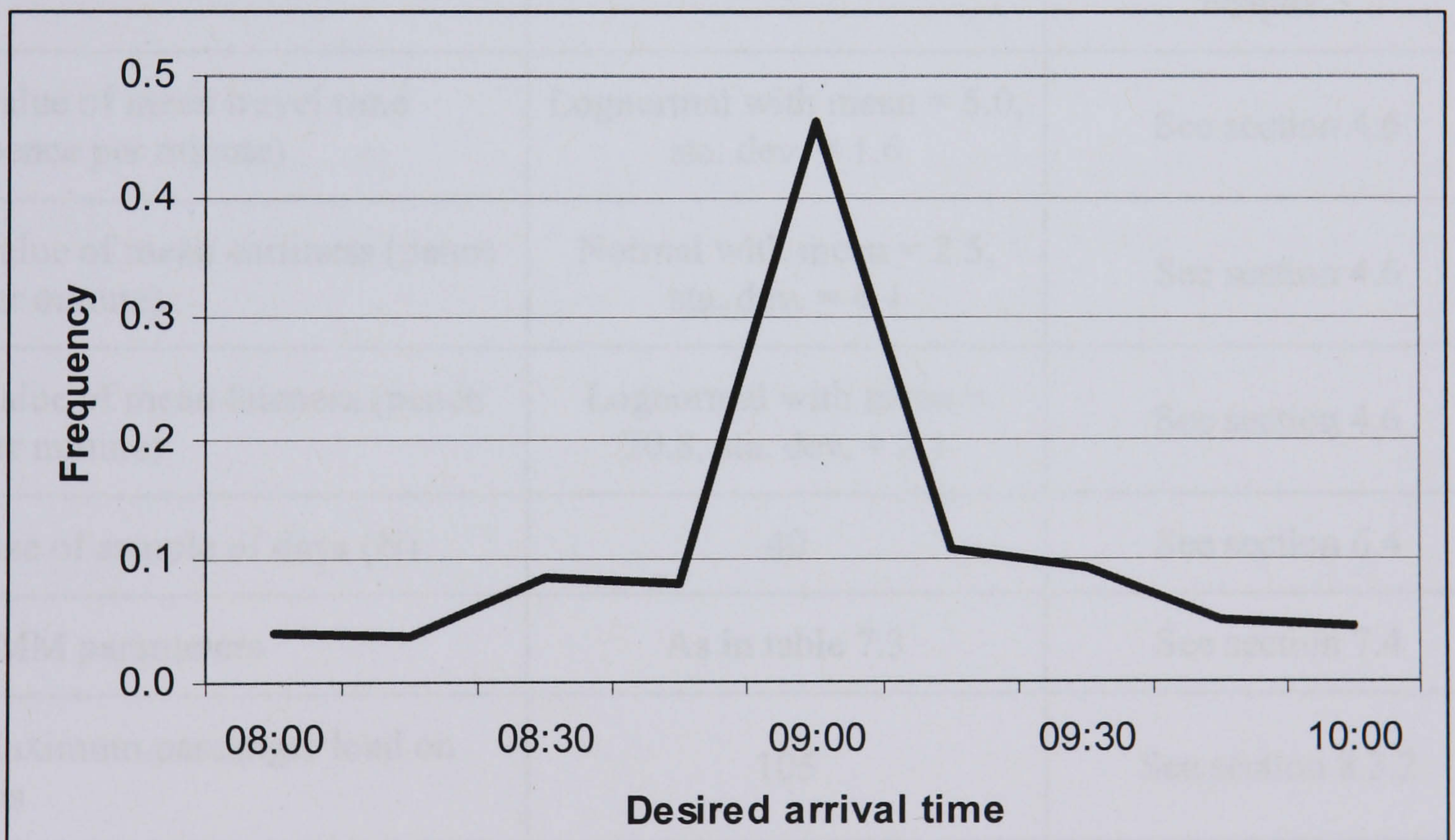


Figure 8.3: The distribution of desired arrival times

Alighting / Boarding	Along Tadcaster Road	Around the Railway Station	Around Clifford Street	Heslington
Acomb	3.2%	7.4%	14.8%	10.3%
Along Tadcaster Road		8.7%	19.2%	16.2%
Around the railway station			10.5%	5.6%
Around Clifford Street				4.1%

Table 8.1: The demand matrix (with the bus stops grouped into zones)

Attribute	Value	Source
Size of sample of bus users (I)	1000	See discussion above
Actual level of demand	As in table 8.1	See section 7.4
Desired arrival times	As in figure 8.3	Survey results – see chapter 3
Value of mean travel time (pence per minute)	Lognormal with mean = 5.0, sta. dev. = 1.6	See section 4.6
Value of mean earliness (pence per minute)	Normal with mean = 2.5, sta. dev. = 4.4	See section 4.6
Value of mean lateness (pence per minute)	Lognormal with mean = 20.8, sta. dev. = 2.1	See section 4.6
Size of sample of days (N)	40	See section 6.4
TMM parameters	As in table 7.3	See section 7.4
Maximum passenger load on bus	105	See section 8.3.2
Convergence test	No change in DTC for 90% of travellers	

Table 8.2: Inputs for an illustrative scenario analysis

8.3.2. Constraining earliness benefits and maximum bus load

The methodology presented earlier in this chapter can be generally used with different sources for the individual WTP and with different TMMs. The specific sources used in the application illustrated here have some drawbacks that required making small modifications to the methodology.

First, we use a distribution of the value of the mean earliness that includes some negative values. As this is based on careful analysis of our survey responses, as described in chapter 4, it does not constitute any problem in itself. However, as mentioned in chapter 4, the dataset that was used to estimate this distribution did not include extremely early departures; therefore, when applying this distribution it is necessary to make sure that the negative WTP is not erroneously interpreted as a justification for some irrational choices. The main risk is that some travellers whose VOE is in the negative tail of the distribution, and whose current DTCs are early enough to make their costs of earliness negative, will keep shifting their departures to earlier times in order to reduce their cost even further. In order to avoid this unreasonable behaviour, a constraint was included in the algorithm, verifying that if the cost of earliness associated with a travellers' DTC is negative already, the option of shifting to an earlier departure can only be chosen if it also offers reduction in the costs other than the cost of earliness. In a mathematical formulation, a departure (j) earlier than the currently chosen one (j_i^0) can be considered only if:

(8.6)

$$COE_i^{j_i^0} \geq 0 \quad and \quad COT_i^j + COE_i^j + COL_i^j < COT_i^{j_i^0} + COE_i^{j_i^0} + COL_i^{j_i^0}$$

$$or$$

$$COE_i^{j_i^0} < 0 \quad and \quad COT_i^j + COL_i^j < COT_i^{j_i^0} + COL_i^{j_i^0}$$

A second amendment to the methodology is needed because the DRACULA version that we use to estimate travel times does not feature any constraining on the maximum passenger load on each bus. In some preliminary test runs of the methodology it became apparent that without such constraint, the number of passengers choosing the same departure might significantly exceed that capacity of a single bus. A bus capacity constraint was therefore introduced externally, as part of our algorithm; it lets a

passenger shift to his/her preferred journey j (which has been chosen using (8.6)) only if the number of passengers on the bus in the simulation is smaller than the bus capacity. Passengers that cannot board their chosen bus are assigned by the algorithm to a slightly earlier or later bus in which the maximum capacity is not reached. This constraint is checked in every iteration, hence some of the passengers who cannot board their preferred bus in one iteration do manage to do so in the following iteration. As shown in table 8.2, the maximum passenger load per bus chosen for the current experiment is 105. Note that this value refers to the physical maximum bus capacity, which is occasionally reached in practice, even if it is much higher than the maximum capacity used for design. While seeking the most appropriate value to use as the physical maximum capacity, we came across a wide range of suggested values, most of which were much lower than 105 (i.e. around 80). Since it was not possible to investigate which of the suggested values was the most realistic, it has been decided that for the current needs even a modest capacity restriction would be sufficient, as it is primarily meant to prevent the algorithm from accepting irrational bus loads; therefore, the relatively high estimate of capacity is used.

8.3.3. Departure time choices and the cost of travel time variability

The DTC and cost estimation methodology presented in the previous section was programmed in C and run twice, once for the "basic" scenario and once for the "extended" scenario. Figure 8.4 presents the gradual changes in the total journey cost, summarised across all 1000 travellers. It shows that in the first few iterations the choices and the respective costs are very unstable. Many travellers move their departure forward or backward, attempting to reduce their generalised cost of travel but sometimes experiencing an actual higher cost, due to the high number of passengers on several particular departures. This is gradually stabilised in later iterations; for both scenarios the process was stopped after 11 iterations. The runtime of the whole process was about 70 hours for each scenario.

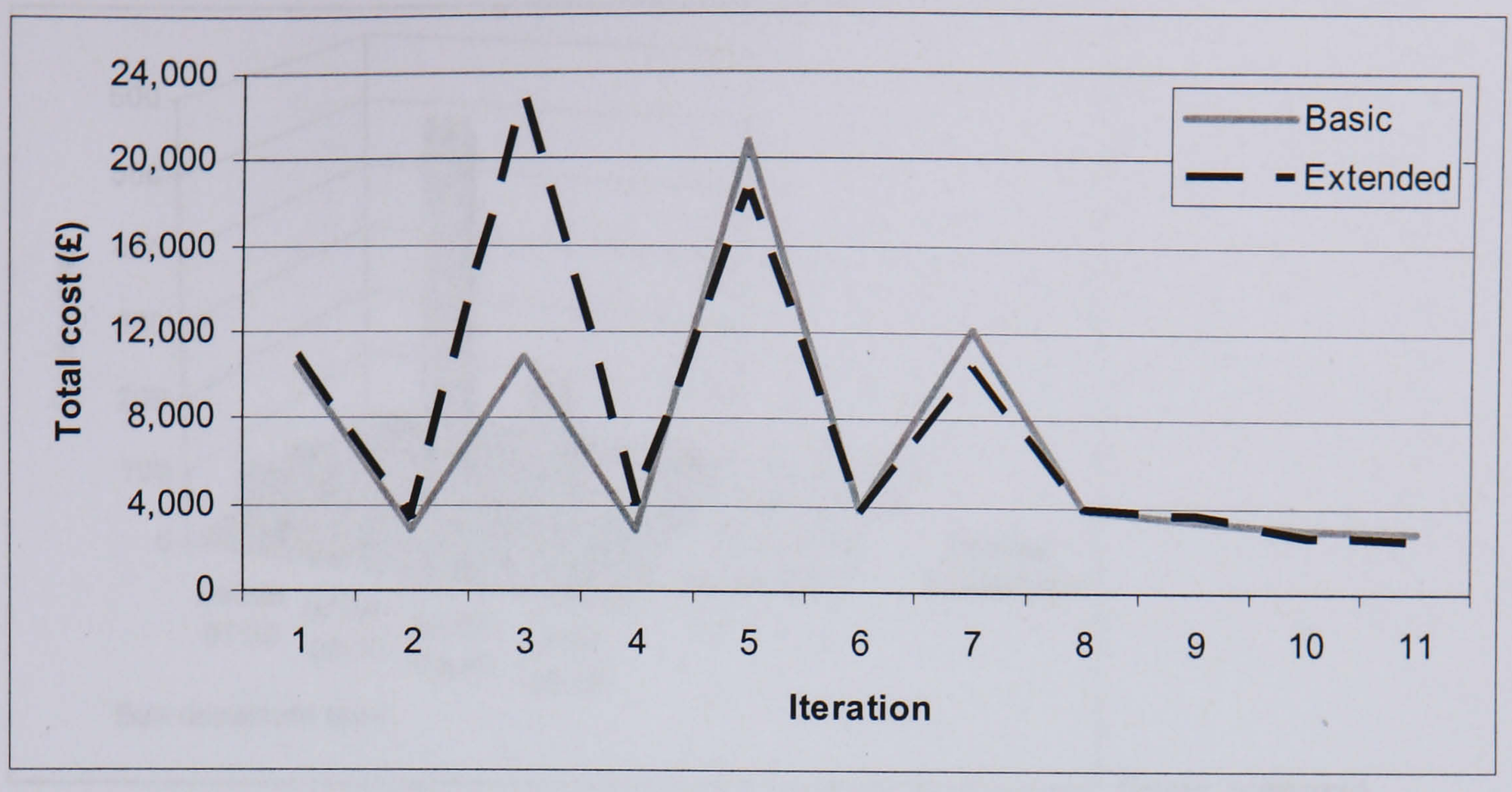
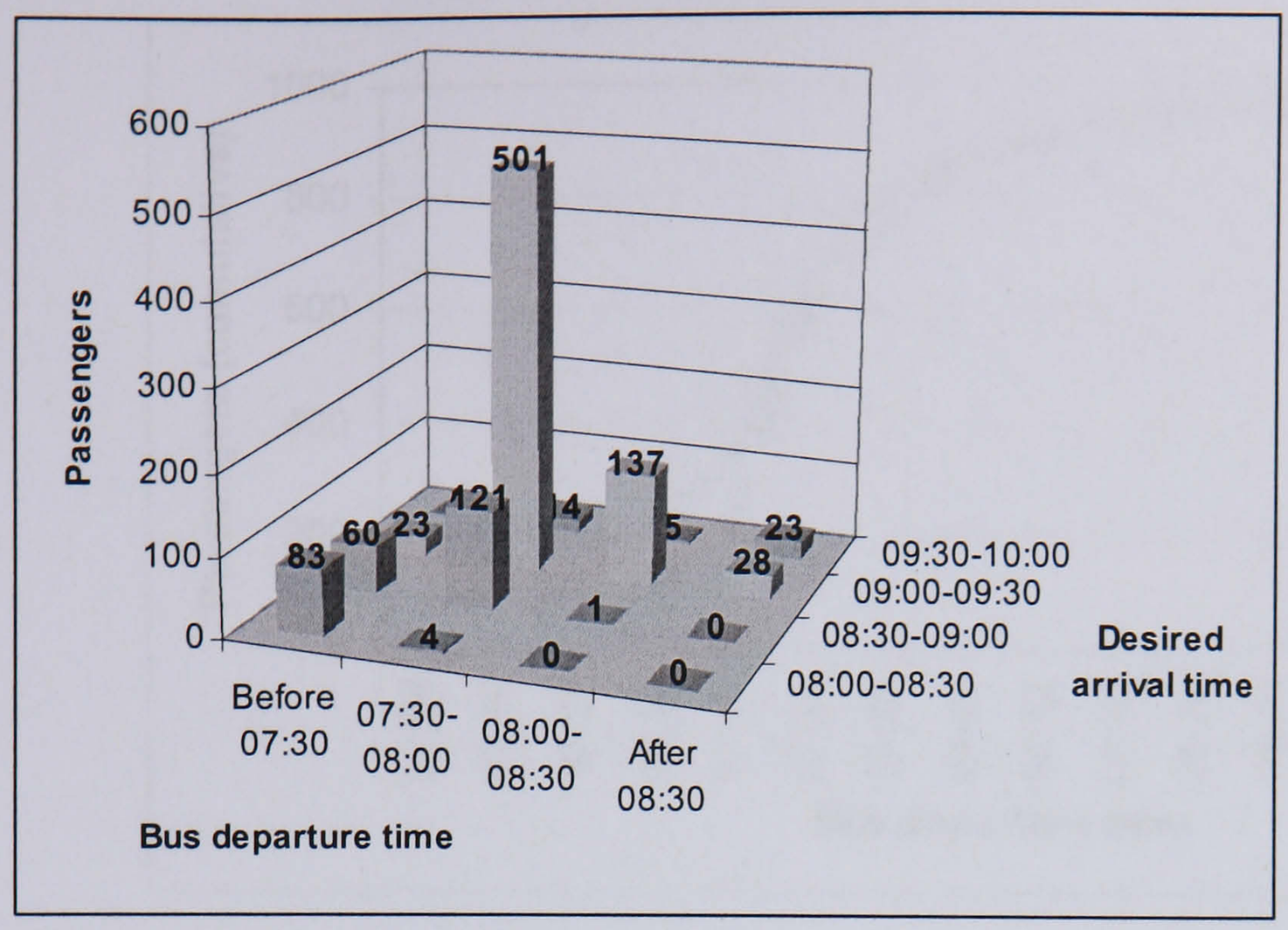


Figure 8.4: Convergence of the total journey cost

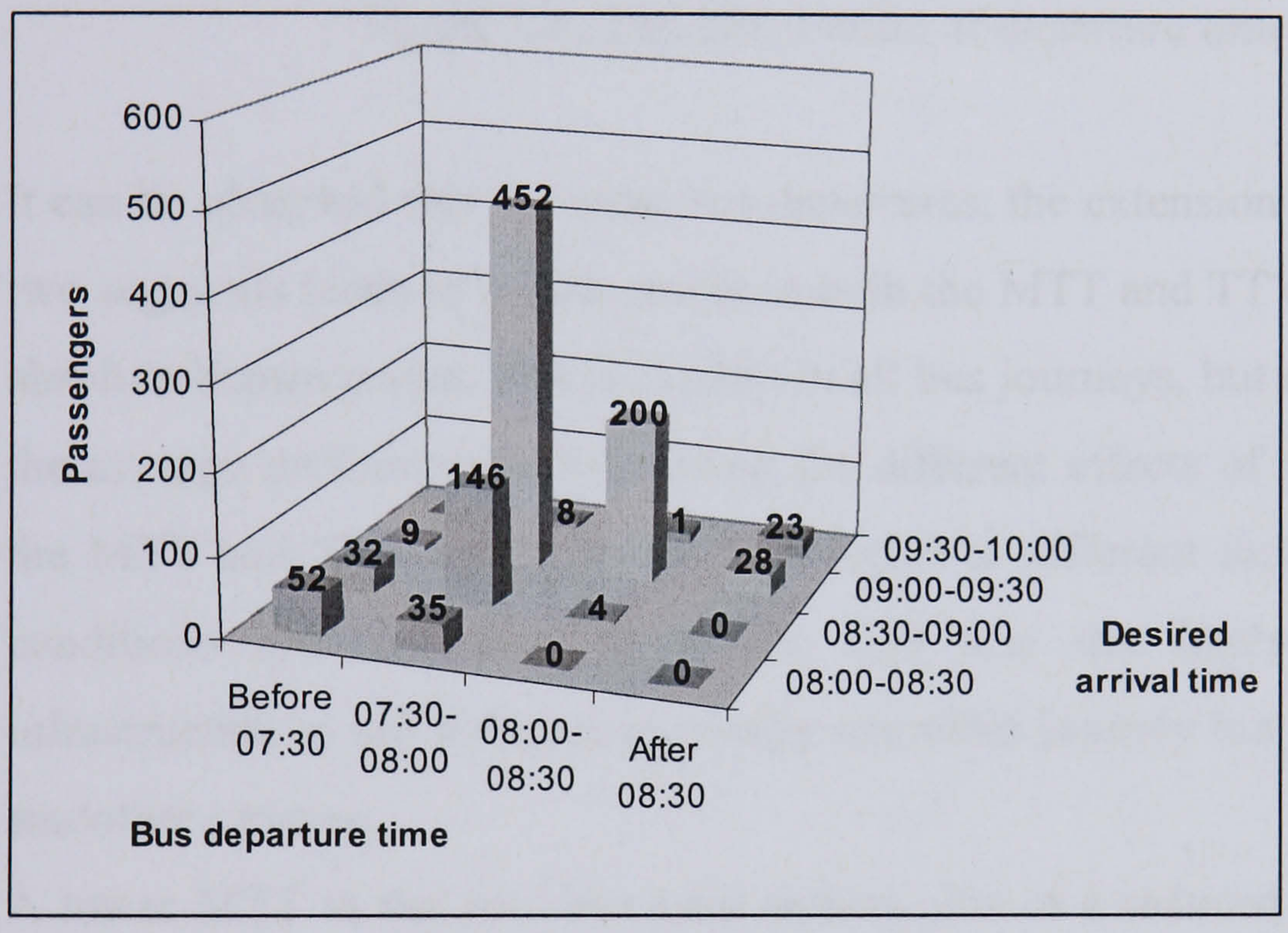
We now focus on the last set of estimates reached for each scenario after convergence. We are primarily interested in comparing the DTCs, travel time distributions and costs between the two scenarios. Figure 8.5 describes the general relationship between desired arrival time to the destination and the chosen departure time, before and after the extension of the bus lane. It shows that with the extended bus lane, travellers generally tend to leave slightly later. This is a direct outcome of the improved travel conditions, as we illustrate later. The shift of some travellers to somewhat later departures is also illustrated in figure 8.6, which depicts the difference in the distribution of DTCs between the scenarios.

Note that the departure times in figures 8.5 and 8.6 are the times when the bus, and not each traveller, leaves its origin. The choices are presented in this aggregated form in order to make the graph easy to read. Travellers board the bus later (or at least not earlier) than the bus departure time, at various points along its route; but since in this example the boarding stop of each traveller is fixed, choosing a bus that departs later also implies that the actual boarding is later.

Tables 8.3 and 8.4 present the changes in the MTT and TTV for each bus departure following the construction of the bus lane. As explained in chapter 1, TTV is measured as the standard deviation of travel times. The tables focus on the section of route 4 in which the bus lane is introduced, i.e. the section B-C in figure 8.2. Naturally, this section is of special interest in the current analysis, since it uses the general-traffic lanes



Basic scenario



Extended scenario

Figure 8.5: Departure time choices of travellers with different desired arrival times

in the basic scenario but has a separate right of way in the extended scenario. The bus lane section includes three bus stops, two at its ends and one at a distance of about 180 meters from the beginning of the section (and about 320 meters from its ending); we therefore present the MTT and TTV separately for each of the two segments between bus stops. The information in table 8.3 is also presented graphically in figure 8.7 (MTT on segment 1) and figure 8.8 (MTT on segment 2). The information in table 8.4 is presented graphically in figure 8.9 (TTV on segment 1) and figure 8.10 (TTV on segment 2).

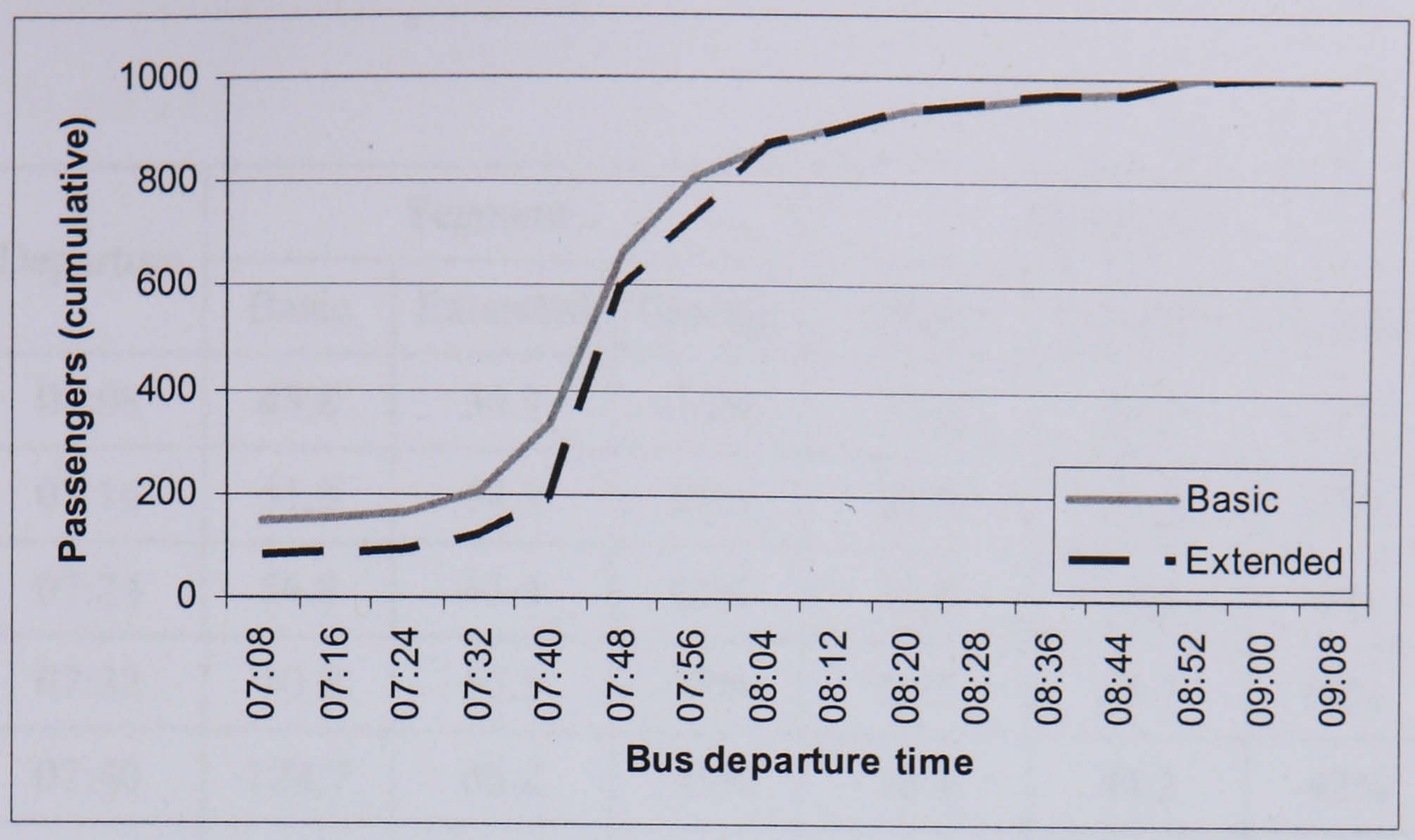


Figure 8.6: The distribution of departure time choices

It can be observed that for most bus departures, the extension of the bus lane into these two segments leads to improvement in both the MTT and TTV. However, this is not an absolute improvement, that is evident in all bus journeys, but a *relative* improvement in the average performance. Examining the different effects of the bus lane extension on the MTT and TTV on the buses that depart at different times reminds us that travel conditions are stochastic in nature, and that the likely outcome of improved infrastructure is not a deterministically smoother journey but an increased *chance* of a smoother journey.

A lower MTT in the new bus lane section, due to a reduced level of friction between buses and cars, is evident for the majority of the bus journeys, on both segments (table 8.3). The average MTT across all departures reduces by 10% on segment 1 and 15% on segment 2. There is also a reduction in the standard deviation of MTT, i.e. in the level of fluctuation between the different departures. Both the departure with the minimum MTT and the one with the maximum MTT have shorter journey duration in the extended scenario. The improvement on segment 2 is generally greater, when expressed as a percent of the original MTT. This is probably an outcome of the fact that segment 1 is only 180 meter long; this length is not sufficient for considerable acceleration, since it is almost immediately necessary to decelerate in order to stop again at the end of the segment. Segment 2 is about twice longer than segment 1, and there is more room in it to take advantage of the segregation from private car traffic.

Departure	Segment 1			Segment 2		
	Basic	Extended	Change	Basic	Extended	Change
07:08	43.8	36.7	-16%	31.8	24.9	-22%
07:16	51.8	58.3	13%	29.3	21.8	-25%
07:24	56.8	63.4	12%	35.6	34.8	-2%
07:32	70.8	55.5	-22%	50.7	55.7	10%
07:40	124.7	68.2	-45%	76.4	44.2	-42%
07:48	79.8	66.7	-16%	78.4	61.7	-21%
07:56	105.6	110.4	5%	69.1	56.3	-18%
08:04	73.1	57.4	-21%	106.6	41.8	-61%
08:12	70.5	43.6	-38%	77.4	44.6	-42%
08:20	57.9	57.8	0%	67.2	48.3	-28%
08:28	65.7	72.6	11%	56.6	50.7	-10%
08:36	59.6	56.7	-5%	53.3	51.8	-3%
08:44	66.7	66.1	-1%	64.7	45.8	-29%
08:52	63.1	67.3	7%	50.4	60.0	19%
09:00	56.2	71.8	28%	38.9	67.6	74%
09:08	47.2	35.5	-25%	35.5	24.0	-32%
Mean	***68.3***	***61.7***	***-10%***	***57.6***	***45.9***	***-15%***
Sta. dev.	***20.9***	***17.2***	***-18%***	***21.2***	***13.7***	***-35%***
Min	***43.8***	***35.5***	***-19%***	***29.3***	***21.8***	***-26%***
Max	***124.7***	***110.4***	***-11%***	***106.6***	***67.6***	***-37%***

Table 8.3: Mean travel time on individual bus journeys (in seconds)

Departure	Segment 1			Segment 2		
	Basic	Extended	Change	Basic	Extended	Change
07:08	20.1	14.8	-26%	20.7	3.0	-85%
07:16	17.4	26.2	51%	23.0	2.2	-90%
07:24	17.9	21.8	22%	28.5	25.1	-12%
07:32	27.4	23.9	-13%	28.2	29.6	5%
07:40	33.6	17.0	-49%	45.8	33.2	-28%
07:48	29.9	15.4	-49%	43.0	29.6	-31%
07:56	34.4	31.3	-9%	31.9	40.8	28%
08:04	39.1	39.7	2%	58.2	34.3	-41%
08:12	41.8	17.2	-59%	59.9	28.9	-52%
08:20	39.8	18.1	-55%	49.9	44.8	-10%
08:28	36.2	39.0	8%	46.9	40.7	-13%
08:36	35.6	13.9	-61%	29.0	30.3	4%
08:44	38.8	20.0	-49%	41.5	26.0	-37%
08:52	17.0	32.3	90%	30.9	46.9	52%
09:00	26.4	9.6	-64%	26.7	40.5	52%
09:08	27.6	32.5	17%	25.1	3.4	-87%
Mean	***30.2***	***23.3***	***-23%***	***36.8***	***28.7***	***-22%***
Sta. dev.	***8.5***	***9.2***	***8%***	***12.5***	***14.4***	***15%***
Min	***17.0***	***9.6***	***-44%***	***20.7***	***2.2***	***-89%***
Max	***41.8***	***39.7***	***-5%***	***59.9***	***46.9***	***-22%***

Table 8.4: Travel time variability on individual bus journeys (sta. dev., in seconds)

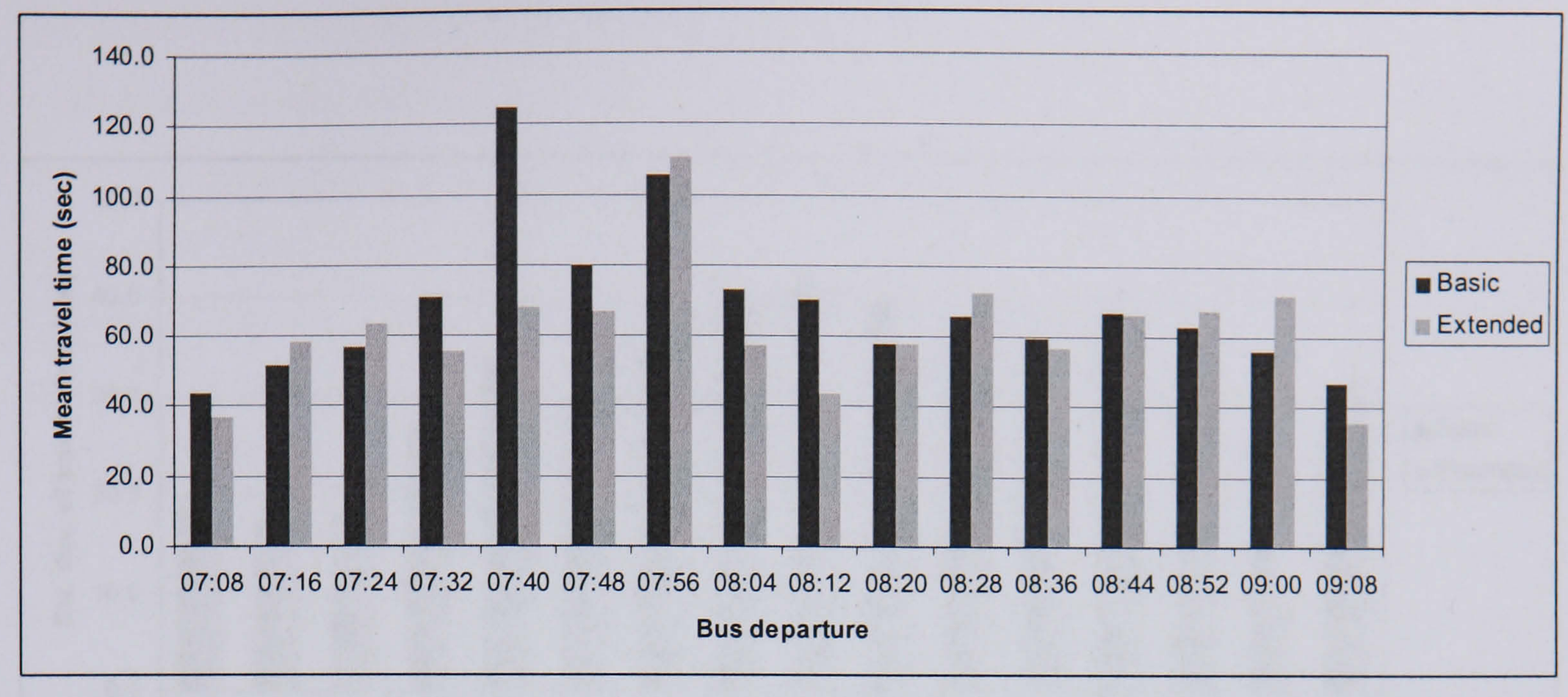


Figure 8.7: Changes in the mean travel time on segment 1

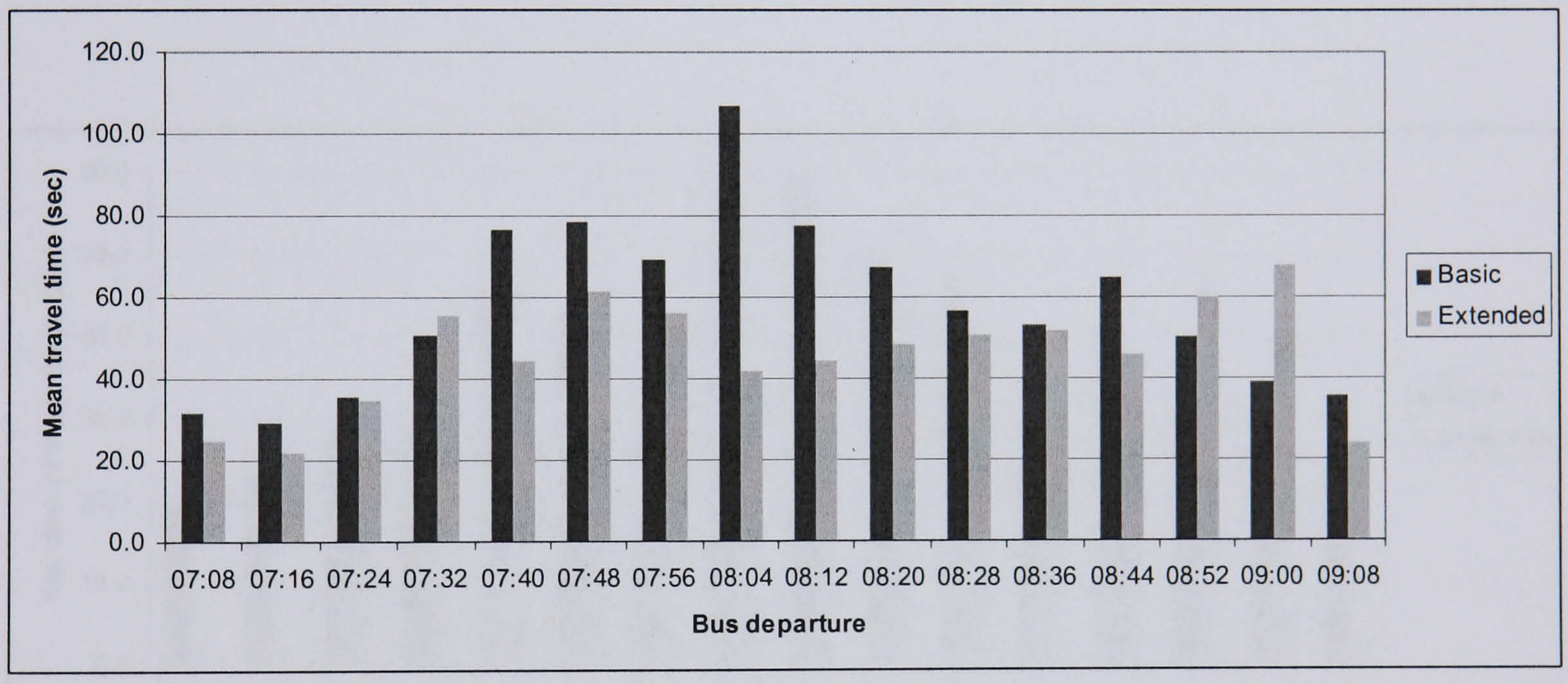


Figure 8.8: Changes in the mean travel time on segment 2

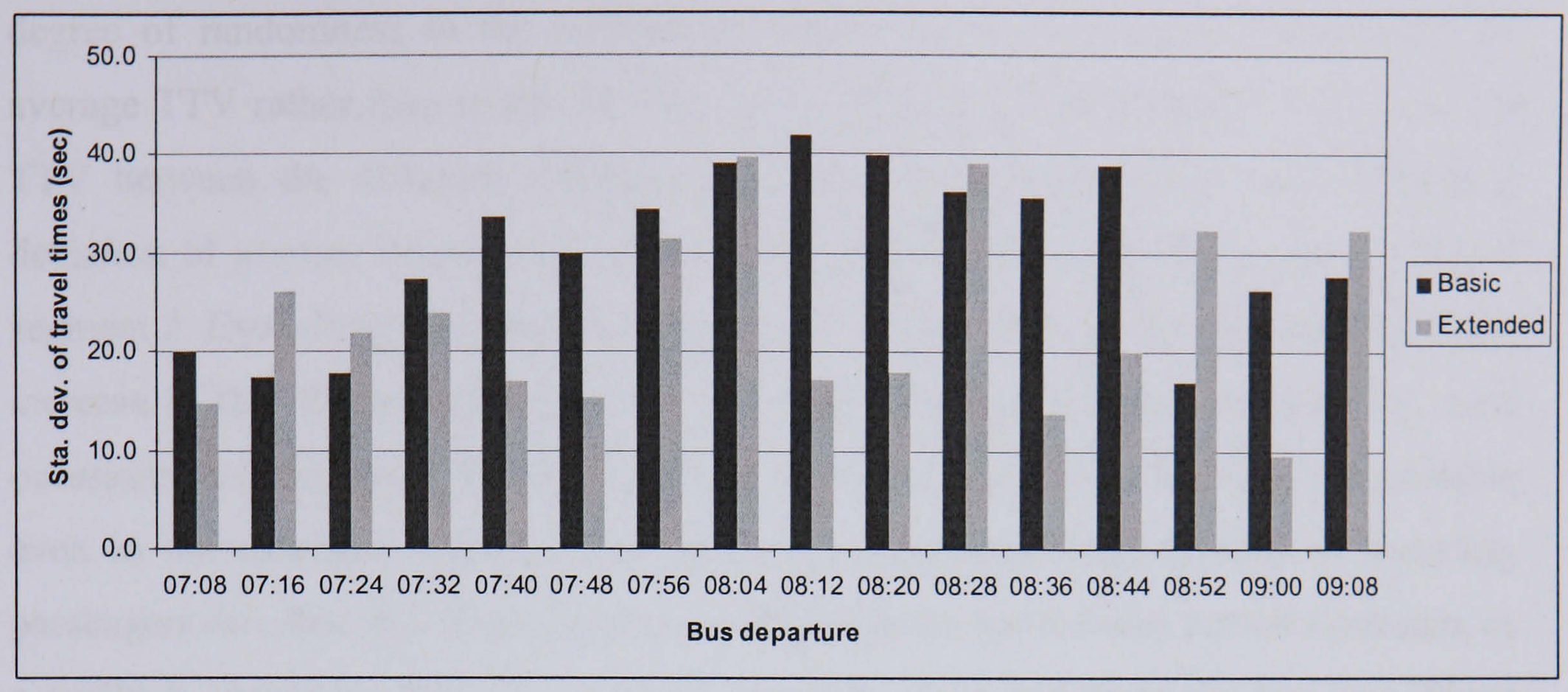


Figure 8.9: Changes in travel time variability on segment 1

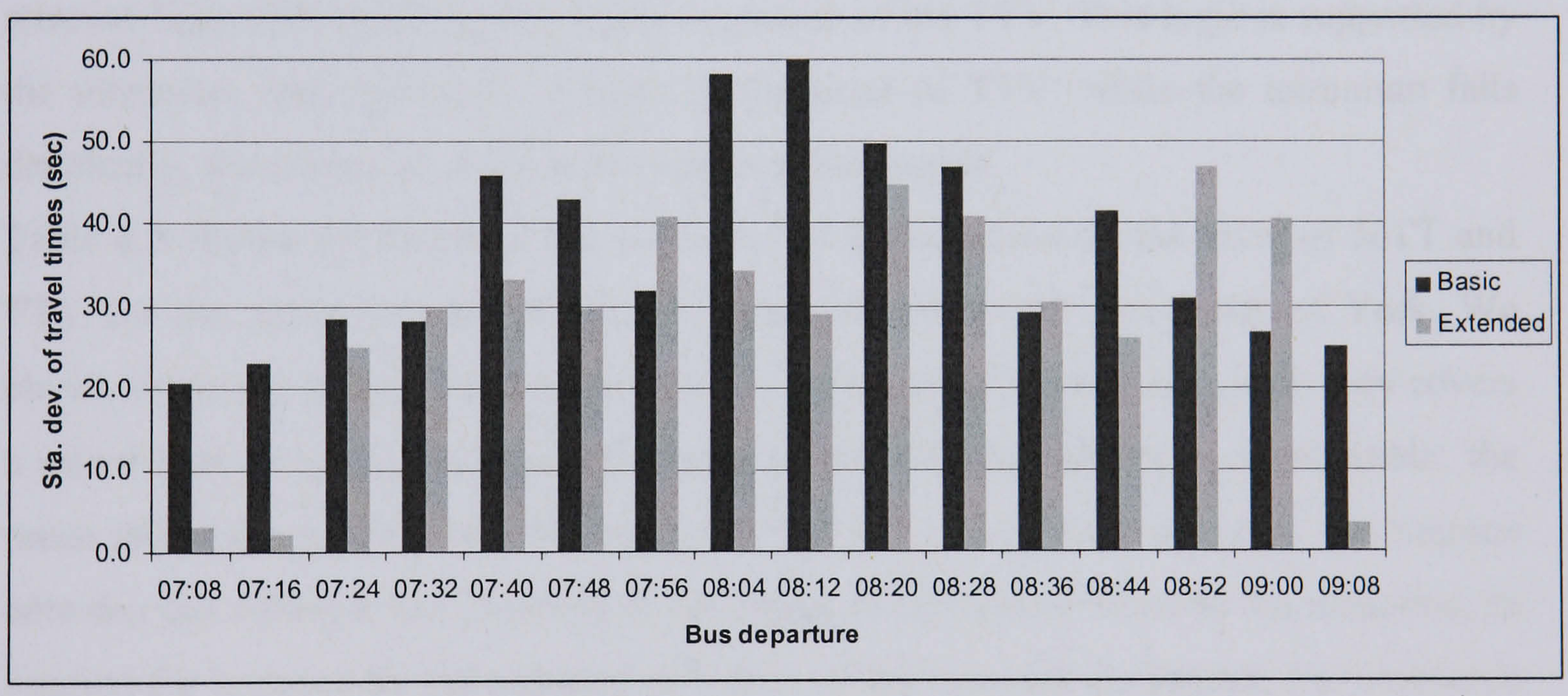


Figure 8.10: Changes in travel time variability on segment 2

The extent of reduction in TTV, in the segments where a bus lane has been constructed, is greater than the reduction of MTT (table 8.4). As with the MTT, there is a high degree of randomness in the network performance, and the improvement is in the average TTV rather than in the TTV on all bus journeys. The variation in the level of TTV between the different departures (i.e. the standard deviation of the standard deviation of journey times) does not fall: it goes up by 8% on segment 1 and 15% on segment 2. Eyeballing the network simulation as it runs confirms that the reason for this increase is that the separate right of way makes the bus journey smoother on most occasions, but not all of them. Occasional serious delays in the bus lane are possible even in the extended scenario, due to reasons such as a high number of boarding passengers on a bus that arrives late following irregular travel times further upstream, or a gridlock caused by private cars at the junction at the exit from the bus lane. Since there are some serious delays, even if less frequent than in the basic scenario, it is likely that on most days and most bus departures TTV decreases, but on some specific once it remains high, thus resulting in a higher variation of the TTV. This logic is supported by the minimum and maximum changes in the level of TTV: while the minimum falls drastically, the reduction of the maximum is more modest.

Table 8.5 shows the effect of the extension of the bus lane on the level of MTT and TTV for the entire journey of route 4, from Acomb to the University of York. We obviously do not expect the new bus lane to cause a major change here, as it only covers a minor part of route. However, the table shows that the change is considerable: the mean reduction in MTT is 8.8% and in TTV it is 11.8%. There are some fluctuations between the different bus journeys in the extent of difference between the scenarios, as implied for instance by the standard deviation of the standard deviations; but compared to the fluctuations in the measurements on individual segments, the results for the entire journey exhibit greater stability. The difference in the level of MTT and TTV fluctuation between a single segment and the whole route results from the ability to partially recover, before the end of the journey, some of the irregular delays that occur on specific segments. We found evidence for such ability in the analysis of real data in chapter 1, and hence, this output of the simulated network makes good sense.

Departure	Mean travel time			Travel time variability		
	Basic	Extended	Change	Basic	Extended	Change
07:08	59.1	57.0	-3.6%	2.75	2.51	-8.7%
07:16	57.9	56.7	-2.2%	2.15	2.04	-4.9%
07:24	57.9	56.9	-1.8%	1.73	1.84	5.9%
07:32	75.0	68.5	-8.7%	4.67	3.58	-23.3%
07:40	104.8	88.0	-16.1%	6.11	5.8	-5.1%
07:48	91.9	86.8	-5.6%	4.55	4.44	-2.5%
07:56	108.3	98.1	-9.4%	3.87	4.38	13.1%
08:04	142.1	129.1	-9.1%	3.19	3.32	3.9%
08:12	162.2	140.0	-13.7%	4.46	3.52	-21.0%
08:20	144.1	132.8	-7.8%	1.71	1.62	-5.6%
08:28	120.4	109.8	-8.8%	6.42	5.16	-19.7%
08:36	116.1	105.4	-9.2%	2.89	2.56	-11.3%
08:44	133.7	118.6	-11.3%	6.46	5.63	-12.8%
08:52	58.1	56.8	-2.2%	1.96	1.87	-4.7%
09:00	72.3	68.1	-5.8%	4.54	3.66	-19.4%
09:08	88.0	79.7	-9.5%	8.19	6.00	-26.7%
Mean	***99.5***	***90.8***	***-8.8%***	***4.1***	***3.6***	***-11.8%***
Sta. dev.	***34.7***	***29.2***	***-15.7%***	***1.9***	***1.5***	***-23.2%***
Min	***57.9***	***56.7***	***-2.2%***	***1.7***	***1.6***	***-5.6%***
Max	***162.2***	***140.0***	***-13.7%***	***8.2***	***6.0***	***-26.7%***

Table 8.5: The effect of the new bus lane on travel time of the entire route (in minutes)

The most striking finding in table 8.5, when it is compared to table 8.3, is that although the direct time saving within the bus lane extension is never more than 2 minutes, the average travel time saving on the whole route is almost 9 minutes. In other words, the simulation outputs imply that most of the reduction in travel time occurs not on the bus lane itself, but elsewhere along the route; this is indeed surprising. In order to understand the source of this result, an attempt was made to follow the movements of many different buses on the simulation screen, in both the basic and extended scenarios, and trace differences in the way delay occurs. This is clearly not a systematic way of investigating whether the effect of the bus lane on travel times in other locations is plausible; but it seems the only possible way, since we do not have real travel time measurements for scenarios with and without a bus lane, and since it is not possible to orient the simulation to only generate time differences that result from the causes that we wish to examine. Our main conclusion from this experiment is that a typical event, which subsequently triggers serious delays, occurs much more frequently in the basic scenario, and leads to the mentioned difference between the scenarios. This typical event involves the arrival of a bus to a stop before the preceding bus has left it; from this stop on, the two buses tend to bunch together, and their frequent arrival at some of the stops at almost the same time hinders their efforts to enter or leave the stop undisturbed. In the extended scenario, slight delays at the first sections of the route have a better chance of recovery in the extended bus lane, before entering the congested city centre, and such deterioration is often avoided. The important insight suggested by this finding is that apart from the direct effect that the introduction of a bus lane has on bus travel conditions on the lane itself, it also has an indirect, but yet strong, positive effect on the way buses perform further downstream, as it gives them a chance to realign (at least partially) with their planned timetable. It can be assumed that the location of the bus lane examined here makes it a critical link of the route, as it is situated late enough along the route to suffer from delays but also just before the city centre, where due to the high number of passengers, small delays might deteriorate to bigger ones. Being a critical link might explain the significant difference between the scenarios and insinuate that this is a good choice of location for a bus lane. It should be reminded, though, that these conclusions are not supported here by real data; it would be important to look for empirical evidence for these effects in future work.

The results presented so far look at the levels of MTT and TTV experienced on the various bus journeys. This gives an idea of the effect of the extension of the bus lane on

the general performance of route 4. However, as our DTC model suggests, bus users are less concerned about the standard deviation of their travel time, and are mostly interested in optimising the extent of late or early arrival to their destinations. The actual cost of travel is therefore derived from these results after applying formulas (8.2) and (8.3), which take into account the individual MTT, earliness and lateness, and the different WTP ascribed by each traveller to changes in these attributes. The distribution of individual WTP was described earlier in the chapter and in more detail in chapter 4; we now combine it with the distribution of the individual MTT, ME and ML that each traveller experiences during his/her bus journey.

Figures 8.11, 8.12 and 8.13 describe the cumulative frequencies of the individual MTT, ME and ML. The mean individual travel times are shorter than the bus travel times presented above, as most passengers board and alight at various points along the route. The figures show reduction in all three time elements, following the extension of the bus lane. They also illustrate that most travellers prefer to arrive at their destination significantly earlier than their formal desired arrival time, in order to reduce their chance of arriving late. This is a direct result of the high lateness penalty found in chapters 3 and 4; a traveller whose VOL is 4 times higher than his/her VOE, would rather arrive up to 4 minutes early than 1 minute late. The strong preference to avoid lateness is also apparent in the fact that for almost a quarter of the travellers, the mean lateness is zero, implying that they are willing to leave early enough to make the risk of late arrival almost nonexistent.

Figures 8.14, 8.15 and 8.16 describe the cumulative frequencies of the elements of the individual trip cost. Figure 8.17 shows the sum of all cost elements. The presented costs are per individual journey. The figures illustrate that the apparent result of the high cost of lateness, or of the strong willingness to avoid it, is that the final cost of lateness is quite low compared to the cost of earliness. Figure 8.18 and table 8.6 present a summary of the total individual costs during the analysis period (namely the morning peak), across the 1000 individual trips in the sample. The absolute numbers presented are less important than the differences between the scenarios (for convenience, table 8.6 also presents the corresponding values in time units). The extension of the bus lane appears to cause a reduction of about 14% in the total journey cost. The effect of the extension on the cost of earliness (-46%) and lateness (-20.4%) is bigger than the effect on the cost of travel time itself (-5.8%). Namely, it seems that the reliability effects of the bus lane scheme are stronger than its time saving effects; this is a crucial finding, as

we have demonstrated in chapter 2 that the common practice of scheme appraisal tends to ignore reliability benefits. Note that the significant fall in the cost of earliness does not result from a relatively uniform reduction across many travellers, but primarily from the existence of a small group of travellers in the basic scenario that due to high penalties on earliness, choose to depart extremely early; there are only very few such travellers in the extended scenario.

The most significant finding is that the introduction of TTV considerations into the calculation of costs has doubled the estimated benefit for the entire journey: from 5.8% when only the MTT is taken into account to 13.8% when all scheduling costs are considered. As this is the effect on the total cost of a relatively minor scheme, that includes a 500-meter-long bus lane without introducing any traffic control measures in intersections, the benefit from reduced TTV seems significant in both absolute and relative terms. This finding has important policy implications, which we discuss at the end of this chapter.

8.4. Conclusions

In this chapter we have demonstrated the feasibility of estimating the cost of bus travel, including the contribution of travellers' trip scheduling behaviour, through the joint application of a departure time choice model with traffic microsimulation. The methodology is based on a disaggregate, microscopic approach; it is sensitive to the diversity of tastes and preferences in the population of bus users, and also to randomness, heterogeneity and many local factors in the urban road network. The methodology was applied in an illustrative case study, in which the extension of an existing bus lane is considered. By examining the outputs of this procedure, it was generally found that the way it converges into estimates of the travel cost makes good sense.

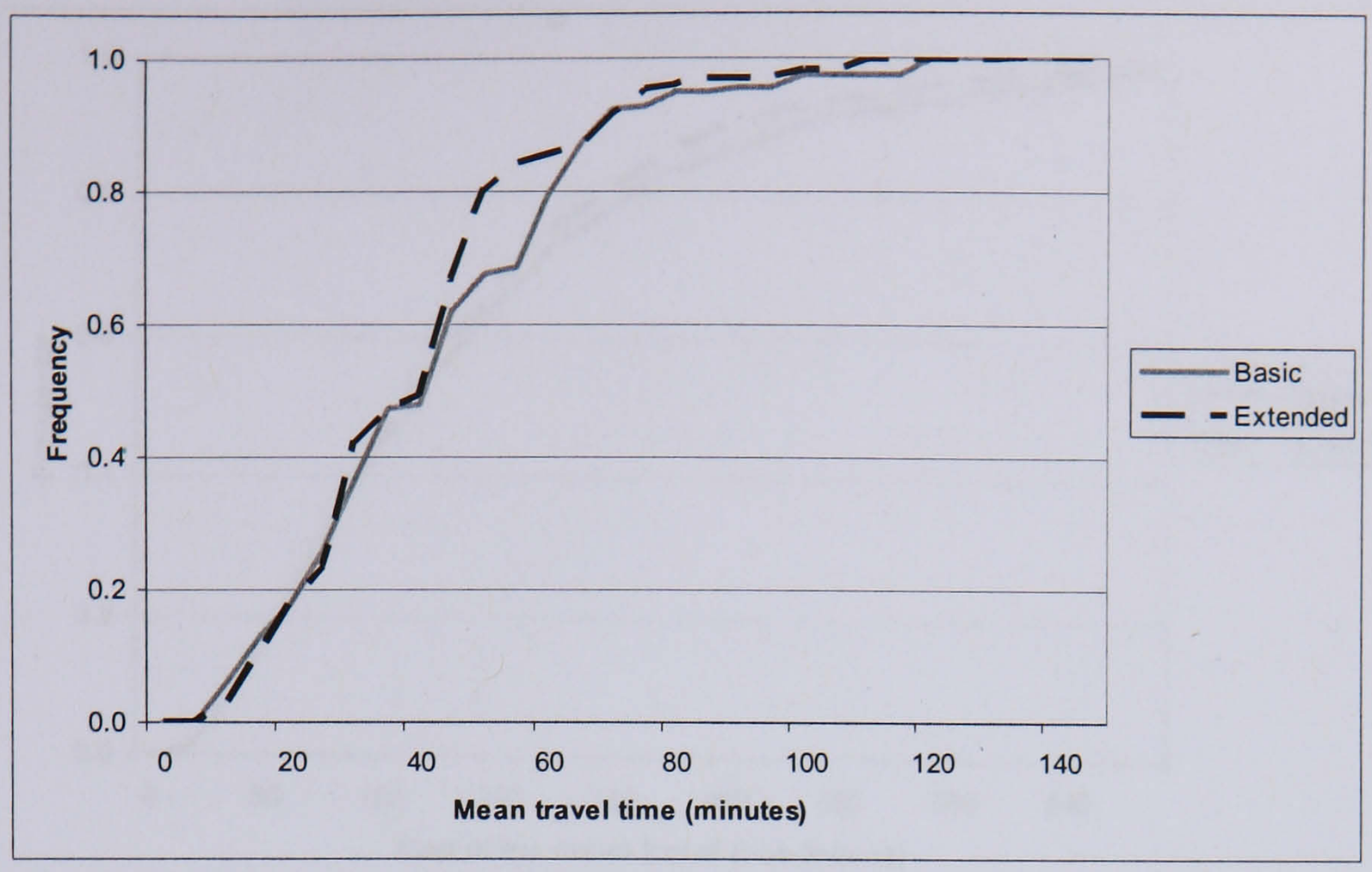


Figure 8.11:
Cumulative frequency curve of the mean travel time

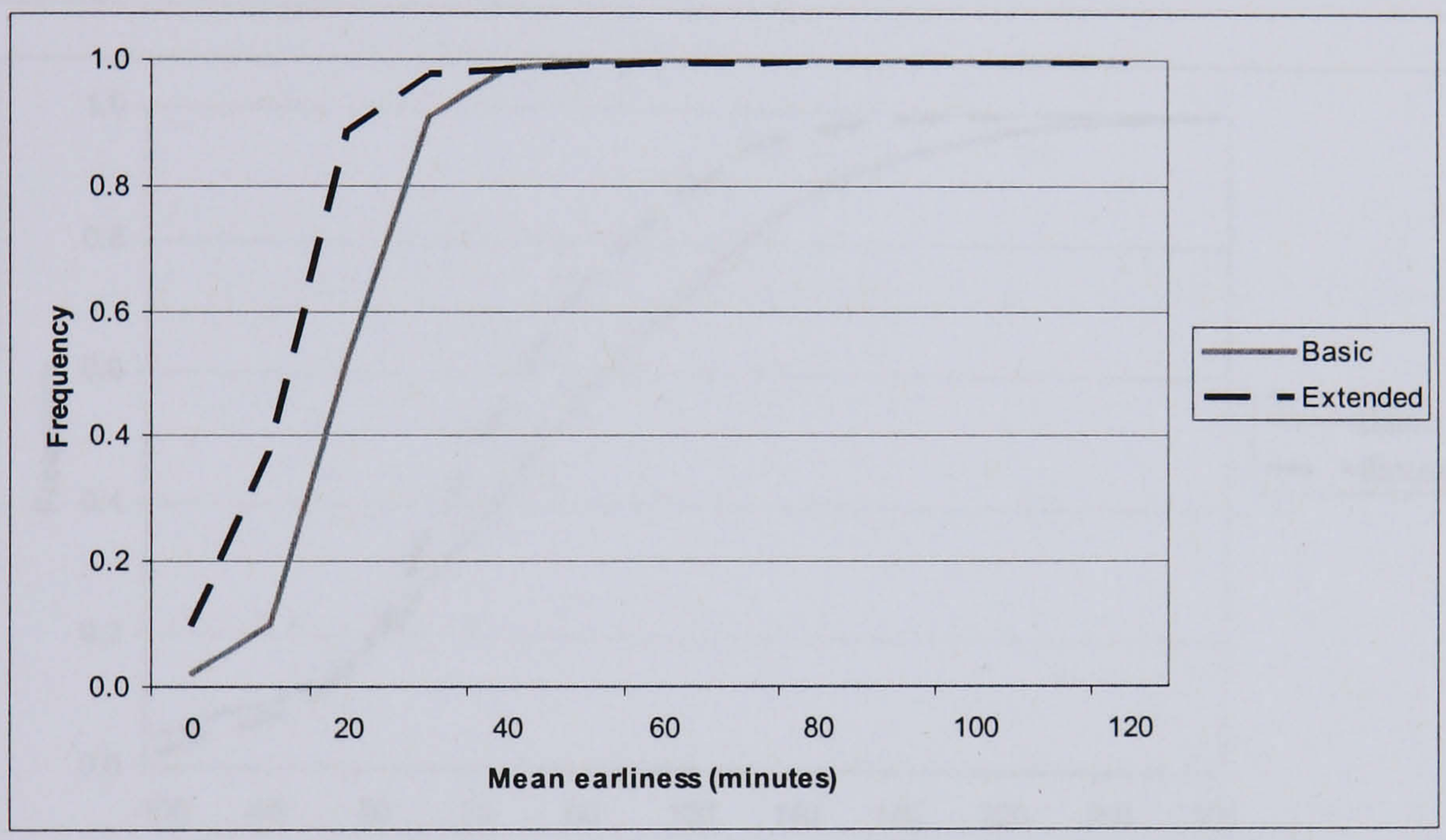


Figure 8.12:
Cumulative frequency curve of the mean earliness

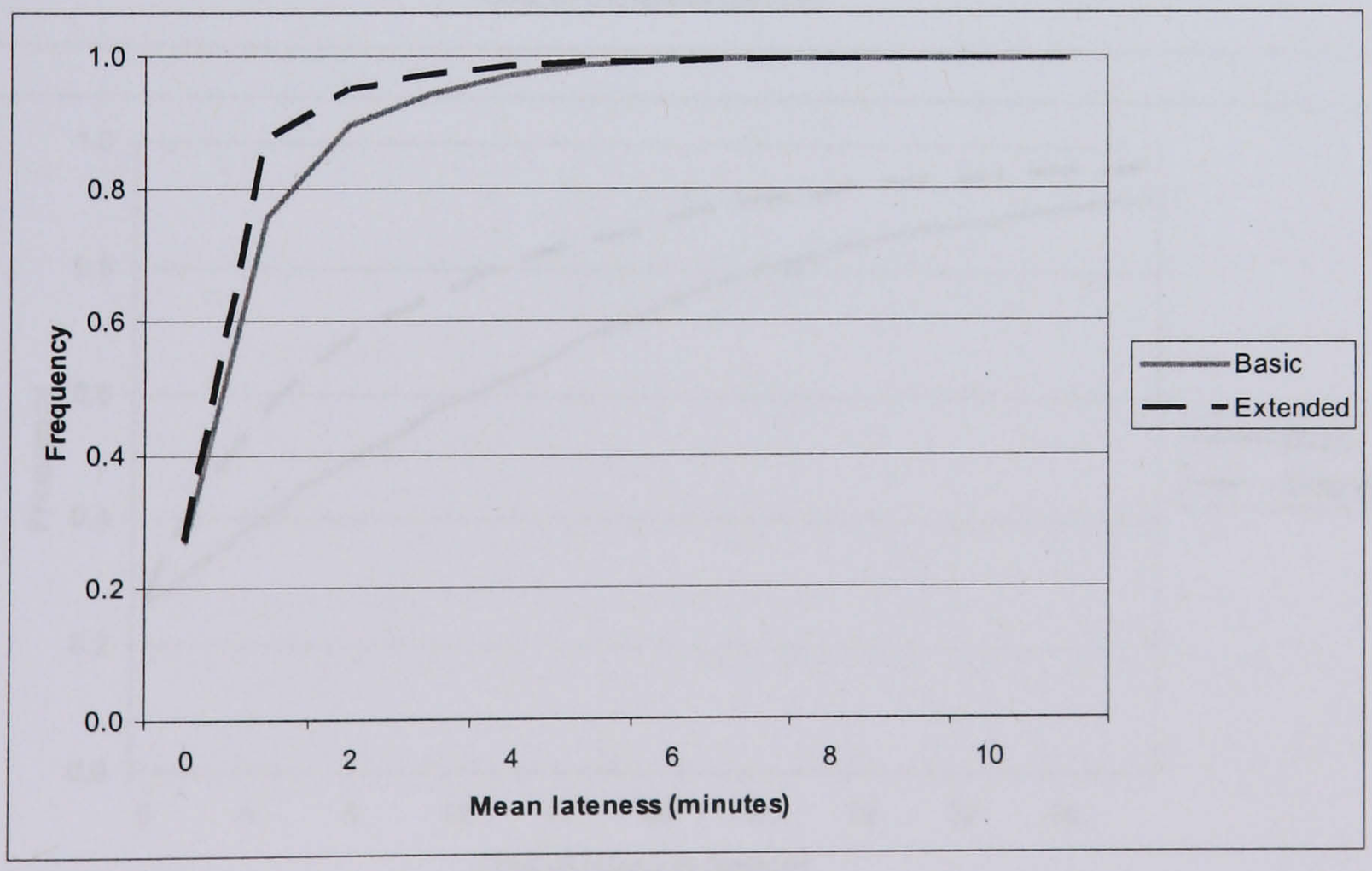


Figure 8.13:
Cumulative frequency curve of the mean lateness

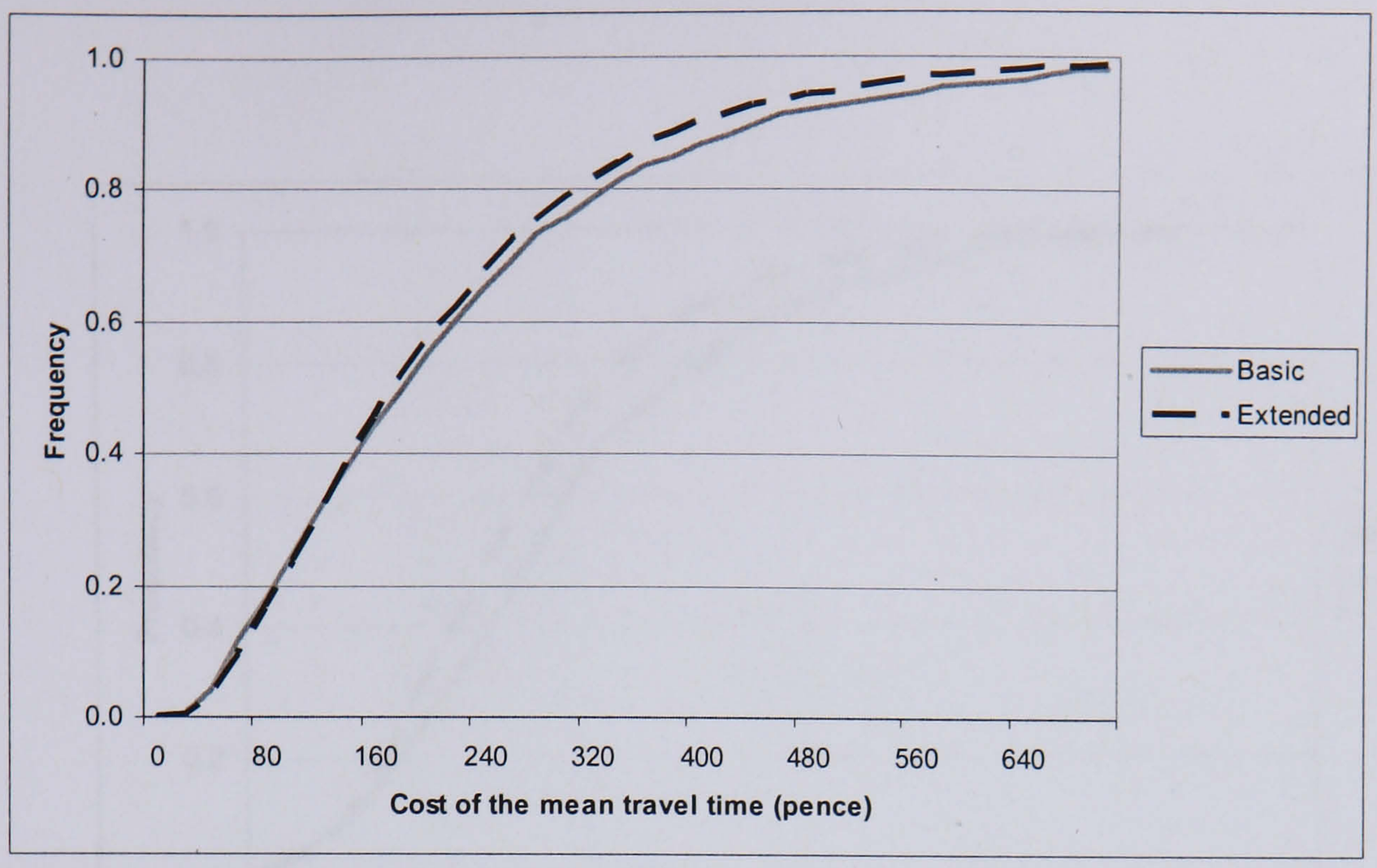


Figure 8.14: Cumulative frequency curve of the cost of the mean travel time

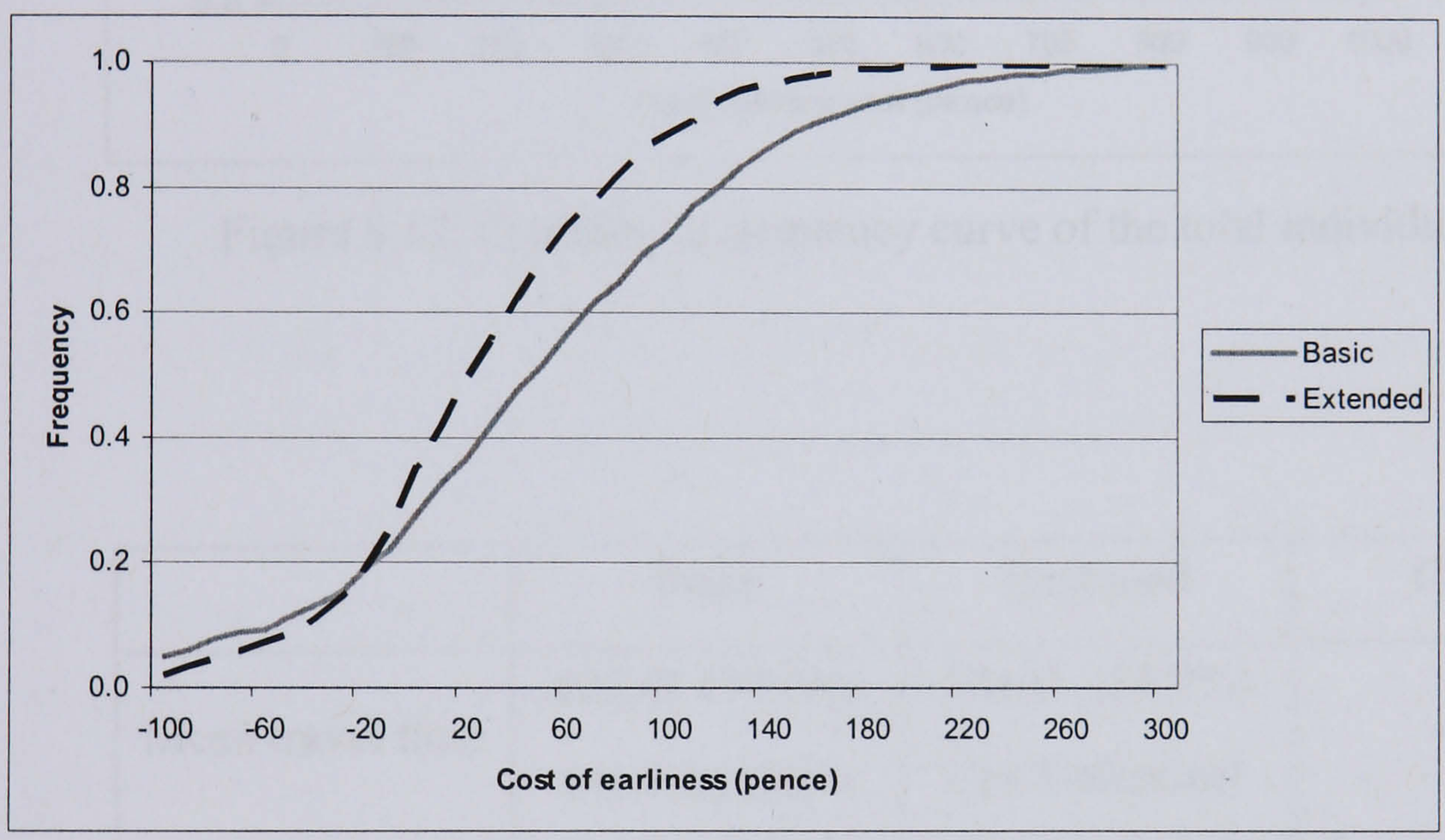


Figure 8.15: Cumulative frequency curve of the cost of the mean earliness

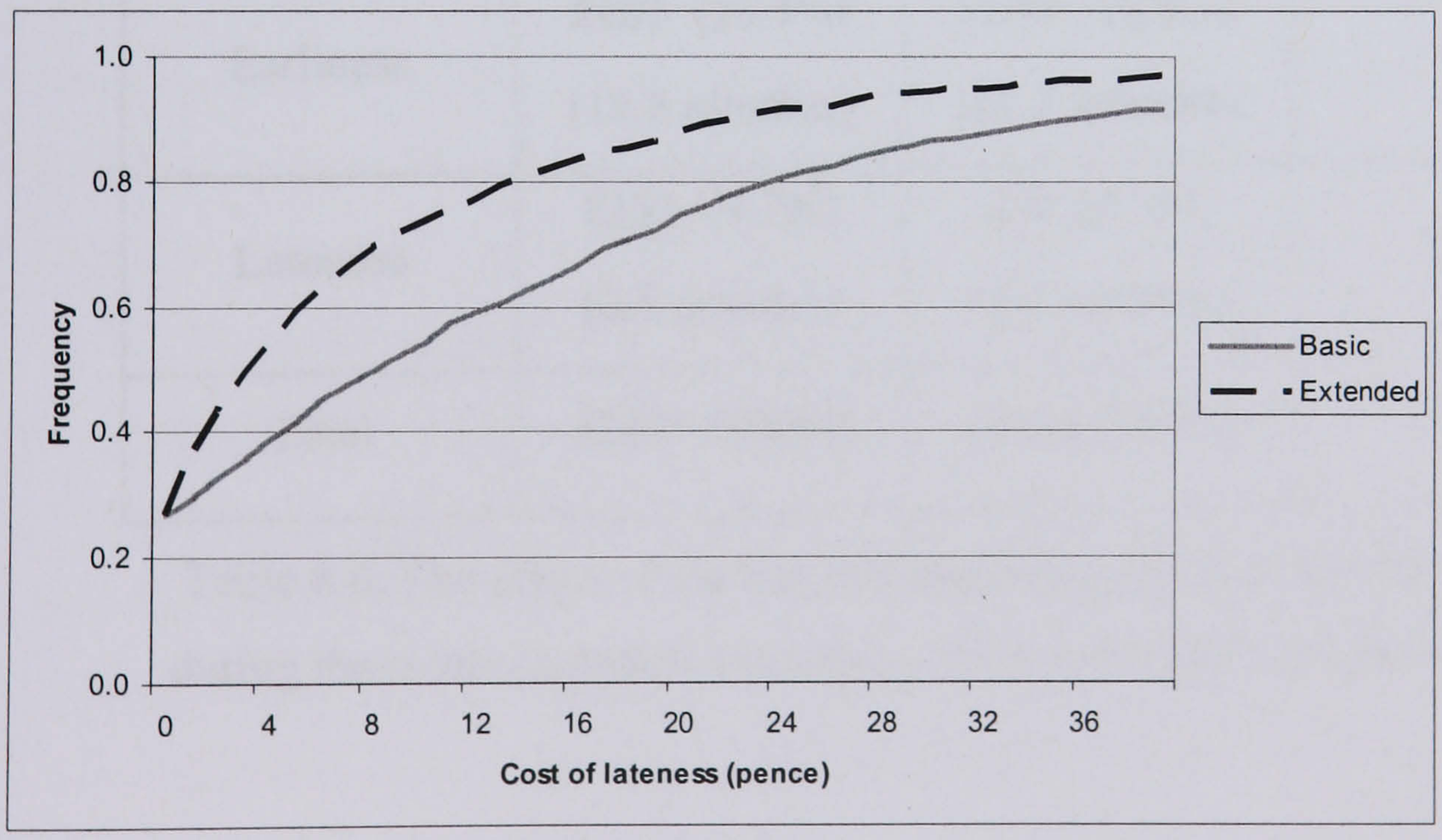


Figure 8.16: Cumulative frequency curve of the cost of the mean lateness

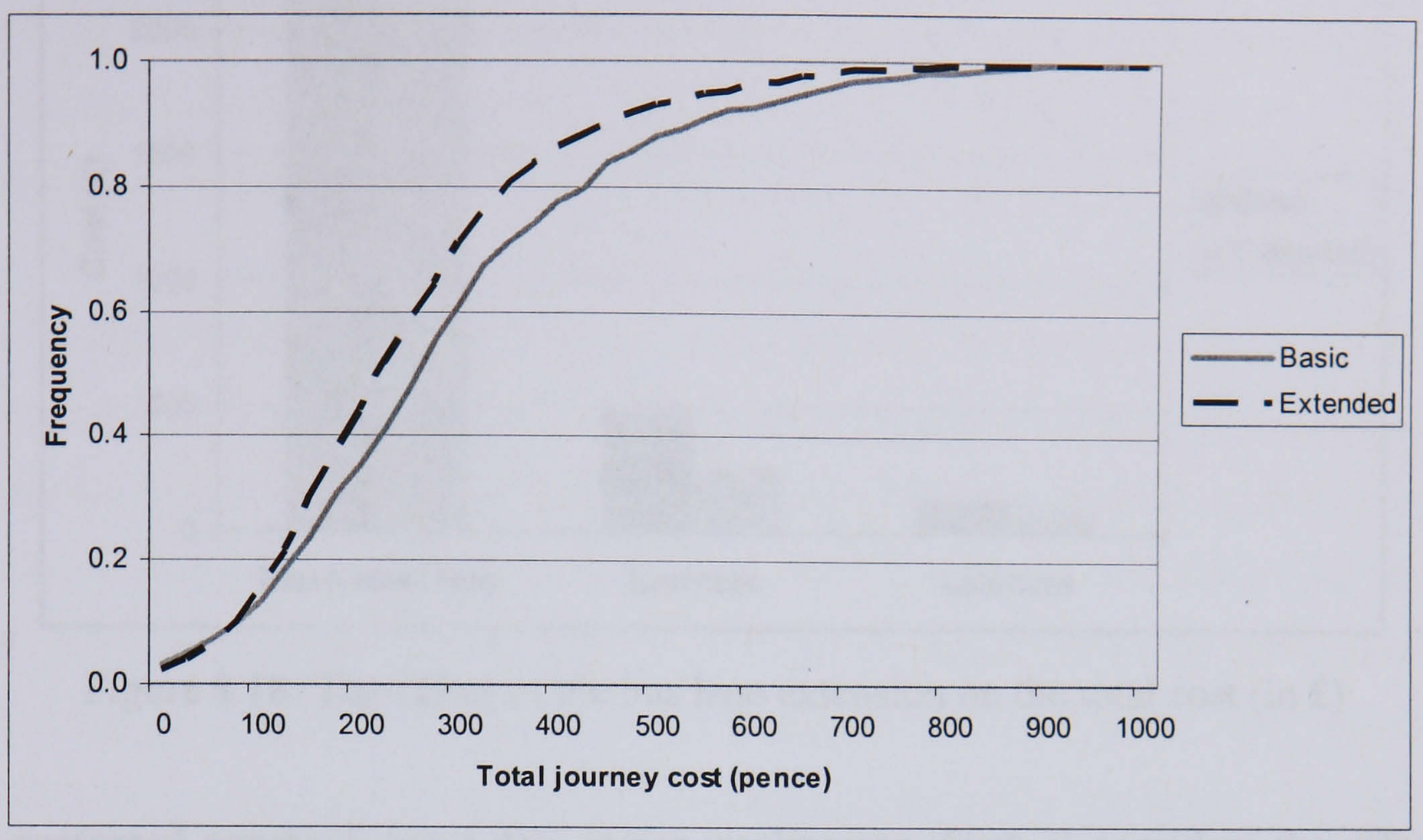


Figure 8.17: Cumulative frequency curve of the total individual cost

	Basic	Extended	Change
Mean travel time	£2232 (78.4%) (41.2 minutes)	£2103 (85.7%) (39.3 minutes)	-5.8%
Earliness	£481 (16.9%) (19.8 minutes)	£260 (10.6%) (11.3 minutes)	-46%
Lateness	£133 (4.7%) (0.8 minutes)	£92 (3.7%) (0.5 minutes)	-20.4%
Total	£2847 (100%)	£2455 (100%)	-13.8%

Table 8.6: The effect of the bus lane extension on the total user-time cost during the morning peak (in brackets: mean value per traveller in minutes)

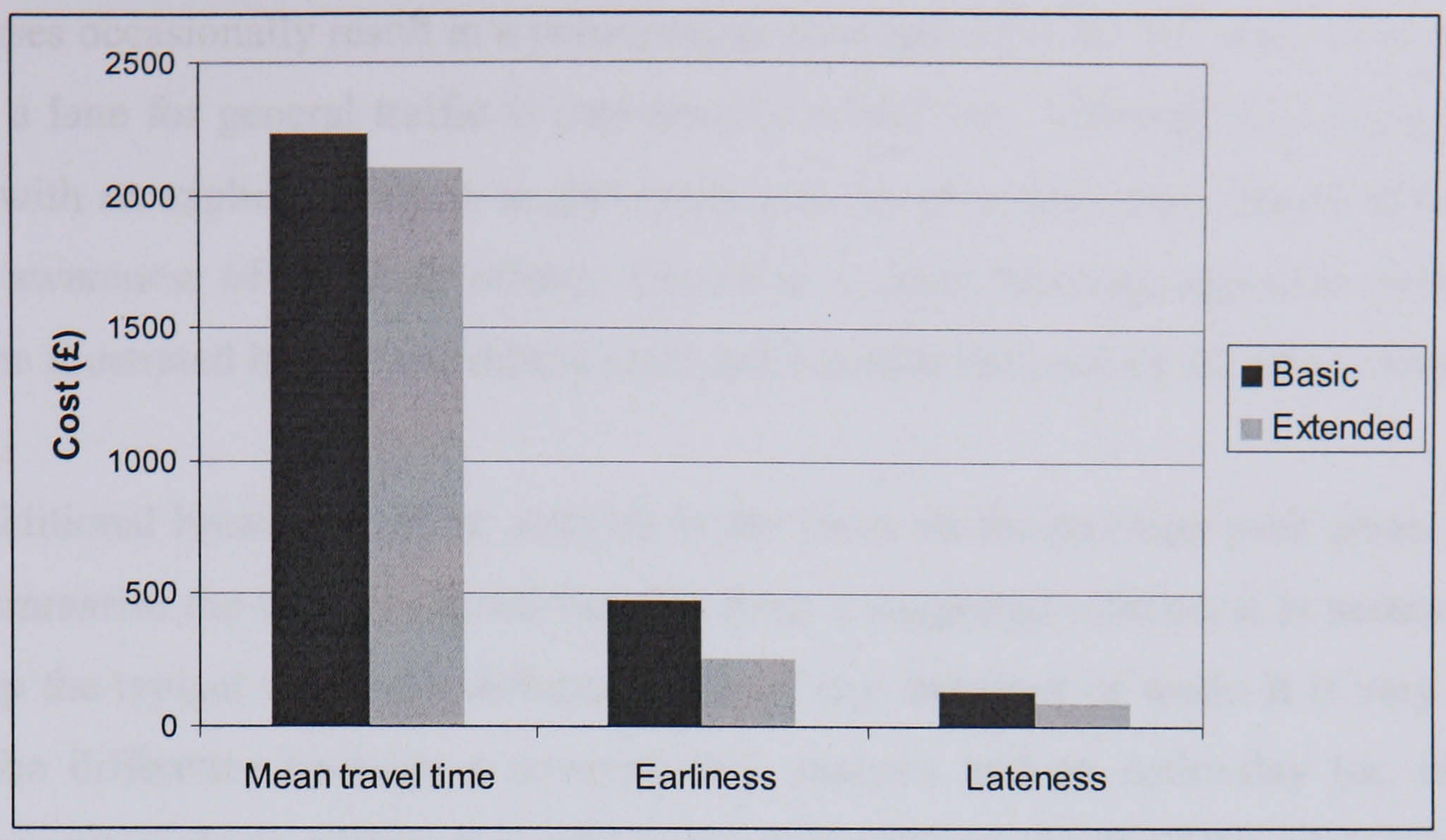


Figure 8.18: The effect of the bus lane extension on the total cost (in £)

The presented approach has a few major weaknesses. First, it considers the effect of TTV on DTC, which we believe is the main effect, but it does not consider other effects that surely exist, such as the way TTV influences mode and route choice. If a change in the level of TTV causes some travellers to shift to other modes or to change their trip itineraries, the equilibrium pattern that should be sought when estimating the cost of TTV is more complicated than the one described here, because there are complex interplays between the different choices that different travellers make. The choice of some travellers to avoid a particular mode of transport or a specific location in the network might lead other travellers to choose the very same mode or location, or to affect the choices of the users of adjacent elements of the transport system; a realistic model of systemwide choice making should be able to accommodate these dynamics. This also relates to the flexibility of the fundamental decision of whether or not to travel: our model assumes that the total number of users is fixed, but in reality, changes in reliability (or in other travel conditions) are also likely to induce new demand, or in other cases, to reduce existing demand. All in all, it would be necessary, in future studies, to allow the repeated loop of decision making, that follows the estimation of the level of TTV in each iteration, to include a more comprehensive choice model.

Another limitation of the proposed methodology is that it only accounts for the effects of a bus scheme on the journey cost for bus users. In practice, launching new infrastructure for buses is often followed by changes in road configuration that directly affect the generalised cost for car users. The attempts to improve the journey conditions

for buses occasionally result in a poorer travel experience for car travellers, for example when a lane for general traffic is converted to a bus lane. Although this is sometimes done with an explicit intention to give buses priority over cars, there should at least be some awareness of all likely effects. Therefore, a more thorough appraisal study than the one illustrated here must address costs and benefits incurred by all affected network users.

An additional limitation of our analysis is the focus on the morning peak period only. To summarise the total costs and benefits from a suggested scheme, it is necessary to add up the typical values for different times of day and days of week. It is very likely that the difference between a morning-only analysis and an entire-day (or, ideally, entire-year) analysis will be not only in the order of magnitude of the total costs, but also in the proportion of the costs of MTT and of TTV out of the total. The DTC considerations that our case study concentrates on only apply in the morning commuting trip, in which the main concern for most travellers is about the desired arrival time at work. In trips on the way back from work, or trips for non-commuting purposes, DTC behaviour is expected to follow different rules, which need to be analysed with a different choice model. It is not possible to estimate how the summation of costs and benefits across multiple periods will alter the conclusions from the morning-only illustration presented here. Clearly, the intensity of travel demand during the morning peak guarantees that the costs of the morning commute will have a major influence on any multi-period analysis; but the exact extent of such influence should be studied in more detail in the future.

It would also be useful to examine whether the estimated costs are seriously affected by our choice of DRACULA as the main tool for modelling network performance. Naturally, the sub-models in DRACULA have features that differ from other tools (such as VISSIM, Paramics and others). For instance, bus capacity constraints and a detailed representation of the individual alighting stop of each passenger are modelled in the version used here in quite a simplistic manner. DRACULA is also simpler than some other packages in the sense that its current version does not have dynamic route choice features. Undertaking similar analysis with other software can help identifying elements of a network model that the cost of TTV is sensitive to.

Despite these weaknesses, our finding concerning direct consideration of reliability costs in the evaluation of a proposed scheme still seems meaningful. Our case study implies that a relatively minor scheme, which gives a separate right of way to buses

along a 500-meter-long corridor, results in a reduction of about 14% in the total user-time cost for the passengers of the analysed route in the morning peak. We find that the introduction of a bus lane not only reduces travel time and its variability within the bus lane boundaries, but also helps to maintain the regular headway along the entire route, and thus prevents the creation of further delays after the bus has left the bus lane. The saving from the bus lane scheme, as calculated here, is more than double the respective saving that would have been implied if only the reduction in the MTT had been accounted for. This finding confirms the concern raised in the introduction to this thesis, that by not including TTV considerations in the common appraisal practice, a major source of benefit is not revealed.

The direct implication of this is that justifying investment in a transport scheme without taking TTV into account might result in a decision to embark on the wrong project. If two alternatives of a proposed schemed are compared, and one of them has a slight advantage over the other in the expected MTT savings, it is likely to be the selected option even if the other option can considerably reduce TTV. The fact that for many years the benefits from improved reliability have not played a role in appraisal studies has probably influenced not only the decisions about specific projects, but the formation of transport policies in general, because the focus on the reduction of MTT is also a focus on the traditional forms of transport investment. Schemes that explicitly aim at improving reliability, such as various control measures or passenger information systems, might still be consistently undervalued. We conclude that if the results of the case study presented here truly represent the outcome of other potential investments in bus infrastructure, then there is a genuine need to rethink the way different time elements compose the total cost savings in many current appraisal studies.

Chapter 9

Conclusions and suggested extensions

9.1. The contribution of this thesis

The reviews, discussions and experiments described in this thesis tackled various problems that transport analysts face when they wish to estimate the benefits from improved bus reliability. We focused on the variability of travel times as a key indicator of unreliability. In a nutshell, the following are the main themes in which this study has added to the experience and the knowledge gathered by others:

1. The study presented estimates of the distribution of the willingness of bus users to pay for reduction in the extent of TTV.
2. The study proposed a methodology for predicting the extent of TTV in hypothetical scenarios, in a way that is sensitive to local factors such as the detailed configuration of the transport network.
3. The study illustrated an estimation framework that combines the two abovementioned methodologies and can be used to incorporate TTV considerations in scheme appraisal.

To the best of the author's knowledge, no previous studies explicitly addressed any of these three issues. Note that these three areas of contribution correspond to the key objectives of this thesis, as specified in section 1.4. In this concluding chapter we bring together some of the main conclusions reached throughout the thesis and point to several potential directions for future research. The following paragraphs elaborate on the three abovementioned areas in which this thesis has gone farther than previous works.

The most straightforward contribution of this study is the values suggested for converting information about the bus journey of an individual into generalised cost estimates, which include the cost of the discomfort caused by TTV. A complete set of such values for bus users did not appear in the literature previously. While seeking these values, we have found new evidence for the advantages of modelling the impact of TTV *indirectly*, through the resulting patterns of earliness and lateness. Such evidence was needed to extend an ongoing discussion in literature, and particularly for public transport users, whose response to TTV has not been sufficiently studied before. We

have also brought evidence for the valuation bias caused by the common use of the *direct* approach for modelling the effects of TTV. We find that this approach underestimates the significance of TTV costs as a proportion of the overall journey cost, and therefore, using it does not do justice the true benefits from improved reliability. The discussion of this issue here might help preventing such undervaluation in future assessment.

On the way to these monetary estimates, we have also introduced new ideas regarding the presentation of the concept of TTV in stated preference (SP) experiments. The surveying methodology suggested here uses a relatively light display in terms of the amount of numerical information presented, and incorporates improved graphical features. This might assist in alleviating the cognitive burden on the participants of such experiments in the future, and hence it might improve the power and fit of future models.

The author is not aware of any previous studies of the attitudes to TTV that take random variations in tastes and preferences between travellers into consideration. We have paid much attention to the derivation of the distribution of the willingness-to-pay (DWP) across different travellers. The thesis does not propose innovative theory in this area, but it presents findings that contribute to the discussion of practical issues in the derivation of the DWP from SP surveys and from choice models, not solely for TTV valuation. We bring in some important insights with respect to the distinction between optimised and non-optimised elements in the specification of Mixed Logit models, and some arguments about different approaches for constraining the distributions of the random parameters. In addition, we raise the need for new statistical tests in the process of estimating the DWP; traditional tools such as the maximum simulated likelihood are efficient and powerful, but as we demonstrate, they can sometimes lead us to prefer an imperfect model specification. We also illustrate here that when comparing different estimates of a distribution, it is important to test not only the somewhat-obvious vertical dimension (of the cumulative frequency curve), but also the horizontal one, as this is likely to detect major errors that the DWP is prone to.

The thesis expands the discussion, which has only recently been launched, of nonparametric estimation of the DWP. It demonstrates how a sub-sampling technique can be involved in the process of estimating the DWP, but also shows that previous studies that used such technique for a similar purpose might have not employed it appropriately, because they used it before the specification of a Mixed Logit model and

then allowed re-estimation of the parameters. The thesis also describes a simple nonparametric experiment which investigates what is the amount of information about the DWP that is explicitly based on the data from the SP survey. An important contribution of this experiment is the conclusion that the ability to reach proper identification of the DWP is related to the dimensionality of the SP survey, and that this must be accounted for in the survey design.

Our analysis of the DWP also contributes to the debate on the more general issue of how to interpret evidence for the existence of negative willingness-to-pay for attributes related to travel time. Unlike previous studies, that either support or contradict the general case for potential negative willingness-to-pay, we find that for some time-related variables (such as the mean earliness or lateness), a small share of negative values should not be as strictly rejected as for the mean or general travel time.

In a research area that at first glance might seem irrelevant to the issues listed above, this study points to the lack of available tools for estimating the level of TTV in hypothetical scenarios. Tools that can generate TTV forecasts, in a way that is sensitive to the network configuration and meets the needs that come up in scheme appraisal, are almost nonexistent. We have discussed the case for using traffic microsimulation for the analysis of variations in the performance of a transport network, owing to the fundamental role that randomness and heterogeneity have in the concept of microscopic modelling. In particular, this study has introduced the concept of establishing an analogy between a single run of a traffic microsimulation model (TMM) and a single day in the real network; such analogy can open an avenue for estimating the level of TTV by analysing variations between runs of the TMM.

As part of our search through the literature in network modelling, we have defined a list of key principles that should be addressed whenever a TMM is to be calibrated. These principles have to do with the scope of the calibration problem, its formulation and automation, the choice of measures of fit, the number of model runs, the dimensionality of the TMM outputs and more. Subsequently, we have developed a full algorithm for TMM calibration. The algorithm attempts to make the distribution of simulated bus travel times resemble the respective distribution in the real transport system. In the development of this algorithm we have tackled various methodological, computational and statistical issues. To enable solution of the calibration problem, this thesis also introduces a modification of the Downhill Simplex Method of optimisation. The modification is suitable for problems in which the objective function is

nondifferentiable and takes a long time to estimate. Although we do not provide a rigorous theoretical background for this modification, we show that according to our experience, it may save precious runtime.

The last methodological section of the thesis proposes a procedure for joint application of the previously-developed economic tool and network tool in an appraisal framework. This iterative procedure seeks a stable pattern of departure time choices and derives the cost of the mean travel time and the cost of unreliability, which can then be compared with the respective costs in other scenarios. Procedures that directly compute the cost of TTV, and applications of departure time choice models in general, are very uncommon in previous literature. We applied this methodology to assess the benefit to the users of a major bus route following the introduction of a bus lane along a 500-meter-long street section, close to the city centre in York. This application suggests that economic saving from the bus scheme, including the contribution of the effects of reduced TTV, is more than double the respective saving calculated in the traditional way that only looks at reduction in the mean travel time. This finding has serious policy implications: it enhances our fear that by ignoring reliability benefits, policy-makers only see part of the full picture of benefits from investment in new infrastructure, and are very likely to allocate money for the wrong schemes. Although these results are based on a limited study, they do indicate that more attention should be given both to schemes that directly aim at improving the reliability of public transport services, and to the secondary reliability effects of schemes that do not see reducing of TTV as their main objective.

9.2. Suggestions for further research

Many of the issues that this thesis raises deserve a more thorough and focused investigation, which was not possible in the current scope. Some of these issues are hereby listed as potential themes for further research.

In the design of the SP survey, we attempted to make the introduction and the presentation of the idea of TTV easy to understand, but have not proved that our chosen way of presentation is optimal. It would be useful to carry out comparison between responses collected using different formats of questionnaires with TTV attributes. In the modelling experiments we have found that the amount of information in the survey database was not sufficient; this emphasizes the need for developing more efficient

formats and surveying techniques, which can be used to extract more information from each respondent, and still avoid the undesirable effects of boredom or fatigue.

Our attempts to estimate the DWP through a Mixed Logit model included a discussion of ways to constrain the distributions of the model parameters, but some of the available ways of controlling these distributions were not attempted. These include the use of various truncated distributions, as well as the entire concept of modelling in the willingness-to-pay space. It is important to investigate whether these alternatives would be able to lead to more reliable and powerful estimates of the DWP. Further discussion is also needed concerning some of the approaches that this study did try, namely constraining by estimation and by imposition, or using distributions that are constrained by definition. There is a need to examine whether we can reach better specification or better identification of models based on any of these concepts. Our analysis also suggests that new tools are needed for testing the fit of Mixed Logit models. Development and testing of such tools in future research would make an important contribution, as the tools we use today have not been purposely designed to be used with Mixed Logit specifications. Furthermore, there is a broad scope to further explore the case for nonparametric estimation of the DWP. The techniques used here, as well as many other nonparametric techniques, bring in the appealing feature of estimation that does not strongly depend on preliminary assumptions. But using these techniques involves new mathematical challenges that future research needs to tackle.

Our experiments in economic modelling only examined the choices made on a commuting trip during the morning peak. Departure time choice considerations would be different on trips made for business or leisure purposes, or in other parts of the day, because constraints such as those imposed here on the arrival time to the destination would be replaced with others. The cost of TTV in other periods but the morning peak has hardly been studied so far, although it is most necessary for realistic estimation of the benefits from improved reliability.

Our choice model is rather simplistic in the sense that it assumes travellers choose the option with the lowest cost. Some recent studies use an alternative assumption to enable more realistic account of habitual behaviour; they assume travellers have a preferred choice that is not necessarily optimal, and that they change this choice only if the added benefit exceeds some threshold. It would be interesting to estimate the cost of TTV and the choice of departure time using such model.

In the area of TMM calibration, there is a clear need to test the effect of the various model parameters on TTV, and the sensitivity of calibration outputs to the values of the different parameters. These are important to assure that when it is not possible to calibrate all model parameters, a smaller-scale calibration effort would be able to concentrate on the most influential ones. The calibration parameters that this study has focused on were not selected in a systematic way, and this means that they do not necessarily represent a general set that we can recommend to use in other studies of simulating TTV. Another reason why the generality of our results should be investigated further is that the methodologies discussed here in the context of network modelling have only been implemented with DRACULA. Although the presented methods were devised as generic concepts, which should apply equally to any TMM, the results might be sensitive to particular features that DRACULA lacks, such as more detailed modelling of the boarding and alighting patterns at bus stops.

Despite the microscopic nature of TMMs, not all causes for TTV in the real transport system are modelled. While it is important to carry out adequate calibration, it is also necessary to continuously improve the explanatory power of the model itself, and keep seeking ways to realistically account for various traffic phenomena. In the analysis of bus performance, in particular, there is a need for more realistic representation of various causes for service cancellation, some of which are related to mechanisms in the public transport operation industry and not necessarily to the traffic itself. There is also need to improve the modelling of the cumulative effects (including phenomena such as bus bunching) of relatively minor delays.

The illustration of the joint application of the demand and supply tools only examined a relatively minor case study. There is great interest in carrying out assessment of the cost of TTV in a more thorough scenario analysis. It is essential to check the consequences of new bus infrastructure not only on bus passengers but also on the other network users, as there might be cases in which the cost for bus users falls but the cost for car users rises. As implied above, this should be examined not only for morning commuters but for the general population of travellers. Moreover, the fact that TTV affects the generalised cost of travel, as demonstrated in this thesis, means that improved reliability can make travellers change not only their choice of departure time but also their mode or route; such changes can indirectly affect the behaviour of users of all modes in all parts of the transport system. It is therefore important to study the impact of changes in

the level of TTV in a fully-elastic intermodal model that allows for these diverse interactions.

When the broader systemwide impacts of TTV are taken into account, it would be important to re-assess the benefits from potential measures of policy intervention. We believe that such re-assessment can strengthen the case for introducing public transport priority measures similar to those examined in this study, and for other measures such as intelligent transport systems and provision of advanced information. We anticipate that a broader model will help establishing our evidence of benefits from improved TTV because it will reveal effects not examined here, like a shift of some demand from cars to public transport.

A key incentive for the entire research presented here is the current lack of tools that are needed to incorporate TTV considerations in scheme appraisal. The tools presented here constitute a step forward in this issue, but additional or alternative steps need to follow. There is a wide scope for developing other estimation instruments that take the dynamic effects of TTV on both supply and demand into account. For instance, there are presently very few documented applications of models of departure time choice that capture the attitudes to unreliability, and very few available tools to choose from for TTV prediction. Moreover, there is still insufficient awareness to the role that the application of such tools should have in practical appraisal. Further work on all these will help making future public transport services more reliable and transforming urban areas into a more vigorous and sustainable environment.

References

1. Abdel-Aty, M. A., Kitamura, R. & Jovanis, P. P. (1995), **Investigating Effect of Travel Time Variability on Route Choice Using Repeated-Measurement Stated Preference Data**, Transportation Research Record, No. 1493, pp. 39-45.
2. Adler, H. A. (1971), Economic Appraisal of Transport Projects: A Manual with Case Studies, Indiana University Press, Great Britain.
3. Andersson, P. A., Hermansson, A., Tengveld, E. & Scalia-Tomba, G. P. (1979), **Analysis and Simulation of an Urban Bus Route**, Transportation Research A, Vol. 13, pp. 439–466.
4. Atkins Consultants LTD (1997), Bus Reliability Study – Stated Preference Research, Great Britain.
5. Avineri, E. & Prashker, J. N. (2002), **Sensitivity to Travel Time Variability: Travellers' Learning Perspective**, paper presented at the 13th mini-EURO conference Handling Uncertainty in the Analysis of Traffic and Transportation Systems held in Bari, Italy.
6. Banister, D. & Berechman, J. (2000), **Economic Evaluation of Transport Projects**, Transport Investment and Economic Development, pp. 161-210, UCL Press, Great Britain.
7. Barcelo, J. & Casas, J. (2004), **Methodological Notes on the Calibration and Validation of Microscopic Traffic Simulation Models**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.
8. Bates, J., Dix, M. & May, T. (1987), **Travel Time Variability and its Effect on Time of Day Choice for the Journey to Work**, Transportation Planning Methods, Proceedings of seminar C held at the PTRC Summer Annual Meeting, University of Bath, Vol. P290, pp. 293-311.
9. Bates, J., Jones, P. & Polak, J. (1995), The Importance of Punctuality and Reliability: A Review of Evidence and Recommendations for Future Work, Research report, Transport Studies Group, University of Westminster, Great Britain.
10. Bates, J., Polak, J., Jones, P. & Cook, A. (2001), **The Valuation of Reliability for Personal Travel**, Transportation Research E, Vol. 37, pp. 191-229.

11. Batley, R., Fowkes, T. & Whelan, G. (2001), **Models for Choice of Departure Time**, Paper presented at the European Transport Conference, Homerton College, Cambridge.

12. Bell, M. G. H. & Cassir, C. (2000), **Introduction**, Reliability of Transport Network, Research Studies Press, Great Britain.

13. Beesley (1973), M. E., Urban Transport: Studies in Economic Policy, Butterworths, London, Great Britain.

14. Ben-Akiva, M., Bolduc, D. & Bradley, M (1993), **Estimation of Travel Choice Models with Randomly Distributed Values of Time**, Transportation Research Record, No. 1413, pp. 88-97.

15. Ben-Akiva, M. E., Darda, D., Jha, M., Koutsopoulos, H. N. & Toledo, T. (2004), **Calibration of Microscopic Traffic Simulation Models with Aggregate Data**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.

16. Ben-Akiva, M. & Lerman, S. T. (1985), Discrete Choice Analyis: Theory and Application to Travel Demand, The MIT Press, Cambridge, Massachusetts.

17. Bhat, C. R. (2001), **Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model**, Transportation Research B, Vol. 35, pp. 677-693.

18. Bhat, C. R. & Sardesai, R. (2005), **On Examining the Impact of Stop-Making and Travel Time Reliability on Commute Mode Choice: An Application to Predict Commuter Rail Transit Mode for Austin, TX**, Proceedings of the 84th TRB annual meeting, Washington, D.C.

19. Bierlaire, M., Bolduc, D. & Godbou, M. H. (2004), An Introduction to BIOGEME (Version 1.0), <*http://roso.epfl.ch/biogeme*>, downloaded 10th Feb 2005.

20. Black, I. G. & Towriss, J. G. (1993), Demand Effects of Travel Time Reliability, Centre for Logistics and Transportation, Cranfield Institute of Technology, Great Britain.

21. Bogers, E. A. I., Viti, F., Hoogendoorn, E. P. & van Zuylen, H. J. (2005), **Valuation of Different Types of Travel Time Reliability in routh Choice – A Large Scale Laboratory Experiment**, Proceedings of the 85th TRB annual meeting, Washington, D.C.

22. Bonsall, P., Liu, R. & Young, W. (2005), **Modelling Safety-Related Driving Behaviour – Impact of Parameter Values**, Transportation Research A, Vol. 39, pp. 425-444.

23. Bookbinder, J. H. & Desilets, A. (1992), **Transfer Optimization in a Transit Network**, Transportation Science, Vol. 26, pp. 106–118.

24. Brownstone, D. & Small, K. A. (2005), **Valuing Time and Reliability: Assessing the Evidence from Road Pricing Demonstrations**, Transportation Research A, Vol. 39, pp. 279-293.

25. Castiglione, J., Freedman, J. & Bradley, M. (2004), **Systematic Investigation of Variability due to Random Simulation Error in an Activity-Based Microsimulation Forecasting Model**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.

26. Caussade, S., Ortuzar, J. D. D., Rizzi, L. I. & Hensher, D. A. (2005), **Assessing the Influence of Design Dimensions on Stated Choice Experiment Estimates**, Transportation Research B, Vol. 38, pp. 621-640.

27. Chu, L., Liu, H. X., Oh, J. S. & Recker, W. (2004), **A Calibration Procedure for Microscopic Traffic Simulation**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.

28. Cirillo, C. (2005), **Optimal Design for Mixed Logit Models**, Paper presented at the European Transport Conference held in Strasbourg, France.

29. Cirillo, C., Daly, A. and Lindveld, K. (2000), **Eliminating Bias Due To the Repeated Measurements Problem in SP Data,** Ortúzar, J. D. (editor), Stated Preference Modelling Techniques: PTRC Perspectives 4, PTRC Education and Research Services.

30. Cirillio, C. and Axhausen, K. W. (2004), **Evidence on the Distribution of Values of Travel Time Savings from a Six-Week Diary**, Arbeitsbericht Verkehrs und Raumplanung, Vol. 212, IVT, ETH, Zurich.

31. Cohen, H. & Southworth, F. (1999), **On the Measurement and Valuation of Travel Time Variability due to Incidents on Freeways**, Journal of Transportation and Statistics, No. 2, pp. 123-131.

32. Cook, A. J., Jones, P., Bates, J. J., Polak, J. & Haigh, M. (1999), **Improved Methods of Representing Travel Time Reliability in SP Experiments**,

Transportation Planning Methods, Proceedings of seminar F held at the European Transport Conference, Homerton College, Cambridge, Vol. P434, pp. 37-49.

33. Dale, H. M., Porter, S. & Wright, I. (1996), **Are There Quantifiable Benefits from Reducing the Variability of Travel Times**?, Transportation Planning Methods, Proceedings of seminar E at the 24th European Transport Forum, Vol. P404, 15pp.

34. De Jong, G., Daly, A., Pieters, M., Vellay, C., Bradley, M. & Hofman, F. (2004), **A Model for Time of Day and Mode Choice using Error Components Logit**, Paper presented at the European Transport Conference held in Strasbourg, France.

35. Dowling, R., Skabardonis, A., Halkias, J., McHale, G. & Zammit, G. (2004), **Guidelines for Calibration of Microsimulation Models: Framework and Application**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.

36. Emam, B. E. & Al-Deek, H. A. (2005), **A New Methodology for Estimating Travel Time Reliability in a Freeway Corridor**, Proceedings of the 84th TRB annual meeting, Washington D.C.

37. Fosgerau, M. (2006), **Investigating the Distribution of the Value of Travel Time Savings**, Transportation Research B, Vol. 40, pp. 687-707.

38. Fosgerau, M. (2005), **Specification of a Model to Measure the Value of Travel Time Savings from Binomial Data**, Paper presented at the European Transport Conference held in Strasbourg, France.

39. Fowkes, T. & Preston. J. (1991), **Novel Approaches to forecasting the Demand for New Local Rail Services**, Transportation Research A, Vol. 25, pp. 209-218.

40. Fowkes, T. & Wardman, M. (1988), **The Design of Stated Preference Travel Choice Experiments with Special Reference to Interpersonal Taste Variations**, Journal of Transport Economics and Policy, Vol. 22, pp. 27-44.

41. Fowkes, A. S. & Watson, S. M. (1989), Sample Size Determination to Evaluate the Impact of Highway Improvement, Working Paper No. 282, Institute for Transport Studies, University of Leeds, Great Britain.

42. Galilea, P. & Ortuzar, J. D. D. (2005), **Valuing Noise Level Reduction in a Residential Location Context**, Transportation Research D, Vol. 10, pp. 305-322.

43. Gaver, D. P. Jr. (1968), **Headstart Strategies for Combating Congestion**, Transportation Science, Vol. 2, No. 2, pp. 172-181.

44. Georgi, H. (1973), Cost-Benefit Analysis and Public Investment in Transport: A Survey, Butterworths, Great Britain.

45. Giuliano, G. (1989), **Incident Characteristics, Frequency, and Duration on a High Volume Urban Freeway**, Transportation Research A, Vol. 23, pp. 387-396.

46. Glaister, S. (1981), **Economic Evaluation, Investment Criteria and Public Enterprise Objectives**, Fundalmentals of Transport Economics, pp. 140-163, Basil Blackwell Oxford, Great Britain.

47. Golob, T., Recker, W. & Leonard, J. (1987), **An Analysis of the Severity and Incident Duration of Truck Involved Freeway Accidents**, Accident Analysis and Prevention, Vol. 19, pp.375-395.

48. Guehthner, R. P. & Hamat, K. (1985), **Distribution of Bus Transit On-Time Performance**, Transportation Research Record, No. 1202, pp. 1–8.

49. Grant-Muller, S. M. (editor) (2000), Assessing the Impacts of Local Transport Policy Instruments, ITS Working Paper 549, University of Leeds, Great Britain.

50. Grant-Muller, S. M., Mackie, P., Nellthorp, J. & Pearman, A. (2001), **Economic Appraisal of European Transport Projects: the State-of-the-art Revisited**, Transport Reviews, Vol. 21, No. 2, pp. 237-261.

51. Hai, W. (2004), Introductory Statistics, School of computing, Dublin city University, *<http://www.compapp.dcu.ie/~wuhai/formulae.pdf>*, downloaded 20th Sept 2004.

52. Hall, R. W. (1983), **Travel Outcome and Performance: The Effect of Uncertainty and Accesibility**, Transportation Research B, Vol. 17, No. 4., pp. 275-290.

53. Hall, R. W. (1985), **Vehicle Scheduling at a Transportation Terminal with Random Delay En Route,** Transportation Science, Vol. 19, pp. 308–320.

54. Harrison, A. J. (1974), The Economics of Transport Appraisal, Croom Helm, Great Britain.

55. Hastings, N. A. J & Peacock (1975), Statistical Distributions, Butterworths, Great Britain.

56. Hendrickson, C. & Plank, E. (1984), **The Flexibility of Departure Times for Work Trips**, Transportation Research A, Vol. 18, No. 1, pp. 25-36.

57. Hensher, D. A. (1994), **Stated Preference Analysis of Travel Choices: the State of the Practice**. Transportation, Vol. 21, No. 2, pp. 107-133.

58. Hensher, D. A. (2004), **Identifying the Influence of State Choice Design Dimensionality on Willingness to Pay for Travel Time Savings**, Journal of Transport Economics and Policy, Vol. 38, Part 3, pp. 425-446.

59. Hensher, D. & Greene, W. H. (2003), **The Mixed Logit Model: The State of Practice**, Transportation, No. 30, pp. 133-176.

60. Herman, R. & Lam, T. (1974), **Trip Time Characteristics of Journeys To and From Work**, Transportation and Traffic Theory, pp. 57-86.

61. Hess, S., Bierlaire, M. & Polak, J. W. (2005), **Estimating of Value of Travel-Time Savings using Mixed Logit Models**, Transportation Research A, Vol. 39, pp. 221-236.

62. Hess, S., Polak, J. W., Daly, A. & Hyman, G. (2004), **Flexible Substitution Patterns in Models of Mode and Time of Day Choice: New Evidence from the UK and the Netherlands**, Paper presented at the European Transport Conference held in Strasbourg, France.

63. Hojman, P., Ortúzar, J. D. & Rizzi, L. (2004), **Internet-Based Surveys to Elicit the Value of Risk Reductions**, Proccedings of the 7th International Conference on Travel Survey Methods, Los Sueños, Costa Rica.

64. Hoogendoorn, S. P. & Ossen, S. (2005), **Parameter Estimation and Analysis of Car-Following Models**, Mahmassani, H. (editor), Transportation and Traffic Theory - Flow, Dynamics and Human Interaction, Proceedings of the 16th International Symposium on Transportation and Traffic Theory, pp. 245-265, Elsevier.

65. Hotchkiss, W. (1977), **Cost-Benefit Analysis**, Hensher, D. A. (editor), Urban Transport Economics, pp. 44-54, Cambridge University Press, Great Britain.

66. Hourdakis, J., Michalopoulos, P. G. & Kottommannil, J. (2003), **Practial Procedure for Calibrating Microscopic Traffic Simulation Models**, Transportation Research Record, No. 1852, pp. 130-139.

67. Hu., T.-Y. & Mahmassani, H. S. (1997), **Day-to-Day Evolution of Network Flows Under Real-Time Information and Reactive Signal Control**, Transportation Research C, Vol. 5, pp. 51-69.

68. Iraguen, P. and Ortuzar, J. D. D. (2004), **Willingness to Pay for Reducing Fatal Accident Risk in Urban Areas: an Internet-Based Web Page Stated**

Preference Survey, Accident Analysis and Prevention, Vol. 36, No. 4, pp. 513-524.

69. Jackson, W. B. & Jucker, J. V. (1982), **An Empirical Study of Travel Time Variability and travel Choice Behavior**, Transportation Science, Vol. 16, No. 4, pp. 460-475.

70. Jayakrishnan, R., Oh, J. S. & Sahraoui, A. E. K. (2001), **Calibration and Path Dynamics Issues in Microscopic Simulation for Advanced Traffic Management and Information Systems**, Transportation Research Record, No. 1771, pp. 9-17.

71. Johnson, N. L. (1949), **Systems of Frequency Curves Generated by Methods of Translation**, Biometrika, Vol. 36, No. 1 / 2, pp. 149-176.

72. Kaczmarczyk, G. (1999), Downhill Simplex Method for Many (~20) Dimensions, <*http://paula.univ.gda.pl/~dokgrk/simplex.html*>, downloaded 21st March 2006.

73. Kendall, B. (2003), **Confidence Intervals**, Data Analysis for Environmental Science and Management, Donald Bren School of Environmental Science & Management - University of California, Santa Barbara, <*http://www.bren.ucsb.edu/academics/course.asp?number=206*>, downloaded 20th Sept 2004.

74. Killi, M. & Nossum, A. (2003), **Internet-Based Stated Preference Surveys – Studies in Traffic Information and Public Transport**, Paper presented at the 4th International Conference on Survey and Statistical Computing, Universtiy of Warwick, Great Britain.

75. Kim, K. O. & Rilett, L. R. (2003), **Simplex-Based Calibration of Traffic Microsimulation Models with Intelligent Transportation Systems Data**, Transportation Research Record, No. 1855, pp. 80-89.

76. Kim, K. O. & Rilett, L. R. (2004), **A Genetic Algorithm Based Approach to Traffic Micro-simulation Calibrating Using ITS Data**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.

77. Kim, S. J., Kim, W. & Rilett, L. R. (2005), **Calibration of Micro-Simulation Models using Non-Parametric Statistical Techniques**, Proceedings of the 84th TRB annual meeting, Washington, D.C.

78. Kleijnen, J. P. C. (1987), Statistical Tools for Simulation Practitioners, Marcel Dekker Inc., New York.

79. Kleijnen, J. P. C. (1995), **Verification and Validation of Simulation Models**, European Journal of Operational Research, Vol. 82, pp. 145-162.

80. Knight, T. E. (1974), **An Approach to the Evaluation of Changes in Travel Unreliability: A "Safety Margin" Hypothesis**, Transportation, No. 3, pp. 393-408.

81. Laidler, J. (1999), Economic Evaluation of Bus Priority Corridors, Internal memo, Greater Manchester Public Transport Executive.

82. Lam, T. (2000), **Route and Scheduling Choice under Travel Time Uncertainty**, Transportation Research Record, No. 1725, pp. 71-78.

83. Lam, T. C. & Small, K. A. (2001), **The value of Time and Reliability: Measurement from a Value Pricing Experiment**, Transportation Research E, Vol. 37, pp. 231-251.

84. Linaritakis, K., **Evaluating the Regularity and Reliability Benefits of Bus Priority Schemes: the London Experience**, Transportation Planning Methods, Proceedings of seminar E at the 23th European Transport Forum, Vol. P392, pp. 249-269.

85. Liu, H. X., Recker, W. & Chen, A. (2004), **Uncovering the Contribution of Travel Time Reliability to Dynamic Route Choice using Real-Time Loop Data**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.

86. Liu, R (2006), DRACULA Microscopic Traffic Simulator, Version 2.3, Working paper, Institute for Transport Studies, University of Leeds, Great Britain.

87. Liu, R., Van Vliet, D. & Watling, D. (2006), **Microsimulation Models Incorporating Both Demand and Supply Dynamics**, Transportation Research A, Vol. 40, pp. 125-150.

88. Ma, T. & Abdulhai, B. (2002), **Genetic Algorithm-Based Optimization Approach and Generic Tool for Calibrating Traffic Microscopic Simulation Parameters**, Transportation Research Record, 1800, pp. 8-15.

89. Mackie, P. J. & Nellthorp, J. (2001), **Cost-Benefit Analysis in Transport**, Button, K. J. & Hensher, D. A. (editors), Handbook of Transport Systems and Traffic Control, Elsevier Science.

90. Mackie, P. J., Wardman, M., Fowkes, A. S., Whelan, G., Nellthorp, J. & Bates, J. (2003), Values of Travel Time Savings in the UK, Report to the Department for

Transport, Institute for Transport Studies, University of Leeds in association with John Bates Services, Great Britain.

91. Mahmassani, H. S. & Stephan, D. G. (1988), **Experimental Investigation of Route and Departure Time Choice Dynamics of Urban Commuters**, Transportation Research Record, No. 1203, pp. 69-84.

92. Maisel, H. & Gnugnoli, G. (1972), Simulation of Discrete Stochastic Systems, Science Research Associates Inc., Chicago.

93. May, A. D. & Montgomery, F. O. (1984), Factors Affecting Travel Times on Urban Radial Routes, ITS working paper 177, University of Leeds, Great Britain.

94. May, A. D., Bonsall, P. W. & Marler, N. W. (1989), Travel Time Variability of a Group of Car Commuters in North London, ITS Working Papers 277 to 279, University of Leeds, Great Britain.

95. McFadden, D. & Train, K. (2000), **Mixed MNL Models for Discrete Response**, Journal of Applied Econometrics, No. 15, pp. 447-470.

96. Mei, M. & Bullen, A. G. R. (1993), **Lognormal Distribution for High Traffic flows**, Transportation Research Record, No. 1398, pp. 125-128.

97. Merritt, E. (2004), **Calibration and Validation of CORSIM for Swedish Road Traffic Conditions**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.

98. Mogridge, M. & Fry, S. (1984), **Variability of Car Journey Times on a Particular Route in Central London**, Traffic Engineering and Control, No. 25, pp. 510-511.

99. Mohammadi, R. (1997a), **Journey Time Variability in the London Area – 1. Journey Time Distribution**, Traffic Engineering and Control, Vol. 38, No.5, pp. 250-256.

100. Mohammadi, R. (1997b), **Journey Time Variability in the London Area – 2. Factors Affecting Journey Time Variability**, Traffic Engineering and Control, Vol. 38, No.6, pp. 337-346.

101. Mora Camino, F. A. C, Lopes Pereira, A. & Pinheiro Moreira, M. E. (1986), **A New Dispersion Model for Traffic Flows**, Road Traffic Control, IEE Conference Publication 260, pp. 140-144.

102. Nagel, E. (1996), **Particle Hopping Model and Traffic Flow Theory**, Physics Review, Vol. E53, pp. 4655-4672.

103. Nash, C. A. (1993), **Cost-Benefit Analysis of Transport Projects**, Oum, T. H., Dodgson, J. S., Hensher, D. A., Morrison, S. A., Nash, C. A., Small, K. A. & Waters, W. G. II (editors), Transport Enocomics: Selected Readings, pp. 627-651.

104. Noland, B. N. (1997), **Commuter Responses to Travel Time Uncertainty under Congested Conditions: Expected Costs and the Provision of Information**, Journal of Urban Economics, No. 41, pp. 377-406.

105. Noland, R. B. & Polak, J. W. (2002), **Travel Time Variability: A Review of Theoretical and Empirical Issues**, Transport Reviews, Vol. 22, No. 1, pp. 39-54.

106. Noland, R. B. & Small, K. A. (1995), **Travel Time Uncertainty, Departure Time Choice, and the Cost of Morning Commutes**, Transportation Research Record, No. 1493, pp. 150-158.

107. Noland, R. B., Small, K. A., Koskenoja, P. M. & Chu, X. (1998), **Simulating Travel Reliability**, Regional Science and Urban Economics, No. 28, pp. 535-564.

108. Oketch, T. & Carrick, M. (2005), **Calibration and Validation of a Micro-Simulation Model in Network Analysis**, Proceedings of the 84th TRB annual meeting, Washington, D.C.

109. Park, B. & Schneeberger, J. D. (2003), **Microscopic Simulation Model Calibration and Validation – Case Study of VISSIM Simulation Model for a Coordinated Signal System**, Transportation Research Record, No. 1856, pp. 185-192.

110. Park, B. & Qi, H. (2005), **Development and Evaluation of Simulation Model Calibration Procedure**, Proceedings of the 84th TRB annual meeting, Washington, D.C.

111. Pells, S. (1987a), **The Evaluation of Reductions in the Variability of Travel Times on the Journey to Work**, Transportation Planning Methods, Proceedings of Seminar C held at the PTRC Summer Annual Meeting, University of Bath, Vol. P290, pp. 313-325.

112. Pells, S. (1987b), The Evaluation of Reductions in Travel Time Variability, Ph.D. thesis in Economics, University of Leeds, Great Britain.

113. Polak, J. (1987a), Travel Time Variability and Departure Time Choice: A Utility Theoretic Approach, Discussion Paper No. 15, Transort Studies Group, Polytechnic of Central London, Great britain.

114. Polak, J. (1987b), **A More General Model of Individual Departure Time Choice**, Transportation Planning Methods, Proceedings of seminar C held at the PTRC Summer Annual Meeting, University of Bath, Vol. P290, pp. 247-258.

115. Polus, A. (1978), **Modeling and Measurements of Bus Service Reliability**, Transportation Research, Vol. 12, pp. 253-256.

116. Polus, A. (1979), **A Study of Travel Time and Reliability on Arterial Routes**, Transportation, pp. 141-151.

117. Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992), **Minimization or maximization of function**, Numerical Recipes in C (second edition), Cambridge University Press, New York, pp. 394-455.

118. Prashker, J. N. (1979), **Direct Analysis of the Perceived Importance of Attributes of Reliability of Travel Modes in Urban Travel**, Transportation, Vol. 8, pp. 329-346.

119. Rao, L. & Owen, L. (2000), **Validation of High-Fidelity Traffic Simulation Models**, Transportation Research Record, No. 1710, pp. 69-78.

120. Revelt, D. & Train, K. (1998), **Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level**, The Review of Economics and Statistics, Vol. 80, No. 4, pp. 647-657.

121. Revelt, D. & Train, K. (1999), Customer-Specific Taste Parameters and Mixed Logit, <*http://elsa.berkeley.edu/wp/train0999.pdf*>, downloaded 5th June 2006.

122. Richardson, A. & Taylor, M. (1978), **Travel Time Variability on Commuter Journeys**, High Speed Ground Transportation Journal, Vol. 12, pp. 77-99.

123. Rohr, C., Daly, A., Hess, S., Bates, J., Polak, J. & Hyman, G. (2005), **Modelling time Period Choice: Experience from the UK and the Netherlands**, Paper presented at the European Transport Conference held in Strasbourg, France.

124. Scheffe, H. (1970), **Practical Solutions of the Behrens-Fisher Problem**, Journal of the American Statistical Association, Vol. 65, No. 332, pp. 1501-1508.

125. Schweiger, C. L. & Marks, J. B. (1999), **Cost/Benefit Analysis for Transit ITS: Using Available Data to Yield Credible Results**, Proceedings of the 6th World Congress on Intelligent Transport Systems Held at Toronto.

126. Senna, L. A. D. S. (1994a), User Response to Travel Time Variability, Ph.D. thesis in Civil Engineering, University of Leeds, Great Britain.

127. Senna, L. A. D. S. (1994b), **The Influence of Travel Time Variability on the Value of Time**, Transportation, No. 21, pp. 203-228.

128. Shaaban, K. S. & Radwan, E. (2005), **A Calibration and Validation Procedure for Microscopic Simulation Model: A Case Study of SimTraffic Arterial Streets**, Proceedings of the 84th TRB annual meeting, Washington, D.C.

129. Siegrist, K. (2004), **Basic Statistics - Interval Estimation**, Virtual Laboratories in Probability and Statistics, <*http://www.fmi.uni-sofia.bg/vesta/Virtual_Labs/index.html*>, downloaded 20th Sept 2004.

130. Sinha, S. (2004), Modelling Public Transport Reliability – Case Study: York, M.Sc. dissertation in Transport Studies, University of Leeds, Great Britain.

131. Small, K. A. (1982), **The Scheduling of Consumer Activities: Work Trips**, American Economic Review, No. 72, pp. 467-479.

132. Small, K. A. (1987), **A Discrete Choice Model for Ordered Alternatives**, Econometrica, Vol. 55, No. 2, pp. 409-424.

133. Small, K. A., Noland, R., Chu, X. & Lewis, D. (1999), Valuation of Travel-Time Savings and Predictability in Congested Conditions for Highway User-Cost Estimation, NCHRP Report 431, Transportation Research Board, U.S.

134. Small, K. A., Winston, C. & Yan, J. (2005), **Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability**, Econometrica, Vol. 73, No. 4, pp. 1367-1382.

135. Sørensen, M. V. (2003), Demand Choice Models – Estimation of Passenger Traffic, Ph.D. Thesis, Centre for Traffic and Transport, Technical University of Denmark, Denmark.

136. Sørensen, M. V. & Nielsen, O. A. (2003), **MSL for Mixed Logit Model Estimation – On Shape of Distributions**, Paper presented at the European Transport Conference held in Strasbourg, France.

137. Steer Davies Gleave (2001), Northen Orbital Quality Bus Corridor, Prepared for Greater Manchester Public Transport Esecutive.

138. Stephenson, D. B. (2004), Data Analysis Methods in Weather and Climate Research, Department of Meteorology, University of Reading, <*http://www.met.rdg.ac.uk/cag/courses/*>, downloaded 20th Sept 2004.

139. Strathman, J. G. & Hopper, J. R. (1993), **Empirical Analysis of Bus Transit On-Time Performance**, Transportation Research A, Vol. 27, pp. 93–100.

140. Talley, W. K. & Becker, A. J. (1987), **On-time Performance and the Exponential Probability Distribution**, Transportation Research Record, No. 1198, pp. 22–26.

141. Taylor, M. A. P. (1982), **Travel Time Variability – the Case of Two Public modes**, Transportaion Science, Vol. 16, pp. 507-521.

142. Thijs, R. & Lindveld, Ch. D. R. (1999), **Impacts of Traffic Control Tools in 'Daccord'**, Proceedings of the 6th World Congress on Intelligent Transport Systems Held at Toronto.

143. Thorburn-Colquhoun (1995), Bus Journey Time Reliability Study – Stage 1, Research report, United Kingdom.

144. Toledo, T. & Koutsopoulos, H. N. (2004), **Statistical Validation of Traffic Simulation Models**, Proceedings of the 83rd TRB annual meeting, Washington, D.C.

145. Toledo, T., Koutsopoulos, H. N., Davol, A., Ben-Akiva, M. E., Burghout, W., Andreasson, I., Johansson, T. & Lundin, C. (2003), **Calibration and Validation of Microscopic Traffic Simulation Tools – Stockholm Case Study**, Transportation Research Record, No. 1831, pp. 65-75.

146. Train, K. (2001), A Comparison of Hierarchical Bayes and Maximum Simulated Likelihood for Mixed Logit, *<http://elsa.berkeley.edu/~train/compare.pdf>*, downloaded 5th June 2006.

147. Train, K., Revelt, D. & Ruud, P. (1999), Mixed Logit Estimation Routine for Panel Data, *<http://elsa.berkeley.edu/Software/abstracts/train0296.html>*, downloaded 1st June 2005.

148. Train, K. and Sonnier, G. (2003) **Mixed Logit with Bounded Distributions of Partworths,** Alberini, A. and Scarpa (eds.), Applications of Simulation Methods in Environmental Resource Economics, Kluwer Academic Publisher.

149. Train, K. & Weeks, M. (2005), **Discrete Choice Models in Preference Space and Willingness-to-Pay Space**, Alberini, A. and Scarpa, R. (eds.), Applications of Simulation Methods in Environmental Resource Economics, Kluwer Academic Publisher, The Netherlands, pp. 1-16.

150. TRL (2004), The Demand for Public Transport: a Practical Guide, Report 593, TRL, Great Britain.

151. Turnquist, M. A. (1978), **A Model For Investigating the Effects of Service Frequency and Reliability on Bus Passenger Waiting Times**, Transportation Research Record, No. 663, pp. 70-73.

152. Van Lint, J. W. C. & van Zuylen, H. J. (2005), **Monitoring and Predicting Freeway Travel Time Reliability,** Proceedings of the 84th TRB annual meeting, Washington, D.C.

153. Vickrey, W. S. (1969), **Congestion Theory and Transport Investment**, The American Economic Review, Vol. 59, No. 2, pp. 251-260.

154. Vovsha, P., Petersen, E. & Donnelly, R. (2002), **Microsimulation in Travel Demand Modeling – Lessons Learned from the New York Best Practice Model**, Transportation Research Record, No. 1805, pp. 68-77.

155. Walton, D., Thomas, J. A. & Cenek, P. D. (2004), **Self and Others' Willingness to Pay for Improvements to the Paved Road Surface**, Transportation Research A, Vol. 38, pp. 483-494.

156. Walker, J. (2002), **Mixed Logit (or Logit Kernel) Model – Dispelling Misconceptions of Identification**, Transportation Research Record, No. 1805, pp. 86-98.

157. Wardman, M. (2004), **Public Transport Values of Time**, Transport Policy, Vol. 11, pp. 363-377.

158. Yang, Q. & Koutsopoulos, H. N. (1996), **A Microscopic Traffic Simulator for Evaluation of Dynamic Traffic Management Systems**, Transportation Research C, Vol. 4, pp. 113-129.

159. Yatchew, A. (1998), **Nonparametric Regression Techniques in Economics**, Journal of Economic Literature, Vol. 36, pp. 669-721.

Appendix A

Models for car and rail users

In parallel with the bus user survey described in chapter 3, similar questionnaires were also presented to some car and rail users. When the link to the survey website was circulated, the identity of the recipients was unknown, and it was therefore not possible to address only bus users. Thus, since users of all modes were contacted, collecting responses from car and rail users, in addition to bus users, did not incur any additional cost. It was decided to also estimate models for those travellers mainly in order to enable comparison with the bus user model and to check its reasonableness. However, only basic analysis of the car and rail user data was possible in the current scope, hence only Multinomial Logit models were estimated, and there was no further analysis of taste heterogeneity.

The sample of non-bus users included responses from 290 car users and, unfortunately, only 20 rail users. Each respondent answered nine questions of a similar structure to the one described in chapter 3. Due to the small number of respondents that commute by rail, the rail user models are primarily judged by common sense and not necessarily by the measures of statistical performance (such as t-test). As mentioned above, the rail model is presented here for the comparative analysis; due to the small sample it is not recommended to use it for other purposes.

The same variables whose potential contribution to the bus user model was examined in chapter 3 were tried again for the car and rail user models. It was found again that the mean travel time, earliness and lateness are sufficient in capturing the response to TTV. For both car and rail users it was found that a direct TTV variable remains significant while the scheduling variables are not included, but does not improve the power of the model once the lateness and earliness variables are introduced.

As in the bus user model, model performance was found better when the mean travel time and earliness were included in the same variable, MTE. Again, this does not prove that the penalties placed on these two separate elements are equal; but it implies that they are close to each other enough to make their treatment as one variable (which makes the model easier to identify) contribute to the statistical significance of the model more than accounting for them separately.

The car and rail user models as presented in table A.1, together with the bus user model of a similar format. Values of the t-statistic appear in brackets. The WTP derived from these models is also presented in figure A.1.

		Car	Bus	Rail
Coefficients	Cost	-0.6996 (-18.1)	-1.375 (-14.2)	-0.1739 (-3.4)
	MTE	-0.05209 (-10.0)	-0.07173 (-11.5)	-0.03229 (-3.2)
	ML	-0.2315 (-5.6)	-0.1974 (-4.1)	-0.1147 (-1.3)
Likelihood	Initial	-1985	-1534	-124
	Final	-1726	-1369	-116
WTP (pence per minute)	MTE	7.4	5.2	18.6
	ML	33.1	14.4	66.0
	Ratio ML / MTE	4.5	2.8	3.5

Table A.1: Departure time choice models for car, bus and rail users

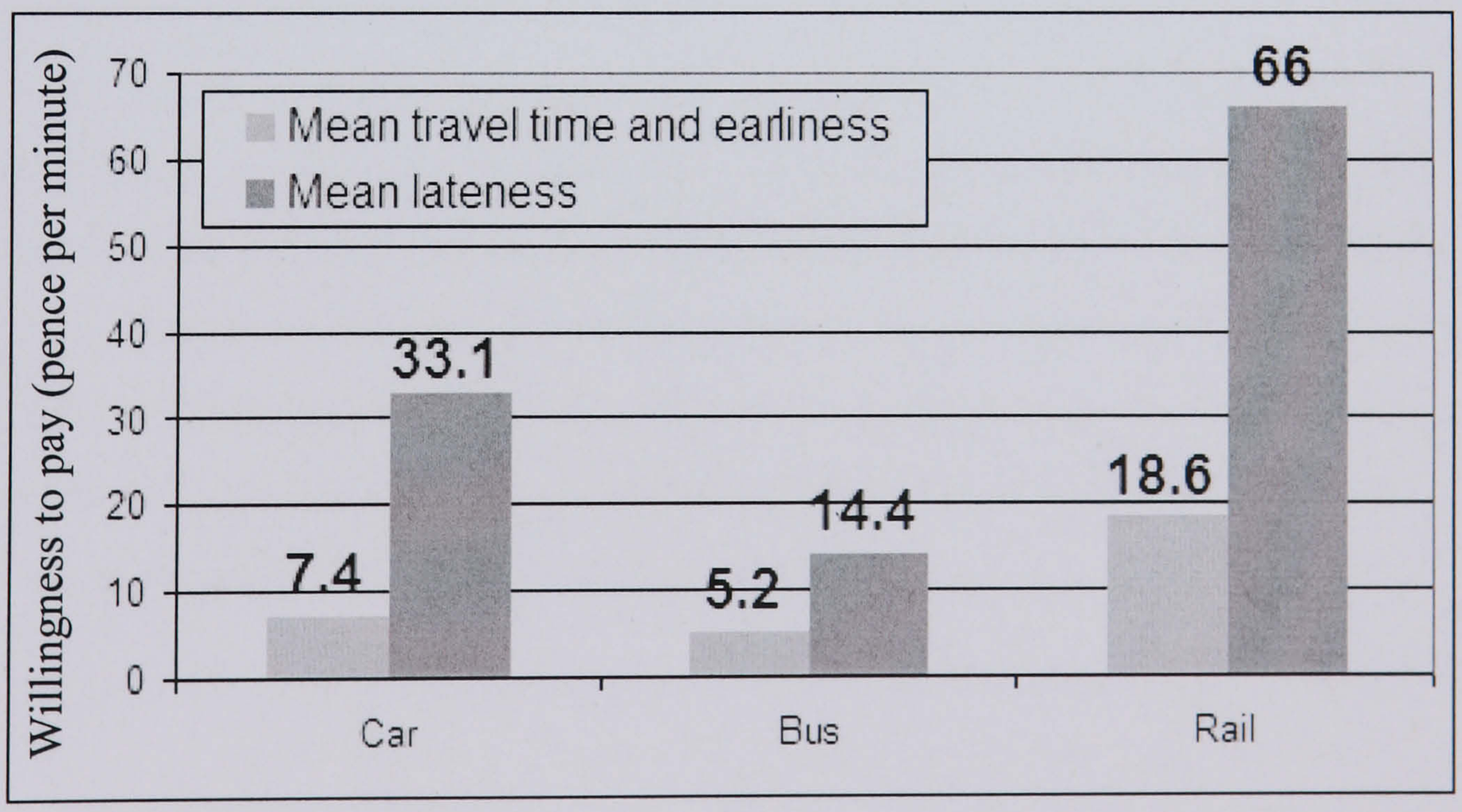


Figure A.1: The willingness of car, bus and rail users to pay for reduced travel time, earliness and lateness

It is not surprising that car users' WTP is found higher than the respective WTP among bus users. It is striking, though, that while the value placed by car users on MTE is 42% higher than for bus users, the value placed on lateness is 130% higher. It seems that bus users do not only exhibit reduced WTP, but they are also less sensitive to lateness, relative to the mean travel time.

The monetary values for rail users are much higher than for car users. This might seem unusual if compared to results from other countries (see, for instance, Dutch study by De Jong et al, 2004, where the WTP among car users is higher than among rail users). Nevertheless, a high level of WTP among rail users is consistent with the findings of recent works that bring inter-modal evaluation of the value of time for travellers in the United Kingdom (Wardman, 2004; TRL, 2004). The rail user model shows that the sensitivity to late arrival is not strictly proportional to the WTP: the number of rail MTE minutes that is equivalent to one minute of lateness (3.5) is higher than for bus users (2.8) bus lower than for car users (4.5).

All in all, the relations between the WTP estimates for the three modes seem plausible. The fact that some of the conclusions reached in the analysis of the bus user data are also found valid for other travellers adds some more confidence about our previous findings.

Appendix B

Structures used in the calibration procedure

The TMM calibration algorithm, developed and applied in chapters 6 and 7, uses various data structures for inputting and outputting information, and for interlinking with either the TMM used (DRACULA) or the user itself. The main data structures are briefly explained here as a supplementary description of the calibration methodology.

The **.rng* file

1	0.1	60	1	5
2	0.1	60	0.2	2
3	0.1	600	20	40
4	0.1	600	40	80
5	0.1	60	1	5
6	0	2	0	0.3
7	0.1	60	2	5
8	0	2	0	0.3
9	0.1	60	1.5	5
10	0	2	0	0.3
11	0.1	60	3.5	6.5
12	0	2	0	0.3
13	0.1	60	0.8	2
14	0	2	0	0.3
15	0.1	60	0.8	2
16	0	2	0	0.3
17	0.1	60	1	4
18	0	2	0	0.3
19	0.1	60	1	4
20	0	2	0	0.3
21	0	0.25	0.01	0.2

The **.rng* file is introduced to define the *feasible range* and the *likely range* for each of the calibration parameters. The parameter index (from 1 to 21) is in the first column. The lower and upper boundaries of the *feasible range* are in the second and third columns. The lower and upper boundaries of the *likely range* are in the forth and fifth columns. All boundaries in the **.rng* file were determined based on commonly used values of the various calibration parameters (see for instance Bonsall et al, 2005, which was one of the main sources).

The *.*par* and *.*rst* files

```
PARAMETERS
 TMAIN=60
 TWARM=5
 TCOOL=5
 TOUTPUT=15
 START_HOUR=07
 START_MIN=55
 NSEED=1200
 GAP=3.0
 GAP_MIN=0.5
 GAP_TSTART=30
 GAP_TEND=60
 FGW_CAR=0.5
 FGW_BUS=0.5
 CIRC_SPEED=30
 QTRAJ=T
 PTRAJ=50
 PHGV=1.0
 QSATNET=T
 QBUS=T
 QNEWDEMAND=T
 PECAR(1)=20
 PECAR(2)=10
 PECAR(3)=20
 PECAR(4)=10
 PECAR(5)=10
 PECAR(6)=10
 PECAR(7)=5
 PECAR(8)=5
 PETRUCK(11)=10
 PETRUCK(12)=90
END
```

The *.*par* file is a standard DRACULA file. Dozens of parameters can be defined using this file, of which the current experiments only focus on a few.

Four of the parameters in the file are calibrated directly:

1. GAP – the normal acceptable gap – it is stored as parameter 1 in the simplex.
2. GAP_MIN – the minimum acceptable gap – parameter 2 in the simplex.
3. GAP_TSTART – the time waited before accepting a lower gap – parameter 3 in the simplex.
4. GAP_TEND – the time waited before accepting the minimum gap – parameter 4 in the simplex.

An additional parameter that is relevant to our experiment is GONZO, a factor by which the entire travel demand is multiplied. We do not wish to calibrate the parameter itself but its standard deviation, which stands for the extent of daily fluctuations in the level of demand; DRACULA does not have an explicit parameter that represents these fluctuations. We store this parameter as parameter 21 in the simplex, and a random

number based on it is written to the GONZO field in the **.par* file whenever this file is being created. The mean value of GONZO remains 1 throughout the whole process.

A random seed number is also defined as a parameter in the **.par* file. Such number is re-generated every time the file is being created.

The **.rst* file (*rst* stands for "the rest") is not a DRACULA file; we use it to store all non-calibration parameters, which are kept fixed throughout the process. The **.rst* file looks exactly like the **.par* file, except that it excludes the lines containing the six parameters mentioned above.

The *veh.tab* file

```
&VEH_PARAM
CAR
 4.50  1.00  1.00  2.69  2.82  1.66  4.84  1.00  1.00
 0.10  0.10  0.00  0.08  0.06  0.22  0.01  0.15  0.10
 3.50  0.80  1.00  1.00  2.00  1.50  3.50  0.80  0.80
 5.50  1.20  1.00  5.00  5.00  5.00  6.50  2.00  2.00
BUS
 7.50  1.00  1.00  1.64  1.69  1.42  3.36  1.00  0.50
 0.10  0.10  0.00  0.02  0.09  0.07  0.00  0.10  0.10
 5.00  0.80  1.00  0.80  0.80  1.00  1.00  0.10  0.20
10.00  1.20  1.00  2.00  2.00  4.00  4.00  1.50  2.00
&END
```

The *veh.tab* file is a standard DRACULA file. For each vehicle type, the user can define in this file the mean (first row), coefficient of variation (second row), and feasible range (third and forth rows) of nine different vehicle characteristic (each described in a different column). In the current experiment we focus on two vehicle types only (car and bus), and use eight of the vehicle parameters of each one, expressing the mean and the coefficient of variation of the following features:

- Normal acceleration
- Maximum acceleration
- Normal deceleration
- Maximum deceleration

All in all, 16 vehicle parameters are calibrated; they are stored as parameters 5 to 20 of the simplex.

When the *veh.tab* file is read, the entire set of car and bus parameters, and not only the calibration parameters, is input into two arrays (one for car parameters and the other for bus parameters). This makes it easier at later stages to output the entire *veh.tab* file in its desired format.

The **.spx* file

1	3.00	0.50	30.00	60.00	1.50	0.10	2.50	0.10	2.00	0.10	5.00	0.10	1.50	0.10	1.60	0.10	1.50	0.10	2.50	0.10	0.15	0.367
2	3.99	1.32	36.37	48.02	2.44	0.26	3.74	0.02	1.78	0.15	4.71	0.06	1.54	0.25	0.83	0.04	3.05	0.28	2.36	0.03	0.11	0.350
3	4.87	0.87	23.26	40.58	3.59	0.13	2.96	0.29	2.56	0.15	5.69	0.12	1.52	0.14	0.80	0.01	1.86	0.24	2.02	0.01	0.05	0.333
4	1.45	0.25	23.74	53.16	2.69	0.08	2.82	0.06	1.66	0.22	4.84	0.01	1.64	0.02	1.69	0.09	1.42	0.07	3.36	0.00	0.12	0.356
5	3.63	1.83	24.95	59.08	2.35	0.18	3.22	0.13	2.10	0.16	5.37	0.17	0.87	0.08	1.88	0.17	1.61	0.13	1.73	0.19	0.18	0.415
6	4.95	1.11	25.38	75.04	1.65	0.06	2.48	0.17	3.26	0.26	6.09	0.13	1.97	0.13	1.20	0.15	1.51	0.05	3.15	0.16	0.18	0.123
7	1.88	1.71	36.36	41.64	4.69	0.17	2.05	0.27	4.28	0.19	3.72	0.03	1.90	0.28	1.44	0.03	1.23	0.29	2.67	0.17	0.06	0.278
8	3.30	1.25	26.17	70.19	2.42	0.07	3.39	0.21	1.96	0.08	4.89	0.02	1.79	0.07	1.02	0.27	3.81	0.20	2.09	0.07	0.06	0.298
9	1.68	0.54	38.94	74.24	4.36	0.13	3.16	0.26	4.34	0.27	5.90	0.10	1.39	0.20	1.25	0.19	2.78	0.21	1.78	0.14	0.03	0.364
10	3.41	1.86	23.30	66.05	4.73	0.16	3.05	0.29	3.61	0.17	4.89	0.15	1.15	0.13	1.31	0.15	1.15	0.06	3.50	0.12	0.02	0.398
11	2.03	1.94	29.57	72.00	1.70	0.00	2.44	0.18	2.33	0.23	5.86	0.22	1.01	0.09	1.78	0.01	2.83	0.25	3.24	0.12	0.11	0.316
12	1.01	1.67	38.01	54.17	2.37	0.04	4.61	0.06	4.12	0.22	3.54	0.16	1.58	0.23	1.45	0.20	1.44	0.24	2.45	0.06	0.08	0.317
13	4.31	0.74	37.70	58.21	4.55	0.24	2.42	0.19	3.77	0.18	4.73	0.13	1.07	0.02	1.85	0.18	1.77	0.26	3.56	0.06	0.15	0.422
14	2.37	1.39	35.85	59.51	3.93	0.01	3.43	0.27	2.01	0.27	6.14	0.22	1.73	0.23	1.54	0.13	1.11	0.08	2.05	0.08	0.02	0.365
15	2.06	0.25	25.37	54.81	2.15	0.13	2.96	0.00	2.00	0.17	5.34	0.04	1.87	0.03	1.60	0.07	1.32	0.06	1.57	0.28	0.05	0.265
16	1.94	1.72	33.20	48.12	3.30	0.11	2.29	0.08	4.38	0.02	5.47	0.22	1.93	0.27	1.90	0.02	2.23	0.19	1.99	0.24	0.02	0.365
17	1.24	1.99	31.80	63.26	1.99	0.29	3.11	0.10	4.30	0.22	4.02	0.18	1.40	0.15	1.05	0.27	1.43	0.26	2.78	0.05	0.09	0.354
18	2.21	0.45	20.49	57.58	2.43	0.24	3.03	0.15	4.58	0.04	5.30	0.04	0.90	0.09	1.38	0.02	3.47	0.29	2.83	0.09	0.18	0.357
19	4.88	0.99	37.55	73.79	2.87	0.27	3.47	0.23	4.57	0.25	5.62	0.29	1.33	0.29	1.61	0.26	1.91	0.10	3.97	0.22	0.05	0.273
20	4.68	0.85	21.61	66.44	2.41	0.27	4.44	0.21	3.62	0.19	4.85	0.29	1.40	0.02	1.17	0.10	3.15	0.18	3.16	0.26	0.09	0.465
21	1.25	1.26	25.67	46.65	4.00	0.04	3.42	0.23	2.24	0.01	5.47	0.16	1.50	0.06	1.48	0.01	2.13	0.27	1.09	0.20	0.04	0.373
22	4.22	1.36	36.00	44.92	3.06	0.23	2.80	0.02	2.30	0.15	3.55	0.19	1.07	0.28	1.87	0.07	2.14	0.12	1.99	0.03	0.18	0.376

A simplex is a multidimensional geometrical shape, employed in procedures such as ours to symbolize a set of decision variables. At each stage of the process, the simplex represents the group of all candidate solutions; each vertex of the simplex stands for a feasible set of values for the calibration parameters, and a value of the objective function corresponds to it.

Each row of the **.spx* file describes one of the vertices of the simplex. Each column, except the first and the last columns, stands for the location of the vertex in a different dimension; each such dimension represents one of the calibration parameters. Parameters 1-4 are gap acceptance parameters, stored in DRACULA in the **.par* file. Parameters 5-20 are vehicle characteristics, stored in DRACULA in the *veh.tab* file. Parameter 21 is a demand fluctuation parameter; it is not stored anywhere in DRACULA, but a demand factor, drawn randomly from a distribution defined by this parameter, is stored in the **.par* file.

The first column in each row of the **.spx* file is the vertex index. The last column in each row shows the value of the objective function that corresponds to the vertex describes in that row. In each stage of the calibration process, the row that includes the vertex with the highest (i.e. worst) objective is changed.

The **.pas* file

Outputs of a simulation run (in DRACULA as well as most other TMMs) include multiple files that list various traffic measures or statistical measures in different formats. Our procedure reads the simulated travel time measurements from the **.pas* file, which is a standard DRACULA output. The **.pas* file is set up as a list of the times when individual buses reach their stops (given in seconds from the beginning of the simulation period). Since the main purpose of our calibration experiments is to examine their potential use in the future in the analysis of bus TTV, it seemed natural to looks directly at bus travel times.

Network: YORK
Summary of Passenger Delays and Bus Dwell Time

Time	Vehicle	Service	Bus_Stop	NPsg	PDelay(m)	TDwell(s)
189	86	5	111	0	0.00	0.00
203	86	5	222	0	0.00	0.00
261	86	5	333	0	0.00	0.00
394	86	5	444	3	2.45	17.00
635	151	5	111	14	36.87	57.00
698	1865	5	111	2	0.97	17.00
737	151	5	222	18	63.60	77.00
743	1865	5	222	0	0.00	0.00
837	1869	5	111	6	7.20	29.00
869	151	5	333	23	108.87	97.00
874	1865	5	333	0	0.00	0.00
892	1869	5	222	6	7.90	29.00
968	159	5	111	5	5.92	25.00
1017	1869	5	333	6	8.40	29.00
1019	159	5	222	6	6.20	25.00
1133	159	5	333	5	5.67	25.00
1151	151	5	444	32	205.87	133.00
1155	1865	5	444	0	0.00	0.00
1270	181	5	111	13	34.02	0.00
1289	1869	5	444	6	6.80	29.00
1294	1870	5	111	1	0.22	61.00
1337	1870	5	222	13	32.93	0.00
1352	181	5	222	0	0.00	61.00
1400	159	5	444	4	3.93	21.00
1425	1870	5	333	12	31.20	53.00
1440	181	5	333	1	0.25	9.00
1608	1864	5	111	13	34.45	57.00
1630	1870	5	444	10	20.17	45.00
1634	181	5	444	0	0.00	0.00
1649	213	5	111	2	0.67	13.00
1694	213	5	222	15	43.25	0.00
1698	1864	5	222	0	0.00	65.00
1792	213	5	333	15	43.75	65.00
1797	1864	5	333	0	0.00	0.00
1834	1862	5	111	7	11.32	33.00
1896	1862	5	222	8	13.20	37.00
1978	1862	5	333	7	11.32	33.00
2002	213	5	444	15	48.25	65.00
2017	1864	5	444	1	0.27	9.00
2043	1861	5	111	9	17.10	0.00

The format of the **.pas* file is relatively convenient for our needs, but still, after reading a set of such files from a series of simulation runs, a complete reorganisation of the data is carried out, for several reasons:

1. The file contains information about travel times in one simulation run, representing one day, whereas for the analysis of TTV we need a series of travel times from multiple days, i.e. we need to combine many such files into a single data structure.
2. The file provides absolute time readings at particular points, whereas we compare route segment travel times, i.e. time differences between points.
3. The file gives a sequence of time readings for the entire simulation period, while our analysis uses several shorter sub-periods. We only compare simulated travel times from any particular sub-period to observed travel times from the same sub-period.
4. In any sub-period on any single simulation day, multiple buses perform a similar journey (given that the frequency is high enough), but our observed time measurements include no more than one measurement per sub-period per day. Segment travel times from each simulation day therefore need to be averaged.

The **.obs* file

The **.obs* file is the medium of inputting real-life, observed bus travel time data into the procedure. In our calibration experiments, each scenarios is represented by a different **.obs* file. The first column in the file gives the time, in seconds after the beginning of the analysed time period. The second column contains the sub-period when the bus has departed from its origin. The third column includes the index of the bus stop where the bus has arrived at the mentioned time point. The forth columns contains the day index. The file is organised such that all timed stops along the route of a daily bus journey, with a particular scheduled departure, are brought in succession. Following is the respective data of the same scheduled journey from the next available day of measurement, and so on. After the entire series of multi-day time measurements of a particular scheduled journey, data of other journeys are recorded in a similar manner.

```
113     1       111     1
179     1       222     1
299     1       333     1
467     1       444     1
93      1       111     2
154     1       222     2
271     1       333     2
417     1       444     2
99      1       111     3
163     1       222     3
266     1       333     3
463     1       444     3
95      1       111     4
156     1       222     4
280     1       333     4
463     1       444     4
95      1       111     5
159     1       222     5
265     1       333     5
453     1       444     5
105     1       111     6
162     1       222     6
265     1       333     6
470     1       444     6
116     1       111     7
178     1       222     7
277     1       333     7
456     1       444     7
92      1       111     8
151     1       222     8
244     1       333     8
446     1       444     8
98      1       111     9
```

The **.prg* file

The **.prg* file gives information about the progress of the calibration process. It is being constantly updated during the process. The file reports on every candidate vertex that comes up throughout the calibration, and shows why it has or has not been chosen, as well as the type of simplex manipulation used.

```
The simplex at the beginning of iteration 9:
1  3.55  2.41 37.36 54.66  4.61  0.05  4.96  0.11  4.98  0.18  4.65  0.29  1.47  0.29  1.65  0.05  1.29  0.
2  2.99  1.75 37.52 50.92  4.28  0.07  3.88  0.17  3.77  0.15  4.51  0.17  1.64  0.24  1.82  0.11  1.35  0.
3  2.23  1.42 37.84 29.59  2.76  0.15  3.03  0.04  4.06  0.20  3.73  0.20  1.19  0.25  2.04  0.02  0.36  0.
4  2.80  1.96 42.05 35.14  2.32  0.16  2.52  0.08  3.94  0.15  3.94  0.16  1.42  0.24  1.78  0.06  1.48  0.
5  4.04  1.81 34.29 43.32  2.39  0.06  3.41  0.12  3.05  0.10  4.98  0.17  1.56  0.21  1.59  0.01  0.96  0.
6  2.82  1.83 33.53 47.08  3.56  0.07  4.48  0.19  3.38  0.25  4.06  0.19  1.52  0.16  1.93  0.08  0.10  0.
7  4.34  1.14 34.43 50.47  2.61  0.01  4.78  0.05  4.57  0.19  3.29  0.21  1.28  0.20  1.40  0.00  0.46  0.
8  4.59  2.50 32.66 20.68  2.32  0.08  3.09  0.15  4.93  0.25  3.39  0.05  1.77  0.20  2.27  0.07  0.10  0.
9  2.92  1.81 34.76 50.58  3.28  0.11  3.38  0.11  3.86  0.14  4.61  0.18  1.56  0.21  1.71  0.08  1.11  0.
10  3.08  1.71 30.68 31.74  4.72  0.12  4.24  0.10  4.35  0.21  3.75  0.03  1.11  0.24  1.48  0.01  2.27  (
11  4.96  1.05 33.94 46.78  3.01  0.18  2.48  0.30  2.99  0.23  3.76  0.18  1.64  0.15  1.27  0.25  1.22  (
12  3.95  2.50 33.58 46.91  3.44  0.21  3.33  0.09  5.89  0.18  2.89  0.21  1.23  0.13  1.50  0.15  0.10  (
13  1.88  2.17 43.90 29.66  3.02  0.17  2.91  0.05  4.80  0.22  3.20  0.11  1.17  0.27  1.94  0.01  1.44  (
14  4.57  1.74 32.45 65.34  4.80  0.00  5.08  0.14  3.37  0.29  1.76  0.25  1.87  0.11  2.32  0.06  1.40  (
15  2.08  1.89 38.06 49.82  3.37  0.15  2.33  0.21  4.52  0.20  4.03  0.17  1.46  0.17  1.73  0.00  2.67  (
16  2.70  1.24 37.29 44.20  1.95  0.00  3.70  0.16  3.68  0.07  3.51  0.28  1.44  0.04  1.75  0.12  1.45  (
17  2.89  1.82 40.62 51.63  3.49  0.14  3.07  0.12  5.00  0.20  3.59  0.22  1.41  0.25  1.77  0.10  0.92  (
18  2.27  1.79 35.95 65.10  3.79  0.09  3.74  0.17  3.20  0.20  4.95  0.12  1.46  0.22  1.79  0.04  1.43  (
19  2.44  1.20 35.54 45.47  2.60  0.11  3.43  0.04  3.13  0.17  4.26  0.24  1.30  0.20  1.93  0.02  0.89  (
20  2.57  2.44 32.03 73.66  3.51  0.07  3.88  0.01  4.90  0.29  4.93  0.07  1.52  0.14  2.09  0.11  1.48  (
21  2.99  1.81 41.58 49.64  3.54  0.13  3.07  0.14  4.54  0.20  3.53  0.21  1.43  0.23  1.87  0.09  0.89  (
22  3.04  1.96 41.85 51.95  4.36  0.19  3.33  0.03  4.91  0.10  4.87  0.13  1.30  0.31  1.61  0.13  1.18  (

Trying simple reflection:
      Candidate objective -    0.275
      Worst objective =    0.250
      Second worst objective -    0.244
      Lowest objective -    0.192

Trying modification with factor   0.5 :
      new candidate objective -    0.233

Vertex 3 was changed to -
  2.72  1.63 36.99 38.72  3.07  0.13  3.30  0.08  4.12  0.19  3.83  0.18  1.32  0.23  1.91  0.05  0.76  0.2
A factor of 0.5 was used.
The new objective value of this vertex is    0.233
End of iteration 9.


The simplex at the beginning of iteration 10:
1  3.55  2.41 37.36 54.66  4.61  0.05  4.96  0.11  4.98  0.18  4.65  0.29  1.47  0.29  1.65  0.05  1.29  0.
2  2.99  1.75 37.52 50.92  4.28  0.07  3.88  0.17  3.77  0.15  4.51  0.17  1.64  0.24  1.82  0.11  1.35  0.
3  2.72  1.63 36.99 38.72  3.07  0.13  3.30  0.08  4.12  0.19  3.83  0.18  1.32  0.23  1.91  0.05  0.76  0.
```

The table of observed versus simulated times

A table of the following format is created at the final stage of preparations towards the calculation of the objective function, every time it is being calculated.

From	To	Period	Day	Obs	Sim
111	222	1	1	81	53
111	222	1	2	41	67
111	222	1	3	63	59
111	222	1	4	31	70
111	222	1	5	62	67
111	222	1	6	47	54
111	222	1	7	58	57
111	222	1	8	38	62
111	222	1	9	71	58
111	222	1	10	68	63
111	222	1	11	38	60
111	222	1	12	37	62
111	222	1	13	56	61
111	222	1	14	36	59
111	222	1	15	67	52
111	222	1	16	47	59
111	222	1	17	85	63
111	222	1	18	59	61
111	222	1	19	53	62
111	222	1	20	83	66
222	333	1	1	122	119
222	333	1	2	53	104
222	333	1	3	81	106
222	333	1	4	90	139
222	333	1	5	105	136
222	333	1	6	110	98
222	333	1	7	108	107
222	333	1	8	134	124
222	333	1	9	115	116
222	333	1	10	85	132
222	333	1	11	122	107
222	333	1	12	112	100
222	333	1	13	118	126
222	333	1	14	140	102
222	333	1	15	121	103
222	333	1	16	77	125
222	333	1	17	91	90
222	333	1	18	145	104
222	333	1	19	82	98
222	333	1	20	99	116
333	444	1	1	193	222
333	444	1	2	182	204
333	444	1	3	156	150
333	444	1	4	192	215
333	444	1	5	164	217

Appendix C

Publications and presentations based on this thesis

Hollander, Y. (2006). Direct Versus Indirect Models for the Effects of Unreliability. Transportation Research, Part A: Policy and Practice, Vol. 40 (9), pp. 699-711.

Hollander, Y. (forthcoming). Fitting Distributions to Random Parameters. Transportation Research Record. Also in: Proceedings of the 85th Annual Meeting of the Transportation Research Board, Washington DC, January 2006.

Hollander, Y. (2005). The Attitudes of Bus Users to Travel Time Variability. Paper presented at the European Transport Conference, October 2005, Strasbourg, France (winner of Neil Mansfield award).

Hollander, Y. & Liu, R. (2005). Calibration of a Traffic Microsimulation Model as a Tool for Estimating the Level of Travel Time Variability. In: Advanced OR and AI Methods in Transportation, Proceedings of the 10th meeting of the EURO Working Group on Transportation, September 2005, Poznan, Poland.

Hollander, Y. (2005). Travellers' Attitudes to Travel Time Variability: Inter-modal and Intra-Modal Analysis. Paper presented at the 3rd international SIIV congress - People, Land, Environment and Transport Infrastructure - Reliability and Development, September 2005, Bari, Italy.

Liu, R. & Hollander, Y. (2005). Modelling and Computational Issues in the Estimation of Travel Time Variability Using Traffic Microsimulation. Presented at: Simulation Models - from the Labs to the Trenches - a Workshop on Traffic Modelling, September 2005, University of Arizona, USA.

Hollander, Y. & Liu, R. Identifying Key Principles in the Calibration of Traffic Microsimulation Models. Currently under review for publication at Transportation.

Hollander, Y. & Liu, R. Inter-run Variation Analysis of Traffic Microsimulation. Currently under review for Transportation Research, Part C: Emerging Technologies.

Hollander, Y. The Distribution of the Willingness to Pay for a Reliable Journey. Paper presented at the 86th Annual Meeting of the Transport Research Board, Washington DC, January 2007.

For the full list of publications of the author, see www.yarhol.com\cv.html.